Data Analysis in Proteomics

Novel computational strategies for modeling and interpreting complex mass spectrometry data.

by

Matthew John Sniatynski

B.Sc. Biology, The University of British Columbia, 2001
B.Sc. Computer Science, The University of British Columbia, 2005

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Experimental Medicine)

The University Of British Columbia

(Vancouver)

August, 2008

Abstract

Contemporary proteomics studies require computational approaches to deal with both the complexity of the data generated, and with the volume of data produced. The amalgamation of mass spectrometry – the analytical tool of choice in proteomics – with the computational and statistical sciences is still recent, and several avenues of exploratory data analysis and statistical methodology remain relatively unexplored. The current study focuses on three broad analytical domains, and develops novel exploratory approaches and practical tools in each.

Data transform approaches are the first explored. These methods reframe data, allowing for the visualization and exploitation of features and trends that are not immediately evident. An exploratory approach making use of the correlation transform is developed, and is used to identify mass-shift signals in mass spectra. This approach is used to identify and map post-translational modifications on individual peptides, and to identify SILAC modification-containing spectra in a full-scale proteomic analysis.

Secondly, matrix decomposition and projection approaches are explored; these use an eigen-decomposition to extract general trends from groups of related spectra. A data visualization approach is demonstrated using these techniques, capable of visualizing trends in large numbers of complex spectra, and a data compression and feature extraction technique is developed suitable for use in spectral modeling.

Finally, a general machine learning approach is developed based on conditional random fields (CRFs). These models are capable of dealing with arbitrary sequence modeling tasks, similar to hidden Markov models (HMMs), but are far more robust to interdependent observational features, and do not require limiting independence assumptions to remain tractable. The theory behind this approach is developed, and a simple machine learning fragmentation model is developed to test the hypothesis that reproducible sequence-specific intensity ratios are present within the distribution of fragment ions originating from a common peptide bond breakage. After training, the model shows very good performance associating peptide sequences and fragment ion intensity information, lending strong support to the hypothesis.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to extend my eternal gratitude to my supervisor, Dr. Juergen Kast, for his support, and for his patience. The freedom I was given to explore made my graduate degree into an academic experience without equal. Thanks for your patience, Juergen, and for believing in me.

A special thanks to Jason C. Rogalski, without whom none of this work would have been possible. Thanks for tolerating the tangents, the demands, the barrages of questions, and the theft of a great deal of your coffee.

Thanks to Ron Beavis and Leonard Foster for the supervisory committee support.

Thanks to the Kast group for the camaraderie, the support, the good times, and the stories. There is no better group of co-workers.

I'd also like to thank my family for being so incredibly supportive and understanding for the past few years, and I'd especially like to thank Charlotte, the kindest and most understanding best-friend anyone could ever hope to have.

Statement of Co-Authorship

As regards the research presented in the first manuscript (Correlation and Convolution Analysis of Peptide Mass Spectra), I assisted in the experimental design, and developed the analysis software used. The manuscript was written by me with input from Jason C. Rogalski, save for the Materials and Methods section, which was written mostly by Jason C. Rogalski. Jason also assisted me with the preparation of the data/spectra in the figures, with the interpretation and labeling of the figures, and with the preparation of the table of mass shifts.

As regards the research presented in the second manuscript (Singular Value Decomposition and Principal Components Analysis for MS/MS Feature Detection and Fragmentation Modeling), I was responsible for the experimental design, the writing of the manuscript, and for the preparation of the figures.

As regards the research presented in the third manuscript (Modeling Peptide Fragmentation Using Conditional Random Fields), I was assisted by Jason C. Rogalski in the experimental design, and was responsible for the implementation of all the algorithms described, the development of the software described, and for the preparation of the manuscript and its accompanying tables and figures.

Chapter 1

Introduction

1.1 Prologue

1.1.1 Biology, Complexity, and Computational Analysis

In recent years, the study of biological systems has undergone a fundamental shift in its goals. In addition to the study of individual components of complex systems, the holistic study of the systems themselves in their entirety has become increasingly popular. This shift is not representative of a philosophical move away from reductionist thinking in the scientific community; rather, it reflects the increasing speed and accuracy with which information on low-level biological system components can be acquired. This increase builds upon the detailed understanding of the systems in question, and is driven mainly by ever-accelerating technological progress on two fronts. The first of these fronts is the instrumentation needed to acquire the vast amounts of biological data necessary, and the second front is the computational machinery needed to interpret the data acquired. Substantial progress in both of these areas has led to the sequencing of the human genome (and multiple other genomes), the establishment of interdisciplinary fields such as "bioinformatics", and has enabled the rise of the new research disciplines of genomics, proteomics, and systems biology. Analytical techniques from these new domains have the potential to revolutionize biological research, as they permit the study of complex systems such as cells, organs, or entire organisms, as holistic entities. Many potential applications for such holistic models exist in clinical diagnosis, where complex and subtle shifts in proteomic state may serve as disease state biomarkers, and convey other therapeutic or prognostic information – this is already a very active area of research. Other applications are to be found in drug discovery, where a more thorough understanding of cell signalling and cellular interaction networks could be used to better direct discovery work, reducing the enormous expenditures that today's trial-and-error approaches typically incur [1, 2].

Many challenges remain before these objectives are realized. Although accurate, high throughput genomic characterizations are becoming more

common (microarray analyses, for example), results generated by these techniques lack sufficient detail to map more intricate and dynamic cellular processes such as cell signaling. For this reason, proteomic approaches are becoming critically important in systems biology [3]. Proteomics in particular has been driven by technological progress in instrumentation for data collection, and computational approaches for data analysis and interpretation. The scope for computational analysis in proteomics is extremely wide, yet the amalgamation of the computer sciences with proteomic analysis was sufficiently recent that many useful and general approaches to data analysis remain relatively unexplored.

The remainder of this chapter introduces the field of proteomics research, describes the instrumentation commonly used, and the data these instruments generate. Computational interpretation of mass spectrometry data is introduced, and common areas where computational strategies are currently employed in data interpretation are surveyed. This discussion is used to motivate the need for further exploratory work in the computational processing of proteomics data, and to introduce several general, novel data analysis strategies that have shown considerable future promise.

1.1.2 Proteomics

Proteomics is the study of the entire complement of proteins present in a particular cell, tissue, or other biological unit, under certain conditions, and at a particular time. Since proteins are critically important functional and structural units of living cells, the full characterization of a cell's proteome, in theory, would provide a complete snapshot of the biochemical state of that cell. In contrast to techniques in genomics, which measure transcribed mRNA molecules before they are translated to protein products, proteomic characterization is direct; the effector molecules that govern the cell are themselves the targets of analysis. This yields richer and more detailed information than the examination of transcribed mRNA molecules, as substantial changes may occur between mRNA transcript and final protein product.

Various forms of mRNA processing often occur between transcription and translation, such as alternate splicing, which enables many functionally different proteins to be produced from the same transcript, and therefore from the same gene. The identity of an mRNA molecule may therefore yield insufficiently detailed information on the identities of the proteins it encodes [4]. Additionally, the amount of mRNA present for a given gene is often a poor proxy for the amount of final protein product, as mRNA molecules

often give rise to multiple translated proteins, often differing widely in the total number of these produced [5, 6, 7]. Finally, protein post-translational modifications – including phosphorylation, methylation, and the addition of complex sugar or lipid moieties – are another very important phenomenon, conferring critical information about the biochemical state of a cell [8, 9, 10]. These modifications affect the localization, activity, and complexing affinity of a diverse array of proteins, and are the principal effectors in cell signaling.

Due to the highly heterogeneous nature of proteins, obtaining a precise and sufficiently detailed characterization of every protein simultaneously is a very complicated problem. Unlike genomic analysis, there is no simple, inexpensive, and high throughput technique equivalent to a microarray analysis. Though many analysis methods have been demonstrated for the quantification and characterization of proteins, the majority of these rely on the use of specific antibodies and stringent purification protocols [11, 12] – these techniques scale poorly to problems of proteomic magnitude due to their speed, variable specificity, and expense. The only currently available technology with sufficient throughput, flexibility, and generality for large proteomics studies is mass spectrometry.

1.2 Mass Spectrometry

Mass spectrometry is a well established, general purpose analytical tool, with many potential uses, and a diverse array of configurations that are amenable to many different experimental designs. It identifies and characterizes unknown analyte molecules by precisely measuring the mass-to-charge ratios of ions produced from the analyte. The accuracy and precision of the instruments used provides mass-to-charge ratios, and therefore determines masses, with sufficient precision to uniquely identify individual molecules, and to identify the components of complex mixtures.

The term "Mass Spectrometry" itself is a general term that describes a diverse array of instrument types and analysis protocols. The techniques used to produce ions from the sample are decoupled from the techniques used to obtain their masses – many different methods for ionization and mass analysis are common, and may be found together in many different combinations. Considerable technological progress has occurred in both areas since the inception of the field of proteomics, and this progress shows no signs of abating. Ion sources are continually becoming more adept at efficiently ionizing molecules of all masses and compositions, and mass analyzers are providing ever higher mass-to-charge resolution. This is enabling

more precise identification of an increasing array of molecules due to ionization and mass-accuracy improvements, and successful analysis of more complex mixtures due to increases in mass analyzer resolution, enabling discrimination between molecules with very similar mass-to-charge ratios.

1.2.1 Mass Spectrometry for Proteomics

Several innovations in mass spectrometry were crucial to its application to proteomics research, as proteomic analysis principally involves characterizing enzymatically-derived protein fragments known as peptides. These are large, complex biological molecules that may range in size from a few hundred Daltons to many thousands of Daltons. An important enabling innovation was the popularization of time-of-flight (TOF) analyzers, as these enabled the accurate mass-to-change determination of these very large molecules at high resolution – a combination that was unachievable using older quadrupole or magnetic sector technology. The most important enabling innovation, however, was the introduction of soft ionization methods. Many older, "hard" ionization techniques such as fast atom bombardment (FAB), or electron impact (EI) proved far too energetic to effectively ionize large biomolecules such as peptides without breaking their chemical bonds, producing mostly indistinguishable fragment ions. Soft ionization methods are able to effectively transfer charge to large molecules without such a destructive effect. The two most popular soft ionization methods currently are electrospray ionization (ESI), and matrix-assisted laser desorption/ionization (MALDI).

The following sections describe two soft ionization methods, and two mass analysis methods that are particularly well suited to high resolution proteomic studies. The addition of chromatographic or gel-based separation before mass spectrometry is also briefly discussed in the context of a standard mass spectrometry-based proteomics workflow. A similar workflow, using the soft ionization and mass analysis approaches described, was used to generate the mass spectrometry data used in this study.

Electrospray Ionization (ESI)

Electrospray ionization was pioneered by Dr. John Bennet Fenn in the 1980s for the analysis of large biological molecules. It forms ions from the solvent protonation of basic regions on the analyte molecules. The analyte is dissolved in suitable solvent (for example, 50:50 methanol:water with a small percentage of acid), and is sprayed through a capillary/needle which is biased

at high potential (several hundred V, to multiple kV) relative to the entry point of the mass spectrometer (or intermediate skimmer plate). This causes appropriately charged droplets to be drawn away from the needle, towards the mass spectrometer. The neutral solvent in the droplet evaporates as it travels (sometimes aided by a nebulizing gas), and as the analyte ions are forced closer together, Coulombic fission divides the droplet repeatedly, until individual, solvent-free ions remain. These lone ions then enter the mass analyzer. The ions produced using this technique are often present at several different charge states. Figure 1.1 (top) shows a simplified electrospray schematic.

Matrix-Assisted Laser Desorption Ionization (MALDI)

Matrix-assisted laser desorption/ionization, first demonstrated for large biological molecules by Koichi Tanaka in 1987, forms ions via a two-stage process: laser volatilization/excitation followed by matrix-mediated charge transfer. The analyte is mixed with a suitable MALDI matrix (such as sinapinic acid, or 2,5-dihydroxybenzoic acid), spotted on a metal plate, and allowed to dry and recrystallize. During analysis, the analyte/matrix spot is subjected to a direct laser pulse at a wavelength chosen to optimally excite the matrix molecules. This pulse volatilizes the matrix/analyte mixture, and ionizes the matrix molecules, which then transfer one or more charges to the analyte molecules while preventing them from overexcitation and destruction. Analyte ions produced are then introduced to the mass spectrometer. This ionization method tends to produce ions possessing only a single or double charge, though higher charge states may occasionally be observed for larger molecules. Figure 1.1 (bottom) shows a simplified MALDI schematic.

Time-of-Flight Mass Analysis

Time-of-flight (TOF) mass analysis is well suited to mass spectrometric studies of large biomolecules, and the wider allowable mass range and high resolution of these instruments are of great benefit in analysis situations where detail and mass-accuracy are as important as speed or throughput. The first step in TOF mass analysis is the acceleration of the analyte ions. This is accomplished by introducing the ions into a region containing a precisely controlled electric field established using grid electrodes. This electric field accelerates the ions, imparting the exact same amount of kinetic energy to ions carrying a similar charge. As kinetic energy is a function of

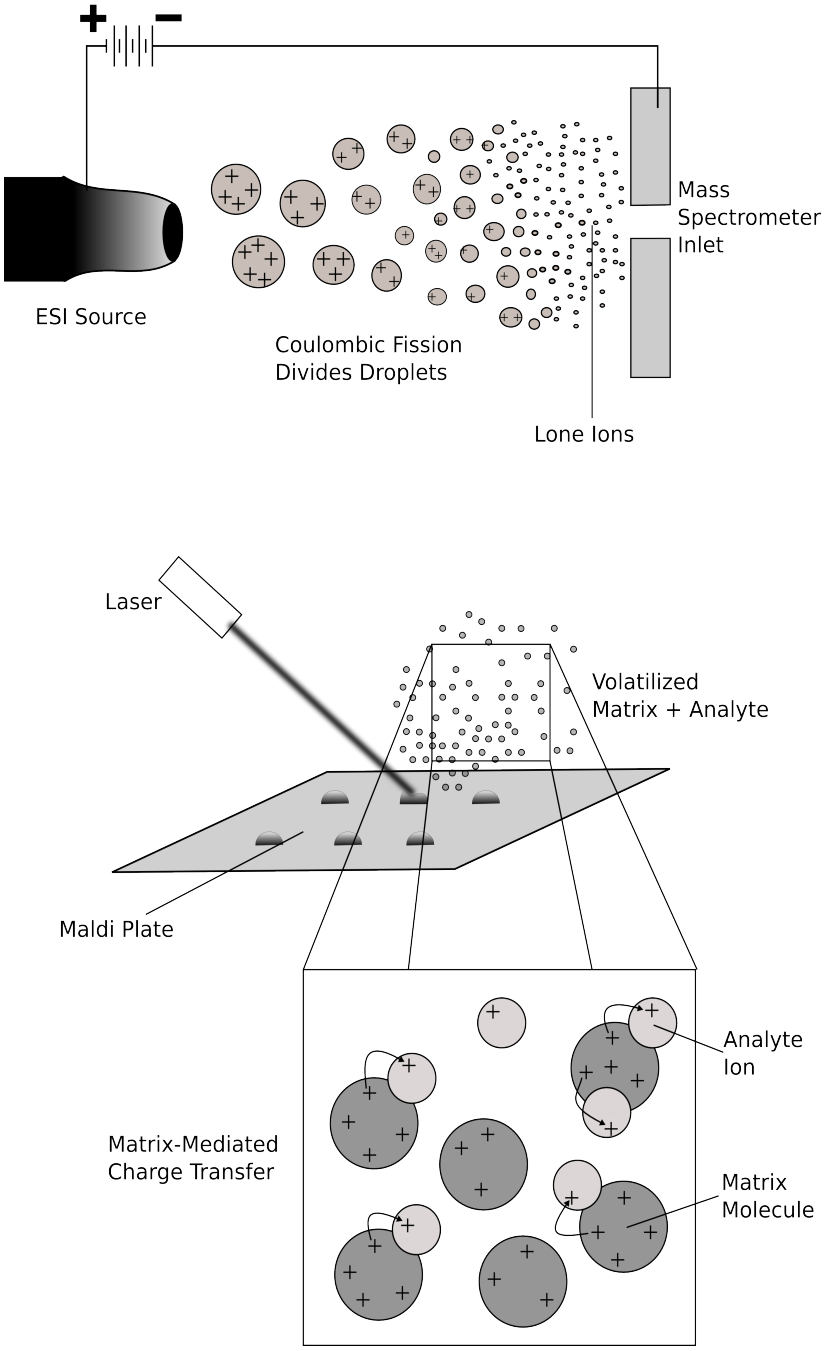


Figure 1.1: Two common ion sources used in the mass spectrometric analysis of peptides. Electrospray ionization (ESI) (top), and Matrix-assisted laser desorption ionization (MALDI) (bottom).

mass and velocity ($E_k = \frac{1}{2}mv^2$), ions of different masses will be accelerated to different velocities. After this acceleration, the ions enter an evacuated, field-free drift tube, where their velocity differences cause them to become spatially separated, and to impinge upon the terminal detector at different times. Ions with identical mass-to-charge ratios will ideally arrive at the detector as a tightly clustered group. The high-precision timing circuitry of the instrument may then be used to determine the mass-to-charge ratios of these ion clusters, based on the time required for each to drift to the detector. A simplified schematic of a TOF-based instrument in shown in figure 1.2.

Fourier Transform Ion Cyclotron Resonance (FT-ICR) Mass Analysis

Pioneered at UBC by Drs. Marshall and Comisarow, Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry is one of the latest hardware innovations to become widely popular in the field, three decades after its initial discovery. This method of mass analysis provides very high resolution and mass-accuracy, and operates in a fundamentally different way compared to other mass analysis methods. Whether scanned from a trap, filtered through a quadrupole, set drifting down a field-free tube, or flung out of a magnetic sector, all other mass analyzers rely on the destructive tallying of ions striking a detector. In contrast, FT-ICR detects ions by measuring their specific cyclotron frequency. To accomplish this, ions are introduced into a Penning trap with a fixed magnetic field, where they naturally orbit at a frequency proportional to their mass-to-charge ratios. Once inside this trap, rapidly alternating radio frequency "chirp" signals excite the ions, pushing them to orbit at larger radii, and causing ions of similar mass-to-charge ratios to coalesce into distinct packets. The frequency of orbit of these packets (the cyclotron frequency) is measured by sensitive detection plates that non-destructively sense proximate ion packet transit via an induced electric current. Thus, the output of the instrument consists of a superposition of sinusoids, one corresponding to each distinct cyclotron frequency – and therefore each distinct mass-to-change ratio – present in the analyte. A standard mass spectrum (intensity vs. mass-to-charge) may then be constructed from this mixture of sine wave functions using Fourier transform methods. A simplified schematic of an FT-ICR instrument is shown in figure 1.2.

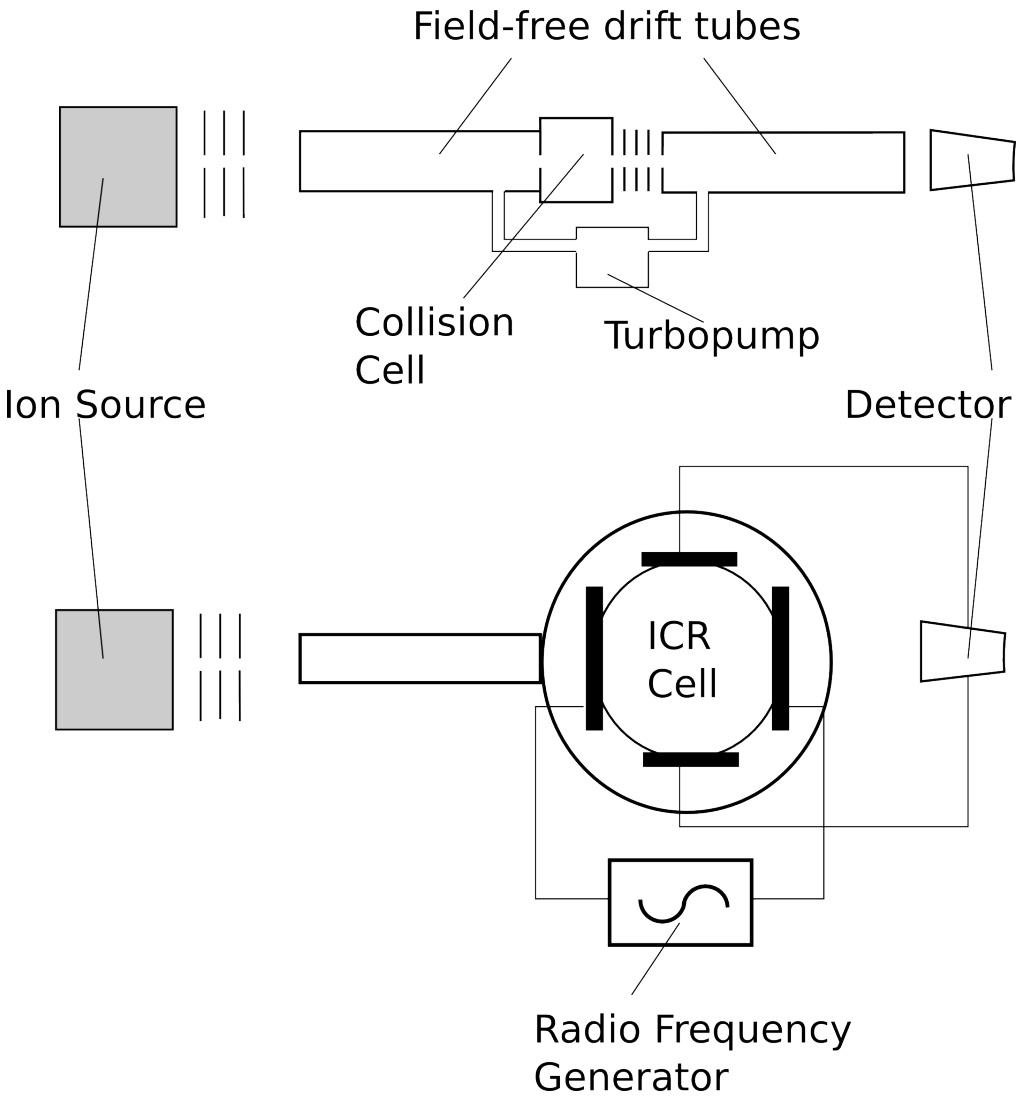


Figure 1.2: Two common mass analysis configurations used for the mass spectrometric study of peptides. A time-of-flight (TOF) mass analyzer (top), and a Fourier transform ion cyclotron resonance (FT-ICR) mass analyzer (bottom).

1.2.2 Separation Technologies

Though the resolution of modern instruments is very high, enabling successful differentiation of the components of complex mixtures, additional separation steps are almost always used. They are particularly necessary when analyzing proteins, or protein fragments, during a proteomics experiment, as the complexity of a typical proteomic analyte, such as a whole cell lysate, can be tremendously high. In addition, differences in protein abundance can span many orders of magnitude. This complexity can overwhelm even the highest resolution instruments currently available, and the differences in abundance can cause the signals coming from high abundance proteins to completely suppress those coming from low abundance proteins. To facilitate the study of these complex samples featuring wide abundance ranges, electrophoresis and chromatographic separation strategies are usually employed to simplify mass spectrometric analysis [13, 14]. These separation strategies generally involve the offline separation of proteins using gel electrophoresis (by size, in one dimension, or by size and isoelectric point, in two dimensions), or the online separation of proteins using high pressure liquid chromatography (HPLC). Often these approaches are combined to yield multidimensional separations, combining gel techniques with online HPLC separations, or stacking several consecutive HPLC separations [15]. Multidimensional separations are slower, but may significantly decrease the complexity of the resulting mass spectra, and allow for specific targeting of particular proteins. A workflow diagram showing a common combined separation strategy is shown in figure 1.3, wherein a one dimensional gel separation allows for the selection of a particular protein by molecular weight, which is then excised and digested with trypsin. An HPLC system then separates the resulting peptides by relative hydrophobicity, so that they may be analyzed one-by-one using mass spectrometry. The HPLC column eluent is often coupled directly to an electrospray ionization source for online analysis, but multiple discrete fractions of eluent may be collected and saved for use in offline analyses, in those using a MALDI ion source, for example.

1.2.3 MS Acquisition

Mass spectrometric (MS) analysis measures the mass-to-charge ratio of ions produced from the analyte introduced into the instrument. As the charge state of a particular ion may be assumed from the nature of the experiment and instrument, or inferred from the mass spectrometric signal (by examining isotope peaks), the mass of each distinct molecular species in the

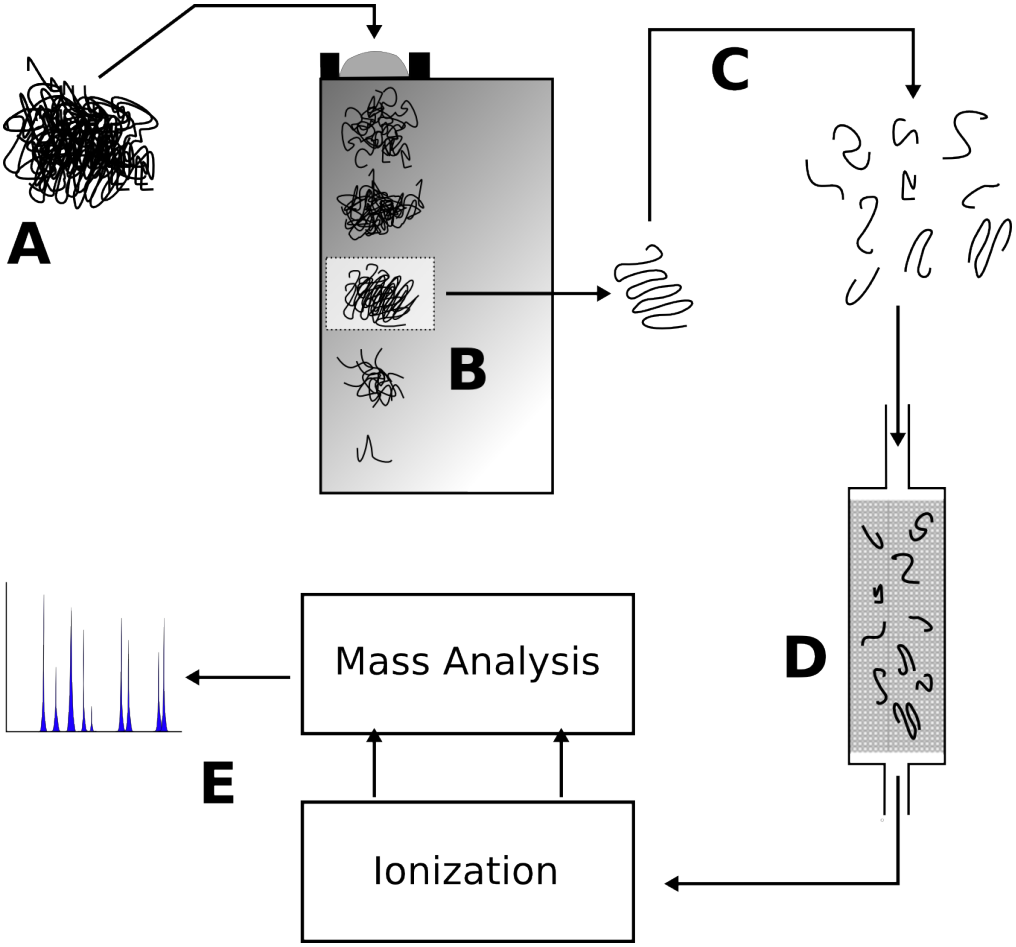


Figure 1.3: A typical two-stage separation strategy used in the characterization of protein mixtures. A mixture of proteins (a) is loaded onto an acrylamide gel, which separates the proteins by molecular weight (b). The mass region of interest is excised from the gel, and the proteins (in this case protein, for representational simplicity) contained therein is digested into peptides by a proteolytic enzyme such as trypsin (c). These peptides are then forced at high pressure through a column packed with hydrophobic material (d), which separates the peptides by their relative hydrophobicities. The output of the column is then analyzed by mass spectrometry (e).

analyte may be determined. As proteins can be extremely large and unwieldy molecules, peptides derived from enzymatic digests of these proteins are usually examined. If a sample is assumed to be free of contaminants, its corresponding mass spectrum will be composed of peaks that correspond to the mass of each peptide present.

The mass spectrum of a raw biological sample may look extremely complex, or uninterpretable, regardless of the resolution of the instrument. For this reason, the separation strategies discussed above are typically employed, and multiple, separate MS analyses are performed on each distinct band excised from a gel, or at many discrete timepoints throughout a chromatographic elution. This has a cost in terms of instrument and analysis time, but reduces the complexity of the mixture at each MS acquisition to a more reasonable level.

1.2.4 MS/MS Acquisition

In the analysis of complex biological molecules such as peptides, the mass of the molecule is often insufficient for its complete characterization. Many of these complex molecules have masses that are very similar to each other, or feature isobaric sequence substitutions. Additionally, chemical modifications can skew the MS mass of a peptide considerably, making a single mass-based identification impossible. For this reason, it is beneficial to acquire additional mass information during mass spectrometric analysis – this may be provided by higher order mass spectrometry, such as tandem mass spectrometry (MS/MS). In tandem MS, molecules of a particular mass are selected, and are fragmented in a specialized collision cell. Several peptide fragmentation strategies are commonly used, including collision induced dissociation (CID), electron capture dissociation (ECD), and electron transfer dissociation (ETD). A second stage of mass spectrometric analysis is performed on the ion fragments produced using these methods, yielding additional mass information characteristic of the specific precursor, as different ions having similar MS masses will usually feature different fragmentation patterns. In peptides, fragmentation occurs primarily between amino acid residues on the peptide backbone, and the collision energy can be adjusted such that fragmentation occurs once per molecule. This, averaged over many molecules, gives detailed peptide sequence information, and is the gold standard for peptide identification in proteomics experiments, using database searching or de novo sequencing methods that will be discussed in later sections. Figure 1.4 shows the selection of a precursor MS peak, reveals the location of putative peptide backbone fragmentations in the precursor, and gives an

example of a possible fragmentation spectrum after MS/MS analysis. In addition to acquiring additional mass information for peptide identification, many peptide modifications are readily dissociated in the collision cell. Monitoring the MS/MS spectrum for the presence of indicative marker ions, or for indications of a neutral loss, are useful techniques in the identification and localization of certain classes of protein post-translational modifications. Studies using MS/MS/MS (MS^3) and higher orders are feasible approaches for obtaining yet more information about the composition and structure of the analyte, and have been successfully applied to proteomics research [16].

1.3 Data Analysis in Proteomics

The acquisition of good quality MS and MS/MS data is the most critical step in a proteomics experiment, as sample amounts are limited, and hardware and analysis time are expensive. The interpretation of the data produced, however, is a second critical step. Though relatively less time-consuming and costly than data acquisition, the objectives are less defined, and harder to enumerate. Specific goals for data acquisition are logical and concrete: high signal-to-noise ratios, lack of contaminant signal, and tight chromatographic elution peaks are examples of these quantifiable objectives. The goal of data analysis, however, is more ephemeral, being simply to elicit the most "meaning" possible from the available data. This builds upon the objectives of data acquisition, but with less specific direction. The objective may be to identify as many proteins as possible in a sample, or only identify a few proteins with very high confidence. Other data analysis objectives include the identification and mapping of post-translational modifications, the identification and characterization of biomarkers, or the relative or absolute quantitation of specifically targeted proteins using spiked-in standards, or stable isotope labeling strategies such as SILAC or ITRAQ. Data analysis in mass spectrometric studies is too often accorded little importance, and does not factor adequately into the overall experimental design. As long as the protein list is sufficiently long, or the expectation values of the desired targets are sufficiently low, the inner workings of the data analysis applications, and their particular strengths and weaknesses, are not considered. This can lead to the generation of erroneous or misleading results [17].

1.3.1 Computational Approaches

Even at very high resolutions, with excellent chromatographic separation, mass spectra can be incredibly complex and difficult to interpret manually.

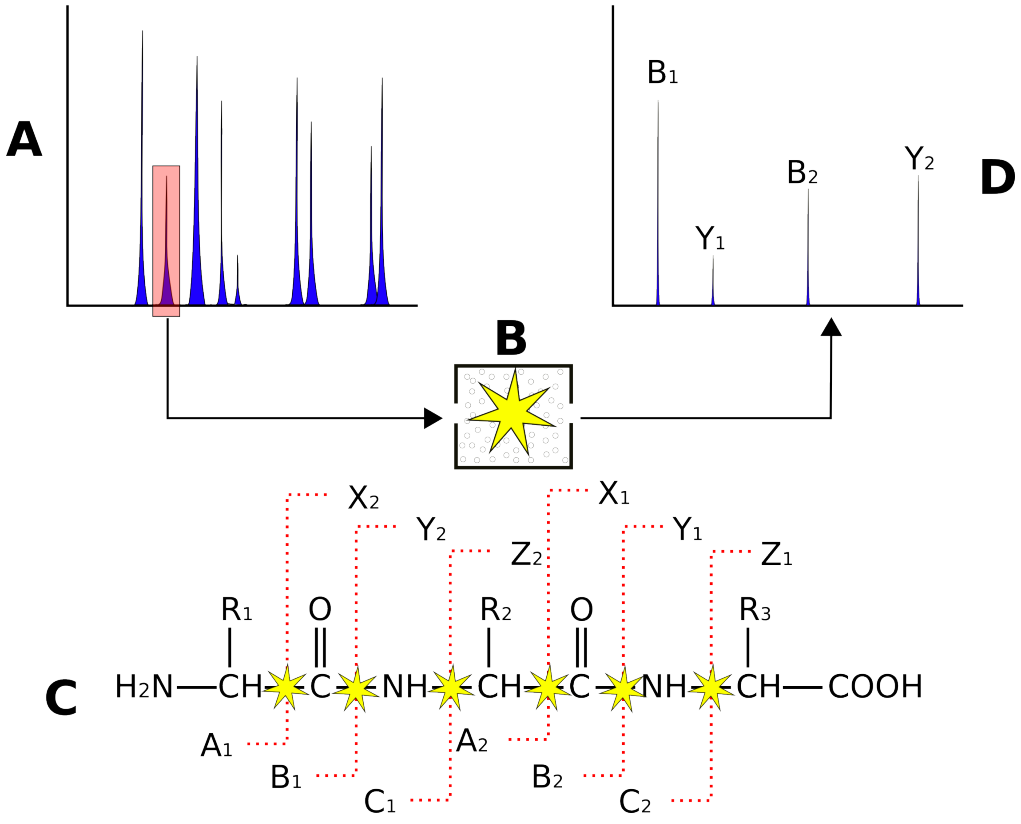


Figure 1.4: In MS/MS studies a precursor peak is selected for further study from an MS scan (a). It is subjected to fragmentation in the collision cell (b), where the collision energy level is calibrated such that the peptide backbone breaks in one location per molecule, creating complementary fragment ions that contain the N- and C-terminus respectively. Possible bond breakage locations are shown, along with the resulting fragment ion types (c). The charged fragment ions produced are then subjected to further mass analysis, producing the MS/MS spectrum (d) – for simplicity only b- and y-type ions derived from the fragmented example peptide shown in (c) are shown in the spectrum (d).

Software has been used in conjugation with analytical mass spectrometry for several decades – not necessarily due to the complexity of the spectra involved, but due to the tremendous size of the reference libraries that related masses (from the mass-to-charge ratios) with chemical compositions and structures, depending on the ion sources used. However, the manual interpretation of data from even a relatively simple proteomics experiment would be impossible. The complexity of individual MS and MS/MS spectra from complex peptide mixtures is one issue, but the sheer number of such spectra produced is another, potentially more serious issue. A typical analysis run using a standard LC-MS/MS analysis protocol can generate several thousand MS spectra acquired at discrete timepoints during the elution, and many MS/MS spectra generated from specific peaks of interest identified in these initial MS analyses. Therefore the use of analysis software to interpret mass spectrometry data is a standard step in proteomics, with manual validation occurring only in specific, select circumstances, such as confirming de novo sequencing results, or determining charge states if isotopic envelopes are poorly formed. Usually these measures are only needed if the data is of sub-optimal quality.

1.3.2 Computer Applications in Proteomics

The field of proteomics is a hybrid field, and although its aims are broad and considerable scope exists for exploratory work, the majority of the analytical tools used for the analysis of mass spectrometry data are motivated by individual problems. Problems targeted by well-known software solutions include peptide sequence determination, protein identification, the identification and localization of protein post-translational modifications, and the automated classification of mass spectra – for quality determination, for instance. Each of these is targeted by multiple high-quality applications that approach the problem in different ways.

Peptide Sequencing/Protein Identification using Databases

Computational approaches for the determination of peptide sequence (and the subsequent identification of the parent proteins) are the most widely used computational applications in contemporary proteomics research, and of these the three most widely used are Sequest [18], Mascot [19], and X!Tandem [20]. Though all three attempt to match experimental spectra with candidate spectra generated *in silico* from stored database sequences, they do so in markedly different ways. Sequest uses a measure of simi-

larity based on spectral correlation, where the experimental and database-candidate spectra are overlaid, and the calculated correlation value is used to assess the similarity. Mascot and X!Tandem both use specifically constructed probabilistic models to compare the experimental spectrum with the *in silico* candidate, and calculate the probability that any particular sequence/spectrum match could have occurred simply by chance. Many other identification approaches using database search are used, including Phenyx [21], OMSSA [22], and the Protein Prospector suite [23] – each one approaches the problem in a slightly different way, using different probabilistic model formulations, scoring functions, and matching criteria. A basic workflow demonstrating two divergent strategies for peptide identification is shown in figure 1.5. Both approaches use the mass of the precursor peak to narrow the range of the database that needs to be searched, but differing strategies are used for the searching process.

De Novo Peptide Sequencing.

The de novo sequencing of peptides is another domain which applies particularly sophisticated computational approaches to the peptide identification problem. Techniques in this domain attempt to extract peptide sequences from the tandem mass spectra alone, without probabilistic comparisons to spectra stored in databases, or constructed *in silico* from stored database sequences. Instead of a direct comparison, these techniques often enumerate long lists of potential sequences constructed from possible combinations of MS/MS peaks, and attempt to find an "optimal" sequence, based on some quantifiable definition of optimality. As a result, the computational techniques that feature prominently in this domain tend to be general strategies for search optimization, and are less specific to MS/MS data. They include mathematical graph construction and traversal, dynamic programming, and some compact probabilistic sequence models, such as hidden Markov models. Well known de novo sequencing applications include Peaks [24], Lutefisk [25], and pepNovo [26].

Modification Discovery

The computational techniques used for finding post-translational protein modifications (PTMs) are mostly very similar to the database matching techniques used for peptide sequencing and protein identification. They usually employ a database filtering approach to produce a smaller list of possible candidate matches, often filtering based on criteria such as de novo

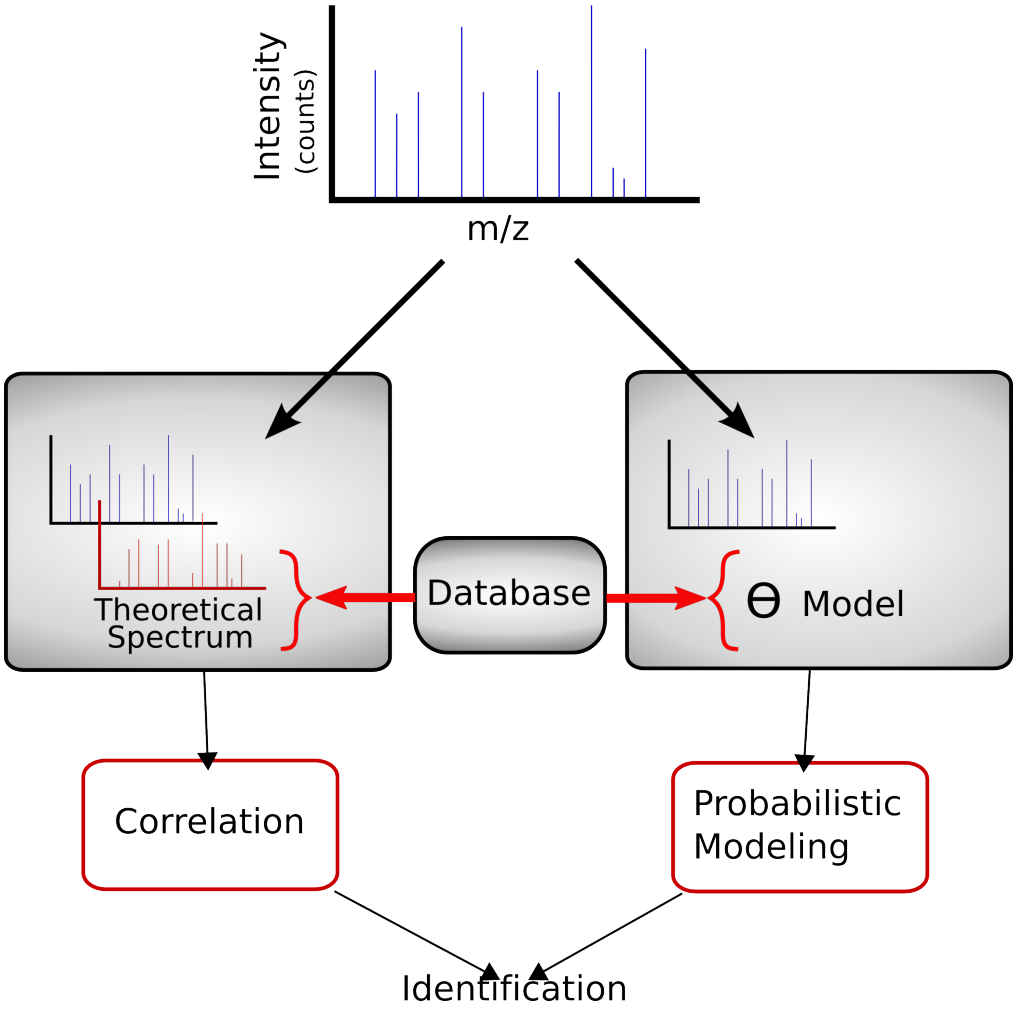


Figure 1.5: A typical workflow for the identification of a peptide using a database. The left branch uses the database sequences to generate theoretical spectra that are compared directly to experimental spectra using correlation (e.g. Sequest). The right branch uses the database sequences and a parametric model to associate experimental spectra with the sequences of the peptides that likely generated them (e.g. Mascot)

sequence tags, or chromatographic retention times. These candidate sequences are then augmented with modification mass information, and a more stringent database matching step then identifies, from this augmented pool, which modification (or set of modifications) a peptide is carrying. Augmented database search strategies include well known and widely utilized applications such as ProSight [27], the VEMS tool [28], and InsPecT [29]. Several other common PTM-finding applications extend de novo sequencing to include modification-containing amino-acid residue masses – as de novo sequencing is already a complicated search problem, this extension makes the problem all the more difficult. The techniques employed by these de-novo approaches – the best known being the openSea alignment algorithm [30], and SPIDER [31] – use partial de-novo sequences, and error-tolerant alignment to effectively generate and sort the list of candidate sequences.

For the applications mentioned above to work, the list of potential modification candidates must be explicitly specified – a smaller set of possible modifications will yield a faster search. A new class of algorithms provides an alternative to this situation. Tools in this domain take a more exploratory approach, and make use of different general purpose data examination tools, such as full spectrum alignment, to search for indications of modification events in the data. Approaches using this methodology include Nuno Bandeira's spectral networks [32], and a blind search identification strategy created by Tsur *et al* [33]. Though they do also make the common assumption that the modification is substoichiometric – that the modified and unmodified forms exist in the sample simultaneously – these approaches allow searching for unspecified modifications, and can scale well to situations involving large numbers of modifications. Figure 1.6 demonstrates the use of spectral alignment to discover unknown modifications in a peptide. If the peptide was unmodified, the line joining the peaks from the two aligned spectra would form a straight diagonal. Aligning a peptide with its modified version produces discontinuities in this line that are indicative of the masses of the potential modifications.

Spectral Classification and Machine Learning

A relatively recent computational approach, swiftly gaining in popularity and useful in many proteomic studies, makes use of statistical machine learning classifiers to group spectra according to characteristics, or to classify or sort spectra on the basis of quality. A set of possible features serving to potentially distinguish spectra from each other is specified, and a training algorithm attempts to learn a model that can discriminate between the

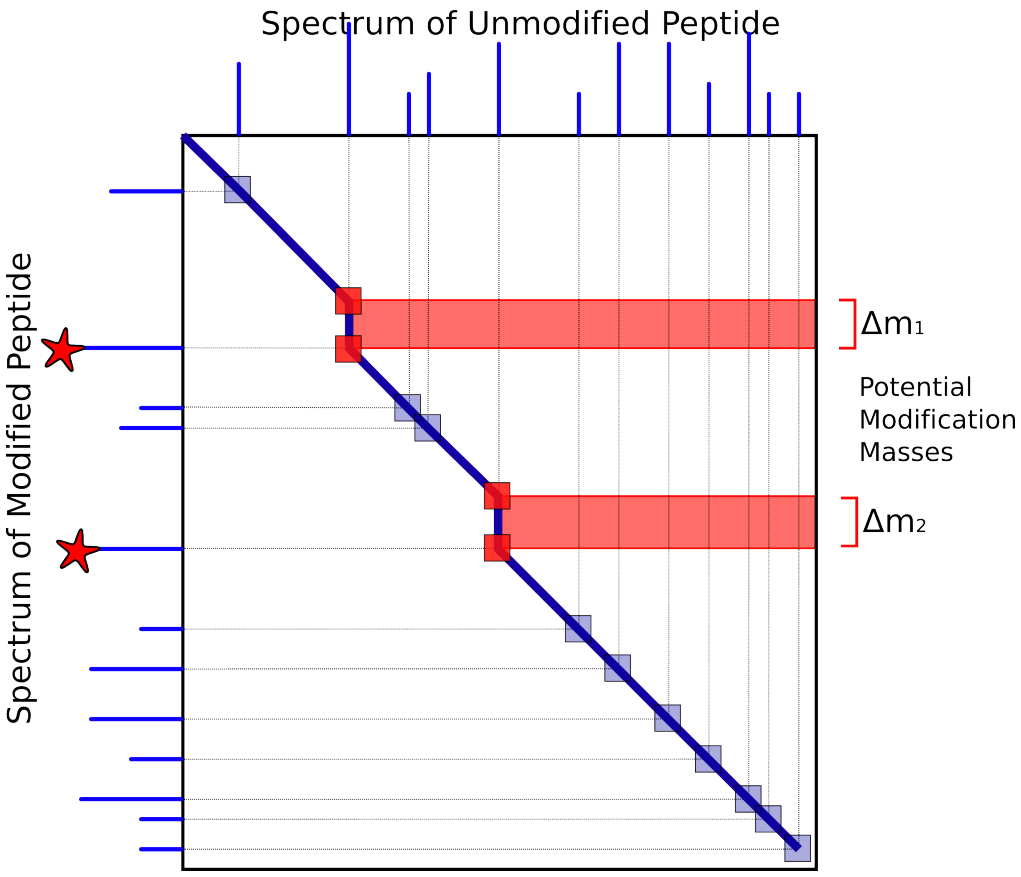


Figure 1.6: An unmodified MS/MS spectrum (top) is aligned with its modified version (side). Discontinuities in the straight diagonal joining peak pairs represent the masses of the modifications. As long as peptide modifications are sub-stoichiometric, this approach can find all modifications present, without advance specification of which to look for.

classes – such models are typically discriminant analysis approaches [34], support vector machines [35], or neural networks [36]. Support vector machines have shown particular promise, as they may easily take advantage of kernel techniques to expand the dimension of the feature space, allowing them to use linear classification techniques on nonlinear data. Figure 1.7 shows an example of this approach – the red and blue points on the plane cannot be separated by a straight line (left), but using a nonlinear kernel to expand the feature space to three dimensions allows the points to be separated by a plane (right). The application of such powerful classifiers to proteomics data enables analytical approaches, such as the development of clinical diagnostic classifiers (known as "clinical decision support systems"), that seem to embody the ultimate diagnostic goals of systems biology. These approaches aim to accurately classify (and diagnose) cancer patients using mass spectra of blood plasma [37], and have shown promising results, though work in this area is at a very early stage. Classification methods can also be used to quickly sort through spectra and determine which are likely to yield poor identifications – these can then be filtered before further analysis takes place, increasing throughput, and reducing the likelihood of false positive identifications [38].

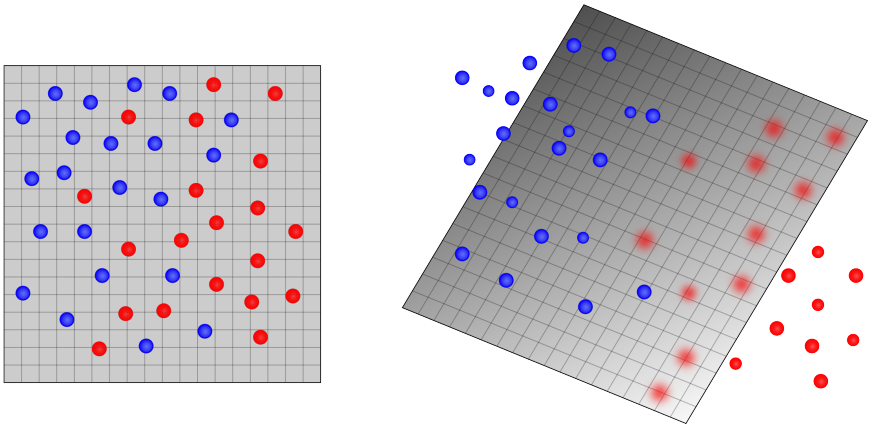


Figure 1.7: Using nonlinear kernels with SVM classification enables inseparable classes (left) to be linearly separated in a higher dimensional space (right).

1.4 Thesis Theme and Hypotheses

From the above discussion, it is obvious that computational techniques play a fundamental role in proteomics research. This role will only grow larger as typical experiments grow ever more complex, and the field of proteomics as a whole approaches the goals of mapping complex cell signaling networks, and of reproducibly identifying the unique signatures of diseases, drugs, genetics, and environmental effects in the human proteome. Though certain of these computational techniques are very well established in proteomics research, and have been the subject of intensive research for many years, it must be noted that the integration of the computer sciences with mass spectrometry, for more than the sake of convenience, is a relatively recent phenomenon. The computational tools that are popular are highly specialized applications oriented towards very specific analytical goals.

Though not currently as widespread, many well established, general computational analysis techniques may be applied to mass spectrometric data, not to answer specific questions, but to re-frame the data, and discover new ways in which it may be exploited. Certain of these general techniques are beginning to find uses in the specific approaches discussed above, such as iterative database filtering, spectral alignment, correlation, and model fitting via machine learning. Even well established data analysis methods are beginng to press these general analyses into service – for example, machine learning systems have recently been applied to the output of the sequest algorithm [39] to yield more predictive power in identifications.

In examining the ensemble of computational approaches available for use on mass spectrometry data, exploratory techniques from different, general exploratory analysis families appear repeatedly, often as preprocessing steps, conjoined to more specific analyses. These include spectral filtering, data transformation, statistical modeling, and techniques recruited from the quickly-evolving fields of statistical machine learning and artificial intelligence. When applied, these discovery-based computational approaches emphasize generality and simplicity in the output of their analyses, which allows for relatively uncomplicated interpretation of patterns and consistencies that emerge in analyzed data, and provides a solid foundation on which to construct useful tools that exploit the information discovered. Advances in machine learning methods are particularly relevant to this area – they are well suited to the extraction of important signal from confounding noise, and provide a graceful mathematical approach to the complexity encountered in mass spectra, without resorting to ad hoc modeling assumptions.

1.4.1 Thesis Statement and Aims

The work presented here aims to establish the validity of discovery-oriented computational work in mass spectrometry and proteomics research, and to show that it may complement more traditional problem-based computational solutions, and have additional utility in the visualization, interpretation, and modeling of contemporary proteomics datasets that are both very large, and very complex. The analysis and modeling tools described will demonstrate how the construction and effective use of general exploratory data analysis methods may lead to the development of simple, yet highly capable tools.

1.4.2 Outline of Approach

This work discusses three separate, general classes of exploratory computational analysis of varying complexity – data transformation, data matrix decomposition, and sequence-based statistical machine learning. For each class, the utility of a specific exploratory method in mass spectrometry data analysis will discussed, and a novel tool will be developed to fully exploit the newly discovered information.

Data Transformation

Data transform methods are commonly used in the analysis of mass spectrometry data. For instance, the fast Fourier transform is used in FT-ICR MS to produce a mass spectrum from a complicated superposition of sine waves generated by the instrument's detectors. However, transforming data from the normal mass spectrum domain (m/z) to other domains can be useful as well. For example, wavelet transforms have proved themselves useful in several areas, including noise filtration [40], peak detection [41], and as a step in complex Bayesian model fitting [42]. The study of data transforms in this study will focus on the correlation transform, and the applications of cross-correlation, autocorrelation, and signal convolution methods to the interpretation of mass spectra. Though correlation is utilized in several areas of proteomics research, including the Sequest algorithm, a novel application of spectrum autocorrelation will be implemented to transform mass spectra into the Δm/z (mass shift) domain. This transformation displays the mass differences between spectrum peaks, rather than the peaks themselves, enabling simpler visual identification of mass-shift signals. To exploit this information, a tool will be developed based on this data transform that will

use the Δm/z information extracted to identify individual MS spectra that contain modified peptides in full scale proteomics experiments.

Matrix Decomposition/Data Compression

Matrix decomposition methods are fundamental techniques originating in linear algebra that are often used in classification and data compression. The application demonstrated here will make use of the singular value decomposition (SVD), which is a generalization of the common eigen-decomposition to non-square matrices. Similar to eigenvectors, the SVD decomposition isolates "singular vectors" that represent the orthogonal, linear combinations of input m/z values (in the case of mass spectrometry data) that, taken together, explain the greatest amount of quantifiable difference between spectra. This technique has been used in the classification of very complex mass spectra obtained from different biological samples of interest [43], and in the elucidation of potential proteomic biomarkers [44]. In this study, the singular value decomposition will be applied to the analysis of peptide fragmentation in the pursuit of two specific aims: the visualization of fragmentation pattern changes due to sequence and peptide-specific effects, and the creation of a feature extraction tool, capable of isolating specific MS/MS features useful for creating a general machine learning model of peptide fragmentation in tandem mass spectrometry experiments.

Statistical Machine Learning

Ever-more powerful computers are ushering in an exciting era of computer science research, particularly in the field of statistical machine learning. Techniques from this field have shown great promise in real-world situations, as they are able to automatically learn complex models from experience, without relying on simplifying model assumptions. These models perform accurate prediction in complex scenarios such as weather pattern prediction, stock market analysis, and image recognition, segmentation, and classification. Machine learning models have been applied to the modeling of mass spectrometry data for some time. Well known examples include classifiers, like the support vector machines mentioned above, decision tree classifiers [45], and generative sequence models such as the hidden Markov model [46, 47]. Though these sequence models have been successfully deployed, the complexity of modeling large amounts of detailed mass spectrometry data is a great challenge. To remain computationally tractable, these models must rely on critical assumptions of probabilistic independence, and

are therefore incapable of the detailed data representation needed for certain useful classes of modeling. The final aim of this study is to assess the performance of a new type of probabilistic sequence model in the interpretation of mass spectrometry data. This model is known as a conditional random field, and was originally developed to model written language. It is a conditionally trained, discriminative model framework, and is thus free from the independence assumptions that cripple traditional, generative sequence models. Although it learns somewhat slower than a hidden Markov model on an equivalent amount of training data, this type of model is capable of representing a greater proportion of the nuance and intricacy present in a mass spectrum, without requiring any debilitating compromises for efficiency. The suitability of this sequence model for use in modeling peptide fragmentation will be assessed, and potential applications in exploratory data analysis and probabilistic sequence/spectrum scoring will be explored.

Bibliography

[1] Schrattenholz, A. and Kic, V. S. *Curr Med Chem* **15**(15), 1520–8 (2008).

[2] Burbaum, J. and Tobal, G. M. *Current Opinion in Chemical Biology* **6**(4), 427–433 (2002).

[3] Weston, A. and Hood, L. *Journal of Proteome Research* **3**(2), 179–196 (2004).

[4] Breitbart, R. E., Andreadis, A., and Nadal-Ginard, B. *Annual Review of Biochemistry* **56**(1), 467–495 (1987).

[5] Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. *Mol. Cell. Biol.* **19**(3), 1720–1730 (1999).

[6] Anderson, L. and Seilhamer, J. *Electrophoresis* **18**(3-4), 533–7 (1997).

[7] Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. *Genome Biology* **4**(9), 117 (2003).

[8] Pevalova, M., Filipcik, P., Novak, M., Avila, J., and Iqbal, K. *Bratisl Lek Listy* **107**(9-10), 346–53 (2006).

[9] Dias, W. B. and Hart, G. W. *Mol Biosyst* **3**(11), 766–72 (2007).

[10] Takahashi, K., Uchida, C., Shin, R.-W., Shimazaki, K., and Uchida, T. *Cell Mol Life Sci* **65**(3), 359–75 (2008).

[11] Uhlen, M. and Ponten, F. *Mol Cell Proteomics* **4**(4), 384–393 April (2005).

[12] Kopf, E. and Zharhary, D. *Int J Biochem Cell Biol* **39**(7-8), 1305–1317 (2007).

[13] Liu, H., Lin, D., and Yates, J. R. *Biotechniques* **32**(4), 898, 900, 902 passim (2002).

[14] Fournier, M. L., Gilmore, J. M., Martin-Brown, S. A., and Washburn, M. P. *Chem Rev* **107**(8), 3654–86 (2007).

[15] Issaq, H. J. *Electrophoresis* **22**(17), 3629–38 (2001).

[16] Olsen, J. V. and Mann, M. *Proceedings of the National Academy of Sciences* **101**(37), 13417–13422 (2004).

[17] Patterson, S. D. *Nat Biotechnol* **21**(3), 221–2 (2003).

[18] Eng, J., McCormack, A., and Yates, J. *Journal of The American Society for Mass Spectrometry* **5**(11), 976 (1994).

[19] Perkins, D., Pappin, D., Creasy, D., and Cottrell, J. *Electrophoresis* **20**(18), 3551 (1999).

[20] Craig, R. and Beavis, R. C. *Bioinformatics* **20**(9), 1466 (2004).

[21] Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. *Proteomics* **3**(8), 1454–1463 (2003).

[22] Geer, L., Markey, S., Kowalak, J., Wagner, L., Xu, M., Maynard, D., Yang, X., Shi, W., and Bryant, S. *Journal of Proteome Research* **3**(5), 958–964 (2004).

[23] Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., and Burlingame, A. L. *Mol Cell Proteomics* **4**(8), 1194–1204 (2005).

[24] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. *Rapid Communications in Mass Spectrometry* **17**(20), 2337–2342 (2003).

[25] Taylor, J. and Johnson, R. *Analytical Chemistry* **73**(11), 2594–2604 (2001).

[26] Frank, A. and Pevzner, P. *Analytical Chemistry* **77**(4), 964–973 (2005).

[27] LeDuc, R. D., Taylor, G. K., Kim, Y.-B., Januszyk, T. E., Bynum, L. H., Sola, J. V., Garavelli, J. S., and Kelleher, N. L. *Nucleic Acids Res* **32**(Web Server issue), W340–5 (2004).

[28] Matthiesen, R., Trelle, M. B., Hojrup, P., Bunkenborg, J., and Jensen, O. N. *J Proteome Res* **4**(6), 2338–47 (2005).

[29] Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. *Anal Chem* **77**(14), 4626–39 (2005).

[30] Searle, B. C., Dasari, S., Wilmarth, P. A., Turner, M., Reddy, A. P., David, L. L., and Nagalla, S. R. *J Proteome Res* **4**(2), 546–54 (2005).

[31] Han, Y., Ma, B., and Zhang, K. *Proc IEEE Comput Syst Bioinform Conf* **NIL**(NIL), 206–15 (2004).

[32] Bandeira, N. *Biotechniques* **42**(6), 687 (2007).

[33] Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE* , 157–166 Aug. (2005).

[34] Lilien, R. H., Farid, H., and Donald, B. R. *J Comput Biol* **10**(6), 925–46 (2003).

[35] Carvalho, P. C., da Gloria Costa Carvalho, M., Degrave, W., Lilla, S., Nucci, G. D., Fonseca, R., Spector, N., Musacchio, J., and Domont, G. B. *J Exp Ther Oncol* **6**(2), 137–45 (2007).

[36] Luk, J. M., Lam, B. Y., Lee, N. P. Y., Ho, D. W., Sham, P. C., Chen, L., Peng, J., Leng, X., Day, P. J., and Fan, S.-T. *Biochem Biophys Res Commun* **361**(1), 68–73 (2007).

[37] Shin, H. and Markey, M. K. *J. of Biomedical Informatics* **39**(2), 227–248 (2006).

[38] Salmi, J., Moulder, R., Filen, J.-J., Nevalainen, O. S., Nyman, T. A., Lahesmaa, R., and Aittokallio, T. *Bioinformatics* **22**(4), 400–406 (2006).

[39] Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. *Nat Meth* **4**(11), 923–925 (2007).

[40] Li, X., Li, J., and Yao, X. *Comput. Biol. Med.* **37**(4), 509–516 (2007).

[41] Du, P., Kibbe, W. A., and Lin, S. M. *Bioinformatics* **22**(17), 2059–65 (2006).

[42] Morris, J. S. and Carroll, R. J. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 179–199(21) (April 2006).

[43] Wagner, M. and Castner, D. *Langmuir* **17**(15), 4649–4660 (2001).

[44] Verhoeckx, K. C. M., Bijlsma, S., de Groene, E. M., Witkamp, R. F., van der Greef, J., and Rodenburg, R. J. T. *Proteomics* **4**(4), 1014–28 (2004).

[45] Vlahou, A., Schorge, J., Gregory, B., and Coleman, R. *Journal of Biomedicine & Biotechnology* **2003**(5), 308–314 (2003).

[46] Wu, X., Tseng, C.-W., and Edwards, N. *J Comput Biol* **14**(8), 1025–43 (2007).

[47] Mirabeau, O., Perlas, E., Severini, C., Audero, E., Gascuel, O., Possenti, R., Birney, E., Rosenthal, N., and Gross, C. *Genome Res* **17**(3), 320–7 (2007).

Chapter 2

Correlation and Convolution Analysis of Peptide Mass Spectra[1]

2.1 Introduction

The nascent field of high-throughput proteomics, though still in a rapid, exploratory stage of growth, has established a few guiding truths that will continue to shape its future development. One of these truths is that high-throughput proteomic analysis produces an enormous amount of data [1, 2]. In particular, the mass spectrometric approaches commonly used to analyze the proteome generate vast numbers of spectra, both normal (MS) and tandem mass spectra (MS/MS); each of these must be processed after acquisition in order to be useful. Technologies such as two-dimensional nanoflow liquid chromatography tandem mass spectrometry [3, 4] and matrix-assisted laser desorption/ionization tandem time-of-flight (MALDI-TOF/TOF) [5] mass spectrometry are increasing the utility of the large quantity of data produced from biological samples – making it more selective and sensitive and increasing its overall quality. These efforts to boost data-generating capability and data quality will prove essential to the development of the often envisioned definitive applications of proteomics, such as fast-turnaround clinical diagnosis and prognosis tools. At present, however, the development of data analysis techniques capable of dealing effectively with such massive quantities of data is lagging behind the hardware and process advances enabling this new data generation capacity. New data analysis techniques, and new ways of thinking about data analysis problems, are needed to transform the sheer abundance of data from a hindrance to a benefit. As it is, a significant amount of data may be overlooked; it is easy to generate more data

[1] A version of this chapter has been published. Sniatynski, M.J., Rogalski, J.C., Hoffman, M.D., and Kast, J. (2006) Correlation and Convolution Analysis of Peptide Mass Spectra. Analytical Chemistry. 78(8), pp2600–2607.

than can be analyzed in depth.

With this generation/analysis asymmetry in mind, researchers have been increasingly focused on general analysis techniques, such as de novo sequencing [6], to extract as much information as possible from collected data. Most of these analyses are done using automated analysis packages such as MASCOT [7] or SEQUEST [8, 9, 10, 11]. These automated techniques use fast algorithms to compare the experimental data to a theoretically possible data set produced by transcribing a genome in silico, digesting the theoretical proteins, and predicting the fragmentation patterns of all possible peptides. These types of database-dependent analyses will find predictable proteins from sequenced genomes but may fail to provide more specific information on those proteins, such as their state of mutation, modification, or truncation. This is an area where additional analysis techniques may be able to supplement these commonly used analysis packages, particularly if they are able to provide information on many potential variants or modifications rapidly and simultaneously. Although there have been improvements in this area, as exemplified by the X!tandem project (www.thegpm.org) [12, 13], additional tools to augment these analyses with minimal researcher involvement are necessary, as such information can have enormous scientific and clinical importance.

Unfortunately, the average normal or tandem mass spectrum is very convoluted, and the evidence for a particular isoform, variant, modification, or sequence tag of interest may be completely obscured by other mass-domain signals. New analytical techniques are therefore required in order to extract and make use of this elusive information. The approach presented here involves the application of signal correlation, in which two linearized spectra are mathematically correlated to each other at a range of offset values. This produces a new, transformed view of the data, with the spectral offset along the x-axis and the correlation coefficient (r-value) on the y-axis. This transformation highlights mass shifts at which spectral signals overlap significantly. Such mass shift signals are important, useful information, as they are often evidence of modification-specific mass increases or losses. This approach is referred to as delay-series correlation, to distinguish it from other correlation-based analyses both in mass spectrometry and in other scientific disciplines.

Correlation-based methods for data analysis and comparison have been used for decades. Correlation analysis is used routinely in digital signal processing [14], machine learning [15], and many branches of analytical physics [16]. It is a robust and versatile analysis, functional for signals in any domain, being simply a linear measure of the spike amplitude rela-

tionship of two variables, reporting a value between 1 (perfect correlation – the signals are exactly the same) and -1 (perfect negative correlation – the signals are the exact inverse of each other). As a commutative operation, it gives a non-directional result, reporting information only on the relationship between two sets of signals. It has been noted previously that an obvious extension exists from the quantized signals in digital signal processing to the mass domain for the analysis of mass spectrometry data and for the identification of mass shift signals from spectra, corresponding to individual amino acid residues [17]. In this study, a potential use of this simple and comprehensive technique in the automated analysis of mass shift signals in mass spectrometry data is demonstrated, focusing on the detection and localization of protein/peptide modifications and the interpretation of isotopically labeled peptide spectra

2.2 Experimental Methods

MS and MS/MS experiments were performed on a quadrupole time- of-flight (Q-TOF) instrument (QStar XL, Applied Biosystems/ MDS Sciex), a matrix-assisted laser desorption/ionization tandem time-of-flight (MALDI-TOF/TOF) instrument (4700 Proteomics Analyzer, Applied Biosystems), or an ion trap Fourier transform ion cyclotron resonance (IT-1) instrument (LTQ-FT, Thermo Electron). The IT-1 was connected to a nano-HPLC system (1100 Series, Agilent) for reversed-phase separation using a gradient of 2-60% acetonitrile, in 0.5% acetic acid. Data were acquired on the IT-1 such that all MS spectra were acquired in the ICR cell, while all MS/MS spectra were acquired in the ion trap.

A peptide (PPGFSPFR) was modified with both the light and heavy versions of a charge directing N-terminal isotope-coded tag as performed previously [18], imparting a mass difference of 9.056 Th. A second peptide (HLLV<u>S</u>NVGGDGEEIER) with an O-linked N-acetylgalactosamine-modified serine (underlined) was synthesized using F-Moc chemistry and purified by HPLC. A third peptide (DRV<u>Y</u>IHPF) with a phosphotyrosine residue (underlined) was purchased (Sigma). These peptides were used at 1 pmol/L for nanospray experiments. Wild-type and ssaR-deficient Salmonella *typhimurium* (SL1344), auxotrophic for arginine biosynthesis, were grown in SILAC conditions, the ssaR line labeled with $_{13}C_6$- Arg, and the wild-type labeled with normal isotopic abundance amino acids. Both populations were induced to secrete, and their secretions were collected and analyzed as described [19, 20].

Spectra were exported out of the mass spectrometers proprietary data format into ASCII format. Due to the characteristics of the instruments that generate TOF and ICR data, bin size varies with location on the m/z axis. However, the correlation functions require all bins to be of identical width. Linearization of TOF and ICR bins was therefore performed prior to any correlation analysis. This was accomplished by constructing new linear bins via interpolation at a lossless resolution, chosen to be half of the smallest nonlinear bin size in the original data.

Equation 2.1 gives the spectral convolution (C_d), where X and Y are two input data series to be analyzed.

$$C_d = \sum_i (X_i - m_x)(Y_{i+d} - m_y) \tag{2.1}$$

X_i is the value of data series X in linear bin number i and Y_{i+d} is the value of data series Y in linear bin number $i + d$, where d refers to a specific offset expressed in bin number. The variables m_x and m_y refer to the average intensities of data series X and Y, respectively. A positive value of C_d indicates that the signals (in vectors X and Y) exhibit similar trends, and a negative value indicates that the signals have opposite trends. A value of zero indicates that there is no similarity between the signals. The convolution thus measures how related two signals are, but it is not normalized. However, the comparative magnitude of the value indicates the degree of similarity. Equation 2.2 gives the cross-correlation coefficient (r), which is essentially a normalized convolution (the convolution as given in eq 1 forms the numerator).

$$r_d = \frac{\sum_i (X_i - m_x)(Y_{i+d} - m_y)}{\sqrt{\sum_i (X_i - m_x)^2}\sqrt{\sum_i (Y_i - m_y)^2}} \tag{2.2}$$

The denominator serves to normalize the convolution and restricts it to a range from -1 to 1 inclusive. By inspection, it can be noted that the denominator yields the theoretical maximum for the convolution. Thus, if the signals are perfectly correlated, the convolution in the numerator will equal this value, and the r-value will be 1. Likewise, if the signals are the exact opposite of each other, the r-value obtained will be -1. It is important to note that the correlation coefficient is not defined if one of the data series is composed entirely of zeros, as this situation leads to a zero denominator.

Equation 2.3 is a simple modification of eq 2.2 and is used when comparing a signal to an offset version of itself. This process is known as autocorrelation.

$$r_d = \frac{\sum_i (X_i - m_x)(X_{i+d} - m_x)}{\sum_i (X_i - m_x)^2} \tag{2.3}$$

Delay series cross-correlations (between two spectra) were performed using eq 2.2. The value of d was varied from 0 to the maximum bin value present in either X or Y, producing a data series of r-value versus delay (d). Delay series autocorrelations were performed using eq 2.3 in a similar manner.

Convolution mapping was performed by calculating the value of the convolution at a particular offset (d) using eq 2.1 and then plotting the individual products instead of performing the summation. These products may be indexed by the value of either X_i or Y_{i+d} (from eq 2.1). All convolution maps performed herein report the heavier of the two.

2.3 Results and Discussion

The transformation of spectra using delay-series correlation is an example of an approach that allows the researcher to visualize data in a new way and, through this, gain additional insight into the experiment. Since mass spectrometric data are typically very complex, such techniques can prove invaluable during the initial stages of the analysis. Due to the inherent flexibility and simplicity of the correlation approach, there are multiple ways it can be applied to acquired mass spectra, each producing a different type of analysis, depending exactly on the need. Four principal correlation-based approaches are readily apparent (Figure 2.1). The first is a comparison of an acquired spectrum to a theoretical spectrum created from a database, as is routinely done by commonly used protein identification database searching tools, which produce theoretical MS/MS spectra in silico [8, 9, 10, 11]. Alternatively, it can be used to compare an acquired spectrum to a known spectrum from a library, as is done for other spectroscopic methods, such as IR and UV spectroscopy. It can also be used to compare the similarity of two different experimental samples or the same sample under different conditions [17]. Finally, it can be used to extract a regularly repeating signal from a single spectrum. The latter two applications will be the focus of the remainder of the discussion.

The example of a specific mass difference between peptides in two otherwise equivalent tandem MS spectra is examined as a preliminary demonstration of this technique. Such a situation could arise if any fixed-mass tag was introduced in one of the samples or if an isotope-coded tag (with multiple possible masses) was used to label peptides from different samples. As a concrete example, two equivalent tryptic peptides are examined, one set labeled with a heavy version of an isotope-coded tag and the other set

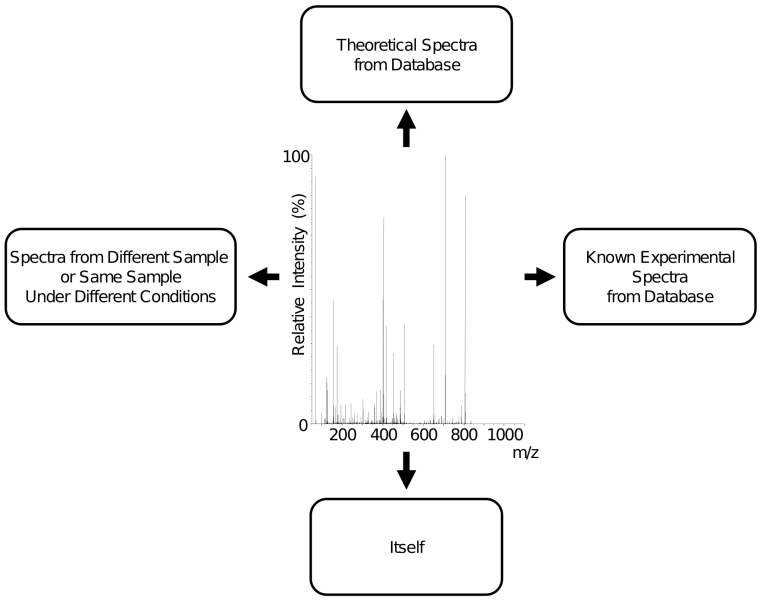


Figure 2.1: Illustration of four different sources of information to which experimentally derived spectra may be compared using correlation based analyses.

labeled with a lighter version of the tag [18]. Such a label introduces a fixed mass shift into the spectra – the position of the light and heavy peaks along the m/z axes of their respective spectra will be offset by the mass difference between the heavy and the light tag, or some multiple thereof, depending on the charge state of the molecule; in the current example, this difference is 9.06 Th for a singly charged peptide. However, the rest of the spectrum will be nearly identical, save for random fluctuations in noise and other small differences that arise due to run-to-run variability. MS/MS spectra of the light (Figure 2.2a) and heavy (Figure 2.2b) tagged versions of bradykinin fragment 2-9 (PPGFSPFR) were acquired, displaying these properties. Many of the peaks in the ion series appear with a 9.06 Th difference between the two spectra. Some fragment ions that are no longer linked to the isotope-coded tag appear identical in the two spectra, while the precursor $[M + 2H]^{2+}$ peaks appear at 531.76 Th for the light tagged version and 536.29 Th for the heavy version, showing a 4.53 Th difference due to its doubly charged state.

Subjecting each of these to delay-series autocorrelation (Figure 2.2c and 2.2d, respectively) shows that the mass shifts predominant in each of the spectra individually can be clearly visualized. Large correlations appear at 0 Th offset because the spectra are perfectly aligned, leading to a perfect correlation score (r-value). There are also significant positive correlations at 1.01 Th due to overlapping isotopes, 17.03 Th due to the overlap of fragments that have undergone a loss of NH_3 with those that have not, and 28.00 Th due to the overlap of a- and b-type ions. It is also interesting to note that the autocorrelations of the spectra shown in Figure 2.2a and 2.2b look similar, indicating that the differences in isotopic coding do not differentially affect fragmentation.

Subjecting the data in Figure 2.2a and 2.2b to a delay-series cross-correlation produces the delay-series plot in Figure 2.2e. A moderate r-value (0.54) at an offset of 0 Th arises due to the nonlabeled fragments both spectra have in common. Since the tag chosen in this example is affixed to the N-terminus, only N-terminal containing fragments have their masses shifted by the weight of the tag (about half the fragments). The rest of the fragments, whether they are C-terminal containing, internal fragmentation products (containing no terminus), or immonium ions, are located at the same m/z in both spectra, leading, as in the autocorrelation, to a significant r-value at an offset of 0 Th. The second largest r-value peak at an offset of 9.06 Th is caused by the overlap of the heavy-labeled and light-labeled fragments, 9.06 being the difference in mass between the heavy and light version of the isotopic tag used in this example. Peaks in the correlation

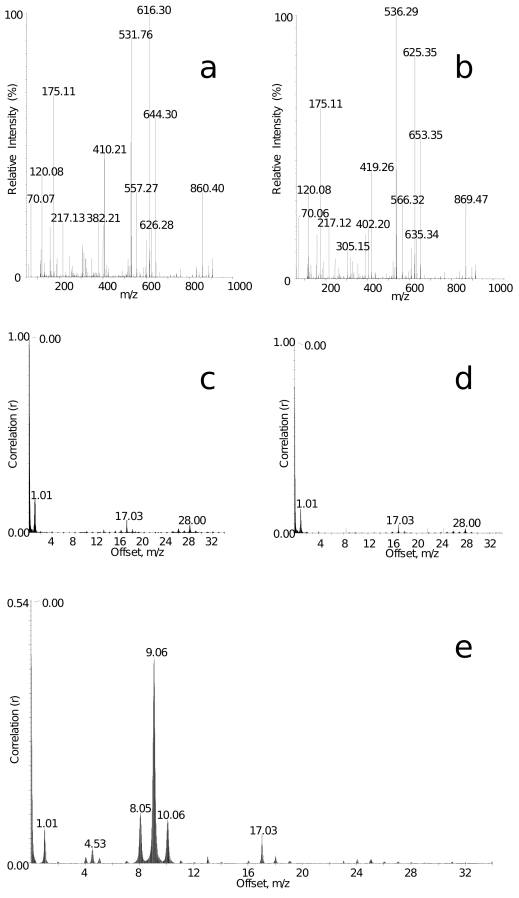


Figure 2.2: Tandem mass spectra of a peptide modified with the light (a) and the heavy (b) version of an N-terminal isotope-coded tag. Delay-series autocorrelation plots of the spectra shown in (a) and (b) are shown in (c) and (d), respectively. Delay-series cross- correlation of the spectra shown in (a) and (b) produces plot (e).

signal representing isotopic overlap are visible at 1.01, 8.05, and 10.06 Th. It is also interesting to note the peak at an offset of 4.53 Th due to the overlap of the remaining doubly charged light precursor with the remaining doubly charged heavy precursor.

The charge state of the ions producing the peaks in these analyses is revealed by examining the distances between the correlation signals caused by overlap of the monoisotopic peak with satellite isotopic peaks, showing that the analysis can integrate information from multiple charge states concurrently.

This example shows how performing a delay-series cross-correlation on two different spectra can reveal the presence of mass shift signals and how these can be interpreted as m/z differences in peak locations occurring between the two spectra. In a similar manner, a spectrum can be correlated with a copy of itself (a delay-series autocorrelation) to reveal the mass shift signals present within a single spectrum. The use of this technique is typified by the detection of a neutral loss modification in MS data. Since some neutral loss modifications are moderately labile, the MS spectrum obtained may feature two peaks, corresponding to the modified and unmodified form of the peptide, in two different regions of the m/z axis. These peaks can arise from a modified peptide losing its modification during the ionization process or in the mass spectrometer itself. As a result of this characteristic, a single spectrum containing an internal mass shift signal corresponding to the mass of the neutral loss of the modification may be produced. If such mass shift information is present within a single spectrum, a delay-series autocorrelation is easily able to extract it.

Figure 2.3a shows an MS spectrum of a peptide containing a phosphorylated tyrosine (DRV$\underline{\text{Y}}$IHPF) with its $[M + H]^+$ at 1126.47 Th and the unmodified version of the same peptide at 1046.50 Th. Autocorrelation of this spectrum produces the delay series seen in Figure 2.3b. The commonly observed elements of autocorrelation are present, with a perfect correlation score (r-value) at a 0 Th delay and a moderate correlation corresponding to the overlap of the monoisotopic peak with the first $_{13}$C isotope peak of each signal. Immediately evident is the large correlation at an offset of 79.98 Th, corresponding to the mass of HPO_3 and the single charge state of the peptide. This means that there is significant information in the spectrum that is separated by the mass of HPO_3.

Several features of these delay-series autocorrelation analyses make them particularly well suited to identifying the neutral loss of peptide modifications compared to other modification screening methods. Traditional neutral loss identification approaches are hardware-based and are lacking in flexibil-

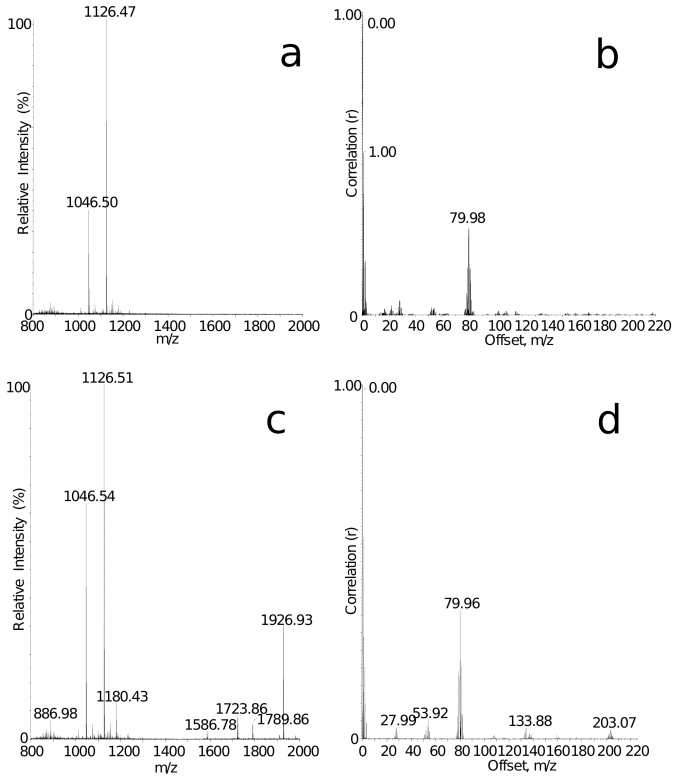


Figure 2.3: MS spectrum of tyrosine-phosphorylated peptide DRVYIHPF (a) and the corresponding delay-series autocorrelation plot (b). The MS spectrum for a mixture of tyrosine-phosphorylated DRVYIHPF, O-linked galactosamine-modified HLLVSNVGGDGEEIER, and O-linked galactosamine-modified LLVSNVGGDGEEIER is shown in (c) with its corresponding delay-series autocorrelation plot (d).

ity. For instance, the popular method of using offset scanning quadrupoles to isolate and detect neutral losses while scanning through a mass range requires specific instrumentation, and the duty cycle of these instruments limits the number of potential modifications that can be considered in each analysis [21]. Replacing these analyses with delay-series correlation circumvents these limitations, allowing the researcher to look for neutral loss peptide modifications off-line, on any data type, for multiple modifications at any charge state [22]. The added benefit of off-line analysis is the ability to utilize old or archived data, which may contain interesting or useful features, but which was collected before the need for a mass shift analysis was apparent. In addition, searching for multiple neutral loss modifications using a single spectrum (or set of spectra) is easily accomplished using this approach (Figure 2.3c, 2.3d).

The spectrum in figure 2.3c is an MS spectrum containing three peptides, each of which appear in their modified and unmodified forms. One peptide (DRV<u>Y</u>IHPF), with a phosphorylated tyrosine, appears at 1126.51 and 1046.54 Th. The second peptide (LLV<u>S</u>NVGGDGEEIER), with an O-linked galactosamine-modified serine, appears at 1789.86 and 1586.78 Th. The final peptide (HLLV<u>S</u>NVGGDGEEIER), with an O-linked galactosamine-modified serine, appears at 1926.93 and 1723.86 Th. It is possible to autocorrelate this single spectrum and simultaneously discover the two different neutral losses that are present. The result of autocorrelation (Figure 2.3d) shows significant correlation at offsets of 79.96 and 203.07 Th, corresponding to the loss of HPO_3 and N-acetylgalactosamine, respectively, from singly charged precursors. Also appearing in the autocorrelation plot are positive correlations, which do not correspond to the mass of known modifications, arising from coincidental overlaps within the spectrum. Two examples are the correlations at an offset of 133.88 and 53.92 Th arising from the overlap of the peak at 1180.43 Th with the peaks at 1046.54 and 1126.51 Th, respectively. As is evident from this example, finding more than one modification requires nothing more in terms of instrument time or data collection; all potential neutral losses are identified in a single analysis step. This is a large improvement over hardware-based methods, where long instrument cycle times can result from multiple modification screening, limiting its utility. An entire analytical strategy has been constructed around this autocorrelation-based analysis, demonstrating its utility in the detection of neutral losses from LC-MS/ MS data [22].

The techniques discussed above use correlation to transform mass spectra from m/z domain signals to mass shift domain signals, revealing mass offsets in the data as peaks in the mass shift domain. These mass shifts in the data

may allude to interesting features in the underlying biological sample, such as post-translational modifications. However, in addition to identifying these mass shifts, it would also be useful to locate the precursor ions producing the mass shift signal in the original, uncorrelated spectrum, as locating the origin of such a signal in peptide MS data would yield information about which peptides are modified, assuming that a signal representing the modification was seen in the delay-series autocorrelation of the peptide MS data. An extension of the correlation-based techniques discussed above can provide this information.

The modification-localization approach presented is referred to as convolution mapping and is performed by decomposing the summation in the numerator of eq 2.2, where the value of d (the delay, indicating the magnitude of the spectral offset) produced a local maximum in the autocorrelation analysis. Thus, the selection of the value of d to use is data-dependent. Instead of summing over all pairwise products of the data vectors (at each bin index incrementally), these products are plotted against the m/z of bin i+d. This allows a direct visualization of the spectral regions that contribute to the large value of the correlation coefficient at that specific offset. Moreover, it is possible to distinguish whether a strong delay-series autocorrelation signal arises from many small contributions originating from many peaks or from a small number of large contributions. This latter case is the most interesting from the point of view of mapping neutral loss modifications, where a large contribution to the convolution value (much larger than average) from a single peak or a small number of peaks could identify the m/z of one or more modified precursors. Compared to other localization approaches attempted, the convolution mapping technique is simple to implement and is computationally trivial.

Figure 2.3a and 2.3b, where a delay-series autocorrelation revealed significant spectral overlap at an offset of 79.98 Th, indicating a potential neutral loss of HPO_3, can now be examined further. Performing convolution mapping on the MS spectrum shown in Figure 2.4a (with a zoom of the $[M + H]^+$ ion of tyrosine-phosphorylated DRV<u>Y</u>IHPF at 1126.47 Th in Figure 2.4b), using 79.98 Th as the value of the delay, reveals the major contributors to the large autocorrelation value seen in the autocorrelation plot; in this case, there is a single large contributor at 1126.5 Th, corresponding to the $[M + H]^+$ of tyrosine-phosphorylated DRV<u>Y</u>IHPF (Figure 2.4c). Closer examination of this result shows that peak shape and resolution are conserved with respect to the original data, allowing isotopic resolution (Figure 2.4d). Thus, the single major contributor to the large autocorrelation value at 79.98 Th can be readily localized.

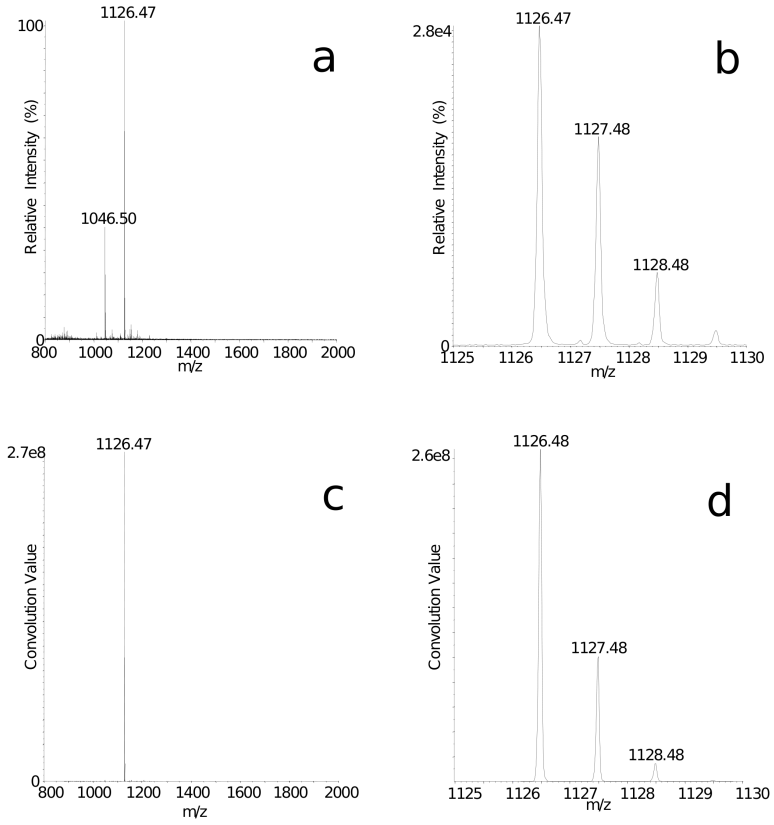


Figure 2.4: MS spectrum of tyrosine-phosphorylated peptide DRVYIHPF (a) with a zoom of the precursor ion (b). A convolution map of the spectrum at an offset of m/z 79.98 is plotted in (c) with a zoom of the major convolution contributor shown in (d).

Performing a delay-series autocorrelation on peptide MS/MS data will yield the exact same information as on MS data – mass shift signals – which in the context of modified peptides may indicate particular modifications within the peptide. These signals arise in the same manner as in MS data, except that the mass shift signals occur between matching pairs of modified and unmodified fragment ions, instead of between matching pairs of modified and unmodified peptide precursor ions. To illustrate this, tandem mass spectra were acquired for DRVYIHPF in its phosphorylated (Figure 2.5a) and unmodified form (Figure 2.5b). Performing the delay-series autocorrelation analysis on these two spectra, as was done for MS data, clearly shows a small but specific positive correlation at 79.97 Th for the phosphorylated peptide (Figure 2.5c) but not for the unmodified peptide (Figure 2.5d). A higher average correlation intensity in the autocorrelation shown in Figure 2.5d is commensurate with the lower signal-to-noise ratio in the spectrum in Figure 2.5b. These results show that although the majority of the MS/MS signals do not overlap at an offset of 79.97 Th, the overlap of modified fragment ions with their counterparts which have undergone a neutral loss of HPO_3 is still readily visualized.

It is very difficult to ascertain whether peaks are involved in mass shifts of 79.97 Th directly from the original source data (Figure 2.5a). However, performing the convolution mapping analysis on these data at the delay values that maximized the correlation signal reveals all fragment ions that have potentially undergone a neutral loss of HPO_3 (Figure 2.5e). Although coincidental overlaps occur between noise signals at all offsets, this targeted approach can readily visualize the ion series, which appears in both its modified and unmodified forms. Fragment ions that do not contain the modification will not appear, allowing a simple readout of the modified ion series, which ends at the modified residue. In this case, the phosphorylated b-ions corresponding to peptide DRV<u>Y</u>IHPF with a phosphotyrosine residue (underlined) are the major components of the convolution map produced at an offset of 79.97 Th. This ion series contains the fragment ions b6 (DRV<u>Y</u>IH) at 864.4 Th, b5 (DRV<u>Y</u>I) at 727.3 Th, and b4 (DRV<u>Y</u>) at 614.2 Th. Since all b-ions in the convolution map are modified, and the b-ion series extracted in the convolution map does not continue below the b4-ion, the phosphorylation can be localized to the tyrosine at position 4. In this manner, convolution mapping can be used to locate the site of any modification that produces a neutral loss. This technique can be used to extract ions that have undergone any secondary fragmentation producing a neutral loss; it is not limited to the discovery of post-translational modifications. For instance, this technique has been used to extract a b-ion series

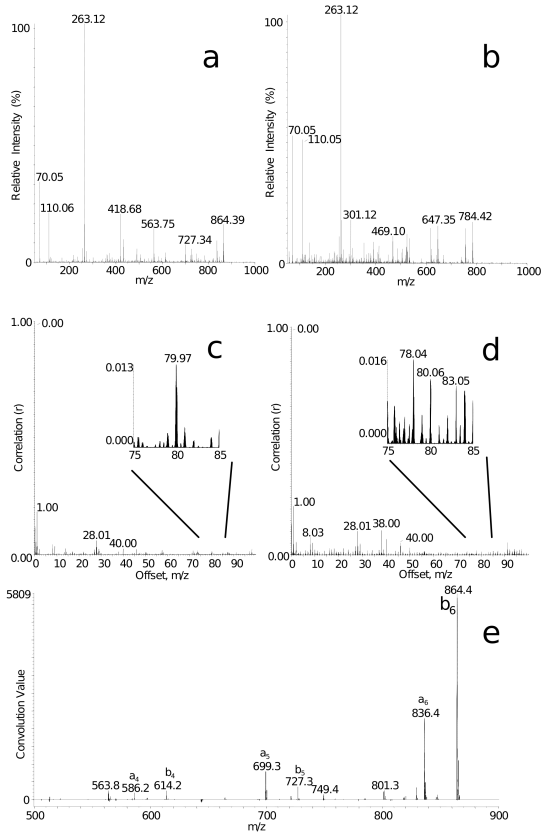


Figure 2.5: Tandem mass spectra of tyrosine phosphorylated peptide DRVY̲IHPF (a) and unmodified peptide DRVYIHPF (b). Delay-series autocorrelations of the spectra shown in (a) and (b) are plotted in (c) and (d), respectively. A convolution map of the spectrum shown in (a) at an offset of m/z 79.97 is shown in (e).

using the 27.99 Th shift between b- and a-ions, as well as a y-ion series using the 17.01 Th shift between y- and y - NH3 ions (data not shown).

The correlation and convolution analysis techniques have so far been demonstrated using example experiments on individual peptides that are labeled or modified. However, the uses envisioned above for these methods will require their application to real-world experiments and mass spectrometry data sets. How these analyses are able to handle complex mixtures, large data sets, large dynamic ranges, differing spectral complexity, and varying data quality will determine their utility in proteomics. As a demonstration, delay-series autocorrelation analysis and convolution mapping were applied to an LC-MS/MS analysis of a tryptic digest of the secreted fraction of S. *typhimurium* cultured under SILAC conditions, to identify SILAC peptides. The sample was first analyzed by LC-MS/MS and produced the total ion chromatogram shown (Figure 2.6a). The autocorrelation of each MS spectrum from this TIC was then calculated using a range of offsets from 2.906 to 3.106 Th, representing the position of expected local maxima from doubly charged peptide pairs separated by 6 Da. The maximum correlation coefficient value from this range was plotted against TIC elution time for each of the spectra, producing an autocorrelation chromatogram (Figure 2.6b). The peaks in the autocorrelation chromatogram were indicative of mass spectra that had significant signal overlap at 3 Th and therefore likely contained doubly charged SILAC peptide pairs. It is important to keep in mind that the correlation value in the autocorrelation chromatogram is not directly related to the total ion intensity of the corresponding spectrum, but is instead a measure of the total similarity of spectral signals when offset by the given amount.

To localize the m/z values of the SILAC peptide pairs from the autocorrelation chromatogram, convolution mapping was performed on each MS spectrum, which generated a maximum autocorrelation value (r-value) greater than 0.02 in the range of offsets used above. This selection threshold was chosen to eliminate spectra with insignificant autocorrelation values at an offset of 3 Th. The resulting convolution maps were pooled, producing a list of 35 putative SILAC doublets. Of these values, three are due to intense triply charged, singly modified SILAC doublets which, as a result of the charge state, are separated by 2 Th, but in which the third isotope peak of the heavy signal overlaps with the monoisotopic peak of the light signal and produces a significant contribution to the convolution. One further isotopic overlap occurs in a similar manner between two intense SILAC signals, which are triply charged and doubly modified, where the third isotope peak of the light signal overlaps the monoisotopic peak of the heavy signal. The

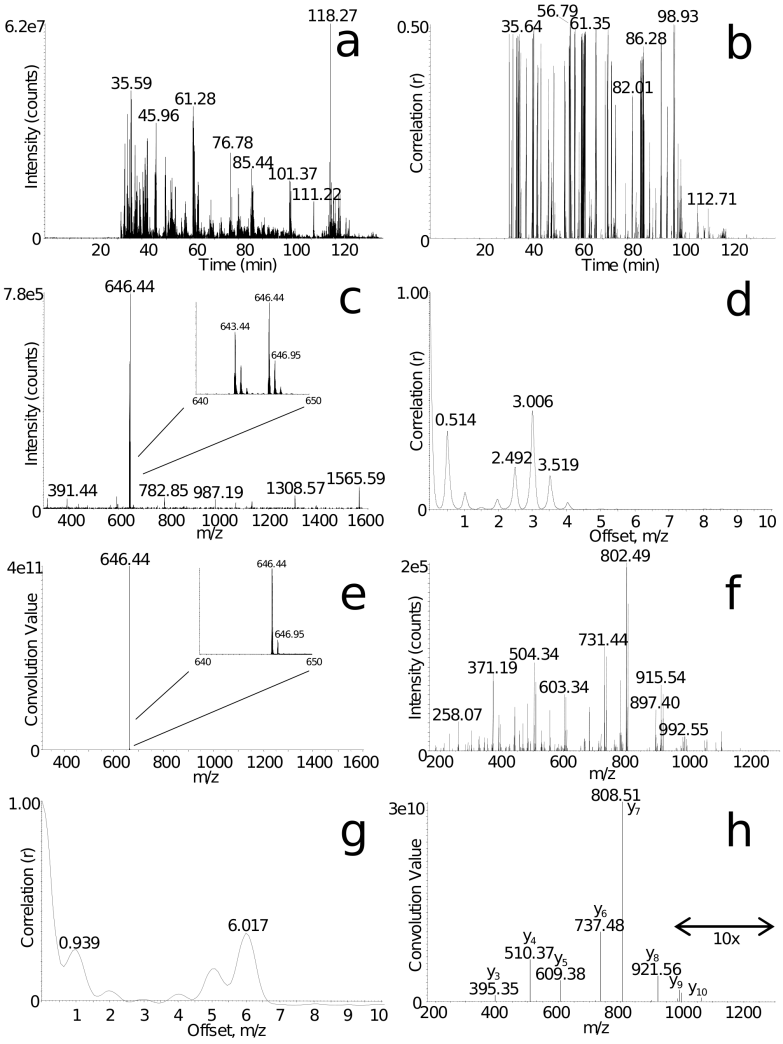


Figure 2.6: Total ion chromatogram (a) of the LC-MS/MS analysis of the secreted fraction of S. *typhimurium*. The autocorrelation chromatogram of all MS spectra from this experiment at a 3 Th mass shift is shown in (b). The MS scan at 61.35 min is shown in (c) with its delay-series autocorrelation plot (d). A convolution map of this MS spectrum at an offset of 3.006 Th is shown in (e). An MS/MS scan collected on both of the ions in the SILAC doublet simultaneously is shown in (f) and its delay-series autocorrelation shown in (g). A convolution map of (g) at an offset of 6.017 Th is shown (h).

remaining 31 m/z values found by convolution mapping correspond to doubly charged, singly modified SILAC doublets, 25 of which match to proteins identified by MASCOT. The remaining six were not identified by MASCOT, while all doublets from the MS scan that were identified by MASCOT were also found by autocorrelation and convolution mapping.

To perform identifications, MASCOT uses MS/MS data from intensity-based, data-dependent scanning, relying on successful peptide matches from MS/MS data to identify SILAC peptides. This method can only identify a SILAC peptide if it produces a dominating signal in the MS spectrum that satisfies the intensity-based selection criteria, allowing an MS/MS spectrum of the peptide to be collected. In contrast, convolution mapping allows peptide precursor masses to be determined specifically for SILAC peptide doublets using the autocorrelation maxima resulting from the SILAC-introduced mass shift. This method does not require the collection of MS/MS spectra and requires only that both peaks in the SILAC doublet have intensities greater than the mean spectral intensity (or a higher, manually selected threshold) in order to produce a significant correlation value.

As an example of what can be accomplished with further analytical processing, one of the 31 potential SILAC peptide signals in the autocorrelation chromatogram was selected for further analysis (the peak at 61.35 min in the autocorrelation chromatogram); its MS spectrum is shown in Figure 2.6c. Autocorrelation of this MS spectrum showed a prominent correlation at a 3.006 Th offset (Figure 2.6d), as it is hypothesized from the previous analytical step that this MS spectrum contains a doubly charged SILAC peptide pair. Performing a convolution mapping analysis using this offset (Figure 2.6e) reveals the major contributor to this high correlation value to be the doublet at 643.44 and 646.44 Th in the MS spectrum, corresponding to the light and heavy versions of IDAALAQVDTLR, respectively (as determined by an independent MASCOT search). The less intense signals from Figure 2.6c do not appear in this convolution map as they are not members of a doublet separated by 3 Th. The convolution mapping technique reports the major contributor as a single peak at the m/z value of the heavier component of the doublet; in this case, the heavier peak is located at 646.44 Th. Simultaneous fragmentation of both the heavy and light peptides in this doublet produces a complex spectrum containing doublets corresponding to y-type ions and single peaks corresponding to b-type and internal fragment ions due to the location of the SILAC isotope label on the C-terminal residue (Figure 2.6f). When an autocorrelation was performed on this MS/MS spectrum, a strong correlation was observed at 6.017 Th, corresponding to the overlap of singly-charged y-type ions offset by the SILAC induced mass shift

(Figure 2.6g), and a convolution analysis using this offset was able to specifically extract the y-ion series from the spectrum (Figure 2.6h). Though the analytical workflow in this example was demonstrated using a specific spectrum, the same procedure has been applied to all of the spectra producing high correlation values in the autocorrelation chromatogram. The ion-series extraction method works well in this case due to the consistent 6 Th mass shift introduced by the incorporated heavy ($_{13}C_6$) and light ($_{12}C_6$) arginine at the C-terminus of the peptide.

2.4 Conclusions

The delay-series autocorrelation and convolution mapping procedures outlined here represent a new data analysis approach involving specific analytical processes to tackle specific analytical problems. This approach cannot replace the comprehensive, sequence identification-based exploratory proteomic analyses provided by applications such as MASCOT or SEQUEST, but may provide a fast and robust supplementary analysis in situations where mass shift signals are known to convey desired information. Several such instances have been demonstrated in this study, including mass shifts introduced by stable isotope labeling, the analysis of neutral loss-associated post-translational modifications, and the extraction of specific ion series using pertinent mass shift information; these are summarized in Table 2.1. In the location of modifications by neutral loss, this is a particular benefit, as modifications may be identified without having to specify which modifications – and therefore which mass shifts – to look for. Though other analysis methods are certainly capable of providing such information, the advantages of delay-series autocorrelation and convolution mapping lie in their algorithmic and mathematical simplicity, making them very quick, and in the inherently comprehensive nature of the analysis they provide, enabling a search for all mass shifts simultaneously.

2.5 Acknowledgments

The authors thank Dr. Leonard Foster at the UBC Centre for Proteomics for his generous donation of the SILAC LC-MS/MS data. This work was funded in part by grants from the Canadian Institutes of Health Research (CIHR), the Protein Engineering Network of Centres of Excellence (PENCE), and the Michael Smith Foundation for Health Research (MSFHR). M.D.H. would like to thank the Natural Sciences and Engineering Research Council of

Canada (NSERC) for a CGS scholarship.

Table 2.1: Summary of Successful Analyses Performed Using Correlation/Convolution Mapping for Specific Mass Offsets.

Analysis Type	**Origin of Mass Shift**		**m/z offset (Th), charge state, data type**
cross-correlation	stable isotope label	in-house tag	9,1+,MS/MS
		SILAC	6,1+,MS/MS
autocorrelation	stable isotope label	in-house tag	4.5,2+,MS 9,1+,MS/MS
		SILAC	3,2+,MS 6,1+,MS/MS
	neutral loss	O-HexNAc serine	101.5,2+,MS 203,1+,MS/MS
		phosphoserine	49,2+,MS/MS
		oxidized methionine	32,2+,MS/MS
		sulfotyrosine	80,2+,MS
		farnesyl cysteine	102,2+,MS 204,1+,MS/MS
		b- and a-ions	28,1+,MS/MS
		y- and y-NH_3 ions	17,1+,MS/MS
cross-convolution mapping	stable isotope label	in-house tag	9,1+,MS/MS
autoconvolution mapping	stable isotope label	in-house tag	4.5,2+,MS 9,1+,MS/MS
		SILAC	3,2+,MS 6,1+,MS/MS
	neutral loss	phosphotyrosine	80,1+,MS/MS
		phosphoserine	49,2+,MS/MS
		O-HexNAc serine	101.5,2+,MS
		oxidized methionine	32,2+,MS/MS
		sulfotyrosine	80,1+,MS
		b- and a-ions	28,1+,MS/MS
		y- and y-NH_3 ions	17,1+,MS/MS

Bibliography

[1] Patterson, S. D. *Nat Biotechnol* **21**(3), 221–2 (2003).

[2] Nesvizhskii, A. I. and Aebersold, R. *Drug discovery today* **9**(4), 173–181 2/15 (2004).

[3] Chervet, J., Ursem, M., and Salzmann, J. *Analytical Chemistry* **68**(9), 1507–1512 (1996).

[4] Wilm, M. and Mann, M. *Analytical Chemistry* **68**(1), 1–8 (1996).

[5] Medzihradszky, K., Campbell, J., Baldwin, M., Falick, A., Juhasz, P., Vestal, M., and Burlingame, A. *Analytical Chemistry* **72**(3), 552–558 (2000).

[6] Zhang, Z. and McElvain, J. *Analytical Chemistry* **72**(11), 2337–2350 (2000).

[7] Perkins, D., Pappin, D., Creasy, D., and Cottrell, J. *Electrophoresis* **20**(18), 3551 (1999).

[8] Eng, J., McCormack, A., and Yates, J. *Journal of The American Society for Mass Spectrometry* **5**(11), 976 (1994).

[9] Sadygov, R., Eng, J., Durr, E., Saraf, A., McDonald, H., MacCoss, M., and Yates, J. *Journal of Proteome Research* **1**(3), 211–215 (2002).

[10] MacCoss, M., Wu, C., and Yates, J. *Analytical Chemistry* **74**(21), 5593–5599 (2002).

[11] Yates, J., Eng, J., McCormack, A., and Schieltz, D. *Analytical Chemistry* **67**(8), 1426–1436 (1995).

[12] D., F. *Trends in Biotechnology* **20**, 35–38(4) (1 December 2002).

[13] Craig, R. and Beavis, R. C. *Bioinformatics* **20**(9), 1466–1467 (2004).

[14] Proakis, J. and Manolakis, D. *Digital signal processing: principles, algorithms, and applications*. Prentice Hall, Upper Saddle River, NJ, (1996).

[15] Hall, M. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, (1999).

[16] Bendat, J. and Piersol, A. *Engineering applications of correlation and spectral analysis*. Wiley, New York, (1980).

[17] Owens, K. *Applied spectroscopy reviews* **27**(1), 1 (1992).

[18] Rogalski, J. C., Taylor, R., Beaudette, P., Lin, M. S., Lin, S., Sniatynski, M. J., and Kast, J. In *J. Am. Soc. Mass Spectrom.*, volume 14, 71S, (2003).

[19] Foster, L. J., de Hoog, C. L., and Mann, M. *Proceedings of the National Academy of Science* **100**, 5813–5818 May (2003).

[20] Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. *Mol Cell Proteomics* **1**(5), 376–386 (2002).

[21] Le Blanc, J., Hager, J., Ilisiu, A., Hunter, C., Zhong, F., and Chu, I. *Proteomics* **3**(6), 859–69 (2003).

[22] Hoffman, M. D., Sniatynski, M. J., Rogalski, J. C., Le Blanc, J. Y., and Kast, J. *Journal of the American Society for Mass Spectrometry* **17**(3), 307–317 3 (2006).

Chapter 3

Singular Value Decomposition and Principal Components Analysis for MS/MS Feature Detection and Fragmentation Modeling[2]

3.1 Introduction

Mass spectrometry datasets may be represented as vector spaces, with the dimensionality of the space corresponding to the number of m/z points in each data vector (spectrum), and each separate spectrum in the database represented by a point in this space. This representation provides new opportunities for data analysis, as it makes powerful data interpretation tools originating in linear algebra compatible with mass spectrometry data.

One of the most popular applications of linear algebra in the area of exploratory data analysis is the eigenvector/eigenvalue matrix decomposition, which has been used widely in scientific endeavours to carry out principal component analysis (or factor analysis) on data. This matrix decomposition extracts general tendencies and trends that occur in the dataset as a whole, but that are obscured by the variability that can occur between data vectors, and the (often vast) abundance of these data vectors themselves.

The eigendecomposition in its pure form is defined only on square matrices, which poses no problems in traditional uses such as factor analysis, which involve generating the eigenvalues and eigenvectors of statistical co-

[2]A version of this chapter will be submitted for publication. Sniatynski, M.J. and Kast, J. Singular Value Decomposition and Principal Components Analysis for MS/MS Feature Detection and Fragmentation Modeling.

variance matrices. The eigenvectors thus produced represent the underlying "factors" that explain the variability observed in the dataset, and the corresponding eigenvalues indicate the amount of variance explained by each factor.

However, matrices of mass spectrometry data are unlikely to be square. A matrix storing 500 spectra, each composed of 2000 m/z values, will have size 500×2000. A traditional eigendecomposition cannot be carried out on this matrix. This limitation was realized early on in the formulation of algebraic axioms by late 19th century mathematicians such as David Hillbert, who coined the term "spectral theory" to describe theoretical approaches that extended the eigenvalue/eigenvector approaches to situations beyond those represented by square matrices.

The most important and useful result from spectral theory is that the data contained in a square matrix may be represented without loss of information using a basis of the matrix's eigenvectors. The eigenvectors of a matrix are vectors assembled from linear combinations of the existing basis vectors (such as the X, Y, and Z axes in a 3-dimensional plot), such that the total variation in the data contained in the columns of this matrix is spread among as few dimensions as possible. The eigenvectors that define the new basis are often called principal component vectors. The construction of such an eigenbasis is shown in a two dimensional case in figure 3.1.

The singular value decomposition (SVD) is a generalization of this eigenbasis construction technique to matrices that are not square, relying on the following theorem: any $m \times n$ dimension matrix A with real-valued entries has a factorization $A = U * S * V^T$ (Called the singular value decomposition), in which U is an $m \times n$ matrix with principal component vectors in the columns (the new orthonormal basis vectors), S is an $n \times n$ diagonal matrix with each diagonal entry expressing the relative importance of the corresponding U vector in describing data spread, and V is an $n \times n$ matrix where the rows contain the datapoint coordinates projected onto the corresponding principal component axis [1]. This is shown graphically in figure 3.2.

The new eigenbasis defined by the columns of the U matrix is constructed from these orthogonal principal component vectors in such a way that the first basis vector lies along the axis of maximal data variance, the second basis vector lies along the axis of maximal data variance orthogonal to the direction of the first vector, and so on, until the new eigenbasis has the same dimensionality as the original matrix. Because these new basis vectors are combinations of the original input variables, linear dependencies within each data vector are removed. The decreasing importance of each consec-

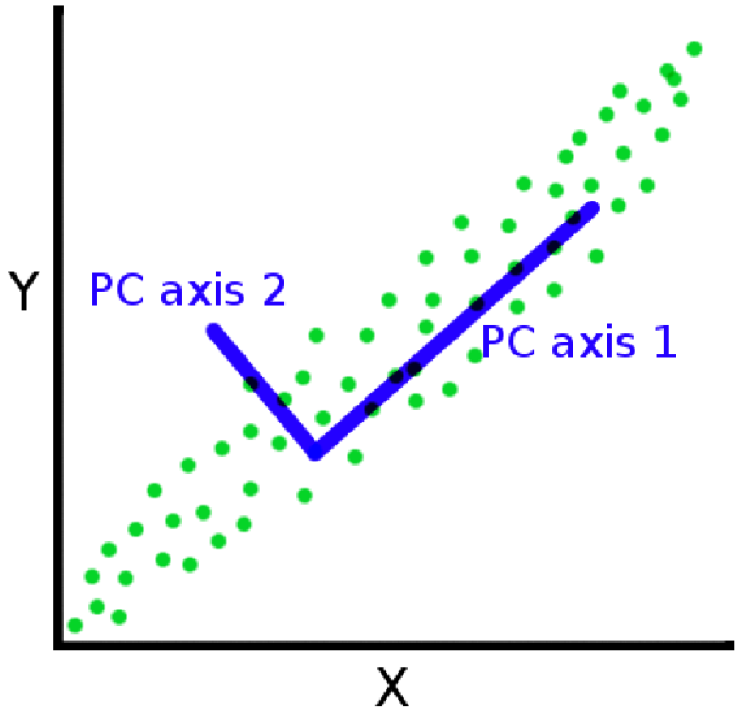


Figure 3.1: A 2-dimensional eigenbasis constructed from a 2-dimensional input space. The first principal component basis vector (PC axis 1) is positioned to span as much data variance as possible, while the second (PC axis 2), spans as much variation as possible, in an orthogonal direction.

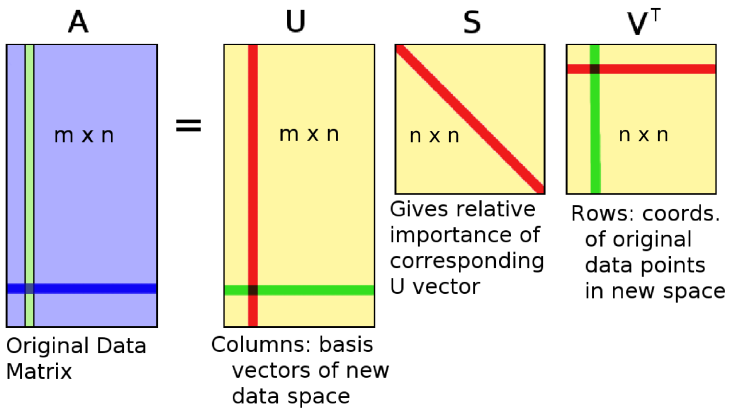


Figure 3.2: A graphical demonstration of the SVD procedure, showing details of the input matrix, and the three output matrices that result.

utively added basis vector also enables data compression approaches – the dimensionality of the new basis space is chosen to optimally compromise between size/storage, and data representation. Data compression is achieved by choosing an appropriate level of truncation – an $m \times n$ data matrix will yield an n dimensional principal component basis space, but before use this matrix is generally truncated to an $m \times k$ matrix, with $k << n$. The error that this incurs in terms of data representation is given precisely by the Frobenius norm of the matrix of differences, U_{diff}, which is constructed by subtracting the truncated matrix from the original full size matrix, after padding the truncated matrix with zeros to give it the same dimension as the full size matrix: $U_{diff} = U_n - U_k^{(padded)}$. The Frobenius norm for an $m \times n$ matrix A is given as follows:

$$\|A\| = \sqrt{\sum_{i}^{m}\sum_{j}^{n}|a_{ij}|^2} \tag{3.1}$$

This matrix factorization is put to use in many fields ranging from image compression [2] to latent semantic analysis in text mining [3], as it is by construction highly suitable for use in situations where differences between data instances are manifested in multiple simultaneous changes in a highly correlated set of features. In biological studies, it has been used for the

clustering of protein alignments [4], and for analyzing genome-wide shifts in gene expression [5]. A situation featuring highly correlated data arises in the analysis of peptide fragmentation in mass spectrometry, where the features may correspond to peak intensities of a-ions, b-ions, y-ions, and derivative secondary ions. By compiling mass spectrometry data into a matrix form and applying the SVD procedure, it is possible to visualize these correlated sets of features that correspond to actual differences occurring in the dataset. Though some generalized classification studies of mass spectra have made use of the SVD [6], The application of the SVD to fragmentation modeling has not yet been demonstrated.

Though basic peptide fragmentation is generally well understood, it serves here as a familiar example of the potential applications of this data analysis technique, and demonstrates its utility in revealing effects of peptide properties on fragmentation patterns, as well as in characterizing uncommon fragment ion effects.

3.2 Methods

A large dataset consisting of tryptic peptide MS/MS spectra from a Q-TOF instrument was collected. MASCOT searches were performed on this data, and the MS/MS spectra that produced the top scoring 200 peptide identifications were retained. The associated sequence identifications were used to extract 100 Th wide windows from the mass spectrum around the predicted masses of each associated a-, b-, and y-ion in these spectra, with each 100 Th window divided linearly into 2000 bins, producing a data vector 6000 elements in length. Depending on the analysis to be performed, up to three additional bins were added to the 6000, that contained either the mass of the a-, b-, or y-ion, or the mass of the intact peptide from which the fragmentation had been extracted (calculated by adding the b- and y-ion masses). These 6000 to 6003 element datavectors were then assembled into a matrix – each column corresponding to a common peptide backbone fragmentation, and each row corresponding to a common feature "bin" in each fragmentation data vector. The SVD procedure was carried out on this matrix, and vectors from the three resulting matrices were analyzed

3.3 Results and Discussion

3.3.1 The U Matrix

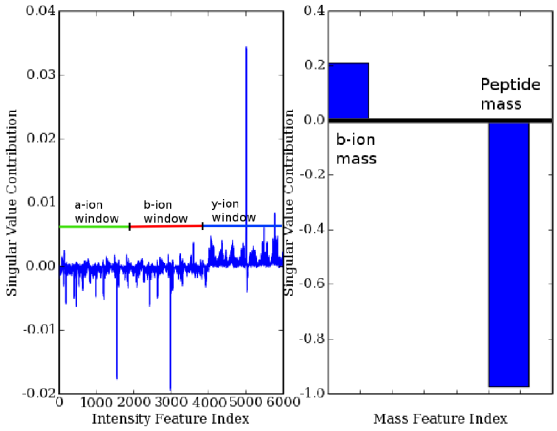
The U matrix that was produced contained the principal component vectors in its columns. These are linear combinations of features that account for the spread of the datapoints in feature space along that principal component axis – in this case, the features were the intensities of the datapoints in the a-, b-, and y-ion windows, and the values of the appended fragment/peptide masses (if present). These vectors defined a new, compacted basis that could optimally span the data space, with each entry in the columns corresponding to the relative importance of that particular feature to the variation present in the data along that particular principal component axis. Figure 3.3 shows two example principal component vectors from two separate analyses. The first analysis included b-mass and peptide mass values along with the spectral intensity values, and is useful for examining the interplay between these masses and key spectral intensity values, such as those representing known primary or secondary fragment ions. The second analysis was similar, but removed the peptide mass from consideration

3.3.2 The S Matrix

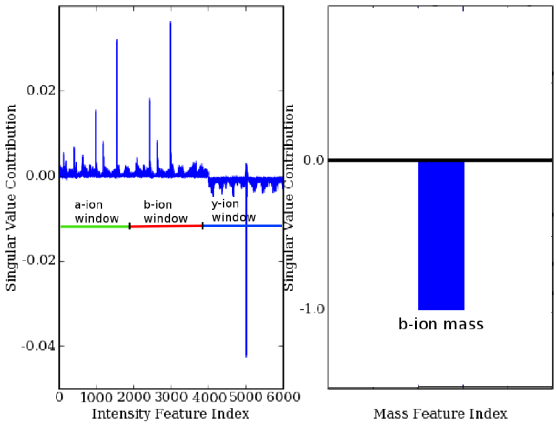
The S matrix is a diagonal matrix that contains the SVD analogue of eigenvalues – the singular values. These values show how much of the global difference between data points in the dataset is spread along their associated principal component vectors in the new, compacted basis space. Singular vectors with higher singular values explain more of the difference in the data than singular vectors with small singular values. Plotting out the singular values in the S matrix provides an intuitive representation of the relative importance of each principal component vector (figure 3.4), and allows for the establishment of a truncation threshold when dataset compression is necessary – such as a truncation to two dimensions for 2-dimensional visual analysis, or to a computationally tractable number of dimensions for model construction. As explained above, the inaccuracy introduced via such a truncation is explicitly given by the Frobenius norm (eqn 3.1).

3.3.3 The V Matrix

The rows of the V matrix contain the coordinates of each datapoint in the newly constructed basis. The first entry of row 0, for example, is a scalar representing the first datapoint's location along the first principal component



(a)



(b)

Figure 3.3: Two example U matrices. Figure (a) shows an example principal component dimension of a data matrix consisting of intensity values, b-mass values, and peptide mass values. Figure (b) shows a similar dimension, but consists of intensity and b-mass values only.

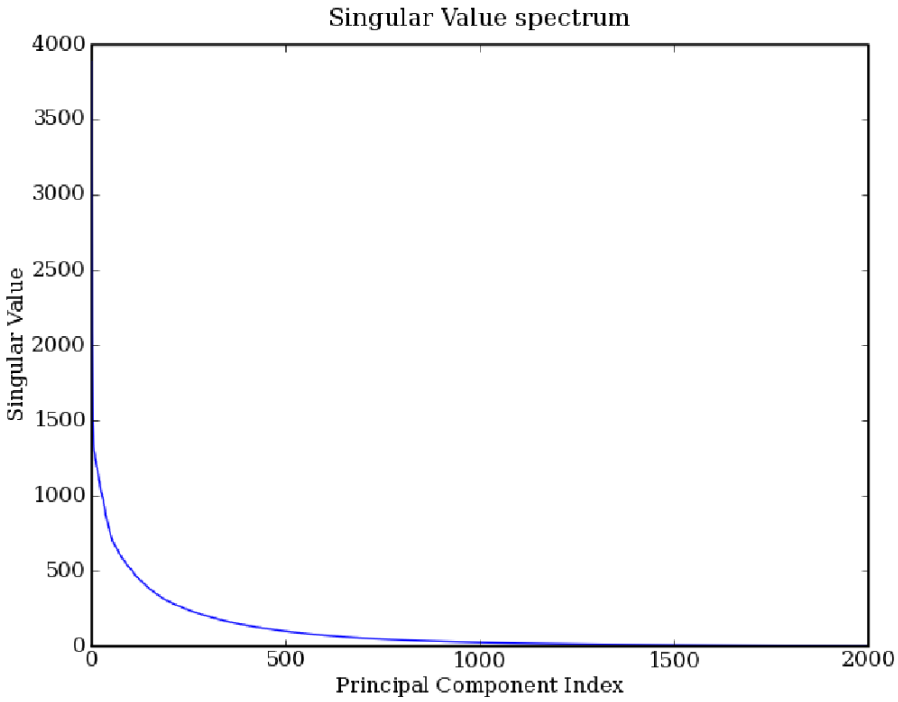


Figure 3.4: Plotting the diagonal elements of the S matrix (all other values are zero) produces the graph shown. The amount of data spread explained by any principal component basis vector (indexed on x-axis) can be found by locating its associated singular value on the y-axis. This importance may be used for establishing truncation thresholds for data compression.

axis, and this row's last entry represents the location of the datapoint along the last principal component axis, which explains the least amount of data spread.

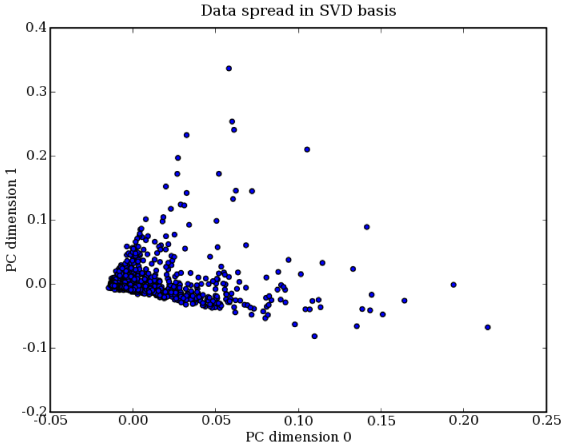
A two dimensional scatterplot showing where each datapoint falls in the new basis space was generated by plotting the first row of matrix V against the second row (Figure 3.5). Since these two dimensions were constructed to contain as much of the variation in the data as possible, this plotting technique was capable of generating a useful picture of a dataset of otherwise intractably high dimensionality.

3.3.4 Exploratory Data Analysis

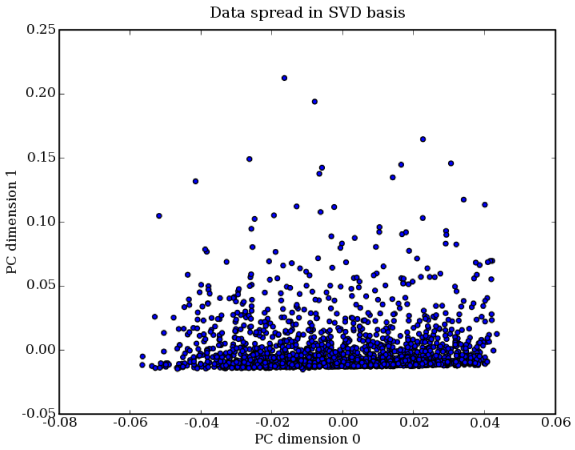
Exploratory Data Analysis (EDA) is an important first step in almost any analysis; human visual processing is extraordinary in its ability to pick out irregular or unexpected anomalies and trends, and its unique ability may be leveraged here by using the SVD to compress an MS/MS dataset into a visually compatible number of dimensions, and then replotting the components of these dimensions. This enables the discovery of trends in the data that would be very difficult to notice by the comparison of individual spectra. An example is shown in figure 3.6, where interesting patterns in the fourth principal component dimension revealed an interesting trend – several of the peaks (visible in the first panel) exhibited inverse changes in their maximum intensities compared to their shoulder intensities (figure 3.6 second panel). This corresponded to a change in general peak shape along this axis, where peak area was retained, but peaks became shorter and wider. This type of high-level qualitative result is made immediately obvious using SVD analysis.

3.3.5 Feature Extraction

Contemporary machine learning models are capable of representing arbitrarily complex relationships among features, and can give high classification/identification accuracy on complex tasks; such tasks are becoming more important as researchers attempt to take advantage of the overabundance of high quality MS and MS/MS data that modern instrumentation and analytical protocol provides. However, the cost of this flexibility is considerable algorithmic complexity, which scales poorly with increasing numbers of features. Using the S matrix to select an appropriate number, n, of principal component basis vectors (columns of the U matrix) and then selecting features to use from these (as in Figure 3.7) will create an optimally truncated

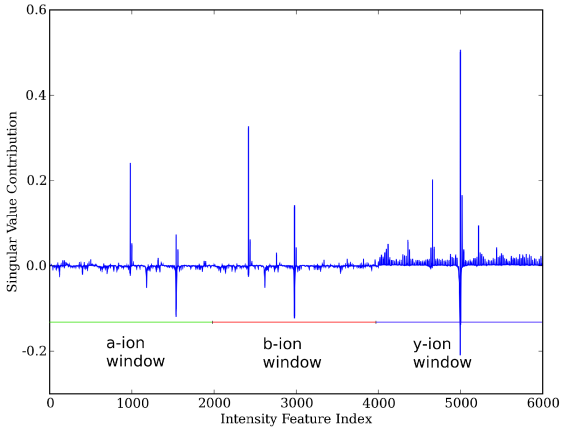


(a)

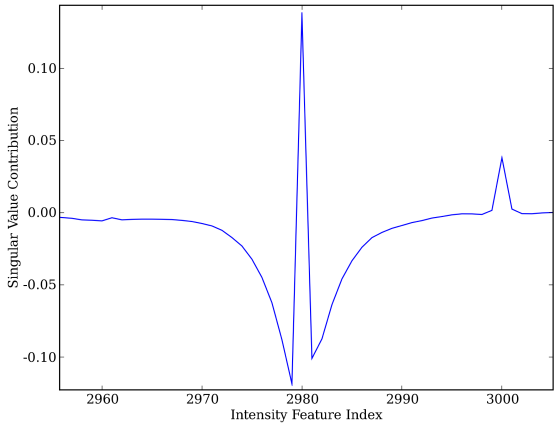


(b)

Figure 3.5: Two example V matrices. The X and Y axes represent principal component basis vectors defined by the column vectors in the U matrix, and the points represent the coordinates of the data in the principal component basis. Both of these plots show the scattering of the fragmentation data vectors in the first two dimensions of the SVD-constructed principal component basis space. Plot (a) shows the spread that occurs when the basis (and the data vectors) do not contain peptide mass information. The visible structure is different compared to (b), where peptide mass information is contained in the basis and the data vectors.



(a)



(b)

Figure 3.6: A potentially interesting principal component dimension is plotted in (a). The accompanying plot, (b), shows a zoom into a region of interesting peak shape.

feature set of size n, which in practice is often far smaller than the size of the full feature set. Alternatively, the translated coordinates of the data points found in the first n rows of the V matrix can be used as features directly when appropriate. This approach can be introduced to any data analysis workflow to provide automated feature selection, which is a crucial step in constructing useful MS and MS/MS models using newer machine learning techniques such as Markov random fields. Though the analysis presented here is limited to cases where feature correlation is linear, related techniques can overcome this restriction. These techniques include functional PCA [7], and a nonlinear PCA variant implemented using neural networks [8].

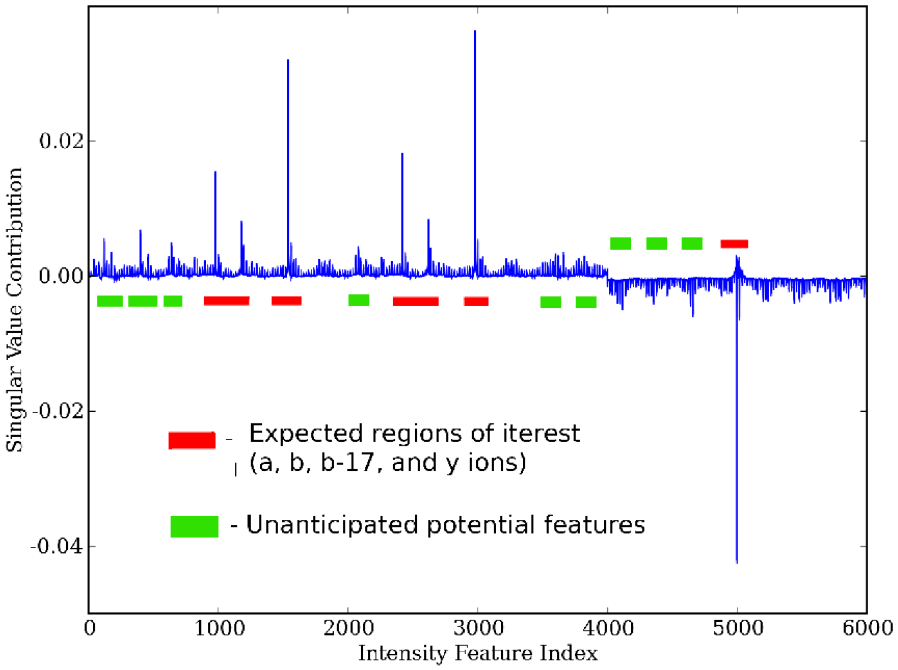


Figure 3.7: Regions that contribute greatly to the magnitude of influential principal component vectors may be selected as potential features for further analysis, or for model construction.

3.4 Conclusion and Future Directions

The approach presented above demonstrates the effectiveness of matrix decomposition techniques in visualizing and interpreting large amounts of very complex data. The utility of this procedure in removing data redundancy and in optimally compressing data was demonstrated, and non-linear extensions to the technique were also briefly introduced. The visualization of trends in very large datasets is of particular concern in mass spectrometry due to the complexity and number of spectra involved, and where the static visualization of trends occurring throughout an entire dataset is extremely useful, and extremely difficult. This general purpose approach also offers a way to optimally compress mass spectrometry data and remove linear correlations, enabling modeling approaches constructed on top of this technique to be smaller in size, and faster in execution time. This approach does assume that spatial separation in data-space correlates with separability in the label space. This is not necessarily true – features causing spectra to be visually different from each other do not necessarily distinguish different spectra from each other. However, other dimensionality reduction techniques have been applied in the classification of mass spectrometry data, including linear discriminant analysis (used in the classification of ovarian cancer data [9]), and a novel combined statistical approach, also used for cancer classification [10]. These techniques are capable of considering class-label and identity information, along with mass spectrum intensities, to identify combinations of optimally discriminative features. The incorporation of class-label and identity information, in a manner similar to these approaches, could provide a powerful extension to the visualization and compression techniques introduced, and merits further study.

Bibliography

[1] Klema, V. and Laub, A. *Automatic Control, IEEE Transactions on* **25**(2), 164–176 (1980).

[2] Andrews, H. and Patterson III, C. *Communications, IEEE Transactions on [legacy, pre-1988]* **24**(4), 425–432 (1976).

[3] Furnas, G., Deerwester, S., Dumais, S., Landauer, T., Harshman, R., Streeter, L., and Lochbaum, K. *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval* , 465–480 (1988).

[4] Ye, J., Janardan, R., and Liu, S. *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on* , 2–9 (2003).

[5] Alter, O., Brown, P., and Botstein, D. *Proc Natl Acad Sci US A* **97**(18), 10101–10106 (2000).

[6] Fung, E., Yip, T., Lomas, L., Wang, Z., Yip, C., Meng, X., Lin, S., Zhang, F., Zhang, Z., Chan, D., et al. *Int J Cancer* **115**(5), 783–789 (2005).

[7] Besse, P., Cardot, H., and Ferraty, F. *Computational Statistics & Data Analysis* **24**(3), 255–270 (1997).

[8] Karhunen, J. and Joutsensalo, J. *Neural Networks* **7**(1), 113–127 (1994).

[9] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2003).

[10] Yu, J., Ongarello, S., Fiedler, R., Chen, X., Toffolo, G., Cobelli, C., and Trajanoski, Z. *Bioinformatics* **21**(10), 2200–2209 (2005).

Chapter 4

Modeling Peptide Fragmentation Using Conditional Random Fields[3]

4.1 Introduction

Pivotal to any successful proteomics research is the ability to accurately identify individual proteins in a sample, or the entire complement of proteins present in a mixture. Accurate, informative protein identification is becoming increasingly challenging as more becomes known about the complexities and intricacies of low-level cellular proteomics. The presence or absence of any of a various number of modifications can influence the function, complexing characteristics, and affinities of proteins for their receptor substrates, and often subtle changes, such as the addition of a single phosphate group act as an on/off switch for a protein's catalytic or enzymatic activity [1, 2, 3]. Many changes occur in proteins in addition to post-translational modifications. Differing protein isoforms and truncations are being discovered at an accelerating rate, and their biological relevance as diagnostic markers, both individually and as a dynamic pattern of interaction, is becoming increasingly appreciated [4, 5, 6, 7, 8, 9].

The availability of this detailed level of information is one key reason to utilize proteomic approaches in biological and biomedical research; a great deal of work in proteomics is now dealing with the study of detailed, low-level protein characteristics, and current analytical approaches are struggling to perform under this new set of demands. Compounding difficulties related to detail are the large variations in abundance typical of cellular proteins. The dynamic range of the normal human proteome spans more than ten orders of magnitude in abundance [10], and significant shifts in the state of the proteome, possibly with useful diagnostic or prognostic correlation,

[3] A version of this chapter will be submitted for publication. Sniatynski, M.J., Rogalski, J.C., and Kast, J. Modeling Peptide Fragmentation Using Conditional Random Fields.

can involve very sparsely abundant proteins. This means that analytical techniques need to make very accurate and detailed determinations using very little evidence; this is becoming a common problem in proteomics research, leading to increasing skepticism regarding single peptide identifications, and focusing attention on the accurate determination of false discovery rates [11, 12, 13]. As our understanding of biological processes reaches unprecedented levels of detail, researchers in proteomics are demanding ever more precise and detailed protein identifications through the entire range of abundance present in the proteome.

Though our understanding of cellular proteomics has changed a great deal since the term was first coined in the early 1990s, the tools used in proteomics experiments have changed very little. Mass spectrometry remains the most powerful and broadly applicable tool in the protein analysis arsenal, and is the only existing approach capable of tackling protein identifications at the whole-proteome level, due to the speed at which it can operate, and the degree to which it may be automated. Newer mass spectrometry hardware is constantly emerging – the recent surge in the popularity of ion cyclotron-based instruments is an example – and performance benchmarks involving the data produced, such as selectivity, sensitivity, resolution, and dynamic range are continually improving.

4.1.1 Protein Identification by Mass Spectrometry

Identifying proteins using mass spectrometry typically involves digesting the protein of interest into its constituent peptides using a proteolytic enzyme, and acquiring precise peptide mass measurements. This peptide mass is referred to as a "peptide mass fingerprint" (PMF), which can be adequate for identifying a highly abundant protein that yields many measurable peptides when digested, if some of these peptides have sufficiently distinct masses to be associated uniquely with the protein in question. In general, however, the PMF alone provides insufficient information to make a unique identification. Proteins frequently yield only a small number of peptides – a problem exacerbated in cases of low protein abundance – and the masses of different peptides can often be nearly the same. Truncations and modifications can also skew the peptide mass considerably. To obtain additional information on the identity of the peptide in the face of this uncertainty, it is typically subjected to a carefully controlled fragmentation that ideally breaks the peptide backbone in one location per molecule, producing a spectrum containing a distribution over each of the possible peptide fragments that can be generated in this manner. This fragmentation spectrum is then associ-

ated with a sequence, typically by direct statistical comparison to predicted fragmentation patterns generated *in silico* from stored database sequences. A model describing how peptides fragment in the mass spectrometer is thus essential to the success of this approach.

However detailed and accurate the raw data from the instrument might be, the detail and accuracy of the final protein identifications also depend on the statistical algorithms and fragmentation models used to interpret the raw data, and to compare the observed fragmentation patterns with the models thereof. Protein identification software and related statistical methods have not generally been keeping pace with enabling advances in hardware in their ability to make use of the detailed information available.

4.1.2 Incorporation of Peptide Fragmentation Information

The principal reason for this is that the models of peptide fragmentation used are deliberately oversimplified to cope with the large amount of variation in fragmentation behaviour that can be present between any two mass spectra of the same peptide. This variation can result from differences in sample preparation, and from differences in the analysis used – particularly from differences in instrument type. Often the experimental spectra are examined only for the presence or absence of the major primary fragment ions resulting from each peptide bond cleavage, with no attempt made to examine derivative ions, or to compare the intensities of fragment ion peaks with model-derived predictions [14]. Restricting the examination of the MS/MS spectrum in this manner provides identifications that are robust against the noise variation introduced by the sources outlined, and provides search tools that work with data of varying quality, produced using many different types of instruments. However, this deliberate simplification discards much of the detailed information reproducibly provided by the latest generation of mass spectrometers – detailed information that could be leveraged to furnish peptide and protein identifications of increased detail and confidence.

This reduction in model detail has particular implications for the identification of sparsely abundant proteins. Due to the lower confidence in the peptide identifications produced by the simplified models, peptide redundancy is employed in protein identification, whereby a protein is only confidently declared as "identified" if multiple peptides enzymatically derived from it are identified during the analysis [15]. In the case of low abundance proteins, the occurrence of more than one peptide in a sample run may have sufficiently low probability as to preclude these single peptide identifications, thereby limiting the utility of mass spectrometric study across an

increasingly important range of protein abundance.

Though underutilized by many popular protein identification methodologies, accurate determinations of the intensity distributions of related fragment ions in MS/MS spectra are increasingly available using modern instrumentation. Ion intensities in these distributions can vary considerably between different fragmentations of the peptide backbone, as the exact mechanism of fragmentation is influenced by many factors both local to the breakage, and non-local, involving the whole peptide [16]. These factors include global peptide characteristics such as relative size, hydrophobicity, and electronegativity, the position of the breakage along the peptide backbone, the chemical properties of the amino acid residues (both local and nonlocal), global sequence motifs/characteristics, and steric effects contributed by residue modifications. These effects have been observed empirically in many studies that focused on the effects of amino acid composition [17, 18], peptide secondary structure and gas-phase conformation [19, 20], and ionization techniques [21] on the fragmentation behaviour of peptides. The effects observed can be reasonably explained using currently popular chemical theories of peptide fragmentation such as the mobile proton model [22], and several of its related derivatives [23, 24].

A new fragmentation modeling approach is needed that can incorporate this type of information in cases where it is reproducibly available, and thus use the available MS/MS fragmentation data to the fullest extent possible. However, extending current theory to provide sufficiently detailed predictions would push established models into domains of unprecedented complexity. Accurate prediction of experimental ion intensities using these models may be too complex a problem in all but the simplest of cases, due to the complexity of the molecular interactions such a large molecule undergoes during mass spectrometric analysis.

The complexity of the physical processes involved in peptide fragmentation, and the need to approximate over many unknowns using limited knowledge, are hallmarks of contemporary proteomics research. These are also the types of systems at which statistical machine learning approaches are targeted. Such systems are able to construct useful models of peptide fragmentation, independently of physics and chemistry theory, by learning de facto fragmentation rules from a set of characterized peptides of known sequence. The challenge lies in identifying the ideal class of model for the situation, to make best use of available information, and to avoid imposing unwarranted assumptions.

The application of statistical machine learning to fragmentation modeling would allow the researcher to build a representation of how peptides

actually fragment during experiments, and use this accumulated knowledge both immediately, to estimate and rank the probabilities of other peptide sequence/spectrum pairs, and as a compilation of observations with which to guide further theoretical research. Though building sufficiently accurate probabilistic models as described requires large amounts of training data, the asymmetry between data generation and data analysis capability in modern, high-throughput proteomics research facilities is very well known; modern instruments and analyses generate a vast abundance of high-quality mass spectra.

4.1.3 Machine Learning Modeling of Peptide Fragmentation

Machine learning models have been successfully applied to fragmentation modeling, but they have mainly been directed at specific data mining tasks, gathering evidence in support of theoretical models. Zhonqui Zhang has used statistical machine learning to estimate the parameters of an explicit multi-pathway fragmentation model based on the mobile proton model [25], and has more recently extended this approach to peptides at higher charge states [26]. Similar machine learning parameterization approaches to versions of the mobile proton model have been undertaken by Kapp *et al* [27]. Other studies have used these statistical techniques independently of fragmentation models, attempting to predict the intensities of specific peaks, in both peptide mass fingerprint identifications [28], and in MS/MS peptide fragmentation [29]. A recent approach by Huang *et al* has demonstrated an elegant implementation of model-free, unsupervised clustering, combining spectra containing similar fragmentation patterns using penalized K-means, then using decision trees to extract the peptide properties that possibly explain the common fragmentation behaviour [30].

The research described here takes a different approach, and describes a general machine learning framework for modeling peptide fragmentation that relies only on very basic theoretical assumptions, and that can be used to probabilistically match experimental MS/MS spectra to candidate sequences without relying on theoretical predictions of fragment ion intensity distributions, and without *a priori* specification of the peaks, or peptide features, to examine. This framework is based on a new type of probabilistic sequence model known as a conditional random field, or CRF, and uses a novel feature elucidation approach based on the singular value decomposition.

4.1.4 Peptide Fragmentation and the Conditional Random Field

Conditional random fields are conditionally trained Markov random fields, which in turn are generalizations of the chain-structured hidden Markov model (HMM) very familiar to bioinformatics researchers. CRF models were first proposed by Lafferty *et al.* in 2001 [31], and have since become popular with artificial intelligence researchers, who found them useful for modelling written language [32], and for automatic labeling and parsing of text [33]. Their unparalleled performance in this domain is due to their ability to integrate observational evidence with its context – unlike HMMs, CRFs are not hamstrung by strict conditional independence assumptions within the model. This means that they are uniquely positioned to incorporate the nonlocal fragmentation effects and global peptide characteristics discussed above into their probability estimates, and thereby take advantage of all the available information in an MS/MS fragment ion spectrum.

In the following three discussion sections, the use of the conditional random field in peptide fragmentation modeling is described. First, theoretical considerations motivating the use of the CRF in peptide fragmentation studies are discussed. The important differences and distinctive characteristics of the conditional random field model are explained by comparing it directly with the hidden Markov model, and by showing how it may be derived from these more traditional probabilistic sequence models. The tradeoffs made by the CRF framework to achieve its high degree of flexibility are then described, and the significance of these tradeoffs to fragmentation modeling in proteomics is discussed. Secondly, practical considerations of modeling peptide fragmentation spectra are discussed – due to the flexibility of the model, there are many ways that this can be accomplished, and the amount of information available means that computational models can rapidly grow unmanageably large. The focus of this second section is on effectively moving the data needed from the spectrum into the model. An approach for extracting relevant fragmentation information from the spectrum using the singular value decomposition (SVD) (chapter 3) is introduced, and paired with a feature selection algorithm based on feature frequencies and feature likelihood estimates that maximizes the information in the CRF model while ensuring computational tractability. In the third section, the interpretation of the output from the CRF model is discussed. To demonstrate the power of this modeling framework, a specific peptide fragmentation hypothesis is tested – that by comparing the relative intensities of fragment ions sharing a common bond breakage, specific information on the peptide sequence may

be obtained.

4.2 Methods

4.2.1 Data Assembly

A total of 100000 peptide spectra were assembled from a laboratory data repository. The sequences corresponding to the spectra were obtained using Mascot, and only the highest scoring matches were retained (Mascot scores ≥ 55), which produced a set of 12000 spectrum/sequence pairs.

4.2.2 Fragmentation Window Extraction

The sequences of each spectrum/sequence pair were used to identify the regions of each spectrum containing primary ions resulting from the breakage of the peptide backbone at each possible position – this study considered the a- and b-type N-terminal ions, and the y-type C-terminal ion. A single 6000-element vector was constructed to represent each putative fragmentation by extracting 100 Th-wide windows centred on each primary ion, interpolating each window to 2000 data points, and concatenating them. This produced a data matrix representing fragmentation information for each spectrum/sequence pair, with dimension 6000 $\times$ (peptide length - 1).

4.2.3 Feature Extraction and Data Compression

The singular value decomposition was then applied to the matrix of the fragmentation vectors, in order to construct a set of principal component variables that optimally represent the recurring features that separate the fragmentation features from each other in feature-space. Because the SVD procedure is extremely memory intensive, a randomly chosen subset of 13000 fragmentation data vectors were chosen to construct the pre-SVD data matrix. These vectors were chosen by randomly selecting a peptide, then randomly selecting one of its corresponding fragmentation data vectors.

After the SVD procedure was carried out (see chapter 3 for a detailed explanation), the diagonal S-matrix was examined to establish the number of principal component variables to retain (out of 6000 total) for an optimal, compressed data representation. Based on the inflection of the plotted S-matrix values, this principal component variable threshold was set at 400. Retaining this many principal component variables retained the vast majority of the information contained in the fragmentation vectors, while

achieving a data compression ratio of 15.0 (6000 ÷ 400). This procedure compressed the original peptide data matrix of size 6000 x (peptide length - 1) to a smaller data matrix of size 400 × (peptide length - 1).

4.2.4 CRF Model training

The 12000 principal component variable vectors and their corresponding sequences were then converted to binary feature functions and used as input data to a conditional random field model, which was trained using iterative maximum likelihood training. Approximately 10 iterations were required for training convergence, which was defined as the point where likelihood values reached a plateau, and did not increase with further training iterations. The final model weights were then examined for evidence of sequence-specific patterns, and were used to probabilistically match sequences from a test set to their corresponding fragmentation spectra.

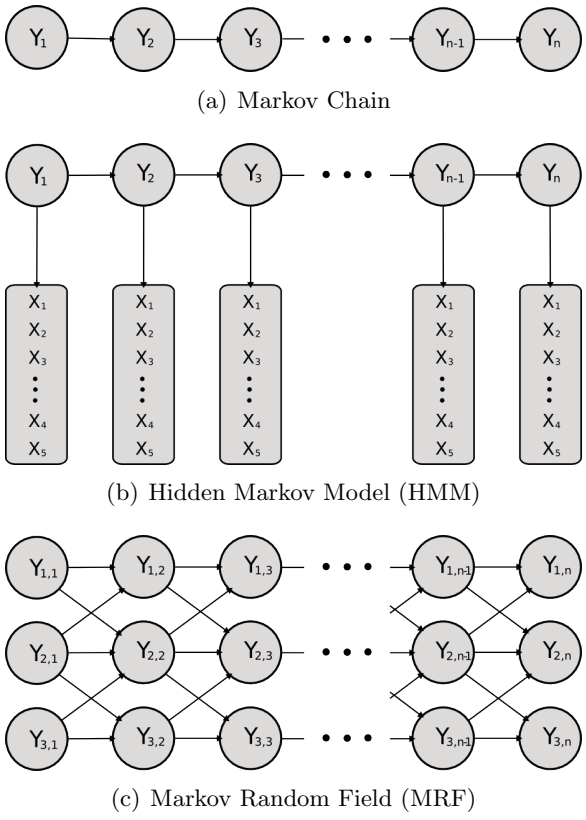
4.3 Results and Discussion

4.3.1 Sequence Models in Biological Research

Probabilistic sequence models have played an important role in the history of machine learning and artificial intelligence. More recently they have been paired with rapid DNA sequencing technologies, and have enabled the rise and success of genome-based bioinformatics research. One of the most popular sequence models in this domain is the Markov model, and the related hidden Markov model (HMM). In genomics research, these models have been used for the elucidation of common sequence motifs from stretches of related DNA [34], for the determination of consensus sequences [35], and for gene-finding [36]. Promising research in proteomics has also made use of these models in unique de novo sequencing applications [37], and in various other situations, including basic fragmentation modeling [38].

Though research into Markov models and the other specific classes of sequence models began separately and in isolation, they are unified within the general-purpose graphical modeling framework, which describes general statistical methods and algorithms for the integrated modeling of probabilistically interrelated variables. Operating within this framework, a mathematical graph is constructed of the situation being modeled where the nodes of the graph represent observations made or unknown quantities to be estimated, and edges between these nodes explicitly map out the probabilistic dependencies relating them. These models include very simple,

one-dimensional Markov models, such as those used for peptide and DNA sequence modeling (figure 4.1(a)), more complex one-dimensional models



(a) Markov Chain

(b) Hidden Markov Model (HMM)

(c) Markov Random Field (MRF)

Figure 4.1: Related probabilistic sequence models formulated within the graphical model framework.

that can deal with observational uncertainty (most hidden Markov models) (figure 4.1(b)), and complex two-dimensional graphical structures such as the Markov random field (MRF) (figure 4.1(c)), which can be used for image segmentation and de-noising [39, 40]. Thanks to the explicit graphical representation of the model, inference algorithms based on dynamic programming may be implemented as graph-traversal algorithms – well known algorithms implemented in this manner include the forward/backward algorithm used in linear chain-structured models, and loopy belief propagation, commonly used in densely connected random field models.

Modeling Peptide Fragmentation With Hidden Markov Models

A straightforward approach to modelling peptide fragmentation might seek to employ a hidden Markov model due to their ubiquity in bioinformatics, their inherent simplicity, and the computational efficiency of their inference and interpretation algorithms. A digression on the mathematics of the HMM will highlight the specific shortcomings of this family of sequence models, and will motivate the derivation of the conditional random field.

An HMM models a sequence of observations $\mathbf{X} = \{x_1, x_2, ..., x_n\}$, assuming that these observations are associated with an underlying sequence $\mathbf{Y} = \{y_1, y_1, ..., y_n\}$, with each state symbol y drawn from a finite set of possible state symbols. To model the fragmentation of a peptide during MS/MS analysis, y_1 will correspond to the first amino acid pair between which a fragmentation occurs, and y_n will correspond to the last such pair – this means that the peptide will have an overall length of $n + 1$. The corresponding x variables represent the spectrum information associated with each fragmentation, and are found by using the associated sequence to extract fragmentation peaks from the relevant areas of the spectrum.

The hidden Markov model approach constructs a model of the joint probability over sequences and observations, $p(\mathbf{y}, \mathbf{x})$. To make this possible (i.e. computationally tractable) two probabilistic independence assumptions are made. The first is that the probability of each state, y_t, depends only its previous neighbour, y_{t-1}; given this state, it is independent of the remainder of the states: y_{t-2}, y_{t-3}, ..., y_1. This is a trademark characteristic of Markov chains. The second assumption is that each set of observations, x_t, depends only on the identity of the directly associated label state y_t. Making use of these assumptions, the joint probability of a given peptide sequence, $\mathbf{y}$, and the corresponding set of fragmentation peaks, $\mathbf{x}$, can be compactly formulated:

$$p(\mathbf{y}, \mathbf{x}) = \prod_{n=1}^{N} p(y_n|y_{n-1})p(x_n|y_n) \tag{4.1}$$

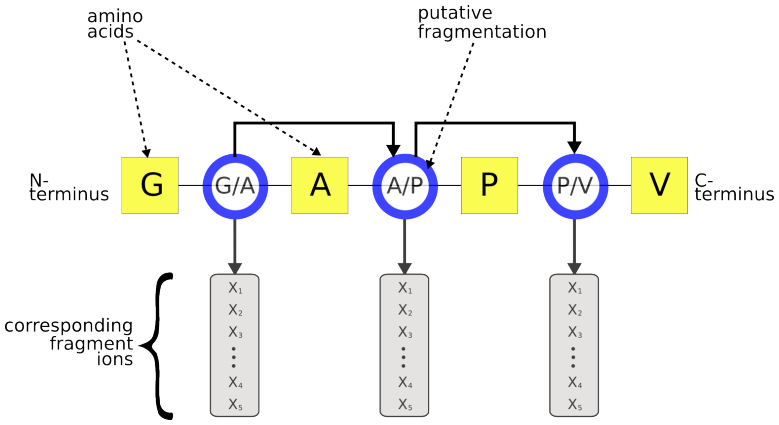
Given this probabilistic structure, the Baum-Welch algorithm can be used to estimate the probabilities involved, the forward algorithm can be used to calculate the probability of a given sequence and associated set of observations, and the viterbi decoding algorithm can be used to generate the most likely sequence given a set of observations having no associated label sequence. These approaches have been widely applied in bioinformatics, and several high-quality implementations of the algorithms are publicly available.

Faced with the task of modeling peptide fragmentation, chain structured sequence models like the HMM seem, initially, like ideal candidates. Pep-

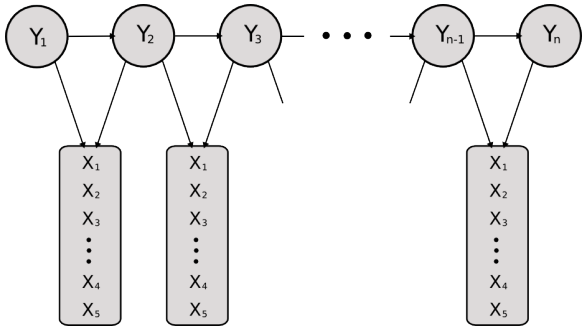
tides are linear chain structures, and a fragmentation event occurring between any two amino acids in the chain produces a set of observations, in this case a set of fragment ion peaks, specific to that fragmentation. These may be easily extracted from the associated spectrum, and be used for computing model probabilities using an HMM structure, like the example shown in figure 4.2(a). However, serious problems crop up when formulating an HMM for fragmentation modeling, stemming from the independence assumptions critical for the tractability of the model.

Firstly, the assumption that $p(y_t)$ may only depend on $p(x_t)$ forces the use of amino acid pairs as the labels, instead of individual amino acids. This greatly increases the cardinality of the label set; HMM training scales very poorly (quadratically) with increasing label set size. It would be optimal if the individual amino acid residues could be labeled (as shown in figure 4.2(b)), which would require labels from a set of size 22 ({start, A, G, D, ..., V, end}), rather than amino acid pairs (figure 4.2(a)), which requires a label set of size $22^2 = 484$ ({start, AA, AG, AD, ..., VV, end}). However, it is evident that the model with the smaller label set in figure 4.2(b) violates the specified independence assumptions, as the sequence states now depend on multiple sets of observations. In addition to requiring a great deal more computing time due to the expanded label set, additional algorithmic contrivances must also be used to enforce logical constraints on the pairwise labeling, and to properly compute the model probabilities. For example, it is illogical for a putative labeling to place a 'GV' label after an 'AD' label – adjacent amino acids in neighbouring label pairs must correspond.

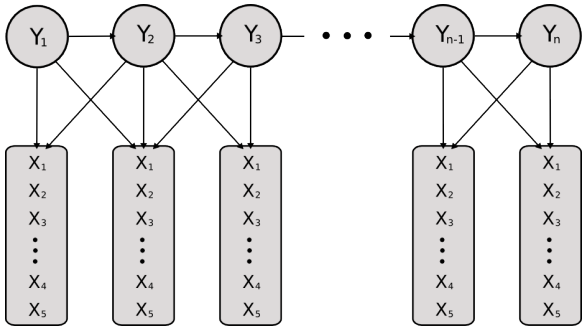
Secondly, and more importantly, the independence assumptions preclude the use of nonlocal fragmentation information readily available in the fragment ion spectrum, and the use of available global peptide information. As discussed, global peptide characteristics such as size, hydrophobicity, charge state, and electronegativity have been shown to influence peptide fragmentation, and sequence-specific characteristics, such as a stretch of basic residues, may affect the subsequent fragmentation in a nonlocal manner. Taking nonlocal fragmentation effects into account violates the independence assumptions in the same manner as modeling single amino acid residues, as shown graphically in figure 4.2(c). The global peptide information (mass, hydrophobicity, etc) referred to above cannot be incorporated into the vector of local observations because it affects the fragmentation distribution directly. Incorporating global information in this manner introduces dependencies within the observation vector $\mathbf{x}$ itself. The model of $p(\mathbf{x})$ constructed by the HMM requires that all such dependencies be specified explicitly, and assumes otherwise that all data is independent and identically distributed.



(a) Fragmentation Model HMM



(b) Single Amino Acid Label HMM



(c) Nonlocal Fragmentation HMM

Figure 4.2: An HMM applied as a fragmentation model (a). The HMM shows evident Markov independence violation when attempting to use single amino acid labels (b), or nonlocal fragmentation information (c).

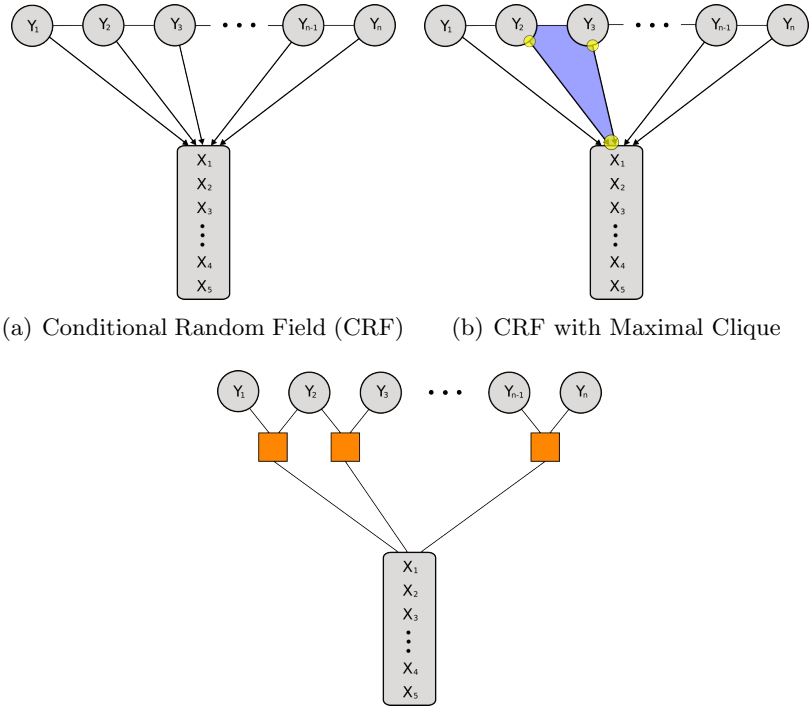
Available global peptide characteristics must therefore be incorporated into the label set used. For example, if peptide mass is discretized into three regions for modeling (low, med, high), then each amino acid pair label will need to be augmented with mass information. Thus, the label 'G-A' will give rise to 'G-A-lowMass', 'G-A-medMass', and 'G-A-highMass'. This further expands the label set cardinality, and quickly increases the algorithmic complexity to an untenable degree. These limitations are by no means proof that an HMM-based fragmentation model could not be useful. However, it is clear that scaling them beyond a certain level of complexity is difficult, or impossible.

Modeling Peptide Fragmentation With Conditional Random Fields

Conditional random fields are conditionally trained variants of the Markov random fields (MRFs) mentioned above, and are also formulated within the graphical model framework. However, CRFs have a degree of inherent flexibility which allows them to tractably incorporate non-local observations, and global peptide information. They are also robust against observational dependencies – when two or more measured variables in the vector of observations co-vary to some degree.

This ability results from the probability distribution modeled – as discussed above, the HMM models the joint probability, $p(\mathbf{y}, \mathbf{x})$, where $\mathbf{y}$ is a peptide sequence, and $\mathbf{x}$ is a matrix of corresponding fragmentation data. Nonlocal fragment information and global peptide information cannot be tractably incorporated into such a model due to Markov independence assumptions. In contrast, the CRF models the conditional distribution – $p(\mathbf{y}|\mathbf{x})$, with sequence probabilities normalized over entire peptides, rather than at each individual fragmentation. This allows the Markov independence assumptions to be relaxed, so that all data available can be used for modeling the probability of any of the individual fragmentations in the peptide. The conditional nature of the probability also means that no independent model of the observations alone, $p(x)$, need be constructed, permitting arbitrary dependencies to exist between observed variables (graphical representation shown in figure 4.3(a)).

The negative consequence of this is that the CRF model is not generative – without an explicit model of $p(x)$, the individual probabilities of the observations may not be obtained, so the probability of a given fragmentation feature occurring given a particular sequence may not be determined. However, probabilities of candidate sequences may be probabilistically matched



(a) Conditional Random Field (CRF) (b) CRF with Maximal Clique

(c) CRF as Factor Graph

Figure 4.3: The CRF model (a) is free of independence assumptions. The probability distributions, $p(\mathbf{y}|\mathbf{x})$, are maintained over the maximal cliques of the CRF graph structure (b), which can be represented as a factor graph (c).

to spectra, and the most likely sequence for a given spectrum may be generated.

This is the inherent tradeoff that separates the HMM model described above from the CRF model. The HMM model gains generative ability at the expense of model flexibility. It automatically learns a full model of the observations only – $p(\mathbf{x})$ – and can be used to generate "synthetic" data by predicting fragmentation patterns given an input sequence. However, it is only capable of learning from data that conforms to a strict set of independence assumptions. In contrast, the CRF model cannot generate synthetic data, and can only make probabilistic predictions from spectra to sequences, and not the reverse. However, it is capable of learning from complex and highly interdependent data. For taking advantage of neglected fragment ion intensity distribution information, a model of $p(\mathbf{x})$ would be of little use, but the data is highly interdependent by the nature of peptide fragmentation. It is clear that the CRF represents the better tradeoff in this case.

Maximum Entropy and the Exponential Family

The CRF model is an exponential family sequence model motivated by the principle of maximum entropy modeling. The notation in the following section is consistent with previous work on CRFs in language modeling by Hanna Wallach [41], and Sutton *et al* [42], and is based on the formulations presented in these excellent works, and the references therein.

The principle of maximum entropy modeling may be usefully paraphrased in the context of sequence modeling as follows: in the face of limitations in available model data (uncertain quality of training data, and limited training data), and incomplete information, the only probabilistically valid model is that which retains the highest information entropy over the modeled variables. This is a sensible and logical modeling assumption in the face of uncertainty. However, a particular mathematical structure is needed to properly develop a maximum entropy formulation. In the graphical model framework, the exponential family model of $p(\mathbf{y}|\mathbf{x})$, can be parameterized as a factor graph, where observations about the data are expressed as a set of feature functions operating over the maximal cliques of the graph (fig 4.3(b) shows a maximal clique) – the factor graph structure is shown in figure 4.3(c). This factor graph representation is generally applicable to graphical models, including the HMM [43], but is a particularly convenient representation for the CRF, as it yields a compact maximum entropy solution under

constrained optimization:

$$p(\mathbf{y}|\mathbf{x},\lambda) = \frac{1}{Z(\mathbf{x})}\exp(\sum_j \lambda_j F_j(\mathbf{y},\mathbf{x})), \tag{4.2}$$

where

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left\{\sum_{k=1}^{K} \lambda_k F_k(\mathbf{y},\mathbf{x})\right\} \tag{4.3}$$

In this formulation $F_j(\mathbf{y},\mathbf{x})$ is the sum of weighted "feature function" j over the sequence observed, $\mathbf{y}$, with $Z(\mathbf{x})$ serving to normalize the exponential sum of weighted feature functions into a proper probability. Note that normalization occurs over entire peptide sequences, $\mathbf{y}$

Using this parameterization, a particular feature function, F_j, represents one observation (feature), summed over the length of a peptide sequence:

$$F_j(\mathbf{y},\mathbf{x}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \mathbf{x}, i), \tag{4.4}$$

where the feature function f_j returns a positive, real value expressing some information about the observation $\mathbf{x}$ in one maximal clique – in other words at each position n in the sequence, in the context of a fragmentation occurring between amino acid residues y_{i-1} and y_i. For instance, one possible feature might be:

$$g(\mathbf{x},i) = \begin{cases} 1 & \text{if peak at m/z = i between 100 and 102 intensity counts} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

This is a general feature representing a single potential "truth" about the data. To incorporate this feature function into the conditional CRF probability shown in equation 4.2, this general feature needs to be expanded to reflect which amino acid pair (or fragmentation) it is associated with. This is done by conjugation with another indicator function – the following is an example of a conjugated, sequence-specific feature function that is specific to Q/G fragmentations:

$$f_j(y_{i-1}, y_i, \mathbf{x}, i) = \begin{cases} g(\mathbf{x},i) & \text{if residue at } y_{i-1} = Q, \text{ and residue at } y_i = G \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

The Meaning of the CRF Model

The factor-graph parameterization of the CRF using feature functions is a natural fit for sequence modeling. Each feature function expresses some constraint regarding the relationship between peptide sequences and the associated fragmentation patterns, and as the model is trained on labeled data, the parameters λ are iteratively adjusted to reflect how useful the corresponding feature is in relating sequences to observations. For a given MS/MS spectrum (set of observations), the correct sequence is likely to have satisfied more of the higher-weighted constraints than an incorrect sequence, giving the correct sequence a relatively higher output probability. In the CRF these probabilities are normalized over entire sequences rather

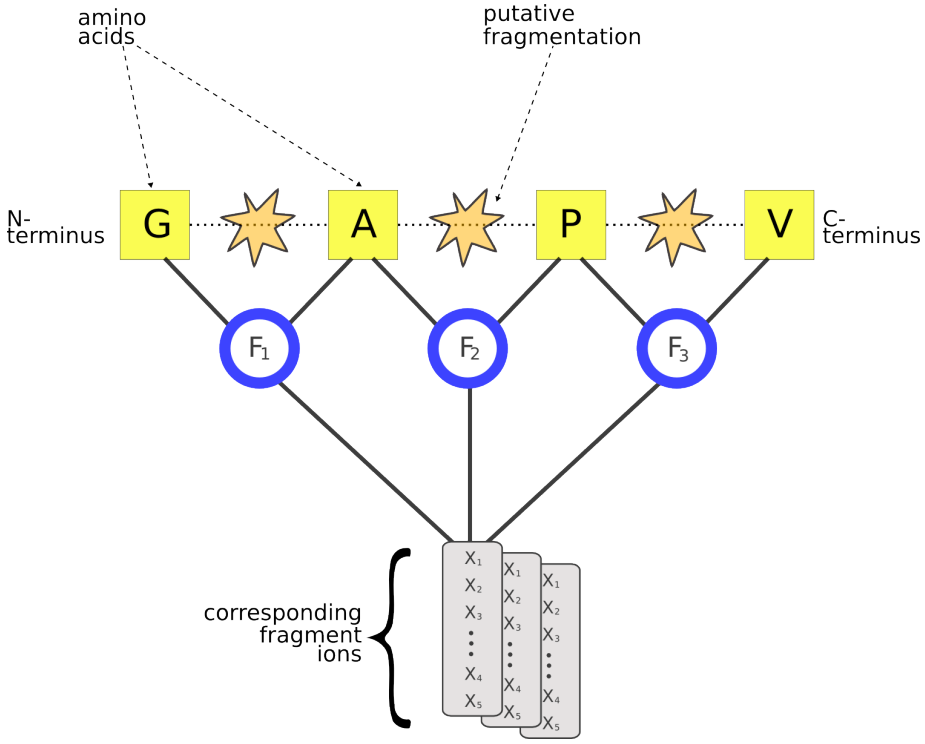


Figure 4.4: A CRF model applied to fragmentation modeling.

than individual sequence positions (as in the HMM), which admits the use of nonlocal fragmentation information, and global peptide information. Figure 4.4 shows a CRF in factor graph format, parameterized for use in modeling

peptide fragmentation.

Training the Model

With the CRF fully parameterized as described above, training on sequenced MS/MS spectra may now proceed. To estimate the optimal feature weights for a particular set of training data, maximum likelihood training is used, whereby the logarithm of the likelihood, $\log p(\{\mathbf{y}^{(k)}\}|\{\mathbf{x}^{(k)}\}, \lambda)$ is maximized as a function of λ. For the CRF, the log-likelihood is formulated as follows:

$$\mathcal{L}(\lambda) = \sum_k \left[\log \frac{1}{Z(\mathbf{x^{(k)}})} + \sum_j \lambda_j F_j(\mathbf{y}^{(k)}, \mathbf{x}^{(k)}) \right] \tag{4.7}$$

To avoid overfitting the model on the training data, a definite possibility for complex models trained on limited data, the likelihood above is regularized by subtracting a penalty term based on the Euclidean norm of the λ vector:

$$\mathcal{L}(\lambda) = \sum_k \left[\log \frac{1}{Z(\mathbf{x^{(k)}})} + \sum_j \lambda_j F_j(\mathbf{y}^{(k)}, \mathbf{x}^{(k)}) \right] - \sum_j \frac{\lambda_j^2}{2\sigma^2} \tag{4.8}$$

This effectively keeps the weights from growing too large during iterative training, producing a smoother, simpler model. From a Bayesian perspective, this is equivalent to performing maximum *a posteriori* estimation of the weights λ, with a spherical gaussian prior on λ. The variance of the prior, σ^2, is a free hyperparameter that must be assigned a value by hand, or must be estimated via cross validation. Previous work has shown that the CRF framework is robust against large changes (up to an order of magnitude) in the value of σ^2 in various language modeling applications [42], which was found to be true in this study as well. As for exponential family models in general, the function $\mathcal{L}(\lambda)$ is strictly concave when regularized, meaning that it has a single global optimum, and no local optima. It can therefore be maximized as a function of λ using any form of gradient descent optimization, with a gradient vector assembled from partial derivatives of the likelihood function with respect to each λ_j,

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda_j} = E_{\tilde{p}(\mathbf{Y},\mathbf{X})}[F_j(\mathbf{Y}, \mathbf{X})] - \sum_k E_{p(\mathbf{Y}|\mathbf{x}^{(k)},\lambda)}[F_j(\mathbf{Y}, \mathbf{x}^{(k)})] - \sum_j \frac{\lambda_j}{\sigma^2} \tag{4.9}$$

Setting the gradient to zero does not yield a closed-form solution for λ, so an iterative gradient descent algorithm must be employed. The limited-memory BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm [44] (L-BFGS)

was employed here, as it has shown good performance in other applications of the CRF. Other gradient descent methods are definitely applicable, and may contribute considerably to the overall modeling approach, particularly methods such as stochastic meta-descent [45], where the batch size is configurable, permitting on-line use of the algorithm, which allows the model to incorporate new data, without needing re-training on the current data.

With the iterative optimization of the weights λ carried out by off-the-shelf gradient-based methods, the remaining implementation concern is the efficient calculation of both the likelihood value and the full gradient vector at each step in the optimization. Calculation of the gradient vector requires the evaluation of the two expectations in equation 4.9. The leftmost of these is the expectation of the feature function F_j under the empirical distribution of the training data. This basically requires enumerating the occurrences of the feature over the training data set – a procedure described with precision by Sutton *et al.* [42]. The rightmost expectation is the expectation of the feature function F_j under the model distribution, and can be computed efficiently by dynamic programming using the forward/backward algorithm. The value of the normalizer $Z(\mathbf{x})$ is required for each sequence in the training data during the likelihood calculation, and computing this quantity – a summation over all possible sequence configurations – is the most computationally intensive step. Fortunately, dynamic programming can again improve the asymptotic complexity of the algorithm by eliminating the need to sum over an exponential number of sequence configurations. In her lucid and thorough treatment of basic CRF mathematics, Wallach [41] describes a matrix-based implementation of the forward/backward algorithm well suited to the CRF, and shows how closed semiring theorems from abstract algebra can be used to compute the normalizer $Z(\mathbf{x})$ efficiently, using matrix multiplication.

Though CRF models may be richer and more descriptive than their generative peers for a given amount of computation, practically useful fragmentation models rapidly grow very large, and in practice an upper limit is reached beyond which the memory requirements become unmanageable. The following sections address key points in managing this model complexity, in the context of constructing a useful CRF model of peptide fragmentation.

4.3.2 Fragmentation Modeling with CRFs: Practical Considerations

Feature Functions for MS/MS data

As discussed above, the CRF fragmentation model of a single peptide can be understood as a globally normalized linear combination of feature functions, each operating in maximal cliques of the chain-structured graph representing the possible fragmentation points of the peptide. These feature functions may be real-valued, taking on any nonzero numerical value. Though this is unproblematic as far as the mathematics of the CRF are concerned, real valued features introduce logical problems to the modeling of fragment ion intensities in a mass spectrum. A single real value cannot adequately represent the information available in a single MS/MS feature, such as a peak. The fact that a particular MS/MS fragment ion peak is unusually large, unusually small, or of medium size may confer equally important meaning about fragmentation in different situations that the model should capture. However, the conditional probability modeled by the CRF is a normalized sum over weighted feature functions (equation 4.2), so any positively weighted feature will contribute more to the overall probability if the feature function returns a larger value. Conversely a negatively weighted feature function will contribute increasingly the smaller the feature is. Multiple feature functions per MS/MS feature are required to adequately represent situations where the average peak intensity distribution, taken over the entire training data set, is multi-modal for a specific fragmentation. Using multiple real-valued features introduces uncertainty regarding their type and parameterization – the real-value may be a raw peak intensity, or any transformation thereof, and it is not immediately apparent which of these will provide maximum discriminative power to the CRF.

To overcome this problem, the CRF used in this study uses binary indicator functions that partition the input space into discrete, narrow ranges. This parameterization retains the ability to model multi-modal fragment ion intensity distributions, but reduces the risk of introducing representational errors due to poor feature function selection. Though potentially less accurate than the optimally representative set of continuous feature functions (due to the discretization process), this reduces the parameterization problem to the finding of an optimal discretization of the input space. This discretization may then be optimized using a technique such as cross-validation, and may also be automated.

An additional benefit of this parameterization is that it simplifies the

interpretation of the final, trained model. Using binary indicator functions gives each discretized region an equivalent amount of probabilistic "weight", so the weight value (λ_j) associated with each discretized unit of the input space (binary feature j) may be then interpreted as the "relative contribution" of that feature to the final model likelihood. Model interpretation will be discussed further in later sections.

Data Compression and Conditioning

Using the discretization process described above enables the modeling of single MS/MS features local to individual fragmentations, such as the intensity of an associated fragment ion peak, or global peptide features such as the peptide mass. Though an optimal discretization needs to be found for each feature when using this approach, the complexity of selecting an optimal set of continuous feature functions is avoided, along with the possibility of introducing systematic model bias due to a sub-optimal parameterization.

However, there are two other major problems involved in adequately representing the data for optimal CRF discrimination that involve the set of available features as a whole, rather than individually. The first of these involves feature importance. MS/MS spectra can be very complex, and contain a tremendous amount of information, particularly spectra originating from high-resolution, ion-cyclotron instruments. Though many potentially useful methods exist for extracting information relevant to fragmentation modeling, unless each feature to be included in the model is selected by hand, it is likely that a great deal of information irrelevant to peptide fragmentation will be included in the model. The second problem is information redundancy, which is a particularly pertinent issue when dealing with fragment ion intensities. As discussed above, well established theory has suggested that sequence- and peptide-specific effects may alter the distribution of fragment ion intensity, meaning that the relative intensities of the fragment ion peaks in the MS/MS spectrum are expected to shift relative to each other. However, relative intensity shifts in fragment ion intensity also occur in the absence of sequence-specific effects, as individual fragment ion peaks intensities are not independent of each other. For instance, if a particular peptide fragmentation event creates an unusually intense primary fragment ion peak (such as an a-ion), it would be reasonable to assume that the associated secondary fragment ion peaks (a-NH_3, a-H_2O) might also be unusually intense. Other N-terminal-containing primary fragment ions might also have elevated intensities, such as the b-ion. Though the CRF does not directly model the observation probabilities $p(\mathbf{x})$, and is therefore

capable of dealing with probabilistic interdependencies in the input observation vectors $\mathbf{x}$, learning these associations that have nothing to do with sequence-specific effects is a waste of modeling effort.

To deal with these problems of MS/MS feature importance and feature independence simultaneously, a data transformation procedure based on the singular value decomposition was used (described in chapter 3). This technique has proved itself useful in many areas of genomic analysis which also deal with highly correlated model input variables, and is a particularly popular adjunct to microarray analysis [46]. It has also been used in the automated clustering of MS/MS spectra [47], along with the other transform-based tools such as wavelet decomposition. In short, this technique constructs linear combinations of the input variables (in this case, the fragment ion intensities) so that differences between the spectra in the dataset may be explained using as few variables as possible. This procedure automatically excludes features that have little explanatory power, and removes linear correlations between input variables. Because the SVD decomposition is carried out using all the spectra to be modeled (or a randomized subset thereof) the only linear dependencies removed are not sequence-specific. The SVD analysis shows a great deal of promise in mass spectrometric modeling, in MS/MS feature extraction, and in data visualization. The reader is referred to chapter 3 for more details.

Features Used In This Study

Rather than explicitly specifying the full set of fragment ions to include in the CRF model, a more general data extraction procedure was employed, whereby the sequences of the dataset peptides were used to extract windows in which important fragment ions were likely to occur. These windows were 100 Th in width (chosen to be conservative), and were centred on each a-, b-, and y- ion of each putative peptide backbone cleavage. Figure 4.5 demonstrates the windowed extraction procedure used to gather the vector of fragmentation pattern information corresponding to each putative peptide backbone cleavage. The MS/MS data was extracted from each window using a conservative interpolation, transforming each windowed region of the spectrum into a vector of 2000 intensity measurements spanning the 100 Th. The three vectors common to a fragmentation were then concatenated into a vector of 6000 intensity variables, and projected into an optimal linear basis constructed using the SVD , after individual normalization of each variable using means and standard deviations calculated over the whole training dataset. This yielded a new vector containing 400 variables, each of which

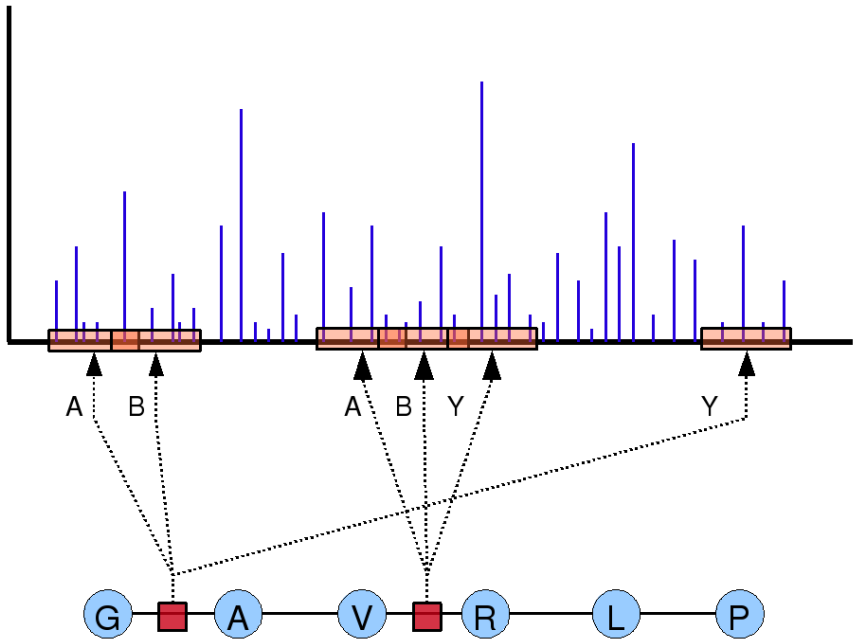
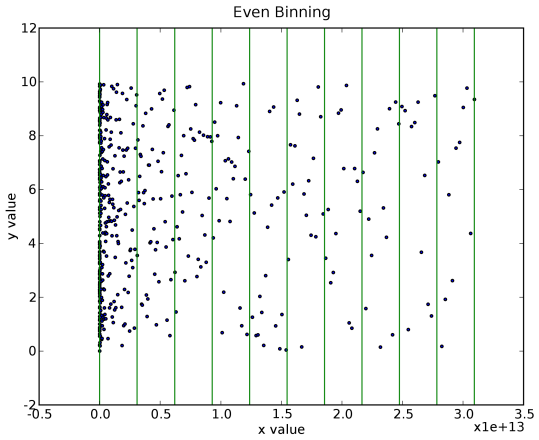


Figure 4.5: A schematic of the sequence-based data extraction procedure. The sequence is used to find 100 Th windows centered on the a-, b-, and y-ions of each putative fragmentation. The windows are then concatenated and vectorized for use in the model.

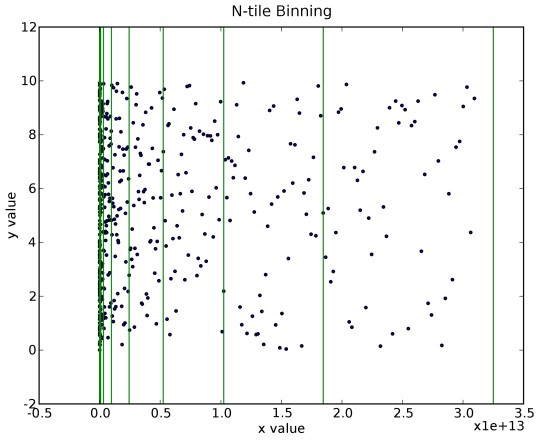
was subjected to discretization as described previously. Though many discretizations were studied, including linear binning, logarithmic binning, and quadratic binning, an "N-tile" binning approach proved the most successful – this approach involved positioning the discretizing bin boundaries using the training dataset such that each bin had an equal number of datapoints (figure 4.6). Optimal model accuracy was observed with $N = 17$. Over the 400 linearly independent variables, this produced a total of 5600 binary features per amino acid pair modelled, for a total of 2710400 features in the entire model. This calculation assumes an amino acid alphabet of size 22, as "start" and "end" amino acids (with distinct, artificial feature vectors) were added to the peptide sequences modeled in this study. This is not essential, but is useful in the calculation of the normalizer. Scatterplots of the first four linearly independent variables of the training set data, shown in figure 4.7, reveal areas of extremely variable data density. Feature functions created using linear binning yielded an extremely sparse model, with most of the points falling into only a fraction of the available bins, and gave very poor recognition performance in testing (data not shown). Feature functions created using N-tile bins were much more successful – their performance is discussed below.

4.3.3 Conjugate Feature Construction

The features and feature functions presented so far consist of linear combinations of individual fragmentation features derived from the SVD procedure. These are referred to as "atomic" features. However, the co-occurrence of two (or more) of these atomic features may have much stronger predictive significance for a particular fragmentation than the features do alone, simply by virtue of the added specificity. For example, atomic feature A might tend to have a high intensity in many fragmentations, including the 'DP' fragmentation, and atomic feature B might tend to have a low intensity in many fragmentations, likewise including the 'DP' fragmentation. However, the co-occurrence of a high intensity A and a low intensity B may be much more specific, occurring in very few other fragmentations apart from the 'DP'. Joining the atomic features already present in the model to represent such specificity creates "superfeatures", referred to as atomic feature conjunctions (or simply feature conjunctions). This is a general way to add expressive, non-linear representation and modeling capability to log-linear models such as the CRF. All features in the model thus far described are binary indicator functions, and these feature conjunctions may be thought of as indicator functions as well, returning 1 (True) if all the indicator func-

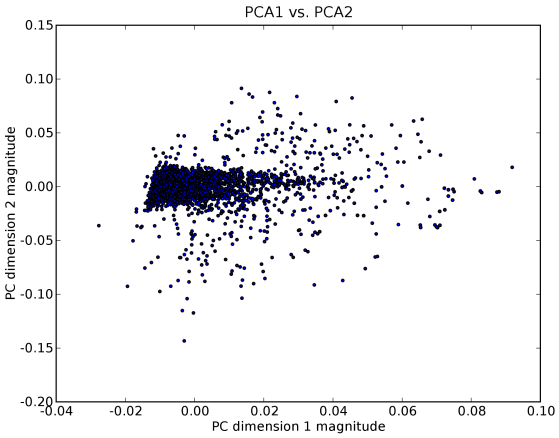


(a) Linear Discretizing Bins

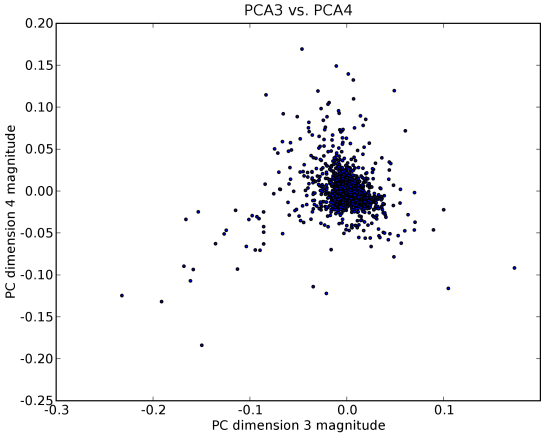


(b) N-tile Discretizing Bins

Figure 4.6: Discretization using even bins (a) creates a much sparser model. This is remedied using N-tile discretization (b), which positions the bin boundaries so that an equal number of points falls into each.



(a) PCA 1 vs PCA 2



(b) PCA 3 vs PCA 4

Figure 4.7: The scattering of the concatenated, vectorized MS/MS fragmentation windows in principal component space, with each point corresponding to a fragmentation data vector. The CRF attempts to correlate the position of the points in this coordinate space with the sequence involved in the fragmentation. The variations in density greatly favour the N-tile discretization approach.

tions of which it is composed return true, and 0 (False) otherwise. Formal specification of feature conjunctions follows a recursive pattern, as can be seen by construction of a binary feature 1 $f_{j,k}$ from atomic features as in equation 4.6:

$$f_{j,k}(y_{i-1}, y_i, \mathbf{x}, i) = \begin{cases} 1 & \text{if } f_j(y_{i-1}, y_i, \mathbf{x}, i) = 1 \text{ and } f_k(y_{i-1}, y_i, \mathbf{x}, i) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

Tertiary, quaternary, and higher order features can be constructed in this manner, representing increasingly complex patterns in the observed data. This type of conjugate feature allows the relative intensities of the various fragment ions to be modeled without the assumption of linearity. It also allows the effects of global peptide characteristics such as mass and electronegativity to be modeled – these peptide characteristics may not interact with ion intensity in a linear fashion, and thus should not be included in the SVD transformed data vectors. Any measurable quantity, properly discretized, can be incorporated into model features using this recursive feature conjugation approach.

Automated construction of feature conjugations from an atomic feature set produces many "nonsense" features that will never actually occur. For instance, it makes no sense to conjugate an atomic feature with itself, or two atomic features that represent different amino acid pairs, or two features that represent different discretization bins of the same fragment peak. These nonsense features are easily pruned from the model during construction – creating pairwise associations from 2000 atomic features yielded roughly 150000 potentially relevant conjugate features in practice. However, this is still an explosive enlargement in the model size; this becomes a major concern given that millions of relevant atomic features may be present originally, and that tertiary or even higher-order feature conjunctions may be desired.

An important goal of the current study was the development of a reliable and convenient method of assessing the relative utility of the model's constituent features, both atomic and conjugate, in order to keep the model's size in check, and to keep the time required for iterative model training within reasonable limits.

4.3.4 Feature Selection/Induction/Pruning.

The extreme flexibility of the CRFs feature functions is a blessing and a curse simultaneously. The feature functions allow for great flexibility in the

input used – any available information can be incorporated in the model, and put to use should it be probabilistically relevant. Unfortunately, covering the entire range of possible model inputs with feature functions requires an enormous number of such functions. In addition to slowing down the training of the overall model and making the construction of feature conjunctions an infeasible task, this effect causes relevant features to be increasingly diluted by meaningless features with no significance in the model. This may slow the convergence of the optimization routines used to adjust the feature function weight values during training, and may reduce the discriminative performance of the model overall. There is also evidence that large numbers of meaningless features may introduce overly high levels of noise to the gradient of the likelihood, causing the optimization algorithms to fail entirely [48].

A feature selection algorithm is the solution to this problem. Feature selection approaches are ubiquitous in machine learning research; as the name implies their purpose is to assess the quality of candidate features and to "select" the best performers for inclusion in the model to ensure its optimality. Alternatively, they may be employed to *remove* poorly performing features from a sub-optimal model in order to improve its performance. This problem has been tackled by researchers using CRFs in other domains – a particularly thorough treatment is given by Andrew MacCallum *et al* [49], in which a mean field approximation (normalizing over individual sequence positions, rather than whole sequences) is elegantly used to tractably estimate the increase in likelihood that would result from the addition of each candidate feature. However, this sophisticated feature selection approach was developed for CRF models of written language featuring much smaller state spaces and feature numbers than described here, and the additional computation needed for calculating the gain of each feature slowed the algorithm to an impractical degree.

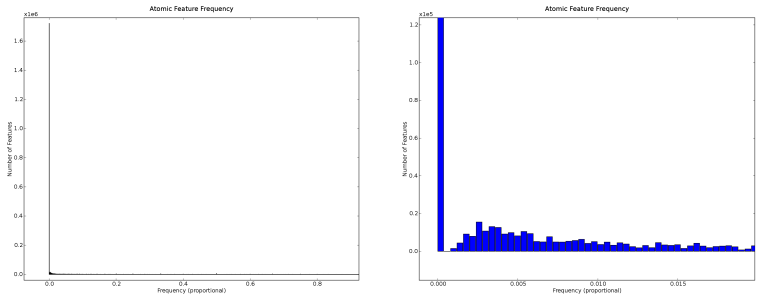
Due to these problems, an alternative feature selection approach was developed that is much less sophisticated, but much less computationally expensive. This approach has two selection phases. First, selection based on feature frequency discards putative features that never, or almost never occur for a given fragmentation. This statistic is calculated simply by observing how many times a given feature occurs, and dividing it by the number of times the associated amino acid combination occurs. As a histogram of feature intensity shows, many features occur very rarely, or not at all. Figure 4.8(a) shows a histogram of relative feature frequency, showing that zero-frequency features vastly outnumber the features that do occur. The removal of zero-frequency features (figure 4.8(c)) allows other frequency pat-

terns to emerge in the histogram. All features selected at this stage are then added to a CRF model which is iteratively optimized until convergence. For the second selection step, features receiving very low scores from the model are dropped – only features associated with the higher CRF weight values are retained in the model. After convergence, many features have small associated weight values. Once this weight value pruning has occurred, feature conjunctions can be constructed. For this step, only a suitably sized subset of the highest scoring features are used, in order to restrict the size of the model. After the creation of the conjugate features the process repeats – to be included in the model, the new conjugate features must pass the frequency selection, and must receive a sufficiently high weight during CRF training. This process repeats until sufficiently complex conjugate features have been constructed – this point can be selected based either on the complexity of features desired, or based on discriminative performance of the model, as assessed by testing.

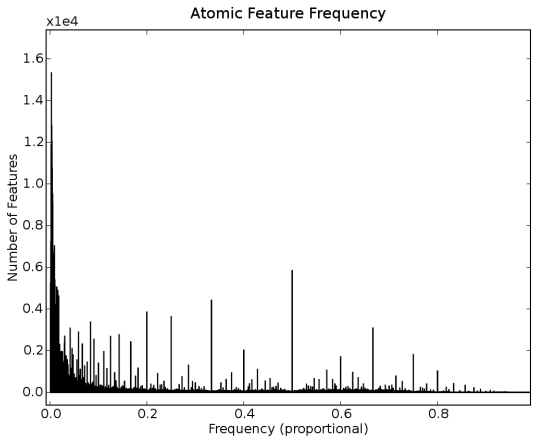
The feature selection approach described trades a certain amount of selectivity for speed when compared to gain-based selection. Indeed, while this approach can reliably identify and exclude features that are truly superfluous (features with 0% frequency, or that are assigned extremely low CRF weights), uncertainty arises in deciding which features to keep and which features to exclude. This is particularly problematic when considering higher order feature conjunctions – atomic features that are only marginally useful by these selection criteria may participate in very useful feature conjunctions, but be "accidentally" eliminated due to their poor atomic feature performance. Though a rigorous threshold determination using cross validation is possible in theory, for this study it was deemed sufficient to err on the side of caution, and lower the feature pruning thresholds, for both frequency and CRF weight, as far as possible while maintaining computational tractability.

Implementation: Parallel Likelihood and Gradient Computation

To use the features selected in the model, the weight associated with each must be optimized to maximize the overall model likelihood. To accomplish this, the equations for the likelihood (equation 4.7) and the gradient vector (equation 4.9) are iteratively solved, with the weights, λ optimized at each step using L-BFGS optimization. The CRF is an additively computed model – the likelihood is computed by summing the likelihood contribution of each feature function in the training data (eqn 4.7), and the gradients for each feature are likewise computed by summation (eqn 4.9). This makes



(a) Feature Frequencies (b) Feature Frequencies (Origin Zoom)



(c) Feature Frequencies (Non-Zero)

Figure 4.8: The feature space is very sparse. In the unfiltered histogram of feature frequency, the zero-frequency features dominate (a). A zoom into the origin region is also shown for comparison (b). Filtering zero-frequency features makes other frequency patterns more apparent (c).

the construction of a highly efficient algorithm to compute these quantities feasible, as it implies that a "divide and conquer" algorithmic approach can be used. The work required to compute the likelihood and the the gradients can be broken into reasonably-sized sections and performed in parallel.

Given the multi-core nature of modern computer server systems on the one hand, and the tightly networked systems of desktop computers found in a contemporary research lab on the other hand, the possibilities for parallelizing the task of CRF optimization in a standard proteomics laboratory become apparent. In the implementation described here, the additive computation of the likelihood and the gradients was broken up into parallel sections in two ways. The first way was a division of the of likelihood and gradient computing functions into separate computational threads; a single computer is thus able to run multiple instances of these functions in parallel on different input data, and is therefore able to fully utilize all of the available processors in a server, or the dual or quad-core desktop processors which are becoming ever more common.

The second method of parallelization is more complex, and involves wrapping the computation to be performed within a distributed computing framework – this allows entirely separate computers (slave nodes) to process independent sections of the input data, with the accumulation of the output and the L-BFGS weight optimization carried out on a central (master) node. The distribution of the input data to the slave computers and the centralization of the computed results takes place over the local network. Both parallelization strategies were implemented for this study, resulting in a distributed computing setup with slave nodes that were themselves multithreaded. The success of this type of threaded/distributed scalability made the CRF an excellent choice for working with extremely large sequence models in a contemporary proteomics laboratory environment, without access to dedicated supercomputing resources.

Implementation: Software Considerations

The sheer size of the full model, the computational complexity of the training and testing algorithms, and the volatility of the various optimization approaches, feature selection approaches, feature sets, and analytical demands combine to create a unique software engineering problem. The speed and efficiency of a low-level compiled programming language (such as fortran, C, or C++) are required in order to perform the needed computations in a timely manner, but the high-level, expressive characteristics of an interpreted language (python, perl, or ruby) are needed as well to enable

flexibility in algorithmic combinations, routine debugging, analysis, and to cope with rapidly changing interface and testing requirements. To effectively manage all these requirements simultaneously, a combined software development approach was taken. All interface, control, and data output and visualization aspects were coded in the Python language, while the algorithmic routines of the CRF (feature selection, likelihood and gradient calculations, and any calculation that involves iterating over the training data) were coded in C++, and compiled as python extension modules using the open-source SWIG interface generation tool. This combined approach allowed complicated and relatively static routines to be coded in a highly efficient, low-level, compiled language for optimal performance, which was then embedded in an expressive, flexible, and highly dynamic control framework.

This combined software development approach made the parallelization optimizations discussed above a great deal easier – the pthreads library for C++ was used for providing threading functionality to the likelihood and gradient calculations, and the Pyro distributed computing framework for Python was used to allow the CRF calculations to be distributed over slave nodes in the network. The pthreads library is cross-platform, being available on any system for which there is an ANSI-compliant C++ compiler, and the Pyro library runs on any system for which a python interpreter exists.

4.3.5 Model Output and Interpretation

As discussed above, the model likelihood of the CRF is expressed as a log-linear sum of weighted feature functions, all of which operate on the maximal cliques of the chain-structured conditional random field graph similar to that shown in figure 4.3(b). Multiple iterations of gradient optimization take place during training, during which the weights associated with each of the feature functions are adjusted to maximize the overall likelihood of the model. The output of the model, therefore, is this set of fully adjusted feature weights.

The weight assigned to an individual feature may be roughly interpreted as the "quality" of that feature. Features with higher weights have more relative influence over the overall model likelihood, and features with lower weights have relatively less influence. Since the features described in this study include the amino acid pair involved in the putative fragmentation, the weight may be interpreted as the "relative strength" of the association between the amino acid pair and the remainder of the information composing the feature function. This interpretation is made possible by the use of

the binary indicator functions described above (equation 4.6), as this parameterization gives each features an equivalent initial contribution to the likelihood.

It must be emphasized at this point that these trained weights are not probabilistic – examination of an individual feature weight cannot yield any information as to the actual probability that the corresponding feature will appear in a fragmentation spectrum. This results from the lack of generative ability in the CRF framework that is a result of modeling the conditional probability distribution ($P(y|x)$) instead of the joint probability distribution ($P(y, x)$), as described earlier. Given a fragmentation spectrum (set of observed features), the CRF can yield probabilistic answers regarding the possible sequences, but given a sequence, no probabilistic statements involving features may be made.

Regardless, comparing relative feature weights is sufficient for use in exploratory data analysis. Examining the trained set of weights using a directed approach – looking only at a specific amino acid combination, or at a particular peak location or intensity bin, will produce a well defined and limited distribution of weights for comparison. This may provide sufficient evidence to be directly useful in theoretical work, or may highlight spectrum areas, or fragmentation instances, that might benefit from further modeling. The isolation of these instances of interest using the CRF will likely reduce the size and complexity of the dataset to the point where limited but fully probabilistic generative models may be employed.

Example: Probabilistic Sequence Scoring Using the CRF

The most compelling use of the CRF model is to assign probabilities to candidate sequences given an MS/MS spectrum. This is accomplished using a variant of the forward/backward algorithm described by earlier CRF work [41, 42]. This is a particularly appealing use of the CRF for detailed, high-throughput proteomics research, as it allows a set of candidate sequences to be ranked according to the quality of the match between the sequence and it's fragmentation spectrum – this makes it an ideal accompaniment to contemporary database search and de novo sequencing approaches that do not consider fragment ion intensity distributions, and can be used to verify results obtained using these tools, or to rescore a list of potential sequences generated via error tolerant searches or sequencing runs.

The probabilistic matching of sequences with fragmentation spectra, for the purposes of this study, is mostly a proof of principle application. As the reader will have ascertained, the flexibility of the CRF framework per-

mits many possible parameterizations of MS/MS fragmentation data, and the approach and results presented here, though quite good, are likely still suboptimal. Feature conjugation was not used in this assessment, and the feature selection approach described above was simplified, and used only to remove zero-frequency features. This means that the only features included were principal component features (from the SVD procedure) consisting of linear combinations of fragment ion intensities – peptide mass, fragment mass, and other global characteristics were not included.

To assess the performance of the model in this regard, 3 test sets of peptides of length 6, 11, and 17 were randomly selected from the training set, and divided into groups of 10, 25, and 50 peptides. The within-group accuracy was assessed, whereby an MS/MS fragmentation spectrum was selected, and the probability, $p(\mathbf{y}|\mathbf{x})$, was calculated for each peptide sequence in the group. A success occurred when the correctly matching sequence was assigned the highest probability, a failure occurred when any of the other non-matching peptide sequences in the group was assigned the highest probability. The percent accuracy was then calculated by dividing the number of successful sequence-picking trials by the total number of such trials. The average percent accuracy is the average of this accuracy over all of the groups of peptides of the same size constructed from the test set. To test the sensitivity of the CRF model to differences in feature discretization, the testing procedure above was repeated with different versions of the model, where the feature space was discretized into different total numbers of bins. The results of this sequence selection are shown in table 4.1

As can be seen from the table, optimal performance (88.75% recognition) was seen when discretizing the feature space into 17 bins exactly. Recognition performance was poorer with fewer bins as the bins were wider, which in this case was a suboptimal representation of the data, conferring insufficient resolving power. Interestingly, recognition performance dropped drastically when adding bins as well. A bin number greater than 17 imposed *too many* distinct feature classes – due to the N-tile binning, this unnaturally forced some equivalent fragmentation vectors into different bins of the feature space, which destroyed model performance. Fortunately, the drop in performance is very steep, making the optimal N-tile discretization easy to find automatically using cross validation.

The length of the test peptides, and the number of test peptides in each test set also influenced the testing performance. Longer test peptide sequences stand a greater statistical chance of being markedly different from each other, and are more likely to feature a unique fragmentation signal that would set them apart from other sequences in the group – this effect

Table 4.1: The Sequence Association Accuracy of the Trained CRF Model, trained on 6000 sample peptides, using 400 PCA features (from SVD), divided into 7, 11, 14, or 21 N-tile regions each.

Number of Discretization Bins	Test Sequence Lengths	Test Partition Size	Avg. % Accuracy
7	12	10	**62.5**
		25	**56.0**
		50	**47.6**
11	12	10	**67.5**
		25	**56.0**
		50	**52.3**
14	6	10	**50.0**
		25	**48.0**
		50	**No Data**
	12	10	**70.0**
		25	**60.0**
		50	**57.14**
	17	10	**88.75**
		25	**81.6**
		50	**No Data**
21	12	10	**20.0**
		25	**16.0**
		50	**9.5**

is clearly visible when examining the model performance, as the recognition on a test set of peptides 17 residues long is more than 30% higher than the recognition achieved on a test set of peptides 6 residues long. The number of peptides in the test set has a similar effect. When assigning probabilities from a larger sequence pool, there is a greater chance of finding a sequence that randomly matches with a probability higher than the target sequence; changing the number of candidate sequences in the pool from 10 to 50 reduced the recognition accuracy by 10 to 15%.

Example: Exploratory Data Analysis Using the CRF

An example of the exploratory data analysis that may be done is provided in figure 4.9, where the top scoring 10% of features (identified by PCA feature number and discretized bin number) are plotted for fragmentations occurring between four different amino acid pairs. From this plot it is possible to see that the four combinations exhibit many similarities in their characteristic fragmentation features, because the patterns look quite similar. However, a closer inspection reveals key differences. Because these are the top scoring features for each amino acid combination in the trained model, these are the features that are indicative of sequence specific effects.

The top scoring features in figure 4.9 are PCA features extracted from the singular value decomposition (described in chapter 3). These features are linear combinations of MS/MS peak intensities – this makes an interpretation of the sequence specific fragmentation effects difficult, as each PCA variable ranges over the entire spectrum. For the purposes of exploratory data analysis, a better strategy would be to parameterize the CRF using direct peak intensities from the MS/MS spectrum. Although this may lower the accuracy of sequence selection, it would mean that a highly weighted feature would correspond to a single peak, rather than a weighted combination of all peaks, making the analysis of the weights more straightforward. However, the focus of this study was on maximizing the discriminative accuracy of the CRF to estimate possible performance. Discriminative accuracy was found to be the highest using SVD-derived features (data not shown), and as a result the direct peak intensity parameterization was only lightly explored in this study.

4.4 Conclusions and Future Directions

The application of statistical machine learning methods to mass spectrometry data is already a well established approach used in many applications within the analysis of proteomics data. It is clear, however, that they also show great future promise in more general purpose modeling approaches as well. The modeling of peptide fragmentation data is one area in proteomics research that is particularly well suited to applications of machine learning – current theory is unable to make predictions of sufficient detail to be practically useful in predicting sequence-specific effects on fragmentation peak intensities, yet data collected from these fragmentations are available in great abundance.

The conditional random field (CRF) represents an ideal sequence model

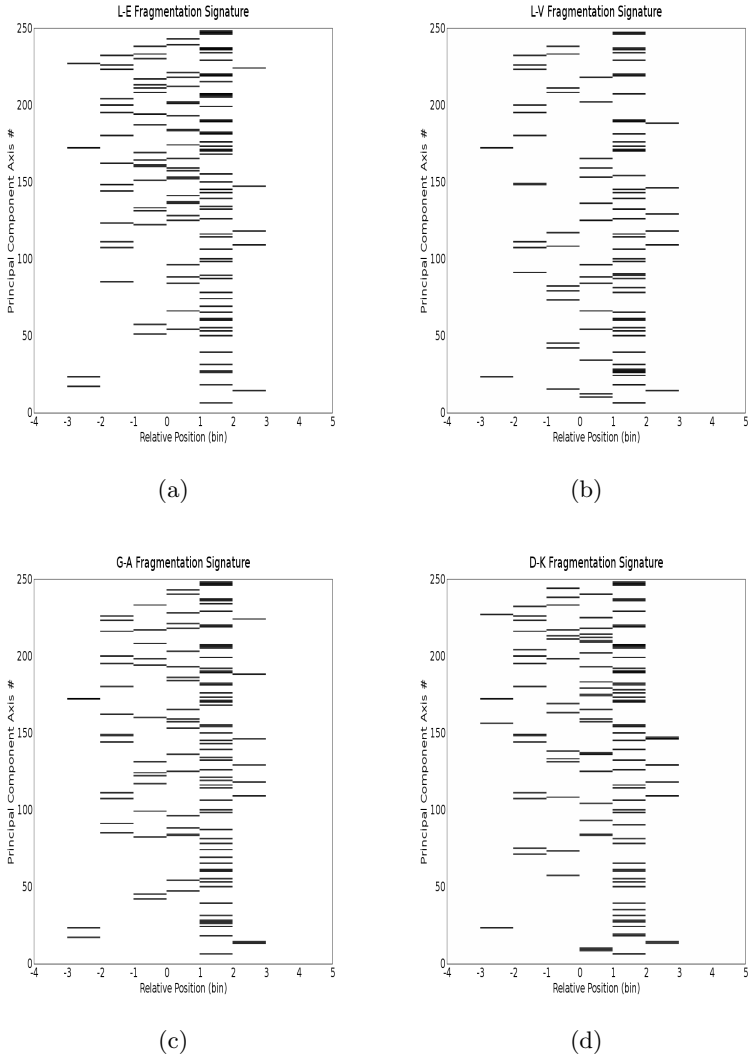


Figure 4.9: The top scoring 10% of features occurring in the fragmentation of four common amino acid pairs: a) L-E, b) L-V, c) G-A, d) D-K, plotted according to principal component feature number and bin number.

for use on peptide fragmentation data. As a large body of theoretical and empirical evidence has suggested, the intensity distribution of fragment ions produced from a specific peptide bond breakage is influenced by several factors; some are local to the breakage point, some are nonlocal, and some are global features of the entire peptide. Generative random field models, including the hidden Markov model (HMM) and the more general Markov random field (MRF), are restricted by Markov independence assumptions, and are therefore incapable of effectively representing these more complicated combinations of interdependent features. Adequate representation involves a higher degree of model connection (higher treewidth), and/or a multiplicative increase in the size of the alphabet size, both of which scale very poorly to large sizes.

The CRF is not encumbered by these independence assumptions. By modeling a peptide sequence using a single conditional probability instead of factoring it into joint distributions on each bond breakage (using a hidden Markov model or a finite state automaton), local, non-local, and global peptide information may be incorporated into the model and used efficiently, even if the data observations are highly correlated.

Modeling peptide fragmentation data with CRFs presents unique challenges, and several parameterization techniques are essential for model tractability and performance.

Representing numerically continuous features using a family of discrete, binary functions is a robust parameterization that requires no *a priori* assumptions about feature behaviour, and that is able to easily represent monotonically changing, unimodal, or multimodal feature densities. Giving each binary feature function equal return values (0 or 1), allows for easier interpretation of the model weights for exploratory analysis. The optimal method of discretization to use for feature discretization was found to be N-tile binning, where the boundaries of the N bins are located such that an equal number of data points fall into each. This avoids the sparsity issues that occur when applying a linear binning scheme to highly nonlinear data. This binning scheme is dataset dependent – the limits must be calculated by averaging over the entire training data set, but an optimal threshold is easily determined via cross validation.

As the size of the feature set needed can grow very rapidly, an inexpensive feature selection algorithm is essential for removing superfluous, redundant information. Though sophisticated gain-based feature selection algorithms have been applied to smaller-scale language modeling problems, they become a bottleneck in applications of this size. Instead, a feature selection approach relying on feature frequency, and on initial CRF-score, was demonstrated

and discussed.

Though the CRF is well suited to huge numbers of features, using as few as possible confers gains in speed, in terms of both iteration time (calculating the likelihood value and the gradient vector) and algorithmic convergence (irrelevant features add noise to the gradient, slowing convergence). Additionally, though the CRF is able to use correlated features, doing so when the correlations are known to be meaningless represents a waste of modeling effort. To maximize performance in this regard, the observed vector of fragmentation features was optimally compressed into a smaller set of variables. This is accomplished by projecting the observed data vectors into a linear basis constructed using the singular value decomposition (SVD). The constructed features are able to represent the information in the input feature vectors to a quantifiable degree of precision, and are free of linear interdependencies.

The hypothesis tested in the example application described in this study was that information on peptide sequence is contained in the relative intensities of fragment ions derived together from a common peptide bond breakage. By training a CRF on features representing this fragmentation information, probabilistic sequence assignments could be made with good accuracy, depending on the length of the test peptides, with sequence recognition performance ranging from 60% to 88%. The fact that the technique demonstrates this level of performance using only a very naive feature set, without considering global peptide characteristics or feature conjunctions, lends strong support to this hypothesis.

The flexibility of the CRF provides several angles to pursue in fragmentation modeling. As mentioned, the set of features used in this study was extremely simple – extending the model with global information such as peptide mass, the relative position of the fragmentation along the peptide backbone, or chemical characteristics of the peptide may increase the discriminative power of the model for use in peptide identification, and may yield useful clues regarding the mechanisms of peptide fragmentation under different conditions, when using the model for directed exploratory data analysis.

There are many ways that a CRF sequence model could be useful in the interpretation of mass spectrometry data, in addition to the approach presented here. Complex data interdependencies, both among observed variables, and between observed variables and those to be estimated, are recurring themes in both general proteomics research and mass spectrometry data. As described in this study, the CRF provides an ideal, general purpose framework for probabilistic sequence modeling in these situations.

Though the conditional nature of the model renders generative predictions impossible, the flexibility and power of the CRF framework in diagnostic applications assures it a place in the future of proteomics research.

Bibliography

[1] Appella, E. and Anderson, C. W. *European Journal of Biochemistry* **268**, 2764–2772(9) (May 2001).

[2] Li, F., Wilkins, P. P., Crawley, S., Weinstein, J., Cummings, R. D., and McEver, R. P. *J. Biol. Chem.* **271**(6), 3255–3264 (1996).

[3] Qin, C., Baba, O., and Butler, W. *Crit Rev Oral Biol Med* **15**(3), 126–136 (2004).

[4] Hammes, A., Guo, J., Lutsch, G., Leheste, J., Landrock, D., Ziegler, U., Gubler, M., and Schedl, A. *Cell* **106**(3), 319–29 (2001).

[5] Fukuda, M. and Mikoshiba, K. *J Biol Chem* **274**(44), 31428–34 (1999).

[6] Grasser, F., Graf, T., and Lipsick, J. *Mol Cell Biol* **11**(8), 3987–96 (1991).

[7] Mears, A. J., Gieser, L., Yan, D., Chen, C., Fahrner, S., Hiriyanna, S., Fujita, R., Jacobson, S. G., Sieving, P. A., and Swaroop, A. *Am J Hum Genet* **64**(3), 897–900 (1999).

[8] Chang, J. S., Yeh, R.-F., Wiencke, J. K., Wiemels, J. L., Smirnov, I., Pico, A. R., Tihan, T., Patoka, J., Miike, R., Sison, J. D., Rice, T., and Wrensch, M. R. *Cancer Epidemiol Biomarkers Prev* **17**(6), 1368–73 (2008).

[9] Riley, J. H., Allan, C. J., Lai, E., and Roses, A. *Pharmacogenomics* **1**(1), 39–47 (2000).

[10] Patterson, S. D. and Aebersold, R. H. *Nat Genet* **33 Suppl**(NIL), 311–23 (2003).

[11] Weatherly, D. B., Atwood, J. A., Minning, T. A., Cavola, C., Tarleton, R. L., and Orlando, R. *Mol Cell Proteomics* **4**(6), 762–772 (2005).

[12] Huttlin, E., Hegeman, A., Harms, A., and Sussman, M. *Journal of Proteome Research* **6**(1), 392–398 (2007).

[13] Higgs, R., Knierman, M., Bonner-Freeman, A., Gelbert, L., Patil, S., and Hale, J. *Journal of Proteome Research* **6**(5), 1758–1767 (2007).

[14] Colinge, J. and Bennett, K. L. *PLoS Comput Biol* **3**(7), e114 (2007).

[15] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. *Mol Cell Proteomics* **3**(6), 531–533 (2004).

[16] Dongre, A., Jones, J., Somogyi, A., and Wysocki, V. *Journal of the American Chemical Society* **118**(35), 8365–8374 (1996).

[17] Schey, K., Schwartz, J., and Cooks, R. *Rapid communications in mass spectrometry* **3**, 305 (1989).

[18] Tabb, D. L., Huang, Y., Wysocki, V. H., and Yates, J. R. *Anal Chem* **76**(5), 1243–8 (2004).

[19] Tsaprailis, G., Nair, H., Somogyi, A., Wysocki, V., Zhong, W., Futrell, J., Summerfield, S., and Gaskell, S. *Journal of the American Chemical Society* **121**(22), 5142–5154 (1999).

[20] Griffiths, W. and Jonsson, A. *Proteomics* **1**(8), 934–45 (2001).

[21] Craig, A., Bennich, H., and Derrick, P. *Australian Journal of Chemistry* **45**(2), 403–416 01/01 (1992).

[22] Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. *J Mass Spectrom* **35**(12), 1399–406 (2000).

[23] Huang, Y., Wysocki, V. H., Tabb, D. L., and Yates, J. R. *International Journal of Mass Spectrometry* **219**(1), 233–244 (2002).

[24] Breci, L. A., Tabb, D. L., Yates, J. R., and Wysocki, V. H. *Anal Chem* **75**(9), 1963–71 (2003).

[25] Zhang, Z. *Anal Chem* **76**(14), 3908–22 (2004).

[26] Zhang, Z. *Anal Chem* **77**(19), 6364–73 (2005).

[27] Kapp, E. A., Schutz, F., Reid, G. E., Eddes, J. S., Moritz, R. L., O'Hair, R. A. J., Speed, T. P., and Simpson, R. J. *Anal Chem* **75**(22), 6251–64 (2003).

[28] Gay, S., Binz, P.-A., Hochstrasses, D. F., and Appel, R. D. *Proteomics* **2**(10), 1374–1391 (2002).

[29] Arnold, R. J., Jayasankar, N., Aggarwal, D., Tang, H., and Radivojac, P. *Pac Symp Biocomput* **NIL**(NIL), 219–30 (2006).

[30] Huang, Y., Tseng, G. C., Yuan, S., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., and Wysocki, V. H. *J Proteome Res* **7**(1), 70–9 (2008).

[31] Lafferty, J., McCallum, A., and Pereira, F. *Proceedings of the Eighteenth International Conference on Machine Learning table of contents* , 282–289 (2001).

[32] McCallum, A. and Li, W. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* , 188–191 (2003).

[33] Sutton, C., McCallum, A., and Rohanimanesh, K. *Journal of Machine Learning Research* **8**, 693–723 mar (2007).

[34] Chunjuan, D. and Yanjun, Z. *Conf Proc IEEE Eng Med Biol Soc* **6**(NIL), 6089–91 (2005).

[35] Garrow, A. G. and Westhead, D. R. *Proteins* **69**(1), 8–18 (2007).

[36] Majoros, W. H., Pertea, M., and Salzberg, S. L. *Bioinformatics* **21**(9), 1782–8 (2005).

[37] Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W., and Buhmann, J. M. *Anal Chem* **77**(22), 7265–73 (2005).

[38] Khatun, J., Hamlett, E., and Giddings, M. C. *Bioinformatics* **24**(5), 674–81 (2008).

[39] Devijver, P. A. *Pattern Recognition* , 131–143 (1988).

[40] Bhatt, M. R. and Desai, U. B. *Graphical Models and Image Processing* **56**(1), 61–yy January (1994).

[41] Wallach, H. M. Technical Report MS-CIS-04-21, University of Pennsylvania, (2004).

[42] Sutton, C. and McCallum, A. Technical report, University of Massachusetts, (2005).

[43] Rosti, A. and Gales, M. *Computer Speech & Language* **18**(2), 181–200 (2004).

[44] Liu, D. and Nocedal, J. *Mathematical Programming B* **45**(3), 503–528 (1989).

[45] Vishwanathan, S. V. N., Schraudolph, N. N., Schmidt, M. W., and Murphy, K. P. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006,* Cohen, W. W. and Moore, A., editors, volume 148 of *ACM International Conference Proceeding Series*, 969–976. ACM, (2006).

[46] Horn, D. and Axel, I. *Bioinformatics* **19**(9), 1110–5 (2003).

[47] Liu, H., Bonner, A. J., and Emili, A. *Conf Proc IEEE Eng Med Biol Soc* **4**(NIL), 3055–9 (2004).

[48] Pal, C., Sutton, C., and McCallum, A. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* **5**, V–V May (2006).

[49] McCallum, A. In *UAI '03, Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence, August 7-10 2003, Acapulco, Mexico,* Meek, C. and Kjærulff, U., editors, 403–410. Morgan Kaufmann, (2003).

Chapter 5

Concluding Chapter

5.1 Discussion and Conclusions

Proteomics, as it currently stands, could not exist without computational analysis, and computational analysis will drive future progress in proteomics. On the instrumentation front, the focus is on obtaining better chromatographic separations, better resolution, sensitivity, and specificity, and higher throughput. The onus is on the data analysis methods used to convert those advances into better, more meaningful biological results. Progress in the automated understanding of mass spectra will be beneficial for most aspects of proteomics research – extracting more relevant and detailed information from the data would permit more detailed questions to be asked of it, and would yield more confident and specific results (such as protein identifications) in everyday usage. In many areas of proteomics research, the consensus is that reliable computational analysis is becoming increasingly essential [1].

The present study examined three general approaches for the automated interpretation of mass spectrometry data, and demonstrated their application to mass spectrometry data produced from proteomics experiments. Three specific techniques based on these general approaches were developed and demonstrated: a correlation-based analysis for finding PTMs, a data visualizer and feature extraction tool based on the singular value decomposition, and a general-purpose peptide fragmentation modeling framework, capable of representing any available information, scalable to very large sizes, and showing promising performance as a peptide sequencing adjunct, able to utilize discarded fragment ion intensity information to increase the confidence of peptide identifications. These techniques show considerable future promise in the computational analysis of proteomics data, being mathematically rigorous, and computationally efficient. The generality of these tools confers them a great deal of flexibility in application and parameterization, and their inherent simplicity facilitates their interpretation.

The first approach explored in this study was the application of data transformation methods to mass spectrometry data. These are general pur-

pose methods that are capable of re-framing the input data to visually express different and (usually) more complex characteristics. These approaches also make these characteristics accessible to further analytical steps, such as modeling or filtering. Though the application of wavelet transforms to such data is well known, as discussed above, the utility of correlation-based transform methods has not been fully explored. This study demonstrated the utility of correlation in extracting mass-shift signals from mass spectra. These signals can be directly visualized for exploratory work, and can also be used as filtering criteria, to simplify downstream analyses. A basic implementation making use of these techniques was demonstrated, that used mass-shift signals isolated from mass spectrometry data to reveal the spectra in a full proteomics experiment that contain evidence of a possible post-translational modification. The modified peaks may then be identified within the highlighted spectra. Due to the simplicity of the implementation used, a high false-positive rate resulted when this approach was used with noisy data. However, methods exist for the identification of many of these false positives, including isotope envelope modeling [2]. The efficiency of the correlation procedure allows it to be stacked in series with these other applications, giving better performance on sub-optimal data.

The second approach explored in this study was the application of matrix decomposition and data projection methods to proteomic mass spectrometry data. These methods are similar to the data transformation methods discussed, but are more sophisticated and powerful. They decompose matrices of mass spectrometry data using techniques based on the eigenvalue decomposition in linear algebra, and can extract information, and trends in that information, that involve all of the variables simultaneously. They are also capable of indicating the relative importance of the information extracted. This study demonstrated the application of the singular value decomposition to mass spectrometry data. This matrix decomposition reveals the combinations of variables (peaks, in the case of mass spectrometry data) that, when combined, best explain the m/z-m/z differences between the spectra. A feature extraction methodology using the SVD was demonstrated, capable of highlighting particularly influential spectral regions, and of assembling "meta-variables" representing linear combinations of mass spectral features. Choosing the most important of these meta-variables created an optimally compressed representation of the data under the linear assumption (valid for representing fragmentation in mass spectrometry data) with a quantifiable rate of error, that could be coupled directly to sophisticated modeling approaches. The use of this compressed representation for visualizing high-dimensional mass spectrometry data was also discussed, along

with the interpretation of the meta-variables involved.

The third approach explored a new and promising application of statistical machine learning, known as a conditional random field, to mass spectrometry data in proteomics research. The model demonstrated is a chain-structured sequence model similar to many found in bioinformatics research, including the common hidden Markov model. The key difference is that this approach models the conditional probability distribution instead of the joint probability distribution. This relaxes the Markov independence assumptions that hold for joint-modeling generative models, which gives the CRF tremendous expressive power, and gives it an unparalleled ability to model noisy, highly correlated input data. To demonstrate the power of the CRF, a probabilistic model of MS/MS peptide fragmentation was constructed. This model is able to incorporate sequence-specific, non-local, and global peptide information into a unified modeling framework. The use of the SVD data extraction tool summarized above allowed the CRF model to be as general as possible – though the regions to examine for information were specified, the specific features to include in the calculation were automatically elicited from the data. The use of the model output for exploratory work in peptide fragmentation modeling was described, as well as its direct use in the probabilistic matching of peptide sequences to the spectra that generated them. Using this tool, it was demonstrated that the relative intensities of associated peptide fragment ions (a-, b-, and y-type primary ions, and their secondary ion derivatives, belonging to a single bond breakage) contain information about the sequence of the peptide, and that this information can be efficiently extracted and used to increase confidence in peptide identifications.

5.2 Future Directions

The association between proteomics and informatics is yet young, and several several key avenues of computational interpretation remain relatively unexplored. Working in a purely exploratory fashion, from a data-driven perspective alone, has not been the path taken during the development of most of the commonly used computational applications used in proteomics today, although the promise of such exploratory analyses are now being recognized [3], particularly in the field of biomarker discovery [4, 5], in diagnostic classifications [6], and in drug-target discovery [7]. This study has confirmed that the application of such general purpose, data-driven exploratory strategies can help to elicit meaning from mass spectrometry data,

and has demonstrated the development of several novel tools to leverage this information. This shows that "computational proteomics", performed in an exploratory fashion, is a relevant complementary approach to biologically motivated hypothesis testing.

Another direction to the computational research presented here lies in the integration of computational tools; combining simple computational tools allows transparent and modular construction of more complicated analyses, and when implemented as screening or filter steps, they may greatly reduce the load on downstream analytical steps, which is crucial in high throughput situations [8]. Combined applications suggested by this research include the use of correlation filters to isolate mass spectra containing post-translational modifications, the use of SVD projection methods to compress and condition data for optimal modeling, and the use of a probabilistic filter as an adjunct to protein identification methods.

Bibliography

[1] Kislinger, T. and Emili, A. *Expert Rev. Proteomics* **2**(1), 27–39 (2005).

[2] McIlwain, S., Page, D., Huttlin, E., and Sussman, M. *Bioinformatics* **23**(13), i328 (2007).

[3] Aliferis, C., Statnikov, A., and Tsamardinos, I. *Cancer Informatics* **2**, 133–162 (2006).

[4] Yasui, Y., Pepe, M., Thompson, M., Adam, B., Wright, G., Qu, Y., Potter, J., Winget, M., Thornquist, M., and Feng, Z. *Biostatistics* **4**(3), 449 (2003).

[5] Wagner, P., Verma, M., and Srivastava, S. *Annals of the New York Academy of Sciences* **1022**(1), 9–16 (2004).

[6] Mertens, B., Noo, M., Tollenaar, R., and Deelder, A. *Journal of Computational Biology* **13**(9), 1591–1605 (2006).

[7] Davidov, E., Holland, J., Marple, E., and Naylor, S. *Drug Discovery Today* **8**(4), 175–183 (2003).

[8] Havre, S., Singhal, M., Gopalan, B., Payne, D., Klicker, K., Kiebel, G., Auberry, K., Stephan, E., Webb-Robertson, B., and Gracio, D. *Proceedings of International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences* (2004).