Title
Spectral networks algorithms for de novo interpretation of tandem mass spectra

Permalink
https://escholarship.org/uc/item/0d09r62g

Author
Bandeira, Nuno Filipe Cabrita

Publication Date
2007

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Spectral Networks Algorithms for De Novo Interpretation of Tandem Mass Spectra

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

in

Computer Science

by

Nuno Filipe Cabrita Bandeira

Committee in charge:

Professor Pavel A. Pevzner, Chair
Professor Vineet Bafna
Professor Trey Ideker
Professor Elizabeth A. Komives
Professor Ramamohan Paturi

2007

The dissertation of Nuno Filipe Cabrita Bandeira is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2007

To my beloved wife and family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I am immensely grateful to my advisor, Pavel Pevzner, for his guidance, inspiration and generous availability of stimulating projects and opportunities. I feel privileged to have studied with him, from whom I have refined the science and art of doing research and effectively presenting it to others. I am also indebted to Professor Vineet Bafna, Haixu Tang and Dekel Tsur for their sharp insights and indispensable contributions that so helped shape my research and quality of the results. Mass spectrometry is an inherently experimental science and the quality of the results obtained from its computational analysis is inextricably bound to the quality of the data - the algorithms described in this dissertation would have been of much less value if not for the outstanding expertise and feedback that we are fortunate to receive from our distinguished collaborators. In particular, this work would not have been possible without Karl Clauser, Phillip Wilmarth, Larry David, Virgil Woods, Chris Gessner, Ebrahim Zandi, Matthias Mann and Jesper Olsen. Much of the research presented in this dissertation would have been much harder to conduct without the invaluable contributions of my fellow students Ari Frank, Stephen Tanner, Samuel Payne and the general support and companionship of my friends and colleagues in the bioinformatics labs and the department as a whole, with a very special mention for Julie Conner, David Bareno, Sheila Manalo, Samira O'Brien and Yuka Nakanishi. In addition, I wish to thank Professor Elizabeth Komives, Professor Trey Ideker and Professor Ramamohan Paturi for their knowledgeable advice and for investing their time and patience in reviewing my dissertation and serving on my committee.

Last but not by any means least, I am more indebted to my wife and family than I am able to express in words. Their unquestioning love and unwavering support for my pursuit of ambitious and faraway goals continues to be a fundamental pillar in my ability to undertake and achieve these goals. You may be sometimes far but you are never absent.

To all of you, I pledge to care for the seeds of knowledge now sown in me and which I'll feed with commensurable dedication and challenges with the hope of

coming to bear the quality of fruit you have educated me to aspire to.

Chapter 1 is, in part, a reprint of the paper “Spectral Networks: A new approach to de novo discovery of protein sequences and post-translational modifications” in BioTechniques vol.42, pp.687-95. The dissertation author was the primary investigator and author of this paper.

Chapters 3 and 5 are, in part, a reprint of the paper “Shotgun Protein Sequencing by tandem mass spectra assembly” co-authored with Haixu Tang, Vineet Bafna and Pavel Pevzner in Analytical Chemistry vol.76, pp.7721-33. The dissertation author was the primary investigator and author of this paper.

Chapter 6 is, in part, a reprint of the papers “Protein identification by spectral networks analysis” in Proceedings of the National Academy of Sciences USA, vol.104, pp.6140-5 and the paper “Protein identification by spectral networks analysis” in the proceedings of RECOMB 2006, both co-authored with Dekel Tsur, Ari Frank and Pavel Pevzner. The dissertation author was the primary investigator and author of these two papers.

Chapter 7 is, in part, a reprint of the paper “Shotgun Protein Sequencing: Assembly of peptide tandem mass spectra from mixtures of modified proteins” co-authored with Karl Clauser and Pavel Pevzner in Molecular and Cellular Proteomics vol.6, pp.1123-34. The dissertation author was the primary investigator and author of this paper.

VITA

1997	Bachelor of Science in Informatics Engineering Universidade Nova de Lisboa, Lisbon, Portugal
2001	Master of Science in Applied Artificial Intelligence Universidade Nova de Lisboa, Lisbon, Portugal
2007	Doctor of Philosophy in Computer Science University of California, San Diego, USA

PUBLICATIONS

Nuno Bandeira, Jesper V. Olsen, Matthias Mann and Pavel A. Pevzner, "Multiplexed De Novo Peptide Sequencing: applications to MS^2/MS^3 analysis," (in preparation).

Nuno Bandeira, Dumitru Brinza, Kim Hixon, Richard D. Smith and Pavel Pevzner, "High-throughput spectral networks reconstruction of a whole proteome," (in preparation).

Victoria Pham, Nuno Bandeira, Pavel A. Pevzner and Jennie R. Lill "De novo Sequencing of a Monoclonal Antibody Light/Heavy Chains Using Multiple Proteases and Shotgun Protein Sequencing," (in preparation)

Ari M. Frank, Nuno Bandeira, Zhouxin Shen, Stephen Tanner, Steven P. Briggs, Richard D. Smith, Pavel A. Pevzner, "Clustering Tandem Mass Spectra: From Spectral Libraries to Spectral Archives," (submitted)

Nuno Bandeira, "Spectral Networks: A New Approach to De Novo Discovery of Protein Sequences and Post-translational Modifications," in BioTechniques Vol. 42, No. 6: pp 687-95 (June 2007)

Nuno Bandeira, Karl R. Clauser and Pavel A. Pevzner, "Shotgunprotein sequencing: Assembly of MS/MS spectra from mixtures of modified proteins," in Mol Cell Proteomics. Vol. 6, pp 1123-34, (July 2007).

Nuno Bandeira, Dekel Tsur, Ari Frank and Pavel A. Pevzner, "Protein identification by spectral networks analysis." in Proc Natl Acad Sci USA. Vol. 104, No. 15: pp 6140-5 (April 2007).

Nuno Bandeira, Dekel Tsur, Ari Frank and Pavel A. Pevzner, "A new approach to protein identification," in Proceedings of the 10th Annual International Conference on Research in Computational Molecular Biology (RECOMB), pp 363-78 (April 2006).

Nuno Bandeira, Haixu Tang, Vineet Bafna and Pavel A. Pevzner, "Shotgun protein sequencing by tandem mass spectra assembly," in Anal Chem. Vol 76, No. 24: pp 7221-33 (Dec. 2004).

Nuno Bandeira, “Analysis of electroencephalograms (EEGs) using machine learning techniques,” Thesis submitted for the degree of Master of Science in Applied Artificial Intelligence, New University of Lisbon (May 2001).

Nuno Bandeira, Victor S. Lobo, and Fernando Moura-Pires, “Analysis of EEG of shooters,” in Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS’2000, 2000.

Nuno Bandeira, Victor S. Lobo, and Fernando Moura-Pires. “EEG/ECG data fusion using self-organising maps,” in Proceedings of EuroFusion99, International Conference on Data Fusion, 1999.

Nuno Bandeira, Victor S. Lobo, and Fernando Moura-Pires. “Training a self-organizing map distributed on a pvm network,” in Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN98, Vol.1, pp:457-461 (May 1998).

Victor S. Lobo, Nuno Bandeira, and Fernando Moura-Pires. “Distributed Kohonen networks for passive sonar based classification,” in Proceedings of the 1998 International Conference on Multisource-Multisensor Information Fusion (FUSION98), 1998.

Victor S. Lobo, Nuno Bandeira, and Fernando Moura-Pires. “Ship recognition using distributed self organizing maps,” in Proceedings of the 1998 International Conference on Engineering Applications of Neural Networks (EANN98), 1998.

FIELDS OF STUDY

Major Field: Computer Science
Studies in Bioinformatics and Computational Mass Spectrometry
Professors Pavel A. Pevzner and Vineet Bafna

ABSTRACT OF THE DISSERTATION

Spectral Networks Algorithms for De Novo Interpretation of
Tandem Mass Spectra

by

Nuno Filipe Cabrita Bandeira
Doctor of Philosophy in Computer Science
University of California, San Diego, 2007
Professor Pavel A. Pevzner, Chair

The ongoing success of the proteomics endeavor is the result of a prolific symbiosis between experimental ingenuity and efficient bioinformatics. But despite valuable contributions, the road to a better understanding of protein behavior is still hurdled by significant difficulties in the extensive identification of post-translational modifications and in the sequencing of novel proteins like cancer fusion proteins or antibody chains.

Recently, tandem mass spectrometry (MS/MS) based approaches seemed to be reaching the limit on the amount of information that could be extracted from MS/MS spectra. However, a closer look reveals that a common limiting procedure is to analyze each spectrum in isolation, even though high throughput mass spectrometry regularly generates many spectra from related peptides.

By capitalizing on this redundancy we show that, similarly to the alignment of protein sequences, unidentified MS/MS spectra can also be aligned for the identification of modified and unmodified variants of the same peptide. Moreover, this alignment procedure can be iterated for the accurate grouping of multiple peptide variants. In fact, when applied to a set of spectra from cataractous lenses proteins from a 93-year old patient, spectral networks were able to capitalize on the highly correlated peaks in spectra from variants of the same peptide to rediscover the modifications identified by database search methods and additionally discovered several novel modification events. Furthermore, the combination of shotgun proteomics with the alignment of spectra from overlapping peptides led to the development of Shotgun Protein Sequencing - similarly to the assembly of DNA reads into whole genomic sequences, we show that assembly of MS/MS spectra enables the highest ever de-novo sequencing accuracy, while recovering large portions of the target proteins sequences. Knowing that novel venom proteins have previously provided essential clues for the design of important drugs, we demonstrate our approach on a mixture of western diamondback rattlesnake venom proteins and recover over 85% of the known protein segments at over 90% sequencing accuracy while additionally sequencing several putative novel peptides and single-nucleotide polymorphism variants.

1

Introduction

Tandem mass spectrometry (MS/MS) is nowadays the technology of choice for the identification of proteins and post-translational modifications [3]. Fast-paced technological developments have delivered high-throughput analysis of thousands of proteins in a mere couple of hours at unprecedented levels of mass resolution and accuracy [143]. However, the major computational approaches to the automated identification of the millions of MS/MS spectra generated on a daily basis still interpret every single MS/MS spectrum in isolation like the original techniques for *de novo sequencing* introduced by Klaus Biemann's group in the 1960's [17] and *database searching* first proposed in the early nineties [58, 144]. In database searching, each MS/MS spectrum is compared against a given database of known peptides and significant matches are selected for protein identification. Elaborate scoring functions have been derived to provide statistical significance to observed identifications and help make this the approach of choice for the analysis of model organisms [74, 97]. However, database search is only applicable when the proteins sequences are obtained in advance through other experimental procedures such as DNA sequencing or Edman degradation. Conversely, de novo sequencing becomes the mass spectrometric approach of choice for studies of unknown proteins. Nevertheless, fully automated de novo analysis has remained an elusive goal due to difficulties in sequencing accuracy - the best algorithms for individual ion trap MS/MS spectra still predict one incorrect amino acid out of every four predictions [41]. In this dissertation, we propose to approach the MS/MS identification problem from a different perspective - first combine uninterpreted MS/MS spectra from overlapping peptides and only then determine consensus identifications (of sequences and modifications) for *sets* of aligned MS/MS spectra.

1.A Experimental setup

Most experimental protocols use enzymatic digestion to generate smaller peptides which are then analyzed by mass spectrometry to identify proteins in the sample. Trypsin digestion is often used because its strong cleavage specificity tends to be reproducible and facilitates the analysis of complex samples by generating only a few different peptides per protein. Alternatively, less specific enzymes or combinations of enzymes may be used to generate extensive protein coverage [34,89]. As illustrated in Figure 1.1, these procedures tend to generate many overlapping peptides covering the same protein regions. While the specificity of trypsin digestion leads to many spectra covering the same protein regions, non-specific digestion tends to generate spectra covering large portions of the protein sequences.

After enzymatic digestion, the substrate consists of a collection of peptides, usually containing sizeable amounts of most peptides. This substrate is then processed through a series of steps such that, in principle, each cycle of tandem mass spectrometry focuses exclusively on multiple instances of the same peptide. The same cycle is then repeated thousands of times subjected to a variety of procedures to maximize the number of spectra from different peptides [3]. After isolating many copies of a particular peptide, an MS/MS spectrum is obtained by inducing breaks at the amide bonds and thus generating peptide fragments whose masses and relative abundances are then measured by a mass analyzer [46]. Most often, the resulting peptide fragments correspond to b (prefix) or y (suffix) ions, although other types of ions may also be generated (see Figure 1.1c for an illustration). Since most amino acids have measurably different masses the ion masses observed in an MS/MS spectrum typically correlate well with the theoretical masses calculated from the peptide sequence. In addition, tandem mass spectrometry can be used to identify post-translational modifications by detecting the characteristic changes in residue mass

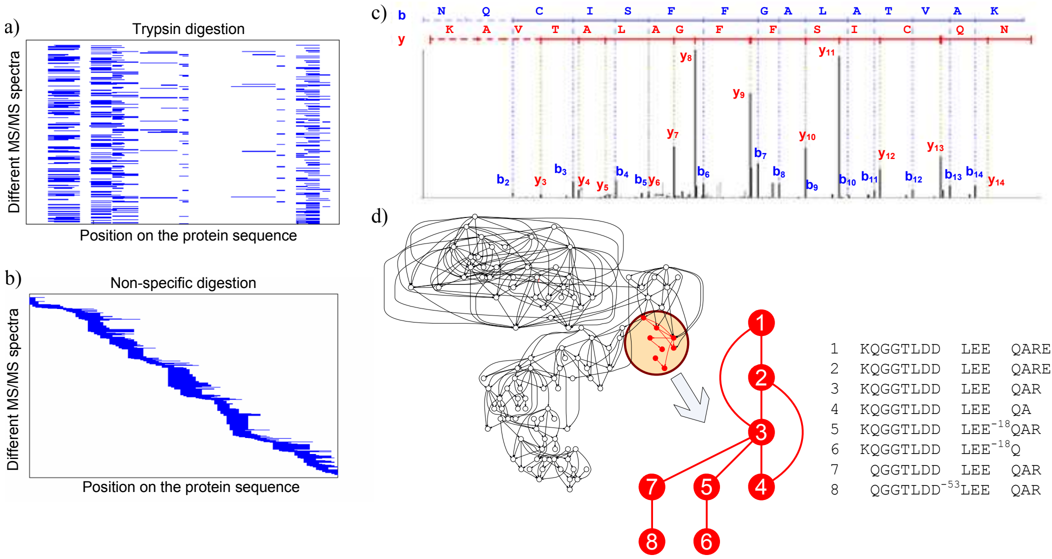


Figure 1.1 Spectral coverage of overlapping peptides resulting from enzymatic digestion of a target protein; horizontal axes represent peptide location on the protein and vertical axes separate different MS/MS spectra: **a)** Spectral coverage resulting from trypsin digestion; **b)** Spectral coverage resulting from non-specific enzymatic digestion or digestion with multiple enzymes of different specificities. **c)** MS/MS spectrum for peptide NQCISFFGALATVAK; b-ions (prefix masses) are shown in blue, y-ions (suffix masses) are shown in red. Note that the b/y peak assignments are not known in advance but can only be determined for identified spectra. **d)** Spectral network formed by a set of 117 IKKβ spectra [11]; each node corresponds to a different spectrum and nodes are connected by an edge if the corresponding spectra were paired by spectral alignment. A subcomponent of the spectral network is shown in red along with the corresponding peptides. For example, the edge between nodes 1 and 3 indicates that the spectrum for peptide 1 was significantly aligned to the spectrum from peptide 3.

due to the addition or loss of particular compounds [66]. In particular, for a modification of mass m, all b and y-ions containing the modified residue will have their mass offset by the same mass m.

1.B Spectral networks

Samples of digested proteins often contain multiple overlapping peptides, i.e. different peptides covering the same region of a protein sequence. The simplest example is the acquisition of multiple spectra from the same peptide (sometimes detected and merged using spectral clustering techniques, such as described in chapter 3 [9, 13, 125]). However, these samples also commonly contain spectra from similar but different peptides such as prefix peptides (e.g. PEPTI/PEPTIDES), suffix peptides (e.g. TIDES/PEPTIDES) or partially-overlapping peptides (e.g. PEPTIDES/TIDESHIGH). If the peptide sequences were known in advance, determining their overlap would be a straightforward application of the standard sequence alignment algorithms [119]. Conversely, spectral alignment is defined as the alignment of matching peaks between spectra from overlapping peptides [10, 105]. This concept described in detail in chapters 5-6 and is illustrated in Figure 1.2a with the matching b-ions highlighted in blue. The surprising outcome of spectral alignment is that even though one does not know the peptide sequences in advance, it turns out that the sequence information encoded in the masses of the b/y-ions suffices to detect pairs of MS/MS spectra from overlapping peptides.

In principle, the score of the spectral alignment between two given spectra could simply be defined as the maximum number of matched ions over all possible offsets of one spectrum in relation to the other. While this would work to a limited extent, we have found that taking into account ion intensities and correlated occurrences of multiple ion types leads to a much more accurate separation between true spectral pairs (spectra from overlapping peptides) and false spectral pairs (spurious matches between spectra from unrelated peptides). In fact, it turns out that the reliability of spectral alignment allows one to discern the high-scoring true spectral pairs from the many millions of possible spectral pairs in high-throughput proteomics experiments [10, 11]. Moreover, since each spectrum may align to several other spectra, the set of detected spectral pairs defines a *spectral network* where each node corresponds to a different spectrum and nodes are connected by an edge if the corresponding spectra were found to to be significantly aligned. This concept is introduced in chapter 6 and illustrated in Figure 1.1d with a particular network found on a set of IKKβ spectra. Note that since most spectra usually come from non-contiguous protein regions, the consequent outcome of this approach is not a single spectral network but rather multiple spectral networks, one for each set of spectra from overlapping peptides.

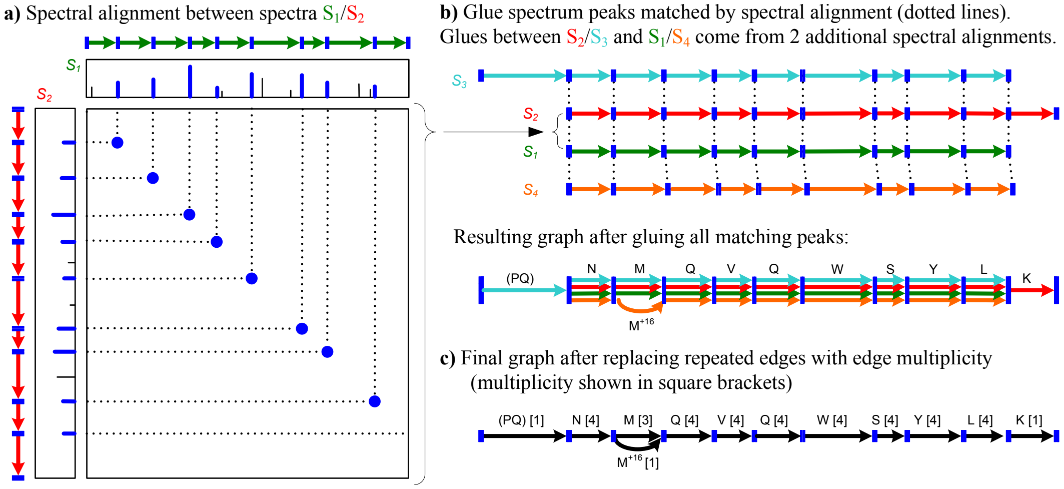
1.C Shotgun Protein Sequencing

The pattern of overlapping peptides illustrated in Figure 1.1b leads to particularly interesting possibilities for computational analysis - as in the assembly of genomic sequences from DNA reads, it now becomes feasible to assemble MS/MS spectra into protein sequences [8, 9] (described in detail in chapters 5 and 7).

The assembly of spectra from overlapping peptides can be likened to a simple allegory - imagine you have a jewelry box containing many copies of a particular model of bead necklaces. In this allegory, all necklaces are made from the same type of bead and thread but different necklace models are characterized by designer-specified varying thread distances between consecutive beads. Thus, any given necklace model is completely defined by a sequence of consecutive inter-bead distances. But what if, after collecting many copies of your favorite necklace model, you one day find that someone cut each necklace multiple times at randomly chosen bead positions? In this context, the necklace-model recovery problem is that of rediscovering the original necklace model given only the leftover pieces in the jewelry box. Although mass spectrometry adds a fair amount of complexity to this problem, this allegory captures the essence of the spectral assembly problem where amino acid masses correspond to inter-bead distances and beads represent the amide bonds between consecutive amino acids.

The Shotgun Protein Sequencing approach to de novo sequencing is a three-stage approach to the assembly of MS/MS spectra into amino acid sequences: *a*) find pairs of spectra from overlapping peptides using spectral alignment, *b*) assemble the aligned spectra and *c*) determine a consensus amino acid sequence for each set of assembled spectra. As illustrated in Figure 1.2, this approach is not unlike *a*) finding necklace pieces with matching inter-bead distances, *b*) gluing the matching beads and *c*) determining the necklace model from the recovered distances between glued beads.

By capitalizing on the correlated ion occurrences in all assembled spectra, shotgun protein sequencing leads to significant improvements in de novo sequencing accuracy and, on average, only makes one mistake out of every 10 amino acid predic-


a) Spectral alignment between spectra S1/S2
S1
S2
b) Glue spectrum peaks matched by spectral alignment (dotted lines).
Glues between S2/S3 and S1/S4 come from 2 additional spectral alignments.
S3
S2
S1
S4
Resulting graph after gluing all matching peaks:
(PQ)
N
M
Q
V
Q
W
S
Y
L
K
M+16
c) Final graph after replacing repeated edges with edge multiplicity
(multiplicity shown in square brackets)
(PQ) [1]
N [4]
M [3]
Q [4]
V [4]
Q [4]
W [4]
S [4]
Y [4]
L [4]
K [1]
M+16 [1]

d) Sequenced portions of the target protein sequence:

```
MSWSPSLTTQTCGAWEMKERLGTGGFGNVIRWHNQETGEQIAIKQCRQELSPRNRERWCLEIQIMRRLTHPNVVAARDVPEGMQNLAPNDLPLLAM
EYCQGGDLRKYLNQFENCCGLREGAILTLLSDIASALRYLHENRIIHRDLKPENIVLQQGEQRLIHKIIDLGYAKELDQGSLCTSFVGTLQYLAPE
LLEQQKYTVTVDYWSFGTLAFECITGFRPFLPNWQPVQWHSKVRQKSEVDIVVSEDLNGTVKFSSSLPYPNNLNSVLAERLEKWLQLMLMWHPRQR
GTDPTYGPNGCFKALDDILNLKLVHILNMVTGTIHTYPVTEDESLQSLKARIQQDTGIPEEDQELLQEAGLALIPDKPATQCISDGKLNEGHTLDM
DLVFLFDNSKITYETQISPRPQPESVSCILQEPKRNLAFFQLRKVWGQVWHSIQTLKEDCNRLQQGQRAAMMNLLRNNSCLSKMKNSMASMSQQLK
AKLDFFKTSIQIDLEKYSEQTEFGITSDKLLLAWREMEQAVELCGRENEVKLLVERMMALQTDIVDLQRSPMGRKQGGTLDDLEEQARELYRRLRE
KPRDQRTEGDSQEMVRLLLQAIQSFEKKVRVIYTQLSKTVVCKQKALELLPKVEEVVSLMNEDEKTVVRLQEKRQKELWNLLKIACSKVRGPVSGS
PDSMNASRLSQPGQLMSQPSTASNSLPEPAKKSEELVAEAHNLCTLLENAIQDTVREQDQSFTALDWSWLQTEEEEHSCLEQAS
```

Figure 1.2 Shotgun Protein Sequencing via assembly of tandem mass spectra; **a)** Spectral alignment between spectrum S_1 (from peptide NMQVQWSYL) and spectrum S_2 (from peptide NMQVQWSYLK) reveals the common sequence information in both spectra. Next to each spectrum is a graph representation of the corresponding peptide sequence with consecutive b-ions represented as nodes connected by arrow edges. **b)** Matching peaks in spectral alignments become pairwise gluing instructions between every pair of aligned spectra. Additional spectra S_3 (from PQNMQVQWSYL) and S_4 (from NM^{+16}QVQWSYL) respectively illustrate assembly of additional types of spectral alignment: partially overlapping peptides and modified/unmodified variants of the same peptide; **c)** Repeated edges are replaced by single edges with weight proportional to their multiplicity and the consensus sequence for all assembled spectra is found by the heaviest path in this graph; **d)** Recovered portions of a target protein in the sample. Correct amino acid predictions are shown in green (93%) and incorrect in orange (7%).

tions, even on low-accuracy ion trap MS/MS spectra. Using this approach, we were able to resequence large portions of multiple proteins in pure venom extract from western diamondback rattlesnake [8]. In addition, compelling evidence was found for novel crotalus atrox peptides featuring strong homology to venom peptides from other species.

1.D Spectral networks from spectra of modified peptides

In traditional DNA sequence alignment, it often happens that query sequences differ from the reference sequences by the insertion or deletion of one or more nucleotides [119]. While the insertion/deletion of amino acids is also usually allowed when aligning protein sequences, an additional factor needs to be considered when aligning peptides from experimental samples - the occurrence of post-translational modifications. From a sequence alignment perspective, a modification could be modeled by following the modified residue with a special character for each type of modification. Thus, the alignment of a modified peptide PEPT*IDE with its unmodified counterpart PEPTIDE would result in a single difference caused by the insertion of the modification '*'.

Although MS/MS spectra represent peptides as a sequence of peaks, computing the spectral alignment between spectra from modified and unmodified variants of the same peptide is substantially similar to the sequence alignment problem. This correspondence can be illustrated by representing each spectrum as sequence of 1/0 symbols respectively corresponding to 'peak'/'no-peak' events at each mass value. Thus, for any integer mass m, let $s(m)$ be a sequence of $m-1$ zeros followed by a single one. For example, if an imaginary peptide of mass 12 was composed by amino acids XYZ (with masses 3,4,5, respectively) then its theoretical spectrum would contain peaks at masses 3,7,12 and the corresponding 0/1-sequence representation would be $s(3)s(4)s(5) = 001000100001$. In this framework, any sequence of masses (such as a peptide or a modified peptide) can be expressed as a sequence of 0/1 symbols and pairs of sequences can then be aligned using standard sequence alignment algorithms [119]. As such, a modification of mass m' corresponds to the insertion of m' additional zeros right before the sequence for the modified residue (i.e. the mass of the residue becomes larger). Conversely, if the modification causes a loss of m'' Daltons (mass units) from the modified residue then the corresponding effect is the deletion of m'' zeros from the sequence for the modified residue. Although the spectral alignment algorithms described in chapters 5-7 do not explicitly convert spectra to sequences of zeros and ones, this model illustrates the essential concepts behind the approach. Figure 1.3a illustrates the spectral alignment between MS/MS spectra from the peptides TETMA and TET^{+80}MA.

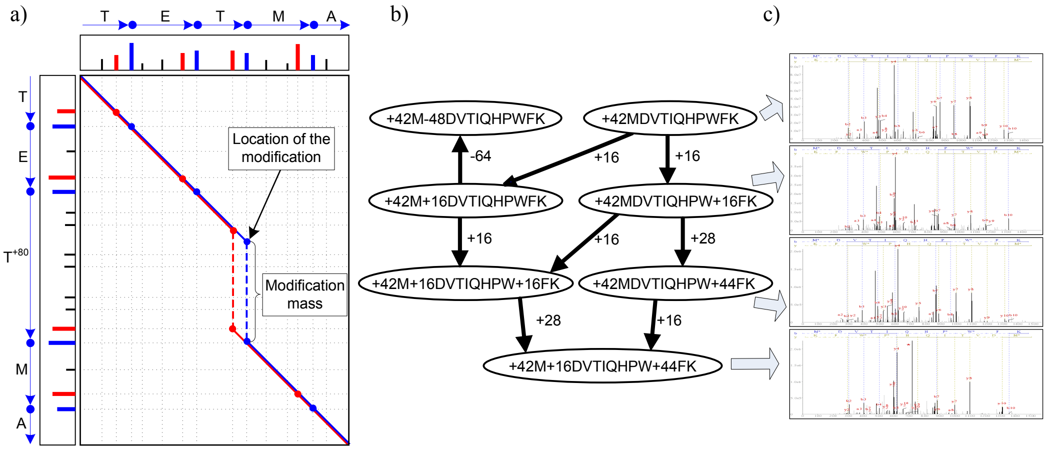


Figure 1.3 Identification of post-translational modifications through spectral networks; **a)** Spectral alignment between modified and unmodified variants of the peptide TETMA (b-ions shown in blue, y-ions in red, blue/red lines track consecutively matched b/y-ions); **b)** Grouped modification states of the peptide MDVTIQHPWFK from a sample of cataractous lenses; **c)** Highly correlated MS/MS spectra from the indicated peptide variants.

When first analyzing a sample possibly containing modified peptides one does not know a priori which residues or peptides will be modified. Thus, spectral alignment considers every possible spectral pair and every possible location for the mass difference (e.g. modification mass) between the aligned spectra. By requiring a significant match between the aligned spectrum peaks [11] but placing no restrictions on which modifications to consider, this approach can be used to discover novel or unexpected modifications. In fact, when applied to a set of spectra from cataractous lenses proteins from a 93-year old patient, spectral networks were able to rediscover the modifications identified by database search methods and additionally discovered several novel modification events [11, 133].

The identification of peptides containing multiple modifications via database search is a challenging problem because of the combinatorial explosion in the number of possible modification variants for all the peptides in a database [133]. Not only can the large number of possible peptide variants make this approach much slower, but the increased number of peptide candidates for any given spectrum significantly increases the risk of incorrect identifications. However, samples containing peptides with two or more modifications often also contain variants of the same peptide with only one or no modification. In these cases, we have found that spectral alignment is able to group these related spectra from multiple modification variants of the same peptide into small spectral networks. Figure 1.3b illustrates the spectral network for a particular peptide in a sample of cataractous lenses proteins.

By grouping together spectra from multiple variants of the same peptide, spectral networks additionally contribute to the reliable identification of highly modified peptides. While database searching is restricted to matching ion masses between theoretical and observed spectra, spectral networks further capitalize on the correlated co-occurrences of ions at corresponding masses and with similar peak intensities (Figure 1.3c). In general terms, it becomes easier to identify a highly modified peptide if one additionally observes highly-similar spectra from the intermediate modification states. Thus, spectral alignment not only allows one to *discover* unexpected modifications (instead of only identifying expected modifications) but additionally defines an alternative way to reliably identify highly modified peptides.

1.E Discussion

Spectra from overlapping peptides or modification-variants of the same peptide deliver a wealth of correlated sequence information that can be explored with a new generation of algorithms based on spectral networks. In a departure from standard procedures, having spectra from modified/unmodified variants of the same peptide allows one to directly discover the modifications in the sample rather than having to guess in advance the list of modifications to search for. Spectra from multiple modification-variants can be combined into spectral networks and correlated ion masses and intensities used to increase the confidence in the identification of highly modified peptides. From a protein sequencing perspective, the extensive sequence coverage achievable with non-specific proteolytic digestion enables the assembly of spectra from overlapping peptides into long *protein contigs*. Moreover, by capitalizing on the correlated sequence information in sets of assembled spectra, the Shotgun Protein Sequencing approach is able to deliver the highest sequencing accuracy ever reported on ion trap MS/MS spectra.

Chapter 1 is, in part, a reprint of the paper “Spectral Networks: A new approach to de novo discovery of protein sequences and post-translational modifications” in BioTechniques vol.42, pp.687-95. The dissertation author was the primary investigator and author of this paper.

2

The *De Novo* Sequencing Problem

Although proteins star as fundamental workhorses of cell biology, the accurate automated determination of their sequences is still an open problem and an active area of research. The most promising technology developed for this purpose is mass spectrometry, which allows for the simultaneous high-throughput analysis of hundreds of proteins sequences. Several generations of computational tools have been developed to interpret the data generated by these instruments and some of them are now commonplace on every biologist's computational toolbox. This chapter focuses on the *de-novo* family of computational tools - used to recover protein sequences directly from mass spectrometry data.

2.A Introduction

In recent years, computer science has radically transformed the study of Biology - landmark achievements like the human genome sequencing project [78,135] would have been unthinkable without the development of new algorithms and tools to efficiently process the immense flood of data. As engineers and biologists constantly strive to develop new ways to inspect biological events it is up to the computer scientist to develop and implement efficient ways to process these new and constantly reinvented types of data.

The typical pattern of algorithm development for the analysis of biological data is one of progression from exponential time algorithms for proof of concept, through several heuristic attempts at containing the runtime, until finally a well designed solution is found that provides optimal results (according to some problem formulation) within a reasonable time frame. By then, of course, the instruments have already evolved and new variants of the same problems emerge with promising new applications.

One of the most dynamic areas of study nowadays is that of Proteomics [3, 37], a field of study that is concerned with the discovery of which proteins act on an organism and in what ways. Understanding these fundamental building blocks is of critical importance and pervasive in the design of many new drugs and therapies for serious health conditions [71,76,93,122,123].

Possibly one of the most basic questions one can ask about a protein is what is its amino acid sequence. In its simplest form proteins can be described as chains of chemical groups called amino acids, each with a distinct chemical structure and most with different molecular masses. Thus, for a computer scientist, a protein sequence can simply be thought of as a string over an alphabet of 20 elements (amino acid letter codes) with each element having a specific known mass. The mass of an amino acid is a particularly important characteristic from an experimental perspective because it is something that can be directly measured using mass spectrometry instruments. Conceptually, these instruments measure the mass of small charged molecules by propelling them over a known distance using a constant electrical field and then converting the time-of-flight back into molecular mass, based on the principle that heavier molecules travel slower than lighter molecules.

The current mainstream application of tandem mass spectrometry is *protein identification* - the protein sequences are previously known and the purpose is to identify which proteins are present in some sample of interest (e.g. diseased tissue vs healthy tissue). Protein identification is an active area of research on its own [69,84] and provides the tools of choice when the sequences of all the relevant proteins are known in advance. But not all protein sequences are known in advance and it is not uncommon for protein identification studies to identify only 15-25% of all the mass spectrometry data generated. Although much of the remaining data is consensually not identifiable, assumptions made by the protein identification software often rule out the correct interpretations.

The complementary way of interpreting mass spectrometry data is *protein sequencing*, where an amino acid sequence is derived directly from the experimental data. If successful, this approach could significantly speed up the rate of discovery of new proteins of potential vital importance - in more than one case, other time and labor intensive methods have been used to directly sequence previously unknown proteins which later led to the development of important drugs for the treatment of serious human conditions such as cancer [93, 122] and blood clotting problems [71, 76, 123].

This dissertation focuses on the computational issues of of the latter - determining protein sequences from mass spectrometry data. Section 2.B formally describes the experimental input data to an adequate level of detail and sets the tone for the description of the multiple attempts made at its automatic interpretation, as presented in section 2.C.

2.B Mass spectrometry data

From a computer scientist perspective, a protein sequence can be thought of as a string over a weighted alphabet of 20 amino acids $\mathcal{A}$, with the mass of each amino acid $a \in \mathcal{A}$ given by $m(a)$ and the set of all amino acid masses denoted by $m(\mathcal{A})$. Substrings of protein sequences are usually referred to as *peptides* and *parent mass* of a peptide $\rho = a_1, \ldots, a_n$ is defined as $m(\rho) = \sum_{i=1}^{n} m(a_i)$. Additionally, the i-th *prefix (suffix) mass* of a peptide, referred to as b_i (y_i), is simply the summed mass of its prefix (suffix) string with i amino acids.

Mass spectrometry instruments measure $\frac{mass}{charge}$ ratios of ionized molecules, or simply measure mass if we make the simplifying assumption that all fragments have charge one[1]. Conceptually, when applied to the analysis of peptides, these instruments proceed through the following three stages:

[1] We remark that the term precursor mass is commonly used to denote the term $\frac{M+18+Z}{Z}$, where M is a peptide's parent mass and Z its parent charge

1. The first MS stage snapshots the parent masses of the peptides passing through the instrument (MS).

2. A parent mass is selected and the many copies of (usually) the same peptide are dissociated into fragments by a collision-induced random process. Peptides tend to break only once and between consecutive amino acids, often generating complementary pairs of detectable fragments: one corresponding to a prefix mass and another corresponding to a suffix mass.

3. The second MS stage determines the masses of the peptide fragments (MS/MS or MS^2).

4. Optionally, steps 2 and 3 may be repeated to generate additional spectra from fragments is the MS^2 spectrum (MS/MS/MS or MS^3).

Because many copies of the same peptide are initially present in the sample the same masses are detected several times with different masses having different relative abundances. As such, a tandem mass spectrum or MS/MS spectrum $S = \{s_1, ..., s_m\}$ of an unspecified peptide ρ with parent mass $m(S) = m(\rho)$ is a list of masses $m(s_i)$, each with relative intensity $I(s_i)$ proportional to the relative abundance of the corresponding fragment mass. Figure 2.1 shows a hypothetical MS/MS spectrum for the peptide CSE.

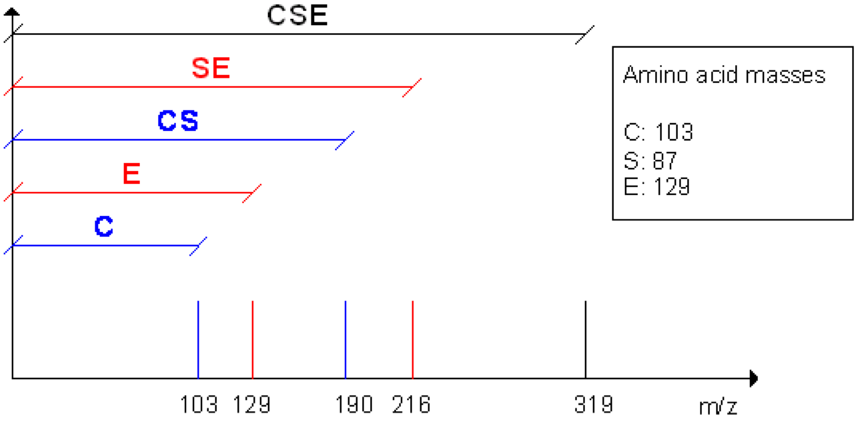


Figure 2.1 Hypothetical MS/MS spectrum for peptide CSE. Prefix masses are shown in blue and suffix masses are shown in red.

In reality, due to physical and experimental constraints, the observed MS/MS spectra tend to convey a much poorer representation of the peptides which originate them - some of the fragment mass peaks are not observed and unexplainable peaks are included in the spectrum. Figure 2.2 illustrates how a reasonable experimental MS/MS spectrum looks like; as introduced above, the i-th prefix mass is denoted by b_i and the i-th suffix mass by y_i.

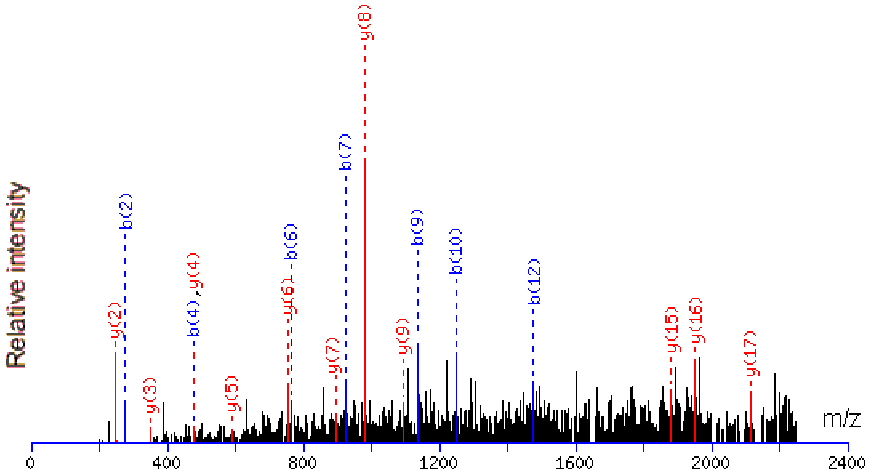


Figure 2.2 Experimental MS/MS spectrum for peptide LYAEERYPILPEYLQCVK. The i-th prefix mass is denoted by b_i and the i-th suffix mass by y_i.

Unfortunately the correspondence between amino acids and amino acid masses is not unique - two pairs of amino acids have indistinguishable (I/L) or nearly indistinguishable (Q/K) masses. On data from most intruments these masses are generally accepted to be the same and the amino acid alphabet is reduced to only 18 different symbols and masses.

2.C ***de-novo*** **interpretation**

Using the definition of tandem mass spectra given above, the *de-novo* interpretation problem can be stated as follows: given an MS/MS spectrum find a peptide with the same parent mass that explains the highest number of observed peaks. A common alternative objective function is one where the target peptide explains the most peak intensity in the MS/MS spectrum.

Formally, given a spectrum $S = \{s_1, \ldots, s_n\}$ and a peptide ρ, the set of *explained peaks* is defined as $EP(S,\rho) = \{s_i \in S : m(s_i) \textit{ is a prefix mass of } \rho \vee m(s_i) \textit{ is a suffix mass of } \rho\}$. Additionally, the *explained intensity* is defined as $EI(S,\rho) = \sum_{s_i \in EP(S,\rho)} I(s_i)$. The MS2 de-novo interpretation problem can now be stated as follows:

MS2 de-novo interpretation problem.

Input:	A spectrum S
Output:	A peptide ρ with $m(\rho) = m(S)$
Objective:	Maximal number of explained peaks $EP(S,\rho)$

2.C.1 Initial attempts

Computer-based *de-novo* interpretations of MS/MS spectra can be traced as far back as 1966 [17], when Biemann et al. first described a search procedure over the space of all peptides with parent mass smaller or equal to that of the spectrum. In essence, their approach corresponds to a depth-first search over all possible prefix sequences stopping *i*) if the parent mass of the candidate peptide sequence matches (report) or exceeds that of the spectrum (reject/backtrack) or *ii*) if a prefix mass is missing from the spectrum (reject/backtrack). This approach was very reasonable for the time and the reported results were very encouraging, but carried the requirement that the quality of the mass spectrometry data had to be very high - no missing prefix masses were tolerated.

Almost 20 years and some hardware generations later, Sakurai et al. [111] introduced PAAS3 - an attempt to compensate for missing peaks by matching each spectrum against every permutation of every set of amino acids whose summed masses equaled the parent mass of the spectrum. The final results were naturally better than those obtained using Biemann et al.'s approach (i.e. higher sensitivity and better accuracy) but at the expense of hugely increased running times and, in practice, of inapplicability to spectra with medium to large parent masses. But in addition, Sakurai et al. made the important observation that when the correct b_i prefix masses are not in the spectrum it may still be the case that the corresponding

y_i suffix masses are present in the spectrum and either case should count as positive evidence for the correct peptide. The exact design of the best scoring function to matche a peptide (with a specific set of prefix/suffix masses) to an experimental MS/MS spectrum is still nowadays a topic of active research [1, 26, 41, 88] and will not be addressed here.

Trying to capitalize on the best of both worlds, several groups [64, 68] extended the original Biemann et al.'s approach to simultaneously account for both b/y masses and tolerate missing masses in the spectrum. The former made its way into the algorithm by changing the way a candidate peptide prefix is scored: for a candidate peptide prefix ρ with mass $m(\rho)$ and a spectrum S with parent mass $m(S)$, search S for peaks with mass $m(\rho)$ or $m(S) - m(\rho)$ and use every found peak to increase the score for ρ. Tolerance for missing spectrum peaks was introduced by allowing candidate prefix sequences to be extended even if a limited number of peaks was not observed in the spectrum. But the increased tolerance for missing peaks combined with the proposed breadth-first search of the space of all possible peptide prefixes brought about the problem of having to keep too many candidate sequences in memory throughout the execution of the search. Both groups of authors [64, 68] addressed this problem by limiting the number of candidate prefix sequences that would be kept in memory at any point in time to a constant number of top scoring sequences (ranging from 100-300). Although theoretically suboptimal, this approach was reasonable enough and both groups reported improvements in accuracy and runtime efficiency. Additional variants of these [118, 147] make the additional assumption that something is known about the amino acid composition of the target peptide and restrict the search space to only those peptides with the specific amino acid compositions.

2.C.2 Spectrum graphs

In January 1990, Christian Bartels [12] detached himself from this procedural way of thinking about the MS2 de-novo interpretation problem and proposed a very different perspective: the concept of *spectrum graphs*. Informally, a spectrum graph for a given spectrum S has a vertex for each peak in S and two vertices are connected by an edge if the masses of the corresponding spectrum peaks differ by the mass of one amino acid. Figure 2.3 illustrates how the spectrum graph would look like for a perfect spectrum of the peptide CSE. The spectrum graphs introduced by the author were a little more elaborate in that a spectrum peak would actually correspond to more that one vertex in the spectrum graph. One important case worth mentioning in the context of this dissertation was that the spectrum graph contained two possible interpretations for every spectrum peak - as a prefix mass and as a suffix mass. Thus, when given an experimental spectrum $S = \{s_1, \ldots, s_n\}$ let the *reversal* $rev(S)$ be defined as $rev(S) = \{s'_1, \ldots, s'_n\}$ where $m(s'_i) = m(S) - m(s_i)$ and $I(s'_i) = I(s_i)$. A spectrum graph $G = (V, E)$ for a spectrum $S = \{s_1, \ldots, s_n\}$, with $S' = S \cup rev(S) = \{s'_1, \ldots, s'_m\}$ is then defined by $V = \{v_1, \ldots, v_m\}$ and a set of directed edges $E = \{(v_i, v_j) : m(s'_j) - m(s'_i) \in m(\mathcal{A})\}$.

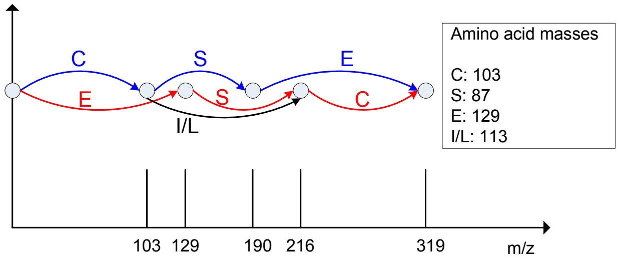


Figure 2.3 Spectrum graph for the peptide CSE; the path corresponding to the sequence of prefix (suffix) masses is shown in blue (red). Also note the spurious edge (shown in black) between the vertices corresponding to masses 103 and 216.

The transformation of a spectrum into a spectrum graph converts the MS2 de-novo interpretation problem into that of finding a highest scoring path in a directed acyclic graph (DAG) - an easy problem to solve. Nevertheless, the results provided were not impressive: the correct peptide corresponded to a maximal scoring path in the spectrum graph on only 6 out of 23 cases, possibly due to a bad way of scoring matches between vertices and spectrum peaks matches. Moreover, the specification of the algorithm used to traverse the graph was not sufficiently detailed as there was no mention on how to attribute a single score to a vertex when reached by two (or more) different paths of different scores. This gap in the algorithm was clearly filled 5 years later by Fernandez-de-Cossio et al. [38–40] by porting traditional dynamic programming graph traversal algorithms to find the highest scoring path in a DAG - the score of each vertex is the score of the corresponding spectrum peak plus the maximum score among all paths ending on it. Equation 2.1 formally defines this dynamic programming recursion with $sc(i)$ representing the score of a single vertex v_i and $Sc[i]$ representing the combined score of the best path up to and including v_i.

$$Sc[i] = sc(i) + \max_{j:m(s_i)-m(s_j)\in\ m(\mathcal{A})} Sc[j] \tag{2.1}$$

The same dynamic programming approach was also reused later by several other groups with the major differences involving the way to score a match between a spectrum peak and a spectrum graph vertex [70, 131, 132]. For a period of time these were the tools used to find a good peptide sequence to explain a spectrum and, depending on the implementation, within a reasonable amount of time. The accuracy of the reconstructions was excellent in some carefully selected cases but generally insufficient for the accurate reconstruction of the complete peptides.

2.C.3 Forbidden pairs

One of the main reasons for the persistent low accuracy was a specific computational issue that was still unaddressed by this dynamic programming approach and was already hinted at in Figure 2.3 - the spectrum graph contains spurious edges between vertices corresponding to prefix and suffix masses. This problem is exacerbated by the scoring complementarity of prefix and suffix masses, which can best be understood through an example. Consider the case of a vertex v_i for a spectrum peak $s_i \in S$, with $m(s_i) = 100$ and $m(S) = 300$. Then, the score $sc(i)$ should take into consideration two peaks in S: one with mass 100 (s_i) and another at the corresponding suffix mass $300 - 100 = 200$ (say s_j, $m(s_j) = 200$). But then, when scoring v_j (as created by the spectrum peak s_j) the same two peaks in the spectrum will be used to determine the score $sc(j)$! These pairs of vertices with symmetric spectrum peak masses are called *complementary* peaks. This symmetry implies that both vertices will tend to have high scores and actually promotes the usage of the spurious edges that we wanted to avoid in the first place. Figure 2.4 illustrates this problem on a hypothetical spectrum for the peptide EDTES; arrows under the spectrum indicate the pairs of prefix (arrow start) and suffix (arrow end) masses.

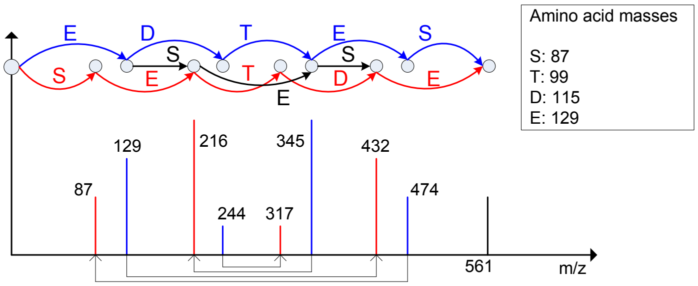


Figure 2.4 Spectrum graph for the peptide EDTES; the path corresponding to the sequence of prefix (suffix) masses is shown in blue (red); arrows under the spectrum indicate the pairs of prefix (arrow start) and suffix (arrow end) masses. Note how the spurious edges (shown in black) between the center vertices promote the greedy reusage of complementary peaks from the same pairs (masses 129/432 and 216/345), making ESESE the highest scoring path in the spectrum graph.

This problem was first addressed by Dančik et al. [26] in a breakthrough that brought the *de-novo* interpretation of MS/MS spectra to a whole new level. According to the authors, two vertices in the spectrum graph constitute a *forbidden pair* if generated from complementary peaks in the corresponding spectrum. In addition, a path in the spectrum graph is said to be *antisymmetric* if it uses at most one vertex from each forbidden pair. Our problem statement then simply becomes that of finding a highest scoring antisymmetric path in a spectrum graph - a problem that is NP-hard in the general case. But an additional crucial insight revealed that the oder of forbidden pairs in spectrum graphs can be exploited to design a polynomial time algorithm for these instances. As shown in Figure 2.4, there are no intersections between the arrows connecting complementary peaks (arrows shown under the m/z-axis). Unfortunately, after describing this special structure of spectrum graphs and arguing that it enables a polynomial time algorithm for these instances, Dančik et al. did *not* provide an algorithm. Two years later, Chen et al. [21] seized this opportunity and published the first dynamic programming algorithm to find a highest scoring antisymmetric path in a spectrum graph. To explore the structure described above for a spectrum S, let $m_{comp}(s_i) = m(S) - m(s_i)$ be the mass of the complement of s_i. Then Eq. 2.2 describes the dynamic programming recursion that finds a highest scoring antisymmetric path in a spectrum graph $G = (\{v_1, \ldots, v_n\}, E)$. The recursive relation $Sc[i, j]$ represents the score of the two best antisymmetric paths in G from v_1 to v_i and from v_j to v_n including v_i and v_j; $sc(i)$ represents the single score of v_i.

$$Sc[i,j] = \begin{cases} -\infty & \text{if } m(s_i) = m_{comp}(s_j) \\ & \textit{(forbidden pair)} \\ \\ sc(j) + \max_k Sc[i,k] & \text{if } m(s_j) < m_{comp}(s_i) \text{ and} \forall_k m(s_k) - m(s_j) \in m(\mathcal{A}) \\ & \textit{(suffix extension)} \\ \\ sc(i) + \max_k Sc[k,j] & \text{if } m(s_i) > m_{comp}(s_j) \text{ and} \forall_k m(s_i) - m(s_k) \in m(\mathcal{A}) \\ & \textit{(prefix extension)} \end{cases} \tag{2.2}$$

After initializing the 2-dimensional matrix Sc to all zeros and executing the recursion relation in Eq. 2.2 we only need to find a maximal entry $Sc[i,j]$ such that $m(s_j) - m(s_i) \in m(\mathcal{A})$ (the connecting edge over the center of the spectrum) and reconstruct a best-scoring antisymmetric path through the graph by tracing back the steps taken to find the score in $Sc[i,j]$. Eq. 2.2 also shows that a best antisymmetric path can be found efficiently, with a worst case running time of $O(|S|^2 \times |A|)$ or simply $O(|S|^2)$ if the standard set of 18 amino acid masses is always used.

Shortly after this algorithm was published, Ma et al [88] complemented it with a more accurate scoring function and Bafna and Edwards [7] provided an extension for an important generalization of forbidden pairs. Recent developments include a bayesian networks- and an HMM-based scoring schemes [1, 41] and an integer programming formulation of the peptide sequencing problem [29].

3

Preprocessing

Although mass spectrometers are very sophisticated instruments capable of measuring minuscule amounts of mass, some amount of noise and uncertainty is unavoidable. In particular, it is not unusual for a mass spectrum to contain far more peaks than can readily be explained by the corresponding peptide species. This chapter describes some of the common issues and the corresponding preprocessing steps necessary to more accurately differentiate signal from noise and enable the proposed algorithmic approaches to peptide sequencing.

3.A Experimental tandem mass (MS/MS) spectra

In terms of numbers of observed masses, MS/MS spectra tend to be extremely noisy sets of measurements. While later generations of instruments have addressed this issue to some extent, it is still the case that most peaks in an average MS/MS spectrum cannot be explained by fragments of the associated ion species. Although chemical and electrical 'noise' are usually advanced as possible explanations, exact models of noise-peak generation are not readily available. As such, the probability of observing a noise peak in a spectrum is most commonly represented using a set of uniform distributions; each for a different mass range [26]. Of course, more elaborate models can be derived to additionally account for peak intensities [1]. As described in the previous chapter, peptides are usually dissociated into prefix/suffix fragments that tend to result[1] in complementary masses in the spectrum, i.e. pairs of masses that add up to the peptide mass. Since this complementarity is unlikely to occur by chance in the uniform model of noise-peak generation, this feature is usually taken as a strong indicator to distinguish true peaks from noise peaks.

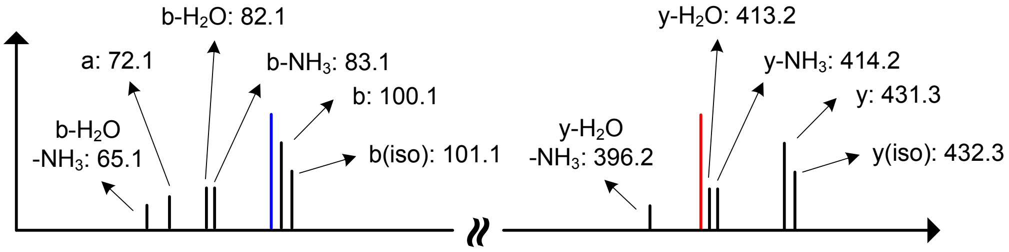


Figure 3.1 Peptide fragment ion types in ion trap mass spectrometers. The indicated mass offsets correspond to some of the possible ion types for the prefix fragment 'V' (shown in blue) and the suffix fragment 'ALID' (shown in red) from the peptide 'VALID'. Note that some of these mass offsets may not be observable for the indicated fragments 'V'/'ALID' - the diagram serves the sole purpose of illustrating each ion type's mass offset in relation to the corresponding prefix/suffix mass. Ion types are the result of the loss of chemical groups either from the amino acid residues (e.g. loss of H_2O) or from the peptide backbone (e.g. a-ion from loss of CO)

Other features commonly used to distinguish true peaks from noise are derived from the compositional properties of peptides: isotopic peaks and neutral losses. Since one of the common atoms in peptides is carbon, with isotopic integer weights 12 Da (propensity $\approx 99\%$) and 13 Da (propensity $\approx 1\%$), it is not uncommon for some peptide fragments to include at least one ^{13}C isotope and thus register at 1 Da heavier than the monoisotopic (all ^{12}C) species. Other common ion types are caused by neutral losses, so called because the loss does not affect the charge of the fragment, such as the loss of H_2O from one of the amino acid residues or the loss of CO from the peptide backbone. Table 3.1 lists the commonly observed ion types and corresponding propensities for a particular ion trap mass spectrometer. The mass offsets induced by some ion types are illustrated in Figure 3.1.

[1] Since peptide fragments must be charged in order to be observable in the mass spectrometer, the probability of observing a particular fragment is directly dependent on its propensity to ionize.

Table 3.1 Peptide fragment ion types in ion trap mass spectrometers. Isotopic ions are denoted with the prefix '(iso)', neutral losses are indicated with a 'minus' sign (-) and doubly-charged fragments are followed by a superscript charge (e.g. b_3^2 for a double-charged b_3 fragment).

Ion type	Estimated percentage of occurrences
y	51.5%
b	50.2%
$y(iso)$	46.5%
$b(iso)$	44.5%
$b - H_20$	30.0%
$b - NH_3$	29.8%
$y - NH_3$	22.0%
$y - H_20$	21.0%
y^2	17.9%
a	17.4%
$b - H_20 - NH_3$	16.3%
b^2	15.3%
$a - NH_3$	15.0%
$b - H_20 - H_20$	14.5%
$b^2 - H_20$	14.0%
$y^2 - H_20$	12.8%
$y - H_20 - NH_3$	12.2%
$a - H_20$	11.3%
$y - H_20 - H_20$	10.1%

3.B Prefix Residue Mass (PRM) spectra

A peptide can be defined as a string $\rho = a_1, ..., a_n$ where a_i is any amino acid with a known residue mass $m(a_i)$. Also, any prefix $\rho_i = a_1, ...a_i$ has a *prefix residue mass (PRM)* $m(\rho_i) = \sum_{j=1}^{i} m(a_j)$; the special case $m(\rho) = m(\rho_n)$ is also referred to as *parent mass*. As such, an equivalent representation of a peptide $\rho = a_1, ..., a_n$ is given by the mass series

$$\mathcal{R} = \{m(\rho_1), ..., m(\rho_n)\}$$

Another equivalent representation is given by the reverse mass series $\mathcal{R}^{REV}$ – the masses of all suffixes of ρ

$$\mathcal{R}^{REV} = \{m(\rho) - m(\rho_{n-1}), ..., m(\rho) - m(\rho_1), m(\rho)\}$$

In general, given a set of masses $X = \{x_1, ..., x_m\}$ with associated parent mass $m(X)$ we define the reverse of X as $X^{REV} = \{m(X) - x_m, ..., m(X) - x_1, m(X)\}$ and the λ-shift of X as $X^{\overset{\lambda}{\to}} = \{x_1 + \lambda, ..., x_m + \lambda\}$.

Using this setup a theoretical MS/MS spectrum S for a peptide ρ is defined as

$$S = \mathcal{R}^{\overset{1}{\to}} \cup (\mathcal{R}^{REV})^{\overset{19}{\to}}$$

where elements of $\mathcal{R}^{\overset{1}{\to}}$ and $(\mathcal{R}^{REV})^{\overset{19}{\to}}$ respectively correspond to b and y ions [2] (neutral losses are considered later while scoring the spectrum). Also, we denote the parent mass of a spectrum S from a peptide ρ as $m(S) = m(\rho)$.

In mass spectrometry one often faces the inverse problem of transforming an experimental spectrum S into the mass series representation of a peptide. The simplest approach to this inverse problem is to reverse the transformation above

$$\text{PRM}(S) = S^{\overset{-1}{\to}} \cup (S^{\overset{-19}{\to}})^{REV}$$

The set PRM(S) represents an attempt to reconstruct the set $\mathcal{R} \cup \mathcal{R}^{REV}$ of the peptide ρ that generated S and defines the peak positions (PRMs) in our PRM spectrum. Figure 3.2 illustrates the steps described above. Ideal PRM spectra could be built from MS/MS spectra containing only b and y ions. This ideal setup can be approximated by selecting peaks from the experimental MS/MS spectra according to their intensity - higher intensity peaks tend to correspond to b and y ions. As such, PRM positions in a PRM spectrum could be determined using only the top 20 highest intensity peaks in each MS/MS spectrum. The choice to keep 20 peaks per MS/MS spectrum is motivated by the analysis of peak annotation histograms which show a very low percentage of b/y ions outside the top 20 intensity peaks (data not shown); b/y ions are the most important peaks in determining the correct positions for the PRMs, could then be scored using all the peaks in the MS/MS spectrum. We further note that the choice of number of peaks to keep per spectrum is a function of the expected peptide length - longer peptides (with higher parent masses) could benefit from the selection of a larger number of peaks.

[2] b-ions include an additional H atom (+1 Da); y-ions include an H_2O molecule (+18 Da) from the C-terminal and also an additional H atom (+1 Da) for a total peak offset of +19 Da.

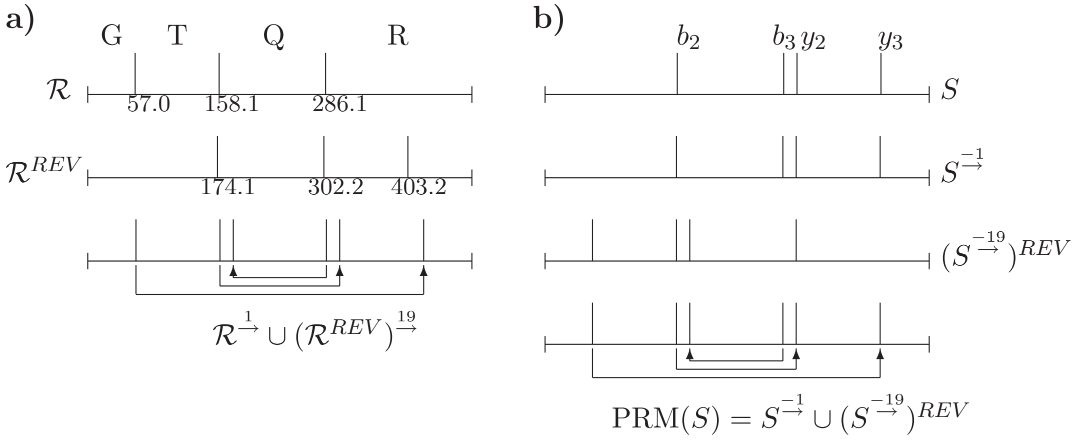

a)
G
T
Q
R
$\mathcal{R}$
57.0
158.1
286.1
$\mathcal{R}^{REV}$
174.1
302.2
403.2
$\mathcal{R}^{\overset{1}{\rightarrow}} \cup (\mathcal{R}^{REV})^{\overset{19}{\rightarrow}}$
b)
b_2
b_3
y_2
y_3
S
$S^{\overset{-1}{\rightarrow}}$
$(S^{\overset{-19}{\rightarrow}})^{REV}$
$\mathrm{PRM}(S) = S^{\overset{-1}{\rightarrow}} \cup (S^{\overset{-19}{\rightarrow}})^{REV}$

Figure 3.2 Part **a)** illustrates how the sets $\mathcal{R}$ and $\mathcal{R}^{REV}$ are used to define the theoretical PRM spectrum for the peptide GTQR. Part **b)** shows how a hypothetical MS/MS spectrum S for GTQR is processed to obtain the PRM spectrum $PRM(S)$. In both **a)** and **b)** arrows indicate the prefix/suffix pairs.

Every peak s in an MS/MS spectrum S generates two *complementary* PRMs ($s - 1$ and $m(S) - s + 19$). Every PRM spectrum P is then necessarily symmetric because every pair of complementary PRMs is symmetric about $\frac{M(P)}{2}$, where $M(P)$ abbreviates $m(P) + 18$.

Scoring PRMs in PRM spectra Not all PRMs are created equal - some have more compelling evidence of being correct than others by having, for example, both corresponding b and y ions and neutral loss peaks present in the MS/MS spectrum. In chapter 5 we reflect how confident we are in a PRM by using the Dančik et al. [26] scoring scheme[3] (see [65, 86, 87, 142] for other applications of the same scoring scheme). Chapters 6 and 7 use the scoring scheme described in [1]. These scoring schemes are particularly adequate for our purposes because they allow us to score putative PRMs without having a putative peptide interpretation (as is the case in database search). Other approaches also take into consideration the relative intensity values of the observed ion types [32, 33, 126, 144] although the best way to incorporate such information is still an open problem under consideration [36,62,127].

The scores defined by Dančik et al. are additive due to a log scaling and have the expected positive premium and negative penalty score changes for ion types

[3] Readers familiar with the scoring defined in [26] may recognize the connection between PRM spectra and scored vertices in the spectrum graph.

with probability of occurrence higher than the probability of background noise. Each PRM $p_i \in P$ is thus assigned a weight $w(p_i)$ by looking for supporting ion peaks in the corresponding MS/MS spectrum S to obtain a PRM spectrum $P = \{p_1, ..., p_n\}$ having associated weights $\{w(p_1), ..., w(p_n)\}$. In the following we assume that in every PRM spectrum the PRMs are sorted by increasing mass.

3.C Clustering MS/MS spectra

Repeated MS/MS spectra (multiple spectra from the same peptide) are common in high-throughput MS/MS experiments. Recent approaches either attempt to discard these to speed up database searches [125] or average over the multiple copies to increase the intensity of correct peaks relative to noise peaks [13] (although averaging could retain high intensity noise peaks in the consensus spectrum). A study by Venable and Yates [134] on the variability of experimental MS/MS spectra from the same peptide provides evidence that peak intensities vary considerably between repeated MS/MS spectra and also argues that although MS/MS spectra averaging improves database search results other approaches may perform better.

We propose to use the redundant information in the repeated MS/MS spectra to filter noise based solely on the principle that real MS/MS spectrum peaks should be present in most MS/MS spectra from the same peptide and the randomly distributed noise peaks should not. But to make use of the redundant information in independently obtained PRM spectra of the same peptide we first need to decide if two PRM spectra originate from the same peptide using *only* the information contained in the PRM spectra.

A naive approach to this problem could be based on the shared peak count between two spectra. However this ignores the fact that some peaks in an MS/MS spectrum have more evidence of being true peaks than others, e.g. by having additional peaks at corresponding neutral loss positions. It may also happen that in one MS/MS spectrum we only observe a b-ion for a given fragmentation point and in the other MS/MS spectrum we only observe a y-ion for the same fragmentation point in which case there are no matching peaks although there is relevant matching information in the spectra. Matching PRM spectra instead of MS/MS spectra addresses both of these points. A match between two PRM spectra P and Q can then be defined as a set $P \cap Q$ of matching PRMs (p_i, q_j) with associated weights $w(p_i) + w(q_j)$. The weight of any set of PRMs $X = \{x_1, ..., x_n\}$ is simply given by $w(X) = \sum_{x_i \in X} w(x_i)$.

Matching PRM spectra: sparse subsets. A subset of a PRM spectrum P is called *sparse* if no two PRMs are less than 57 Da apart (i.e. the mass of the lightest amino acid Glycine). In PRM spectra, peaks are supposed to correspond to prefix residue masses. Therefore, closely located PRMs (i.e. less than 57 Da apart) cannot be both correct. These closely positioned PRMs can be avoided by finding sparse subsets of PRMs.

To find a maximum weight sparse subset of a PRM spectrum $P = \{p_1, ..., p_n\}$ we define a simple dynamic programming recursion where $D(i)$ is the maximum weight of a sparse subset of $\{p_1, ..., p_i\}$ that includes p_i. Then

$$D(i) = w(p_i) + \max_{j:\ p_j \leq p_i - 57} D(j)$$

Matching PRM spectra: anti-symmetric subsets. Although computing maximum weight sparse subsets would already impose tighter conditions for PRM spectra matching, it may still happen that MS/MS spectrum peaks are double counted in the matching process. As described in the previous section, an MS/MS spectrum peak $s \in S$ generates a pair of complementary PRMs: $s - 1$ and $m(S) - s + 19$. Since both these PRMs are scored using the same MS/MS spectrum peaks, including both PRMs in a match effectively counts the same MS/MS spectrum peaks twice and should be avoided. As such, a subset X of a PRM spectrum P is defined as *anti-symmetric* if is has no complementary PRMs, i.e. no two PRMs in X add up to $M(P)$.

Matching PRM spectra: optimal subsets. A subset of a PRM spectrum P is *optimal* if it is a sparse and anti-symmetric subset of maximum weight. Computing an optimal subset of a set of PRMs is algorithmically the same problem as the *de-novo* problem of finding the peptide that best explains a spectrum [7,21,26]. The only difference between these two problems is that in the latter there is only a limited set of valid jumps between PRMs (corresponding to amino acid masses) while in the former any jump ≥ 57 Da is a valid jump. A detailed description of the implemented algorithm for the computation of optimal subsets can be found in Section A.2 of our supplementary material.

Match score between PRM spectra. An *optimal match* between two PRM spectra P and Q is simply an optimal subset of their overlap $P \cap Q$. Although the weight of an optimal match between PRM spectra is already a good measure of similarity, we observed that sometimes spurious high scoring matches occur when only a few PRMs match in a small mass range, simply by chance or due to local sequence similarities. On the other hand, repeated PRM spectra from the same peptide tend to match most high-scoring PRMs in a large mass range. To account for this effect we introduce a correction factor α - the percentage of mass range covered by the restricted match. Using d_{PQ} as the difference between the maximum and minimum masses of the matched PRMs (i.e. match range) and m_{PQ} as the parent mass of the matched PRM spectra, this correction factor is thus defined as $\alpha = \frac{d_{PQ}}{m_{PQ}}$.

The *match score* $\mathcal{M}$ between P and Q is then defined as $\mathcal{M} = \alpha \times w(Y)$, where Y is an optimal match between P and Q.

Constructing clusters of PRM spectra. After computing the match scores $\mathcal{M}$ we consider two spectra as similar if their match score is above a chosen threshold. A natural extension of the pairwise similarity concept when considering clusters of PRM spectra is to require each spectrum to coherently match at least two other spectra which must also match each other. A simple example of a cluster rejected through this condition is a star-like cluster of size n where $n-1$ spectra match only a single spectrum. This requirement is thus enforced as the *triangle* condition: a match between PRM spectra P and Q is retained if and only if there is some other PRM spectrum R such that P matches R and Q matches R. Clearly this step removes any cluster of PRM spectra of size < 3. The remaining matches define connected components interpreted as clusters. This approach was recently integrated in a hierarchical clustering framework to allow for efficient large-scale clustering of millions of spectra [44].

Chapter 3 is, in part, a reprint of the paper "Shotgun Protein Sequencing by tandem mass spectra assembly" co-authored with Haixu Tang, Vineet Bafna and Pavel Pevzner in Analytical Chemistry vol.76, pp.7721-33. The dissertation author was the primary investigator and author of this paper.

Clustering results. To evaluate the performance of our clustering procedure we used a set of 1455 `Sequest` annotated MS/MS spectra. `Sequest` annotations were used only for validation purposes and are not used by our algorithm at any point. A match between two PRM spectra is considered correct if the peptide annotations are the same for the MS/MS spectra originating the matched PRM spectra. Every spectrum was matched against every other spectrum with a parent mass difference not exceeding 2 Da; on average each spectrum was matched against 7.3 other spectra.

Figure 3.4 shows how true/false positives (TP/FP) and true/false negatives (TN/FN) vary for different thresholds on the match score $\mathcal{M}$; a Receiver Operating Characteristic (ROC) curve is shown on the left and, because the number of true positives is only around 10% of the number of true negatives, the precision vs. sensitivity curve is also shown on the right. For comparison purposes, Figure 3.4 also includes curves for the normalized dot-product approach proposed in [13, 125] as a similarity metric between MS/MS spectra.

As shown in Figure 3.4, our method clearly outperforms the normalized dot-product approach. One possible reason why our match score approach performs better than the normalized dot-product is the variability in peak intensity between different MS/MS spectra of the same peptide (see Venable and Yates [134]). The match score $\mathcal{M}$ thus allows us to separate between correct and incorrect pairwise matches with the choice of adequate threshold conditioned by the instrument parameters and the level of different peptides with the same precursor mass expected from the experiment. For our alignment and assembly purposes we selected a subset of matches as detailed in Table 3.2.

`Sequest` peptide annotations were also used to estimate the quality of the clusters obtained - the median percentage of non-matching peptide annotations in a cluster was found to be 11%. The retained 617 matches result in a total of 39 clusters - 29 annotated as coming from the protein in our sample and 10 where the peptide

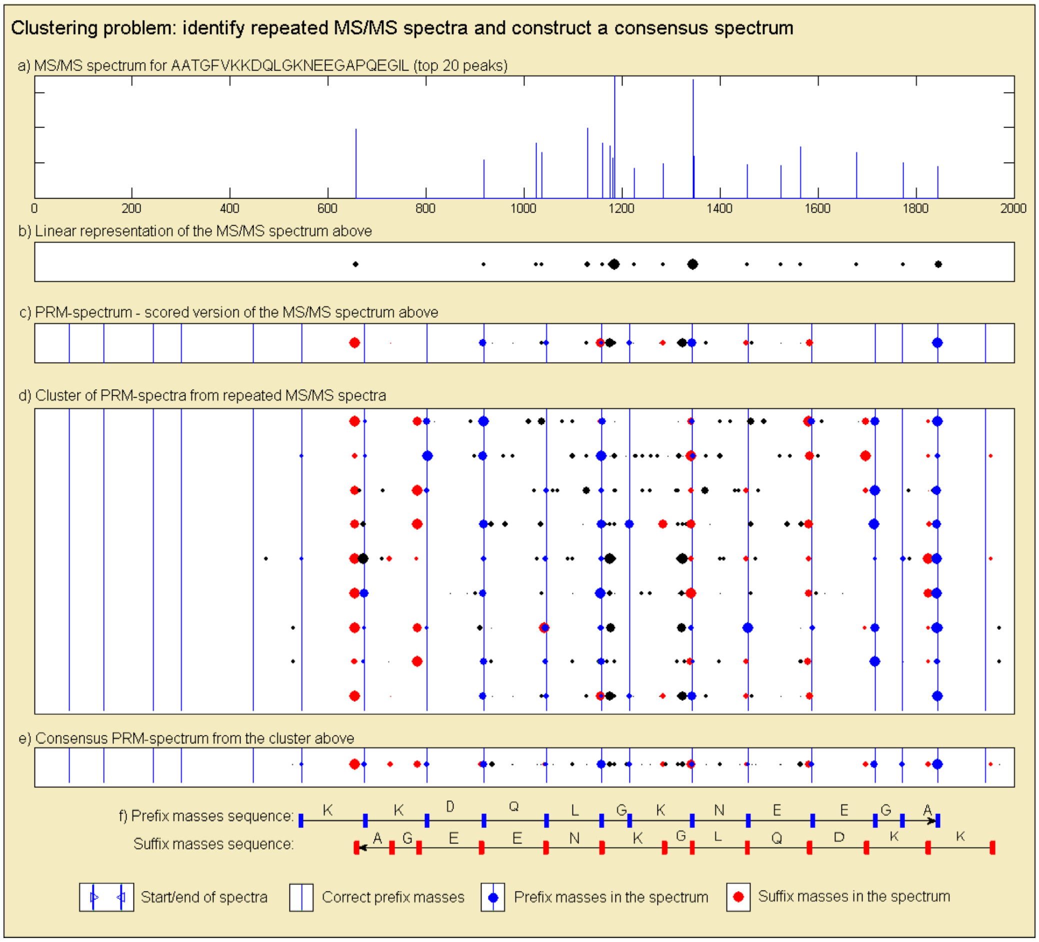


Figure 3.3 **Clustering phase**; **a)** and **b)** illustrate our linear representation of spectra where a dot indicates a peak and the dot size is proportional to the peak height (used to save space when showing multiple alignments of several spectra). Part **c)** shows the corresponding PRM spectrum (our preprocessed and scored version of an MS/MS spectrum). For convenience of the reader, prefix masses are colored green and suffix masses are colored red although this distinction is not known in advance. Spurious masses (that do not correspond to prefix or suffix masses) are shown as black dots. **d)** Clustering is then used to take advantage of redundant information in multiple spectra from the same peptide and **e)** obtain a single, more reliable, *consensus PRM spectrum* (some of the red colored dots are hidden by green colored dots). All black dots still present in e) correspond either to neutral losses or to doubly charged fragments. The increased number and significance of red/green dots in the consensus PRM spectrum as compared to individual spectra would already yield a reliable *de-novo* peptide sequence (as illustrated in **f)**), although we refrain from interpreting the spectra until the overlapping spectra are further processed.

annotations do not match. While it is possible that these 10 clusters are retained because our match score threshold was not aggressively selective it may also be the case

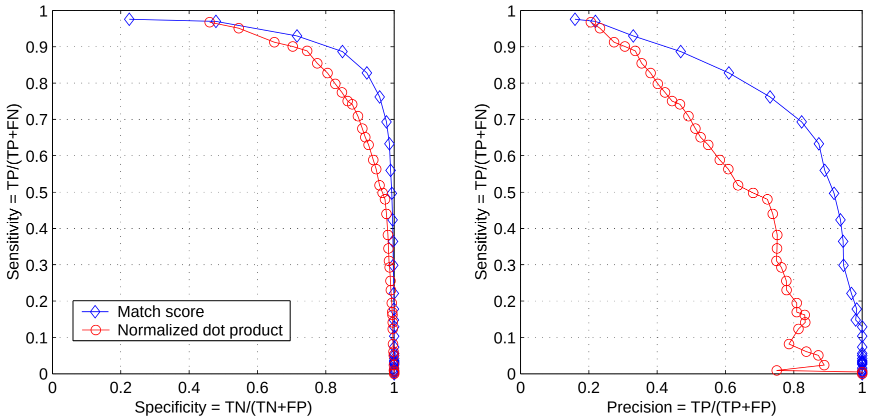


Figure 3.4 ROC curve (left) and precision vs sensitivity (right). The triangle condition is enforced on both methods.

Table 3.2 Match results in the clustering phase

	Number of matches	Number of correct matches	% correct
Total	5322	697	13%
After thresholding $\mathcal{M}$	823	545	66%
After triangle condition	617	501	80%

that the annotations are incorrect - only the highest scoring peptide annotation was retained from the database search procedure. In any event, the clustering procedure can be made as selective as desired (depending on the experiment requirements), with an acceptable penalty in sensitivity. Our choice of sensitivity/selectivity tradeoff reflects the fact that the obtained clusters are not our final goal but rather a preprocessing step for the alignment procedure where some amount of noise (incorrect PRM spectra) is tolerable. Also, a minor amount of incorrect MS/MS spectra in any single cluster does not produce a significant amount of noise in the corresponding consensus spectrum (see next section).

Building consensus PRM spectra. The usefulness of any spectral clustering technique is defined by how well the consensus spectrum reflects the true peaks in all spectra originating from the same peptide. As mentioned above, our approach to this problem is to score the putative PRMs across *all* clustered spectra - real peaks should appear in most MS/MS spectra (albeit with varying intensities) and noise peaks should not. As such, when given a cluster $C = \{P_1, ..., P_k\}$, a single consensus PRM spectrum can be constructed by a direct extension of the scoring procedure described above. The weight $w(t, C)$ of a putative PRM t over the cluster C is given by

$$w(t, C) = \sum_{i=1}^{k} w(t, P_i)$$

where $w(t, P_i)$ is the PRM weight (positive or negative) for the mass t in the i-th PRM spectrum in the cluster. Negative PRM scores occur whenever there is little or no evidence that a putative PRM represents a real prefix residue mass, e.g. by not having corresponding b or y ion peaks in the MS/MS spectrum. This is a common event when scoring a putative PRM t originating in a noise peak in one of the clustered spectra - most other spectra will have no peaks supporting t and the overall score for t will thus be negative. The consensus PRM spectrum for a cluster considers all putative PRMs in all PRM spectra in the cluster but retains only the PRMs with a positive summed score. In this way, PRMs generated by high intensity unexplained peaks in any MS/MS spectrum are not likely to be present in the consensus PRM spectrum because its absence in all other spectra will make its summed score negative. Although relative intensities vary across multiple MS/MS spectra from the same peptide [134] the presence or absence of real fragment peaks tends to be stable. As shown in Table 3.3 our resulting consensus PRM spectra are dominated by high scoring PRMs at the correct prefix and suffix mass positions and almost half of the remaining PRMs correspond to either doubly charged fragment masses or neutral losses (Figure 3.3). As a result, the *de-novo* interpretation of the consensus PRM spectra is greatly simplified as compared to individual spectra . However, we refrain from *de-novo* interpretation at this stage to take advantage of overlapping MS/MS spectra as described in the following chapters.

Table 3.3 Quality of the PRMs in the consensus PRM spectra

Type of fragment originating PRM	Median % of PRM spectrum score
b/y	77%
Neutral-loss or doubly charged	10%
Unexplained	12%

4

De novo sequencing of MS^2/MS^3 spectra

Mass spectrometry-based analysis of proteins is usually conducted by collecting MS/MS (MS^2) spectra and matching them against a database of known protein sequences. The experimental protocol can be extended to also collect multiple MS/MS/MS (MS^3) spectra for each MS^2 spectrum and thus increase the reliability of peptide identifications. But, as discussed in chapter 2, spectra from peptides generated by combinatorial assembly and rearrangements (e.g. peptides from novel immunoglobulins or fusion proteins in cancer) or peptides from unsequenced species are not amenable to database search. In such cases, the correlated information in MS^2/MS^3 spectra from the same peptide can be combined to increase the accuracy of de novo peptide sequencing and attenuate the difficulties in still unreliable de novo sequencing of individual MS^2 spectra. Nevertheless, the absence of algorithms and software for MS^2/MS^3 analysis has limited the utility of this straightforward experimental approach. In this chapter we analyze de novo peptide sequencing from multiple related spectra (like reconstructing amino acid sequence PEPTIDE from spectra of peptides PEPTIDE, PEPTID, and EPTIDE) and develop a probabilistic framework for solving this problem. We further apply this framework to develop an efficient algorithm for the MS^2/MS^3 de novo sequencing problem. The gain in sequencing accuracy is demonstrated on a dataset of yeast MS^2/MS^3 spectra and shown to achieve nearly perfect accuracy when enough ‘usable’ MS^3 are available. We additionally evaluate the impact of the number of usable MS^3 spectra on the sequencing accuracy and discuss the underlying tradeoff with instrument-cycle time.

4.A Introduction

The coupling of tandem mass (MS^2) spectrometry with database search tools [103, 130, 144] is the enabling core behind high-throughput protein identification [3]. Unfortunately, this successful strategy is not applicable whenever the protein sequences are not known in advance. Particularly important examples of value derived from initially unknown proteins include antibody drugs such as Herceptin$^{\text{TM}}$or Avastin$^{\text{TM}}$ [56, 139] and drugs derived from venom proteins [82, 108]. Antibodies illustrate the scenario where the universe of possible protein sequences is very large and constantly altered by recombination and somatic hypermutation [91]. Additionally, drugs derived from venom proteins exemplify the potential benefits that can be derived from exploring the wide range of viable proteins already probed by natural biodiversity.

Mass spectrometry-based studies of unknown proteins often have to resort to de novo peptide sequencing techniques that attempt to recover the amino acid sequences directly from the spectra. However, de novo peptide sequencing from an experimental MS^2 spectrum remains a challenging problem. While algorithms have been developed to find a best-‘scoring’ peptide for a given spectrum [1,7,21,26,41,88], their sequencing accuracy is still strongly affected by incomplete fragmentation, noise and ambiguity in ion-type assignments. In abundant, low-complexity samples, these difficulties may be attenuated by generating overlapping peptides and combining the resulting MS^2 spectra to yield higher accuracy de novo sequences - examples include ^{16}O/^{18}O labeling [116] or Shotgun Protein Sequencing [8, 9] (described in chapters 5 and 7). Intuitively, combined spectra from overlapping peptides can be used to increase the signal-to-noise ratio (noise peaks are scattered while b/y-ions match consistently) and it becomes possible to separate b-ions from y-ions, therefore simplifying the de novo sequencing task.

In this chapter we explore an alternative approach to the acquisition of spectra from overlapping peptides by generating up to 5 MS/MS/MS (MS^3) spectra per MS^2 spectrum as described in chapter 2. Although MS^3 spectra typically have lower quality than MS^2 spectra, the additional fragmentation has already been shown to increase the confidence in database search results [99]. The manual usage of MS^3 spectra as an aid to de novo sequencing dates back to almost a decade ago [85] and an automated approach was previously proposed by Zhang&McElvain [145]. However, Zhang&McElvain were limited to 42 sets of MS^2+MS^3 spectra and thus the sequencing accuracy gains from the proposed heuristics was described qualitatively rather than quantitatively. Moreover, this approach has not enthused further MS^2/MS^3 sequencing efforts, in part, because no implementation is publicly available.

Sequencing an MS^2 spectrum in conjunction with k dependent MS^3 spectra entails searching for the best-scoring peptide while considering every possible combination of fragment types for the MS^3 spectra (i.e. was the MS^3 generated from a b- or a y-ion). Our approach explores this search space by using dynamic programming to find the best peptide and exhaustive search to consider all 2^k possible configurations for MS^3 fragment-type assignments. Moreover, we build on a probabilistic model [1] to score peptide-spectrum matches and extend it to include the particularities of MS^2/MS^3 sequencing. Using this approach, we show that MS^3 spectra can significantly increase de novo sequencing accuracy and even make it almost error free when enough ‘usable’ MS^3 spectra are available.

4.B Methods

Dataset

The results described in this chapter were obtained using a dataset containing 3184 MS^2 spectra with 15770 dependent MS^3 spectra. In addition, this dataset contained 2181 MS^2 spectra that did not generate any MS^3 spectra and were thus not investigated here. InsPecT [130] was used to search a database containing 7517 *Saccharomyces cerevisiae* protein sequences (SwissPROT, Oct.8, 2006) and sequences of common contaminant proteins. The database additionally contained 7517 decoy protein sequences used to enforce the selected 5% false discovery rate. InsPecT was configured to allow for 0.5 Da fragment mass tolerance and 2.5 Da precursor mass tolerance; the high accuracy of the experimental precursor masses (measured on a Fourier Transform instrument) was used to confirm identifications rather than to restrict the space of possible peptides. The set of allowed modifications was oxidation (M), phosphorylation (S,T,Y), acetylation (N-term), deamidation (Q) and 13C(6)15N(2) Silac label (K). While our approach does not address sequencing with post-translational modifications [1], these were known to be present in the sample and the unknowing utilization of spectra from these peptides would have resulted in distorted estimates of de novo sequencing accuracy.

The search identified 890 out of all 3148 MS^2 spectra. However, only 1282 MS^3 spectra (out of 15770) resulted in a significant match to the database. To further increase the number of annotated MS^3 spectra we matched each MS^3 precursor mass to the theoretical fragment masses from the peptide assigned to the parent MS^2 spectrum. Note that this procedure does not increase the number of identified peptides but rather allows us to better characterize the data and evaluate the algorithms described below. The selection of unambiguous MS^3 precursor mass matches resulted in 1039 MS^3 spectra annotated as prefix fragments (b-ions), 2592 MS^3 spectra annotated as suffix fragments (y-ions) and 320 spectra annotated as prefix/suffix fragments after loss of H_2O or NH_3. The remaining non-annotated 496 MS^3 spectra either did not match a theoretical fragment mass or could be interpreted as two different types of fragment. Nevertheless, all dependent MS^3 spectra from all identified MS^2 spectra were used for de novo sequencing.

The ion statistics for all identified spectra (shown in Table 4.2) allow us to quantify the differences between MS^2 and MS^3 spectra. In general, the latter tend to have less explained intensity and less b/y-ion peaks. Note that these observations are not entirely surprising because we only consider the MS^2 spectra that had a strong match to the database while most MS^3 spectra were identified only by their precursor mass.

[1] With the single exception of the very abundant Silac-K whose mass was added as a 21^{st} amino acid.

De novo peptide sequencing problem

As discussed in chapter 2, the simplest way to score a spectrum S against a peptide P is to count the number of peaks in common between S and the theoretical spectrum of peptide P. However, the best de novo sequencing results have been obtained using probabilistic models that capture multiple features such as peak intensities and expected propensities of the different ion types [1, 26, 41, 88]. In this chapter, we start by introducing a model that seemingly has nothing to do with de novo peptide sequencing but rather describes a very general probabilistic process that transforms one Boolean string into another. We will show later that this process not only generalizes the probabilistic model for de novo peptide sequencing from chapter 2 but also allows one to study de novo peptide sequencing from multiple spectra.

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

Table 4.1 Probability $P(x|y)$ of a peptide symbol y generating a spectrum symbol x.

y x	1	0
1	ρ	θ
0	$1-\rho$	$1-\theta$

Given a spectrum s and a set of strings Π, we are interested in solving the optimization problem $\max_{\pi\in\Pi} P(s|\pi)$. Below we focus on the sets Π that are relevant in the context of tandem mass spectrometry. Let $V = \{1, \dots, n\}$ and $G(V, E)$ be a topological ordering of a DAG (Directed Acyclic Graph [50]) such that $i < j$ for every directed edge (i, j) in E. Every path from 1 to n in graph G corresponds to a *G-peptide* $\pi = \pi_1 \dots \pi_n$ such that $\pi_i = 1$ iff vertex i belongs to the path (see Figure 4.1). We are interested in the following Peptide Sequencing (PS) Problem:

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

We impose no restrictions on the graph $G(V, E)$ but in practical applications it is usually assumed that $(i, j) \in E$ iff $(j-i)$ equals the integer mass of an amino acid. Such graphs are usually referred to as *spectrum graphs*, as described in chapter 2.

The relation between this abstract Boolean strings model and de novo peptide sequencing is straightforward. In reality, an MS/MS spectrum can be represented as a string of ones (peak present) and zeros (no peak present), with a 0/1 for every consecutive 1 Da interval. Similarly, sequences of amino acid masses (peptides) can also be represented as strings of zeros and ones: every amino acid can be represented as a string of $\alpha - 1$ zeros followed by a single one, where α is the integer amino acid mass. Then, a peptide is simply a concatenation of the Boolean strings corresponding to its sequence of amino acids. In this context, $\theta \approx 0.05$ (probability of observing a noise peak) and $\rho \approx 0.7$ (probability of observing a b-ion) represent typical values of θ and ρ for ion-trap MS/MS spectra (Table 4.1). This somewhat simplistic Boolean string model can be modified for any mass resolution, peptide fragmentation rules and peak intensities [1, 7, 21]. Moreover making this model more realistic typically does not affect the algorithmic solution. In particular, chapters 5 and 7 describe how spectral alignment can be used to separate between b/y-ions and thus generate strings as in this model.

The PS Problem has an easy solution first described by Dančik et. al [26]. We will find it convenient to cast the approach in [26] as an application of the Viterbi algorithm [2, 110] in an appropriately constructed Hidden Markov Model $\mathcal{G}$. Figure 4.1 shows a graph G with every edge (i, j) substituted by a path with new $j - i - 1$ *traversal* vertices that starts at i and ends at j. The resulting graph with vertex set $V \cup T$ (where $V = \{1, \ldots, n\}$ and T is the set of all traversal vertices) represents the hidden states of the HMM $\mathcal{G}$ and all possible transitions between the states. The emission probabilities of the HMM are defined by the matrix $P(x|y)$ with $P(x|y = 0)$ for T states and $P(x|y = 1)$ for all other states. The transition probabilities of all edges in this HMM are defined to be 1[2]. Finding an optimal

[2] Although these transition probabilities do not add up to 1, we prefer not to normalize them. This keeps the resulting probabilities of hidden paths consistent with the Dančik et al. model [26] and does not affect the Viterbi algorithm for finding an optimal path.

path in this HMM is a straightforward application of the Viterbi algorithm. In mass spectrometry, the graph G usually encodes all peptides of a given *parent mass* n. As such, even though all models below work for an arbitrary graph G, we will henceforth assume that G is a spectrum graph.

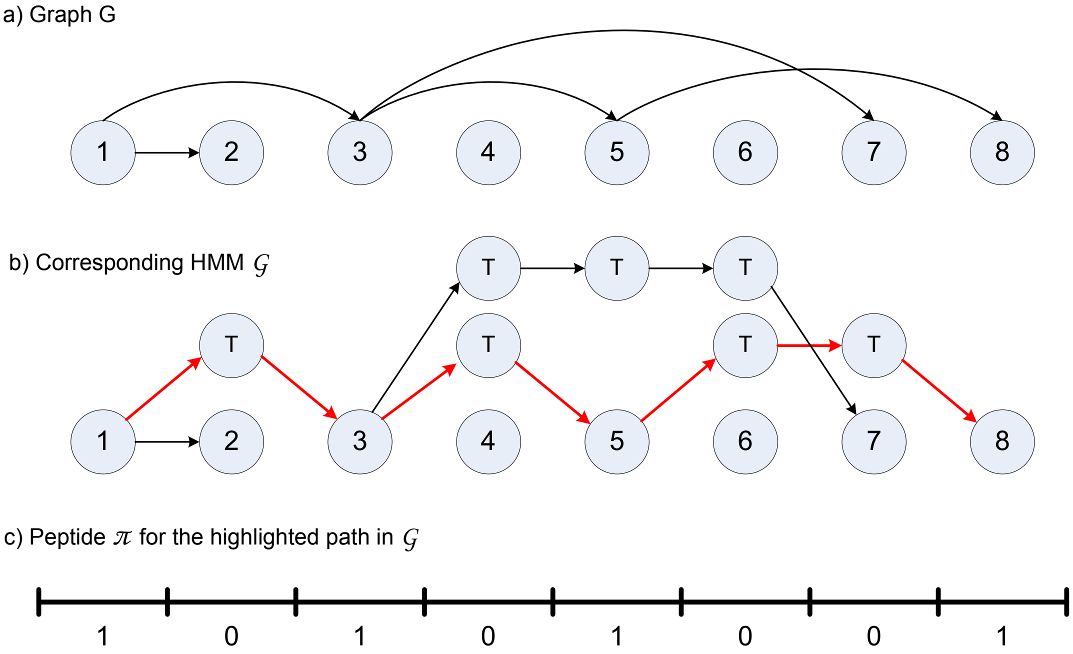

a) Graph G
1
2
3
4
5
6
7
8
b) Corresponding HMM $\mathcal{G}$
T
T
T
T
T
T
T
T
1
2
3
4
5
6
7
8
c) Peptide π for the highlighted path in $\mathcal{G}$
1
0
1
0
1
0
0
1

Figure 4.1 Construction of a Hidden Markov Model $\mathcal{G}$ for a graph G. The red (hidden) path corresponds to the $\mathcal{G}$-peptide 10101001.

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

However, this formulation encodes a particular bias towards peptides that have pairs of b/y-ions such that the mass of b_i equals the mass of y_j. In these cases, both ions use the same masses in the spectrum to artificially increase the score of the peptide but do so with conflicting ion type assignments. Peptides that do not have such pairs of b-ions are referred to as *anti-symmetric* peptides and efficient algorithms are available to find the maximum scoring anti-symmetric peptide for any given spectrum $[7, 21]^3$. Nevertheless, ambiguous b/y-ion assignments remain one of the main sources of de novo sequencing errors. We show below how MS^3 spectra can help resolve these ambiguities.

Although this model still essentially amounts to maximizing the weighted number of matched masses between a spectrum and a peptide, it already captures enough detail to allow us to describe the proposed extensions for combining $\mathrm{MS}^2/\mathrm{MS}^3$ spectra. In practice, this same framework is used to model more elaborate events that take into consideration the intensity of the peaks in the spectrum and to account for the presence/absence of other ion types (e.g. $b - H_2O$ ions). As shown in Table 4.2, using the more elaborate scoring terms from [1], one can replace raw intensities with peak scores to significantly increase the signal-to-noise ratio over all collected spectra (see chapter 3). The resulting scored spectra have more intensity assigned to true fragment masses and feature a much smaller number of noise peaks while simultaneously retaining almost all b/y-ions.

Multi-spectra sequencing problem and $\mathrm{MS}^2/\mathrm{MS}^3$ analysis

The simultaneous sequencing of spectra from multiple peptides (e.g. PEPTIDE, PEPTID, EPTIDE) requires solving two problems: *i*) finding the correct multiple alignment between all spectra (described in the next section) and *ii*) reconstructing a maximal-scoring peptide from the aligned spectra. In the following we assume that the spectra are already aligned (as shown in Figure 4.2) and describe the problem of peptide sequencing from multiple *aligned* spectra.

MS^3 spectra contain valuable additional information in the form of corroborating fragmentation - peaks at the expected fragment masses in the MS^2 and all MS^3 spectra. When the MS^3 spectra are correctly aligned with the MS^2 spectrum the corroborating b-ions match 'vertically' while the matching y-ions are found at different positions depending on the parent mass of each MS^3 spectrum (see Figure 4.2). As such, we now consider the problem of finding the most probable peptide $\pi = \pi_1, ..., \pi_n$

[3] In practice, peptides that are not anti-symmetric are not excluded from consideration but special care is taken to avoid multiple ion-type assignments to the same spectrum peaks in the scoring function.

for a set of multiple (possibly overlapping) spectra $s^1, ..., s^k$. Thus, the alignment between the MS^2/MS^3 spectra defines a *substring mapping* where $\phi(s^i) = \{a, ..., b\}$ is the sequence of *consecutive* numbers from a to b ($1 \leq a \leq b \leq n$) such that the substring $\pi_{\phi(s^i)} = \pi_a, \ldots, \pi_b$ *generates* the spectrum s^i.

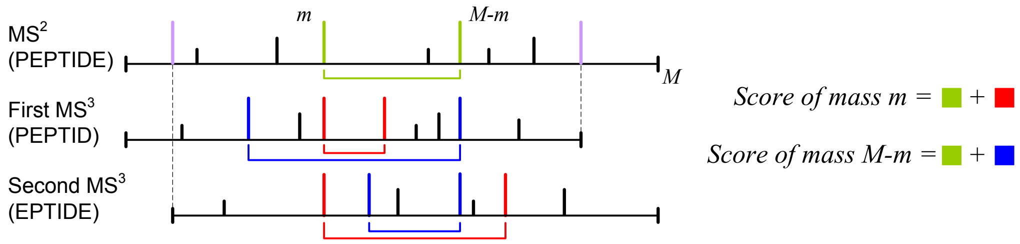


Figure 4.2 Conceptual illustration of peak scoring on an MS^2 spectrum (e.g. from PEPTIDE) aligned with its dependent MS^3 spectra (e.g. S_1 from PEPTID and S_2 from EPTIDE). Since the correct alignment between MS^2/MS^3 spectra is not known a priori, our algorithm finds the highest scoring peptide over all possible alignments. M indicates the parent mass of the MS^2 spectrum and complementary masses (e.g. b/y ions whose masses add up to the corresponding parent mass) are connected by arcs under each spectrum. Colors indicate the sets of masses accounted for when scoring different peaks; the colored squares in the equations represent the summed scores of all peaks of the same color; violet peaks indicate the b-ion masses dictated by the parent masses of the MS^3 spectra. Intuitively, the score of a mass m summarizes all corroborating evidence for it being generated from a prefix fragment. The red and blue peaks illustrate the contribution of the MS^3 spectra to the separation of b/y-ions in the MS^2 spectrum - intuitively, if m is a b-ion and $M - m$ is not then the red peaks should be more prominent than the blue peaks.

We now consider the following Multiple Spectra Peptide Sequencing (MSPS) Problem:

Multi-Spectra Peptide Sequencing (MSPS) Problem.

Given spectra $s^1, ..., s^k$ and a substring mapping ϕ, find a G-peptide π maximizing $P(s^1, ..., s^k|\pi) = \prod_{i=1}^{k} P(s^i|\pi_{\phi(s^i)})$.

It is easy to see that the HMM constructed for the PS Problem (described above) can also solve the MSPS Problem. The only difference between the two HMMs resides in the types of symbols emitted by the hidden states. Let $S(i)$ be the set of all spectrum values generated from π_i (defined by spectra $s^1, \ldots, s^k$ and the mapping ϕ). Then, each hidden state now emits a set of independent values $S(i) = \{v_j\}$, with

$P(v_j|\pi_i)$ given by Table 4.1 as before: $P(x|y = 0)$ for T states and $P(x|y = 1)$ for all other states. Assuming all v_j are independent observations, the probability of observing any given set $S(i)$ is $\prod_{v_j \in S(i)} P(v_j, \pi_i)$. Once again, the MSPS problem can be solved by an anti-symmetric sequencing algorithm on $\mathcal{G}^*$ [7, 21]. We note that although multiple ion-type assignments can be avoided in the MS^2 spectrum, the resulting peptide may be somewhat biased by conflicting ion-type assignments in the MS^3 spectra. While these conflicts could be readily avoided when only one MS^3 spectrum is available (see chapter 6), extending the algorithm to more MS^3 spectra would lead to a significant computational burden unlikely to result in relevant gains in terms of sequencing accuracy.

A possible change to the MSPS Problem would be to require the start/end positions of each spectrum as mandatory 1s in the returned peptide π. This modified problem can also be solved using the strategy described above if we modify the HMM $\mathcal{G}^*$ by removing all T states at the start/end positions of every spectrum. As such, every path through this modified HMM $\mathcal{G}^*$ will be forced to use the states corresponding to these start/end positions and thus results only in peptides that are consistent with the alignment of MS^3 spectra. Note that since MS^3 spectra have different ion statistics than MS^2 spectra (see Table 4.2) , it makes sense to use different scoring models for each type of spectra [1]. However, due to the small sample size of MS^3 spectra, it is difficult to retrain the scoring model. Therefore, the peaks in MS^3 spectra were scored using a model trained on MS^2 spectra and the resulting scores were somewhat arbitrarily divided by 2 to reflect the lower expectation of finding true fragment masses in the MS^3 spectra.

Aligning MS^2/MS^3 spectra

Under common experimental conditions, the highest intensity peaks in MS^2 spectra typically correspond to *y*- and, to a lesser extent, *b*-ions. Another common source of high intensity peaks are doubly charged ions. However, one can avoid doubly-charged fragment ions by considering only peaks with a mass higher than that of the precursor ion. By restricting the selection of MS^3 precursor ions to this high-mass region our experimental setup *i*) implicitly selected singly-charged MS^3

precursors from doubly-charged MS^2 precursors and *ii*) biased towards MS^3 spectra generated from b/y-ions [4]. As such, determining the correct alignment between an MS^3 spectrum and its parent MS^2 spectrum essentially reduces to determining the b/y-ion type of the MS^3 precursor ion.

Given an MS^2 spectrum from a particular peptide (e.g. PEPTIDE), the generation of an MS^3 spectrum from one of its b-ions yields additional information about the corresponding prefix peptide. For example, a dependent MS^3 spectrum from b_6 would contain additional fragment masses from the prefix peptide PEPTID. The converse reasoning applies to MS^3 spectra from y-ions and suffix peptides (e.g. y_5 contains fragment masses for PTIDE). As such, the assignment of an MS^3 spectrum to the correct ion type allows one to match the corroborating fragmentation from the same peptide regions and thus reinforce the confidence in co-occurring fragment masses. However, the correct ion-type assignments are not known in advance and the direct consequence of this uncertainty is that whenever an MS^2 spectrum results in the generation of k MS^3 spectra one needs to explore 2^k possible combinations of assignments - an easy task since k is usually small. We note that k is often less than the number of acquired MS^3 spectra; since many MS^3 spectra bear little resemblance to the parent MS^2 spectrum. Using the set of peaks with matching masses between the MS^2 and MS^3 spectrum (in either ion-type assignment), we define an MS^3 spectrum as *usable* if the summed scores of the matched peaks include at least 25% of the total summed peak scores (in each spectrum).

Using the peptide sequencing framework described above, our approach explores all 2^k possible ion-type assignments for an MS^2 spectrum with k usable dependent MS^3 spectra and selects the combination of assignments resulting in the highest scoring peptide.

4.C Results

When analyzed in isolation, single MS^3 spectra are less useful than single MS^2 spectra. Lower amounts of substrate and a bias towards shorter peptides re-

[4] The only notable exceptions (≈7% of all MS^3 spectra) were the selection of MS^3 precursors from neutral-loss ions (e.g. $b - H_2O$, $y - NH_3$).

sult in spectra with generally inferior ion statistics (see Table 4.2) and consequently a much smaller percentage of identified spectra - 8% of all MS^3 spectra vs 28% of all MS^2 spectra (using traditional database search). However, the combination of dependent MS^3 spectra with the parent MS^2 spectrum promptly reveals matching peaks from true peptide fragments and non-matching peaks from unexplained noise masses. Capitalizing on this corroborating fragmentation leads to a significant increase in signal-to-noise ratio with 19% intensity in non-explained peaks vs 31%/49% in MS^2/MS^3 spectra. Also, the distinct locations of b/y-ions in the aligned MS^2/MS^3 spectra (see Figure 4.2) allow one to separate between these ion types. Separating b/y-ions alleviates the uncertainties in peak ion-type assignments and thus reduces the probability of high-scoring incorrect peptide identifications [5]. Furthermore, the parent masses of the usable MS^3 spectra create a strong bias towards the corresponding b-ion masses. Since the overwhelming majority of all usable MS^3 spectra come from true fragment masses, the resulting set of essentially mandatory b-ions significantly reinforces the score of the correct peptide while severely reducing the probability of an incorrect high-scoring peptide match (see Figure 4.3). The gains obtained from merging MS^3 spectra with the corresponding parent MS^2 spectra are summarized in Table 4.2c. The different number of spectra in Table 4.2b-c stems from the fact that 202 MS^2 spectra did not generate any usable MS^3 spectrum and also because 86 MS^2 spectra contained modified residues not considered for our current de novo sequencing purposes. The only exception to the latter was Silac-labeled Lysine because of the high number of peptides containing this modification.

Our analysis revealed that scored MS^3 spectra sometimes have b/y-ions absent in MS^2 spectra. Such additional fragmentation is especially important in a de novo sequencing context because one implicity searches for the best peptide match over the space of all possible peptides of the observed parent mass. The combined contributions of the factors described above results in a strong bias towards the correct peptide sequence reflected in a significant increase in the average percentage of correctly predicted amino acids (from 85.7% to 90.7%, Table 4.3). Moreover,

[5] Either via de novo sequencing or database searching, although the latter is not explicitly evaluated here

we note that this increase in sequencing accuracy is not achieved at the cost of making less predictions - sequencing MS^2/MS^3 spectra resulted in 4922 amino acid predictions vs 4772 for MS^2 spectra only. As expected, having more usable MS^3 spectra results in increasing de novo sequencing accuracy, generating almost-error-free sequences as soon as 4 usable MS^3 spectra are available.

Table 4.2 Spectrum ion statistics. **a)** Ion statistics for RAW spectra. **b)** Ion statistics after replacing each peak's raw intensity with a likelihood score [1]. **c)** Ion statistics of the consensus spectra obtained by merging each MS^2 spectrum with its dependent MS^3 spectra (after scoring). This category has less spectra (602) than the others (890) mainly because not all MS^2 spectra had at least one 'usable' MS^3 spectrum (details in main text). **d)** Ion statistics for the peaks selected by de novo sequencing on the merged spectra.

	Type of spectra	*# spectra*	*% intensity*				*# peaks*	*p(b)*	*p(y)*
			b	*y*	*satellite*	*unexplained*			
a)	MS/MS (MS^2)	890	13%	30%	20%	37%	542	0.81	0.84
	MS/MS/MS (MS^3)	4447							
	from y-ions	2592	11%	17%	23%	49%	62	0.41	0.51
	from b-ions	1039	11%	1%	25%	64%	47	0.27	0.09
b)	Scored MS^2	890	28%	28%	23%	31%	80	0.81	0.81
	Scored MS^3	4447							
	from y-ions	2592	20%	20%	16%	44%	41	0.53	0.53
	from b-ions	1039	14%	3%	22%	61%	36	0.30	0.12
c)	Merged (MS^2+MS^3)	602	66%	8%	7%	19%	83	0.89	0.77
d)	Merged + sequenced	602	92%	1%	1%	6%	11	0.82	0.01

Table 4.3 De novo sequencing accuracy; $X \rightarrow Y$ indicates the gains in sequencing accuracy from combined MS^2/MS^3 spectra (Y) vs MS^2 spectra in isolation (X). MS^3 spectra are said to be usable if found to have a significant match to the parent MS^2 spectrum (details in the text). Each line summarizes the average sequencing results when combining an MS^2 spectrum with k dependent usable MS^3 spectra, $1 \leq k \leq 5$. *Sequencing accuracy* is defined as the number of correctly predicted amino acids divided by the total number of predicted amino acids; a *tag* is a sequence of consecutive amino acid predictions. Note that tags may occur in the middle of the spectrum and do not necessarily connect to the start/end of the spectrum. As expected, the gains in sequencing accuracy increase with the number of usable MS^3 spectra and result in almost-error-free sequences when 4 usable MS^3 spectra are available. † although there were no sequencing errors whenever 5 usable MS^3 spectra were available, one set of MS^2/MS^3 spectra did not generate a long tag due to missing peaks in the spectra.

# usable	# cases	Sequencing	Mean tag length		% cases with a correct
MS^3 spectra		accuracy	Uninterrupted	One jump of 2 aa	uninterrupted tag of length ≥ 6
1	151	81.8→ 84.1	5.9→ 6.0	7.5→7.5	61→64
2	187	85.1→ 90.4	6.2→ 6.7	7.6→8.2	66→76
3	150	88.8→ 93.1	6.3→ 6.8	7.8→8.2	67→77
4	91	88.6→ 96.0	6.3→ 7.0	7.4→7.9	74→86
5	23	85.0→ 100.0	6.2→ 7.4	7.6→8.3	78→96†
Overall	602	85.7→ 90.7	6.2→ 6.6	7.6→8.0	67→76

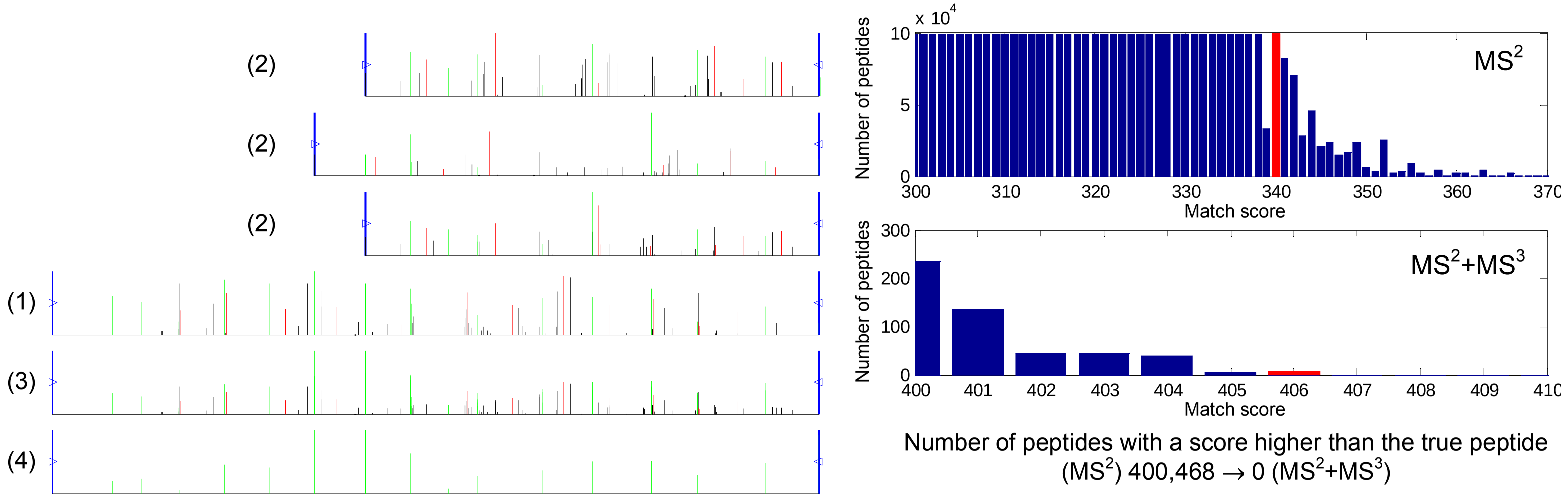


Figure 4.3 Spectra for peptide GPAPLNLEIPAYEFDGDK. **Left**) Scored MS^2 spectrum for GPAPLNLEIPAYEFDGDK (1), dependent scored MS^3 spectra (2), merged spectrum (3) and peaks selected in the de novo sequence (4). Each spectrum is shown with *b*-ions colored in green and *y*-ions colored in red. **Right**) Histograms of peptide scores for the MS^2 (upper half) and combined MS^2/MS^3 spectra (lower half). On both cases, the x-axis represents the peptide-spectrum match score and the y-axis shows the number of different peptides resulting in the same score (out of all possible peptides); the bin containing the peptide identified by database search is highlighted in red. As illustrated, combining the MS^2/MS^3 spectra into a merged spectrum (3) dramatically reduced the number of incorrect peptides with a match score higher than the correct peptide and thus made it possible to accurately recover the correct sequence.

The higher sequencing accuracy in combined MS^2/MS^3 spectra increases the number of spectra for which one can recover long subsequences (i.e. *tags*) of at least 6 consecutive amino acids. The availability of MS^3 ameliorates but does not completely eliminate the difficulties in de novo sequencing of complete peptides caused by missing fragment masses and high-intensity noise peaks. Nevertheless, one is often able to confidently recover tags that may uniquely identify the peptide using a simple (and very efficient) text-based database search. Even when these long tags happen to match multiple locations in larger databases, the number of possibilities tends to be very small and can be quickly resolved by matching the N/C-terminal masses (tags are usually recovered from the middle of the spectrum) and additional peaks not included in the tag. These factors are especially relevant when de novo sequencing is followed by homology-tolerant searches such as those enabled by MS-Alignment [133]. In these cases, the tag-based efficiency gains [10,130] should combine with the improved ion statistics shown in Table 4.2d to simultaneously deliver faster and more accurate results.

While Table 4.3 shows a strong connection between the number of usable MS^3 spectra and de novo sequencing accuracy, it turned out that only one third of all MS^3 spectra was found to be usable (see Table 4.4). Out of these, the overwhelming majority was generated from y-ions in the parent MS^2 spectrum - most likely a direct consequence of the higher quality ion statistics in this type of MS^3 spectra. In contrast with dependent spectra from other ion types, the percentage of usable MS^3 spectra from y-ions drops sharply with increasing numbers of collected MS^3 spectra. While this effect can be partially explained by decreasing amounts of substrate, Table 4.4 also shows a complementary increase in the number of non-usable spectra from other ion types, thus suggesting the alternative explanation that this effect may be exacerbated by the progressive selection of precursor ions from other ion types. Offline analysis of the collected MS^2 spectra reveals that one third of the top 5 peaks have masses below that of the precursor MS^2 ion. As a consequence, peaks of lower intensity but with masses in the upper half of the MS^2 spectrum were sometimes selected for additional fragmentation. Although the lower halves of the MS^2 spectra have different propensities of each ion type (e.g. only 30% are y-ions instead of almost

60% in the upper halves), the higher amounts of substrate may result in more usable MS^3 spectra. In particular, doubly-charged precursor ions have the potential to yield higher quality MS^3 spectra and could eventually justify extending our approach to support MS^3 spectra from doubly-charged fragments.

4.D Discussion

With this approach, we have demonstrated the utility of MS^3 spectra for de novo peptide sequencing. Although the ion statistics in individual MS^3 spectra are usually too weak to even allow for reliable identification via database searching, the combination of multiple MS^3 spectra with the parent MS^2 spectrum results in higher signal-to-noise ratios and much improved separability of b/y-ions. By combining the corroborating fragmentation in multiple spectra our approach leads to increased accuracy with increasing numbers of usable MS^3 spectra and even achieves almost error-free sequencing as soon as 4 usable MS^3 spectra are available. Also, while the approaches described in the following chapters increase the sequencing accuracy by combining multiple MS^2 spectra from overlapping peptides [8, 9, 116] , the experimental setup described here requires fewer sample handling steps and should thus be applicable to smaller amounts of substrate.

In general, the major reason why database search is more accurate than de novo sequencing is that the former only matches each spectrum to a relatively small number of peptides in the database while the latter always searches the space of all possible peptides. This distinction is especially relevant because peptide fragmentation is usually incomplete and generally confounded by noise peaks. Rather than requiring an a priori guess of the set of possible peptides, the simultaneous analysis of sets of spectra from distinct fragments of the same peptide provides an independent experimental bias towards the correct peptide and largely reduces the set of possible high-scoring alternative explanations. Since more spectra lead to increasingly restricted sets of high-scoring peptides, it follows that increasing the number of usable MS^3 spectra should result in yet higher sequencing accuracy. Experimentally, higher fragment mass accuracy would significantly reduce the chances of spurious peak matches and should seamlessly increase the percentage of usable MS^3 spectra. Alternatively, more elaborate statistical models could be used to predict the ion-types of MS^3 precursor ions and eventually allow the utilization of MS^3 spectra from doubly-charged precursor ions. We further note that the approach described here is applicable to peptides of any length and could, in principle, be used in the context of top-down or middle-down (via limited proteolysis) proteomics experiments. However, algorithmic extensions may be necessary to account for MS^3 spectra of internal peptide fragments.

Table 4.4 Statistics of MS^3 precursor ions. In our instrument setup each MS^2 spectrum generated up to 5 MS^3 spectra selected by the top 5 most intense peaks with a mass higher than the MS^2 precursor ion. For peak ranks 1-5, each line shows the percentage of MS^3 spectra with a precursor ion of the indicated type; an MS^3 spectrum is said to be useful if it has a good match to the parent MS^2 spectrum; neutral loss includes loss of H_2O or NH_3 from either b or y-ions. Overall, only a third of all MS^3 spectra was found to be usable for de novo sequencing purposes, with a large proportion of these ($\approx$80%) coming from y-ions in the parent MS^2 spectrum.

	Usable MS^3 spectra				Non-usable MS^3 spectra			
Peak rank	y-ion	b-ion	neutral loss	other	y-ion	b-ion	neutral loss	other
1	48.1%	5.4%	0.2%	0.9%	25.6%	15.5%	1.1%	3.1%
2	36.6%	5.2%	0.1%	0.8%	32.4%	14.9%	2.9%	7.1%
3	28.8%	6.1%	0.4%	1.1%	29.8%	16.6%	6.4%	10.8%
4	18.3%	6.1%	1.1%	1.5%	29.9%	20.1%	11.0%	12.0%
5	14.7%	5.6%	0.6%	1.1%	27.3%	21.9%	11.6%	17.4%
Overall	29.3%	5.6%	0.6%	1.1%	29.0%	17.8%	6.6%	10.1%

5

Shotgun Protein Sequencing

5.A Introduction

Traditional MS/MS-based protein analysis starts from a specific digestion of a protein into *non-overlapping* (usually tryptic) peptides. The non-specific digestion into *overlapping* peptides is hardly ever used in MS/MS studies and the common perception is that non-specific digestion only complicates the already difficult protein identification problem and should thus be avoided. However, in a pioneering experiment back in 1989 Hopper et al. [59] took advantage of spectra from overlapping peptides to *de-novo* sequence a whole protein from the rabbit bone marrow. Today, it is relatively easy to run experiments where the proteins are separately digested with different enzymes such as trypsin and pepsin, resulting in the acquisition of MS/MS data from more, partially or completely overlapping (i.e. identical) peptides from the proteins in the sample. This type of data has a clear parallel with the type of data obtained in whole genome sequencing where overlapping DNA reads were collected and assembled into whole genomes. However, it is not clear how to take advantage of overlapping spectra in MS/MS analysis and in the 15 years since the Hopper et al. paper [59] there was no attempt to assemble uninterpreted spectra from overlapping peptides. In this chapter, we show that MS/MS spectra assembly is feasible and demonstrate that it leads to a highly accurate approach to *de-novo* sequencing of entire proteins.

The feasibility of generating and the benefits of using rich peptide ladders were demonstrated in two different contexts. Woods and co-workers [18, 34, 51–53, 101, 102, 141] demonstrated that rich peptide ladders can be generated by non-specific proteolytic digestion in the context of hydrogen exchange (DXMS) studies of protein structure. Also, MacCoss et al. [89] recognized the potential of non-specific proteolytic digestion in improving the procedures for database search of post-translationally modified proteins. In the latter, the richer set of peptides generates enough MS/MS spectra from non-modified peptides to create a smaller protein sequence database that is then searched for post-translational modifications. Promising results were presented but the methodology faces difficulties in that it depends on having at least one (or more for reliable identification) good MS/MS spectrum from an unmodified peptide to first identify the protein in the sample. Moreover, there is a delicate balance between choosing too many protein candidates or choosing less candidates but taking the risk of not including the correct protein sequence in the subsequent search for post-translational modifications. Neither of these approaches attempts to assemble non-interpreted MS/MS spectra.

In this pilot experiment, we capitalized on the principle *'Pairwise alignments whisper while multiple alignments shout out loud'* that was well explored in genomics but had not been applied to MS/MS studies. Our approach provides a *proof of concept* by showing that the de novo interpretation of unknown protein sequences can be significantly improved by detecting overlaps between uninterpreted MS/MS spectra and used to increase the quality and extent of de novo interpretations. By making absolutely no use of any database information we avoid the pitfalls of current methods in that we do not require prior knowledge of the protein sequence and do not face the same exponential growth in running time when considering post-translational modifications. Experimental results are provided using a data set of 2646 alpha-synuclein MS/MS spectra, 303 of which were identified as 83 overlapping alpha-synuclein peptides. This proof of concept further shows the potential of using non-specific digestion enzymes in proteomics experiments and motivates the collection of more and larger such data sets.

Similarly to the *overlap→layout→consensus* approach in DNA fragment assembly, we propose a *alignment→layout→de-novo* interpretation approach for MS/MS analysis (Figures 5.1 and 5.2). In the *alignment* stage (Figure 5.1), we address the pairwise alignment of PRM spectra to detect overlaps and describe the construction of the overlap graphs. Our *assembly* stage uses the overlap graph to assemble spectra and subsequently determine the best amino acid sequence (Figure 5.2).

5.B Aligning MS/MS spectra

The purpose of PRM spectrum alignment is to determine how much overlap, if any, exists between two peptides given only two uninterpreted PRM spectra, one from each peptide. When a large overlap exists then there is some shift of the PRMs in one spectrum such that these match the PRMs in the other spectrum and the sum of the scores of the matched PRMs is high (Figure 5.1).

Every shift λ between two PRM spectra P and Q defines a partial overlap region with a corresponding set of matching PRMs ($P \cap Q^{\xrightarrow{\lambda}}$). As such, scoring a shift is almost the same as scoring full spectrum matches (as described in chapter 3). The only difference is that the requirement to exclude complementary PRMs can now be dropped because these are not expected to match simultaneously in partial overlaps. Thus, in this context, it suffices to compute a maximum weight sparse subset Y of ($P \cap Q^{\xrightarrow{\lambda}}$) and set the shift score to $w(Y)$. Moreover, due to the inherent symmetry of MS/MS and PRM spectra, every shift λ has a *symmetric* shift λ_S with exactly the same score; in correct alignments one of the shifts matches the prefix masses and its symmetric shift matches the suffix masses (Figure 5.1a,b). The center of symmetry when aligning P to Q is given by $c = \frac{m(P)-m(Q)}{2}$ and, as such, any shift λ has a symmetric shift λ_S given by $\lambda_S = 2c - \lambda$. Therefore, the best alignment between two PRM spectra is now defined not by a *single* shift but by a *pair* of symmetric shifts (λ, λ_S)

Overlap graph. The best alignments between PRM spectra define a directed *overlap* graph where each vertex corresponds to a PRM spectrum and each edge corresponds to a shift between two PRM spectra. Only the highest scoring alignment is used to define edges between two vertices and edge directionality is used to represent the sign of the shifts: a positive shift λ from P to Q defines an edge (P, Q) and a negative shift λ' from P to Q defines an edge (Q, P). Every edge $e = (P, Q)$ is characterized by $\lambda(e)$ (the shift between P and Q) and $w(e)$ (the shift score as defined above). Figure 5.3 shows an example of an overlap graph on three imaginary PRM spectra from the peptides listed in the vertices. Note that blue edges represent the shifts where prefix masses match and red edges represent the symmetric shifts where suffix masses match.

Filtering edges in the overlap graph. Since we compute alignments for all pairs of PRM spectra and every such pair will have some best symmetric shift pair we are bound to have many incorrect pairwise alignments that need to be filtered out. We address this issue by building on the principle that a correct alignment should match most of the high scoring PRMs in the overlap region and define a quality score β as the ratio between the matched and unmatched PRM scores.

Given a pair of PRM spectra P and Q for which the best alignment is (λ, λ_S) let M_λ (M_{λ_S}) be the maximum weight sparse subset of $P \cap Q^{\overset{\lambda}{\rightarrow}}$ $(P \cap Q^{\overset{\lambda_S}{\rightarrow}})$ and let $M = M_\lambda \cup M_{\lambda_S}$. Conversely, let U_P be the set of all the unmatched PRMs in the overlapped regions of P when shifting by λ and λ_S. The quality score is then defined as $\beta_P = \frac{w(M)}{w(U_P)}$ (similarly for β_Q). Figure 5.4 shows the ROC and precision/sensitivity curves obtained by varying a threshold t and selecting edges $e = (P, Q)$ from the overlap graph where both $\beta_P \geq t$ and $\beta_Q \geq t$.

As in the clustering section, the triangle condition is also enforced in the overlap graph but only whenever applicable. A *valid triangle* is defined by three edges $e_{PQ} = (P, Q), e_{QR} = (Q, R), e_{PR} = (P, R)$ if $\lambda(e_{PR}) = \lambda(e_{PQ}) + \lambda(e_{QR})$ and is invalid otherwise. Therefore, *if* an edge in the overlap graph is part of some triangle then either it belongs to at least one valid triangle or it is removed from the overlap graph. If the edge does not belong in some triangle (e.g. a set of only two PRM spectra) then this restriction does not apply.

Pairwise alignments were computed for all 39 PRM spectra obtained from the clustering phase as described in the previous section; results are shown in Table 5.1.

The 114 pairwise alignments define 5 connected components in the overlap graph and are the input to the assembly stage of our method.

Table 5.1 Pairwise alignment results; two symmetric shifts per pairwise alignment

	Number of shifts	Number of correct shifts	% correct
Total	$2 \times 39 \times 38 = 2964$	147	5.0%
After filtering	114	114	100%

5.C Assembling MS/MS spectra

The first step in going from an overlap graph to an assembly and interpretation of the partially overlapping spectra is to make the distinction between red and blue edges as illustrated in Figure 5.3. In reality, after building the overlap graph the colors of the edges are unknown.

Decomposing the overlap graph. In any acceptable solution to the assembly problem each vertex in the overlap graph has a unique position on the assembled sequence. The conventional fragment assembly problem assigns a coordinate (e.g. a starting position in the genome) to every read while trying to optimize some target function. Similarly, the MS/MS assembly problem attempts to assign a coordinate to every MS/MS spectrum. The difference is that the coordinate of an MS/MS spectrum from to a peptide starting at position i of a protein $\rho_1, \ldots, \rho_n$ corresponds to the mass of the first $i-1$ amino acids. Thus, let G be an overlap graph and let Ψ be a function (called *vertex potential*) that assign a coordinate to every vertex in G. Figure 5.5 shows an overlap graph with assigned vertex potentials.

An edge $e = (v, u)$ is called *coherent* with respect to a potential Ψ if $\lambda(e) = \Psi(u) - \Psi(v)$. The vertex potentials Ψ in Figure 5.5 define 6 coherent edges of overall weight $w(\Psi)$ given by

$$\begin{aligned} w(\Psi) &= w(A \to B) + w(A \to C) + w(A \to D) + w(B \to C) \\ &\quad + w(B \to D) + w(C \to D) \\ &= 3.0 + 3.5 + 2.9 + 2.4 + 1.8 + 2.2 \end{aligned}$$

Given an overlap graph we are interested in finding a potential Ψ of maximal weight $w(\Psi)$ - *Maximal Coherent Edgeset Problem.* Similarly to the DNA fragment assembly problem, the Maximal Coherent Edgeset Problem is NP-complete. However, the overlap graphs arising in MS/MS assembly are rather small (in contrast to overlap graphs arising in DNA fragment assembly) rendering the MS/MS assembly problem simpler in practice. We construct the potential function Ψ using a greedy algorithm: start with the highest scoring triangle in the overlap graph and iteratively add vertices that increase the overall weight of coherent edges by the maximum amount possible at each step. Although the weight of the coherent edgeset returned by this procedure is not guaranteed to be maximal, it is likely to be so after an adequate threshold is imposed on β to select which edges to retain in the overlap graph.

Once a coherent edgeset E is found, one can construct another coherent edgeset from the edges symmetric to those in E. If the first of these corresponded to the alignment of the prefix masses then the other would correspond to the alignment of the suffix masses (and vice-versa).

Multiple alignment of MS/MS spectra. A set of coherent edges defines a multiple alignment (assembly) of PRM spectra. In this context, a multiple alignment is thus defined as a pair $A = (\mathcal{P}, \Psi)$ where $\mathcal{P} = \{P_1, \ldots, P_n\}$ denotes a set of PRM spectra and $\Psi = \{\psi_1, \ldots, \psi_n\}$ denotes the positions (potentials) of the PRM spectra. Then, scoring the individual PRMs over a multiple alignment A is very similar to what was described before for scoring over a cluster - for a putative PRM t, score it in each overlapped spectrum and set its consensus score $w(t, A)$ to the sum of the obtained per-spectrum PRM scores. The only difference is that in this case there are different starting positions Ψ that need to be taken into account; scoring a PRM t over a multiple alignment $A = (\{P_1, \ldots, P_n\}, \{\psi_1, \ldots, \psi_n\})$ then becomes

$$w(t, A) = \sum_{t-\psi_i \in P_i} w(t - \psi_i, P_i)$$

Thus, a consensus PRM spectrum P for a multiple alignment A could be defined as the set $P = \{t : w(t, A) > 0\}$ - the set of all PRMs with a positive summed score. Although this is a reasonable first approach, it is sometimes the case that MS/MS peaks corresponding to neutral losses also generate PRMs in the aligned PRM spectra and could, therefore, also generate PRMs in the consensus PRM spectrum. This effect is minimized by requiring a minimum 57 Da distance between PRMs in a consensus (sparse subset)

> A *consensus* PRM spectrum for a multiple alignment A is a sparse subset of P, where $P = \{t : w(t, A) > 0\}$ is the set of all PRMs with a positive score over A.

Avoiding double-counting in the consensus PRM spectrum. In a correct multiple alignment of ideal PRM spectra it would be likely (although not certain) that only same-type PRMs would match, i.e. either all matched PRMs would be prefix masses or all would be suffix masses. In reality, this is not the case due to two reasons. The first is that MS/MS spectrum peaks from neutral losses may also generate PRMs in a PRM spectrum - this was already minimized above by the definition of consensus as a sparse set. The second reason is that sometimes, due to random chance or local similarities, complementary PRMs will both match other PRMs in a multiple alignment – if both were used in the consensus spectrum, the same MS/MS peaks would be counted twice (as described in section 3.B, complementary PRMs are generated and scored by the same MS/MS spectrum peaks). To avoid this, let $\oplus$ be the positive score of a correct PRM and let $\ominus$ be the negative score of a PRM with no supporting MS/MS spectrum peaks. We define an *orientation* for every pair of complementary PRMs (p, q) as

Orientation $\oplus/\ominus$ when p is the prefix mass and q is its complementary PRM; represented as

$$p \oplus\!\!-\!\!\!-\!\!\!-\!\!\ominus\ q$$

Orientation $\ominus/\oplus$ when q is the prefix mass and p is its complementary PRM; represented as

$$p \ominus\!\!-\!\!\!-\!\!\!-\!\!\oplus\ q$$

The score of the consensus PRMs is computed in exactly the same way – the sum of the matched PRM scores. Figure 5.6 illustrates this for a pair of PRM spectra with given complementary PRM orientations and considering $\oplus = +1$, $\ominus = -2$. The observed weights of the putative PRMs define a consensus PRM spectrum C with PRMs at positions $\{a, c\}$ and $w(C) = w(a) + w(c) = 4$.

Our problem then becomes: given a multiple alignment find a set of complementary PRM orientations that yield a consensus PRM spectrum of maximal weight – *Maximal Oriented Consensus Problem.* In practice, this problem can be solved using a greedy approach – consider a multiple alignment A with unknown complementary PRM orientations and assume that the score of non-oriented PRMs is given by $\oplus$. Then proceed as follows:

1. Select the putative PRM t with the highest aggregate score $w(t, A)$ (as described above) and assign it to the consensus PRM spectrum.

2. Mark the PRMs matching t in the multiple alignment as $\oplus$ and their corresponding complementary PRMs as $\ominus$.

3. Repeat from 1 until all aggregate scores are negative.

Step 2 above guarantees that there is no double counting of MS/MS spectrum peaks – whenever a PRM is selected as part of the consensus PRM spectrum its complementary PRM is marked as $\ominus$ and thus will not contribute positively to the score of any other consensus PRM.

The preferential match of same-type PRMs (all prefix or all suffix) in a multiple alignment leads to the selective retention of same-type PRMs (as can be seen in Figures 5.8 and 5.7) – the mean percentage of PRM spectrum scores assigned to same-type PRMs is 95%. Even more interesting, these PRMs tend to form very clear ladders where the sequential mass differences correspond to amino acid masses, turning the *de-novo* interpretation of the PRM spectra into a simple problem.

5.D Results

We evaluated this sequencing approach in a pilot study of a sample containing purified alpha-synuclein. The sample was digested using pepsin and a total of 2646 MS/MS spectra with precursor charge +2/+3 were obtained using an ESI/IonTrap mass spectrometer. After parent mass correction and precursor charge selection we retained 1748 of these MS/MS spectra, all believed to be of charge +2. We chose not to include MS/MS spectra with precursor charge +3 due to the high percentage of charge +2 ion types (e.g. b^2) which would not align with the predominant single charge ion types in MS/MS spectra with parent charge +2.

In the absence of an expert-annotated dataset we annotated the MS/MS spectra using `Sequest`. The 1748 MS/MS spectra were searched against a database of 10000 protein sequences randomly selected from the NCBInr database plus our sequence for alpha-synuclein. `Sequest` was configured to allow for a peptide precursor mass tolerance of 2 Da, spectrum peak tolerance of 0.5 Da and for non-specific enzymatic digestion. This procedure identified 303 MS/MS spectra as peptides from alpha-synuclein (by considering the top peptide assignment only), which corresponds to a rate of 17% positive IDs or correct spectra. Once again we stress that these annotations and database search results are *not* used by our method in any way; these are used only to evaluate the quality of the results. This annotation strategy is of course biased towards what `Sequest` could do on a set of MS/MS spectra from a non-trypsin specific digestion, but, in absence of an adequate and curated data set, it is a reasonable approximation to the true performance of our method.

In the clustering phase, match scores were computed over the set of 1748 spectra for every pair of PRM spectra with an absolute parent mass difference not larger than 2 Da. Around 83% of the spectra were matched to at least one other spectrum resulting in 236 spectra being retained in the obtained 39 clusters (the remaining spectra did not meet the clustering criteria):

Clusters	**Number of spectra**	**Spectra from alpha-synuclein**	**% correct**
All	236	183	77.5%
Top 29 clusters	201	183	91.0%
Bottom 10 clusters	35	0	0%

As can be seen form the table above, most of the 'incorrect' spectra retained were concentrated in 10 small clusters which were later ignored in the alignment and assembly phases - the consensus PRM spectra obtained from these 10 clusters did not align to any other PRM spectrum.

The 39 clusters obtained from the clustering phase produced 39 consensus PRM spectra which were then aligned using our pairwise alignment procedure as described in section 5.B. After adequately thresholding the alignment quality score

β we retained 114 relative shifts, all from correct alignments between alpha-synuclein spectra. These pairwise alignments defined 5 connected components in the overlap graph with consensus spectra and interpretations as shown in Figures 5.7 and 5.8. As shown in these, we were able to accurately recover large portions of the overlapped peptide regions. Another major advantage of our approach is also shown – the differences between the shifts (right-pointing triangles) and the parent masses (left-pointing triangles) of the aligned PRM spectra also correspond to amino acid masses. This fact allows us to reconstruct the amino acid sequences near the ends of the consensus PRM spectra even when absolutely no MS/MS spectrum contains any peaks for these fragments - we call these *end-point sequences*. In 4 out of 5 cases (Figure 5.8) at least one end-point matches an internal PRM in the consensus PRM spectrum, either directly or by looking for PRMs at valid amino acid mass distances. In the single occasion where this is not the case (Figure 5.7), the end-point sequence yields additional peptide sequence information but the orientation is not known - the shift of 71 Da only indicates that the peptide either starts (correct answer) or ends with Alanine.

Figure 5.9 illustrates the position of the retained MS/MS spectra relative to the alpha-synuclein protein sequence. Boxes MA1-MA5 contain spectra participating in multiple alignments and boxes C1-C3 contain spectra that clustered together but did not successfully align to other spectra. The recovered amino acid sequences are shown together in Figure 5.10 – the identified sequence blocks (multiple alignments MA1-MA5 and clusters C1-C3) cover 90% of the whole protein and accurately recover 60% of the whole amino acid content.

The coverage gap near the end of the protein sequence is not caused by our method but rather a consequence of the very low MS/MS spectrum coverage in that specific area, observed even when using Sequest to search the database with the correct protein sequence. This was possibly an area of high enzymatic cleavage by pepsin which did not generate enough MS/MS spectra from peptides covering this and adjacent areas and also, the two Proline amino acids near the center of the gap promote the absence of valuable MS/MS peaks when attempting clustering or alignment in this region.

5.E Discussion

The method presented in this chapter builds on strengths from previous approaches to generate larger and more reliable peptide sequences without requiring an existing database of protein sequences. In our approach, spectra are compared against each other (similar to the comparison of experimental spectra to theoretical spectra in database search) to detect repeated MS/MS spectra from the same peptide which are then used to effectively increase the signal-to-noise ratio. The same principle is also applied to detect partial overlaps between spectra and assemble them into multiple alignments where the evidence for real fragment masses becomes overwhelming when compared to that available in single spectra. Furthermore, the multiple alignments themselves provide additional valuable information in that the endpoints of the aligned spectra must necessarily correspond to inter-residue points in the protein sequence and provide, for the first time, a way to recover sequence information where absolutely no MS/MS spectrum peaks are available. Altogether, we build on the ideas previously applied to DNA sequencing to significantly improve the de novo analysis of amino acid sequences and take it from single peptide sequencing to the level of protein sequencing.

Moreover, our approach is directly applicable to sets of MS/MS spectra from post-translationally modified proteins. Because we make no assumptions on the set of residue masses (other than a minimum residue mass of 57 Da) the same procedure can be used to seamlessly assemble spectra from modified peptides and directly determine the modified protein sequence (see the following chapters). Related work by MacCoss et al. [89] has shown that the analysis of partially overlapping peptides provides valuable evidence towards confirming the presence of post-translational modifications. Along the same lines of reasoning, even when complete protein coverage is not available our method can be used to increase the confidence of *de-novo* interpretations by supporting the peptide sequences reconstruction with several partially overlapping spectra.

From an experimental perspective our approach does not require any new developments or significant changes to the currently known protocols; the single difference is that instead of using only trypsin as a digestion enzyme (for which database search tools are specifically tailored), non-specific enzymes (or sets of enzymes with different specificity) should be used. As Woods and co-workers have shown [34, 52, 102, 141], the generation of rich peptide ladders is feasible and within reach of readily available technology. The fact that our results were produced using data from ESI/IonTrap mass spectrometers further reinforces this point: although higher mass accuracy instruments such as MALDI-qTOF and MALDI-TOFTOF should greatly enhance the quality of the sequence reconstruction they are not required for our method to be applicable.

The major difficulty faced by our method was the quality of the experimental MS/MS spectra. This was circumvented by the application of clustering and filtering techniques but at the cost of reduced protein sequence coverage. The availability of larger datasets generated by adequate experimental protocols would allow us to better estimate both the necessary peptide coverage for complete protein sequencing and rigorous thresholds for statistically significant matches. Also, although the occurrence of long repeats in the protein sequence could be an issue this does not seem to be a frequent event. Most repeated subsequences tend to be very short and completely covered by several longer peptides and thus do not significantly affect our approach.

Chapter 5 is, in part, a reprint of the paper “Shotgun Protein Sequencing by tandem mass spectra assembly” co-authored with Haixu Tang, Vineet Bafna and Pavel Pevzner in Analytical Chemistry vol.76, pp.7721-33. The dissertation author was the primary investigator and author of this paper.

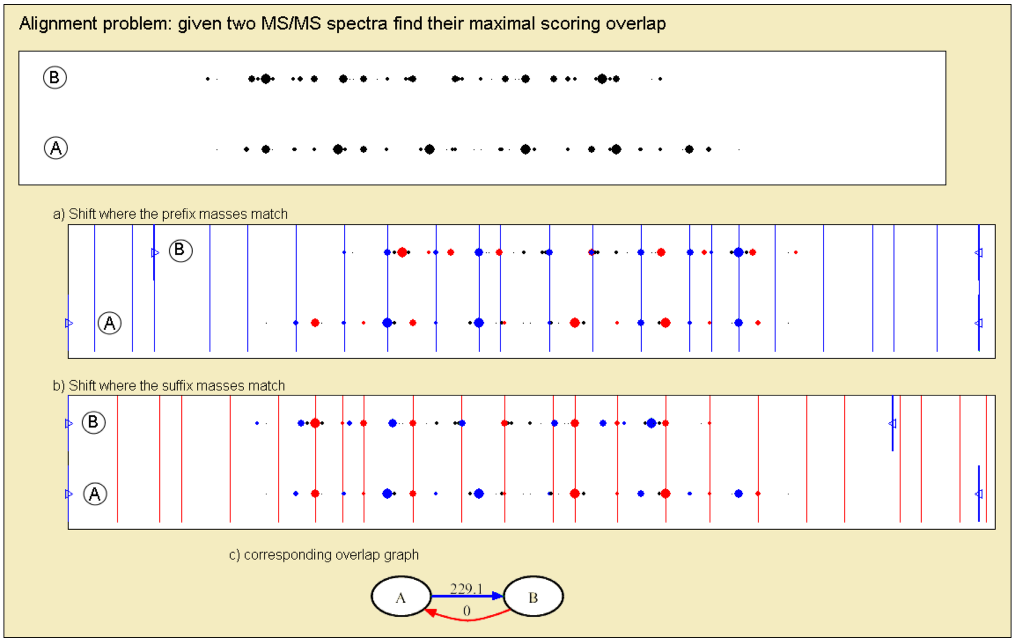

Alignment problem: given two MS/MS spectra find their maximal scoring overlap
B
A
a) Shift where the prefix masses match
B
A
b) Shift where the suffix masses match
B
A
c) corresponding overlap graph
A
229.1
0
B

Figure 5.1 **Pairwise alignment phase**; The two MS/MS spectra used in this example correspond to the peptides ATGFVKKDQLGKNEEGAPQEGIL (spectrum A) and FVKKDQLGKNEEGAPQEGIL (spectrum B) - the peptide themselves are unknown to the algorithm. This example illustrates the case where one peptide is contained in the other but it does not have to be so; our method detects partially overlapping spectra as well. When aligning two PRM spectra we look for a maximal scoring shift between them (score is proportional to the number of matched PRMs). Since PRM spectra are symmetric, we always have two such shifts: **a)** the shift 229.1 (total mass of ATG) in the case when the prefix masses match, **b)** the shift 0 in the case when the suffix masses match (peptide B is a suffix of peptide A). The resulting pairwise alignment is represented as two edges between vertices A and B as shown in **c)** where the shift with matching prefix (suffix) masses is shown in blue (red). This representation is further used in the assembly phase (Figure 5.2).

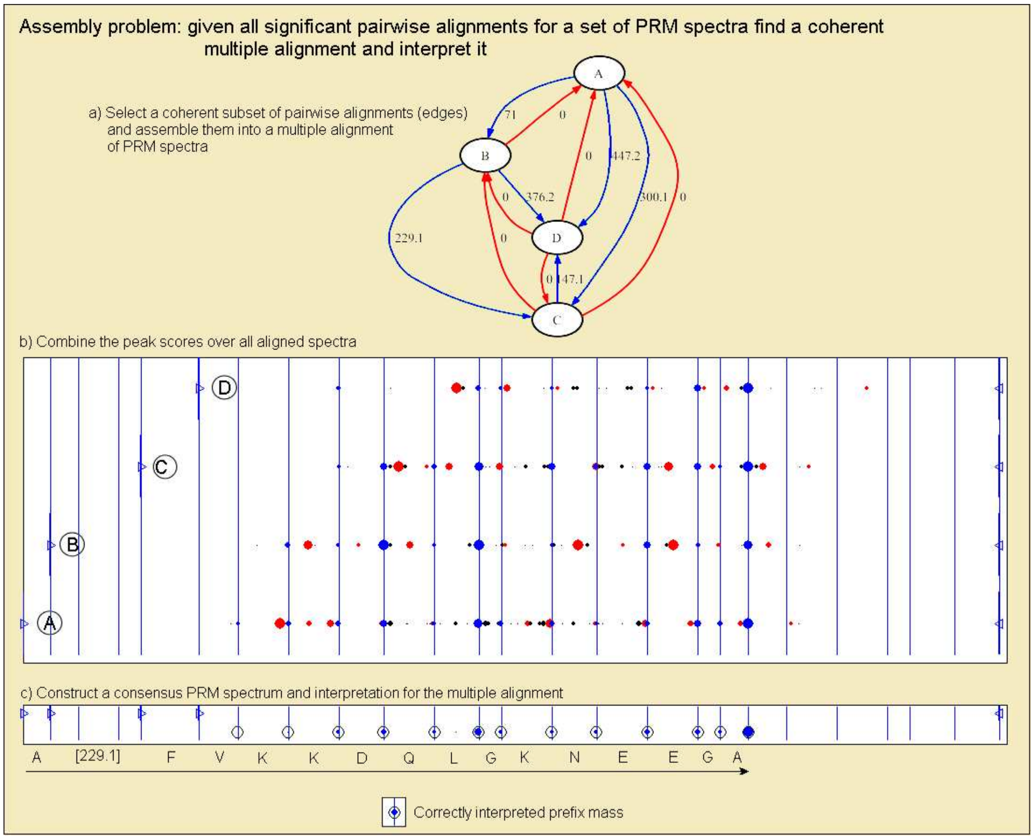

Assembly problem: given all significant pairwise alignments for a set of PRM spectra find a coherent multiple alignment and interpret it
a) Select a coherent subset of pairwise alignments (edges) and assemble them into a multiple alignment of PRM spectra
A
B
C
D
71
0
0
447.2
0
376.2
300.1
0
229.1
0
0
147.1
b) Combine the peak scores over all aligned spectra
D
C
B
A
c) Construct a consensus PRM spectrum and interpretation for the multiple alignment
A [229.1] F V K K D Q L G K N E E G A
Correctly interpreted prefix mass

Figure 5.2 **Assembly phase**; Part **a)** shows the overlap graph constructed for four PRM-spectra where edges represent optimal pairwise alignments. Our assembly algorithm finds the optimal coherent subset of edges that defines the path $A \overset{71}{\to} B \overset{229.1}{\to} C \overset{147.1}{\to} D$. The edges $A \overset{300.1}{\to} C$, $A \overset{447.2}{\to} D$ and $B \overset{376.2}{\to} D$ provide additional support for this path ($300.1 = 71 + 229.1$, $376.2 = 229.1 + 147.1$, $447.2 = 71 + 376.2$) and are thus also included in the selected set of edges (the blue edges). The corresponding multiple alignment shown in **b)** is used to construct the consensus PRM spectrum shown in **c)** and recover the indicated amino acid sequence. *De-novo* interpretation of the assembled MS/MS spectra becomes much simpler because noise was completely removed from the consensus PRM spectrum.

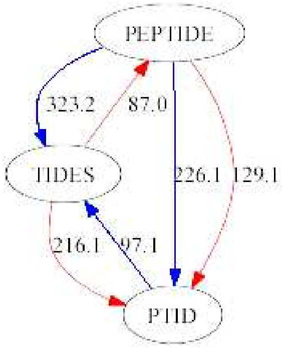


Figure 5.3 Example overlap graph; each vertex represents a PRM spectrum from the listed peptide and edges represent shifts corresponding to the highest scoring alignment (red/blue pairs) between spectra. For example, 323.2 corresponds to the mass of `PEP` while 87.0 corresponds to the mass of `S`.

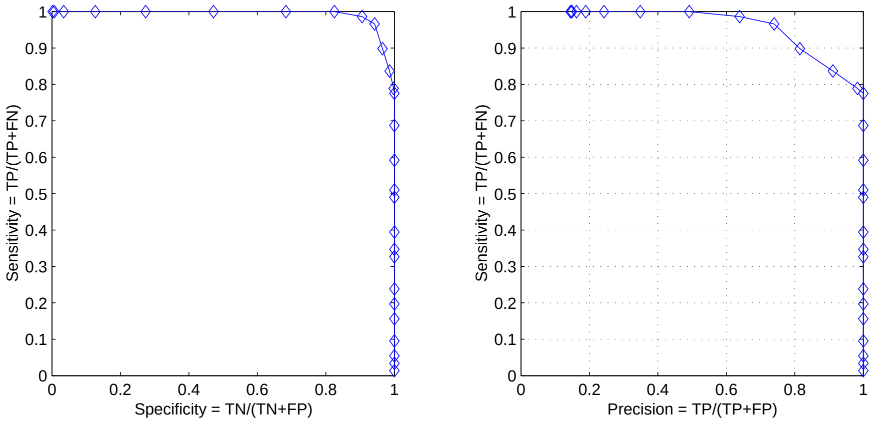


Figure 5.4 ROC curve (left) and precision vs sensitivity (right)

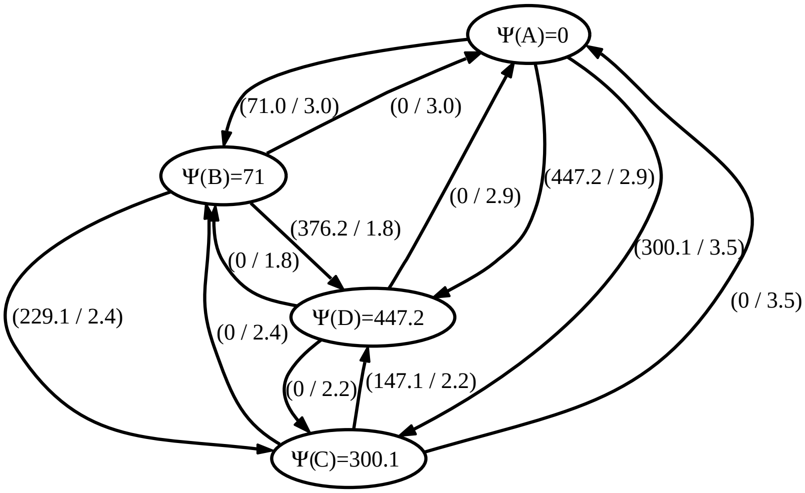
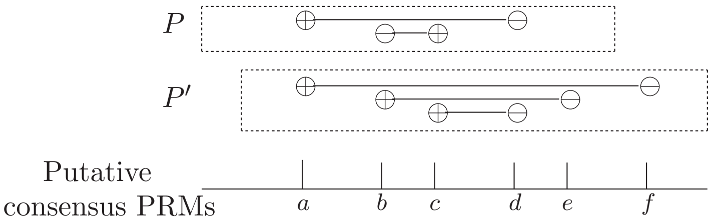


Figure 5.5 Overlap graph from Figure 5.1 with assigned vertex potentials $\Psi(v)$. Edges are labeled by (shift λ, shift score $w(\lambda)$) pairs.



	a	b	c	d	e	f
$w(\cdot)$	+2	-1	+2	-4	-4	-2

Figure 5.6 Putative consensus PRM scores for two aligned PRM spectra P (2 complementary PRM pairs) and P' (3 complementary PRM pairs) with fixed complementary PRM orientations.

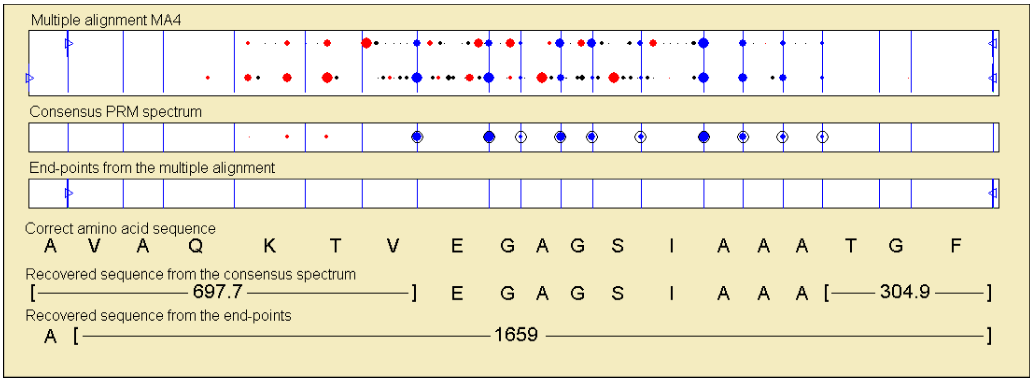

Multiple alignment MA4
Consensus PRM spectrum
End-points from the multiple alignment
Correct amino acid sequence
A V A Q K T V E G A G S I A A A T G F
Recovered sequence from the consensus spectrum
[697.7] E G A G S I A A A [304.9]
Recovered sequence from the end-points
A [1659]

Figure 5.7 Resulting interpretation of the assembled PRM spectra in multiple alignment MA4 (Fig. 5.9). In this case, unlike those shown in Figure 5.8, there is no match between internal PRMs in the consensus spectrum and the endpoints. As such, the endpoints only contribute that there is an `Alanine` either at the start (correct answer) or end of the peptide but not its exact location.

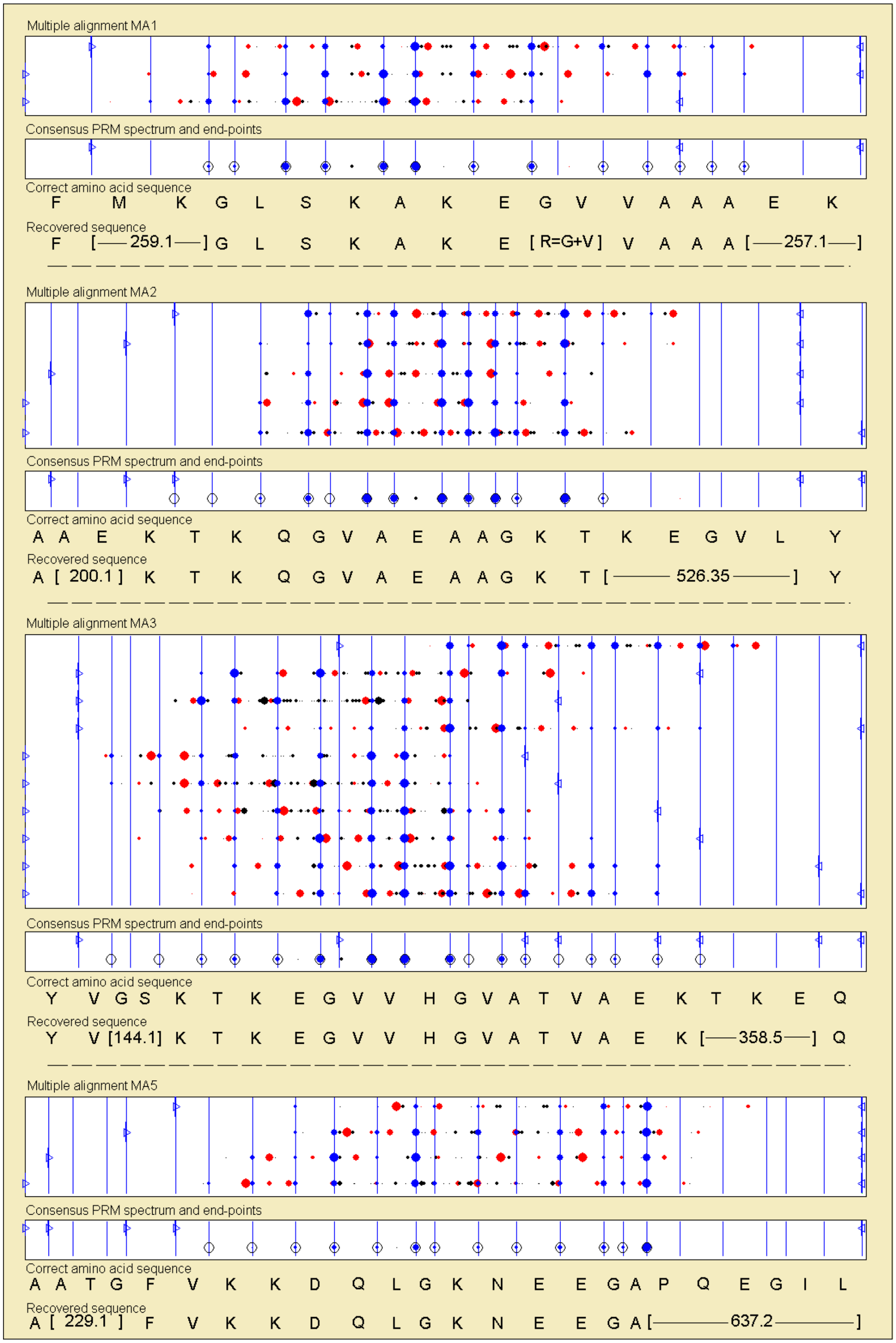


Figure 5.8 Resulting interpretation of the assembled PRM spectra in multiple alignments MA1, MA2, MA3 and MA5.

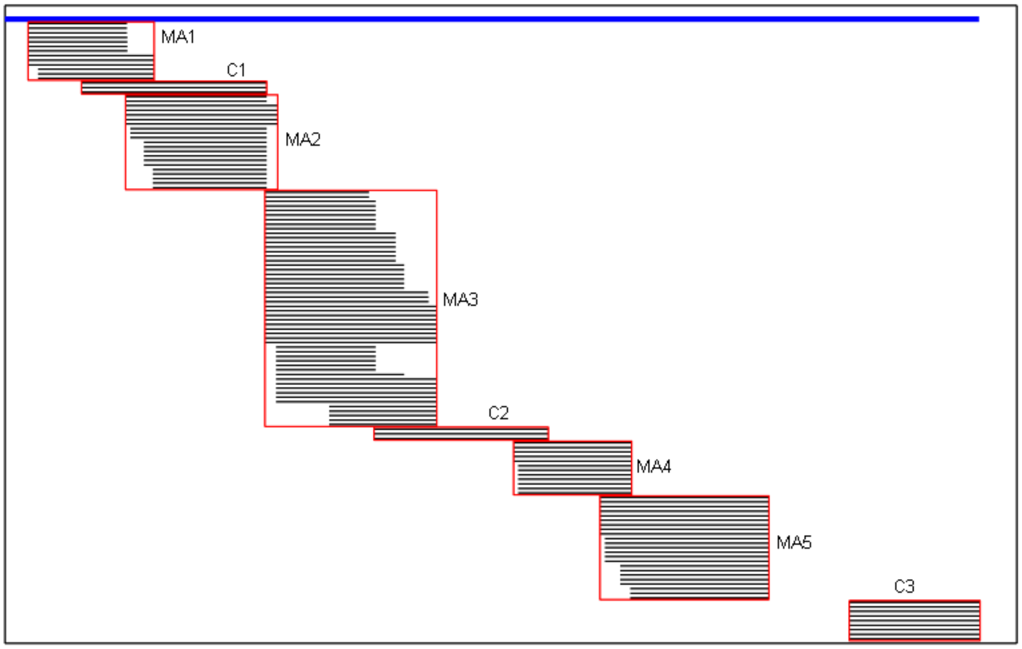


Figure 5.9 Alpha-synuclein spectra clustered (C1-C3) and assembled (MA1-MA5). The blue line at the top represents the complete protein sequence.

```
      MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAV
MA1:     F--GLSKAKE--VAAA--
MA2:                     A--KTKQGVAEAAGKT-----Y
 C1:             ----------EKTKQG--EAA--------
MA3:                                          YV--KTKEGVVHGVATVAEK---Q
C2a:                                                          ------EQV--VGGAV

      VTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA
C2b:  VTD-------
MA4:       A------EGAGSIAAA---
MA5:                         A---FVKKDQLGKNEEGA------
 C3:                                                          -----EEGYQ--E---
```

Figure 5.10 Recovered portions of the alpha-synuclein protein sequence. The recovered sequences are shown under the correct (underlined) protein sequence; dashes indicate portions where the mass intervals are known but the exact amino acid sequences are not. The only incorrectly identified amino acid is shown in red. The interpretation of cluster C2 was split into two lines as C2a and C2b; the full interpretation is the concatenation of these two.

6

Discovery of Modifications using Spectral Networks

Advances in tandem mass-spectrometry (MS/MS) steadily increase the rate of generation of MS/MS spectra. As a result, the existing approaches that compare spectra against databases are already facing a bottleneck, particularly when interpreting spectra of modified peptides. This chapter introduces a new idea that allows one to perform MS/MS database search ... without ever comparing a spectrum against a database. We propose to take advantage of *spectral pairs* - pairs of spectra obtained from overlapping (often non-tryptic) peptides or from unmodified and modified versions of the same peptide. Having a spectrum of a modified peptide paired with a spectrum of an unmodified peptide, allows one to separate the prefix (b-ion) and suffix (y-ion) ladders, to greatly reduce the number of noise peaks, and to generate a small number of peptide reconstructions that are likely to contain the correct one. The MS/MS database search is thus reduced to extremely fast pattern matching (rather than time-consuming matching of spectra against databases). In addition to speed, our approach provides a new paradigm for identifying post-translational modifications via spectral networks analysis.

6.A Introduction

Most protein identifications today are performed by matching spectra against databases using programs like SEQUEST [33] or Mascot [103]. While these tools are invaluable, they are already too slow for matching large MS/MS datasets against large protein databases, particularly when one performs a time-consuming search for post-translational modifications. We argue that new solutions are needed to deal with the stream of data produced by shotgun proteomics projects. Beavis et al [23] and Tanner et al. [130] have developed the X!Tandem and InsPecT algorithms to prune (X!Tandem) and filter (InsPecT) the sequence databases and thus speed-up the search. However, these tools still have to compare every spectrum against a (smaller) database.

In this chapter we explore a new idea that allows one to perform MS/MS database search without ever comparing a spectrum against a database. We propose to take advantage of *spectral pairs* - pairs of spectra obtained from overlapping (often non-tryptic) peptides or from unmodified and modified versions of the same peptide. Most current protocols try to minimize the number of spectral pairs since non-tryptic and chemically modified peptides further complicate the spectral interpretations and lead to higher running times. MacCoss et al. [89] were the first to realize the potential of overlapping peptides for the identification of post-translationally modified proteins and have recently demonstrated the increased throughput of modified digestion schemes on the identification of proteins from complex mixtures (Klammer and MacCoss [77]). Also, even samples digested with trypsin typically have many peptides that differ from each other by a deletion of terminal amino acids (semi-tryptic peptides). In addition, the existing experimental protocols already unintentionally generate many chemical modifications (sodium, potassium, Fe(III), etc.) and it has been shown that existing MS/MS datasets often contain modified versions for many peptides [63, 130, 133, 140].

While seemingly redundant, spectral pairs open up computational avenues that were never explored before. Having a pair of spectra (one of a modified and another of an unmodified peptide) allows one (i) to separate the b (prefix) and y

(suffix) ion mass ladders, (ii) to greatly reduce the number of noise peaks, and (iii) to propagate modification identification from spectrum to spectrum thereby detecting unanticipated and multiple modifications. Thus, spectral pairs allow one to generate a small number of peptide reconstructions that are very likely to contain the correct one. Instead of generating *covering sets* of short 3–4 amino acid *tags* [43, 90, 92, 121, 126, 130], this approach generates a covering set of *peptides* 7–9 amino acids long. This set typically has a single perfect hit in the database that can be instantly found by hashing and thus eliminates the need to ever compare a spectrum against the database.[1] Other approaches [54,83,115,117] that compare *de-novo* peptide sequences against a database of protein sequences obtain their query sequences from *individual* MS/MS spectra (instead of from *spectral pairs*) and thus suffer from relatively low accuracy of *de novo* peptide sequencing [1,7,21,26,88,131]. In addition to improvements in de novo peptide sequencing, spectra de-noising (ii) and propagation of annotations (iii) also improve the standard MS/MS database search.

Let $S(P)$ and $S(P^*)$ be spectra of an unmodified peptide P and of its modified version P^* (spectral pair). The crux of our computational idea is a simple observation that a "database" consisting of a single peptide P is everything one needs to interpret the spectrum $S(P^*)$.[2] Thus, if one knows P there is no need to scan $S(P^*)$ over the database of all proteins! Of course, in reality one does not know P and only $S(P)$ is readily available. Below we show that a spectrum $S(P)$ is nearly as good as the peptide P for interpreting $S(P^*)$ and can thus eliminate the need for database search. This observation opens the possibility of substituting MS/MS database search with finding spectral pairs and further interpreting the peptides that produced them. We show that these problems can be solved using a new combination of de-novo and spectral alignment techniques [21,105] to transform any given spectral pair (S_1, S_2) into virtual spectra $S_{1,2}$ and $S_{2,1}$ of extremely high

[1] We remark that the Peptide Sequence Tag approach reduces the number of considered peptides but does not eliminates the need to match spectra against the *filtered* database. For example, Tanner et al., 2005 [130] describe a dynamic programming approach for matching spectra against a filtered database.

[2] In *blind* database search the list of possible modifications is not known in advance and P suffices to interpret $S(P^*)$ [133]. In *restrictive* database search one also needs the list of possible modifications to interpret $S(P^*)$.

quality; with nearly perfect b and y ion separation and the number of noisy peaks reduced twelvefold. These spectra (albeit virtual) are arguably the highest quality spectra mass-spectrometrists ever saw.

In addition to fast peptide identification, our approach also provides a new paradigm for the identification of chemical and post-translational modifications (PTMs) *without* any use of a database. Recently, Tsur et al., 2005 [133] and Savitski et al., 2006 [114] argued that the phenomenon of modifications is much more widespread than previously thought and advocated blind database search for the identification of these modifications. In particular, blind database search recently resulted in the most comprehensive set of PTMs identified in aged human lenses (Wilmarth et al., 2006 [140]). The surprising conclusion of our approach is that we can discover almost all modifications in cataractous lenses (previously identified via blind database search) and even detect some PTMs missed in [133] and [140].

We further combine spectral pairs into a *spectral network* where each vertex corresponds to a spectrum and each edge to a spectral pair. Figure 6.1 shows a spectral network of 945 MS/MS spectra (corresponding to different peptides from an IKKb protein sample) illustrating the key advantage of spectral networks over the traditional MS/MS database search. Traditional approaches to peptide identification consider each of these spectra separately without attempting to correlate different spectra from related peptides. As a result, the important insights that can be derived from the structure of the spectral network are lost. Our approach consolidates all these spectra into 117 clusters (vertices of the network) and reveals many spectral pairs (edges of the network). This results in the analysis of all spectra “at once” and thus increases the confidence of peptide identifications, reinforces predictions of modifications by using correlated spectra, and eliminates the need to “guess” modifications in advance. Moreover, the spectral network even allows one to assemble these spectra into an intact 34 amino acid long segment of the IKKb protein, thus opening the door toward Shotgun Protein Sequencing [8].

6.B Results

6.B.1 Interpretation of spectral pairs/stars

A set of spectra incident to a spectrum S_1 in the spectral network is called a *spectral star*. For example, the spectral star for the spectrum derived from peptide 3 in Figure 6.1 consists of multiple spectra from five different peptides. The high quality of the virtual spectra derived from spectral pairs and spectral stars makes *de-novo* interpretation of these spectra straightforward (see Table 6.1 and Figure 6.3). Since these spectra feature excellent separation of b and y-ion ladders and only a small number of noise peaks, de-novo reconstructions of these spectra produce reliable (gapped) sequences that usually contain long correct tags.[3] On average, de-novo reconstructions of our consensus spectra correctly identify 72% of all possible “cuts” in a peptide (i.e., on average, $0.72 \cdot (n-1)$ b-ions (y-ions) in a peptide of length n are explained). This is a very high number since the first (e.g., b_1) and the last (e.g., b_{n-1}) b-ions are rarely present in the MS/MS spectra and thus make it nearly impossible to explain more than 80% of all “cuts” in the IKKb sample. Moreover, on average, the recovered peaks account for 95% of the total score of the de-novo reconstruction implying that unexplained peaks usually have very low scores.[4]

Benchmarking in mass-spectrometry is inherently difficult due to a shortage of manually validated large MS/MS samples that represent “gold standards”. While the ISB dataset [75] represents such a gold standard for unmodified peptides, large validated samples of spectra from modified peptides are not currently available. As a compromise, we benchmarked our algorithm using a set of 11,760 spectra from the IKKb dataset that were annotated using InsPecT and extensively studied in recent publications [130,133], including comparisons with SEQUEST, Mascot and X!Tandem. Our entire spectral networks analysis (starting from clustering and end-

[3] We use the standard longest path algorithm to find the highest scoring path (and a set of suboptimal paths) in the *spectrum graph* of spectra $S_{i,j}$ and S_i^* (see [26] for using spectrum graphs in de novo peptide sequencing). In difference from the standard de-novo algorithms we do not insist on reconstructing the entire peptide and often shorten the found path by removing its prefix/suffix if it does not explain any peaks. As a result, the found path does not necessarily start/end at the beginning/end of the peptide.

[4] We realize that our terminology may be confusing since, in reality, it is not known whether a spectrum $S_{i,j}^b$ describes b or y-ions. Therefore, in reality we average between b and y-ion ladders while referring only to b-ions.

ing with interpretations) of this IKKb dataset took 9 minutes on a regular desktop machine (Intel® Pentium® 4, 2.8GHz clock speed)[5]. We compared our performance to that of InsPecT, which was previously shown to be two orders of magnitude faster than SEQUEST for restricted database search [130]. Even when searching against a moderately sized database, such as Swiss-Prot's set of 13,749 human proteins, InsPecT's running time was 55 minutes Thus, our spectral networks approach (that finds both unmodified *and* modified peptides) was 6 times faster than InsPecT (in the mode that only searches for unmodified peptides). Below we give identification results for both spectral pairs and spectral stars.

InsPecT identified 515 unmodified peptides in the IKKb sample, 413 of which have some other prefix/suffix or modified variant in the sample and are thus amenable to pairing. We were able to find spectral pairs for 386 out of these 413 peptides. Moreover, 339 out of these 386 peptides had spectral pairs coming from two (or more) different peptides, i.e., pairs (S_1, S_2) and (S_1, S_3) such that spectra S_2 and S_3 come from different peptides.

The average number of (gapped) de-novo reconstructions (explaining at least 85% of the optimal score) for spectral stars was 10.4. While spectral stars generate a small number of gapped reconstructions, these gapped sequences are not well suited for fast membership queries in the database. We therefore transform every gapped de-novo reconstruction into an ungapped reconstruction by substituting every gap with all possible combinations of amino acids.[6] On average, it results in 165 sequences of length 9.5 per spectrum – for 86% of all peptides, one of these tags is correct.

While checking the membership queries for 165 sequences can be done very quickly with database indexing (at most one of these sequences is expected to be present in the database), there are no particular advantages in using such super-long

[5] While clustering of IKKb and Lens dataset is a simple task, the problem may appear to be extremely time-consuming for larger MS/MS datasets. We emphasize that clustering of MS/MS datasets with millions of spectra is not unlike clustering of Internet pages (with every peak corresponding to a keyword from a dictionary). It does not require "all-against-all" spectra comparison and becomes very fast with a variation of the Locality-Sensitive Hashing approach developed in the context of Internet clustering [22, 57] (Dutta and Chen, 2006 [31] pioneered applications of these ideas to MS/MS clustering).

[6] In rare cases the number of continuous sequences becomes too large. In such cases we limit the number of reconstructions to 500.

tags (9.5 amino acids on average) for standard database search: a tag of length 6-7 will also typically have an unique hit in the database. However, the long 9-10 amino acid tags have distinct advantages in difficult non-standard database searches, e.g., discovery of new alternatively spliced variants via MS/MS analysis. Moreover, for standard search one can generate a smaller set of shorter (6-7 amino acids) tags based on the original gapped reconstruction and use them for membership queries. We used the obtained gapped reconstruction to generate such short 6-mer tags. On average, each consensus spectrum generates about 50 6-mer tags. It turned out that 82% of spectra derived from spectral stars contain at least one correct 6-mer tag.

6.B.2 Using spectral networks for PTM identification

Our approach, for the first time, allows one to detect modifications without any reference to a database. The difference in parent masses within a spectral pair corresponds either to a modification offset or to a sum of amino acid masses. While not every difference in parent mass corresponds to a modification offset (some spectral pairs may be artifacts), the histogram of parent mass differences (Figure 6.4a) reveals the modifications present in the IKKb sample. Indeed, 7 out of the 8 most frequent parent mass differences in Figure 6.4a are listed among the 8 most common modifications in the IKKb dataset [133]. We emphasize that Figure 6.4a was obtained without any reference to a database while Tsur et al., 2005 [133] found these modifications via database search. The only frequent modification identified by Tsur et al., 2005 [133] and not represented in Figure 6.4a is deamidation with a small mass offset of 1 Da that is difficult to distinguish from parent mass errors and isotopic peaks artifacts. Interestingly enough, our approach reveals an offset of +34 (present in thousands of spectral pairs) that was not mentioned by Tsur et al., 2005 [133].

Additionally, spectral networks can also make a contribution for the detection of rare modifications. These modifications usually occur on only a very small number of peptides and are thus unlikely to be detected by the PTM frequency matrix approach from [133]. Furthermore, these can co-occur with other more frequent modifications and thus completely escape identification. We addressed these cases by focusing on *modification networks* – subnetworks of the spectral network connecting multiple modification states of the same peptide.

We illustrate our modification networks approach to PTM identification using the Lens dataset. As an initial preprocessing step we removed low quality quality spectra using an approach similar to Bern et al., 2004 [16]. By applying our clustering procedure to the remaining spectra, we identified 938 clusters (including 6319 spectra) and obtained a combined dataset of 11,932 spectra (938 consensus spectra from the clusters and 10,994 non-clustered spectra). Out of these, 2001 spectra were found to be paired, resulting in the identification of 280 unmodified peptides. (88% of all unmodified peptides that have some pair in the dataset).

Although at a first glance the number of annotations (280) may seem small when compared to the number of paired spectra (2001), it should be noted that many of these paired spectra come from modified peptides and thus may not generate long enough tags to match the correct peptide in the database. However, most spectra from modified peptides were correctly paired with their unmodified counterpart and were thus already linked to the correct peptide. Additionally, as described in A, the spectral alignment between any two spectra promptly provides both the location and mass of the modification. Thus, suppose that an identified spectrum S was annotated with a peptide $p_1 \ldots p_n$ and paired with a non-annotated spectrum S'. Using our spectral alignment approach we can derive on which amino acid p_i the modification occurred and readily annotate S' with $p_1 \ldots p_{i-1} p_i^* p_{i+1} \ldots p_n$, where p_i^* stands for a modification of p_i (see Figure 6.2e). This operation is defined as the *propagation* of a peptide annotation via spectral pairs. In order to use propagation on any given spectral network we need to consider two additional conditions: (i) some non-annotated spectra may not be directly connected to an annotated spectrum (e.g. spectra with two modifications) and (ii) some non-annotated spectra may be connected to multiple annotated spectra (e.g. different prefix/suffix variants). We therefore use an iterative procedure that, at each step, propagates peptide annotations from every annotated spectrum onto all its non-annotated neighbors. If a non-annotated spectrum happens to gain more than one putative annotation then we simply choose that which best explains the spectrum. The neighbors are then marked as annotated and are allowed to propagate their annotations on the next iteration. This procedure stops when there are no more annotated spectra paired with non-annotated spectra. For example, the propagation procedure starts from 58 (out of 117) annotations of unmodified peptides in the spectral network shown in Figure 6.1, adds 53 annotations with a single modification on the next iteration, and finally adds 6 annotations with 2 modifications on the final iteration. Figure 6.4 illustrates this iterative propagation on the Lens dataset with the modification network for peptide MDVTIQHPWFK. We remark that the existing peptide identification tools have difficulties identifying and validating peptides with multiple modifications. Modification networks open up the possibility of reliably identifying such heavily modified peptides (which may be common in heavily modified proteins involved in cell signalling like the IKK complex) via cross-validation with other modified peptides as exemplified in Fig 6.4.

Overall, the spectral networks analysis of the Lens dataset found all but one of the modification types previously identified by blind database search (see Table 6.2) and provided evidence for 6 previously undetected modifications types (see Table 6.3). The only modification listed by Tsur et al. [133] and not rediscovered using spectral networks was again deamidation on N,Q, due to the same reasons described above for the IKKb dataset.

Two out of the six new putative modifications were independently identified in cataractous lenses by other groups [128,137] thus reinforcing our predictions. One more modification was previously reported as a loss of methane sulfenic acid on the same site [79]. The newly found N-terminal modification with an offset of 57 Da is potentially interesting: it occurs only on two semi-tryptic peptides whose non-tryptic ends were previously reported as degraded N-terminii of betaB1-crystallins [28] thus also reinforcing our predictions. Moreover, since all protein N-terminii are expected to be acetylated (as it is mostly observed) this could hypothetically correspond to a previously undetected in-vivo modification of the degraded N-terminii. Note that this 57 Da offset would normally be attributed to a common experimental artifact caused by the cysteine alkylation [24]. However, the fact that this 57 Da is not observed on any other peptides and the lack of corroborating peptide fragmentation evidence (i.e. characteristic loss of 57 Da from precursor mass) suggest that this is a localized event that could warrant further investigation. As an alternative confirmation step, we modified the traditional database search parameters to consider all newly discovered putative modifications and observed a complete agreement with our proposed annotations with large X-corr and ΔCn scores.

It should be noted that all of these new putative modification types occur on peptides that had been previously identified in this dataset [133, 140]. However, most of these modifications are rare in that they occur only at specific sites and thus tend to have low spectral counts – the major reason why these are hard to detect through blind database search. By independently comparing each MS/MS spectrum against a database, blind database search generates many false positives that are usually filtered by requiring a minimum number of occurrences of each modification. While a successful approach in detecting multiple-site modifications this leads to difficulties in the detection of single-site and less-common modifications.

The spectral networks approach remedies this limitation of blind database search by being more selective on the assignment of modified peptide annotations. Spectral pairs provide additional evidence that two spectra were derived from the same peptide (in the form of correlated ion peaks and intensities) and thus add significance to otherwise difficult spectrum identifications. As illustrated in Figure 6.4, this increased sensitivity is particularly evidenced on modification networks by the grouping of multiple spectra from different modification states of the same peptide.

6.C Methods

6.C.1 Datasets

We describe our algorithm using MS/MS spectra from human IKKb and lens proteins, two particularly challenging samples for PTM analysis. The IKKb dataset consists of 45,500 spectra acquired from a digestion of the inhibitor of nuclear factor kappa B kinase beta subunit (IKKb) protein by multiple proteases, thereby producing overlapping peptides (spectra were acquired on a ThermoFinnigan LTQ mass spectrometer). The activation of the inhibitor kappaB kinase (IKK) complex and its relationships to insulin resistance were the subject of recent intensive studies [5, 20]. This complex represents an ideal test case for algorithms that search for post-translationally modified (PTM) peptides. Until recently, phosphorylations were the only known PTMs in IKK, insufficient to explain all mechanisms of signaling and activation/inactivation of IKK by over 200 different stimuli, including cytokines, chemicals, ionization and UV radiation, oxidative stress, etc. It is likely that different stimuli use different mechanisms of signaling involving different PTM sites. Revealing the combinatorial code responsible for PTM-controlled signalling in IKK remains an open problem.

The IKKb dataset was studied in Tanner et al., 2005 [130] and Tsur et al., 2005 [133] resulting in 11,760 identified spectra and 1154 annotated peptides. This IKKb sample presents an excellent test case for our protocol since 77% of all peptides in this sample have spectral pairs.

The Lens dataset [115] consists of 27,154 MS/MS spectra from a trypsin digestion of lenses from a 93-year-old male (spectra were obtained on a ThermoFinnigan LCQ Classic ion trap mass spectrometer). Lens proteins, due to a very low turnover, tend to accumulate many post-translational modifications over time and often result in increased opaqueness and cataracts [115, 133]. This dataset was extensively studied in [115, 133] and peptide identifications were subjected to manual validation in Wilmarth et al., 2006 [140], resulting in the identification of 416 unmodified peptides and 450 modified peptides. Furthermore, 318 unmodified peptides had spectral pairs and 343 modified peptides had an unmodified version in the sample.

6.C.2 Clustering spectra

The clustering approach used here was as described in chapter 3 with some improvements outlined below.

While spectral similarity via optimal matching largely succeeds in identifying related spectra, it may sometimes pair non-related spectra. Although such false pairings are rare, they may cause problems if they connect two unrelated clusters. To remove false pairs we use a heuristic approach proposed by Ben-Dor et al. [14]. On the IKKb dataset, this clustering procedure resulted in 567 clusters representing 98% of all unmodified and 96% of all modified peptides with three or more spectra in the sample.

Each cluster of spectra is then collapsed into a single *consensus spectrum* that contains peaks present in at least k spectra in the cluster. The parameter k is chosen in such a way that the probability of seeing a peak in k spectra by chance is below 0.01. We model the noise peak generation as a Bernoulli trial and the occurrence of k matching peaks in a cluster of n spectra as random variable with a Binomial distribution. We further sum up the scores of matching peaks to score the peaks in the consensus spectrum. As shown in Table 6.1, the resulting consensus spectra have unusually high signal-to-noise ratio (the number of unexplained peaks in the consensus spectra is reduced by a factor of 2.5). We also observed some consistently co-occurring unexplained peaks possibly due to co-eluting peptides or unexplained fragment ions (e.g., internal ions). After clustering we end up with 567 consensus spectra (that cover 93% of all individual spectra) and 862 unclustered spectra.

6.C.3 Spectral pairs

Peptides P_1 and P_2 form a *peptide pair* if either (i) P_1 differs from P_2 by a single modification/mutation, or (ii) P_1 is either a prefix or suffix of P_2.[7] Two spectra form a *spectral pair* if their corresponding peptides are paired. Although the peptides that give rise to a spectral pair are not known in advance, we show below that spectral pairs can be detected with high confidence using uninterpreted spectra.

Our approach for detecting spectral pairs is similar in spirit to the blind search for modified peptides first described by Pevzner et al. [105,106] and further developed by Tsur et al. [133]. Hansen et al. [55] and Tang et al. [129] have alternatively proposed enumeration- and preindexing-based approaches to blind database search and Savitski et al. [114] recently complemented blind database search by taking into account the retention time. It should be noted that the retention time analysis im-

[7] Condition (ii) can be viewed as a variation of (i) if one considers extending a peptide by a few residues as a single "mutation" (such variations are common in MS/MS samples). More generally, peptides P_1 and P_2 form a peptide pair if either (i) P_1 is a modified/mutated version of P_2, or (ii) P_1 and P_2 overlap. While our techniques also work for this generalization, we decided to limit our analysis to simple peptide pairs described above. We found that such simple pairs alone allow one to interpret most spectra. Adding pairs of spectra with more subtle similarities further increases the number of spectral pairs but slows down the algorithm.

poses the constraint that both spectra must come from the same sample while our approach seamlessly enables detection of spectral pairs from multiple MS/MS sample runs (e.g., different cell states or diseased/healthy tissue samples).

For two spectra S_1 and S_2, the *spectral product* of S_1 and S_2 is the set of points (x, y) in 2-D for every $x \in S_1$ and $y \in S_2$ (S_1 and S_2 are represented as sets of masses) Figure 6.2a shows the spectral product for the theoretical spectra of two peptides. The similarity between the two spectra is revealed by two diagonals in the spectral product: one is formed by matching b-ions (blue) and another one is formed by matching y-ions (red).

Figures 6.2b,d show pairs of uninterpreted spectra, denoted S_1 and S_2, and their spectral product. Although the "colors" of the peaks are not known in this case, we still take the liberty of naming one diagonal blue and the other red. One can use circles (matching peak masses) on the blue diagonal to transform the original spectrum S_1 into a new spectrum $S^b_{1,2}$ (Figure 6.2c) with a much smaller number of peaks (a peak in S_1 is retained in $S^b_{1,2}$ only if it generates a circle on the blue diagonal). Similarly, one can transform S_1 into a spectrum $S^y_{1,2}$ using circles on the red diagonal. The peak scores in both spectra $S^b_{1,2}$ and $S^y_{1,2}$ are inherited from spectrum S_1. Similarly, the spectrum S_2 is transformed into spectra $S^b_{2,1}$ and $S^y_{2,1}$.[8]

Intuitively, if two spectra are unrelated, the blue and red diagonals represent random matches and the number of circles appearing on these diagonals is small. Paired spectra, on the contrary, are expected to have many circles on these diagonals. Although this simple criterion (number of circles on the diagonals) would already allow one to roughly distinguish paired spectra from unrelated spectra, we describe below a more accurate *spectral alignment* test for finding spectral pairs.

Figure 6.2b illustrates case (ii) in the definition of spectral pairs. The situation becomes less transparent in case (i), namely when modification/mutation occurs in the middle of the peptide (Figure 6.2d). In this case both detecting spectral pairs (S_i, S_j) and further processing them into spectra $S^b_{i,j}$ and $S^y_{i,j}$ is rather complicated. In A we describe the anti-symmetric spectral alignment algorithm for

[8] We remark that the assignments of upper indexes to spectra $S^b_{1,i}$ and $S^y_{1,i}$ are arbitrary and it is not known in advance which of these spectra represents b ions and which represents y ions.

deriving virtual spectra $S_{i,j}$ from spectral pairs that also covers this case of internal modifications/mutations.

For the sake of simplicity, the above description hides many details that turn interpretation of spectral pairs into a rather difficult algorithmic problem. For example, the red points on the red diagonal in Figure 6.2d are actually located slightly off the diagonal due to mass measurement errors and their deviation from the diagonal may increase with the mass. More importantly, the original algorithm from Pevzner et al. [105] considered only b-b (or y-y) pairs of matching peaks and was not able to consider all three types of matching peaks (b-b, y-y, and b-y) when computing the spectral alignment. This complication was addressed in Tsur et al. [133] for the case "spectrum vs. peptide" comparison. Here we face a more difficult case of "spectrum vs. spectrum comparison"[9] and take into account the anti-symmetric path condition [21, 26] that further complicates the spectral alignment algorithm (even in the case of a single internal modification).

6.C.4 Spectral networks

The *correlation score* of spectra S_1 and S_2 is defined as the total score of all peaks in spectra $S^b_{1,2}$ and $S^y_{1,2}$: $score(S_1, S_2) = score(S^b_{1,2}) + score(S^y_{1,2})$. Similarly, $score(S_2, S_1) = score(S^b_{2,1}) + score(S^y_{2,1})$. We accept S_1 and S_2 as a putative spectral pair if both the ratio $\frac{score(S_1,S_2)}{score(S_1)}$ and $\frac{score(S_2,S_1)}{score(S_2)}$ exceed a predefined threshold (0.4 in examples below), where $score(S_i)$ is the summed score of all peaks in S_i.

In addition to the correlation score test described above, we also use a test that takes into account the size of the MS/MS sample. The larger the set of spectra under consideration the larger the chance that a certain correlation score can be achieved by chance. To account for this phenomenon we assume that the correlation score between a given spectrum S and any unrelated spectrum S' approximately follows a Gaussian distribution. Thus, a correlation score is only considered significant if the probability of this score appearing by chance is below 0.05. The combined filtering efficiency of these criteria allowed us to retain 78.4% of all correct spectral pairs

[9] In the case "spectrum vs. peptide" one knows the sets of b and y-ions in the theoretical spectrum of the peptide while in the "spectrum vs. spectrum" case this partition is unknown. A similar problem was considered by Zhang and McElvain [145] in case of MS2 and MS3 spectra comparison.

at a precision level of 95% and find several different variants for most unmodified peptides. The main reason why the remaining spectral pairs were not detected by our alignment procedure was the change in fragmentation patterns between these closely related peptides. The prediction of peptide fragmentation patterns is still an active area of research [61, 134] and a comprehensive study of how instrument variability, peptide extensions and modifications affect the observed fragment propensities is beyond the scope of this dissertation.

The spectral pairs that satisfy the tests above form the *spectral network* on the set of all spectra (see Figure 6.1 for an example). The spectral network for the whole IKKb dataset has 43 connected components with 1021 vertices and 1569 edges. The small number of connected components is not surprising since overlapping peptides in this dataset can be assembled into a small number of contigs (an effect explored in the context of shotgun protein sequencing as described in chapters 5 and 7).

Table 6.1 describes the statistics of spectra $S_{i,j}$ and Figure 6.3 shows the dramatic increase in signal-to-noise ratio as compared to consensus spectra (let alone individual spectra). Moreover, spectral pairs provide a nearly perfect separation between b and y-ion ladders, the key condition for successful de novo reconstruction [15]. When compared to EigenMS's [15] average performance on single LTQ MS/MS spectra, spectral pairs reduce the contamination of y-ions in b-ion ladders (and vice-versa) from their reported level of 11% to only 2%.

6.C.5 Spectral stars

Even though for a single spectral pair (S_1, S_2), the spectra $S^b_{1,2}$ and $S^y_{1,2}$ already have high signal-to-noise ratio, below we show that spectral stars allow one to further enrich the b and y-ion ladders (see Table 6.1). A spectral star consisting of spectral pairs (S_1, S_2), (S_1, S_3), $\ldots$, (S_1, S_n) allows one to increase the signal-to-noise ratio by considering $2(n-1)$ spectra $S^b_{1,i}$ and $S^y_{1,i}$ for $2 \leq i \leq n$. We combine all these spectra into a *star spectrum* S^*_1 as in our clustering approach. This needs to be done with caution since spectra $S^b_{1,i}$ and $S^y_{1,i}$ represent separate b and y-ion ladders. Therefore, one of these ladders needs to be reversed to avoid mixing b and

y-ion ladders in the star spectrum. The difficulty is that the assignments of upper indexes to spectra $S^b_{1,i}$ and $S^y_{1,i}$ are arbitrary and it is not known in advance which of these spectra represents b-ions and which represents y-ions (i.e., it may be that $S^b_{1,i}$ represents the y-ion ladder while $S^y_{1,i}$ represents the b-ion ladder).

A similar problem of reversing DNA maps arises in *optical mapping* (Karp and Shamir, 2000 [73], Lee et al., 1998 [80]). It was formalized as the *Binary Flip-Cut* (BFC) Problem [27] where the input is a set of n 0-1 strings (each string represents a snapshot of a DNA molecule with 1s corresponding to restriction sites). The problem is to assign a *flip* or *no-flip* state to each string so that the number of consensus sites is maximized. We found that for the case of spectral stars, a simple greedy approach to the BFC problem works well. In this approach, we arbitrarily select one of the spectra $S^b_{1,i}$ and $S^y_{1,i}$ and denote it $S_{1,i}$. We select $S_{1,2}$ as an initial consensus spectrum. For every other spectrum $S_{1,i}$ $(2 < i \leq n)$, we find whether $S_{1,i}$ or its reversed copy $S^{rev}_{1,i}$ better fits the consensus spectrum. In the former case we add $S_{1,i}$ to the growing consensus, in the latter case we do it with $S^{rev}_{1,i}$.

After solving the BFC problem we know the orientations of all spectra in the spectral star. The final step in constructing a *star spectrum* S^* from the resulting collection of $S_{1,i}$ spectra is identical to the consensus spectrum approach described above for clusters. Table 6.1 illustrates the power of spectral stars for the enrichment of b/y-ion ladders.

6.D Discussion

We have demonstrated the utility of using spectral networks for protein identification. The key idea of this approach is that correlations between MS/MS spectra of modified and unmodified peptides allow one to greatly reduce noise in individual MS/MS spectra and, for the first time, make de-novo interpretations so reliable that they can substitute the time-consuming matching of spectra against databases. We have also shown how the correlated spectral content on modification networks can provide consistent evidence to support the identification of rare modifications and highly modified peptides. A current limitation of our approach is its restricted applicability to spectra with parent charges 1 and 2; two further algorithmic developments are necessary to allow for the integration of spectra with higher parent charges into spectral networks. First, while spectral alignment works for two spectra of precursor charge 3 (or higher), it generally does not work for comparison of a spectrum of precursor charge 1 or 2 to a spectrum of precursor charge 3. The main reason is that spectra of higher precursor charges tend to generate *b*- and *y*-ions of higher charge that do not align to the singly-charged variants predominant in spectra of precursor charge 1 or 2. Second, even if two spectra with parent mass 3 (or higher) are aligned, reliable de novo algorithms for interpreting multi-charged spectra are still unknown.

Tandem mass-spectra are inherently noisy and mass-spectrometrists have long been trying to reduce the noise and achieve reliable de novo interpretations by advancing both instrumentation and experimental protocols. In particular, Zubarev and colleagues [112, 113] recently demonstrated the power of using both CAD and ECD spectra. We emphasize that, in difference from our approach, this technique as well as the recent approach described in Frank et al., 2006 [45] require special instrumentation or highly accurate Fourier transform mass-spectrometry. Another approach to reduce the complexity of spectra involves stable isotope labeling [116]. However, the impact of this approach (for peptide identification) has been restricted, in part by the cost of the isotope and the high mass resolution required. Alternative end-labeling chemical modification approaches have disadvantages such as low yield, complicated reaction conditions, and unpredictable changes in ionization and fragmentation. As a result, the impact of these important techniques is mainly in protein quantification rather than identification [116]. The key difference between our approach and labeling techniques is that, instead of trying to introduce a specific modification in a controlled fashion, we take advantage of multiple modifications naturally present in the sample. Our spectral networks approach allows one to decode these modifications (without knowing in advance what they are) and thus provide a computational (rather than instrumentation-based) solution to the problem of MS/MS spectra identification.

Chapter 6 is, in part, a reprint of the papers “Protein identification by spectral networks analysis" in Proceedings of the National Academy of Sciences USA, vol.104, pp.6140-5 and the paper "Protein identification by spectral networks analysis" in the proceedings of RECOMB 2006, both co-authored with Dekel Tsur, Ari Frank and Pavel Pevzner. The dissertation author was the primary investigator and author of these two papers.

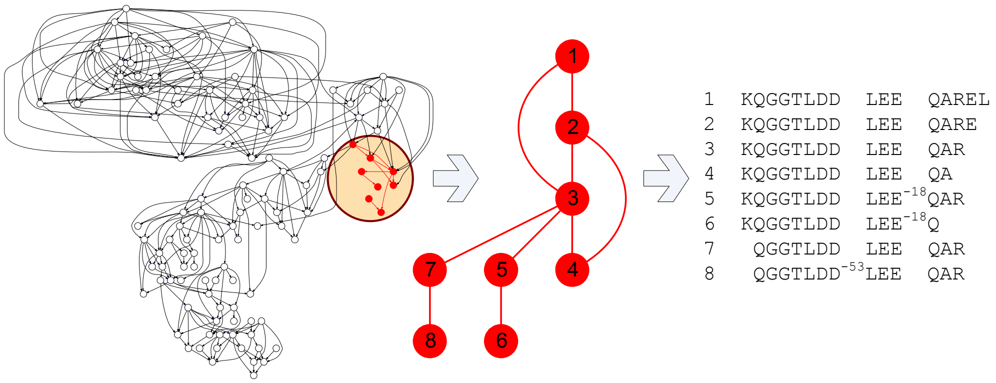


Figure 6.1 (Left) Spectral network for 945 spectra representing different peptides from the fragment IVDLQRSPMGRKQGGTLDDLEEQARELYRRLREK of the human IKKb protein. The spectral network is constructed without any knowledge of the peptide annotations. Each of 117 vertices in the spectral network corresponds to either a single MS/MS spectrum or to a consensus spectrum of multiple MS/MS spectra from the same peptide (derived by clustering). Two vertices are connected by an edge whenever the corresponding spectra form a spectral pair. (Middle) A sub-network of the entire spectral network spanning the fragment KQGGTLDDLEEQAREL (shown by red vertices on the left). (Right) Paired peptides found by analyzing the spectral sub-network in the middle with our paired spectra detection procedure.

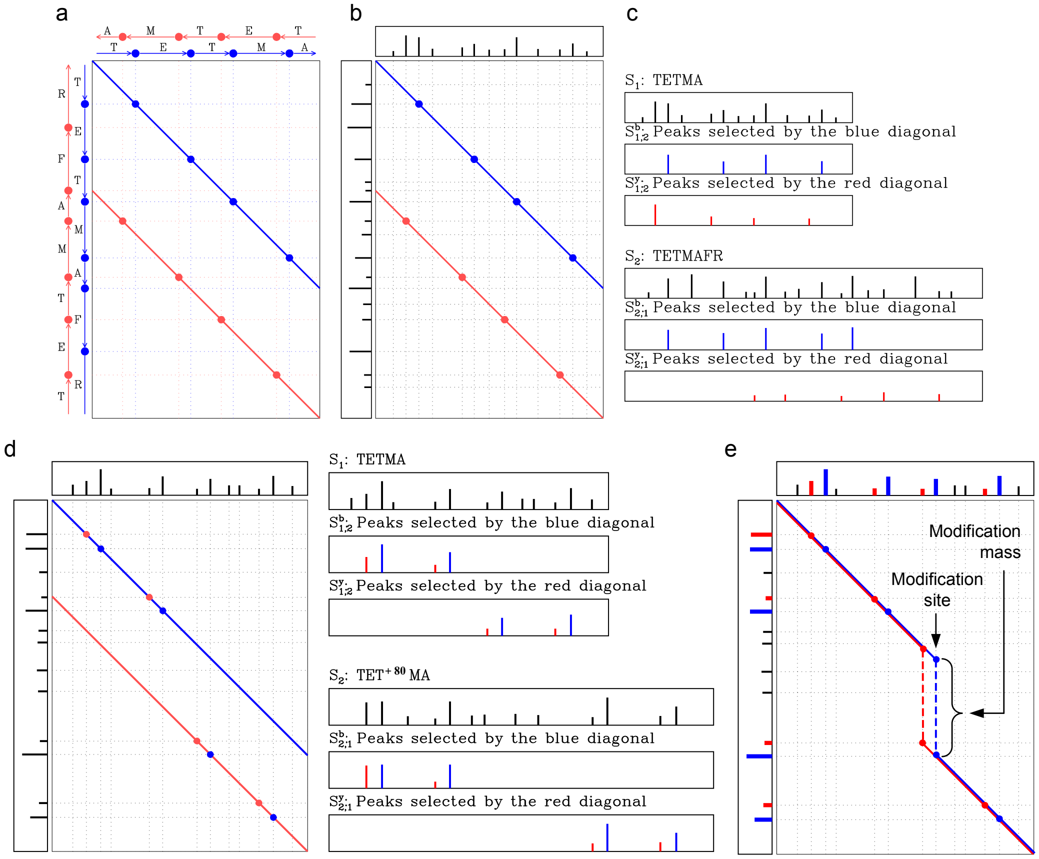

a
A M T E T
T E T M A
R T E F T A M M A T F E R T
b
c
S1: TETMA
S1,2 b: Peaks selected by the blue diagonal
S1,2 y: Peaks selected by the red diagonal
S2: TETMAFR
S2,1 b: Peaks selected by the blue diagonal
S2,1 y: Peaks selected by the red diagonal
d
S1: TETMA
S1,2 b: Peaks selected by the blue diagonal
S1,2 y: Peaks selected by the red diagonal
S2: TET+80 MA
S2,1 b: Peaks selected by the blue diagonal
S2,1 y: Peaks selected by the red diagonal
e
Modification mass
Modification site

Figure 6.2 Spectral products for terminal and internal modifications. **a)** spectral product for the theoretical spectra of the peptides TETMA and TETMAFR (all points at the intersections between the vertical and horizontal lines). The blue (resp., red) circles correspond to matching *b*-ions (resp., *y*-ions) in the two spectra. The blue and red circles are located on the blue and red diagonals. **b)** spectral product for uninterpreted spectra of the peptides TETMA and TETMAFR. The two diagonals in the spectral product matrix still reveal the points where peaks from the spectrum at the top match peaks from the spectrum on the left. **c)** spectra $S^b_{1,2}$ and $S^y_{1,2}$ defined by the blue and red diagonals. **d)** spectral product for uninterpreted spectra with one internal modification: The top spectrum corresponds to an unmodified peptide and the left-side spectrum corresponds to a modified peptide. In these cases it is not appropriate to construct $S^b_{i,j}/S^y_{i,j}$ by simply selecting peaks on the diagonals. **e)** the algorithm described in the text allows for modifications to occur in the middle of the peptide and separates the overlapping series of b and y-ions (resp. blue and red diagonals). The peaks selected from each spectrum by the blue/red diagonals are shown in the corresponding color.

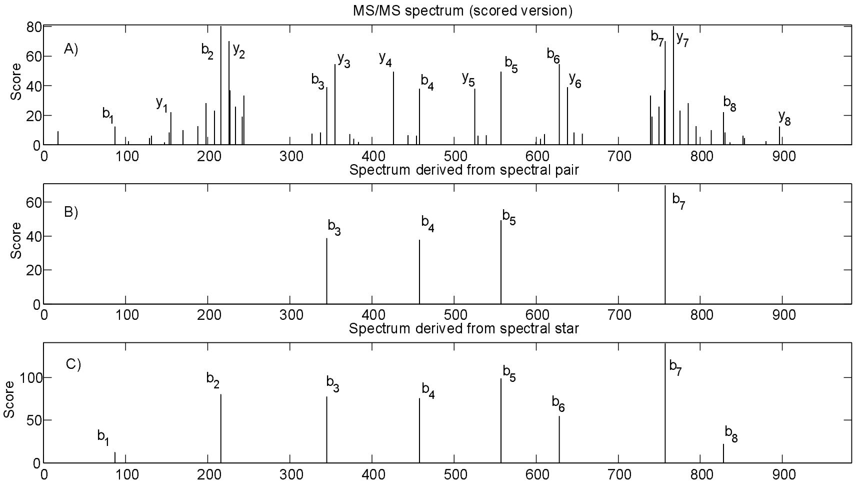


Figure 6.3 Improvements in signal-to-noise and separation of b/y-ion ladders. The scored MS/MS spectrum for peptide SEELVAEAH (from the IKKb dataset) has both b and y-ions along with several noise peaks (top). Using the spectral alignment of a pair of spectra (e.g., with KSEELVAEAH) many of the y-ions and noise peaks that do not reside on the selected diagonal are eliminated. Though paired spectra provide very good separation of b/y-ion ladders they may sometimes be too selective (e.g. causing the loss of the b_1, b_2, b_6, b_8 ions) (middle). By incorporating more paired spectra to form a spectral star, all noise peaks are removed and all missing b-ions are adequately recovered (bottom).

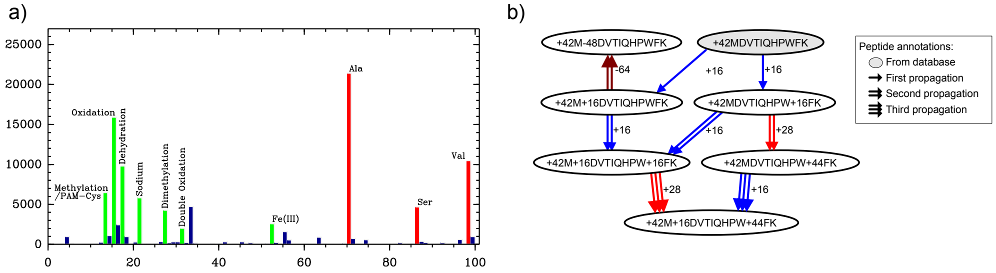


Figure 6.4 Discovery of modifications using spectral networks. **a)** histogram of absolute parent mass differences for all detected spectral pairs on the IKKb dataset; the y-axis represents the number of spectral pairs with a given difference in parent mass. For clarity, we only show the mass range 1–100 Da. The peaks at masses 71, 87, and 99 correspond to amino acid masses, and the peaks at masses 14, 16, 18, 22, 28, 32, and 53 correspond to known modifications which were also found by Tsur et al. [133] using blind database search. The peak at mass 34 corresponds to a putative modification that remains unexplained to date. **b)** modification network for peptide MDVTIQHPWFK from the Lens dataset. The gray node was annotated as peptide +42MDVTIQHPWFK by database search of the tag VTIQHP; the remaining nodes were annotated by iterative propagation. On each propagation, the source peptide annotation is combined with the modification determined by the spectral product to yield a new peptide annotation (different modifications are shown as edges with different colors).

Table 6.1 Statistics of single spectra, consensus spectra, spectral pairs, and star spectra. Satellite peaks include fragment ions correlated with b and y-ions ($b - H_2O$, $b - NH_3$, a, b^2, etc.). Signal-to-noise ratio is defined as $\frac{\#b{-}ions}{\#unexplained\ peaks}$. Spectral pairs separate b and y-ion ladders and make interpretations of resulting spectra $S^b_{i,j}$ straightforward. Spectral stars further increase the number of b and y peaks in the resulting spectra. Note that b peaks are responsible for about 90% of the score in both paired and star spectra. The results are given only for the $S^b_{i,j}$ spectra since the $S^y_{i,j}$ spectra have the same statistics.

Type of spectra		*#Explained*			*#Unexplained*	*#Total*	*signal-to-noise*
		b	*y*	*Satellite*			*ratio*
Single spectra	# peaks:	9.48	9.26	20.07	35.25	74.05	**0.27**
(11760 spectra)	% peaks:	13%	13%	26%	48%		
	% score:	28%	28%	19%	25%		
Consensus spectra	# peaks:	9.47	9.39	10.42	13.74	43.06	**0.69**
(567 spectra)	% peaks:	22%	22%	24%	32%		
	% score:	37%	36%	13%	14%		
Spectral pairs $S^b_{i,j}$	# peaks:	6.47	0.2	0.38	1.69	8.64	**3.83**
(1569 pairs)	% peaks:	75%	2%	4%	19%		
	% score:	87%	2%	4%	7%		
Star spectra	# peaks:	8.38	0.52	0.92	2.9	12.72	**2.89**
(745 stars)	% peaks:	66%	4%	7%	23%		
	% score:	88%	3%	2%	7%		

Table 6.2 Modifications on the Lens dataset as identified by blind database search and independently rediscovered with spectral networks. Only two modifications were not rediscovered: (i) deamidation on N,Q because the corresponding +1 Da mass offset is smaller than our minimum absolute modification mass of 2 Da and phosphorylation on S,T because these modifications were only present in spectra from an additional lens dataset acquired with a different instrument (Micromass QTOF) and thus not analyzed here.

Location	Modification mass	Putative annotation
S,T	-18	dehydration
Q	-17	deamidation
W	-2	cross-linking
H	14	methylation
M,W	16	oxidation
S,H	28	double methylation
N-term	42	acetylation
N-term	43	carbamylation
K, non-terminal	43	carbamylation
W	44	carboxylation
R	55	unknown
K	58	carboxymethylation
K	72	carboxyethylation

Table 6.3 New putative modifications identified by spectral networks on the Lens dataset. All of these modifications occur on peptides that had been previously identified in this sample. However, most of these modifications are rare in that they occur only on specific sites and thus tend to have low spectral counts. Two of these modifications can be explained as artifacts, two are known to occur in the context of lens and two remain unexplained to date (see discussion in main text). Spectral networks and annotated MS/MS spectra figures supporting these modifications can be found in our supplementary materials. Our approach identified all modifications found in [133] except deamidation on N,Q (Tsur et al. [133] also found phosphorylation on S,T in an additional lens dataset acquired with a QTOF instrument but not on the ion-trap dataset considered here).

Location	Modification mass	Type	Putative annotation	Comment
M	-48	neutral loss	loss of methane sulfenic acid	Reported on the same site [79]
W	4	PTM	kynurenine	Reported in cataractous lenses [128]
S	30/73	unknown	unknown	
W	32	PTM	formylkynurenine	Reported in cataractous lenses [137]
N-term	57	unknown	carboxyamidomethylation	Possible chem. artifact [24]
N-term	229/271	unknown	unknown	

7

Shotgun Protein Sequencing of Modified Proteins

7.A Introduction

The limited availability of sequenced genomes and multiple mechanisms of protein variation often refute the common assumption that all proteins of interest are known and present in a database. Well known mechanisms of protein diversity include variable recombination and somatic hypermutation of immunoglobulin genes [47]. The vital importance of some of these novel proteins is directly reflected in the success of monoclonal antibody drugs such as Rituxan$^{\mathrm{TM}}$, Herceptin$^{\mathrm{TM}}$and Avastin$^{\mathrm{TM}}$ [56, 139], all derived from proteins that are not directly inscribed in any genome. Similarly, multiple commercial drugs have been developed from proteins obtained from species whose genomes are not known. In particular, peptides and proteins isolated from venom have provided essential clues for drug design [82, 108] - examples include drugs for controlling blood coagulation [71, 76, 123] and drugs for breast [100, 122] and ovarian [93] cancer treatment. Even so, the genomes of the venomous snakes, scorpions, and snails are unlikely to become available anytime soon.

Despite this vital importance of novel proteins, the mainstream method for protein sequencing is still initiated by restrictive and low-throughput Edman degradation [98,148] - a task made difficult by protein purification procedures, post-translational modifications and blocked protein N-termini. These problems gain additional relevance when one considers the unusually high level of variability and post-translational modifications in venom proteins [19,109]. Moreover, the common laborsome approach of DNA cloning and sequencing from Edman-derived primers requires the additional availability of expensive instrumentation and expertise.

The primary function of venom is to immobilize prey and prey animals vary in their susceptibility to venom. As a result, venom composition within snake species shows considerable geographical variation, an important consideration because snake bites (even by snakes of the same species) may require different treatments. Moreover, the amount and number of different proteins and isoforms varies with gender, diet, etc. [25,30,94]. These difficulties have been widely acknowledged [35,42] and have motivated several attempts at de novo sequencing of tandem mass spectra (MS/MS) from venom proteins [120,138]. However, all such attempts were made using traditional approaches that consider each MS/MS spectrum in isolation and thus face difficulties in the reliable interpretation of individual spectra [21,26,41].

Conceptually, sequencing a protein from a set of MS/MS spectra can be described by a simple analogy. Imagine a jewelry box with many identical copies of a specific model of bead necklaces. Although all the beads are identical, this model is characterized by having irregular distances between consecutive beads - the set of inter-bead distances is initially chosen by the designer and all necklaces are then made using exactly the same specification. Now assume that one day you open your jewelry box and realize that someone has vandalized all the necklaces by cutting them to fragments at randomly chosen bead positions. Can you recover the original design of this model of necklaces, as specified by the set of consecutive inter-bead distances? In this allegory inter-bead distances correspond to amino acid masses and beads correspond to MS/MS fragmentation points (between consecutive amino acids). MS/MS data add more than a few difficulties to this necklace assembly problem; for example, most peaks in MS/MS spectra do not correspond to any fragment ions (extra beads) and many fragment ions do not result in any peaks (missing beads). Nevertheless Figure 7.1 presents an example of assembled MS/MS spectra resulting in an error-free 25 amino acid long segment of Catrocollastatin from western diamondback rattlesnake venom.

As mentioned in chapter 2, Klaus Biemann's group [67] first recognized the potential of tandem mass spectrometry for protein sequencing and manually sequenced a complete protein from rabbit bone marrow. In 2006, this approach was resurrected by Genentech researchers who were able to sequence antibodies by a combination of MS/MS and Edman degradation [107]. With the same purpose in mind we introduced the approach described in chapter 5 that utilizes multiple MS/MS spectra from overlapping peptides generated using non-specific proteases or multiple proteases with different specificities [34, 77, 89]. While this approach proved to be efficient for the assembly of a single purified unmodified protein, practical applications (like sequencing snake venoms) require applicability to mixtures of modified proteins. In fact, most MS/MS samples contain both modified and unmodified versions for many peptides, including both biological or chemical modifications introduced during sample preparation. However, it turned out that modifications present a formidable algorithmic challenge for assembly algorithms and the performance of the approach in chapter 5 degraded as soon as even a small percentage of the spectra come from modified peptides. To use the beads analogy, the necklace puzzle becomes very difficult if in addition to the canonical necklaces (non-modified proteins), the jewelry box also contains some necklaces that deviate from the designer's specification (modified proteins). In genomics, this challenge is not unlike that of assembling a highly polymorphic genome (like *Ciona* [136]) - still an unsolved problem in bioinformatics.

Using the algorithm from chapter 6 for the alignment of spectra from modified and unmodified peptide variants [10, 11], we now show that the integration of these alignments into Shotgun Protein Sequencing is not trivial and indeed requires a completely new form of spectral assembly. To this end, we introduce a generalized notion of $\mathcal{A}$-*Bruijn graphs* (originally proposed in the context of DNA fragment assembly [104]) for the assembly of MS/MS spectra from overlapping, modified and unmodified peptides into *contigs*. We further show how each contig then capitalizes on the corroborating evidence from the assembled spectra to yield a high-quality de novo consensus sequence. In fact, comparison of our contig sequences to the protein sequences identified by standard database search reveals that Shotgun Protein Sequencing results in the highest quality de novo interpretations ever reported for ion-trap spectra from a mixture of modified proteins. Combined with an extensive contig coverage of the target proteins, our results indicate that the major remaining obstacle to high-throughput protein sequencing is experimental rather than computational.

In genomics, DNA fragment assembly hardly ever produces a contiguous genome - even for small bacterial genomes it typically results in hundred(s) of disconnected contigs. While these contigs cover almost the entire genomes, they are subject to *finishing* procedures that order and join contigs together using additional experiments. Similarly, limitations in proteolytic cleavage restrict Shotgun Protein Sequencing to multiple contigs rather than contiguous proteins and motivate a quest for MS/MS-based (e.g., analysis of long multi-charged peptides that connect different contigs) finishing experiments that would allow one to connect these contigs. Alternatively, exploratory results suggest that homology-tolerant comparison of contig sequences to known protein sequences may also be a viable approach for contig ordering (i.e. comparative protein sequencing).

Even in the absence of finishing experiments, our modification-tolerant approach readily generates much more information about western diamondback rattlesnake venom proteins than some of the most laborious Edman-degradation/cloning studies [146]. We obtained de novo sequences featuring 96% average coverage at an average sequencing accuracy of 90% and identified several polymorphisms and putative novel sequences with strong homology to known venom proteins from other snake species. We therefore argue that Shotgun Protein Sequencing has the potential to overcome the limitations of current protein sequencing approaches and deliver a proteomics-based platform for studies of unknown proteins.

7.B Methods

The human inhibitor of nuclear factor kappa B kinase beta (IKKb) dataset is a set of MS/MS spectra collected from multiple IKKb samples and previously described in detail [95, 130]. Briefly, each sample was separately digested with different proteases (trypsin, elastase, Glu-C) resulting in a rich ladder of spectra from overlapping peptides. IKK is known to be a key signaling complex involved in controlling cell proliferation, survival, and tumorigenesis [60]. This IKKb dataset was extensively analyzed with SEQUEST, Mascot, X!Tandem, and InsPecT [11,130,133] resulting in many reliably identified peptides and thus constitutes a gold standard against which to benchmark the performance of our sequencing approach. The IKKb dataset contains 6126 reliably identified spectra from 524 unmodified peptides and 1383 reliably identified spectra from 346 modified peptides, out of a total of 45,500 MS/MS spectra. We consider a spectrum to be reliably identified if it meets 3 criteria: a) its InsPecT score is below the p-value threshold for 5% false-positives, b) the spectrum contains at least 50% of all true b or y ions and c) at least 50% of the spectrum intensity is in b/y ions. The unusually high percentage of modified peptides (40% of all identified peptides were found to be modified) makes this a challenging dataset in our sequencing context. Beyond the usual artefactual modifications, this dataset additionally contains evidence [133] for Fe(III) adduct on E, sodium adducts on multiple residues, including Q, dehydration of T, a putative mutation of S to D, etc.

7.B.1 Venom digestion and mass spectrometry.

Our second dataset was generated from a sample of lyophilized *crotalus atrox* western diamondback rattlesnake venom (Sigma-Aldrich, St Louis, MO). This venom was chosen for benchmarking our approach because it is relatively well studied and several of its approximately two dozen proteins, ranging from 5-70 kDa, have been previously sequenced. The complexity of our sample is illustrated in an SDS-PAGE snapshot provided in the supplementary materials (see Figure C-1). Briefly, the sample was reduced with DTT and the cysteines were alkylated with iodoacetamide. The proteins which had not already precipitated were further precipitated with 60% ice-cold ethanol. After centrifugation, the supernatant was removed and discarded. The pellet was washed several times with 95% cold ethanol, then resuspended in 0.1% Rapigest (acid-labile SDS-like detergent). 4 aliquots were created and diluted for 2hr digestions at ph 8.0 in 100 mM NH_4HCO_3; trypsin and Lys-C digests were performed in 0.085% Rapigest; chymotrypsin and Asp-N digests were performed in 0.01% Rapigest. Digestions were stopped and the detergent cleaved by acidifying with trifluoro acetic acid (TFA) pH ~ 2. LC/MS/MS data was collected twice for each digest with an automated nano LC/MS/MS system, using an 1100 series autosampler and nano pump (AgilentTechnologies, Wilmington, DE) coupled to either an LTQ or an LTQ-FT hybrid ion trap Fourier transform mass spectrometer (Thermo Electron, San Jose, CA) equipped with a nanoflow ionization source. Peptides were eluted from a (75 μm x 10 cm) PicoFrit (New Objective, Woburn, MA) column packed with (5 μm x 200 Å) Magic C-18AQ reversed-phase beads (Michrom Bioresources, Inc., Auburn, CA) using a 100 min acetonitrile/0.1% formic acid gradient at a flow rate of 250 nl/min to yield 30 sec peak widths. Centroid mode data-dependent LC/MS/MS spectra were acquired in 3 sec cycles; each cycle was of the form: 1 full MS scan followed by 8 MS/MS scans in the ion trap using normal scan rate on the most abundant precursor ions subject to dynamic exclusion for a period of 120 sec after 2 repeats. For the LTQ data set the acquisition software was LTQ v1.0 SP1, the full IT MS survey scan was at the normal scan rate, and charge state screening was not employed. For the LTQ-FT data set the acquisition software was LTQ-FT v1.0, the full FT MS survey scan was at 100K resolution with an automatic gain control (AGC) target of 200K ions, and precursor ions of unassigned charge were excluded from triggering MS/MS. Spectrum Mill v 3.02 b was used to extract all MS/MS spectra from each LC/MS/MS run including the spectral processing steps of merging replicates having a precursor mass within +/- 1.4 m/z and eluting within +/- 15 sec, quality filtering to retain spectra with a sequence tag length > 1, assigning precursor charges, and correcting ^{13}C precursor m/z misassignments. Precursor charges were assigned by Spectrum Mill for 62% of LTQ spectra using a combination of additional precursor charge states present in the MS spectra, b/y pairing in MS/MS spectra, and absence of peaks above the precursor mass in MS/MS spectra. This yielded 21,520 LTQ MS/MS spectra and 29,223 LTQ-FT MS/MS spectra. All LTQ-FT precursor charge assignments were done by the Thermo acquisition software using isotope spacing in the high resolution MS spectra. Further peak detection and de-isotoping for each spectrum was done independently in subsequent programs as needed.

7.B.2 Interpretation of venom spectra using database search.

A database of 5510 snake proteins was obtained from SwissPROT (August 3rd, 2006) by selecting all proteins from the taxa *Serpentes*, including 33 proteins and fragments from *Crotalus Atrox*. These *Crotalus Atrox* proteins were sequenced over the years in various labs using laborious Edman degradation as the first step. The obtained peptides were often used to design probes for further cloning and DNA sequencing. This database was extended with 19 protein sequences from common contaminants and proteases and 5529 “decoy” shuffled versions of all protein sequences. MS/MS spectra were searched against the database using InsPecT [130] with a peptide mass tolerance of 2.5 Da, fragment peak tolerance of 0.5 Da and allowing for oxidation on Methionine, deamidation on Asparagine, Pyro-Glu from N-terminal Glutamine and Pyro-carbamidomethylcys from N-terminal Cysteine [48]. The “decoy” database was used to enforce a false discovery rate of 1% and all retained peptides had an InsPecT-assigned p-value of 0.01 or less. Proteins were identified by iteratively selecting the protein sequence that explained the most identified spectra (minimum 10 spectra per protein).

7.B.3 Pairwise Spectral Alignment.

As usual in the analysis of MS/MS spectra, we employed several preprocessing steps. In particular, we used parent mass correction, parent charge estimation and clustering of multiple spectra from the same peptide as described in chapter 3. Furthermore, we replaced every peak with its likelihood score [1]. This scoring combines each peak’s intensity, b/y complementarity and presence/absence of neutral losses into a single likelihood score. Also, it has the additional effect of making every spectrum symmetric - a desirable transformation because we often can't tell *ab initio* which peaks come from prefix fragments (e.g. b-ions) and which come from suffix fragments (e.g. y-ions).

In our necklace problem, one can only rely on matching inter-bead distances from overlapping fragments to reconstruct the original sequence of consecutive inter-bead distances. This matching is the exact purpose of the spectral alignment described here - to find pairs of spectra from overlapping peptides (spectral pairs). Conceptually this procedure is akin to aligning inter-bead distances in that we need to detect overlaps between MS/MS spectra without knowing the corresponding peptides.

The algorithm for detection of spectra from overlapping peptides follows the previous approaches described in chapters 5-6 (see Figure 7.2). Briefly, spectral alignment translates the powerful Smith-Waterman sequence alignment technique [119] to the realm of MS/MS analysis. Like the dynamic programming matrix used in sequence alignment we construct a *spectral matrix* and find an *optimal path* in this matrix. Intuitively, the spectral matrix of spectra S and S', is the set of pairs of peaks ($p \in S, p' \in S'$) called *matching peaks* (Figure 7.2). Pairs of matching peaks may be connected by *jumps* as described in Figure 7.2 with oblique jumps corresponding to putative modifications. As in classical sequence alignment, the optimal path (i.e. sequence of jumps) in the spectral matrix reveals the relationships between spectra. If spectra S and S' originate from overlapping peptides then there exists a path in this graph containing a large number of matching peaks, otherwise spectra S and S' are likely to be unrelated (in reality, peaks are scored by intensities as described in [1]). Algorithmically, spectral alignment is more complex than sequence alignment since in the former case one optimizes two correlated paths in the spectral matrix (one corresponding to b-ions, illustrated in blue, and another corresponding to y-ions, illustrated in red) while in the latter case one is only concerned with a single path. While these paths are referred to as “blue” and “red” paths, in reality, the colors of the paths are not known in advance. We further note that although pairs of related spectra can also be identified by chemical tagging procedures [49, 81] or special instrumentation [112], these approaches do not consider overlapping peptides and cannot match spectra from multiple samples.

Figure 7.2 presents three cases where spectral alignments help reveal overlapping and modified peptides from the IKKb dataset without even trying to interpret the spectra: a) SVSCILQEPK and SVSCILQEPKR (suffix extension), b) SVSCILQEPK and SVSCILQ^{+22}EPK (modified variant), and c) PESVSCILQEPK and SVSCILQEPKR (partial-overlap). The corresponding optimal paths (shown in blue for *b*-ions and red for *y*-ions) and selected matching peaks between the different spectral pairs are illustrated in Figure 7.2 . Note that choosing where to place the jumps implicitly defines the type of spectral pair - modified/unmodified pair if there is an oblique jump in the middle, prefix/suffix pair if there is a single horizontal/vertical jump at the end/start or overlap pair if there is one horizontal/vertical jump at the start and another at the end. The spectral alignment places the jump(s) in a position that maximizes the total scores of all matching peaks [11, 133]. On a single desktop machine (Pentium4 at 2.8 GHz with 1Gb of memory) our pairwise spectral alignment step executed in 46 minutes for the *Crotalus Atrox* dataset. However, the computation of pairwise spectral alignments can easily be executed in parallel and completes in only a few minutes when run on UCSD's FWGrid 64-node Linux cluster.

As a final step in our spectral alignment stage, we capitalize on a useful byproduct of spectral alignment - the separation of *b* and *y*-ions in the aligned spectra. Even though the colors of the paths are unknown to the algorithm it turns out that, with high probability, the blue and red paths cleanly separate *b* and *y*-ions (see chapter 6). This separation is used to transform every aligned spectrum S into a *star spectrum* - a subset of S composed of mostly *b*-ions or mostly *y*-ions, but not both. As shown in Table 6.1, s tar spectra contain very few noise peaks while retaining most b-ions (or y-ions) and are extremely selective of same-type ions (i.e. only *b* or only *y*).

7.B.4 Shotgun Protein Sequencing.

It is widely accepted that *pairwise alignment whispers while multiple alignment shouts out loud* - combining pairwise spectral alignments into a single multiple alignment reveals peaks that are simultaneously supported by all or most of the aligned spectra. The high quality of star spectra may create the impression that the standard “overlap-layout-consensus” (OLC) approach [96] for DNA fragment assembly should work for spectra assembly. In fact, we originally pursued this approach (as described in chapter 5) but it turned out to face serious difficulties as soon as even a small fraction of spectra represent modified peptides [9]. The problem is that MS/MS spectra often come in *both* modified and unmodified versions thus posing a formidable challenge for assembly algorithms. In particular, the OLC approach simply projects all aligned peaks to a consensus spectrum and scores each consensus peak according to its co-occurrence in all overlapped spectra. Unfortunately this approach does not work when a set of overlapping spectra contains modifications since a simple projection of peaks onto a consensus spectrum would generate “shadow” peaks for each modification state. This shadowing effect would become even more severe if the alignment happened to include spectra from peptides with multiple modifications.

Note that although a spectral alignment is able to identify the mass and location of a modification, it is not immediately obvious which spectrum comes from the modified peptide, i.e. whether the modification corresponds to a loss or gain of residue mass. The situation becomes even more complex in the case of multiple modifications on the same peptide. Similar reasons help explain why assembly of de novo interpretations from the aligned spectra would lead to limited success at best. Even when no modifications are present, accurate de novo sequencing of MS/MS spectra is a difficult problem, often resulting in several possible peptides that explain the spectrum almost equally well. Thus, while committing any spectrum to a particular peptide would ignore the multiple alignment, considering all possible combinations of all top peptide interpretations would quickly lead to a combinatorial explosion of possible assembly configurations. However, the set all possible interpretations for any given spectrum can be represented in a very compact way by a *spectrum graph* - each peak in the spectrum defines a vertex and two vertices are connected by an edge if their peak masses differ by one or two amino acid masses (see chapter 2). Also, each vertex is assigned a score equal to the intensity of the corresponding spectrum peak. In this representation, every possible peptide interpretation corresponds to a path from zero to the spectrum's parent mass (since there is a one-to-one correspondence between spectrum peaks and spectrum graph vertices, these terms will be used interchangeably). Figure 7.3a illustrates two simplified spectrum graphs for the aligned spectra S_1/S_2, showing only the vertices for the true b-ions (in blue) and edges for the correct peptide path (in orange for S_1 and purple for S_2).

In the terms of the bead necklace analogy, each of these peptide paths would correspond to a necklace fragment from one of the original necklaces. Thus, we propose to reconstruct the original sequence of beads by finding similar pairs of overlapping fragments and "gluing" the matching beads to form a long chain identical to the original necklace model. Figure 7.3 illustrates how this intuitive notion can be applied in the realm of spectral assembly: use spectral alignment to find the set of matching peaks between S_1/S_2 (Figure 2a) and use these matches to glue the corresponding spectrum graph vertices (Figure 2b). When applied to the simplified spectrum graphs in Figure 7.3a, this would result in a merged spectrum graph with a single peptide path spelling the consensus sequence of S_1 and S_2. These merged spectrum graphs will be referred to as $\mathcal{A}$-Bruijn graphs.

$\mathcal{A}$-Bruijn graphs were first proposed by Pevzner et al. [104] in the context of repeat analysis and DNA fragment assembly. The key idea in their approach is to represent every DNA read as a path through nucleotides and "glue" all paths (reads) using matching nucleotides as pairwise gluing instructions. However, while each DNA read defines a single path through its nucleotide sequence, any given spectrum will correspond to a spectrum graph encoding many possible paths through its peaks. In fact, if genomic sequences did not contain so many similar and long repetitive regions, they would be much easier to assemble than protein sequences from MS/MS spectra! In particular, MS/MS spectra are intrinsically more error-prone than DNA reads - while reads are 98% accurate, MS/MS spectra contain mostly noise peaks and the best known de novo peptide sequencing algorithms are only ≈75% accurate [41].

The process of using matching peaks to glue spectrum graphs into a single $\mathcal{A}$-Bruijn graph is illustrated in Figure 7.3. Note that edges between glued vertices are also glued if originally labeled with the same amino acid. Formally, an

$\mathcal{A}$-Bruijn graph is constructed as follows: given a spectral alignment $\mathcal{S}(S, S')$ on two spectra S and S' and two corresponding spectrum graphs G and G', output a single $\mathcal{A}$-Bruijn graph $\mathcal{G}$ having G and G' as subgraphs. The specific gluing procedure is defined by the following operations:

1. Vertices in $\mathcal{G}$; vertices $v_i \in G$ and $v'_j \in G'$ are glued into a single vertex in $\mathcal{G}$ if the corresponding peaks $p_i \in S$ and $p'_j \in S'$ are matched in $\mathcal{S}(S, S')$. All remaining non-matched vertices are imported directly into $\mathcal{G}$. Each $\mathcal{A}$-Bruijn vertex is scored by the sum of its grouped peaks' intensities.

2. Edges in $\mathcal{G}$; all edges in G and G' are imported directly into $\mathcal{G}$. However, edges are also glued if the endpoint vertices in G are glued to the endpoint vertices in G' and the edges are labeled with the same mass. Such pairs of edges, say e and e', are replaced by a single edge e'' of the same mass.

The construction of an $\mathcal{A}$-Bruijn graph for a set of spectra and a set of spectral alignments is a straightforward iteration of the gluing operations described above. An example of a long sequence obtained from a set of 24 assembled spectra is illustrated in Figure 7.1. However, errors in the spectral alignments may lead to the incorrect gluing of some peaks and generate inconsistent vertices in the $\mathcal{A}$-Bruijn graph. In particular, it sometimes happens that multiple peaks from the same spectrum end up glued in the same vertex. Fortunately, these inconsistencies are easily detected and techniques are provided to resolve them (see appendix B).

After an $\mathcal{A}$-Bruijn graph is constructed, the consensus sequence is defined as the heaviest path in the resulting directed graph. On most occasions, the resulting $\mathcal{A}$-Bruijn graph is a directed acyclic graph and thus standard algorithms are readily available to solve this problem. On the rare occasions when incorrect spectral alignments induce directed cycles in the $\mathcal{A}$-Bruijn graph, we find that a simple greedy modification to the standard heaviest path algorithm works well on our $\mathcal{A}$-Bruijn graphs (described in detail in appendix B).

7.C Results

In the spirit of DNA fragment assembly [96], each set of overlapping spectra assembled by our approach is referred to as a *contig*. Table 7.1 lists the number of contigs assembled from each dataset along with some statistics on $\mathcal{A}$-Bruijn graph construction and sequencing; de novo sequences obtained from the contigs are referred to as *contig sequences*. Note that contig sequences may be shorter than the span of amino acids covered by MS/MS spectra within a contig (some amino acids at the beginning/end of the contigs may not be recoverable). Overall, these contig sequences covered 96% of all assembled regions in the venom dataset and 85% in the IKKb dataset. Table 7.1 also shows the sequencing accuracy and coverage for the most abundant proteins in each dataset. It may appear that sequencing proteins is an easier task than sequencing DNA since protein sequences have few repeats or palindromes (the major source of difficulties in whole-genome assembly). However, not only are MS/MS spectra intrinsically more error-prone than DNA reads but peptide sampling is strongly biased and results in some portions of the proteins being represented in many spectra while others are not seen at all. As a result, the observed peptides often correspond to isolated sets of overlapping spectra separated by coverage gaps or sometimes connected by only one or two spectra. Figure 7.4 shows the spectrum coverage observed for the IKKb and Catrocollastatin proteins (see appendix C for spectrum and contig coverage of all venom proteins).

Figure 7.4 and Table 7.1 demonstrate that Shotgun Protein Sequencing is a modification-tolerant approach applicable to protein mixtures. On the IKKb protein, 100 different amino acids were found to be modified in at least one spectrum and the whole dataset contained over a thousand spectra from hundreds of modified peptides. Nevertheless, we were able to assemble 87% of all regions covered by at least 3 spectra and to derive de novo sequences that were found to be over 90% correct. Moreover, we observed that errors predominantly fall into the initial/terminal regions of the contigs where there are fewer peaks to reliably call amino acids. Similar results were obtained on the venom dataset even though it contained almost 3000 different peptides from a mixture of crotalus atrox venom proteins. This 3.5-fold increase in the number of different peptides did not affect our sequencing accuracy and resulted in a 2-fold increase in the number of sequenced amino acids (IKKb vs venom). Although the total length of all proteins identified on the venom dataset is approximately 4 times that of the IKKb protein, much of the additional peptide diversity in the former is actually coming from the same protein regions. This is evidenced both by a larger number of peptides per contig and by the increase in sequencing coverage - more peptides per contig lead to an increased probability of finding spectrum peaks for all amino acids.

The majority of all contig sequences was readily identifiable as a peptide from the corresponding database - 84% for the IKKb dataset and 70% for the venom dataset. However, the latter also resulted in a significant number of contig sequences that did not match any proteins from the target species but had a significant match to other related species when matched against the database (using blastp [4] and SPIDER [54]). These are listed in Table 7.2 as *homologous* peptides and represent 14% of all de novo sequences obtained in the venom dataset. As it turned out, for 19 out of the 28 homologous contigs the assembled spectra could also be identified by database search (i.e. the peptide existed in a protein from a different species) and the found peptides matched our de novo sequence. On the remaining 9 cases the assembled spectra did not match any peptide in the database and thus this step neither confirmed nor refuted the putative homologies. All of these novel homologies were derived from contigs assembling multiple peptides where the annotated MS/MS spectra strongly supported the recovered sequences (see supplementary materials). It should also be noted that all *crotalus atrox* homologies were either matched to a different snake species or can be explained by single nucleotide polymorphisms of the original sequences, which were also detected in our sample.

Together with the 13 homologous peptides that matched only venom proteins from other species, these results suggest that some crotalus atrox venom proteins still remain unknown. Moreover, all homologous peptides were found among proteins from other snakes thus reinforcing our predictions.

In addition to homologous peptides, some contig sequences showed no similarity to any peptide in UniProtKB. Moreover, these contigs contained only spectra that were not identified by traditional database search of the individual spectra. In the venom dataset, it turned out that 6 out of 18 such unidentified contigs yielded highly reliable de novo sequences containing a long tag of 10 or more amino acids (allowing for one summed mass of 2 amino acids), thus again suggesting a few still unknown proteins in *crotalus atrox* venom (see Table 7.3 for sequences and supplementary materials for annotated MS/MS spectra).

A small number of the assembled contigs turned out to be incorrect (due to incorrect alignments of spectra from different peptides) or to yield mostly incorrect de novo sequences that did not match the peptide sequences assigned to the assembled spectra by traditional database search. These were mostly caused by spuriously matching both b and y peaks or high intensity unexplained peaks in the assembled spectra and account for less than 5% of all assembled contigs.

7.D Discussion

Shotgun Protein Sequencing is a modification-tolerant approach to the interpretation of tandem mass spectra that enables de novo sequencing of protein mixtures, even on ion trap instruments. This approach, for the first time, demonstrated the feasibility of very accurate de novo sequencing of modified proteins into contigs (20 aa and longer) covering contiguous sequence regions up to 108 amino acids long. In fact, the extensive contig coverage of all regions with three or more overlapping peptides indicates that the major difficulty preventing the assembly of whole proteins is the strong bias in proteolytic digestion. Thus, one straightforward route towards the production of longer contigs is through the generation of richer peptides ladders using proteases with diminished cleavage specificity. Indeed, the coverage observed in the venom dataset (based on a slightly improved digestion protocol) is already much better than the fragmented coverage of IKKb (compare Figures 7.4a and 7.4b). In the context of deuterium exchange (DXMS) studies [34,101], much progress has been achieved with controlled pepsin digests.

In general, u sing mass spectrometry for Shotgun Protein Sequencing results in certain limitations that are without counterpart in the DNA sequencing realm. The sampling frequency of the amino acids across a protein sequence is not uniform and is dictated by local sequence context. The coverage of a protein by its peptides is biased by the specificity and distribution of cleavage sites of the proteases employed. The ionizability and extent of fragmentation of individual peptides are biased by the presence/absence of basic, charge-bearing residues (Arg, Lys, His) and Pro, whose constrained side-chain is covalently bound to the peptide backbone. Certain combinations of amino acids have identical elemental compositions that are indistinguishable by mass and may leave ambiguity in the draft (or even finished) sequences depending on the completeness of fragmentation in the MS/MS spectra (I=L=113, GG=N=114, GA=Q=128). Others have the same nominal mass, but not elemental composition, and are distinguishable only in MS/MS from high resolution instruments (Q=K=128 and W=DA=VS=186). Distinguishing the identical elemental composition of Isoleucine and Leucine may be achievable by performing MSn to further fragment the Ile/Leu specific immonium ion at m/z 86 [6] or, to a limited extent, by capitalizing on the cleavage specificity of chymotrypsin.

High-resolution mass spectrometers, such as Thermo's LTQ-Orbitrap, may seamlessly elevate Shotgun Protein Sequencing to a whole new level of productivity. In principle, higher mass accuracy should be directly translatable into much more sensitive detection of overlaps between spectra with poor b/y-ion ladders. This increased sensitivity would be particularly relevant for the case of MS/MS spectra from highly charged (3+) peptides, which usually feature poor b/y-ion fragmentation – these peptides tend to span more than one contig and could thus serve as "connectors" between adjacent contigs. Also, when datasets from LC time-scale compatible, electron-transfer dissociation (ETD [124]) becomes available, CNBr-derived long peptides may yield near complete, contiguous sequences.

Nonetheless, even with a standard experimental setup and using only a relatively small MS/MS dataset from a modest resolution mass spectrometer, our approach very rapidly generated much more information about western diamondback rattlesnake venom proteins than some of the most laborious Edman-degradation/cloning studies [146]. Moreover, these contigs can be easily produced with minimal experimental and computational effort while Edman degradation projects often take months to complete. Furthermore, our contigs may be readily aligned and ordered by comparative protein sequencing that, akin to comparative DNA sequencing, utilizes previously determined protein sequences from evolutionarily close species. For example, one can use the crotalus durissus durissus catrocollastatin protein sequence to map and order our crotalus atrox catrocollastatin contigs and obtain long sequences up to 96 aa in length.

While defining the termini of mature proteins could be accomplished by employing amine and carboxyl reactive labeling agents prior to enzymatic digestion, determining the signal peptides that are post-translationally cleaved would require gene cloning. To this end, the readily available contigs can be used to design degenerate DNA primers/probes to enable subsequent gene cloning efforts from venom gland cDNA libraries.

Chapter 7 is, in part, a reprint of the paper "Shotgun Protein Sequencing: Assembly of peptide tandem mass spectra from mixtures of modified proteins" co-authored with Karl Clauser and Pavel Pevzner in Molecular and Cellular Proteomics vol.6, pp.1123-34. The dissertation author was the primary investigator and author of this paper.

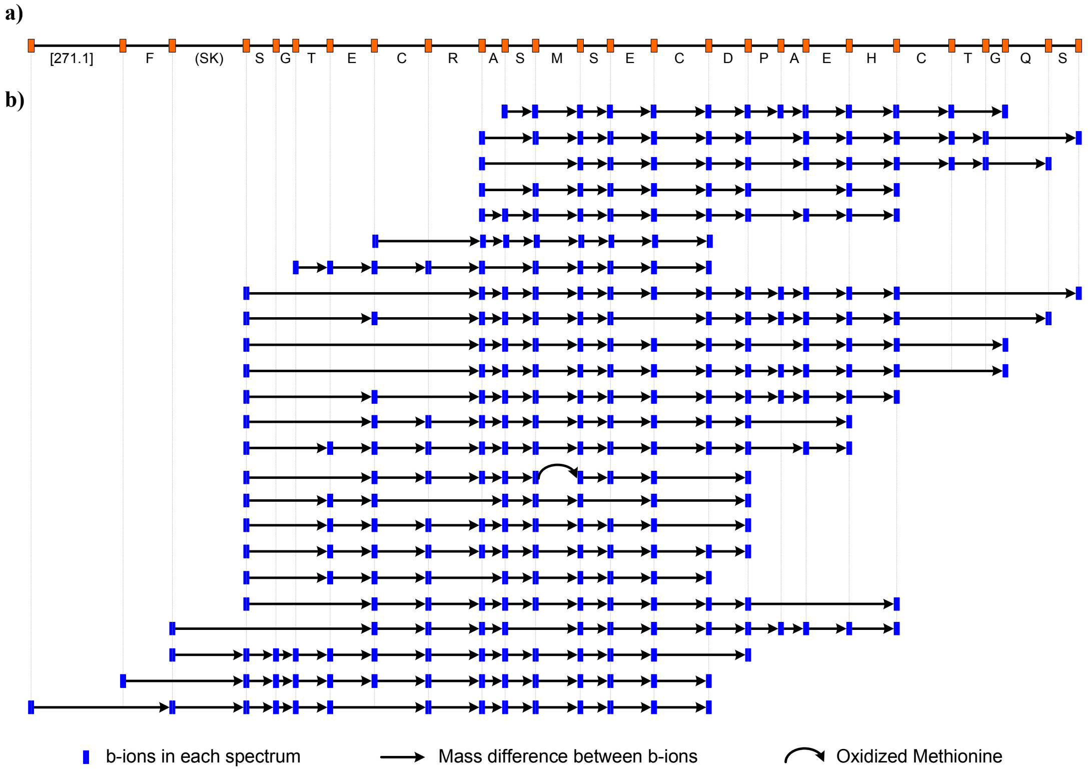
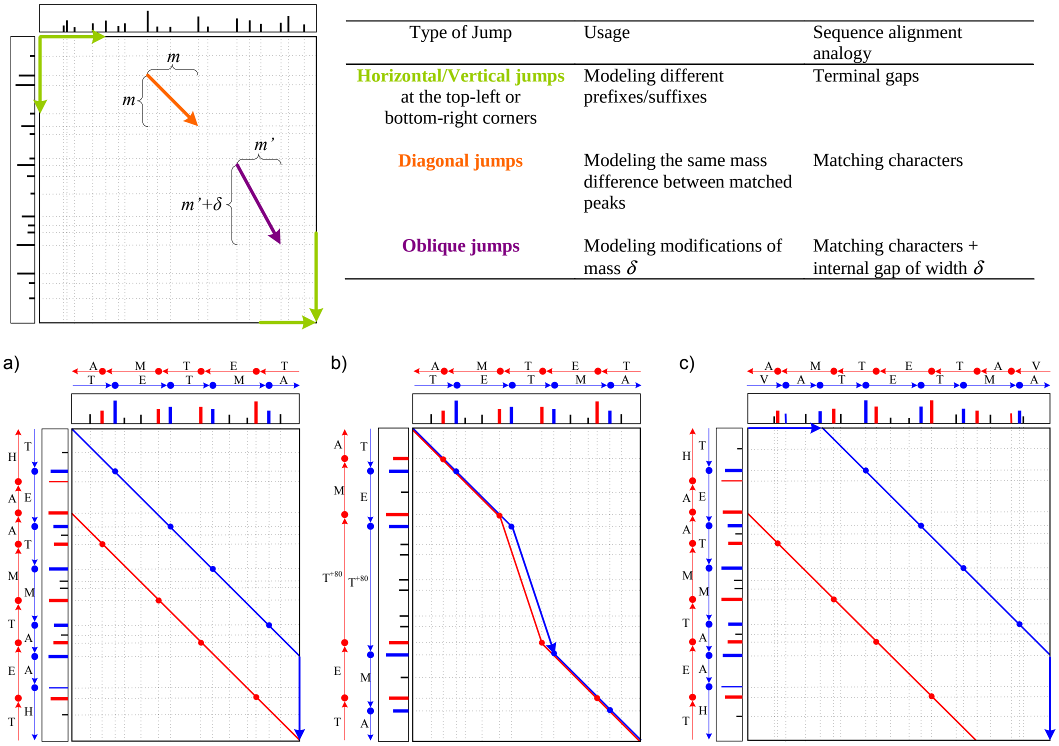
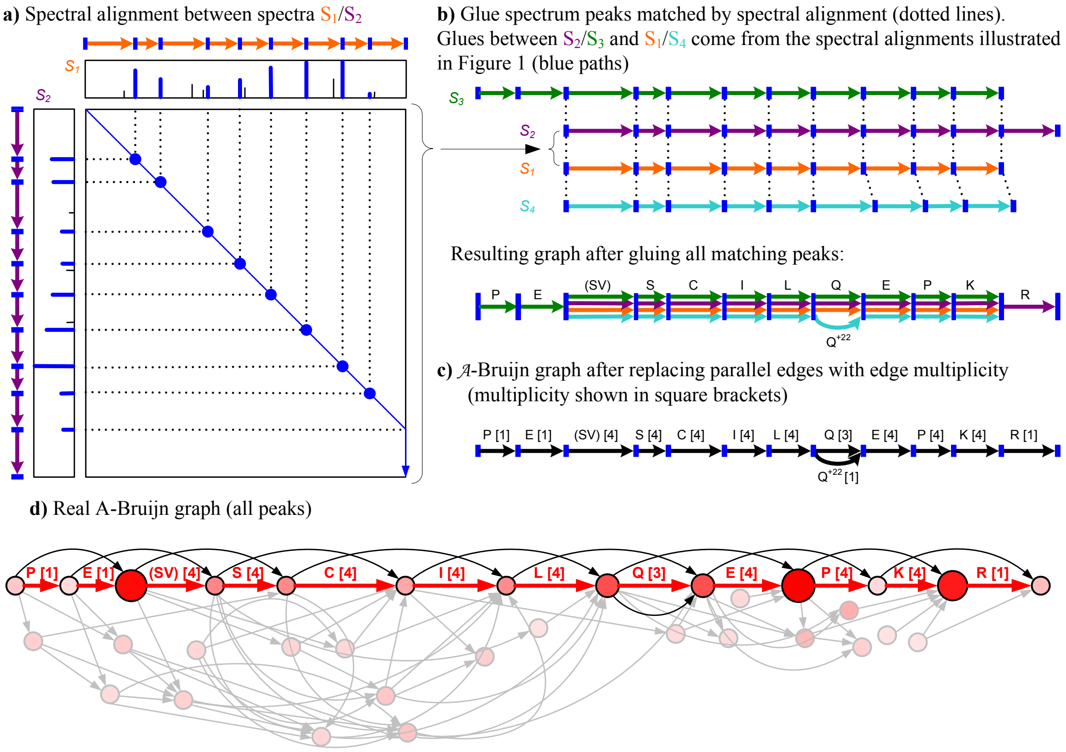


Figure 7.1 Contig assembling 24 spectra covering a 25 amino acid portion of *Crotalus Atrox* Catrocollastatin. Note that no single spectrum contains all the *b*-ions for the recovered sequence, even after we recovered missing *b*-ions from correlated ion types (e.g. *y*-ions). **a)** De novo contig sequence reconstructed from the assembled spectra. **b)** MS/MS spectra assembled in the contig. Each line corresponds to a different spectrum where matched *b*-ions are shown as blue rectangles connected by arrows.



Type of Jump	Usage	Sequence alignment analogy
Horizontal/Vertical jumps at the top-left or bottom-right corners	Modeling different prefixes/suffixes	Terminal gaps
Diagonal jumps	Modeling the same mass difference between matched peaks	Matching characters
Oblique jumps	Modeling modifications of mass δ	Matching characters + internal gap of width δ

Figure 7.2 Pairwise spectral alignments [10, 11, 105] are computed with a dynamic programming algorithm similar to the Smith-Waterman sequence alignment algorithm [119]; the corresponding intuitive interpretations are given in the table. The alignment of two spectra is defined on the set of all matching peaks - each pair of matching peaks is represented as an intersection of vertical and horizontal dotted lines on the spectral matrix (top left). 18 peaks in the first spectrum and 17 peaks in the second spectrum result in 17 × 18 matching peaks in the spectral matrix. Matching peaks may be connected by three types of jumps: Horizontal/vertical[†], Diagonal and Oblique jumps. A spectral alignment is defined as a sequence of jumps from the top-left corner to the bottom-right corner. We consider spectral alignments with any number of diagonal jumps but a limited number of other jumps and distinguish between three types of spectral alignments: a) Prefix/suffix alignments use a single horizontal/vertical jump (either at the top-left or bottom-right); b) Modified/unmodified alignments use a single oblique jump; c) Partial-overlap alignments use one horizontal/vertical jump at the top-left corner and another at the bottom-right corner. The optimal alignment of two spectra is an alignment with the longest sequence of valid jumps on the spectral matrix (the implemented scoring function is described in the main text). The alignment of b-ions is shown in blue and y-ions in red. [†]Since MS/MS spectra commonly lack peaks in the low/high mass regions, we also accept Horizontal/vertical jumps to locations where no peaks are matched.



a) Spectral alignment between spectra S_1/S_2

b) Glue spectrum peaks matched by spectral alignment (dotted lines). Glues between S_2/S_3 and S_1/S_4 come from the spectral alignments illustrated in Figure 1 (blue paths)

Resulting graph after gluing all matching peaks:

c) $\mathcal{A}$-Bruijn graph after replacing parallel edges with edge multiplicity (multiplicity shown in square brackets)

d) Real A-Bruijn graph (all peaks)

Figure 7.3 Construction of an $\mathcal{A}$-Bruijn graph from MS/MS data. Star spectra of peptides SVSCILQEPK (S_1), SVSCILQEPKR (S_2), PESVSCILQEPK (S_3) and SVSCILQ^{+22}EPK (S_4) are “glued”’ together into an $\mathcal{A}$-Bruijn graph using gluing instructions provided by pairwise spectral alignments shown in Figure 7.2. **a)** The spectral alignment of spectra S_1 and S_2 shown in Figure 7.2a reveals matching peaks in these spectra (only the blue path is shown). The peaks corresponding to b-ions are shown in blue while other peaks are shown in black. Simplified spectrum graphs are shown next to each spectrum as paths through b-ions. **b)** Matching peaks in spectral alignments shown in Figure 7.2a,b,c generate pairwise gluing instructions between every pair of aligned spectra. Thus, dotted lines are used to represent both matching peaks in a) and gluing instructions in b). **c)** Parallel edges are replaced by a single edge with weight proportional to its multiplicity. In reality, edge weights are determined from peak intensities. **d)** Real $\mathcal{A}$-Bruijn graph using all peaks in the aligned spectra. Vertex scores are represented as vertex size and color intensity; edges to noise peaks are shown in grey. The path found by Shotgun Protein Sequencing is shown in red, with edge labels for the identified amino acids (numbers in square brackets indicate edge multiplicity).

Figure 7.4 Assembled sets of spectra (*contigs*) for the most abundant protein in the IKKb(a) and venom(b) samples; horizontal axes represent amino acid positions, vertical axes represents the multiple spectra assigned to peptides from the corresponding protein; each spectrum is represented as a short blue/red horizontal line for unmodified/modified peptides. **a)** 619 spectra from IKKb resulted in 41 contigs. **b)** 1019 spectra from Catrocollastatin precursor resulted in 34 contigs. **c)** Recovered portions of the IKKb protein sequence; correct portions are shown in green (430 amino acids), incorrect are shown in orange (33 amino acids). The longest contiguous portion is 87 amino acids long and 95% of its amino acids were correctly predicted. Amino acids found to be modified (oxydation, deamidation, dehydration, etc.) in at least one spectrum are shown underlined and a bar over each amino acid indicates how often it occurred in the central portion (i.e. 20-80%) of all identified peptides - note that most de novo errors occur on non-central amino acids for which b/y peaks are often missing. **d)** Recovered portions of the catrocollastatin protein sequence; correct portions are shown in green (321 amino acids), incorrect are shown in orange (12 amino acids). The longest contiguous portion is 108 amino acids long and all of its amino acids were correctly predicted. Note that the Catrocollastatin protein in our sample is most likely a cleaved form of the sequence currently listed in SwissPROT [146].

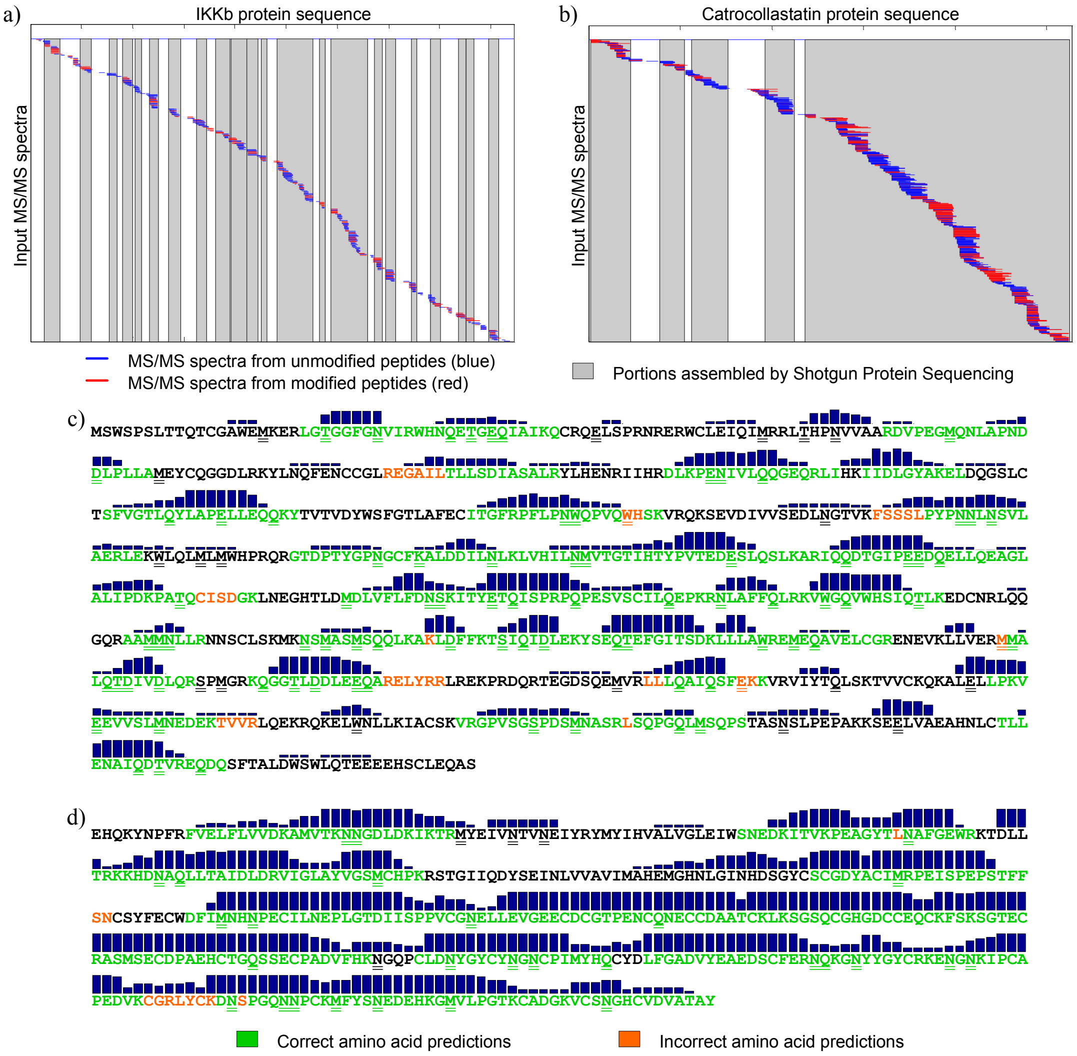

a)
IKKb protein sequence
Input MS/MS spectra
b)
Catrocollastatin protein sequence
Input MS/MS spectra
MS/MS spectra from unmodified peptides (blue)
MS/MS spectra from modified peptides (red)
Portions assembled by Shotgun Protein Sequencing
c)
MSWSPSLTTQTCGAWEMKERLGTGGFGNVIRWHNQETGEQIAIKQCRQELSPRNRERWCLEIQIMRRLTHPNVVAARDVPEGMQNLAPND
DLPLLAMEYCQGGDLRKYLNQFENCCGLREGAILTLLSDIASALRYLHENRIIHRDLKPENIVLQQGEQRLIHKIIDLGYAKELDQGSLC
TSFVGTLQYLAPELLEQQKYTVTVDYWSFGTLAFECITGFRPFLPNWQPVQWHSKVRQKSEVDIVVSEDLNGTVKFSSSLPYPNNLNSVL
AERLEKWLQLMLMWHPRQRGTDPTYGPNGCFKALDDILNLKLVHILNMVTGTIHTYPVTEDESLQSLKARIQQDTGIPEEDQELLQEAGL
ALIPDKPATQCISDGKLNEGHTLDMDLVFLFDNSKITYETQISPRPQPESVSCILQEPKRNLAFFQLRKVWGQVWHSIQTLKEDCNRLQQ
GQRAAMMNLLRNNSCLSKMKNSMASMSQQLKAKLDFFKTSIQIDLEKYSEQTEFGITSDKLLLAWREMEQAVELCGRENEVKLLVERMMA
LQTDIVDLQRSPMGRKQGGTLDDLEEQARELYRRLREKPRDQRTEGDSQEMVRLLLQAIQSFEKKVRVIYTQLSKTVVCKQKALELLPKV
EEVVSLMNEDEKTVVRLQEKRQKELWNLLKIACSKVRGPVSGSPDSMNASRLSQPGQLMSQPSTASNSLPEPAKKSEELVAEAHNLCTLL
ENAIQDTVREQDQSFTALDWSWLQTEEEEHSCLEQAS
d)
EHQKYNPFRFVELFLVVDKAMVTKNNGDLDKIKTRMYEIVNTVNEIYRYMYIHVALVGLEIWSNEDKITVKPEAGYTLNAFGEWRKTDLL
TRKKHDNAQLLTAIDLDRVIGLAYVGSMCHPKRSTGIIQDYSEINLVVAVIMAHEMGHNLGINHDSGYCSCGDYACIMRPEISPEPSTFF
SNCSYFECWDFIMNHNPECILNEPLGTDIISPPVCGNELLEVGEECDCGTPENCQNECCDAATCKLKSGSQCGHGDCCEQCKFSKSGTEC
RASMSECDPAEHCTGQSSECPADVFHKNGQPCLDNYGYCYNGNCPIMYHQCYDLFGADVYEAEDSCFERNQKGNYYGYCRKENGNKIPCA
PEDVKCGRLYCKDNSPGQNNPCKMFYSNEDEHKGMVLPGTKCADGKVCSNGHCVDVATAY
Correct amino acid predictions
Incorrect amino acid predictions

Table 7.1 Contigs obtained by Shotgun Protein Sequencing; ‡spectrum coverage is the percentage of protein sequence represented in at least 3 spectra;§contig coverage is defined as the assembled percentage of protein sequence represented in at least 3 spectra; †sequencing coverage is the percentage of contig regions that could be sequenced (in some instances there were not enough peaks in the assembled spectra to determine a complete amino acid sequence); ¶sequencing accuracy is defined as the percentage of correctly predicted amino acids; $^{\diamondsuit}$the venom dataset contained 14 reliably identified *crotalus atrox* proteins and provided strong evidence of containing additional, currently unknown venom proteins (described in the main text). Types of contig sequences listed: **a)** the contig sequence matched a protein that was expected to be in the sample; **b)** the contig sequence matched a peptide from an unexpected protein or suggested mutation of the target proteins; **c.1)** the contig sequence contains a tag of length ≥ 10 but did not match any peptide in UniProtKB and the individual MS/MS spectra were not identified by database search (SpectrumMill,InsPecT); **c.2)** like c.1) but containing only shorter tags; **d)** erroneous contigs (assembled spectra from non-overlapping peptides or de novo sequence was incorrect).

	IKKb	venom
Number of contigs	104	194
Spectrum coverage‡	57%	54%
Contig coverage§	87%	75%
Sequencing coverage†	85%	96%
Average counts per contig:		
# assembled spectra	11.4	15.1
# assembled peptides	6.5	7.3
De novo sequencing:		
a) matched the database	87 (84%)	141 (73%)
b) matched a homologous peptide	2 (2%)	28 (14%)
c.1) suggests a new peptide	0	6 (3%)
c.2) from unidentified contig	11 (11%)	12 (6%)
d) incorrect	4 (4%)	7 (4%)

	Sequencing coverage†	Sequencing accuracy¶
IKKb dataset		
IKKb	82%	92%
Overall (12 proteins)	85%	92%
Venom dataset		
Catrocollastatin (Q90288)	87%	90%
Hemorrhagic metalloproteinase (P34182)	90%	87%
Vascular apoptosis-inducing protein 1 (Q9DGB9)	100%	99%
Phospholipase A2 homolog Cax-K49 (Q8UVZ7)	100%	92%
Phospholipase A2 precursor (Q90391)	92%	94%
Overall (14+ proteins)$^{\diamondsuit}$	96%	90%

Table 7.2 Homologous contig sequences from the venom dataset. The segments identical to the de novo reconstructions are shown underlined. On the de novo sequences, parentheses indicate sequences where the order of the amino acids was not determined; square brackets indicate indistinguishable amino acid masses (on ion trap spectra). A homologous sequence is confirmed (√) if it matches the peptides obtained by independent traditional database search of the assembled MS/MS spectra. This confirmation step turned out positive whenever the homologous peptide was present in the database (albeit on a protein from a different snake species); assembled spectra in the remaining homologous contig sequences had no significant match to the database and were thus neither confirmed nor refuted. Also, all *crotalus atrox* homologies were either matched to a different snake species or can be explained by single nucleotide polymorphisms of the original sequences, which were also detected in our sample. The complete list of all putative homolog peptides can be found in our supplementary materials, as well as annotated MS/MS spectra for all novel homologies.

De novo sequence	Homologous matches	Homologous protein	Species, protein name
L(TP)GSQCAD(GV)CCDQCRF[Q,K]	LTPGSQCADGVCCDQCRFT	O42138 √	Agkistrodon contortrix laticinctus, Metalloproteinase-disintegrin-like protein
	LRPGSQCAEGMCCDQCRFM	Q2QA03	Crotalus durissus durissus, Metalloproteinase P-II
	LRPGAQCADGLCCDQCRFI	P68520	Crotalus atrox, Platelet aggregation activation inhibitor
KVLNEDEQTRD(PK)	KVLNEDEQTRDPK	Q9DF66 √	Trimeresurus jerdonii, Venom serine proteinase 3
	KVPNEDEQTRNPK	Q8QHK2	Crotalus atrox, Serine protease catroxase II
(LTNCSPK)(TD)IYSYSWKR	LTNCSPKTDIYSYSWKR	Q71QE8 √	Crotalus viridis viridis, Phospholipase A2
Y(MF)(YL)DFLCTDPSEKC	YMFYLDFLCTDPSEK	Q71QE8 √	Crotalus viridis viridis, Phospholipase A2
(IVS)WGGDI(CA)Q(PH)EPGVY(TK)	IVSWGGDICAQPHEPGHYTK	Q9I961	Agkistrodon acutus, Acubin2
	IVSWGGDPCAQPREPGVYTK	Q71QH8	Trimeresurus stejnegeri, Serine protease CL4
	IVSWGGDICAQPREPEPYTK	Q2QA04	Crotalus durissus durissus, Serine proteinase

Table 7.3 Putative new crotalus atrox peptides with no homologous matches in Swiss-Prot/UniProtKB. Parentheses indicate portions where the order of the amino acids was not determined; square brackets indicate indistinguishable amino acid masses (on ion trap spectra); numbers in square brackets indicate mass intervals that could be explained by different amino acid compositions. The annotated MS/MS spectra for these contigs can be found in our supplementary materials.

De novo sequence	Number of assembled spectra
[Q,K]FGP[Q,K]NPFCF[I,L]VQK	7
QRAV[218.0][I,L]DEYPESVAHNF	5
(MT)TGDSE[I,L]SVCW	4
YWPNTD[Q,K]E[I,L]G[I,L]DK	5
AAYPWNPVASTTLCASAE[371.0]	10
[242.3]D[I,L]SED[Q,K]D[I,L][Q,K]AEVNK	3

Appendix A

Algorithmic details of anti-symmetric spectra alignment

Orienting spectral pairs in a spectral star

A spectral star consisting of spectral pairs (S_1, S_2), (S_1, S_3), $\dots$, (S_1, S_n) allows one to increase the signal-to-noise ratio by considering $2(n-1)$ spectra $S^b_{1,i}$ and $S^y_{1,i}$ for $2 \leq i \leq n$. However, the orientation of spectral pairs in a spectral star needs to be done with caution since for each (S_1, S_i) either $S^b_{1,i}$ or $S^y_{1,i}$ needs to be reversed to avoid mixing b and y-ion ladders in the star spectrum. The difficulty is that the assignments of upper indexes to spectra $S^b_{1,i}$ and $S^y_{1,i}$ are arbitrary and it is not known in advance which of these spectra represents b-ions and which represents y-ions (i.e., it may be that $S^b_{1,i}$ represents the y-ion ladder while $S^y_{1,i}$ represents the b-ion ladder). A similar problem of reversing DNA maps arises in *optical mapping* (Karp and Shamir, 2000 [73], Lee et al., 1998 [80]). It was formalized as the *Binary Flip-Cut* (BFC) Problem [27] where the input is a set of n 0-1 strings (each string represents a snapshot of a DNA molecule with 1s corresponding to restriction sites). The problem is to assign a *flip* or *no-flip* state to each string so that the number of consensus sites is maximized. We found that for the case of spectral stars, a simple greedy approach to the BFC problem works well. In this approach, we arbitrarily select one of the spectra $S^b_{1,i}$ and $S^y_{1,i}$ and denote it $S_{1,i}$. We select $S_{1,2}$ as an initial consensus spectrum. For every other spectrum $S_{1,i}$ $(2 < i \leq n)$, we find whether $S_{1,i}$ or its reversed copy $S^{rev}_{1,i}$ better fits the consensus spectrum. In the former case we add $S_{1,i}$ to the growing consensus, in the latter case we do it with $S^{rev}_{1,i}$.

After solving the BFC problem we know the orientations of all spectra in the spectral star. The final step in constructing a *star spectrum* S^* from the resulting collection of $S_{1,i}$ spectra is identical to the consensus spectrum approach described above for clusters.

The anti-symmetric spectral alignment algorithm

Let S_1 and S_2 be two spectra, and assume w.l.o.g. that $M(S_1) < M(S_2)$, where $M(S)$ denotes the parent mass of S. Let $\Delta = M(S_2) - M(S_1)$. For simplicity, we shall assume in the following that the masses in S_1 and S_2 are integers. Furthermore, we assume that S_i $(i = 1, 2)$ contains the masses 0 and $M(S_i)$.

Denote by $\mathcal{M}(S_1, S_2)$ the spectral product matrix of S_1 and S_2. We define a *path* in $\mathcal{M}(S_1, S_2)$ to be a set of points in $\mathbb{R}^2$ that is composed of two diagonal segments $\{(x, x) : a \leq x < b\}$ and $\{(x, x + \Delta) : b < x \leq c\}$ for some $a \leq b \leq c$. Note that the first segment is on the blue diagonal and the second segment is on the red diagonal (one of the segments is empty when $a = b$ or $b = c$). We say that the *endpoints* of the path are the leftmost and rightmost points on the path.

The spectral alignment algorithm, as described in [105], finds the path from $(0, 0)$ to $(M(S_1), M(S_1) + \Delta)$ that contains the maximum number of points from $\mathcal{M}(S_1, S_2)$. For the optimal path P, the projection of P onto S_i (i.e. the set $\{x : (x, y) \in P\}$ for S_1 or $\{y : (x, y) \in P\}$ for S_2) gives a subset of S_i which usually contains many b-ion peaks. However, this set can also contain many peaks corresponding to y and neutral loss ion peaks. In order to obtain better b/y separation, we change the spectral alignment problem by selecting only a subset of the points in P: (1) Since the minimum mass of an amino acid is 57 Da, we will choose peaks with distance at least 57 between every two peaks, and (2) We will not select two points that are generated by a peak and its complement peak in S_1 or S_2.

Formally, we say that two peaks x and x' in a spectrum S are *complements* if $x + x' = M(S) + 18$. A subset A of a spectrum S is called *anti-symmetric* if it does not contain a complement pair. A set A is called *sparse* if $|x - x'| \geq 57$ for every $x, x' \in A$. Given a path P, a set $A \subseteq P$ is called sparse if the projection of A onto S_1 is sparse, and it is called anti-symmetric if the projections of A onto S_1 and S_2 are anti-symmetric (w.r.t. S_1 and S_2, respectively). Our goal is to find the largest sparse anti-symmetric subset of $\mathcal{M}(S_1, S_2)$ that is contained in some path from $(0, 0)$ to $(M(S_1), M(S_1) + \Delta)$, and contains the points $(0, 0)$ and $(M(S_1), M(S_1) + \Delta)$.

Our algorithm for solving the problem above is similar to the algorithm of Chen et al. [21] for de-novo peptide sequencing. But unlike de-novo peptide sequencing, our problem is two-dimensional, and this adds additional complication to the algorithm. We use dynamic programing to compute optimal sets of points that are contained in two paths, one path starting at $(0, 0)$ and the other path starting at $(M(S_1), M(S_1) + \Delta)$. By keeping two paths, we make sure that for each set of points we build, its projection on S_1 is anti-symmetric. In order to keep the projection on S_2 anti-symmetric, we need additional information which is kept in a third dimension of the dynamic programming table.

The input to the problem are two spectra S_1 and S_2 and the goal is to find largest sparse anti-symmetric subset of $\mathcal{M}(S_1, S_2)$ that is contained in some path from $(0, 0)$ to $(M(S_1), M(S_1) + \Delta)$, and contains the points $(0, 0)$ and $(M(S_1), M(S_1) + \Delta)$.

In a preprocessing stage, we remove every element x of S_1 if $x \notin S_2$ and $x + \Delta \notin S_2$. Denote $S_1 = \{x_1, \ldots, x_n\}$ and $S_2 = \{y_1, \ldots, y_m\}$, where $x_1 < x_2 < \cdots < x_n$ and $y_1 < y_2 < \cdots < y_m$. Let N be the largest index such that $x_N \leq (M(S) + 18)/2$.

A peak x_i in S_1 will be called *left-critical* (resp., *right-critical*) if $x_i + \Delta \in S_1$ (resp., $x_i - \Delta \in S_1$). Denote by S_1^L and S_1^R the left-critical and right-critical peaks in S_1, respectively.

For $i \leq n$, let $\text{Left}(i)$ be the set of all sparse anti-symmetric subsets of $S_1^L \cap [x_i - \Delta, x_i - 57]$, and let $\text{Right}(i)$ be the set of all sparse anti-symmetric subsets of $S_1^R \cap [x_i + 57, x_i + \Delta]$. Note that if $\Delta < 57$ then $\text{Left}(i) = \text{Right}(i) = \phi$ for all i, which simplifies the algorithm. In the following, we shall assume that $\Delta \geq 57$.

For $i \leq N$ and $j > N$, define $D_1(i, j)$ to be the maximum size of a sparse anti-symmetric set $A \subseteq \mathcal{M}(S_1, S_2)$ such that

1. A is contained in the union of a path from $(0,0)$ to (x_i, x_i) and a path from $(x_j, x_j + \Delta)$ to $(M(S_1), M(S_1) + \Delta)$.

2. A contains the points $(0,0)$, $(M(S_1), M(S_1) + \Delta)$, (x_i, x_i), and $(x_j, x_j + \Delta)$.

If there is no set that satisfies the requirements above, $D_1(i,j) = 0$.

We define tables D_2 and D_3 in a similar way: For $i \leq N < j$ and $S \in \text{Left}(i)$, $D_2(i,j,S)$ is the maximum size of a sparse anti-symmetric set $A \subseteq \mathcal{M}(S_1, S_2)$ such that

1. A is contained in the union of a path from $(0,0)$ to $(x_i, x_i + \Delta)$ and a path from $(x_j, x_j + \Delta)$ to $(M(S_1), M(S_1) + \Delta)$.

2. A contains the points $(0,0)$, $(M(S_1), M(S_1) + \Delta)$, and $(x_j, x_j + \Delta)$. Moreover, if $i > 1$ then A contains the point $(x_i, x_i + \Delta)$.

3. $\{x \in S_1^L : x_i - \Delta \leq x \leq x_i - 57 \text{ and } (x, x + \Delta) \in A\} = S$.

For $i \leq N < j$ and $S \in \text{Right}(j)$, $D_3(i,j,S)$ is the maximum size of a sparse anti-symmetric set $A \subseteq \mathcal{M}(S_1, S_2)$ such that

1. A is contained in the union of a path from $(0,0)$ to (x_i, x_i) and a path from (x_j, x_j) to $(M(S_1), M(S_1) + \Delta)$.

2. A contains the points $(0,0)$, $(M(S_1), M(S_1) + \Delta)$, and (x_i, x_i). If $j < n$ then A also contains the point (x_j, x_j).

3. $\{x \in S_1^R : x_j + 57 \leq x \leq x_j + \Delta \text{ and } (x, x) \in A\} = S$.

We also need the following definitions: For $i \leq n$, $\text{prev}(i) = i'$, where i' is the maximum index such that $x_{i'} \leq x_i - 57$. If no such index exists then $\text{prev}(i) = 1$. Similarly, $\text{next}(i) = i'$, where i' is the minimum index such that $x_{i'} \geq x_i + 57$. If no such index exists then $\text{next}(i) = n$. Define

$$M_1^L(i,j) = \max_{i' \leq i} D_1(i', j)$$

$$M_1^R(i,j) = \max_{j' \geq j} D_1(i, j')$$

$$M_2^R(i,j,S) = \max_{j' \geq j} D_2(i, j', S)$$

and

$$M_3^L(i, j, S) = \max_{i' \le i} D_3(i', j, S).$$

We also define

$$M_2^L(i, j, S) = \max_{i' \le i} \max_{S'} D_2(i', j, S'),$$

where the second maximum is taken over all sets $S' \in \text{Left}(i')$ that are consistent with S, namely $S' \cap [x_i - \Delta, x_i - 57] = S$. Similarly,

$$M_3^L(i, j, S) = \max_{j' \ge j} \max_{S'} D_3(i, j', S'),$$

where the second maximum is taken over all sets $S' \in \text{Right}(j')$ such that $S' \cap [x_j + 57, x_j + \Delta] = S$. We now show how to efficiently compute $D_1(i, j)$, $D_2(i, j, S)$, and $D_3(i, j, S)$ for all i, j, and S.

Computing $D_1(i, j)$

If either $x_i \notin S_2$ or $x_j + \Delta \notin S_2$, then by definition, $D_1(i, j) = 0$. We also have $D_1(i, j) = 0$ when x_i and x_j are complements or when $x_j - x_i < 57$. Furthermore, if $i = 1$ and $j = n$ then $D_1(i, j) = 2$. Now, suppose that none of the cases above occurs. Then,

$$D_1(i, j) = \begin{cases} M_1^L(\text{prev}(i), j) + 1 & \text{if } x_i > M(S_1) + 18 - x_j \\ M_1^R(i, \text{next}(j)) + 1 & \text{otherwise} \end{cases}.$$

Computing $D_2(i, j, S)$

Suppose that $x_i + \Delta, x_j + \Delta \in S_2$, x_i and x_j are not complements, and $x_j - x_i \ge 57$. If $x_{i'} + \Delta$ is complement of $x_{j'} + \Delta$ (w.r.t. S_2) for some $i' \in \{i, j\}$ and $j' \in S \cup \{j\}$, then $D_2(i, j, S) = 0$. Otherwise,

$$D_2(i, j, S) = \begin{cases} M_2^L(\text{prev}(i), j, S) + 1 & \text{if } x_i > M(S_1) + 18 - x_j \\ M_2^R(i, \text{next}(j), S) + 1 & \text{otherwise} \end{cases}.$$

Computing $D_3(i, j, S)$

Suppose that $x_i, x_j \in S_2$, x_i and x_j are not complements, and $x_j - x_i \ge 57$. If $x_{i'}$ is complement of $x_{j'}$ (w.r.t. S_2) for some $i' \in \{i, j\}$ and $j' \in S \cup \{j\}$, then

$D_3(i, j, S) = 0$. Otherwise,

$$D_3(i, j, S) = \begin{cases} M_3^L(\mathrm{prev}(i), j, S) + 1 & \text{if } x_i > M(S_1) + 18 - x_j \\ M_3^R(i, \mathrm{next}(j), S) + 1 & \text{otherwise} \end{cases} .$$

Computing $M_1^L(i, j)$

The recurrence formula for M_1^L is straightforward: For $i = 1$, $M_1^L(i, j) = D_1(i, j)$, and for $i > 1$,

$$M_1^L(i, j) = \max \left\{ D_1(i, j), M_1^L(i - 1, j) \right\} .$$

The recurrence formulae of M_1^R, M_2^R, and M_3^L are similar.

Computing $M_2^L(i, j, S)$

For $i > 1$,

$$M_2^L(i, j, S) = \max \left\{ D_2(i, j, S), \max_{S'} M_2^L(i - 1, j, S') \right\} ,$$

where the second maximum is taken over all sets $S' \in \mathrm{Left}(i - 1)$ that are consistent with S. The computation of $M_3^R(i, j, S)$ is similar.

Finding the optimal solution After filling the tables D_1, D_2, and D_3, we can find the size of the optimal set of points by taking the maximum value in these tables. The corresponding optimal set can be found by traversing the dynamic programming tables starting from the cell containing the maximum value.

Time complexity Using additional data structures, each cell of D_1, D_2, and D_3 can be computed in constant time (we omit the details). Thus, the time complexity of the algorithm is $O(kn^2)$, where

$$k = \max\{|\mathrm{Left}(1)|, \ldots, |\mathrm{Left}(N)|, |\mathrm{Right}(N + 1)|, \ldots, |\mathrm{Right}(n)|\}.$$

Although k can be exponential in n, in practice, k has small values.

Improvements The algorithm described above can be improved in two areas. First, the accuracy can be improved by considering a variant of the maximum sparse anti-symmetric subset problem which differ from the original problem in the following aspects: (1) Each point (x, y) has a score which is equal to $score(x) + score(y)$. The goal is to find maximum weight subset that satisfies the requirement. (2) The sparse requirement is replaced by the following requirement: For every two points (x_1, y_1) and (x_2, y_2) in A, $|x_1 - x_2|$ is either greater than 200, or is equal to to the parent mass of either 1 or 2 amino acids. The algorithm described above can be easily modified to solve the new problem.

The time complexity of the algorithm can be improved by filling only part of the tables D_1, D_2, and D_3. More precisely, after some changes to the algorithm, we can fill these tables only for i and j such that $|x_i - x_j| \leq 200$. We omit the details.

Appendix B

Algorithmic details of Shotgun Protein Sequencing

The *Binary Flip Cut* Problem

Before a set of aligned spectra can be assembled into a single multiple alignment we first need to ensure that all spectra are “oriented” in the same way, i.e. all aligned spectra contain predominantly b-ions or predominantly y-ions. Note that although every star spectrum is already expected to contain mostly b or mostly y-ions, some star spectra may be composed of predominantly b-ions, while others may contain predominantly y-ions. Thus, some of the aligned spectra may need to be reversed to avoid mixing b and y-ions in the consensus multiple alignment. This orientation problem is akin to the *Binary Flip Cut* Problem previously addressed in the context of optical mapping [73] and for MS/MS spectra in [9,10]. Our approach is essentially a combination of the approaches in [9,10]. After this procedure, one can assume that all spectra were oriented in the conventional left-to-right order corresponding to amino acid sequence (i.e., assume that all spectra are composed of mostly b-ions).

The input to the BFC problem is a set of n 0-1 strings where each string represents a snapshot of a DNA molecule with 1s corresponding to restriction sites. The problem is to assign a *flip* or *no-flip* state to each string so that the number of consensus sites is maximized. We have found that a simple greedy approach to the spectral orientation problem works quite well. First we choose the highest scoring spectral pair (S_i, S_j) and use it to define the consensus orientations of S_i and S_j (as indicated by the matching peaks in their spectral product). Then we iterate an expansion procedure: find a non-oriented spectrum S_k with maximal connectivity[1] to the set of oriented spectra, choose the orientation that maximizes the agreement with the consensus and add S_k to the set of oriented spectra. This expansion procedure is repeated until all spectra are oriented.

Constructing $\mathcal{A}$-Bruijn graphs from MS/MS spectra

When constructing an $\mathcal{A}$-Bruijn graph, some complications may arise due to spurious matches between spectrum peaks. As is illustrated in Figure B-1a, it may happen that an incorrect spectrum alignment leads to an incorrect gluing of spectrum peaks. In this case, multiple peaks from the same spectrum end up glued in the same *composite* vertex (e.g. peaks p_i and p_{i+1} in Figure B-1a)). Composite vertices were also observed by Pevzner et al. when $\mathcal{A}$-Bruijn graphs were first proposed [104] and we have chosen to use the exact same strategy that was then proposed to separate composite vertices into regular vertices. Very briefly, the composite vertex splitting procedure is as follows: find the highest scoring edge e connecting a composite vertex v^c to a non-composite vertex v. Then, split the set of peaks in v^c into two disjoint sets of peaks v^c and v^c_e such that $p \in v^c_e$ if e is incident on p or $p \in v^c$ otherwise. Since v^c_e is guaranteed to be non-composite (only one peak per spectrum can match the mass difference in e) we simply repeat these steps until no composite vertex remains.

Most of the times, finding the heaviest path in our $\mathcal{A}$-Bruijn graphs is a straightforward procedure because these graphs are usually acyclic. In such cases, a simple dynamic programming algorithm solves the problem very efficiently. In short, each vertex v keeps track of the the score $ps(v)$ of the highest scoring path reaching it. Then, for every source v_s and every edge $e = (v_s, v)$, update $ps(v)$ to $\max(ps(v), ps(v_s) + score(e))$ and remove e from the graph. After all edges have

[1] Connectivity is the number of reliable pairings between a non-oriented spectrum S_k and the spectra in the oriented set. In our case, we weigh each spectral pair $(\_\ , S_k)$ by the percentage of matching peaks' scores in S_k and define the connectivity of S_k as the summed weights of all considered spectral pairs.

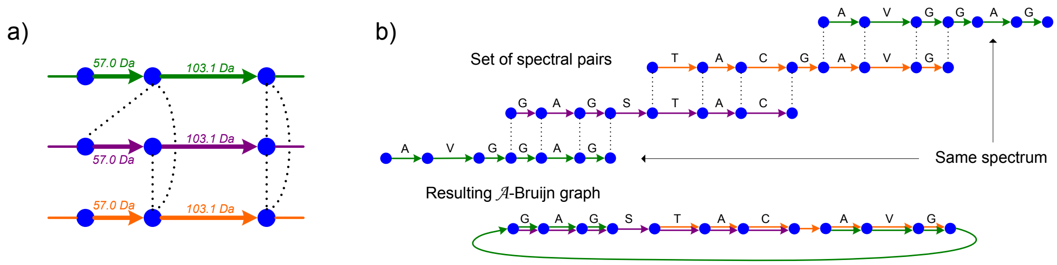


Figure B-1 Possible complications in the 𝒜-Bruijn graph. a) Example of a composite vertex - the incorrect alignment between the green and purple spectra results in two peaks from the same purple spectrum being glued into the same 𝒜-Bruijn vertex. b) The green spectrum creates a cycle in the A-Bruijn graph by connecting the AVG suffix of the orange spectrum back to the GAG prefix of the purple spectrum.

been processed we simply locate the vertex with the highest score and trace back the highest scoring path. There are, however, some cases where the 𝒜-Bruijn graph contains cycles due to an incorrect alignment of one or more spectra - one such case is illustrated in Figure B-1b). Even though, in general, finding the heaviest path in a graph with cycles is a very hard problem [72], we have found that a small change to the standard algorithm performs well for 𝒜-Bruijn graphs generated from spectral alignments. Essentially, a cycle in the 𝒜-Bruijn graph would cause the algorithm to fail because at some point there would still be unprocessed vertices in the graph but no more sources to iterate over (cycles contain no source vertices). Whenever we faced this problem we located the vertex in the graph with the lowest percentage of unprocessed incoming edges and converted it into a source by removing all such edges from the graph.

Appendix C

Assembly coverage of Crotalus Atrox proteins

Figure C-1 illustrates the complexity of the venom extract analyzed using our approach.

The coverage figures below illustrate the protein coverage of our resulting *Crotalus Atrox* contigs - portions covered by assembled spectra are shown in gray. On every coverage figure, the small blue (red) horizontal lines correspond to MS/MS spectra of identified unmodified (modified) peptides, as determined by traditional database searching. Vertical axes separate the identified MS/MS spectra and horizontal axes show where the spectra matched the protein sequence. For increased visibility of the areas covered by identified spectra, some proteins are shown without the initial portion of their protein sequence. Most likely, the non-observed prefixes of these proteins correspond to cleaved signal peptides but further experiments would be required to confirm this conjecture (e.g. by modifying all protein N-termini prior to proteolytic digestion). On some cases, there are regions of the protein covered by assembled spectra that were not identified by traditional database search but whose de novo interpretation matched the corresponding protein.

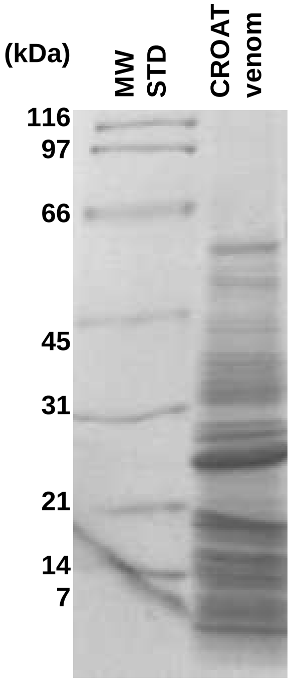


Figure C-1 SDS-PAGE snapshot of the *Crotalus Atrox* venom sample. A 37.5 mg aliquot of the reduced/alkylated crotalus atrox venom resolubilized in 0.1%rapigest was taken out of the sample prior to in-solution proteolytic digestions and separated by SDS-PAGE. The 10-20% gradient tris-HCL gel was Coomassie stained to visualize the approx. two dozen proteins present in the venom.

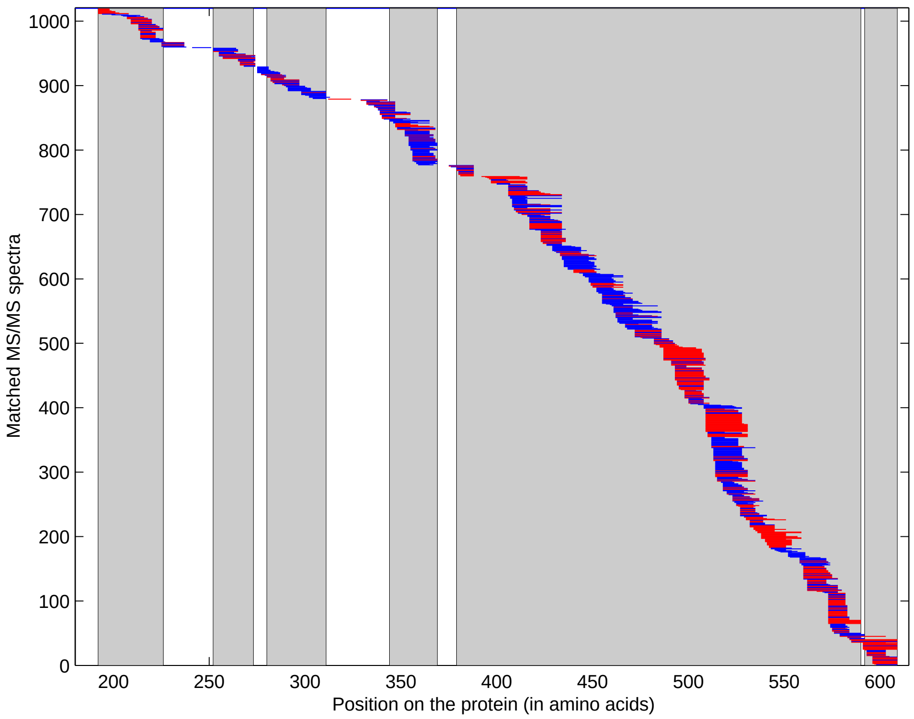


Figure C-2 Protein regions covered by assembled spectra from Catrocollastatin (Q90282_CROAT).

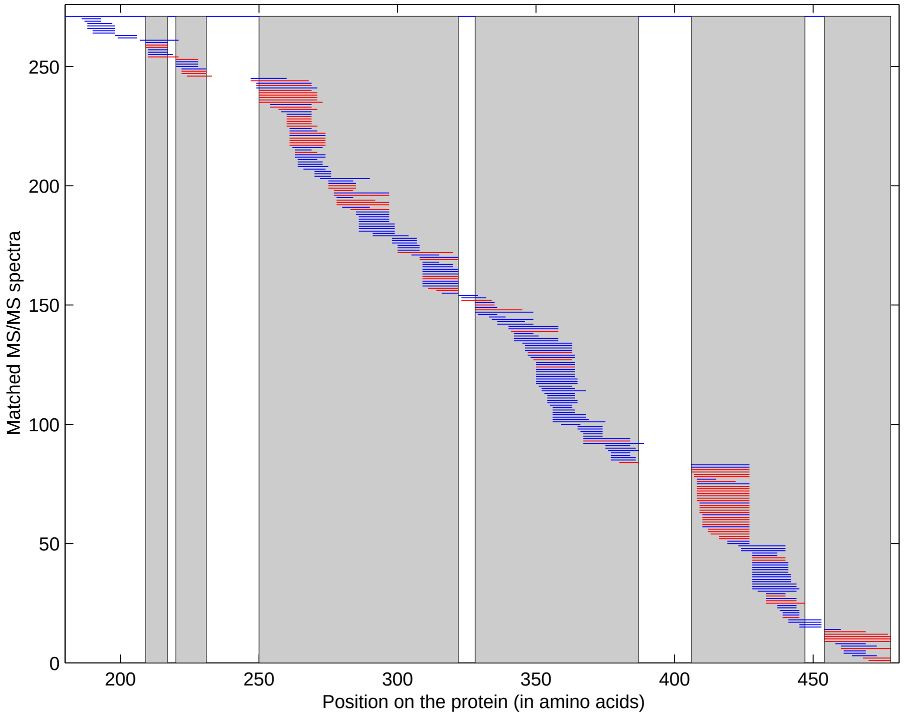


Figure C-3 Protein regions covered by assembled spectra from Hemorrhagic metalloproteinase HT-E (HRTE_CROAT).

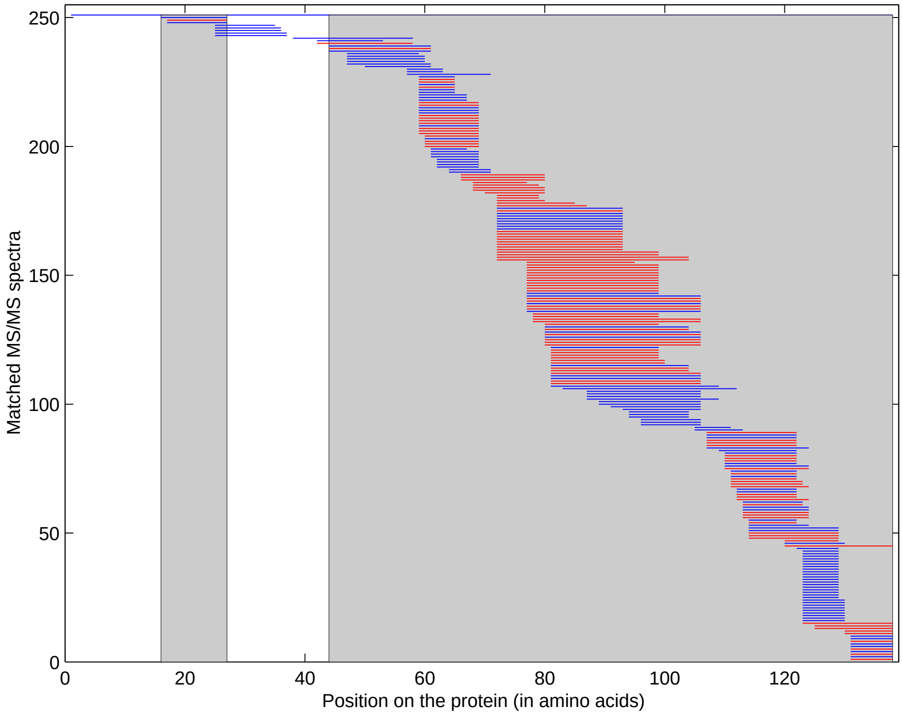


Figure C-4 Protein regions covered by assembled spectra from Phospholipase A2 (PA2_CROAT).

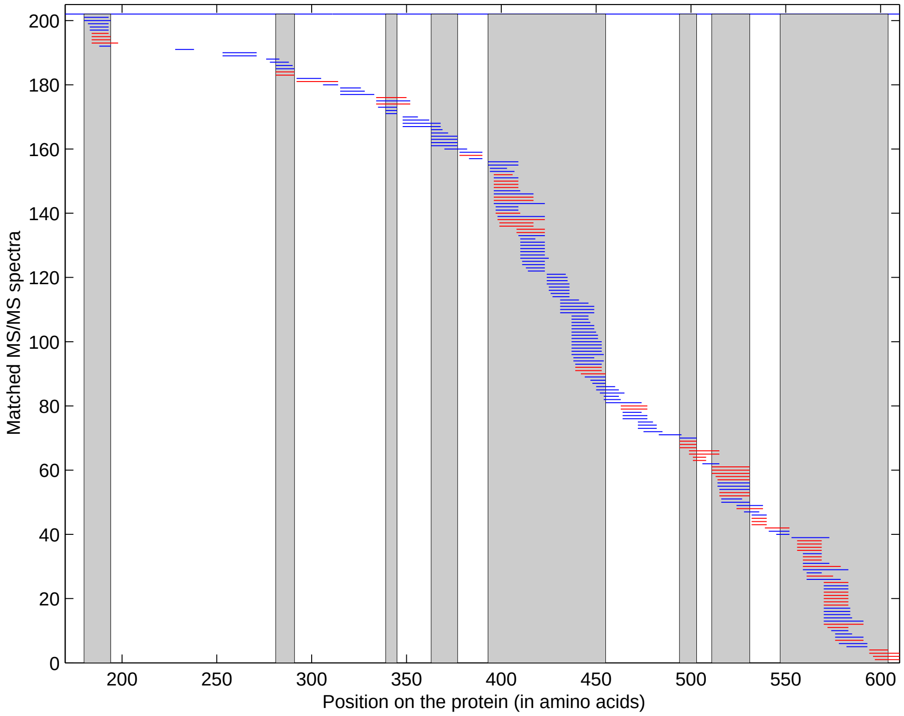


Figure C-5 Protein regions covered by assembled spectra from Vascular apoptosis-inducing protein 1 (Q9DGB9_CROAT).

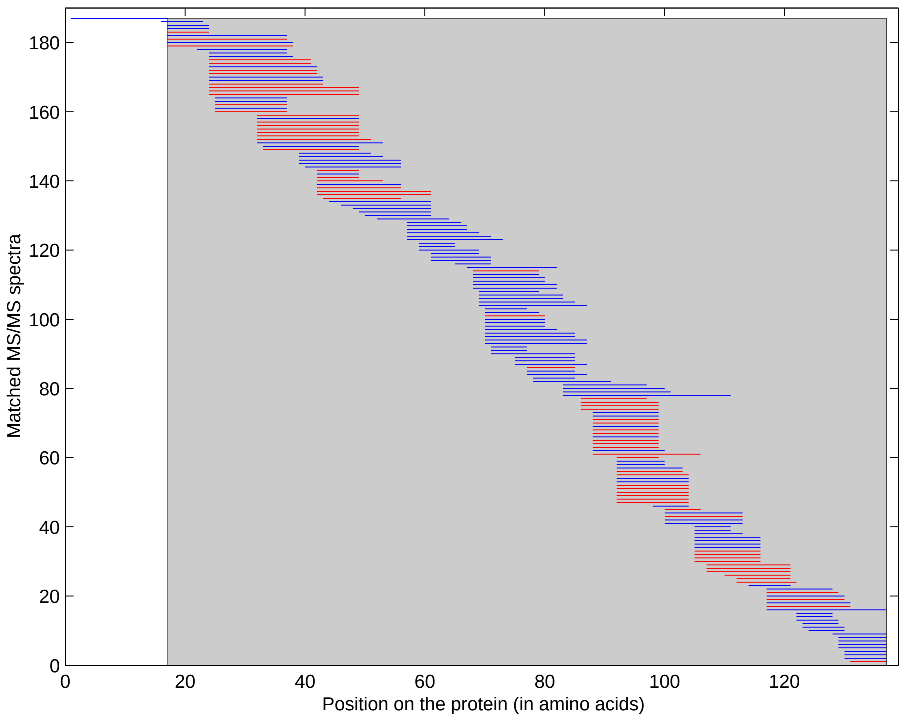


Figure C-6 Protein regions covered by assembled spectra from Phospholipase A2 homolog Cax-K49 (PA2H_CROAT).

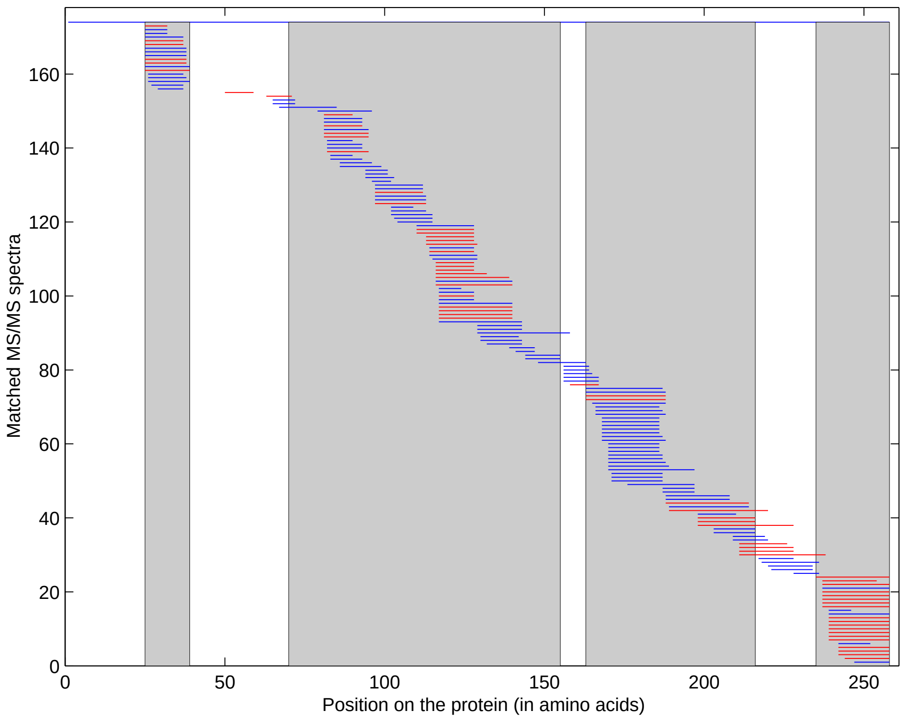


Figure C-7 Protein regions covered by assembled spectra from Serine protease catroxase II (Q8QHK2_CROAT).

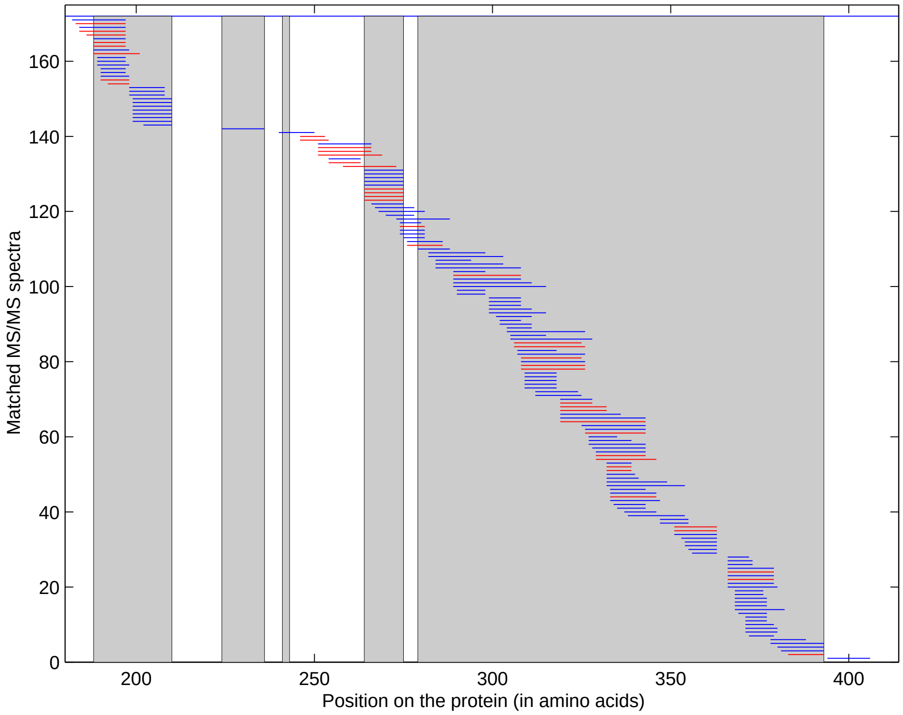


Figure C-8 Protein regions covered by assembled spectra from Hemorrhagic metalloproteinase HT-D/HT-C (HRTD_CROAT).

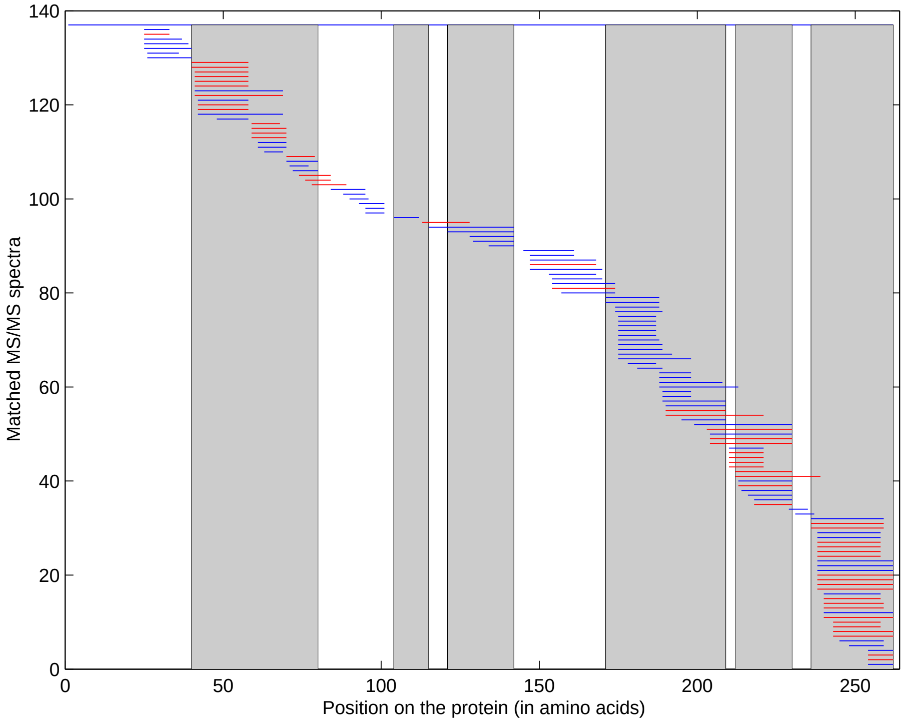


Figure C-9 Protein regions covered by assembled spectra from Serine protease catroxase I (Q8QHK3_CROAT).

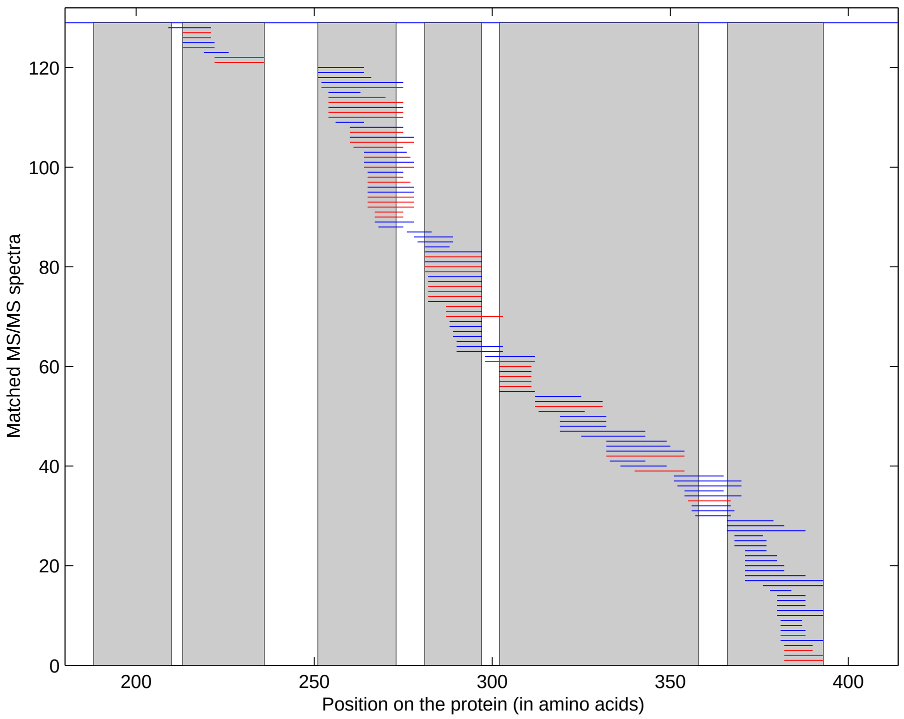


Figure C-10 Protein regions covered by assembled spectra from Prepro-hemorrhagic toxin b (Q90391_CROAT).

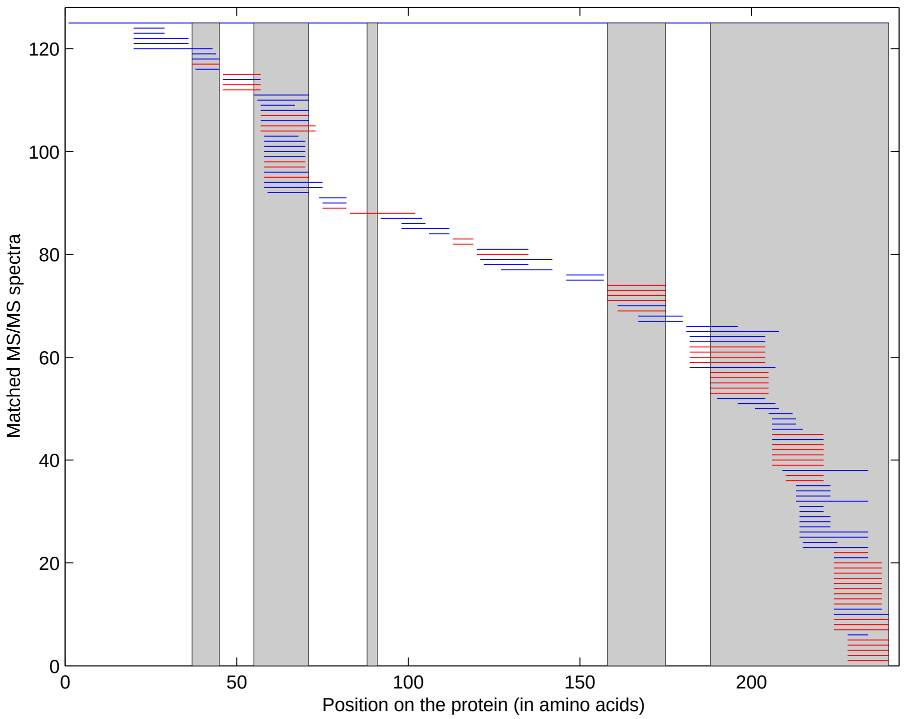


Figure C-11 Protein regions covered by assembled spectra from Catrin-1/2 (CRVP_CROAT).

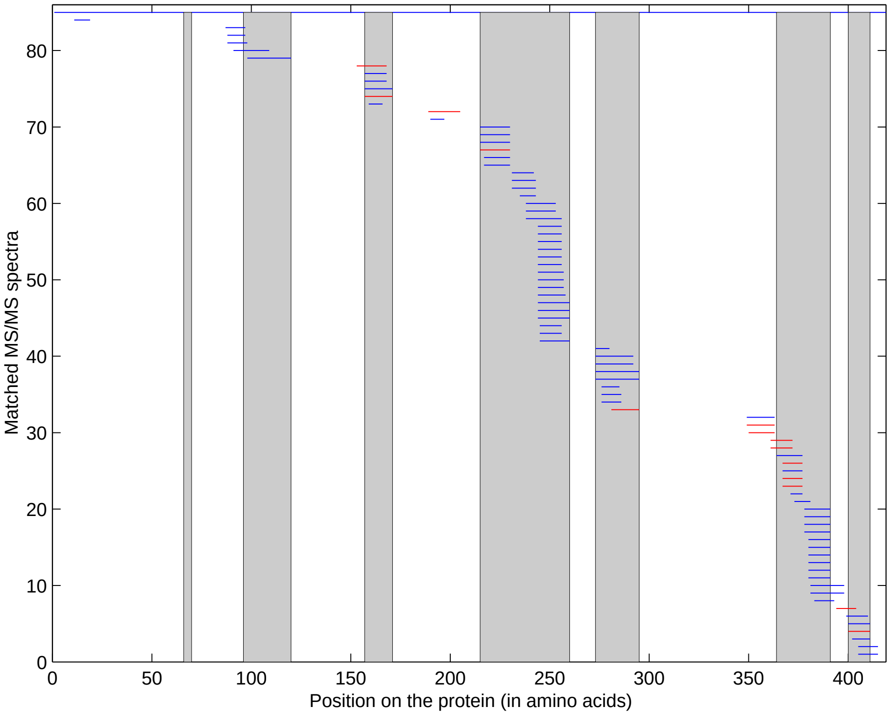


Figure C-12 Protein regions covered by assembled spectra from Hemorrhagic toxin a (Q92043_CROAT).

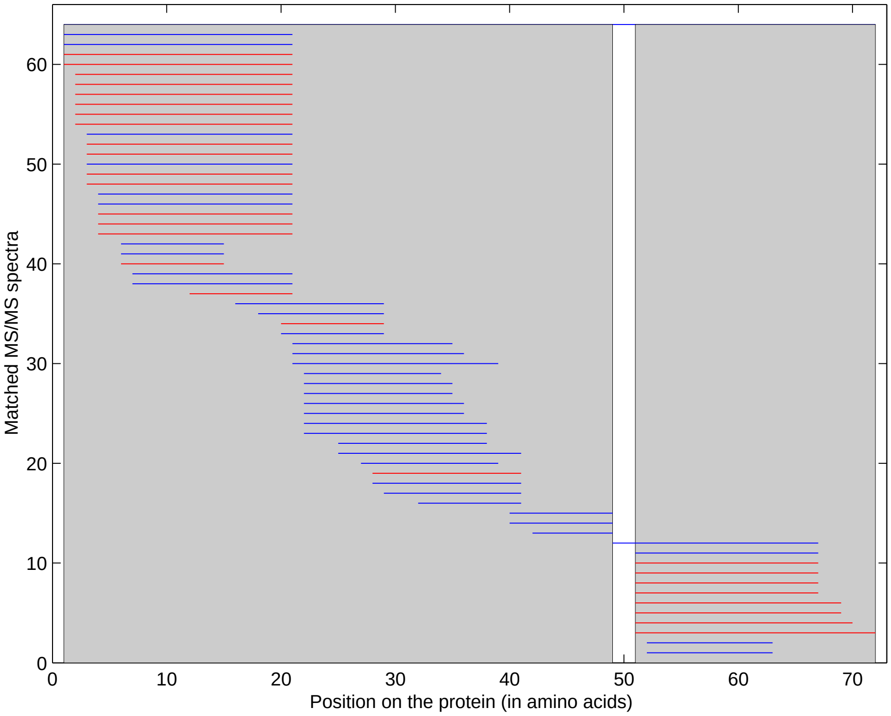


Figure C-13 Protein regions covered by assembled spectra from Platelet aggregation activation inhibitor (DISI_CROAT).

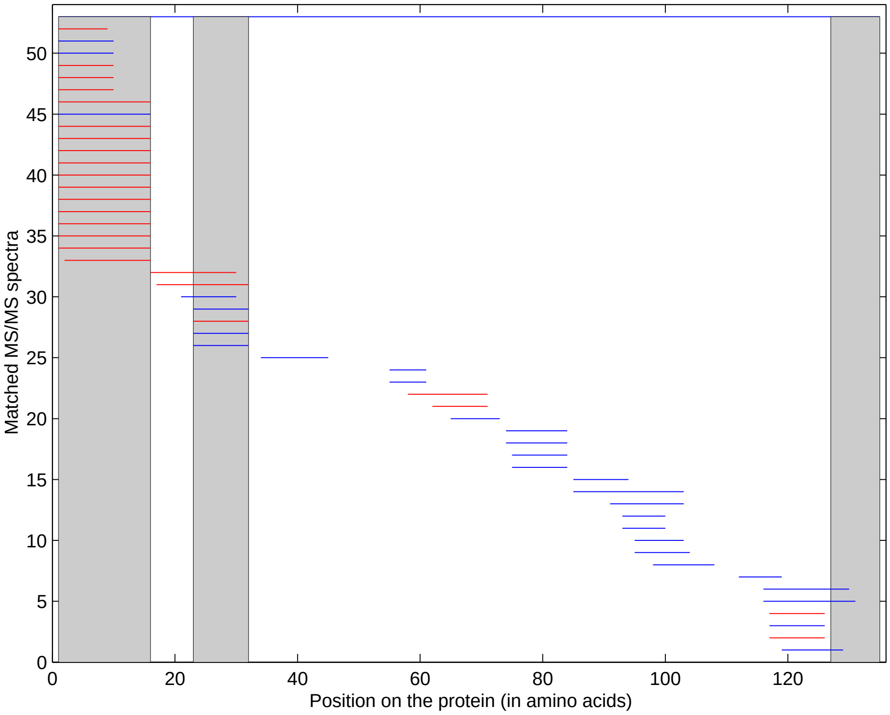


Figure C-14 Protein regions covered by assembled spectra from Galactose-specific lectin (LECG_CROAT).

Bibliography

[1] Frank A. and Pevzner P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.*, 77:964–973, 2005.

[2] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

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

[4] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, 1997.

[5] M C Arkan, A L Hevener, F R Greten, S Maeda, Z W Li, J M Long, A Wynshaw-Boris, G Poli, J Olefsky, and M Karin. Ikk-beta links inflammation to obesity-induced insulin resistance. *Nat Med*, 11:191–198, 2005.

[6] A Armirotti, E Millo, and G Damonte. How to discriminate between leucine and isoleucine by low energy esi-trap msn. *J Am Soc Mass Spectrom*, 18:57–63, 2007.

[7] V. Bafna and N. Edwards. On de-novo interpretation of tandem mass spectra for peptide identification. *Proceedings of the seventh annual international conference on Computational molecular biology*, pages 9–18, 2003.

[8] N. Bandeira, K. Clauser, and P.A. Pevzner. Shotgun Protein Sequencing: Assembly of Tandem Mass Spectra from Mixtures of Modified Proteins. *Mol Cell Proteomics*, 6:1123–34, 2007.

[9] N. Bandeira, H. Tang, V. Bafna, and P. Pevzner. Shotgun protein sequencing by tandem mass spectra assembly. *Analytical Chemistry*, 76:7221–7233, 2004.

[10] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. A New Approach to Protein Identification. In Apostolico A., Guerra C., Istrail S., Pevzner P.A., and Waterman M., editors, *Proceeding of the Tenth Annual International Conference in Research in Computational Molecular Biology (RECOMB 2006)*, volume 3909 of *Lecture Notes in Computer Science*, pages 363 – 378, 2006.

[11] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. Protein Identification via Spectral Networks Analysis. *Proc Natl Acad Sci U S A*, 104:6140–6145, 2007.

[12] C. Bartels. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomedical and Environmental Mass Spectrometry*, 19:363–8, 1990.

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

[14] A Ben-Dor, R Shamir, and Z Yakhini. Clustering gene expression patterns. *J Comput Biol*, 6:281–297, 1999.

[15] M Bern and D Goldberg. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J Comput Biol*, 13:364–378, 2006.

[16] M Bern, D Goldberg, W H McDonald, and J R Yates. Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, 20 Suppl 1:49–49, 2004.

[17] K. Biemann, C. Cone, BR. Webster, and GP. Arsenault. Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J Am Chem Soc*, 88:5598–5606, 1966.

[18] BE. Black, DR. Foltz, S. Chakravarthy, K. Luger, VL. Woods, and DW. Cleveland. Structural determinants for generating centromeric chromatin. *Nature*, 430:578–582, 2004.

[19] O Buczek, G Bulaj, and B M Olivera. Conotoxins and the posttranslational modification of secreted gene products. *Cell Mol Life Sci*, 62:3067–3079, 2005.

[20] D Cai, M Yuan, D F Frantz, P A Melendez, L Hansen, J Lee, and S E Shoelson. Local and systemic insulin resistance resulting from hepatic activation of ikk-beta and nf-kappab. *Nat Med*, 11:183–190, 2005.

[21] T. Chen, MY. Kao, M. Tepel, J. Rush, and GM. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol*, 8:325–337, 2001.

[22] E Cohen, M Datar, S Fujiwara, A Gionis, P Indyk, R Motwani, J D Ullman, and C Yang. Finding interesting associations without support pruning. *IEEE Trans Knowl Data Eng*, 13:64–78, 2001.

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

[24] D M Creasy and J S Cottrell. Unimod: Protein modifications for mass spectrometry. *Proteomics*, 4:1534–1536, 2004.

[25] J C Daltry, W Wüster, and R S Thorpe. Diet and snake venom evolution. *Nature*, 379:537–540, 1996.

[26] V. Dancík, TA. Addona, KR. Clauser, JE. Vath, and PA. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol*, 6:327–342, 1999.

[27] V. Dancík, S. Hannenhalli, and S. Muthukrishnan. Hardness of flip-cut problems from optical mapping. *Journal of Computational Biology*, 4:119–126, 1997.

[28] L L David, K J Lampi, A L Lund, and J B Smith. The sequence of human betab1-crystallin cdna allows mass spectrometric detection of betab1 protein missing portions of its n-terminal extension. *J Biol Chem*, 271:4273–4279, 1996.

[29] P A DiMaggio and C A Floudas. De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal Chem*, 79:1433–1446, 2007.

[30] M C Dos-Santos, E B Assis, T D Moreira, J Pinheiro, and C L Fortes-Dias. Individual venom variability in crotalus durissus ruruima snakes, a subspecies of crotalus durissus from the amazonian region. *Toxicon*, 46:958–961, 2005.

[31] D Dutta and T Chen. Speeding up tandem mass spectrometry database search: metric embeddings and fast near neighbor search. *Bioinformatics*, 23:612–618, 2007.

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

[33] JK. Eng, McCormack AL., and JR. Yates. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal Of The American Society For Mass Spectrometry*, 5:976–989, 1994.

[34] JJ. Englander, C. Del Mar, W. Li, SW. Englander, JS. Kim, DD. Stranz, Y. Hamuro, and VL. Woods. Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry. *Proc Natl Acad Sci U S A*, 100:7057–7062, 2003.

[35] P Escoubas. Mass spectrometry in toxinology: A 21st-century technology for the study of biopolymers from venoms. *Toxicon*, 47:609–613, 2006.

[36] Schutz F, Kapp EA, Simpson RJ, and Speed TP. Deriving statistical models for predicting peptide tandem ms product ion intensities. *Biochem Soc Trans*, pages 1479–83, 2003.

[37] John B. Fenn, Koichi Tanaka, and Kurt Wthrich. The nobel prize in chemistry 2002. http://nobelprize.org/chemistry/laureates/2002/chemadv02.pdf, 2002.

[38] J. Fernández-de Cossío, J. Gonzalez, and V. Besada. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Comput Appl Biosci*, 11:427–434, 1995.

[39] J. Fernandez-de Cossio, J. Gonzalez, L. Betancourt, V. Besada, G. Padron, Y. Shimonishi, and T. Takao. Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by 'SeqMS', a software aid for de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*, 12:1867–1878, 1998.

[40] J. Fernandez-de Cossio, J. Gonzalez, Y. Satomi, T. Shima, N. Okumura, V. Besada, L. Betancourt, G. Padron, Y. Shimonishi, and T. Takao. Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry. *Electrophoresis*, 21:1694–1699, 2000.

[41] B Fischer, V Roth, F Roos, J Grossmann, S Baginsky, P Widmayer, W Gruissem, and J M Buhmann. Novohmm: a hidden markov model for de novo peptide sequencing. *Anal Chem*, 77:7265–7273, 2005.

[42] J W Fox, L Ma, K Nelson, N E Sherman, and S M Serrano. Comparison of indirect and direct approaches using ion-trap and fourier transform ion cyclotron resonance mass spectrometry for exploring viperid venom proteomes. *Toxicon*, 47:700–714, 2006.

[43] A. Frank, S.W. Tanner, V. Bafna, and P.A. Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *J. of Proteome Research*, 4:1287–95, 2005.

[44] A M Frank, N Bandeira, Z Shen, S Tanner, S P Briggs, R D Smith, and P A Pevzner. Clustering tandem mass spectra: From spectral libraries to spectral archives. *J Proteome Res*, 6:114–123, 2007.

[45] A M Frank, M M Savitski, M L Nielsen, R A Zubarev, and P A Pevzner. De novo peptide sequencing and identification with precision mass spectrometry. *J Proteome Res*, 6:114–123, 2007.

[46] Siuzdak G. *Mass Spectrometry in Biotechnology.* MCC Press, San Diego, 2003.

[47] P J Gearhart. Immunology: the roots of antibody diversity. *Nature*, 419:29–31, 2002.

[48] KF Geoghegan, LR Hoth, DH Tan, KA Borzilleri, JM Withka, and JG Boyd. Cyclization of n-terminal s-carbamoylmethylcysteine causing loss of 17 da from peptides and extra peaks in peptide maps. *J Proteome Res.*, 1:181–7, 2002.

[49] I C Guerrera and O Kleiner. Application of mass spectrometry in proteomics. *Biosci Rep*, 25:71–93, 2005.

[50] Cormen T H, Charles C E, Rivest R L, and Stein C. *Introduction to Algorithms.* The MIT Pres, 2001.

[51] Y. Hamuro, GS. Anand, JS. Kim, C. Juliano, DD. Stranz, SS. Taylor, and VL. Woods. Mapping intersubunit interactions of the regulatory subunit (RIalpha) in the type I holoenzyme of protein kinase A by amide hydrogen/deuterium exchange mass spectrometry (DXMS). *J Mol Biol*, 340:1185–1196, 2004.

[52] Y. Hamuro, L. Burns, J. Canaves, R. Hoffman, S. Taylor, and V. Woods. Domain organization of D-AKAP2 revealed by enhanced deuterium exchange-mass spectrometry (DXMS). *J Mol Biol*, 321:703–714, 2002.

[53] Y. Hamuro, KM. Zawadzki, JS. Kim, DD. Stranz, SS. Taylor, and VL. Woods. Dynamics of cAPK type IIbeta activation revealed by enhanced amide H/2H exchange mass spectrometry (DXMS). *J Mol Biol*, 327:1065–1076, 2003.

[54] Y Han, B Ma, and K Zhang. Spider: software for protein identification from sequence tags with de novo sequencing error. *J Bioinform Comput Biol*, 3:697–716, 2005.

[55] B.T. Hansen, S.W. Davey, A.J. Ham, and Liebler. D.C. P-mod: an algorithm and software to map modifications to peptide sequences using tandem ms data. *J Proteome Res.*, 4:358–68, 2005.

[56] J S Haurum. Recombinant polyclonal antibodies: the next generation of antibody therapeutics? *Drug Discov Today*, 11:655–660, 2006.

[57] T H Haveliwala, A Gionis, and P Indyk. Scalable techniques for clustering the web. *WebDB 2000*, pages 129–134, 2000.

[58] W J Henzel, T M Billeci, J T Stults, S C Wong, C Grimley, and C Watanabe. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A*, 90:5011–5015, 1993.

[59] S. Hopper, RS. Johnson, JE. Vath, and K. Biemann. Glutaredoxin from rabbit bone marrow. Purification, characterization, and amino acid sequence determined by tandem mass spectrometry. *J Biol Chem*, 264:20438–20447, 1989.

[60] M C Hu and M C Hung. Role of IkappaB kinase in tumorigenesis. *Future Oncol*, 1:67–78, 2005.

[61] Y Huang, J M Triscari, G C Tseng, L Pasa-Tolic, M S Lipton, R D Smith, and V H Wysocki. Statistical characterization of the charge state and residue dependence of low-energy cid peptide dissociation patterns. *Anal Chem*, 77:5800–5813, 2005.

[62] Y. Huang, V.H. Wysocki, D.L. Tabb, and J.R. Yates III. Histidine effect on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *Int. J. Mass Spectrom*, 2002.

[63] E. Hunyadi-Gulyas and K. Medzihradszky. Factors that contribute to the complexity of protein digests. *Drug Discovey Today: Targets - mass spectrometry in proteomics supplement*, 3(2):3–10, 2004.

[64] K. Ishikawa and Y. Niwa. Computeraided peptide sequencing by fastatombombardment massspectrometry. *Biomed Environ Mass Spectrom*, 13:373–380, 1986.

[65] Colinge J, Magnin J, Dessingy T, Giron M, and Masselot A. Improved peptide charge state assignment. *Proteomics*, 3:1434–1440, 2003.

[66] O N Jensen. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol*, 7:391–403, 2006.

[67] RS. Johnson and K. Biemann. The primary structure of thioredoxin from Chromatium vinosum determined by high-performance tandem mass spectrometry. *Biochemistry*, 26:1209–14, 1987.

[68] RS. Johnson and K. Biemann. Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed Environ Mass Spectrom*, 18:945–957, 1989.

[69] RS. Johnson, MT. Davis, JA. Taylor, and SD. Patterson. Informatics for protein identification by mass spectrometry. *Methods*, 35:223–236, 2005.

[70] RS. Johnson and JA. Taylor. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Methods Mol Biol*, 146:41–61, 2000.

[71] JS. Joseph and RM. Kini. Snake venom prothrombin activators similar to blood coagulation factor Xa. *Curr Drug Targets Cardiovasc Haematol Disord*, 4:397–416, 2004.

[72] David R. Karger, Rajeev Motwani, and G. D. S. Ramkumar. On approximating the longest path in a graph. *Algorithmica*, 18:82–98, 1997.

[73] R. Karp and R. Shamir. Algorithms for optical mapping. *J Comput Biol*, 7:303–316, 2000.

[74] A. Keller, AI. 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, 2002.

[75] A. Keller, S. Purvine, AI. Nesvizhskii, S. Stolyar, DR. Goodlett, and E. Kolker. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, 6:207–212, 2002.

[76] RM. Kini, VS. Rao, and JS. Joseph. Procoagulant proteins from snake venoms. *Haemostasis*, 31:218–224, 2001.

[77] A A Klammer and M J MacCoss. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J Proteome Res*, 5:695–700, 2006.

[78] ES. Lander and et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[79] V N Lapko, D L Smith, and J B Smith. Identification of an artifact in the mass spectrometry of proteins derivatized with iodoacetamide. *J Mass Spectrom*, 35:572–575, 2000.

[80] J. K. Lee, V. Dancík, and M. S. Waterman. Estimation for restriction sites observed by optical mapping using reversible-jump Markov Chain Monte Carlo. *J Comput Biol*, 5:505–515, 1998.

[81] A Leitner and W Lindner. Current chemical tagging strategies for proteome analysis by mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*, 813:1–26, 2004.

[82] R J Lewis and M L Garcia. Therapeutic potential of venom peptides. *Nat Rev Drug Discov*, 2:790–802, 2003.

[83] DC. Liebler, BT. Hansen, SW. Davey, L. Tiscareno, and DE. Mason. Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. *Anal Chem*, 74:203–210, 2002.

[84] D. Lin, DL. Tabb, and JR. Yates. Large-scale protein identification using mass spectrometry. *Biochim Biophys Acta*, 1646:1–10, 2003.

[85] T Lin and G L Glish. C-terminal peptide sequencing via multistage mass spectrometry. *Anal Chem*, 70:5162–5165, 1998.

[86] O. Lubeck, C. Sewell, S. Gu, X. Chen, and D. Cai. New computational approaches for de novo peptide sequencing from MS/MS experiments. *PROCEEDINGS OF THE IEEE*, 90:1868–1874, 2002.

[87] Havilio M, Haddad Y, and Smilansky Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem*, 75:435–444, 2003.

[88] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*, 17:2337–2342, 2003.

[89] MJ. MacCoss, WH. McDonald, A. Saraf, R. Sadygov, JM. Clark, JJ. Tasto, KL. Gould, D. Wolters, M. Washburn, A. Weiss, JI. Clark, and JR. Yates. Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc Natl Acad Sci U S A*, 99:7900–7905, 2002.

[90] AJ. Mackey, TA. Haystead, and WR. Pearson. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol Cell Proteomics*, 1:139–147, 2002.

[91] N Maizels. Immunoglobulin gene diversification. *Annu Rev Genet*, 39:23–46, 2005.

[92] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66:4390–4399, 1994.

[93] FS. Markland, K. Shieh, Q. Zhou, V. Golubkov, RP. Sherwin, V. Richters, and R. Sposto. A novel snake venom disintegrin that inhibits human ovarian cancer dissemination and angiogenesis in an orthotopic nude mouse model. *Haemostasis*, 31:183–191, 2001.

[94] M C Menezes, M F Furtado, S R Travaglia-Cardoso, A C Camargo, and S M Serrano. Sex-based individual variation of snake venom proteome among eighteen bothrops jararaca siblings. *Toxicon*, 47:304–312, 2006.

[95] B S Miller and E Zandi. Complete reconstitution of human ikappab kinase (ikk) complex in yeast. assessment of its stoichiometry and the role of ikkgamma on the complex activity in the absence of stimulation. *J Biol Chem*, 276:36320–36326, 2001.

[96] E W Myers. Toward simplifying and accurately formulating fragment assembly. *J Comput Biol*, 2:275–290, 1995.

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

[98] Y. Ogawa, R. Yanoshita, U. Kuch, Y. Samejima, and D. Mebs. Complete amino acid sequence and phylogenetic analysis of a long-chain neurotoxin from the venom of the African banded water cobra, Boulengerina annulata. *Toxicon*, 43:855–858, 2004.

[99] J V Olsen and M Mann. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A*, 101:13417–13422, 2004.

[100] S K Pal, A Gomes, S C Dasgupta, and A Gomes. Snake venom as therapeutic agents: from toxin to drug development. *Indian J Exp Biol*, 40:1353–1358, 2002.

[101] D. Pantazatos, JS. Kim, HE. Klock, RC. Stevens, IA. Wilson, SA. Lesley, and VL. Woods. Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. *Proc Natl Acad Sci U S A*, 101:751–756, 2004.

[102] D Pantazatos, J.S. Kim, H.E. Klock, R.C. Stevens, I.A. Wilson, S.A. Lesley, and V.L. Woods. Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange ms. *Proc Natl Acad Sci USA*, 101(3):751–6, 2004.

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

[104] P A Pevzner, H Tang, and G Tesler. De novo repeat classification and fragment assembly. *Genome Res*, 14:1786–1796, 2004.

[105] PA. Pevzner, V. Dancík, and CL. Tang. Mutation-tolerant protein identification by mass spectrometry. *J Comput Biol*, 7:777–787, 2000.

[106] PA. Pevzner, Z. Mulyukov, V. Dancik, and CL. Tang. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res*, 11:290–299, 2001.

[107] V Pham, W J Henzel, D Arnott, S Hymowitz, W N Sandoval, B T Truong, H Lowman, and J R Lill. De novo proteomic sequencing of a monoclonal antibody raised against ox40 ligand. *Anal Biochem*, 352:77–86, 2006.

[108] A M Pimenta and M E De Lima. Small peptides, big world: biotechnological potential in neglected bioactive peptides from arthropod venoms. *J Pept Sci*, 11:670–676, 2005.

[109] A M Pimenta, B Rates, C Bloch, P C Gomes, M M Santoro, M E de Lima, M Richardson, and M d o N Cordeiro. Electrospray ionization quadrupole time-of-flight and matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometric analyses to solve micro-heterogeneity in post-translationally modified peptides from phoneutria nigriventer (aranea, ctenidae) venom. *Rapid Commun Mass Spectrom*, 19:31–37, 2005.

[110] Durbin R., Eddy S. R., Krogh A., and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1999.

[111] T. Sakurai et al. PAAS 3, a computer program to determine probable sequence of peptides from mass spectrometric data. *Biomed Mass Spectrom*, 11:396–399, 1984.

[112] M M Savitski, M L Nielsen, F Kjeldsen, and R A Zubarev. Proteomics-grade de novo sequencing approach. *J Proteome Res*, 4:2348–2354, 2005.

[113] M M Savitski, M L Nielsen, and R A Zubarev. New data base-independent, sequence tag-based scoring of peptide ms/ms data validates mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of ms/ms techniques. *Mol Cell Proteomics*, 4:1180–1188, 2005.

[114] 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.

[115] B C Searle, S Dasari, P A Wilmarth, M Turner, A P Reddy, L L David, and S R Nagalla. Identification of protein modifications using ms/ms de novo sequencing and the opensea alignment algorithm. *J Proteome Res*, 4:546–554, 2005.

[116] A Shevchenko, I Chernushevich, W Ens, K G Standing, B Thomson, M Wilm, and M Mann. Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom*, 11:1015–1024, 1997.

[117] A. Shevchenko, A. Loboda, S. Sunyaev, A. Shevchenko, P. Bork, W. Ens, and K.G. Standing. Charting the proteomes of organisms with unsequenced genomes by MALDI-Quadrupole Time-of Flight Mass Spectrometry and BLAST homology searching. *Analytical Chemistry*, 73:1917–1926, 2001.

[118] MM. Siegel and N. Bauman. An efficient algorithm for sequencing peptides using fast atom bombardment mass spectral data. *Biomed Environ Mass Spectrom*, 15:333–343, 1988.

[119] T F Smith and M S Waterman. Identification of common molecular subsequences. *J Mol Biol*, pages 195–197, 1981.

[120] M R Soares, A L Oliveira-Carvalho, L S Wermelinger, R B Zingali, P L Ho, I d e L Junqueira-de Azevedo, and M R Diniz. Identification of novel bradykinin-potentiating peptides and c-type natriuretic peptide from lachesis muta venom. *Toxicon*, 46:31–38, 2005.

[121] S. Sunyaev, AJ. Liska, A. Golod, A. Shevchenko, and A. Shevchenko. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem*, 75:1307–1315, 2003.

[122] S. Swenson, F. Costa, R. Minea, RP. Sherwin, W. Ernst, G. Fujii, D. Yang, and FS. Markland. Intravenous liposomal delivery of the snake venom disintegrin contortrostatin limits breast cancer progression. *Mol Cancer Ther*, 3:499–511, 2004.

[123] S. Swenson, CF. Toombs, L. Pena, J. Johansson, and FS. Markland. Alpha-fibrinogenases. *Curr Drug Targets Cardiovasc Haematol Disord*, 4:417–435, 2004.

[124] J E Syka, J J Coon, M J Schroeder, J Shabanowitz, and D F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A*, 101:9528–9533, 2004.

[125] DL Tabb, MJ MacCoss, CC Wu, SD Anderson, and JR 3rd. Yates. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem*, 75:2470–7, 2003.

[126] DL. Tabb, A. Saraf, and JR. Yates. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem*, 75:6415–6421, 2003.

[127] DL. Tabb, LL. Smith, LA. Breci, VH. Wysocki, D. Lin, and JR. Yates. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem*, 75:1155–1163, 2003.

[128] O Takikawa, R J Truscott, M Fukao, and S Miwa. Age-related nuclear cataract and indoleamine 2,3-dioxygenase-initiated tryptophan metabolism in the human lens. *Adv Exp Med Biol*, 527:277–285, 2003.

[129] W H Tang, B R Halpern, I V Shilov, S L Seymour, S P Keating, A Loboda, A A Patel, D A Schaeffer, and L M Nuwaysir. Discovering known and unanticipated protein modifications using ms/ms database searching. *Anal Chem*, 77:3931–3946, 2005.

[130] S. Tanner, H. Shu, A. Frank, LC. Wang, E. Zandi, M. Mumby, PA. Pevzner, and V. Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem*, 77:4626–4639, 2005.

[131] JA. Taylor and RS. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*, 11:1067–1075, 1997.

[132] JA. Taylor, KA. Walsh, and RS. Johnson. Sherpa: a Macintosh-based expert system for the interpretation of electrospray ionization LC/MS and MS/MS data from protein digests. *Rapid Commun Mass Spectrom*, 10:679–687, 1996.

[133] D Tsur, S Tanner, E Zandi, V Bafna, and P A Pevzner. Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol*, 23:1562–1567, 2005.

[134] JD. Venable and JR. Yates. Impact of ion trap tandem mass spectra variability on the identification of peptides. *Anal Chem*, 76:2928–2937, 2004.

[135] JC. Venter and et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[136] J P Vinson, D B Jaffe, K O'Neill, E K Karlsson, N Stange-Thomann, S Anderson, J P Mesirov, N Satoh, Y Satou, C Nusbaum, B Birren, J E Galagan, and E S Lander. Assembly of polymorphic genomes: algorithms and application to ciona savignyi. *Genome Res*, 15:1127–1135, 2005.

[137] G F Vrensen, J van Marle, R Jonges, W Voorhout, W Breipohl, and A R Wegener. Tryptophan deficiency arrests chromatin breakdown in secondary lens fibers of rats. *Exp Eye Res*, 78:661–672, 2004.

[138] L S Wermelinger, D L Dutra, A L Oliveira-Carvalho, M R Soares, C Bloch, and R B Zingali. Fast analysis of low molecular mass compounds present in snake venom: identification of ten new pyroglutamate-containing peptides. *Rapid Commun Mass Spectrom*, 19:1703–1708, 2005.

[139] M Wiles and P Andreassen. Monoclonals - the billion dollar molecules of the future. *Drug Discov World*, Fall 2006:17–23, 2006.

[140] 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.

[141] VL. Woods and Y. Hamuro. High resolution, high-throughput amide deuterium exchange-mass spectrometry (DXMS) determination of protein binding site structure and dynamics: utility in pharmaceutical design. *J Cell Biochem Suppl*, Suppl 37:89–98, 2001.

[142] Cannon WR and Jarman KD. Improved peptide sequencing using isotope information inherent in tandem mass spectra. *Rapid Commun Mass Spectrom*, 17:1793–1801, 2003.

[143] J R Yates. Mass spectrometry as an emerging tool for systems biology. *Biotechniques*, 36:917–919, 2004.

[144] JR. Yates, JK. Eng, and AL. McCormack. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*, 67:3202–3210, 1995.

[145] Z. Zhang and JS. McElvain. De novo peptide sequencing by two-dimensional fragment correlation mass spectrometry. *Anal Chem*, 72:2337–2350, 2000.

[146] Q Zhou, J B Smith, and M H Grossman. Molecular cloning and expression of catrocollastatin, a snake-venom protein from crotalus atrox (western diamondback rattlesnake) which inhibits platelet adhesion to collagen. *Biochem J*, 307 (Pt 2):411–417, 1995.

[147] D. Zidarov, P. Thibault, MJ. Evans, and MJ. Bertrand. Determination of the primary structure of peptides using fast atom bombardment mass spectrometry. *Biomed Environ Mass Spectrom*, 19:13–26, 1990.

[148] A Zugasti-Cruz, M Maillo, E López-Vera, A Falcón, E P Heimer de la Cotera, B M Olivera, and M B Aguilar. Amino acid sequence and biological activity of a gamma-conotoxin-like peptide from the worm-hunting snail conus austini. *Peptides*, 27:506–511, 2006.