# Bioinformatics of proteomic tandem mass spectra: selection, characterization, and identification

David L. Tabb

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2003

Program Authorized to Offer Degree: Molecular Biotechnology

UMI Number: 3102723

UMI Microform 3102723



ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature David L. Tabb

Date June 28, 2003

University of Washington

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

David L. Tabb

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Co-Chairs of Supervisory Committee:

John R. Yates, III

Philip Green

Reading Committee:

Philip Green

Walter L. Ruzzo

John R. Yates, III

Date: July 18, 2003

University of Washington

Abstract

# Bioinformatics of proteomic tandem mass spectra: selection, characterization, and identification

by David L. Tabb

Co-Chairs of Supervisory Committee:

Affiliate Professor John R. Yates, III
Molecular Biotechnology

Professor Philip Green
Molecular Biotechnology

Tandem mass spectrometry is a powerful technology for proteomics. Quadrupole ion traps can isolate ions of a particular peptide, fragment them through collision-induced dissociation, and catalog the fragment ions in tandem mass spectra. Database search algorithms such as Mascot and SEQUEST can then identify the peptides represented by a collection of these spectra. These spectra, however, have not been extensively characterized, leading to inaccuracies in the ways these algorithms model fragment ions. In this body of research, a new algorithm, "DTASelect," was created to summarize, filter, and compare the identifications produced by database search algorithms. The extent and significance of spectral similarity in proteomic collections was explored. A set of well-identified peptides was statistically characterized to demonstrate the impact of peptide sequence on fragmentation. This information led to the creation of a new fragmentation model, which made possible a new algorithm, "GutenTag," to identify peptides via an automated, accurate sequence tagging approach. Taken together, this research shows that more accurate models of fragmentation can both improve existing algorithms and make new classes of algorithms feasible.

## TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF TABLES

## ACKNOWLEDGMENTS

Proteosome samples in Chapters 2 and 3 were provided by Rati Verma and Raymond Deshaies of the Division of Biology and Howard Hughes Medical Institute at Caltech. W. Hayes McDonald, supported by the Merck Genome Research Institute, helped structure the examples of DTASelect's and Contrast's uses. Special thanks go to Laurence Florens and Mike Washburn, who encouraged me to publish DTASelect and adapt it for use by other labs.

Special thanks go to Linda Breci, Yingying Huang, Lori Smith, and Vicki Wysocki at the University of Arizona Department of Chemistry for their help in framing the issues explored in Chapter 4. Without their patient tutelage, I would still think *b* ions look like short peptides. The spectra used in this analysis were produced by Dayin Lin, who was supported by National Institutes of Health grant R33 CA81665-04. I thank Dayin for helping me produce my own MudPIT data for a yeast proteome. His data was also used for training GutenTag in Chapter 6.

I thank Ianessa Morantte and Marc Montminy for allowing me to use their gel band data in Chapter 5. Ryoma Ohi and Timothy Mitchinson supplied the MAP sample. Chris Wu and Mike MacCoss produced the rat hippocampal lysate used in this research and were supported by the American Cancer Society's grant PF-03-065-01-MG0 and NIH grant F32 DK59731, respectively. Michael Gross and Ilan Vidavsky aided me with their correspondence in the creation of NoDupe.

Chris Wu generated the defined protein mixture spectra analyzed in Chapter 6. Anita Saraf provided the cataract sample under NIH grant R01 EY13288-03. David Goldberg of Xerox Parc made many helpful suggestions in the development of GutenTag.

During the course of this research, I was supported by an NSF Graduate Research Fellowship, an NIH Training Grant, NIH grant R33 CA81665, and NIH joint funding shared between the Universities of Arizona and Washington.

Of course, I must thank my advisor, John Yates, who allowed me enough independence and had just about enough patience.

## DEDICATION

To my parents, who insisted.

How shall I go in peace and without sorrow? Nay, not without a wound in the spirit shall I leave this city. Long were the days of pain I have spent within its walls, and long were the nights of aloneness; and who can depart from his pain and his aloneness without regret?

*Kahlil Gibran, The Prophet*

# Chapter 1

# INTRODUCTION TO PROTEOMIC TANDEM MASS SPECTROMETRY

## *1.1 Overview*

Proteomics represents a rearrangement and augmentation of classical protein biochemistry . Where traditional techniques have focused on a few proteins per analysis, proteomics attempts to conduct the comprehensive analysis of complex protein mixtures. Proteomics could be defined as the science of assessing the protein state for a complex, cell, or tissue under a particular set of conditions. Depending on the techniques used, this may entail giving a catalog of proteins present [1], characterizing the post-translational modifications of those proteins [2], or evaluating their relative quantities [3]. Several technologies have been combined to yield this capacity. The general categories into which these tools fall include protein separation, mass analysis, and spectral identification.

Because proteomics deals with more complex mixtures than most protein biochemistry experiments, the science relies more heavily upon the separation of components in protein mixtures. The physical separation of proteins is usually conducted by two dimensional gel electrophoresis (2DGE) [4] or through liquid chromatography (LC) [5]. Denaturation and enyzmatic digestion are used in some methodologies to cleave the intact proteins into peptides. In these protocols, peptides are separated rather than proteins. Sample handling prior to mass analysis is handled in Section 1.2.

Mass analysis has been the driving engine of proteomics. Essentially, the adaptation of mass spectrometry to proteins has greatly increased the amount of data resulting from these experiments. Each mass spectrometer has three major elements: ion source, mass analyzer, and detector. Ionization in proteomics is generally handled via matrix-assisted laser desorption ionization (MALDI) [6] or electrospray ioninization (ESI) [7]. Mass analysis is typically managed through time-of-flight (TOF) [8] or quadrupole analyzers [9]. A detector records the ion current resulting from ions exiting the analyzer. Because the experiments reported in this research were all the result of ESI quadrupole

ion trap tandem mass spectrometry, these technologies will be emphasized in this introduction.

Mass spectrometry produces proteomic data, but proteomic information requires that the data be interpreted. Spectral analysis has been managed through several software packages. First, mass spectrometry is conducted via instrument control software; this control impacts which spectra will be produced and how they will be represented in files. When tandem mass spectrometry has been employed, the spectra must be extracted from the instrument capture files, filtered, and assigned to particular precursor ion charge states. Then the peptides may be identified, typically by sequence database search algorithms. Section 1.4 describes the handling of data after its production in the tandem mass spectrometer.

## *1.2 Sample separation*

Proteins can vary widely in mass, hydrophobicity, structure, and other characteristics. As a result, many techniques have evolved for separating them [10]. The primary techniques used for proteomics are gel-based separations and liquid chromatography separations. In this section, the use of gel separations for proteomics are first considered, and then a powerful two-dimensional liquid chromatography separation technique is described. These techniques can separate proteins or peptides such that particular portions of a sample can be analyzed independently of the rest.

### *1.2.1 Gel-based proteomic protocol*

Gel separations are common experiments for evaluating the components of protein complexes. Early proteomics efforts were grafted onto existing gel-separation technologies, such as in Lawrence et al's discovery of a regulatory factor of Insulin-like Growth Factor [11]. The protocol for gel-based proteomics experiments begins with the denaturation of proteins. The proteins are separated by one or two-dimensional gel electrophoresis, and bands or spots containing proteins of interest are cut from the gel. Disulfide bridges are reduced and alklyated, and proteolytic enzymes cleave proteins to peptides. These peptides are separated by reversed-phase liquid chromatography en route to a tandem mass spectrometer (see Figure 1.1).

Gel electrophoresis is a standard technique for separating proteins [12]. Typically, the proteins are first denatured by a detergent such as SDS. In one-dimensional electrophoresis, they are separated

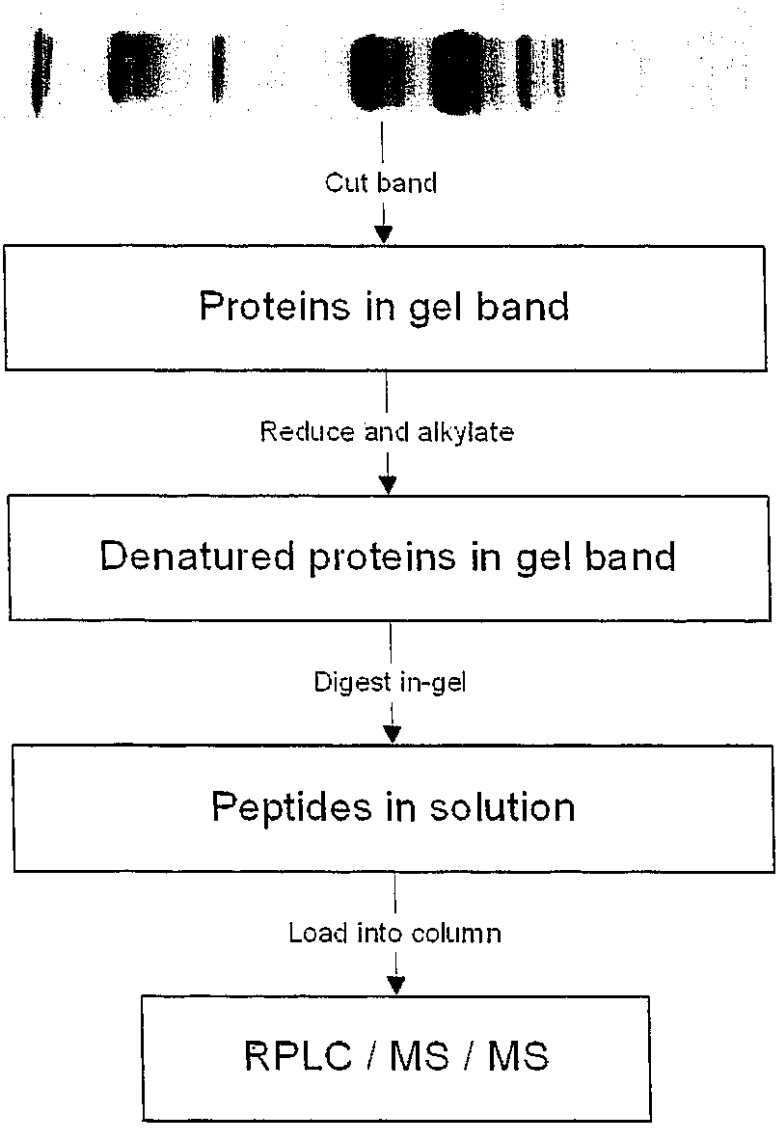



Figure 1.1: An overview of the steps in gel-based proteomics

by size by their sedimentation velocities through polyacrylamide. Because proteins contain different amino acid residues, the pH values at which they are neutral (called the isoelectric point) differ. In two-dimensional gel electrophoresis, the proteins are separated by size in one dimension and by isoelectric focusing in another. Such gels have been shown effective for resolving up to 10,000 proteins in a single experiment.

The positions of proteins are marked by applying a dye to the gel [13]; Coomassie blue dye is widely used but only highlights proteins for which at least 30-50 ng are present (this represents a best-case scenario). Silver staining, on the other hand, can mark positions for proteins of which at least 1 ng are present. Fluorescent dyes have roughly the same sensitivity as silver staining but may be more easily removed. Once proteins have been separated and positions are marked, gel bands (for 1D gels) or spots (for 2D gels) can be excised for extraction.

Protein tertiary structures may be maintained by disulfide cross-links between cysteine residues. Reduction of these bridges by dithiothreitol or TCEP can break these links, allowing the protein to be fully denatured. Subsequent alkylation by iodoacetamide blocks off the cysteine side chains and adds 57 Da to their masses. This protocol is useful when the proteins are to be digested to peptides; by breaking the disulfide bonds, reduction and alkylation prevents pairs of peptides from being linked together.

Several enzymatic digestions of proteins are available [14]. The most commonly used protocol in proteomics is the trypsin digest. This enzyme, which cleaves proteins after arginine and lysine residues, is available in a form bound to beads for removal from the sample after the digest. EndoK-C has the same cleavage specificity as trypsin, but shows better efficiency in the presence of urea and other denaturants, and use of the enzyme increases experiment cost. Alternative enzymatic cleavages may increase the diversity of peptides produced; subtilisin, elastase, themolysin, and proteinase K can be employed to create peptides covering different portions of a protein sequence.

These digestions yield peptides which can be separated en route to the mass spectrometer by reversed-phase liquid chromatography [15]. In the experiments reported herein, fused silica capillaries with inner diameters of 100 microns were pulled to form tips with inner diameters of 5 microns. These columns were then loaded with 5 micron C18-coated beads, and the sample's peptides were loaded into the column under pressure. A gradient of increasingly hydrophobic solvents was used to elute the peptides progressively from the column into the mass spectrometer.

This protocol can be effective, but many disadvantages prevent it from being an optimal proteomics strategy. First, gel electrophoresis may fail to retain proteins of extreme pI, molecular weight, or hydrophobicity. Proteins of low concentration may fail to be selected for removal from the gel if the staining technique is of insufficient sensitivity. Cutting gel bands or spots is a tedious process for which automation has only recently become available. Some reagents for gel electrophoresis of proteins are not compatible with mass spectrometry. On the other hand, these techniques enable existing separation methodologies to be used for proteomic analyses.

### *1.2.2 MudPIT*

A technique created in the Yates Lab attempts to automate separation and overcome the limitations of gel electrophoresis. The essential modifications in the multidimensional protein identification ("MudPIT") experiment are that proteins are reduced, alkylated, and digested to peptides prior to separation and that separation takes place via two-dimensional liquid chromatography rather than gel electrophoresis [16]. The resulting technique automates the analysis of even very complex samples (such as cellular lysates). See Figure 1.2 for an overview.

The MudPIT separation uses a biphasic column which elutes directly into a mass spectrometer. The first separation uses strong cation exchange (SCX) material to separate peptides by their charges at acidic pH. The second separation employs hydrophobic C18 material for a reversed phase (RP) gradient. MudPIT separations proceed in cycles: for each salt concentration in the SCX separation, a separate RP separation is conducted. As sample complexities increase, larger numbers of cycles can be employed.

In a typical MudPIT separation, twelve cycles might be conducted. The peptides are loaded into the biphasic column, and the peptides which initially passed through the SCX material and into the RP material are eluted during the first cycle of the MudPIT. For subsequent cycles, a few minutes' flow of a specified percentage of ammonium acetate moves a subset of the peptides from the SCX material to the RP material. This subset of peptides is then eluted via an acetonitrile gradient. The salinity for each cycle increases throughout the MudPIT. For example, these percentages of 500 mM ammonium acetate were used for separating the proteins of a rat hippocampus (see Chapter 5): 0%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 60%, and 100%. A linear gradient would yield

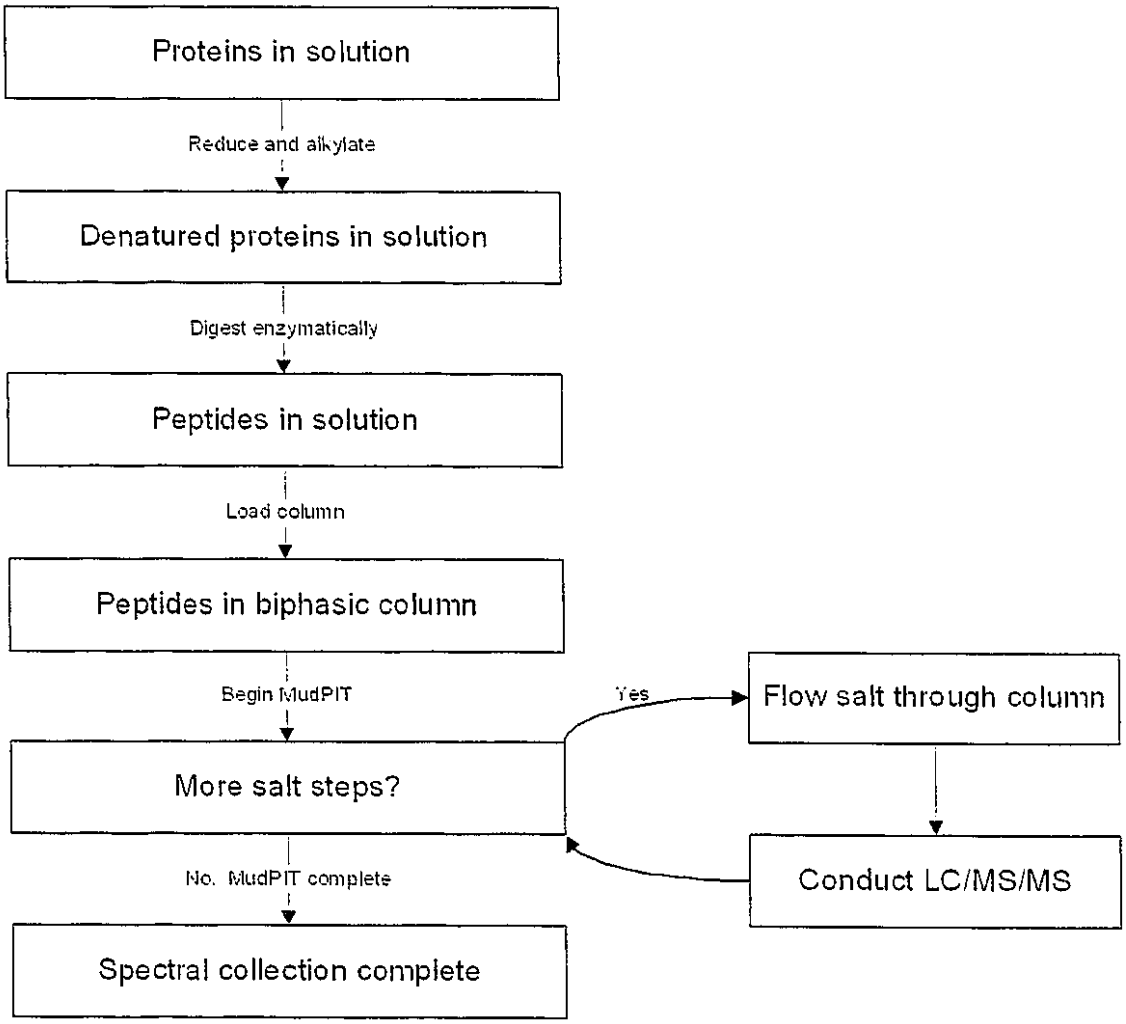



Figure 1.2: An overview of the steps in MudPIT proteomics

better separation than this type of step gradient, but the intent of the SCX is to segregate the peptides into aliquots for reversed phase separation rather than resolve them in a single dimension.

MudPIT separations help to automate separations for proteomics experiments. Their key advantage is that the eluent from the column flows directly into the mass spectrometer, obviating the need for tedious excision of gel bands or spots. In the next section, the fate of eluting peptides is examined in detail as they move through the mass spectrometer.

## *1.3 Mass analysis*

The use of mass spectrometry for proteomics yields much more detailed information than does a 2D gel image or UV trace from reversed phase liquid chromatography. If intact proteins are analyzed, mass spectrometry can give the molecular weights of the proteins to high precision. Peptides, on the other hand, can be subsequently fragmented to reveal primary structure. Only tandem mass spectrometry of peptides will be addressed in this treatement since the other techniques were not used for the described research.

### *1.3.1 ESI/MS/MS*

Proteins and peptides were not obvious candidates for mass spectrometry due to their high masses relative to other compounds. The first challenge of proteomic mass spectrometry was to introduce protein or peptide ions to the vaccuum of a mass spectrometer. John Fenn solved this problem by use of the electrospray ionization (ESI) technique [17]. Direct injection or liquid chromatography elutes proteins directly into an electrospray source, where a high voltage is applied, causing the proteins to separate from each other into smaller and smaller droplets. Pressure and voltage differences then force the charged proteins into the mass analyzer.

The most common type of mass analyzer paired with ESI is the quadrupole [18]. The traditional format of a quadrupole is four linear electrodes arranged in a square to form a channel down which ions fly (see Figure 1.3, but a rearrangement of the electrodes to form the quadrupole ion trap (see Figure 1.4) can lend extra versatility to the mass analyzer [19]. In either case, the way in which voltage is applied to the electrodes controls which ions are stabilized and which are destabilized in the area among the electrodes.

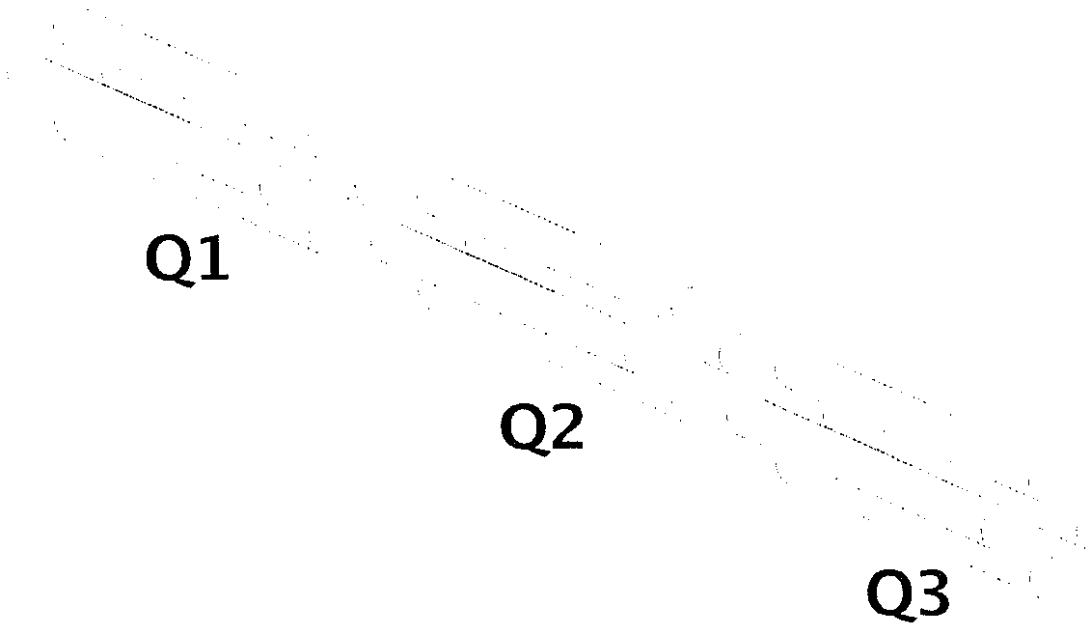



Figure 1.3: In triple quadrupole CID, mixed peptide ions enter the first quadrupole (Q1) from the ion source. In Q1, peptide ions of a particular sequence are retained while others are filtered out. In Q2, the peptide ions are energized by collision with gas molecules, causing them to fragment. Q3 is employed to scan through the mass range, cataloging the produced fragment ions in a tandem mass spectrum.

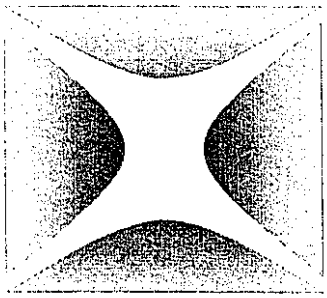


Figure 1.4: The hyperbolic cross-section of the quadrupole is repeated for the ion trap, but in three dimensions rather than two. The left and right electrodes shown above are part of a single "ring" electrode, while the top and bottom electrodes are electrically paired but separate "endcap" electrodes. In quadrupole ion trap CID, mixed peptide ions enter the trap, but only the selected peptide's ions are retained. Once enough ions are present, the trap energizes the ions, causing them to collide with gas molecules and dissociate. The trap then scans through the produced fragment ions to produce the tandem mass spectrum.

In the traditional quadrupole, a tandem mass analyzer consists of three sets of four electrodes– a triple quadrupole. Collision induced dissociation (CID) selects a particular peptide passing through the first quadrupole by m/z ratio, fragments the peptide in the second quadrupole, and scans the fragment ions in the third quadrupole. In this way, a triple quadrupole separates the processes of tandem mass spectrometry in space.

A quadrupole ion trap mass analyzer uses the same set of electrodes for all three processes. First, the trap collects ions of the selected peptide. Next, energy is applied to the peptide ions to cause them to oscillate more rapidly, resulting in more energetic collisions with the noble gas filling the trap and thus fragmentation of the peptide. Finally, the fragment ions resulting from fragmentation are ejected from the trap in order of m/z ratio to constitute the tandem mass spectrom. The three steps of CID are separated in space for triple quadrupoles but by time in quadrupole ion traps.

### *1.3.2 Collision induced dissociation*

Since CID plays such a key role in the research presented here, a closer examination is warranted. While many aspects of these reactions are poorly understood, the basics have been established through experiments spanning the last two decades. The key actors in CID are the surplus protons on each peptide ion and the energy applied through collisions with noble gas molecules.

Typically, proteomic samples are acidified before separation, resulting in positive charges for peptides. The added proton is most likely to be found at a site of high proton affinity or basicity. Among the amino acids, the side chains of arginine, lysine, and histidine show the greatest gas-phase basicity, with arginine the most basic of the three [20]. In addition, the N-terminus of the peptide may play host to protons. Because trypsin is commonly used to digest proteins in proteomic protocols, generated peptides often possess C-terminal arginine or lysine residues, which can retain one proton. Additional protons may be solvated among other side chains or at the N-terminus of the peptide.

The energy applied to peptide precursor ions in ion trap CID accelerates their motions within the trap. Noble gas molecules are always present in the trap at low pressure, but at normal energies the peptide collisions with them do not cause fragmentation. When energy is applied in CID, though, these collisions substantially increase the internal energy of the peptides, causing them to adopt unusual conformations and causing the additional protons to sample less accomodating locations

within the peptides (as explained in the Mobile Proton Model of peptide fragmentation [21]).

If a proton comes to rest near the oxygen of a peptide bond carbonyl, it may draw the bond electrons toward the oxygen, leaving the carbon partially positively charged (see Figure 1.5). In some cases, the preceding carbonyl's oxygen may attack the partially positive carbon, forming an unstable intermediate centering on a ring structure. This ion then breaks, generally forming one of three possibilities: an N-terminal *b* ion and a neutral leaving group, a C-terminal *y* ion and a neutral leaving group, or both *b* and *y* ions. Ion formation depends upon the number of protons carried by the precursor ion and where they are located; in the resolution of the intermediate structure, protons may transfer from one part to another. Typically, the *b* ion adopts an oxazolone structure. The *y* ion, however, generally adopts the linear structure of an ordinary peptide. Once a precursor ion has dissociated, its fragments are no longer energized by ion trap CID, which targets a narrow m/z range.

In some cases, alternate mechanisms are at work in forming fragment ions. If a peptide carries only one additional proton and it is immobilized at an arginine side chain, aspartic acid's side chain may attack its carbonyl to produce a dominant cleavage [22]. If proline is present in a peptide's sequence, cleavage may occur dominantly to its N-terminus but almost never to its C-terminus [23]. The chemical diversity of peptides implies a diverse chemistry for their fragmentation. In addition, fragment ions may subsequently fragment; *b* ions in particular are unstable and may lose carbon monoxide to form *a* ions.

In the final analysis, mass spectrometry of peptides meshes chemical processes (CID) with physical processes (mass analysis) to yield spectra giving structural information for peptides. The data produced, however, require analysis to give proteomic information. In the next section, we examine how the tandem mass spectrum can be identified.

### *1.4 Spectral analysis*

Initially, tandem mass spectra were matched to peptide sequences manually [24]. In contemporary usage, the path from spectrum to identification is managed by several algorithms. The first stage of this process is the acquisition of the spectra via the instrument control software. Next, the tandem mass spectra are separated from the other data and preprocessed by utility programs. Finally, the

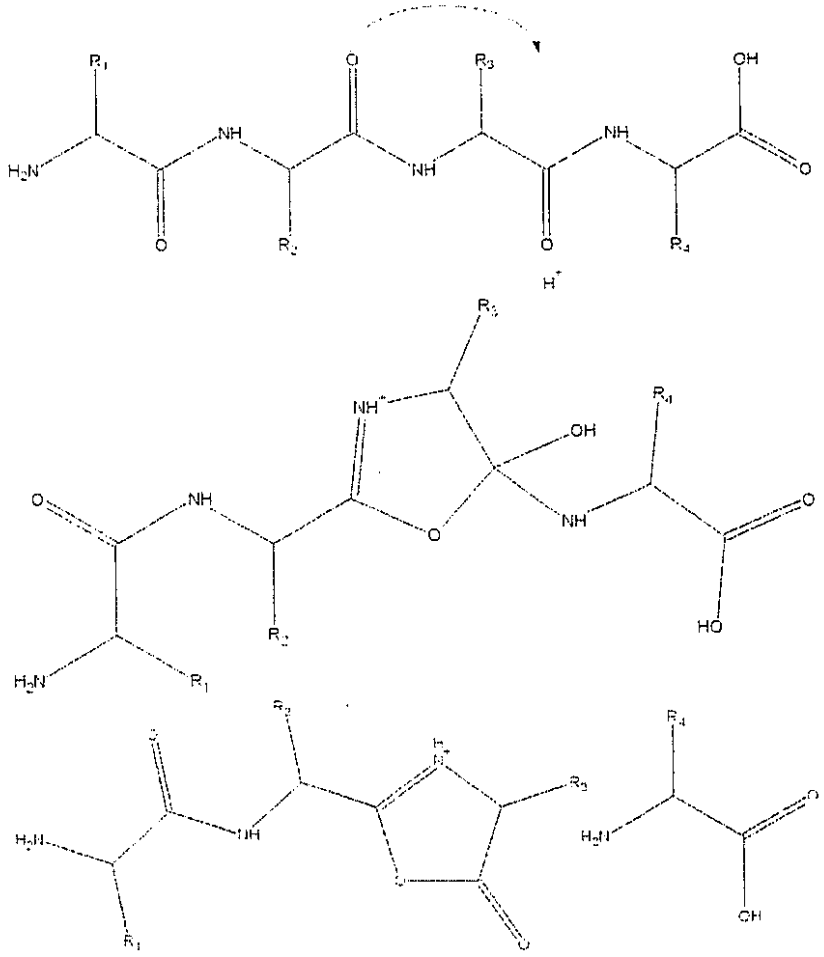


Figure 1.5: Chemistry of collision-induced dissociation

spectra are identified, usually by a database search algorithm such as SEQUEST [25]. An examination of each of these steps reveals many areas of active research.

### *1.4.1 Instrument control software*

The configuration of the software controling the mass spectrometer and pumps can affect the data produced in many ways. It can determine the gradient used in the liquid chromatography and thus which peptides are eluted when. It determines which peptides will be selected for fragmentation. This software controls the amount of energy applied to accelerate the peptides to begin CID. It sets the number of times each peptide's CID is repeated to be averaged for each spectrum. Importantly, the instrument control software handles data export to other software. A first-rate mass spectrometer can be hamstrung by instrument control software that is unstable or limits export of data to manipulable formats.

For the experiments described in this text, the Thermo Finnigan XCalibur instrument control software was used. XCalibur manages the pumps required for producing MudPIT gradients, automates the processes of peptide selection and CID, and exports data via an ActiveX control or to standard file formats. Although many XCalibur installations incorporate SEQUEST software for peptide identification, all applications in this document used external, independently managed installs of SEQUEST. Thermo Finnigan was the first to market an ion trap instrument to automate these processes and thus is used widely in the proteomics field.

### *1.4.2 Separation and preprocessing of tandem mass spectra*

When XCalibur captures data for a tandem mass spectrometry experiment, its capture file contains more than just tandem mass spectra. These data must be extracted to a separate file and then preprocessed. The ExtractMS program (http://fields.scripps.edu/sequest/extractms.html) was created for this purpose, creating a separate "DTA" text file for each tandem mass spectrum. Spectra which contain insufficient information (for example, spectra which represent failed fragmentation) are not extracted to files. In addition, the software attempts to deduce the charge state of the peptide fragmented for each spectrum. ExtractMS can distinguish spectra from singly-charged peptides from those which were multiply-charged, but it cannot distinguish between doubly-charged and triply-

charged peptide spectra, so two files are created for each multiply-charged spectrum, one representing each of these two charge states.

The 2to3 program [26] attempts to determine the charge state for the multiply charged spectra in a collection. When the software determines the precursor ion was triply-charged, it removes the doubly-charged spectrum copy, and vice versa. The software works by finding pairs of fragment ions; if +1 fragment ions pair with +1 fragment ions, the spectrum is likely from a doubly-charged precursor, but if +1 fragment ions pair with +2 fragment ions, the spectrum is likely from a triply-charged precursor. By removing these superfluous copies of the spectrum, 2to3 reduces the number of spectra to be processed by the peptide identification algorithm, generally the most computationally expensive part of this process.

### *1.4.3 SEQUEST, a database search algorithm*

SEQUEST, originally published in 1994 [27], introduced the concept of the database search peptide identification algorithm. Similar algorithms have also been published and made commercially available, notably Matrix Sciences' Mascot [28]. Software of this type uses a sequence database as the source of candidate sequences to explain each spectrum.

The algorithms begin by assembling a candidate sequence list. The mass of the peptide which was fragmented to produce a particular spectrum can be derived from the observed m/z value of the precursor ion and the charge state which has been assigned to the spectrum (in some cases, spectral copies at multiple charge states may be present). Since there may be some error in the peptide's calculated mass, an error of a few Da may be permitted. The database is queried for peptide sequences which have masses which match the precursor peptide's mass within this error range. If a specific enzymatic cleavage is indicated, only those peptides beginning or ending in appropriate amino acid residues may be considered. The collection of peptides which match the precursor peptide's mass constitutes the candidate sequence list for the spectrum.

In SEQUEST, preliminary scoring is used to prune the candidate list. The positions of fragment ions for each sequence are calculated, and the observed spectrum is inspected for the presence of these ions. A simple scoring function ranks the candidate sequences, and the top 500 are retained. Preliminary scoring plays a more significant role with large databases, when candidate lists may

include more than a thousand peptides.

Each of the remining sequence candidates is compared to the spectrum more rigorously. Once again, the positions of fragment ions for each sequence are calculated. A theoretical spectrum is generated containing peaks at the fragment ion positions, and then each theoretical spectrum is compared to the observed spectrum via cross-correlation. The sequences which generate the closest matching spectra are ranked most highly in SEQUEST's output.

Algorithms of this type can be differentiated based on the way in which they compare observed and model spectra. SEQUEST makes use of the cross-correlation algorithm for its comparisons, while Mascot and others use statistically-derived algorithms. Mascot's comparison starts with a few peaks and progressively uses more to refine its analysis, while SEQUEST uses a fast Fourier transform of the entire spectrum for each cross-correlation. These implementation differences can obscure the fact that fundamentally, these algorithms are achieving the same thing, comparison of a theoretical spectrum to the experimental one.

Less attention has focused on the way in which the model spectra are constructed. Fragment ion masses are generated from the candidate sequences. For N-terminal *b* ions, low intensity peaks are put into the model spectrum. For C-terminal *y* ions, high intensity peaks are inserted. This technique addresses the horizontal dimension (m/z) of the spectrum adequately, but it represents a simplistic approach to the vertical dimension (intensity). The ions in experimental spectra vary widely in intensity, even with individual ion series. Improving the techniques for intensity modeling could potentially increase the accuracy of database search algorithms.

### *1.5 Research goals*

This research was designed to construct a statistical foundation upon which improved proteomic tools could be built. As originally planned, the aims included:

- Select a set of representative peptide identifications,
- Characterize statistically the associated spectra,
- Develop a model by which accurate spectra could be predicted from sequences, and

- Create software which could infer partial sequences from spectra on the basis of this model.

Superficially, the first of these aims appeared to be the simplest. The Yates Lab had developed expertise in processing complex protein samples via MudPIT separation and subsequent SEQUEST analysis. The execution of this aim, however, quickly revealed unexpected challenges. Selecting a set of heuristics for consistently and automatically selecting reliable identifications was the first difficulty. Next, the "DTASelect" software was created to conduct this selection. Chapter 2 describes the process by which this aim was achieved. Several other capabilities became possible by the creation of the selection software, and these are described in Chapter 3.

The aim of selecting reliable identifications was to create a set of spectra from which statistical trends could be inferred. A set of yeast proteome spectra were acquired by a Yates Lab post-doc, and DTASelect chose a subset for which reliable identifications had been made. The primary challenge was to determine which aspects of the spectra were related to peptide sequences. After exploratory data analysis, some determinants of fragment ion intensity were published in the article which has been included as Chapter 4.

Review of proteomic datasets via DTASelect had made it apparent that spectra are commonly duplicated in these collections. In an effort to characterize the large-scale structure of proteomic spectral collections, the "NoDupe" algorithm was created for finding spectral pairs showing high similarity. The ubiquity and significance of spectra duplication were published in the article which was included as Chapter 5.

A spectral model was derived from the statistical analysis, and it was implemented in GutenTag, software for infering short sequence "tags" directly from spectra. These tag sequences were shown to be useful for identifying both modified and unmodified peptide sequences from a database. Chapter 6 describes this implementation of sequence tagging and demonstrates its effectiveness for peptide identification.

Chapter 2

# PEPTIDE IDENTIFICATION ASSEMBLY AND SELECTION: THE DTASELECT ALGORITHM

## *2.1 Introduction*

The first challenge in statistically characterizing peptide tandem mass spectra was identifying a suitable set for study. The tools available for summarizing SEQUEST results (SEQUEST Summary / Autoquest [1]) provided some basic capabilities for reporting identified peptides and filtering them on the basis of score cutoffs, but several significant capabilities were not implemented.

First, identifications from precursors of different charge states were given XCorrs from different distributions. A solid spectrum identification for a +1 charged peptide might receive a SEQUEST XCorr of 1.8, while equivalently identified doubly- and triply-charged peptides were likely to score approximately 2.5 and 3.5. To apply a single cutoff score to all classes of peptides discriminated against identifications of singly-charged precursors and be biased toward triply-charged precursors.

Secondly, the existing summary tools did not allow users to set rules for *protein* inclusion. A researcher might, for example, choose to see only those proteins for which at least two different peptides were present. To complicate the situation further, an individual peptide might have resulted from the presence of any of several proteins from a database. To work optimally, a summary tool would need rules for which proteins should be included and which excluded.

Thirdly, the existing tools did not handle redundancy in the data. Proteomic separations quite commonly produce multiple spectra for peptides (see Chapter 5). In analyzing a group of identifications, the best representative from similar sets should be chosen. Likewise, a set of peptides may be found in each of multiple proteins. This occurence was not noted in SEQUEST Summary or Autoquest.

Fourthly, Autoquest and Summary were not scaling effectively to handle the growing size of proteomic data sets. Summary was capable of collating the results for an individual reversed phase

separation. Several subsequent programs (collectively called Autoquest) were written to expand the program's capabilities, but each contributed its own layers of complexity. The problems inherent in proteomic summary necessitated the development of new software designed with large data sets in mind.

A new program was created in the Java programming language to select sets of spectra which met specific criteria. Since spectra were, at the time, stored in files with DTA extensions, the program was named DTASelect. The software allowed criteria customization at both peptide and protein levels, with reports designed to ease subsequent identification confirmation and biological analysis.

## *2.2 Experimental section*

### *2.2.1 DTASelect algorithm*

DTASelect functions in three phases: summarization, evaluation, and reporting. The summarization phase reads individual spectrum identifications from the SEQUEST output (OUT) files and assembles information for each locus (the database identity of the protein or gene from which the peptide derives). The evaluation phase applies the criteria indicated by the user for spectra and proteins, removing those that do not qualify. The reporting phase creates files to be read in a web browser, spreadsheet, or database. At each step, the user is notified of DTASelect's progress.

Summarization collects the most important information from SEQUEST identification. Each identification's output file is read to acquire the cross-correlation score (XCorr), normalized difference between first and second match scores (DeltCN), score and rank by preliminary score (Sp), identified sequence, observed and calculated precursor mass, protein or genetic locus from which this peptide derives, intensity, and percentage of fragment ions matched between predicited and observed spectra. The peptides are sorted by locus, and peptides for each locus are sorted by sequence. Next, the same database used by SEQUEST is consulted for the full protein sequence, descriptive name, and peptide positions. The information from the output files and database lookup is stored to the DTASelect.txt file so that this step can by bypassed in subsequent analyses of this sample.

Evaluation applies user-selected criteria to the matches. An identification is included only when it passes all specified spectrum criteria (see Tables 2.1 and 2.2). Selection is next conducted at a

higher level, retaining proteins which have a sufficient number of different peptides or which have at least one peptide showing up several times (see Table 2.3). Specific reports can be activated by the utility options (see Table 2.4).

Two filters reduce the redundancy of results. DTASelect can choose the best example of multiple spectra matching the same sequence by score or by signal intensity. In addition, many proteins may be indicated by exactly the same sets of peptides; DTASelect groups these together as identical. These selections highlight peptides and proteins which represent the entire sample.

The reporting phase generates several different files. The primary report is the DTASelect.html file, which enumerates the proteins and the spectra in evidence for each. Hyperlinks from this report lead to sequence coverage for individual proteins, a view of each spectrum, and the output file for each spectrum. This information is also recorded in DTASelect-filter.txt, a tab-delimited text file suitable for review in a spreadsheet. (See Figures 2.3 and 2.5.) For better data portability, DTASelect includes a graphical user interface, which can be used independently of SEQUEST's support programs. (See Figure 2.1.) This interface provides superior support for identifying sequence ions from spectra, especially those that come from triply-charged precursors. In another mode, it displays the selected protein's sequence coverage. These files and the graphical user interface present the primary information produced by DTASelect.

Supplementary information is presented in a series of auxiliary files. The first contains information about protein similarity. Multiple proteins may feature a particular peptide, but this is not always an indicator of substantial similarity among the proteins. Similarity is calculated to be the number of peptides shared between two proteins minus the number of nonmatching peptides; a score of zero indicates that the number of peptides that match equals the number that do not match. DTASelect notes all protein pairs which share at least one peptide in its primary report and in a separate similarity table.

Another DTASelect report evaluates chromatography. After filters have been applied, the number of singly, doubly, and triply charged peptides remaining from each round of chromatography is given, along with a profile of where in each cycle the peptides eluted. This information can help users optimize their separations for both completeness and efficiency.

Research focusing on post-translation modifications can benefit from DTASelect's modification report. This report, generated when the `--mods` option is used, yields information about

Table 2.1: DTASelect Spectrum Filters: # symbols indicate a numerical parameter follows while @ symbols indicate "true" or "false" should follow. Changes can be made at the command line or in a DTASelect.params file. Each spectrum must pass all filters.

| Option | Default | INDIVIDUAL SPECTRUM FILTERS: |
|---|---|---|
| `-1 #` | 1.8 | Set lowest +1 XCorr |
| `-2 #` | 2.5 | Set lowest +2 XCorr |
| `-3 #` | 3.5 | Set lowest +3 XCorr |
| `-d #` | 0.08 | Set lowest DeltCN |
| `-c #` | 1 | Set lowest charge state |
| `-C #` | 3 | Set highest charge state |
| `--mz #` | | Set minimum precursor m/z |
| `--MZ #` | | Set maximum precursor m/z |
| `-i #` | 0.0 | Set lowest proportion of fragment ions observed |
| `-s #` | 1000 | Set maximum Sp ranking |
| `-S #` | | Set minimium Sp score |
| `-a @` | true | Should ambiguous identifications be shown? |
| `-m 0` | | Require peptides to be modified |
| `-m 1` | X | Include peptides regardless of modification |
| `-m 2` | | Exclude modified peptides |
| `-y 0` | X | Include peptides regardless of tryptic status |
| `-y 1` | | Include only half- or fully tryptic peptides |
| `-y 2` | | Include only fully tryptic peptides |
| `-v -1` | | Keep "N" peptides, discard all others |
| `-v 0` | X | Ignore manual validation info |
| `-v 1` | | Keep "Y" peptides, discard "N" peptides |
| `-v 2` | | Keep "Y" and "M" peptides, discard "N" peptides |
| `-v 3` | | Keep "Y" and "M" peptides, discard "N" and "U" peptides |

Table 2.2: Extended Spectrum Filters: # symbols indicate a numerical parameter follows while $ symbols indicate character or string parameters follow. Extended filters only apply if they are specified).

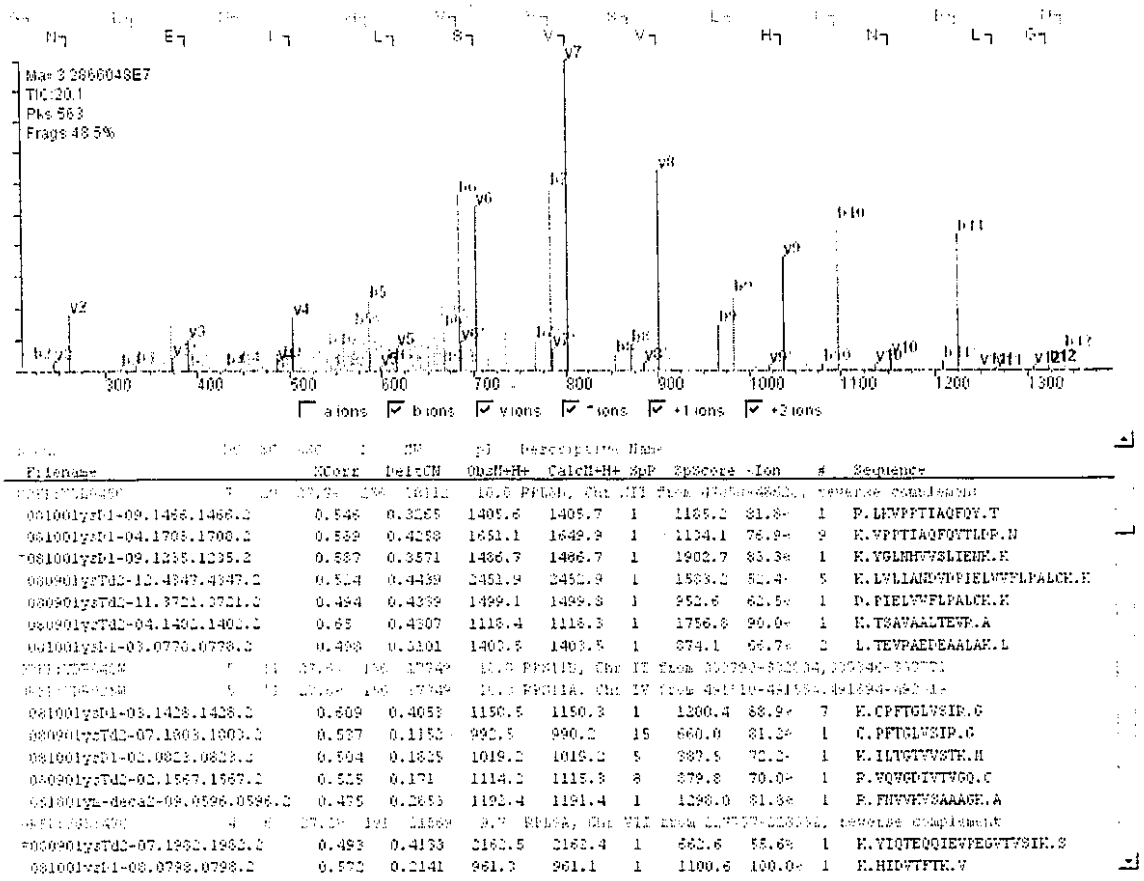
| Option | EXTENDED FILTERS (off by default): |
|---|---|
| -Sic $ | Sequences must contain all of these characters (excludes C terminal residue) |
| -Sip $ | Sequences must contain this pattern |
| -Sec $ | Sequences must not contain any of these characters (excludes C terminal residue) |
| -Stn $ | Preceding residue must be one of these |
| -Stc $ | C terminal residue must be one of these |
| -Smn # | Sequence must be at least this length |
| -Smx # | Sequence must be no longer than this |
| -Ser 0 | (default) No sequence end requirement |
| -Ser 1 | Sequences must be one-ended |
| -Ser 2 | Sequences must be complete to both ends |
| -X1 # | Set highest +1 XCorr |
| -X2 # | Set highest +2 XCorr |
| -X3 # | Set highest +3 XCorr |

Table 2.3: DTASelect Locus Filters: # symbols indicate a numerical parameter follows while $ symbols indicate character or string parameters follow. Filter changes can be made at the command line or in a DTASelect.params file. Each locus must pass either `-r` or `-p` and then any `-e`, `-E`, `-l`, `-L`, `-u`, and `-o` options specified.

| Option | Default | LOCUS FILTERS: |
|---|---|---|
| `-t 0` | | Do not purge duplicate spectra for each sequence |
| `-t 1` | | Purge duplicate spectra on basis of total intensity |
| `-t 2` | X | Purge duplicate spectra on basis of XCorr |
| `-V -1` | | Keep "N" proteins, discard all others |
| `-V 0` | X | Ignore manual validation info |
| `-V 1` | | Keep "Y" proteins, discard "N" proteins |
| `-V 2` | | Keep "Y" and "M" proteins, discard "N" proteins |
| `-V 3` | | Keep "Y" and "M" proteins, discard "N" and "U" proteins |
| `-u` | false | Include only loci with uniquely matching peptides |
| `-o` | false | Remove proteins that are subsets of others |
| `--mw #` | | Set minimum protein molecular weight |
| `--MW #` | | Set maximum protein molecular weight |
| `--pi #` | | Set minimum protein isoelectric point |
| `--PI #` | | Set maximum protein isoelectric point |
| `-e $` | | Remove proteins with IDs matching this string |
| `-E $` | | Include only proteins with IDs matching this string |
| `-l $` | | Remove proteins with descriptions including this word |
| `-L $` | | Include only proteins with descriptions including this word |
| `-M #` | 0 | Set minimum modified peptides per locus criterion |
| `-r #` | 10 | Show all loci with peptides that appear this many times |
| `-p #` | 2 | Set minimum peptides per locus criterion |

Table 2.4: DTASelect utilities

| Option | UTILITIES: |
|---|---|
| `--nofilter, -n` | Do not apply any criteria |
| `--copy` | Create script to copy selected spectra and IDs (or subset SQT and MS2 files) |
| `--GUI` | Report through GUI instead of output files |
| `--compress` | Create .IDX and .SPM files from spectra |
| `--Mascot` | Draw peptide IDs from Mascot .dat files instead |
| `--BE` | Produce Bird's Eye view of proteins found |
| `--class` | Classify proteins according to Classifications.txt |
| `--aux` | Incorporate auxiliary protein information from AuxInfo.txt |
| `--XML` | Save XML report of filtered results |
| `--DB` | Save in format for database import |
| `--chroma` | Save chromatography report |
| `--similar` | Save protein similarity table |
| `--align` | Save sequence alignment report |
| `--mods` | Save modification report |
| `--help, -h` | Print this list of options |
| `--here, -.` | Include only IDs in current directory |




| Filename | XCorr | DeltCN | ObsM+H+ | CalcM+H+ | SpP | SpScore | Ion | # | Sequence |
|---|---|---|---|---|---|---|---|---|---|
| 081001ysD1-09.1466.1466.2 | 0.546 | 0.3265 | 1465.6 | 1465.7 | 1 | 1185.2 | 81.8* | 1 | P.LEVPPTIAQFQY.T |
| 081001ysD1-04.1708.1708.2 | 0.569 | 0.4258 | 1651.1 | 1649.9 | 1 | 1134.1 | 76.9* | 9 | K.VPPTIAQFQYTLDP.N |
| 081001ysD1-09.1235.1235.2 | 0.587 | 0.3571 | 1486.7 | 1486.7 | 1 | 1902.7 | 83.3* | 1 | K.YGLNHVYSLIENK.E |
| 080901ysTd2-10.4847.4847.2 | 0.514 | 0.4439 | 2451.9 | 2452.9 | 1 | 1583.2 | 82.4* | 5 | K.LVLLAHVVPIELVYPLPALCK.E |
| 080901ysTd2-11.3721.3721.2 | 0.494 | 0.4399 | 1499.1 | 1499.8 | 1 | 952.6 | 62.5* | 1 | D.PIELVYPLPALCK.K |
| 080901ysTd2-04.1401.1401.2 | 0.65 | 0.4307 | 1118.4 | 1118.3 | 1 | 1756.8 | 90.0* | 1 | K.TSAVAALTEVP.A |
| 081001ysD1-03.0778.0778.2 | 0.498 | 0.3201 | 1403.5 | 1403.5 | 1 | 874.1 | 66.7* | 2 | L.TEVPAEIEAALAK.L |
| 081001ysD1-08.1428.1428.2 | 0.609 | 0.4053 | 1150.5 | 1150.3 | 1 | 1200.4 | 88.9* | 7 | K.CPFTGLVSIP.G |
| 080901ysTd2-07.1803.1803.2 | 0.527 | 0.1153 | 992.5 | 990.2 | 15 | 660.0 | 81.3* | 1 | C.PFTGLVSIP.G |
| 081001ysD1-02.0823.0823.2 | 0.504 | 0.1825 | 1019.2 | 1019.2 | 5 | 387.5 | 70.2* | 1 | K.ILTGTVVSTK.H |
| 080901ysTd2-02.1567.1567.2 | 0.535 | 0.171 | 1114.2 | 1115.3 | 8 | 879.8 | 70.0* | 1 | P.VQVGDIVTVGQ.C |
| 081001ys-decal-09.0596.0596.2 | 0.475 | 0.2693 | 1192.4 | 1191.4 | 1 | 1296.0 | 81.8* | 1 | P.FHVVKVSAAAGK.A |
| 080901ysTd2-07.1962.1962.2 | 0.493 | 0.4133 | 2162.5 | 2162.4 | 1 | 662.6 | 55.6% | 1 | K.YLQTEQQIEVPEGVTVSIK.S |
| 081001ysD1-08.0798.0798.2 | 0.572 | 0.2141 | 961.3 | 961.1 | 1 | 1100.6 | 100.0* | 1 | K.HIDVTFTK.V |

Figure 2.1: Example of DTASelect's GUI displaying a spectrum: Spectra can be viewed directly from the compressed files or from DTA or MS2 files. Clicking on the name of a spectrum in the list will display it in the viewing window.

gi—117385—sp—P02511—CRAB_HUMAN ALPHA CRYSTALLIN B CHAIN

| | | | | | |
|---|---|---|---|---|---|
| 19 | RRPFFPFHSPS | 2.8212 | 07160134Sela-18.1284.1284.2 | 1 | 1 |
| 80.0 | RPFFPFHSPSR | 3.2679 | 07160134Sela-18.1217.1217.3 | 1 | 1 |
| | HSPSRLFDQFF | 4.738 | 06130134Ssub-14.1615.1615.2 | 1 | 4 |
| | HSPSRLFDQFFGE | 5.301 | 06130134Ssub-14.1462.1462.2 | 1 | 47 |
| | * | | | | |
| 21 | HSPSRLFDQFF | 2.7263 | 06130134Ssub-16.1923.1923.2 | 1 | 2 |
| 80.0 | HSPSRLFDQFFGE | 4.4624 | 06130134Ssub-16.1853.1853.2 | 1 | 4 |
| | * | | | | |
| 35 | RLFDQFFGEHLLESDLFP | 2.599 | 07160134Sela-14.1842.1842.2 | 1 | 1 |
| 80.0 | * | | | | |

Figure 2.2: Example of DTASelect-mods.html Fragment

proteins with at least one modified peptide present (see Figure 2.2). Available in web (DTASelect-mods.html) and text (DTASelect-mods.txt) formats, the report enumerates the residues in each protein for which at least one dynamic modification was observed. When multiple peptides are observed for a particular modification, a sequence alignment of the identified sequences is produced with a marker to show the modification's position.

The alignment report can be helpful in showing which regions of a protein's sequence are most amenable to mass spectrometry. The report, generated when the `--align` option is in effect, creates a table for each protein. Each row in the tables represents a region of contiguous sequence coverage and shows how the peptides in these regions align against each other. Modified residues are indicated by being colored red (or are shown in lowercase in the text report). The starting and ending sequence positions of each contiguous region are included in the first cell of each row. After each protein table, four statistics are listed. The "Maximum Depth" field shows the maximum number of peptides listed in any contiguous region for this protein (not necessarily the largest number of peptides containing a particular residue of the protein). The "Peptide Residues Observed" field sums together the length of each peptide observed for the protein and can be used to determine

Table 2.5: Summary tables from DTASelect output for LC/MS/MS and MudPIT analysis of purified 26S proteosomes: (A) DTASelect summary output for LC/MS/MS analysis on 4 $\mu$g of purified 26S proteosome. Shown are total counts for proteins, peptides, and spectra. The difference between the nonredundant and redundant protein counts reflects that some proteins have been grouped together because of identical sequence coverage. When used with databases that contain a large number of related proteins (such as the human database), DTASelect's grouping functionality is a timesaver. (B) As in (A) except that results are for a MudPIT analysis of 40 $\mu$g of purified 26S proteosome.

| (A) | Proteins | Peptides | Spectra |
|---|---|---|---|
| Redundant | 38 | 345 | 414 |
| Nonredundant | 37 | 342 | 411 |
| (B) | Proteins | Peptides | Spectra |
| Redundant | 90 | 978 | 4189 |
| Nonredundant | 77 | 968 | 4175 |

the mulitiplicity of sequence coverage for each protein. The "Sequence Coverage" and "Sequence Length" show the same numbers as in the locus lines of DTASelect.html.

## *2.3 Results and discussion*

Analyses of purified yeast 26S proteosomes by both single dimension LC/MS/MS [2] and by MudPIT [3] have been reported previously. All experiments were performed on a Thermo Finnigan LCQ or LCQ Deca using standard procedures [4]. The resulting peptide identifications were processed by DTASelect.

### *2.3.1 DTASelect application: single dimension LC/MS/MS*

Digested peptides from 4 $\mu$g of purified 26S proteosome were resolved and analyzed during a 2 hour LC/MS/MS experiment [5]. After a SEQUEST search against a database of predicted yeast open reading frames and common contaminants, the resulting 6587 output files were filtered and summarized using the default criteria of DTASelect. (See Tables 2.1 and 2.3.) The filtering yielded 411 spectra corresponding to 342 peptides and 37 distinct proteins. (See Figure 2.5A.)

U YGL147C 2 5 7.9% 191 21604 9.7 RPL9A

U YNL067W 2 5 7.9% 191 21692 9.7 RPL9B

yeast.2052.2052.2 4.0341 0.3996 1807.64 1 67.9% 2 R.YVYAHFPINVNIVEK.D 2

yeast.1646.1646.2 2.5635 0.272 1381.86 1 72.7% 3 Y.AHFPINVNIVEK.D 22

Similarities:

YLL024C(1:1)

U YGR034W 2 2 15.5% 129 14639 10.3 RPL26B

* yeast.0695.0695.2 2.865 0.2723 1257.73 1 75.0% 1 R.KAYFTAPSSER.R

yeast.0342.0342.2 3.5671 0.2422 1101.44 1 93.8% 1 R.RDDEVLVVR.G 2

Figure 2.3: Example DTASelect.html fragment: Locus lines supply the current assigned manual validation status ("U" means no status assigned), the locus name, the number of peptides listed below, the total number of spectrum copies representing this locus, the percentage of sequence coverage, the length of the full protein sequence, the average molecular weight of the protein, the pI,and the descriptive name of the locus from the database. NOTE: DTASelect.html also contains the calculated precursor mass and Sp score for each peptide, omitted here for space reasons). If a peptide has multiple matching sequences with equal XCorrs, the number of ambiguous sequences is shown at the left end of the peptide line.

All but 2 of these 37 proteins had been observed before: 32 had been identified as part of the 26S proteosome, 1 was known to associate with the complex [6], and 2 were exogenous contaminants [7]. DTASelect prepared its report in minutes. (See Figure 2.3.) The DTASelect output compiled the important metrics for each spectrum into a single file, enabling further study through hypertext links.

### *2.3.2 DTASelect application: MudPIT*

Digested peptides from 36 $\mu$g of the same preparation of 26S proteosome as for the single dimension experiment were analyzed by MudPIT [8] and searched against the database. After filtering and summarization, the 33,841 output files were pared down to 4157 spectra representing 968 distinct peptides and 77 proteins. (See Figure 2.5B.) Upon manual evaluation of the results, only 4 of these

77 proteins (5.2%) passed through the filters as false positives. Such percentages are typical of the results for these experiments. The percentage of false postives, however, is a function of the filters employed by the user; DTASelect allows considerable latitude in the stringency applied to the analysis.

### *2.4 Conclusions*

DTASelect assembles protein-level information from peptide data more quickly, flexibly, and uniformly than existing tools. DTASelect focuses attention on peptides of interest by sweeping away less likely identifications. By streamlining data analysis for proteomics, DTASelect makes more complex experiments feasible. Most users of SEQUEST can benefit from DTASelect. Just as SEQUEST automated the time-intensive process of spectrum identification, DTASelect automates the process of SEQUEST result interpretation.

Chapter 3

# SAMPLE DIFFERENTIATION AND OTHER OUTGROWTHS OF DTASELECT

## *3.1 Introduction*

DTASelect started life as an identification selector for a specific project. Very quickly, though, it became the standard way for the Yates Lab to process SEQUEST identifications. By replacing the existing summary software, DTASelect opened several avenues for development.

First, by automating and standardizing the procedures by which identifications were selected, DTASelect made it possible to apply the same criteria to two or more samples, yielding the opportunity to compare protein content between samples. An earlier attempt at this software, the "mix" program, was limited to comparing pairs of samples and featured few filters for removing spurious identifications. DTASelect was expanded to incorporate the capability of comparing multiple samples in the program Contrast.

The Yates Lab was increasing the pace at which spectra were collected, and DTASelect was making it possible to assemble the results for samples more rapidly as well. The growth of the Lab's data collection was putting a significant strain on the file servers. Because DTASelect could be modified at will (in contrast to existing tools, which were effectively static), it became possible to change to a new system for storing data which scaled to large sample sizes much more successfully.

Changing to a new format for storing spectra and identifications necessitated new tools for reviewing these data. DTASelect was modified to incorporate a novel spectrum viewer with many powerful features. Likewise, DTASelect's protein sequence assemblies enabled a new viewer to show, for the first time, depth of protein sequence coverage rather than simple percentage of residues covered.

In this chapter, each of these features will be examined in detail. By making significant changes in how day-to-day proteomics can be conducted, each has improved the speed and accuracy of data analysis in the Yates Lab.

## *3.2 Contrast: flexible and fast sample differentiation*

### *3.2.1 Algorithm*

Contrast extends DTASelect across multiple samples or multiple sets of criteria. Its algorithm proceeds in three phases: reading, comparison, and reporting. The program is configured by changes in the Contrast.params file, and the specified DTASelect results are read from disk. The presence of each protein is compared between the described sets, and the similarities and differences are reported in HTML and text formats.

The reading phase begins with the Contrast.params file, which allows the user to specify the DTASelect results of interest and to enumerate criteria sets to be applied to the results. Contrast requires at least one sample and set of criteria to be specified, but more can be included as long as the number of samples multiplied by the number of criteria sets does not exceed 63. Each DTASelect result is processed under each set of criteria, so four samples and two criteria sets will result in eight "data sets," corresponding to eight columns in the output. Once the Contrast.params file sets up the run, Contrast will read the DTASelect.txt file for each sample and apply the specified filters as in DTASelect.

To compare the data sets, Contrast develops a master list of proteins including each locus appearing in any of the data sets. Each data set is assigned a power of two. Each locus' pattern of presence or absence is stored by summing the powers of two for the data sets in which the protein is found. Finally, the master list of loci is sorted by these power sums and then by protein name. The effect is that loci are grouped by patterns of appearance across data sets.

Contrast produces extensive output. The major report is Contrast.html, which lists the proteins in each group and includes links to the corresponding DTASelect.html files for each data set. (See Figures 3.3 and 3.2.) To permit Contrast results to be analyzed by spreadsheet, they are exported in a tab-delimited text file. For users who want to see how individual peptides for each protein differ among data sets, Contrast offers a "verbose" mode, which provides this level of detail. (See Figure 3.1.) Although a complex comparison can yield a lengthy report, Contrast's output organization ensures that details can be found in predictable locations.

| Locus | prev def | new def | Total | Description |
|---|---|---|---|---|
| YDR471W | 17.6 | 21.3 | 30.9 | RPL27B |
| YHR010W | | | 30.9 | RPL27A |
| K.KVVIVKPHDEGSK.S +2 | 3.6802 | | | |
| K.SVVSTETFEQPSQREEAK.K +2 | | 2.9577 | | |
| K.VVIVKPHDEGSK.S +2 | 3.0628 | | | |
| K.VVNYNHLLPTR.Y +2 | 3.2914 | 2.9249 | | |

Figure 3.1: Sample Verbose Contrast.html fragment: Proteins YDR471W and YHR010W were found in both samples under this criteria set, though with different sequence coverages (17.6% and 21.3%, respectively). One peptide was found in both samples, but the other peptides were found in only one. The highest XCorr for each peptide in each sample is shown beside its sequence. Cumulatively these peptides add up to 30.9% sequence coverage. The sequence coverage percentages for each sample lead to the relevant sections in the respective DTASelect output files. The cumulative sequence coverage links to a view of the protein's sequence overlaid with the peptide sequences.

### 3.2.2 *Contrast application: MudPIT versus MudPIT*

The most common application for Contrast is the comparison of multiple LC/MS/MS or MudPIT experiments. The recent MudPIT analysis was compared to a previous MudPIT analysis of 60 $\mu$g of purified proteosome [1]. The summary table generated by Contrast allows the researcher to evaluate how many proteins were found in both samples. (See Figure 3.2.) Manual evaluation confirmed that none of the 60 shared proteins were passed through the filters as false positives. For this group of proteins, all were either components of the proteosome (32), associated with the proteosome (8), highly abundant yeast proteins that are found in numerous purifications (18), or exogenous contaminants such as immunoglobulin and trypsin (2).

Contrast produces an analysis within a few minutes that would take a few hours to produce manually. As in DTASelect output, important information is directly available, including locus name, percent sequence coverage, and locus description / protein name. (See Figure 3.3.) The proteins are sorted by the experiments in which they are found, and hypertext links provide access to more detailed information for evaluative purposes. When multiple positive and negative controls

| Count | Percent | prev def | new def |
|---|---|---|---|
| 60 | 50.8% | X | X |
| 18 | 15.3% | | X |
| 40 | 33.9% | X | |
| 118 | | 100 | 78 |

Figure 3.2: Sample Contrast.html Summary: Each row in this table represents a particular combination of presence and absence in each of the data sets, with the "X" marks indicating this pattern. Each row's count links back to the appearance of the group above it in the Contrast.html file. Of the 118 proteins appearing, 60 were present in both samples, 18 were present only in the "new" analysis, and 40 were found only in the "prev" experiment.

are part of the experiment, Contrast's automation becomes even more valuable.

### *3.2.3 Miscellaneous Contrast applications*

Since a large variety of customizable filters are available in DTASelect and thus Contrast, individual researchers can adapt the programs to their own needs. Contrast can be used for a "step analysis" in which progressively more stringent criteria are applied to data from a single experiment. Those proteins that meet the highest stringency requirements are sorted to the top of the output file, and those passing only the lowest criteria are sorted to the end. This approach is well suited for stratifying the protein identifications in a complex sample. For projects focusing on a particular protein, the verbose mode of Contrast can show how individual peptides vary among samples. The open-ended design of the software accomodates many purposes.

## *3.3 Unification of spectral and identification file formats*

When SEQUEST was created, the peptide mixtures it was processing might contain hundreds of peptides. As a result, it made sense to store each spectrum in a separate DTA file, with identifications stored in separate OUT files. As separation techniques have improved, however, samples of greater complexity are being analyzed. A typical MudPIT analysis could produce 50,000 spectra. This

| Locus | prev def | new def | Total | Description |
|---|---|---|---|---|
| KERATIN03 | | 9.1 | 9.1 | |
| KERATIN13 | | 3.3 | 3.3 | |
| KERATIN22 | | | 3.3 | |
| KERATIN20 | | 6.8 | 6.8 | |
| YBL027W | | 15.3 | 15.3 | RPL19B |
| YBR084C-A | | | 15.3 | RPL19A |
| YCR076C | | 18.0 | 18.0 | |
| YDR143C | | 5.7 | 5.7 | SAN1 |
| YDR418W | | 23.0 | 23.0 | RPL12B |
| YEL054C | | | 23.0 | RPL12A |
| YEL026W | | 19.0 | 19.0 | |
| YEL037C | | 10.1 | 10.1 | RAD23 |
| YHR111W | | 5.0 | 5.0 | |
| YJL141C | | 3.6 | 3.6 | YAK1 |
| YMR076C | | 2.9 | 2.9 | PDS5 |
| YNL069C | | 8.1 | 8.1 | RPL16B |
| YNL207W | | 3.3 | 3.3 | |
| YOL081W | | 1.6 | 1.6 | IRA2 |
| YOR123C | | 7.8 | 7.8 | LEO1 |
| 18 | | X | | |

Figure 3.3: Sample Contrast.html fragment: This represents a group of proteins that appear in the new MudPIT sample but not the previous experiment when the same criteria are used against each. Each row in the table represents one protein, and the numbers in the columns are the sequence coverage percentages found in each data set (or, in the Total column, the cumulative sequence coverage across multiple columns). The percentages link to each protein's location in a corresponding DTASelect.html file. If multiple proteins have identical sequence coverage, they are grouped together (for example, RPL19A and RPL19B). Several such sections appear in each Contrast output file, one for each combination of presence and absence.

growth required the Yates Lab to develop a new system for storing spectra and identifications.

Individual DTA and OUT files do not use disk space efficiently. These files, which typically store one kilobyte of data, generally take four or even eight kilobytes of hard disk space, depending upon the filesystem used. The overhead is not necessary for system data; instead, this extra space is simply not used. The filesystems' minimum file sizes are often larger than those of DTA and OUT files.

Filesystems perform poorly when thousands of files are created in a single directory. Microsoft Windows Explorer becomes unresponsive if such a directory is viewed. Linux may take several seconds to prepare such a directory listing. The problem with these directories is not that too much data is stored there but the way in which that data is stored.

Finally, reading identifications from OUT files is problematic. The data is formatted in these files for human review, not software parsing. Column headings and other labels are repeated in each file. Different variants of SEQUEST produce varied OUT file formats. Because each OUT file had to be opened individually, identification summary is slower than if the identifications were stored in a single file. Shifting away from the DTA and OUT formats would ease software analysis of these files.

The Yates Lab adopted the MS2 file format for storing spectra [2]. These tab-delimited files are essentially concatenations of DTA files, with labels to mark the start of each spectrum. When a MudPIT separation uses twelve different salt concentrations, twelve reversed phase gradients are run. Each results in an individual MS2 file. An unanticipated advantage of this change was that multiple SEQUEST analyses could be run on a single set of spectral files through symbolic linking; previously, the lab had re-extracted the DTA files for each analysis.

Changing the OUT file format constituted a greater challenge. The data in OUT files are structured; some values apply to the analysis of the spectrum, while others apply to individual sequences matched to the spectrum. Likewise, some sequences are found in individual database proteins while others are present in large numbers. In effect, three classes of information are present in each OUT: spectral analysis, sequence match description, and proteins containing matched sequence.

The Yates Lab adopted the tab-delimited SQT file format to store this information. S lines store spectral analyses, M lines store sequence analyses, and L lines store the proteins corresponding to each sequence. In essence, a SQT file is a flat file representation of a database.

New software was created for managing data in these formats. SQTSort, for example, was created to move the best-scoring identifications to the top of the SQT and MS2 files for improved performance during manual evaluation. PepGrep was designed to match the sequences presented in a SQT file with entries in a specified database, making it possible to predict a proportion of identifications against a new database from the identifications in an old one. SQTCollator is useful when different SEQUEST searches have been produced for a single set of spectra; the program can determine which identification across all the searches was the best for each spectrum. While many of these aims could be achieved for DTA and OUT files, the difficulty of parsing OUT files stood in the way of these developments.

The transition from DTA to MS2 and OUT to SQT has enabled the Yates Lab to increase the scale of its experiments. Disk space, which had been a constant concern, is now more efficiently used. New tools can now be created with greater ease. This greater efficiency has made large-scale proteomics a more reasonable plan of attack for biological problems.

### *3.4 Associated CGI development*

The shift from DTA to MS2 and OUT to SQT required substantial redevelopment of lab tools. The software for producing DTA files [3] was modified to produce MS2 files, and SEQUEST itself was altered to read MS2s and generate SQTs. Because many of the CGI programs previously in use were specific to DTA and OUT file formats, a host of new ones were created. The development work had the benefit of causing the group to re-think how the evaluative tools should work, and improved software resulted.

#### *3.4.1 Show / SpectrumApplet*

Display_Ions, the spectrum viewer previously used by the lab, was specific to DTA files. DTASelect featured a spectrum viewer as part of its graphical user interface, and it was adapted to present an applet interface. Hayes McDonald produced a CGI named "Show" to send this applet the spectral data in the form of an HTML file. In addition, Show integrated the identification from the SQT file with the spectrum, allowing the replacement of Show_Out, a separate CGI for viewing SEQUEST identifications. The end result was a CGI which would replace both of these existing CGIs (see

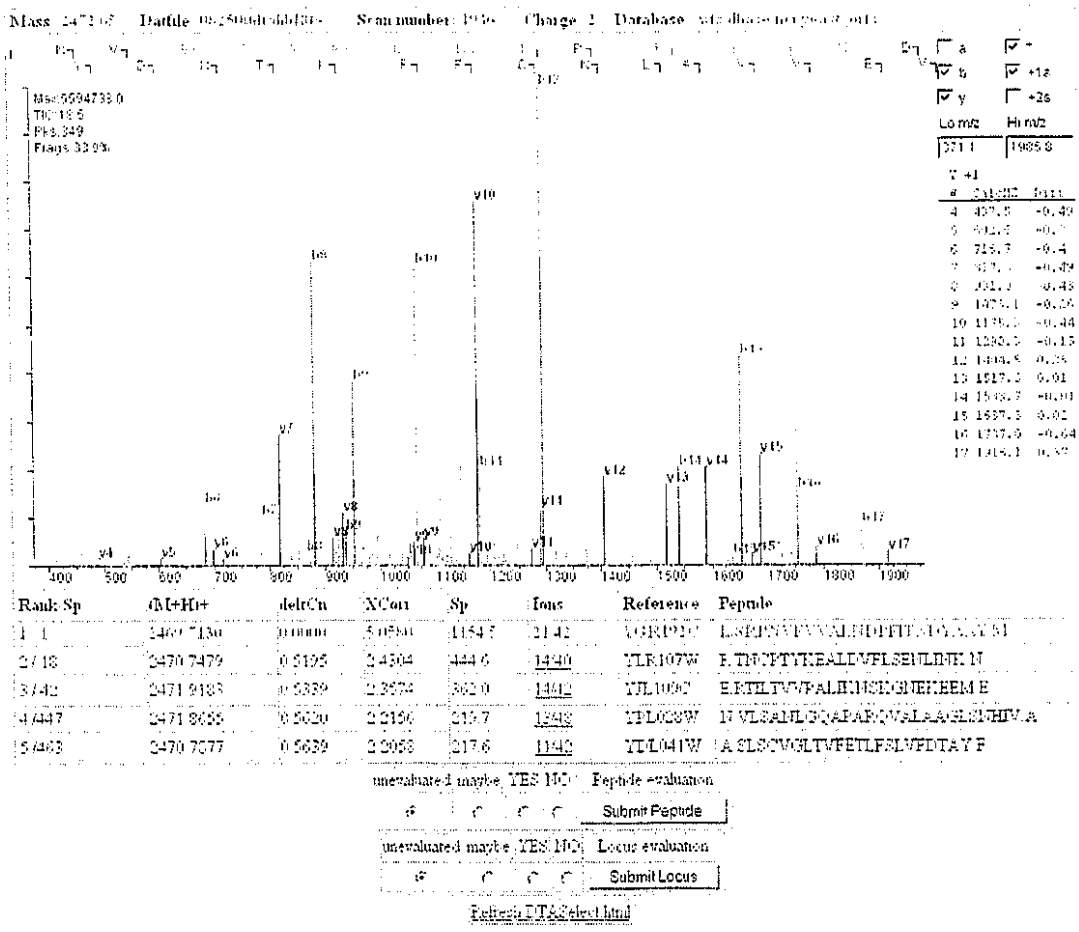



Figure 3.4: Example of Show CGI output using SpectrumApplet viewer

Figure 3.4).

The transition to the DTASelect spectrum viewer led to many important changes. First, the applet was modified to allow for zooming particular regions of the spectrum. A color-coding system was used to label peaks that corresponded to fragment ions. Neutral losses from the precursor ion and fragment ions were identified and labelled. Doubly-charged fragment ions were displayed in a way that made triply-charged peptides much more easily validated. The development of a new spectrum viewer has made a contribution to results analysis in the Yates Lab.

### *3.4.2 SeqCov*

A CGI named Consensus ships with SEQUEST to show which amino acid residues of a protein sequence were observed in peptides and which were not. Since other CGIs were being rewritten, this one was also revisited. Johannes Graumann created a Perl script, entitled SeqCov, which would display the *depth* of sequence coverage throughout a protein's sequence. DTASelect was modified to determine the coverage depths for each protein and to communicate with the SeqCov CGI (see Figure 3.5).

Using SeqCov rather than Consensus has helped to reveal the extent to which peptide identifications tend to group in some areas of the protein sequence while others are not observed. This information can be useful in tuning proteolytic digests to observe sequences of interest. This display and DTASelect's sequence alignment may be also useful in revealing database sequence errors.

### *3.4.3 Evalocus and EvalPeptide*

The unified SEQUEST result format included a field to store a researcher's manual evaluation of a peptide identification. Johannes Graumann wrote Evalocus to handle evaluation of proteins, and Hayes McDonald created Evalpeptide to manage this for peptides. Users can evaluate either proteins or peptides as legitimate matches, possible matches, unlikely matches, or unknown (the default). The capability to store these assessments can then be filtered by DTASelect. An annotation that was not possible with individual OUT files became possible with the file format unification.

## *3.5 Conclusion*

The creation of new proteomic summary software had powerful implications for the Yates Lab. The capability to differentiate samples' protein contents with customizable criteria has been significant in lab publications [4]. The transition to unified file formats has made it possible to store proteomic data more compactly and to analyze it more rapidly. The associated replacement of accessory programs has contributed to the assessment of proteomic identifications. As these changes propagate to other labs, we believe that they will benefit as well.

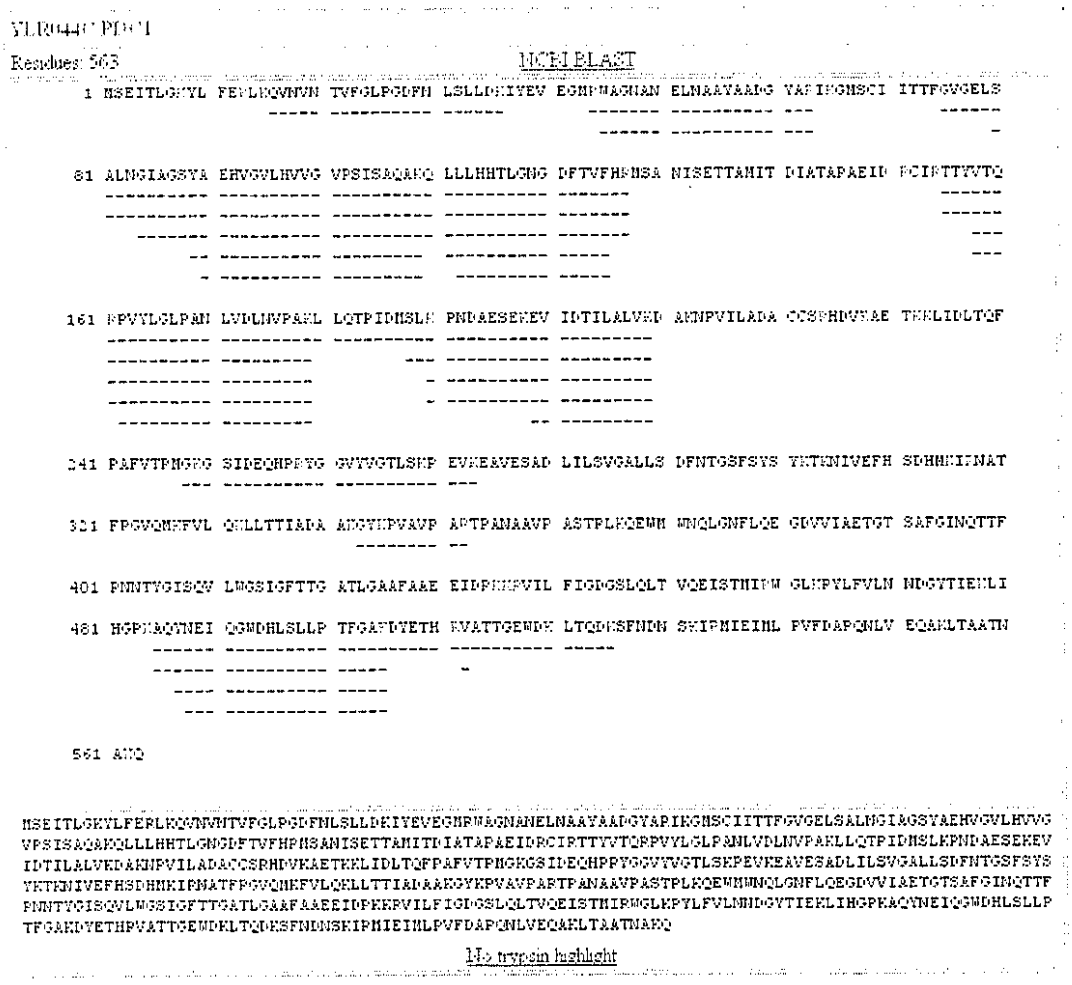



Figure 3.5: Example of SeqCov CGI output showing DTASelect's assembly of peptide sequences

Chapter 4

# STATISTICAL CHARACTERIZATION OF ION TRAP TANDEM MASS SPECTRA

## *4.1 Introduction*

DTASelect's initial purpose was the selection of spectra which had been identified confidently. A statistical analysis of this collection could then reveal the influences at work during collision-induced dissociation of peptides. These statistics, in turn, could be used to model spectra more accurately for peptide identification by *de novo* techniques.

Previous efforts to statistically analyze fragment ion spectra had limited success. Van Dongen and coworkers assembled a collection of 138 peptides for their analysis but focused on singly charged precursor ions and used high energy CID [1]. Dančík and coworkers studied a collection of low energy CID spectra while developing the SHERENGA algorithm, but the analysis assumed that all ions from a series would exhibit similar characteristics [2]. A promising recent effort applied a kinetic model to simulate fragment ion spectra as a function of several mechanisms [3]. Such studies have begun to establish a statistical foundation for future algorithms.

This research attempted to identify statistical trends in fragment spectrum peak intensity and to put these trends into chemical context. The relationship of fragment ion peak intensity to ion series origin and fragment mass were explored first. Next, the influence of amino acid residues on neighboring amide bond cleavages was evaluated. Finally, the link between amino acid composition and neutral loss fragmentation was tested.

## *4.2 Experimental section*

### *4.2.1 Preparation of yeast proteome for MS analysis*

A culture of *Saccharomyces cerevisiae* (strain 1560) was grown to an optical density at 600 nm of 1.2 in 2 L of YPD media. Cells were lysed and proteins were extracted and digested according to a

protocol similar to that described for the insoluble fraction by Washburn *et al.* [4]. Briefly, the total yeast lysate was centrifuged and the pellet portion was dissolved in 90% formic acid and treated with cyanogen bromide, CNBr, (Sigma, St. Louis, MO) overnight while the soluble portion was directly reduced with tris(2-carboxyethyl)phosphine (Pierce Chemical Company, Rockford, IL) and alkylated with iodoacetamide (Sigma, St. Louis, MO). For the CNBr treated fraction, the pH of the solution was brought up to 8.0 with cold 30% $NH_4OH$ and saturated $(NH_4)_2HCO_3$. The peptides of the pellet fraction were reduced and alkylated like those of the soluble fraction. For both soluble and pellet fractions, proteins were digested sequentially with Endoproteinase Lys-C and trypsin. The resulting peptides were purified once using SPEC solid phase extraction C18 pipette tips (Ansys Diagnostics, Inc., Lake Forest, CA).

### *4.2.2 Analysis of proteome by MudPIT*

Multidimensional Protein Identification Technology (MudPIT) was used to determine the protein content of this complex mixture [5]. The soluble fraction was divided and analyzed twice, while the pellet fraction was analyzed in a single experiment. The liquid chromatography columns were constructed at the time of use. A piece of fused silica capillary (100 $\mu$m ID / 363 $\mu$m OD) (Polymicro) was pulled to have an opening of 5 $\mu$m. A biphasic LC column was prepared by packing the capillary with AQUA C18 particles (Phenomenex, Torrance, CA) followed by Whatman SCX beads (VWR) using a high-pressure cell [6]. The yeast peptides were loaded onto the column using the same pressure cell.

The columns were positioned to elute directly into the ion source on a custom-built stage [7]. An Agilent HP1100 LC pump (Palo Alto, CA) produced a flow of 100 $\mu$L / min., which was split to produce a flow rate of 200 - 300 nL / min. There were 12 LC cycles in the completely automated LC/LC/MS/MS analysis. The initial cycle used no salt in order to collect data on peptides that bypassed the SCX media. Subsequent cycles used 4%, 8%, 10%, 12%, 15%, 20%, 30%, 40%, 50%, 75%, and 100% concentrations of 500 mM ammonium acetate. Each reversed phase gradient used an "A" solvent of 5% acetonitrile / 0.1% formic acid and a "B" solvent of 80% acetonitrile / 0.1% formic acid. The gradients themselves lasted 100 minutes, ramping from 88% A, 12% B to 45% A, 55% B. The final cycle of the MudPIT increased to a maximum of 30% A, 70% B over 110 minutes

to elute the most hydrophobic peptides from the column.

Two different Thermo Finnigan LCQ Deca ion trap mass spectrometers (San Jose, CA) were used for the soluble fraction samples, and one of the pair was used for analyzing the pellet sample. XCalibur instrument control software, version 1.2, handled the instruments. The software's "dynamic exclusion" feature reduced the extent to which high abundance peptides were sampled preferentially over low abundance ones, thus increasing the instrument's effective sensitivity. A normalized collision energy of 35% was applied to the peptides, a setting which typically fragments all of the precursor peptide ions.

### *4.2.3 Preliminary informatics*

The SEQUEST algorithm [8] was used to process the spectra against the yeast open reading frame database [9]. The program was not configured to search specifically for tryptic peptides, so all peptides from the yeast genome were considered in assessing the identifications. The algorithm was configured to assume that all cysteine residues had been modified by reduction and alkylation. The configuration allowed matching to database peptides with masses three Da higher and three Da lower than the observed peptide mass.

The DTASelect algorithm [10] filters SEQUEST results and assembles protein-level information from peptide data. Its default settings indicated that more than 2500 proteins were identified in the assembled results of the three MudPIT analyses. The program's filtering capacity was used in two stages to isolate identifications for further analysis. The first filtering step was designed to isolate peptides coming from reliably identified proteins. SEQUEST's primary score for each peptide identification is the XCorr, a measure of how well the theoretical spectrum cross-correlates to the observed spectrum. XCorr cutoffs were found which would retain the top 10% of singly and triply charged peptides and 25% of the doubly charged peptides. The XCorr cutoffs for each charge state corresponding to these percentages were as follows: 1.699 (+1), 2.290 (+2), 3.083 (+3). Only spectra for which the first sequence match scored at least 8% better than than second were retained (corresponding to a SEQUEST-assigned "DeltCN" of 0.08). When multiple copies of a particular spectrum were found, DTASelect retained only the one with the highest XCorr. Two different sequence identifications were required for any protein to be retained. Cumulatively, this

filtering reduced the set of 129,282 original spectra to 6,417.

The second filtering step was designed to reduce the pool of identifications to the subset targeted for statistical analysis. DTASelect's charge filters isolated the doubly charged peptides remaining after the first round of filtering. Of these, only the peptides with more than 50% of expected ions appearing were retained. In addition, peptides were required to have sequences ending in arginine or lysine, with no internal arginine or lysine residues. The second round of filtering selected a final set of 1465 spectra.

A new algorithm, entitled DaughterDB, created databases describing these spectra [11]. The program creates three reports. The first includes a row of information for each spectrum, including the number of peaks appearing in the spectrum and the percentage of total intensity from singly charged fragment ions. The second report includes a row for each theoretical fragment ion expected within the scan range of each spectrum, describing the most intense matching ion in the experimental spectrum within 0.75 m/z to either side of the calculated fragment ion position. Ion intensities were reported as a fraction of the sum of the spectrum's intensities in order to normalize the intensity differences between spectra (see Figure 4.1 for examples). Although DaughterDB enumerates both singly and doubly charged fragment ions, only the singly charged fragments were included. Fragment ions of fragment ions are not identified by DaughterDB. The energy applied to the peptide precursors was sufficient to remove these ions' peaks completely from the fragment ion mass spectra. The third DaughterDB report is similar to the second but enumerates pairs of consecutive fragment ions which are both within the scan range. These data were analyzed statistically in a free statistics package called R [12] which is based upon Insightful's S and S-Plus software.

## *4.3 Results and discussion*

### *4.3.1 Peptide diversity and representation*

The 1465 doubly charged tryptic peptides selected for this study had a median length of 15 residues. The middle 50% of peptide lengths ranged from 13 to 18 residues; the shortest peptide was seven residues, and the longest was 27 residues. The median peptide mass was 1643 Da. The middle 50% of the masses ranged from 1397 to 1945 Da, while the full range stretched from 843 to 2747 Da. The median XCorr assigned by SEQUEST was 3.64. The middle 50% of XCorrs ranged from 3.12

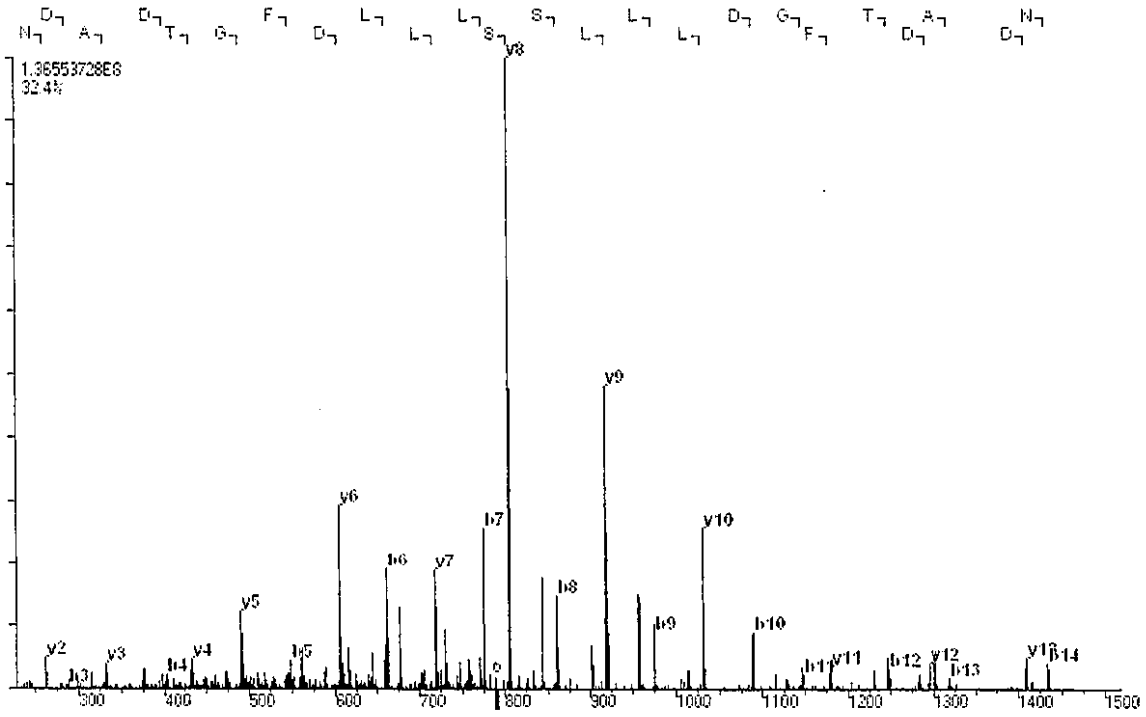



Figure 4.1: This mass spectrum represents the sequence AVDDFLLSLDGTANK. The $y5$ ion corresponds to the intensity of the median $y$ ion for all spectra in this analysis. The base peak in the above spectrum, which was identified as $y8$, embodies 9% of the spectrum's total intensity. Identified $b$ and $y$ ions account for 32.4% of this spectrum's total intensity.

to 4.28, with a minimum 2.29 (the cutoff applied by DTASelect) to a maximum of 6.63.

Figure 4.2 compares the composition of the full *Saccharomyces cerevisiae* database to the peptide content of the identifications included in this analysis. Each peptide was required to terminate in a basic residue; 650 (44%) had C-terminal Arg residues, while 815 (55%) ended in Lys. Amino acid residues with alkyl side chains (Leu, Val, Ala, and Ile) were the four most common residues found in the peptides.

The cross-correlation scores produced by SEQUEST for this set of peptides correlated to both peptide mass (correlation coefficient r = 0.59) and sequence length (r = 0.61). DaughterDB's report included several measures coordinating SEQUEST's match to the peptide. SEQUEST scores correlated most closely (r = 0.66) to the length of the longest contiguous series of singly charged fragment ions for each spectrum. SEQUEST appears to give higher scores to longer sequences and more massive peptides, especially favoring spectra with long contiguous series of matching fragment ions.

### *4.3.2 Series-specific characterization*

The cleavage of a peptide at an amide bond in low-energy CID results in several series of ions (see Figure 4.3). The *b* and *y* ions directly result from cleavage, while *a* ions result from the formal loss of carbon monoxide from *b* ions. While *b* and *y* ions are found at the vast majority of locations where they are predicted, *a* ions are less common (see Table 4.1). The energy of fragmentation in low-energy CID is insufficient to break the bond between the alpha carbon and the carbonyl, and so *x* ions are not typically produced; in this study they are included only as a measure of background noise. In these spectra, predicted *x* ions can be matched to observed peaks 23% of the time, indicating that noise peaks in the spectra may be contributing substantially to the percentages of identified ions. Ions predicted to fall outside the observed scan range were excluded from this analysis; because the spectra were collected on an ion trap mass spectrometer, low m/z peaks were truncated from the spectrum. On average, the smallest m/z in each spectrum was 30% of the precursor's m/z.

In this spectral collection, *y* ions were found only slightly more often than *b* ions, but their peaks were typically more than twice as intense. The distributions of each series' intensities are shown in Figure 4.4. The percentage of intensity in each spectrum accounted for by *b* and *y* ions

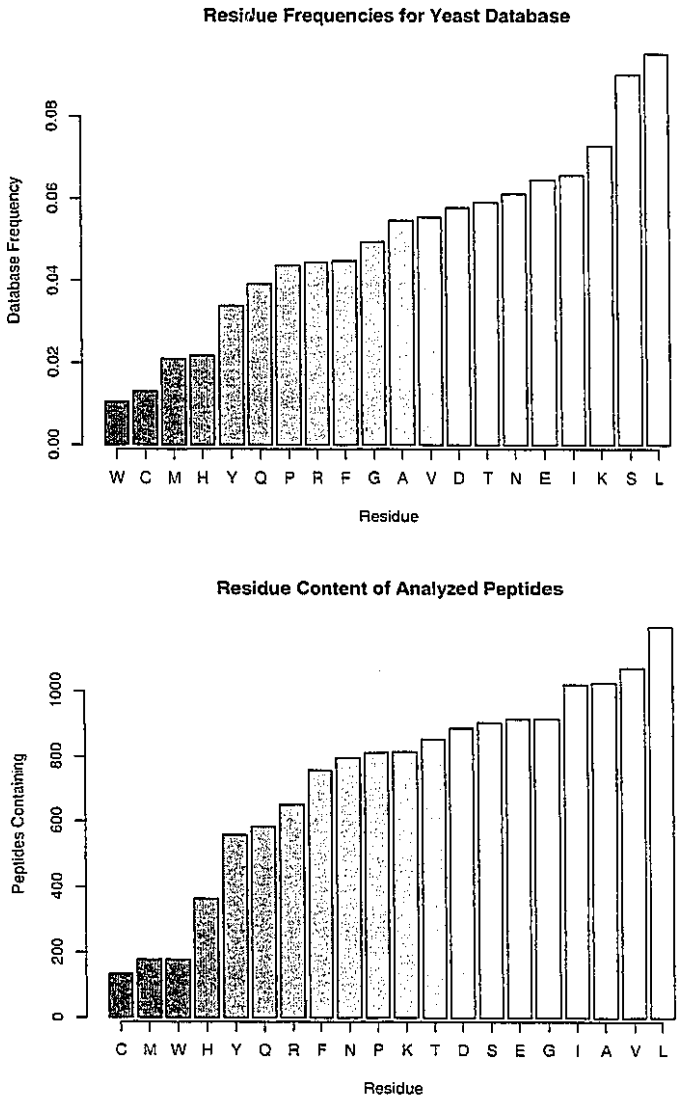



Figure 4.2: The peptides included in this analysis have a somewhat modified composition with respect to the sequence database by which they were identified. The residues identified most often among the peptides had alkyl sidechains. The six rarest residues in the identified peptides were the same as those seen least often in the sequence database. Cysteine was present in only 134 peptides, and Met and Trp were each present in 178 peptides.

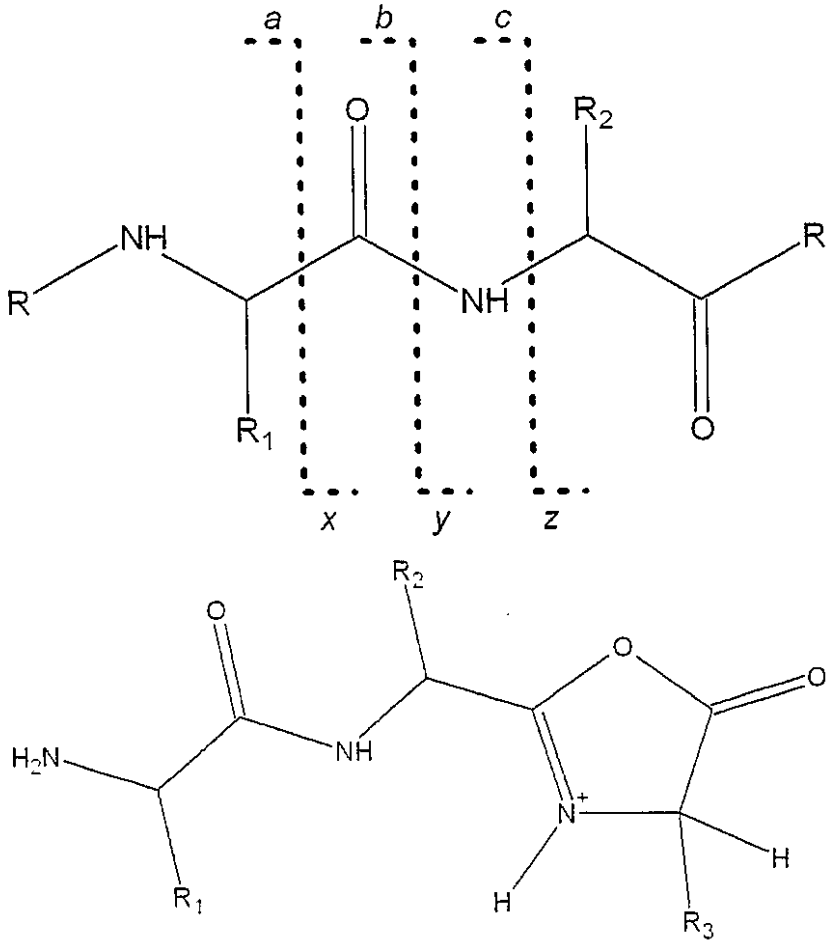



Figure 4.3: Low energy CID primarily cleaves peptide bonds, resulting in *b* ions (which contain the N-terminus and the atoms to the left of the dotted line) and *y* ions (which contain the C-terminus and the atoms to the right of the dotted line). *b* ions (pictured) generally take on an oxazolone structure, which may subsequently fragment to produce smaller *b* ions or lose carbon monoxide to form *a* ions. The remaining possible backbone ions (*c*, *x*, and *z*) do not typically form under low energy conditions.

Table 4.1: Peptide fragment ion series comparison. "Number Found" shows the number of predicted ions for which a peak was observed within 0.75 of predicted m/z. "Percent Found" compares the number of ions found to the number expected within the scan range for each ion series. The median intensity for the ions identified for each series appears under "Individual Intensity" (leaving out ions that could not be matched to an observed peak). "Spectrum Intensity" shows the average percentage of each spectrum's intensity that can be accounted for by the ions from each series. The *x* ion series is included as a measure of noise; these ions are not expected to form in low-energy CID.

| | Number Found | Percent Found | Individual Intensity | Spectrum Intensity |
|---|---|---|---|---|
| a | 6863 | 40% | 0.20% | 1.98% |
| b | 14823 | 84% | 0.47% | 8.59% |
| y | 15729 | 90% | 1.03% | 21.82% |
| x | 3971 | 23% | 0.15% | 0.75% |

was calculated. The median percentage was 28.8%, with the range between the 25th and 75th percentiles spanning from 23.1% to 35.9%. See Figure 4.1 for a sample spectrum in which 32.4% of the intensity was accounted for by *b* and *y* ions. In its search for each fragment ion, DaughterDB counts only the largest peak found within 0.75 m/z of the calculated position. As a result, peaks representing isotopic variants of the fragment ions will not be counted as part of the identified intensity for each spectrum. Low-intensity peaks can account for a sizeable portion of the spectrum's intensity because they are numerous. The remaining peaks may include unfragmented precursor ion, neutral losses from the fragment or precursor ions, or other rearrangement ions.

Figure 4.5 illustrates the relationship between fragment peak intensity and relative mass. The relative mass for each fragment ion is its mass divided by that of the intact peptide. The peaks for the *y* series reach a distinct peak in intensity when the fragments are approximately two-thirds the mass of the precusor ion. The *b* series peaks are highest at approximately 45% the mass of the precursor. The *a* series is excluded from this figure because fewer than half of these ions can be associated with spectral peaks outside the region ranging from 30% to 50% the mass of the precursor. The regions where the peaks are most intense are also the regions in which the intensity variation is greatest.

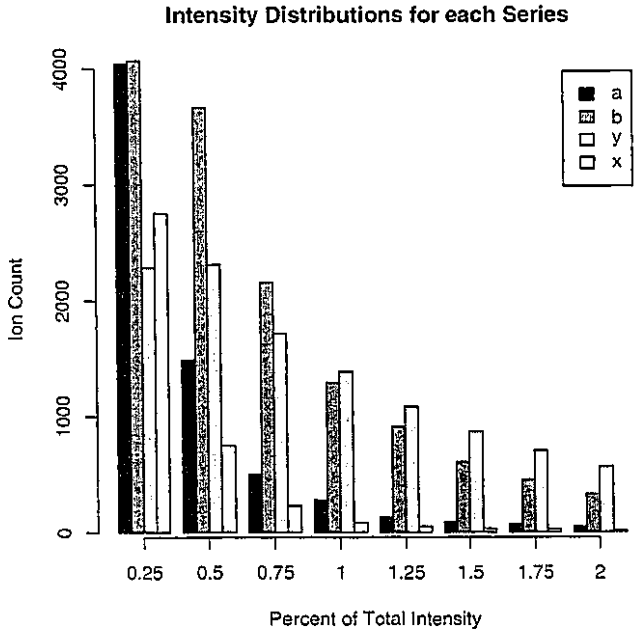



Figure 4.4: Small peaks are more common than intense ones for each series, but the more intense the peak, the more likely it is to represent an ion from the $y$ series. The distribution of peak intensities extends beyond the most intense category in this graph: 31% of $y$ peaks are more intense than 2% of the spectrum's summed intensity, as compared to 9% of $b$ peaks, 3% of $a$ peaks, and 1% of background ($x$) peaks.

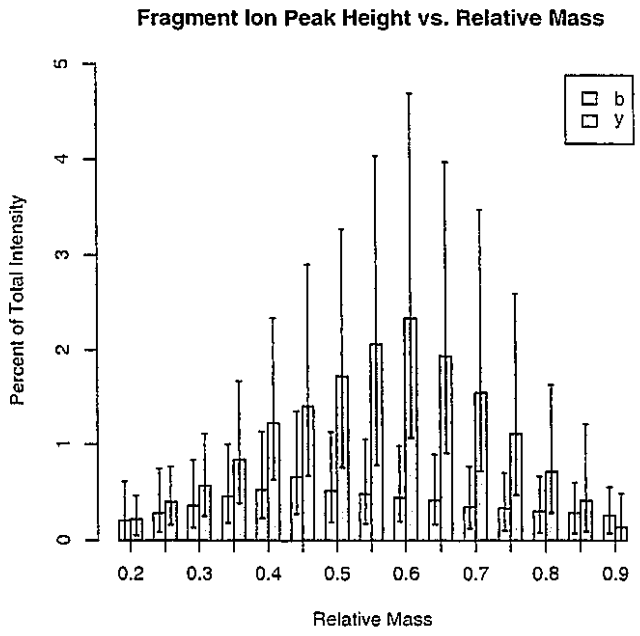



Figure 4.5: Peak intensities are related to the relative masses of the fragment ions they represent. The horizontal axis gives masses of fragments as a proportion of precursor mass. The bar shows the intensity of the median peak for the collection of ions in a particular relative mass bin, with a line extending above and below to show the 75th and 25th percentile intensities. Missing peaks are assigned intensities of zero. The *y* series shows a distinct peak at approximately 60% of precursor mass, while *b* peaks crest at 45%.

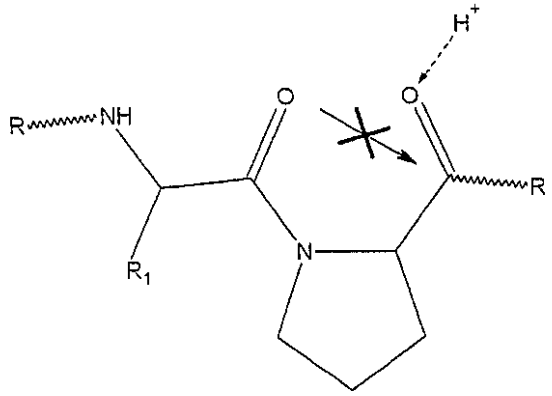



Figure 4.6: The chemical structure of Pro produces an unusual fragmentation. Because Pro's side chain forms a ring to its nitrogen, attack on its carbonyl carbon by the preceding carbonyl oxygen would result in a strained 5-5 bicyclic ring. Cleavage to its C-terminus is reduced, and cleavage to its N-terminus is encouraged, yielding a large differential between the fragment ion peaks adjacent the residue.

### *4.3.3 Residue-specific behavior*

*N-bias by residue*

Individual amino acid residues influence which of the two adjacent amide bonds (N-terminal or C-terminal) break in the process of collision-induced dissociation. The structure of Pro, for example, prevents the cleavage of the peptide bond C-terminal to the residue by hindering the attack of the N-terminal carbonyl [13] (see Figure 4.6). The extent to which each residue directionally enhances cleavage was measured by comparing the intensities of fragment ion peaks adjacent to each residue. The intensity of the C-terminal fragment peak was subtracted from the intensity of the N-terminal fragment peak. These differences were divided by the sum of the two peak intensities to yield the "N-bias." A subset of the fragment peaks was calculated to fall within the scan range but did not appear in the tandem mass spectra. Each of these peaks was assigned an intensity of zero and included in the analysis. If neither peak was observed for a pair, it was excluded. If either of the two fragment ions fell outside the scan range, the pair was excluded from the analysis.

Examples from the spectrum shown in Figure 4.1 may clarify this measure. Three *y* ion peak pairs show the presence of Leu residues. In all three cases, the lower m/z peak is more intense than

the higher m/z peak ($y6 > y7$, $y8 > y9$, and $y9 > y10$). For *y* ions, the higher m/z peak of each pair corresponds to the N-terminal cleavage. All three examples of Leu in this spectrum yield a negative N-bias (or positive "C-bias"); the C-terminal peaks are more intense than the N-terminal peaks. This spectrum's results for Leu are in line with the median behavior observed for the broader set of spectra, which showed an N-bias of -0.33 for this residue.

For *b* ions, the three residues with the highest N-bias were Pro, Gly, and Ser (see Figure 4.7). At the median, Pro's N-bias was 1.0; the peak representing the C-terminal cleavage product did not appear in the spectrum. Gly and Ser were more moderate influences on fragmentation, with N-biases of 0.47 and 0.35, respectively. Showing the opposite impact was His, which yielded a C-bias of -0.64. The phenomenon, however, was quite variable; the mean interquartile difference for *b* ion N-biases was 0.86. The variability for residues did not appear to bear a relationship to sample size: the interquartile differences for Cys, Met, and Trp were not substantially different than for the other residues.

Figure 4.7 shows the N-bias measurements for *y* ions. The highest N-bias residues for *y* ions were Pro (0.93), Gly (0.59), and Ser (0.35). Although His was the strongest C-bias influence in *b* ion intensities, it appeared to have little effect in the *y* ion series. A C-bias was found for Ile (-0.45), Val (-0.44), and Leu (-0.33). The average interquartile difference for *y* ion N-biases was 0.74. The variability may have decreased relative to the *b* series due to the greater intensity (and thus signal-to-noise) of *y* series peaks. A visual inspection of measured N-biases and cleavage position did not reveal a relationship between the two.

Information from the literature of peptide dissociation mechanisms may explain the strong N-bias calculated for Pro cleavage and the strong C-bias calculated for His cleavage in *b* ions. Pro is the only cyclic amino acid of the 20 commonly occurring amino acids. It has been noted in the literature that there is a strong preference for Pro to cleave at its N-terminal side, whether the product formed is the *b* ion or *y* ion [14]. Because *b* ions formed at most residues are accepted to have oxazolone structures (see Figure 4.3) [15] and because the *b* ion that would be formed at the C-terminal side of Pro would require a transition-state involving formation of an unstable strained 5-5 bicyclic ring (see Figure 4.6) [16], cleavage typically occurs at the N-terminal side of Pro rather than the C-terminal side. This selective cleavage also provides information on the mechanism of the cleavage to form *b* and *y* ions. If the mechanisms of formation of *b* and *y* ions have no common intermediates, one would not

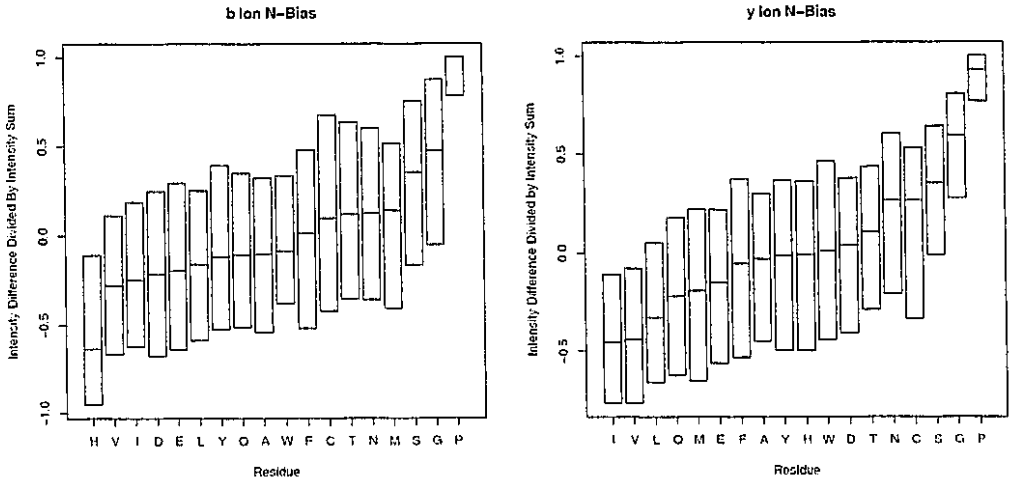



Figure 4.7: N-bias measures the extent to which each residue directs local fragmentation. The statistic measures the difference between the N-terminal peak intensity and the C-terminal peak intensity, normalizing this difference by the sum of the two peak intensities. Residues with N-biases of greater absolute value impact local fragmentation to a greater extent. The median bias for each residue is marked by the line across each box, and the upper and lower edges of the boxes represent the 75th and 25th percentiles, respectively. The most distinctive bias toward N-terminal fragmentation is that of Pro. A smaller N-bias appears for Gly and Ser. Hydrophobics Ile, Leu, and Val show a bias toward C-terminal cleavage in *y* ions, and His shows a pronounced bias towards C-terminal cleavage in *b* ions.

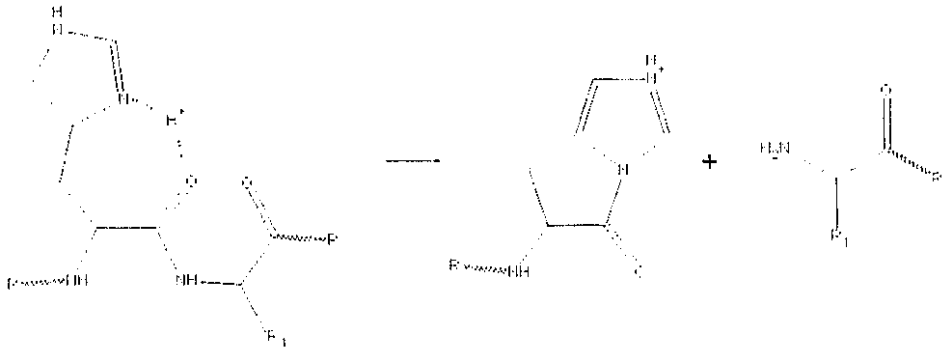


Figure 4.8: Histidine can form unusual *b* ions. In normal fragment ion formation, the carbonyl N-terminal to a residue attacks the carbonyl at the residue's C-terminus, resulting in an intermediate which produces a *b* ion with a single, five-membered ring. His' side chain may short-circuit that process by attacking its own carbonyl, yielding a *b* ion with a double ring structure.

expect that both the *b* and *y* ions would show a strong N-bias. The strong N-bias shown in Figure 4.7 is indirect evidence that a coupled *b*-*y* cleavage mechanism [17] occurs for Pro: ring closure to form an oxazolone N-terminal to Pro leads to an ion-molecule complex that can dissociate to either a *b* or a *y* ion, depending on whether a proton transfer to the Pro occurs before the two fragments separate.

The C-bias for fragmentation at His has not been reported directly in the literature for a large body of compounds although mechanistic data have been reported that are consistent with this bias. Wysocki and coworkers suggested that fragmentation at His occurs via formation of a *b* ion structure resulting from the side chain's attack on the backbone carbonyl (see Figure 4.8) [18]. O'Hair and coworkers calculated the stability of this 5-5-ring bicyclic structure for the N-acetylated methyl ester of His and found that this structure is more stable than the corresponding protonated oxazolone [19]. A C-bias for cleavage at His that is greater than that at the other residues in Figure 4.7 is consistent with the suggestion that the His side chain produces a unique *b* ion structure not formed by the other residues in the analysis [20]. To test the hypothesis that other basic side chains may attack the neighboring carbonyl, a set of peptides including internal Arg and Lys residues was produced. The C-bias exhibited by Arg was comparable to that of His (data not shown), suggesting that its side chain may also be capable of nucleophilic attack on the carbonyl oxygen, as suggested by Glish [21].

*Neutral loss by residue*

Neutral losses from fragment ions can complicate tandem mass spectra. A loss of ammonia either before or after fragmentation may reduce fragment ion masses by 17 mass units (often called "*" ions). The median intensity among observed *b*-17 ions was 0.24% of the spectrum's intensity, corresponding to approximately half the intensity of the median *b* ion. 55.5% (9744) of predicted ions matched to a peak. The median intensity for observed *y*-17 ions accounted for 0.18% of the spectrum's intensity; the average *y* peak is five times the size of its ammonia loss peak. Of the predicted ammonia losses, 44.7% (7850) could be matched to observed peaks.

To determine the influence of residue content on the production of these neutral loss ions, the fragment ions associated with neutral loss peaks more intense than the observed median were segregated from the other fragment ions of the series. The residue content of these ions was compared to the residue content of fragment ions of the same type. The *b5* ion of ASGEIVSINQINEAHPTK, for example, has the sequence ASGEI. If this ion showed a neutral loss of 17 mass units more intense than the median, this *b5* ion would contribute to the composition ratios of Ala, Ser, Gly, Glu, and Ile. The results for Ala may clarify this process. The residue is found in 49.3% of predicted *y* ions (8650 of 17538). Of the *y* ions with intense identifiable ammonia loss peaks, 48.3% (1894 of 3925) contain Ala. The frequency of Ala content for ions with substantial ammonia losses is not substantially different from the frequency of Ala content for all ions. There appears to be no relationship between Ala content and neutral loss of ammonia. The graphs in Figures 4.9 and 4.10 show the ratios of these percentages, so that a ratio of 1.0 indicates that a particular residue was found neither more nor less often than expected.

For *b* ions, the result shows that prominent ammonia-losing fragment ions are 32% more likely to contain Asn and 7% more likely to contain Gln than *b* ions in general (see Figure 4.9). On the other extreme, His (-30%) and Pro (-29%) are underrepresented in *b* ions that lose ammonia. The reliability for Met, Trp, and Cys may be limited by the number of peptides containing these residues (178, 178, and 134, respectively).

The case of *y* ions shows similarities and differences to that of *b* ions (see Figure 4.9). Fragments that are 17 Da less massive than *y* ions are 28% more likely to contain Asn and 26% more likely to contain Gln than *y* ions in general. Ions from the *y* series that lose ammonia are 16% more likely to

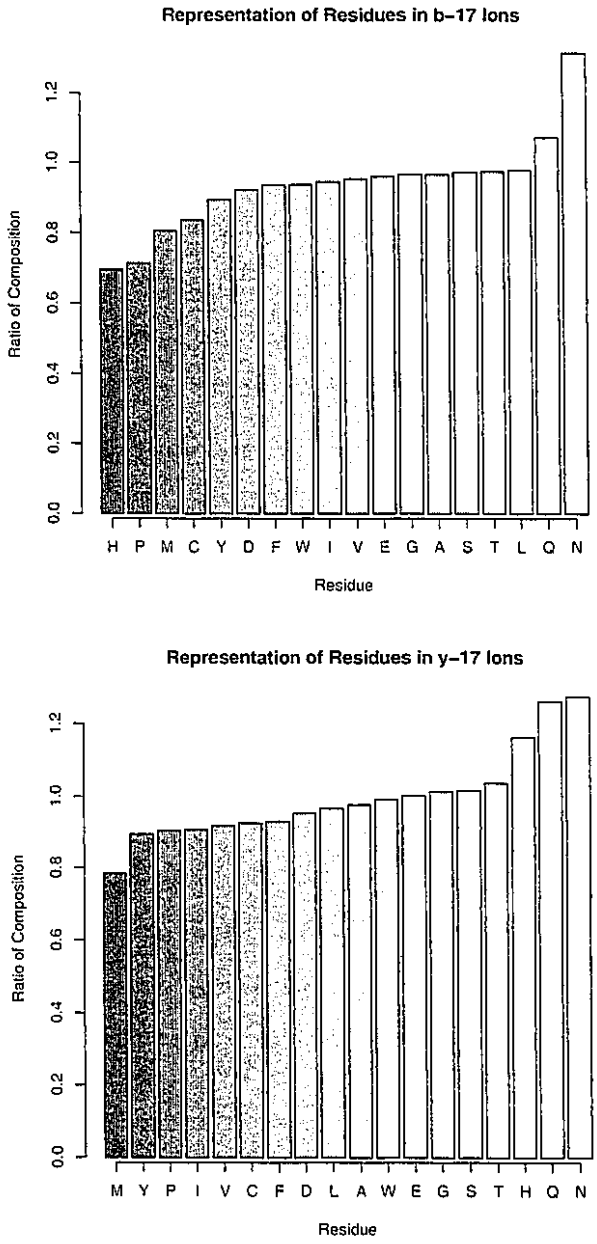



Figure 4.9: Fragment ions which exhibit prominent loss of ammonia are more likely to contain Asn or Gln than fragment ions in general. Ammonia loss is enhanced by His for $y$ ions, but His suppresses the loss for $b$ ions. Neutral loss of ammonia may be diminished by the presence of Pro and Met. The vertical axis shows the ratio between the sequence composition of the ions displaying intense loss peaks and the sequence composition of fragment ions from the appropriate series.

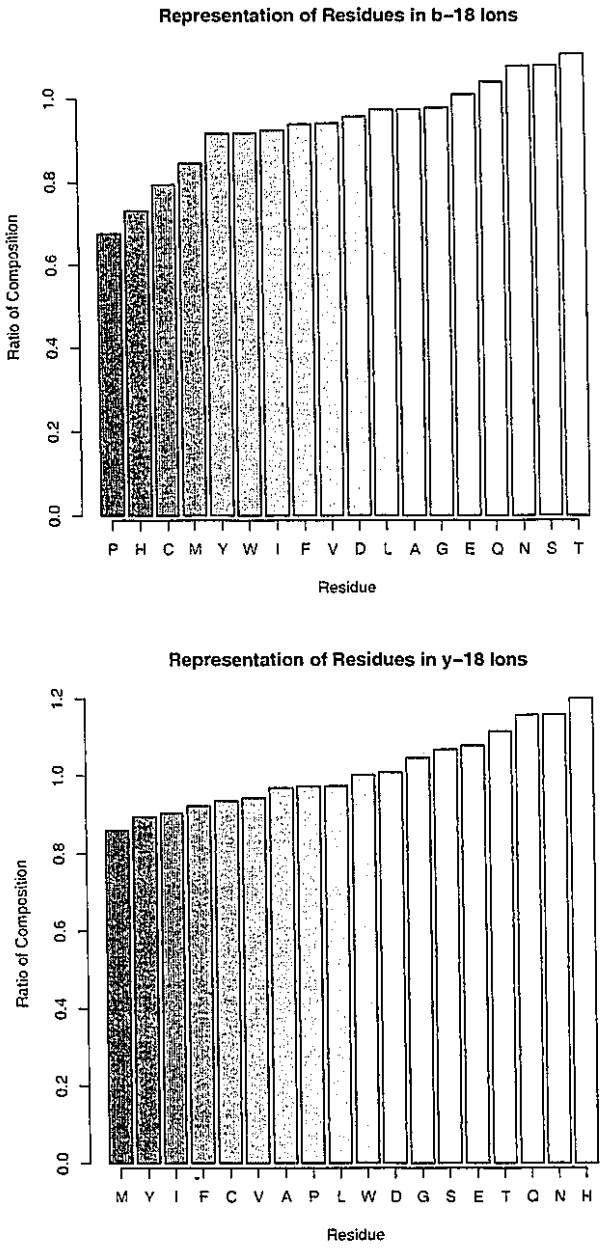



Figure 4.10: Fragment ions which exhibit prominent loss of water are more likely to contain Ser, Thr, or Glu. As seen in ammonia loss, His appears to enhance loss in *y* ions but diminish it in *b* ions. The mass accuracy of the ion trap, combined with the window for identifying peaks in DaughterDB, may result in the misidentification of some ammonia loss peaks as water loss peaks, thus resulting in Asn and Gln's ranking above.

contain His, the opposite effect seen in *b* ions. The effect of His was observed to depend upon the C-terminal residue of the peptide; fragment ions ending in Lys and containing His were more likely to produce neutral losses than those ending in Arg (data not shown). Pro content remains lower (-9%) in ammonia losing *y* ions, but not to the extent seen in *b* ions.

Fragment ions can also lose water molecules, a mass shift of 18 Da (sometimes referred to as "°" ions). Although ion trap mass spectrometers putatively feature unit mass resolution, some amount of conflation between ammonia and water loss ion peaks is to be expected. The median intensity of observed *b*-18 ions was 0.15% of the spectrum's intensity, with 66% (11,612) of calculated *b*-18 ions matching to a peak. The corresponding intensity for *y*-18 ions was 0.10%, with 51% (9025) of these ions matching to a peak. The intensity medians for ions showing a loss of 18 mass units are well below the corresponding medians for the ions exhibiting 17 mass unit losses, though a larger proportion of these calculated ions match to observed peaks.

The differences in residue content for water loss ions do not appear to be as marked as those in ammonia losses. Intense *b*-18 ions are 11% more likely to contain Thr and 8% more likely to contain Ser than *b* fragment ions in general (see Figure 4.10). Asn and Gln are also more common among these ions (8% and 4%, respectively). The appearance of these residues for both water and ammonia losses may have resulted from the limited mass accuracy of the ion trap, which can yield m/z measurement errors sufficient to cause DaughterDB to identify a particular peak as both a water loss and as an ammonia loss. Another point of similarity between both classes of neutral loss is that both *b*-18 ions and *b*-17 ions diminish by the content of Pro or His.

Compositions of *y*-18 ions do not appear to be substantially different from those of *y*-17 ions (see Figure 4.10). Again, His is prominent in ions losing water (an increase of 20%). As with ammonia losses, His content's contribution to water neutral losses was observed to be dependent upon Lys' presence at the C-terminal residue of the peptide. Asn and Gln are also common (16% each). Thr and Ser, which are found preferentially in *b*-18 ions, are among the residues increasing the presence of *y*-18 ions, but they do not stand out substantially.

Peptide dissociation trends reported in the literature can help explain and support the dependence of neutral losses of water and ammonia from sequence ions on amino acid residue composition. The formation of sequence ions competes with neutral loss ion formation due to side chain interactions. Research has corroborated the relationship of Asn and Gln content to *b*-17 and *y*-17 ion formation

(see Figure 4.9). The roles of Ser and Thr content in the formation of *b*-18 ions and of His in the formal loss of water from *y* ions (see Figure 4.10) have also been explored.

The reported loss of ammonia from Asn and Gln side chains in peptides corresponds to reported results for individual amino acids. O'Hair and coworkers [22] reported the fragmentation patterns for a variety of protonated N-acyl amino acid methyl esters, including Asn and Gln. Both of these ions underwent the neutral loss of ammonia readily, the source of which could not be from the N-terminus due to N-acylation. The only remaining possible sources of the ammonia loss were the side chains of Asn and Gln residues.

Gaskell and coworkers [23] investigated the neutral loss of water from singly-protonated peptide ions and provided evidence for three dehydration reactions through [$^{18}$O]-labeling studies: (i) loss of water from the C-terminal carboxylic acid or acidic residues, (ii) loss of water from the side chain hydroxyl group of Ser or Thr residues, and (iii) loss of water from the amide carbonyl of the peptide bond. Their study concluded that a single peptide may follow any of several competing dehydration pathways. Due to the variety of peptide sequences included in the statistics reported here, some residue-specific trends may be masked by the presence of other amino acids. All tryptic peptides included in this study contained a C-terminal carboxylic acid and several amide carbonyls in peptide bonds, complicating the determination of which pathway led to the formation of observed water loss ions.

Water loss may precede the formation of fragment ions. Separate [$^{18}$O]-labeling MS/MS experiments of Thr-containing peptides and corresponding *ab initio* calculations for these molecules have indicated that the neutral loss of water occurs from the precursor ion [M+H$^+$] exclusively from the side chain hydroxyl group [24]. The interaction of Ser's side chain with the backbone is a potential explanation for the neutral loss of water in low energy CID MS/MS experiments. The currently accepted mechanism for protein splicing could be considered a solution phase counterpart to this mechanism in which the hydroxyl oxygen of the Ser side chain attacks the carbonyl carbon to form a five-membered intermediate [25]. The presence of His in a peptide sequence has also been reported to enhance the neutral loss of water from sequence ions, perhaps through a neighboring group pathway involving induction of water loss by the nucleophilic side chain on the adjacent protonated carbonyl [26]. The presence of His in peptides could promote the neutral loss of water (as seen in Figure 4.10).

## *4.4 Conclusion*

The intensities of fragment ion peaks are a function of their fragment types, the locations of cleavage within the peptide, and the residues that are adjacent to the cleavage. The residue content of fragment ions, in turn, plays a role in the formation of secondary fragment ions such as ions formed by the neutral loss of water or ammonia. While intensity may be quite variable, particularly for the most prominent peaks in spectra, the information encoded in fragment ion intensity should be used in algorithms which process these spectra. The trends observed in fragment ion intensity can be developed into improved models of peptide fragmentation.

Existing algorithms for identifying database peptides from tandem mass spectra employ simple fragmentation models to generate spectra for comparison to observed spectra. SEQUEST, for example, models all *y* ions to be a uniform intensity, with *b* ions a uniform lower intensity. The above results establish that even spectra that SEQUEST can correctly identify show distinctive patterns in intensity for each ion series and marked differences in fragmentation neighboring particular amino acid residues. An algorithm that uses these trends in intensity as part of its fragmentation model should offer improved accuracy in peptide identification, correctly identifying spectra that deviate from simple fragmentation models.

The challenge of inferring sequence directly from tandem mass spectra (*de novo* rather than from a database) requires that all information present in the spectrum be used optimally. The measures calculated in this research can be applied directly to this problem. Instead of relying solely upon m/z data to consider sequence possibilities, algorithms based on improved fragmentation models can use the information present in recorded fragment ion intensity to aid the process of sequence validation.

This analysis is limited to spectra for doubly-charged peptides resulting from a complete trypsin digest. While this class comprises the identifications most common in proteomic experiments, other classes of peptides are also significant and produce distinct intensity patterns. Triply charged peptide spectra, in particular, differ in having doubly charged fragment ions intermixed with the singly charged ones. The locations of basic residues within the peptide sequence instead of or in addition to the terminal basic residue of tryptic peptides can also change the pattern of intensities observed in a spectrum. Characterizing these classes of spectra will help highlight the mechanisms by which

peptides fragment and set the stage for a second generation of sequence identification software.

Chapter 5

# SIMILARITY AMONG TANDEM MASS SPECTRA FROM PROTEOMIC EXPERIMENTS

## *5.1 Introduction*

Liquid chromatography paired with tandem mass spectrometry can produce thousands of fragment ion spectra for a complex mixture of peptides [1]. The mass spectrometer's control software catalogs intact peptide ions eluting from chromatography to produce an MS scan, isolates ions of a particular peptide for fragmentation, and collects the produced fragment ions in a fragment ion (MS/MS) spectrum. The control software attempts to prevent the repeated isolation and fragmentation of particular peptides in order to increase the diversity of spectra acquired. Thermo Finnigan's "dynamic exclusion" feature [2], for example, maintains a list of the precursor m/z values which have been fragmented during the last several seconds. Peptide ions that are listed will not be fragmented. Fragment ion spectra may be repeated despite these features. For example, peptides which elute over a period of several minutes exceed the duration of exclusion and may be duplicated. In addition, a peptide mixture may be of sufficient complexity to yield more peptide ions within the exclusion duration than the list can hold. As a result, some degree of duplication can be expected among the MS/MS spectra for these experiments.

The spectra captured during an experiment pass through several steps before sequence identification. The instrument control software will first compose the spectra by averaging multiple microscans and centroiding the peaks. The MS/MS spectra must then be separated from the MS spectra produced in the process of liquid chromatography. For SEQUEST [3] users, this task is completed by the ExtractMS program [4], which filters out unusable spectra on the basis of criteria such as peak count and separates singly-charged peptide MS/MS spectra from those resulting from multiply charged peptides. A recently published program, 2to3, extends on ExtractMS by determining the charge state for multiply-charged spectra [5]. With these processes complete, the SEQUEST algo-

rithm can begin its task.

MS/MS spectra from the same peptide may differ from each other for several reasons. If a higher concentration of a peptide is present at one isolation than at another, the signal-to-noise difference between the spectra may cause one to be of higher quality. Variations in collision energy may produce different levels of peptide fragmentation. In addition, the mass spectrometer's detector may register higher or lower intensities due to random noise. All of these sources may yield spectra that appear different even though they represent the same peptide precursor ion.

If spectra show significant similarity to each other, they generally represent the same peptide. The intensities of individual peaks may vary considerably from spectrum to spectrum, but the m/z values of fragment ions can be measured to within a single m/z in most mass spectrometers. If the primary fragment ions in a pair of spectra are at the same m/z locations but vary in intensity, the spectra can be judged as resulting from the same peptide. In some cases, such as the presence of labile post-translational modifications, neutral losses from the precursor ion are the dominant peaks in spectra, reducing the amount of intensity to be found in informative fragment ion peaks. Such spectra may appear as quite similar while representing different peptides.

Identifying peptides from spectral collections is often the rate limiting step for proteomic experiments. Peptide identification algorithms may consume several seconds or minutes for each spectrum. If duplication is ubiquitous among proteomic data, peptide identification algorithms should be modified to explicitly recognize and handle similarity. The time taken to analyze spectra could be reduced by handling similar spectra simultaneously. Likewise, spectra which are similar to each other could be combined to improve the overall signal-to-noise for peptide identification. Spectral similarity has implications for both the performance and accuracy of peptide identification algorithms.

Techniques for determining the degree of similarity between spectra have been used for searching libraries of reference spectra to identify experimental ones [6]. The technique of cross correlation is employed by LIBQUEST [7] with high sensitivity, but the algorithm is CPU-intensive. The approaches used in electron ionization mass spectrometry for library searching are generally faster than cross correlation but may not be as sensitive. The dot-product comparison (also called the "spectral contrast angle") fared best in Stein and Scott's library search algorithm comparison [8]. In a study by Wan *et al* [9], the technique was shown to be effective at differentiating very similar

oligonucleotide fragment ion spectra.

The dot-product comparison builds a vector in multidimensional space for each of two spectra being compared and determines the angle between the vectors. Higher angles imply greater differences between the spectra, and angles approaching zero indicate considerable similarity between the spectra. Peptide fragment ion spectra contain more peaks than the spectra typically used with dot-product comparison. The algorithm loses discriminatory power if large numbers of peaks are included in the spectra to be compared, necessitating a peak selection process to reduce spectral complexity prior to similarity analysis. When the peptide sequence corresponding to each spectrum is known, selection of significant fragment ions is relatively straightforward. Seeking similarity among uninterpreted spectra, however, requires different means for peak selection.

We adapted the dot-product algorithm to group uninterpreted peptide tandem mass spectra by similarity. The resulting software automatically selects a subset of peaks from each spectrum for use in comparison. It infers clusters of similar spectra and can retain a representative from each group while removing duplicates. The algorithm was used to analyze spectra resulting from 1D and 2D liquid chromatography (MudPIT). We show the impact of removing duplicate spectra on SEQUEST results and propose a modification for peptide identification algorithms which would improve the speed of identification while diminishing the occurence of false positive matches.

## *5.2 Experimental section*

### *5.2.1 Materials*

All chemicals were purchased from Sigma (St. Louis, MO) unless otherwise noted.

#### *Gel band samples*

Cells of a stable HEK 293 (Human Embryonic Kidney) cell line were lysed and subjected to two-step affinity chromatography to purify a multiprotein complex [10]. A 1D gel separated the proteins, and Coomassie dye stained the complex constituents. The protein bands were cut from the gel, reduced by dithiothreitol, and alkylated with iodoacetamide. An in-gel digest with trypsin (Promega sequence-grade trypsin) eluted peptides from the gel pieces.

*Microtubule associated proteins sample*

The microtubule associated proteins (MAP) sample was purified from bovine brains by a published protocol [11]. Proteins were denatured by 8 M urea. Disulfide bridges were reduced with TCEP and alkylated with iodoacetamide. The proteins were digested initially with EndoK-C (Roche, Basel, Switzerland) and then by trypsin (Perceptive, Foster City, CA).

*Rat hippocampus sample*

Three rat brains were dissected, and regions enriched in the hippocampus were pooled, homogenized, and centrifuged [12]. Proteins were denatured with 8 M urea, and disulfides were reduced with dithiothreitol and alkylated with iodoacetamide. Proteinase K was employed to digest the proteins to peptides, and the resulting mixture was desalted.

### 5.2.2 *Separation and mass spectrometry*

The liquid chromatographic separations used for each of the three described samples differed, but the same basic materials were used. Buffer A, the low hydrophobicity buffer, was 5% acetonitrile / 0.1% formic acid. Buffer B, the high hydrophobicity buffer, was 80% acetonitrile / 0.1% formic acid. Buffer C, for producing the salt steps, was 500 mM ammonium acetate / 5% acetonitrile / 0.1% formic acid. The columns used for separations were produced from fused silica capillaries with outer diameters of 365 $\mu$m and inner diameters of 100 $\mu$m (Polymicro, Phoenix, Arizona), with tips drawn to inner diameters of 5 $\mu$m. The reversed phase separations were produced by 5 $\mu$m Aqua C18 material (Phenomenex, Ventura, CA), and 5 $\mu$m Partisphere strong cation exchanger (Whatman, Clifton, NJ) separated peptides into steps for the MudPIT analyses.

The gel band protein digests were analyzed by 1D liquid chromatography. The columns contained 7-10 cm of reversed phase material. A Surveyor pump (Thermo Finnigan, San Jose, CA) produced the 35 min gradients. An LCQ ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) acquired spectra for the sample.

The MudPIT separations employed three phase columns [13]. Each consisted of 7 cm of reversed phase material, 3 cm of strong cation exchanger, and 3 cm of material for the initial loading of the peptides. The MAP sample was separated in six cycles of chromatography, produced by a Thermo

Finnigan Surveyor pump (San Jose, CA). Buffer C percentages for this separation are given in Table 5.2. Peptides were electrosprayed directly into a Thermo Finnigan LCQ mass spectrometer. The hippocampal homogenate was analyzed by a twelve cycle MudPIT, driven by an Agilent 1100 quaternary HPLC pump (Palo Alto, CA). The concentrations of buffer C appear in Table 5.3. Peptides were electrosprayed directly into an LCQ Deca mass spectrometer (Thermo Finnigan, San Jose, CA). The XCalibur control software managed the acquisition of spectra on both the LCQ and LCQ Deca instruments. Dynamic exclusion was configured to maintain a list of the last 50 masses fragmented, with masses excluded for a duration of 25 MS scans. The pattern of acquisition was set to produce an MS scan followed by three MS/MS scans.

### *5.2.3 Peptide identification and assembly*

The 2to3 algorithm [14] was applied to the obtained MS/MS spectra to remove spectral copies with incorrect charge state assignments. The normalized version of SEQUEST [15] was used to identify the spectra, using monoisotopic masses for fragment ions. Because all the samples had been reduced and alkylated, a mass of 160 was used for all cysteines rather than 103 in the SEQUEST search. The gel band spectra were analyzed against the RefSeq Homo sapiens sequence database [16]. The MAP sample was identified with a custom database including 1180 proteins drawn from RefSeq. The hippocampal homogenate was matched to a database consisting of RefSeq's Homo sapiens, rat, and mouse databases.

DTASelect [17] assembled and filtered the identifications, removing spectra with normalized XCorrs below 0.3, retaining spectra that matched their identifications well and removing those that were identified poorly. This threshold passes approximately 12% of the identifications in MudPIT results. In addition, DTASelect was configured to require identifications to have sequences of at least six residues, and the top sequence score for each spectrum was required to exceed the second best score by 8%. Proteins with only one peptide passing these criteria were included; multiple peptides were not required for protein inclusion. If multiple spectral copies of the same sequence, precursor charge, and modification were retained, they were counted as a single peptide.

### 5.2.4 *NoDupe algorithm*

Software named "NoDupe" was created in the Java programming language to analyze spectral similarity. To reduce the complexity of the spectra, NoDupe preprocesses the spectra before comparing them. The spectra are grouped on the basis of their similarities, and a report is created for review via spreadsheet. NoDupe can optionally remove the duplicate spectra from each liquid chromatography run, keeping one spectrum from each cluster of similar spectra.

NoDupe reads spectra from SEQUEST's unified MS/MS file format [18]. All fragment ion spectra produced for a cycle of chromatography are read into memory. The fragment ions are assigned to bins 1.00057 m/z wide to abstract away minor variations in recorded m/z values for peaks. If two succeeding peaks fall within the same m/z bin, their intensities are added together. Each peak's intensity is normalized by the sum of the peak intensities for the spectrum. To emphasize smaller peaks, all intensities are square-rooted [19].

Because a large proportion of the peaks in peptide fragment ion spectra are very low in intensity, NoDupe removes these peaks to prevent dilution of the similarity measurements. The sum of the intensities' square roots is calculated, and the peaks are sorted in order of decreasing intensity. Peaks are accepted into the final list of peaks until their sum is greater than half the sum of square roots. Spectra in which the intensity is concentrated in very few peaks will have a larger proportion of peaks removed than those in which intensity is spread over a larger number of major peaks. See Figure 5.1 for examples of the preprocessing results.

Once all spectra are preprocessed, the scans are sorted by precursor m/z. The spectral contrast angle is computed for pairs of spectra with precursors within 3 m/z of each other. The angle equation is defined as:

$$\cdot \; cos\theta = \frac{\sum i_A i_B}{\sqrt{\sum i_A^2 \sum i_B^2}}$$

where $\theta$ is the spectral contrast angle, $i_A$ is a peak intensity from spectrum $A$, and $i_B$ is a peak intensity from spectrum $B$. In essence, if both spectra have a peak at a particular m/z, the intensities are multiplied together and added to the top sum. For all peaks in either spectrum, the intensity is squared and added to one or the other sum on the bottom. Only those peaks found in both spectra will contribute to the top sum.

If two spectra are identical, their angle will be zero, while two completely dissimilar spectra

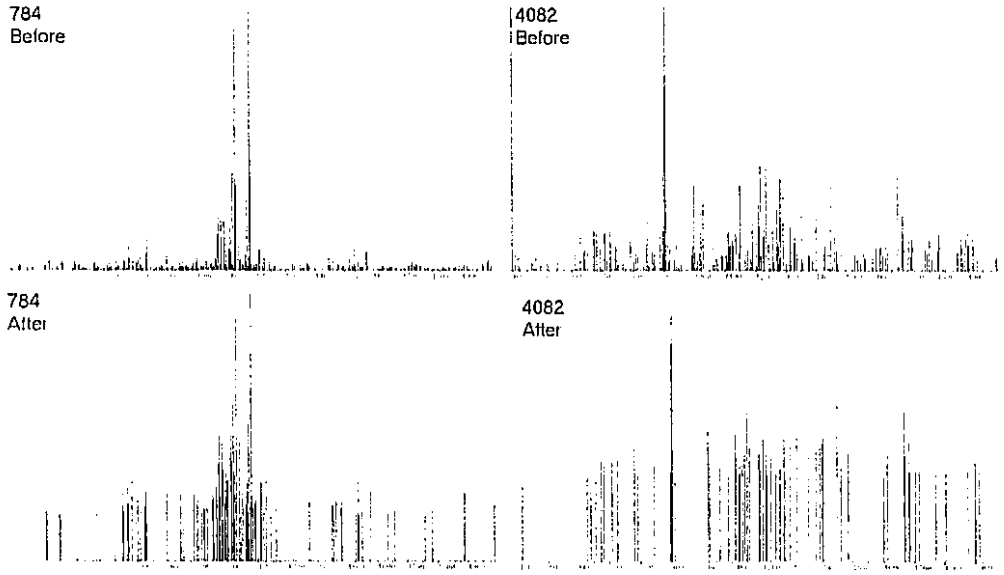
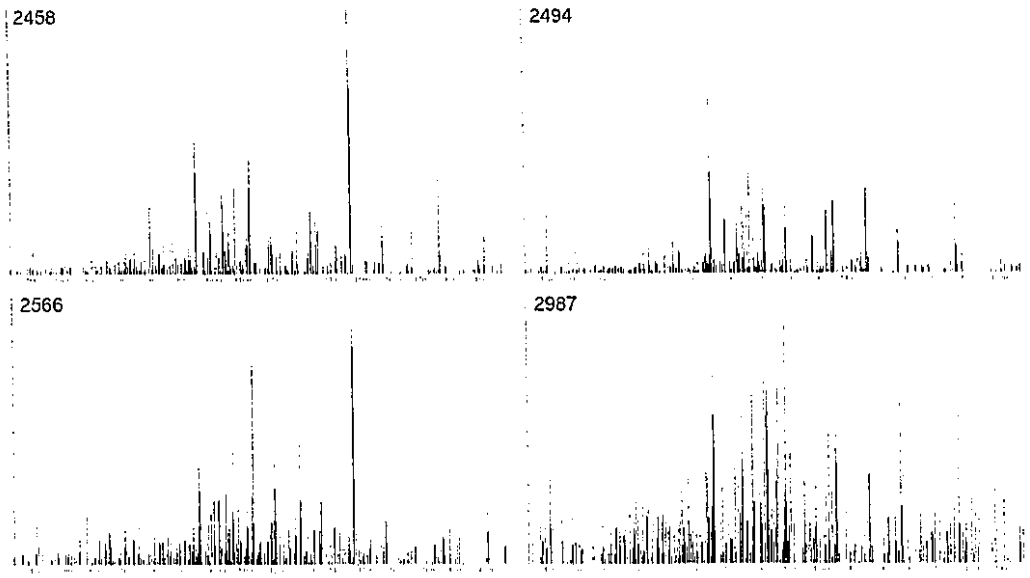



Figure 5.1: Preprocessing may alter spectra considerably. Intensities are square-rooted, resulting in increased significance for less intense peaks. Because only the most intense peaks are retained after the intensity quota, peak counts are reduced. Scan 784 reduced from 478 peaks to 120. Scan 4082 reduced from 235 to 77 peaks. All other figures show spectra prior to preprocessing.




Spectral Contrast Angles

| | 2494 | 2566 | 2987 |
|---|---|---|---|
| 2458 | 0.812 | 0.952 | 1.083 |
| 2494 | | 0.967 | 1.098 |
| 2566 | | | 1.205 |

Figure 5.2: The four spectra above were found in the sixth cycle of the MAP sample MudPIT. Scans 2494 and 2458 are the closest match with an angle of 0.812 radians. Scan 2987 is different enough from the others that it bears only marginal similarity to scans 2458 and 2494. Although 2566 rates as similar to both 2458 and 2494, it is not similar enough to scan 2987 to be grouped with it.

give a right angle ($\pi/2$ radians). The spectral contrast angle is commutative; a pair of spectra will yield the same angle whether A is compared to B or vice versa. The comparison can be written as a cosequential algorithm; the angle can be computed in time proportional to the number of peaks included. See Figure 5.2 for examples of this measure.

As shown in Figure 5.3, there was not a clear delineation between significant and insignificant spectral contrast angles. Spectra that are dissimilar to all others form a mass of high spectral contrast angles at the top of the figure, but the extent of this mass is ambiguous. Groups of spectra which SEQUEST identified as representing the same peptide sequence were analyzed visually to

determine the maximum spectral contrast angle likely to indicate genuine similarity. A similarity angle cutoff of 1.1 radians (approximately $60^{o}$) was chosen to divide significant from insignificant spectral contrast angles. A lower cutoff would be more selective about which spectral pairs are named as similar, but the normal variation of fragment ion spectra sometimes produces pairs which yield 1.1 radian spectral contrast angles.

For each spectrum, the number of spectra matching with a similarity angle below 1.1 radians is recorded as the spectrum's "match count." A spectrum is marked as a duplicate if a similar spectrum has a higher match count. If the two most representative spectra for a group have the same match counts, the one for which the larger proportion of peaks was removed during preprocessing is retained.

## *5.3 Results and discussion*

Because the extent of duplication among acquired spectra may depend upon sample type and chromatographic technique, the similarity algorithm was tried on three different sets of data. The simplest sample included 18 reversed-phase gradients on in-gel digests of 1D gel bands (see Table 5.1). For a sample of intermediate complexity, a six-step MudPIT of more than 200 microtubule-associated proteins was processed (see Table 5.2). To gauge results for a complex sample, an unfractionated rat hippocampal homogenate was analyzed by twelve-step MudPIT (see Table 5.3).

### *5.3.1 Preprocessing*

The three samples differed in complexity, chromatography, and mass spectrometry. One way to compare them is by the number of tandem mass spectra which were produced per minute of separation. In the gel bands, an average of 18.3 spectra were produced for each minute of separation. This average was 24.2 for the six-cycle MudPIT of the MAP proteins. The rat hippocampus 12-cycle MudPIT yielded a higher density of sampling at 34.6 spectra per minute. One cause of this diversity was the sample complexity; among the gel bands, the sample with the fewest peptides present also yielded the fewest spectra, averaging 9.6 spectra a minute. The least complex MudPIT cycles were equivalent to the most complex gel band separations in spectra produced per minute.

The spectra varied considerably in the numbers of peaks they included. The doubly-charged

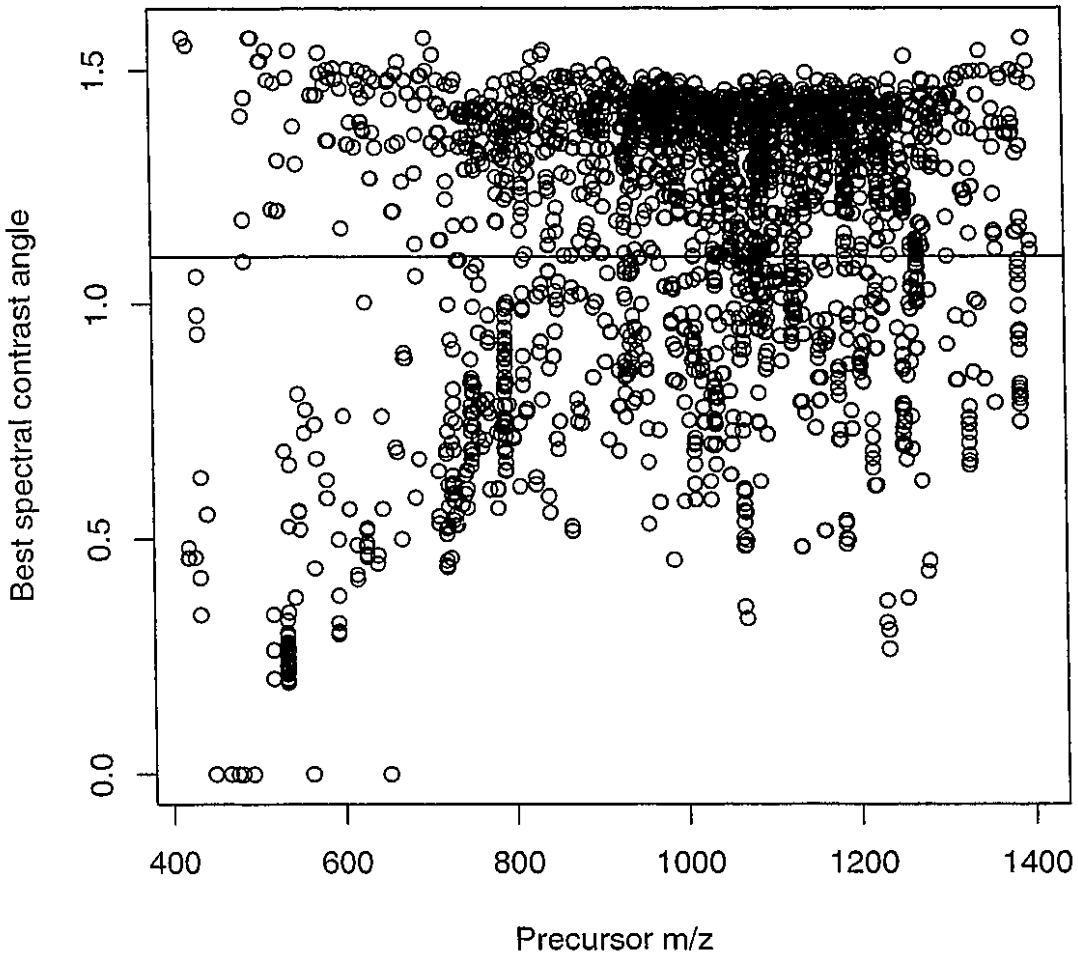



Figure 5.3: The second of six cycles in the MAP sample MudPIT analysis was analyzed by NoDupe. Each of 2761 spectra is represented by a circle, positioned to indicate the angle to the best matching spectrum in the cycle and the m/z of the peptide ion which yielded the spectrum. Large groups of spectra for the same peptide may form vertical streaks of overlapping circles. Only the matches below the line at 1.1 radians are considered significant by NoDupe. Spectra too distant from neighboring spectra may not be compared to any spectra and are shown here as having zero angles.

Table 5.1: Eighteen gel bands were analyzed by RPLC / MS / SEQUEST. A total of 776 different peptides were confidently identified from these spectra (284 were found in multiple bands). When NoDupe removed duplicate spectra prior to SEQUEST's use, 33 peptides were not identfied. This loss of 4% of the sequences resulted from removing 18% of the spectra judged to be duplicates. The retained representative spectra did not score as highly as the removed duplicates.

| Gel Band | Spectra | % Matched | % Duplicate | Pep Before | Pep After |
|---|---|---|---|---|---|
| 1 | 606 | 28% | 23% | 65 | 64 |
| 2 | 684 | 27% | 22% | 59 | 56 |
| 3 | 560 | 26% | 22% | 64 | 62 |
| 4 | 665 | 15% | 11% | 62 | 60 |
| 5 | 660 | 8% | 5% | 66 | 66 |
| 6 | 576 | 22% | 16% | 62 | 59 |
| 7 | 690 | 18% | 12% | 44 | 43 |
| 8 | 687 | 11% | 8% | 53 | 53 |
| 9 | 615 | 12% | 8% | 90 | 90 |
| 10 | 740 | 22% | 16% | 55 | 53 |
| 11 | 724 | 27% | 20% | 60 | 55 |
| 12 | 336 | 49% | 45% | 22 | 20 |
| 13 | 583 | 23% | 16% | 56 | 55 |
| 14 | 758 | 30% | 22% | 62 | 57 |
| 15 | 559 | 19% | 16% | 42 | 41 |
| 16 | 621 | 20% | 15% | 84 | 79 |
| 17 | 782 | 32% | 24% | 78 | 72 |
| 18 | 685 | 38% | 30% | 36 | 32 |
| All | 11531 | 23% | 18% | 776 | 743 |

Table 5.2: A sample of approximately 200 microtubule associated proteins was separated by a six-cycle MudPIT. Of the 1044 identified peptides, 209 were found in multiple cycles of chromatography. Twelve percent of the identifications were lost if only one representative spectrum was retained from each group of similar spectra. The initial cycle of the MudPIT included many more duplicate spectra than other cycles.

| Cycle | Salt Conc. | Spectra | % Matched | % Duplicate | Pep Before | Pep After |
|---|---|---|---|---|---|---|
| 1 | 0% | 2891 | 60% | 47% | 174 | 114 |
| 2 | 10% | 2761 | 34% | 26% | 204 | 177 |
| 3 | 25% | 2445 | 42% | 36% | 147 | 123 |
| 4 | 50% | 2770 | 27% | 19% | 296 | 288 |
| 5 | 80% | 2810 | 25% | 19% | 274 | 255 |
| 6 | 100% | 2956 | 26% | 21% | 158 | 142 |
| All | | 16633 | 36% | 28% | 1044 | 916 |

actin peptide VAPEEHPVLLTEAPLNPK appeared 35 times among the gel band data, the most for any peptide. The average number of peaks for these spectra was 303 with a standard deviation of 77. After the preprocessing in NoDupe, the average number of peaks in these spectra decreased to 74, and the relative standard deviation diminished from 26% to 20%. The preprocessing step removed approximately 70% of the peaks in spectra. See Figure 5.1 for two additional examples of the effects of preprocessing.

### *5.3.2 Spectral similarity characterization*

Spectral duplication was common in these collections. Among the gel bands, 23% of spectra yielded similarity angles below 1.1 radians to at least one other spectrum in the same band. For MudPIT results, the rate was even higher. In the MAP sample, 36% of the spectra were similar to another within the same salt cycle, and 33% of the spectra in the hippocampus sample were similar to others within the same MudPIT cycle. Individual reversed phase gradients varied considerably from these percentages; gel band five contained only 8% similar spectra, while 60% of the spectra from the first cycle of the MAP sample were similar to others within the cycle.

Table 5.3: A rat hippocampus was analyzed by a twelve-cycle MudPIT. Fourteen percent of the peptide identifications were lost when duplicate spectra were removed. Of the 4308 peptide identifications, 563 were found in multiple cycles. As seen in the MAP sample MudPIT, the initial cycles contained a higher proportion of duplicates. The final cycle shows an increase in duplication relative to the preceding cycles.

| Cycle | Salt Conc. | Spectra | % Matched | % Duplicates | Pep Before | Pep After |
|---|---|---|---|---|---|---|
| 1 | 0% | 3528 | 37% | 31% | 214 | 166 |
| 2 | 10% | 3219 | 39% | 35% | 154 | 110 |
| 3 | 15% | 3269 | 40% | 36% | 174 | 133 |
| 4 | 20% | 4041 | 37% | 29% | 565 | 503 |
| 5 | 25% | 4175 | 37% | 28% | 555 | 507 |
| 6 | 30% | 4241 | 30% | 21% | 467 | 425 |
| 7 | 35% | 4233 | 28% | 20% | 449 | 393 |
| 8 | 40% | 4191 | 30% | 21% | 510 | 449 |
| 9 | 45% | 4153 | 30% | 21% | 526 | 448 |
| 10 | 50% | 4072 | 26% | 18% | 467 | 407 |
| 11 | 60% | 4049 | 24% | 17% | 426 | 381 |
| 12 | 100% | 3894 | 37% | 32% | 364 | 252 |
| All | | 47065 | 33% | 25% | 4308 | 3685 |

Clusters were assessed among the similar spectra. On average, there were 4.2 spectra per group in the gel band data, though the mean ranged from 2.7 to 12.5 spectra in individual bands. The average cluster size in the MAP sample was 4.6 spectra, and the average group size for the hippocampus sample was 4.5 spectra. The most common type of group was the spectral pair; among spectra matching at least one other in the gel band data, 35% were in pairs. In the MAP sample, 27% of spectra showing similarity were in pairs, and 30% were members of pairs in the hippocampus spectra. Excluding groups consisting of single spectra, the gel band data averaged 36 groups per band. The larger numbers of spectra in the MudPIT separations corresponded to larger numbers of groups: 215 groups per cycle in the MAP sample and 285 in the hippocampus sample. The spectrum representing the peptide ELGGY was the most common overall, appearing 120 times in the third cycle of the hippocampus analysis.

The second cycle of the MAP sample MudPIT showed typical similarity for this sample. Of its spectra, 34% were similar to others, and 26% could be removed as duplicates. The best similarity angle for each of the 2761 spectra in this cycle is displayed in Figure 5.3. Approximately two-thirds of the spectra showed insignificant similarity to any spectrum and formed a mass at the top of the Figure.

### *5.3.3 Peptide identification and similarity clustering*

NoDupe was employed to test the effect of removing duplicate spectra on SEQUEST results. If retaining only the most representative spectra from spectral clusters did not result in peptide identification loss, a considerable amount of processing time could be saved. The results, however, showed that high-confidence peptide identifications are lost when duplicates are removed. Of the identifications with normalized XCorrs above 0.3, 4% were removed among the duplicates in the gel band data, 12% no longer appeared in the MAP sample MudPIT, and 14% were removed in the rat hippocampus MudPIT. A comparison of proteins identified with and without duplicate spectrum removal shows a similar trend; 5% of the gel band proteins were no longer identified, while 9% and 19% of MAP and rat hippocampus proteins were lost. Simple removal of duplicate spectra resulted in lost identifications.

Figure 5.4 gives an example of an identification which is lost if NoDupe removes duplicates.

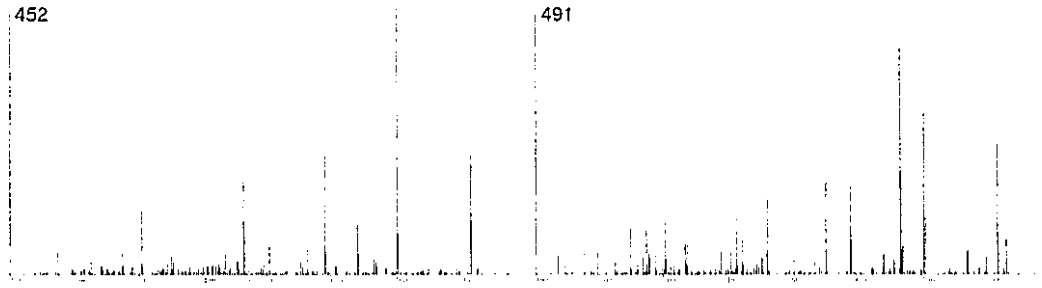



Figure 5.4: Scans 452 and 491 yield a spectral contrast angle of 0.847 radians. Of the pair, 491 is judged by NoDupe to be the better representative. When 452 is removed from the collection, however, the peptide KLLSAEER is no longer identified (this sequence ranks third for scan 491). The differences between these spectra may be the result of a co-eluting peptide in scan 491. A peptide identification algorithm which takes similarity into account could process these spectra simultaneously, saving time and retaining this peptide identification.

Scans 452 and 491 score as similar to each other with an angle of 0.847 radians. Since there are only two spectra in this group, choosing the more representative is arbitrary. NoDupe retains the spectrum with the largest proportion of peaks removed, and so scan 491 is preferred to 452 (21% of peaks remaining vs. 24%). Since pairs are the most common group size, the means by which ties are broken is significant in determining the loss of peptide identifications.

The peptide IVQVVTAEAVAVLK is represented by 185 spectra in the MAP sample (see Figure 5.5). Five of the six cycles of chromatography include this peptide: 9 in cycle two, 11 in cycle three, 19 in cycle four, 58 in cycle five, and 88 in cycle six. Within cycle six, the spectrum assigned the best normalized XCorr (0.754) is scan 2302. NoDupe, however, selects scan 2014 as most representative (both 2014 and 2302 match to 86 other spectra in the cycle, and so the proportion of peaks retained after preprocessing is used to break the tie). Scan 4892 is assigned the same sequence as the others by SEQUEST, but NoDupe did not find it sufficiently similar to any of the other spectra for grouping. In this example, NoDupe's grouping corresponded very closely to SEQUEST's results.

Although NoDupe compares spectra within an individual liquid chromatography separation, it is apparent that spectra in multiple separation cycles can be similar to each other. In the MAP sample MudPIT (see Table 5.2), 1044 different peptides were observed overall, but the sum of peptide identifications for each cycle was 1253; 209 of the identifications were identical to those in other

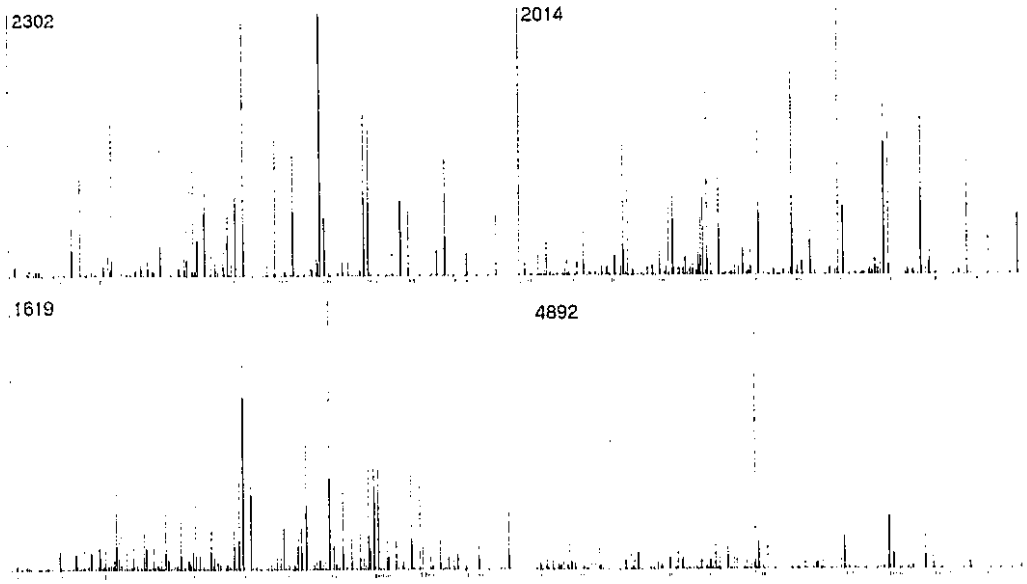



Figure 5.5: Scan 2302 is the spectrum representing IVQVVTAEAVAVLK which yields the best normalized XCorr (0.754). The spectrum NoDupe choses as most representative of the group of 88 spectra is scan 2014. Scan 1619 is shown as a variant form of this spectrum, grouped correctly by NoDupe and identified by SEQUEST as representing the same peptide sequence. SEQUEST identifies scan 4892 as being the same sequence as the others, but its similarity was insufficient for NoDupe to include it in the group.

cycles. Similar results were observed for the hippocampus MudPIT (see Table 5.3), where 4308 different peptides were observed overall, but the sum of identifications for individual cycles was 4871. Surprisingly, the gel band data also showed considerable overlap. A total of 776 different peptides were observed overall, but the sum of peptide counts for each gel band was 1060 peptides. Similarity may exist between spectra which resulted from different separation conditions.

Peptides may elute in multiple ammonium acetate salt steps due to several causes. A particular peptide's superabundance may result in chromatographic peak broadening. On a larger scale, overloading a column may reduce resolution for all peptides. Dead volume between the pumps and the column may result in apparent carryover between multiple steps of a MudPIT through delay of the highest hydrophobicity solvent mixture. The use of a step gradient rather than a linear gradient for the first dimension of separation may diminish its separative capacity [20]; the pI ranges of peptides eluting in adjacent salt steps generally overlap [21]. Taken together, these causes may result in multiple elutions of individual peptides.

## *5.4 Conclusion*

This application of the dot-product algorithm reveals the degree of duplication present among spectra resulting from liquid chromatographic separations of peptides for tandem mass spectrometry. In gel bands, the proportion of duplicate spectra ranged from 5 to 45%, and MudPIT separations varied from 17 to 47%. Spectra bearing similarity to others are often members of pairs, but some clusters may be far larger in size. This use of the dot-product algorithm treats experimental spectra as the library, revealing the structure of the data before peptide identification software is employed.

Currently, SEQUEST and other peptide identification algorithms handle each spectrum as an independent entity, whether or not a spectrum is similar to others. A more efficient way to deal with duplication among proteomic spectra is to process similar spectra in parallel. If spectra are initially grouped by similarity, they can be treated as a unit during identification for substantial time savings. For example, several spectra may result from a peptide ion with a particular m/z. The precursor m/z measurements for each spectrum may vary slightly, but a similarity algorithm could note their fragment ion similarities and group them. A more accurate precursor m/z could be calculated from the multiple spectra. The peptide identification algorithm could then draw its candidate sequences

from the database. Instead of finding the best sequence for each spectrum independently, though, the algorithm would find the sequence which produces the best hit against any of the spectra and then assign that sequence to all of the spectra in the group.

Such an algorithm would outperform traditional peptide identification algorithms in several ways. In a yeast database search, approximately half of SEQUEST's time is taken in searching the database for candidate peptides and calculating preliminary scores for them. This proportion increases with database size. If similar spectra are grouped, the candidate peptides can be selected from the database once for each group rather than once for each spectrum. Because the precursor m/z measurements would be more accurate, a narrower mass window could be employed for selecting candidate sequences, reducing their number. In addition, calculating preliminary scoring in parallel for the group rather than serially for spectra would be more efficient. As database sizes increase, greater amounts of processing time could be saved by this technique.

An important result of processing similar spectra as groups would be that lower quality spectra are assigned sequences which correspond to the higher quality spectra in their groups. For example, if a peptide has a prominent neutral loss, some spectra may be so dominated by the precursor neutral loss that little fragment ion information is present. If these spectra are associated with spectra with more informative fragment ions, however, their sequences can be correctly assigned. As spectral collections increase in size, the chance of a hit to any protein in a database increases. Grouping the spectra by similarity before identification would help alleviate this random matching.

The differences between similar spectra may be useful for peptide identification. If one variant of a spectrum shows fragment ions more clearly at low m/z values and another shows them more clearly in the high m/z region, these two spectra together could yield a more accurate peptide identification than either could separately. The creation of such an algorithm would require more sophistication than the modification described above, but it may be the case that such an algorithm would yield higher accuracy in peptide identifications than is currently possible.

Similarity matching among uninterpreted spectra has other possible applications. For example, spectral libraries would be most effective if they contained representative spectra rather than randomly chosen ones. Another use would flag spectra which group by similarity but receive different sequence identifications for subsequent manual or *de novo* examination. Pevzner *et al* suggested that similarity algorithms could match modified and unmodified variants of peptide spectra or match

peptide spectra which have overlapping sequences [22]. The extent of similarity among proteomic spectral collections is a feature which proteomic software should exploit.

# Chapter 6

# CREATION OF HIGH-THROUGHPUT, ACCURATE SEQUENCE TAGGING ALGORITHM

## *6.1 Introduction*

Algorithms which search for peptide sequences to identify tandem mass spectra can be classified by the way in which they limit the search. By far the most commonly used class of algorithms limits the sequence search to peptides found in protein sequence databases. The database searching approach, exemplified in the SEQUEST [1] and Mascot [2] algorithms, is limited to those spectra that correspond to known sequences. Another class of algorithms combinatorially assmbles amino acid residue masses until the appropriate peptide mass is reached. The list of candidates is compared to the spectrum and ranked accordingly. Because the numbers of possible sequences grows very rapidly with peptide mass, this technique is applicable only to spectra of the smallest peptides [3]. *De novo* algorithms attempt to deduce the amino acid sequence based on the fragment ion peak-to-peak mass differences. Mann proposed the "sequence tagging" approach to peptide identification [4], which uses elements of both database search and *de novo* approaches; a short sequence "tag" is inferred manually from the spectrum, and then database lookup finds complete peptide sequences which match this sequence and the sequence masses flanking it. Because algorithms to infer tag sequences have had limited accuracy, this technique has been applicable only to spectra where the partial sequences can be derived manually.

All of these algorithms rely upon models which relate a tandem mass spectrum to a particular sequence. In database searching, theoretical spectra are constructed from candidate sequences, and then the theoretical spectra are compared to the observed one. In *de novo* algorithms, a sequence is inferred from the spectrum. Most models currently in use, however, do not reflect the complex chemistry underlying the formation of tandem mass spectra, described in Chapter 4. One common approach expects all ions from the *y* series to be uniformly intense, while all ions from the *b* series

are modeled at a uniform lower intensity. Chemical and statistical research, however, has established that instrumentation and sequence composition can influence fragmentation, causing some ions to be very intense while occulting others [5].

Bartels's publication of a peptide sequence inference algorithm based upon the concept of a "sequence spectrum" [6] has been influential in shaping subsequent *de novo* algorithms. Sherenga [7], Lutefisk [8], and SeqMS [9] have all used this approach, which transforms the observed spectrum into a directed graph from which the sequence can be deduced. The sequence graph contains nodes which represent collections of peaks in the original spectrum; for example, if a spectrum for a peptide of mass 1000.0 Da contains ions at m/z 172.0, 200.0, and 800.0, these may represent *a*, *b*, and *y* ions, respectively, leading to a node in the sequence graph. Sequences which explain the nodes with higest probability are those most likely to correspond to the sequence of the peptide. This "sequence spectrum" approach attempts to assess the series for each observed ion before possible sequences are generated for the spectrum. As a result, the influence of sequence on the intensity and presence of ions cannot be part of this series assessment. Statistical analyses, however, have shown that particular amino acid residues can influence the intensities of neighboring fragment ions significantly (again, see Chapter 4). Attempting to determine fragment ion identity without the sequence context is prone to error.

In this chapter, I describe GutenTag, a new sequence tagging algorithm with several advantages over existing approaches. First, the software automates the inference of sequence tags from the spectrum using a more accurate model of spectral intensity, making it possible to apply sequence tagging in large-scale analyses. In addition, it simultaneously queries the sequence database for the top tag sequences with improved performance. In most cases, multiple database sequences will match sequence tags for each spectrum, but the algorithm selects the best match by an efficient scorer. The program is capable of comparing partial sequences to spectra, enabling it to identify post-translationally modified peptide spectra or to identify peptides for proteins which have incorrect or variant database sequences. I compare GutenTag identifications to those of SEQUEST for a defined mixture of proteins including both modified and unmodified peptides. An examination of human lens crystallins helps to demonstrate the importance of matching peptides with unspecified modifications and / or sequence errors.

## 6.2 Experimental section

### 6.2.1 Training set preparation

The extraction, digestion, separation, and mass spectrometry of the peptides used as GutenTag's training set were previously described in Chapter 4. Proteins from a culture of *Saccharomyces cerevisiae* (strain 1560) were divided into pellet and supernatant fractions and then reduced and alkylated. The pellet fraction was digested by CNBr, Endoproteinase Lys-C, and trypsin, while the supernatant fraction was digested by Endoproteinase Lys-C and trypsin. Twelve-cycle MudPITs [10] analyzed the soluble fraction twice and the pellet fraction once. Spectra were identified with Normalized SEQUEST [11], and identifications were assembled by DTASelect [12]. The 1437 identifications allowed into the training set met each of the following criteria: the precursor was doubly-charged, XCorr > 0.45, and the peptide was not post-translationally modified.

### 6.2.2 Defined mixture preparation

The defined protein mixture was previously described by MacCoss *et al.* [13]. A mixture of the five proteins was reduced and alklyated before being divided into four aliquots. Aliquot 1 was digested with tryptin, aliquot 2 was digested by elastase, aliquot 3 was digested by subtilisin, and aliquot 4 was digested by proteinase K. Only aliquots 1 and 4 were used for the comparison between GutenTag and SEQUEST. Each aliquot was then separated by a six-cycle MudPIT analysis. A database consisting of proteins identified in the sample (both contaminants and known contents) was appended to the *S. cerevisiae* ORF database. Normalized SEQUEST and GutenTag both used this database for identification.

### 6.2.3 Lens sample preparation

The lens sample preparation was previously described by MacCoss *et al.* [14]. Tissues were obtained from a 4.25-year-old congenital cataract patient. After extraction and solubilization, proteins were reduced and carboxyamidomethylated. The mixture was digested with trypsin. The sample was analyzed via an 18-cycle MudPIT separation. Prior SEQUEST analysis included the following modifications: phosphorylation, oxidation, methylation, and acetylation. GutenTag identifications

employed a database consisting of previously identified protein components and the *S. cerevisiae* ORF database (as distractors). Follow-up SEQUEST searches used a database consisting of previously identified protein components.

### *6.2.4 GutenTag algorithm*

Software in the Java programming language was created to preprocess spectra, infer sequence tags from them, search a database for these tag sequences, and evaluate the peptide sequences resulting from these searches (see Figure 6.1). Although the software was developed under Microsoft Windows 2000, it runs equally well on Linux and other operating systems which support Java Virtual Machines. With a database of 6500 proteins (such as the *S. cerevisiae* database), the software requires approximately one second per spectrum on an AMD Athlon XP 1700+, and the time required scales roughly linearly with the size of the database. GutenTag relies upon an analysis of its training set [15]; the model it uses for estimating fragment ion peak intensities is derived from spectra produced in a tryptic digest of a yeast proteome.

#### *Preprocessing*

Deisotoping the spectrum is necessary to remove peaks which would introduce spurious sequences. In essence, the fewer peaks in a spectrum, the fewer the possible sequences. Fragment ion isotope peaks result from the natural isotope distributions of the elements. Since approximately 1% of carbon atoms contain seven neutrons rather than six neutrons, a fragment ion containing tens of these atoms has a reasonable chance of containing at least one more neutron than others. GutenTag's deisotoper calculates how much of each peak's intensity is likely to be the result of peaks which appear at lower m/z values and removes that amount of intensity from the peak. Peaks which are less intense than estimated in the isotope distribution are removed. The technique used for estimating intensity distributions was published by [16].

Because ion trap mass spectrometers have relatively low mass accuracies, the precursor peptide's observed mass may be higher than the monoisotopic mass of the precursor due to additional neutrons or due to measurement error. Knowing the mass accurately is important to align $b$ ions with $y$ ions correctly. Because each pair of $b$ and $y$ ion constitutes an independent measurement of the

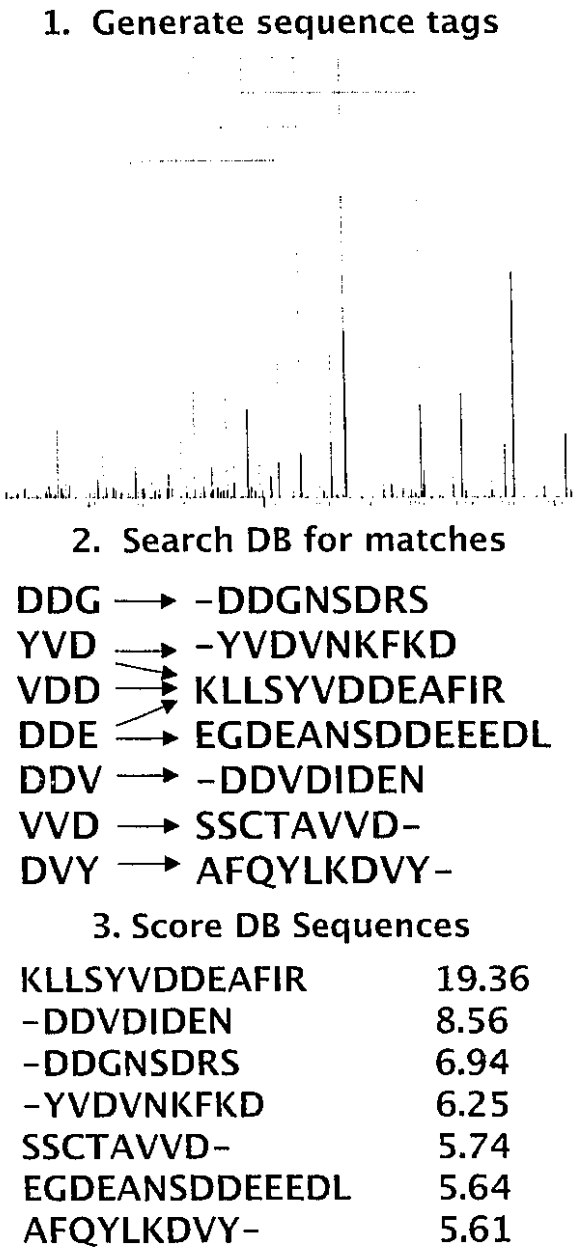



Figure 6.1: GutenTag infers short sequences directly from each spectrum. The best sequence tags are sought in a sequence database to find peptide sequences which match a tag sequence and at least one flanking sequence mass. These partial and complete peptide sequences are ranked by dot product score. In the above example, three tag sequences match to the same complete peptide sequence. This complete sequence scores more highly by dot product than the other sequences.

precursor mass, it is possible to improve the mass calculation from the complementary peak pairs in the spectrum. GutenTag serially subtracts 0, 1, and 2 neutrons from the observed precursor mass to determine which causes *b* and *y* ions to align optimally, and then the complementary pairs are summed to re-estimate the monoisotopic, singly-charged mass of the peptide ion.

GutenTag uses the re-estimated peptide mass to pair *b* and *y* fragment ions. The mass of a *b* ion plus the mass of the complementary *y* ion is equal to the mass of the singly-charged peptide precursor ion plus one proton. For each peak, the most intense complementary peak within m/z tolerance is assigned as the complement ion. Prior studies have shown that 84% of *b* ions and 90% of *y* ions are present at measurable intensity for these spectra [17], and so the presence of *b* ions can corroborate the information present in *y* ions.

The final step in preprocessing finds pairs of peaks separated by the residue masses of amino acids. The spectrum is looped through once for each amino acid mass. The software finds pairs of fragment ions separated by the masses of amino acid residues and stores the links for later sequence generation. In essence, the peaks in the spectrum act as nodes in a graph, with edges linking them to other peaks which appear at appropriate mass gaps. Since GutenTag is designed to handle only doubly charged peptide spectra, it assumes all fragment ions are singly charged, and m/z can be assessed as identical with mass.

*Sequence construction*

Generating possible sequences from the peaks and gaps observed in the spectrum is a recursive process. The example in Figure 6.2 gives an example of six peaks with five links among them. Peak A is an amino acid mass away from peaks B and C. Peak B shows appopriate gaps to peaks D and E, and peak C is a proper distance away from peak F. When the recursion begins at peak A, it first uses the link to peak B for construction. Peak B's link to D completes the first two-residue sequence tag constructed (A-B-D), and then peak B's link to peak E is followed to produce the second tag (A-B-E). After these possibilities are exhausted, the recursion proceeds from peak A through peak C, constructing the A-C-E tag and then the A-C-F tag. Since three peaks are incorporated in each tag, the sequences contain two amino acids.

GutenTag is configured to accept a minimum and maximum sequence length. Whenever a tag

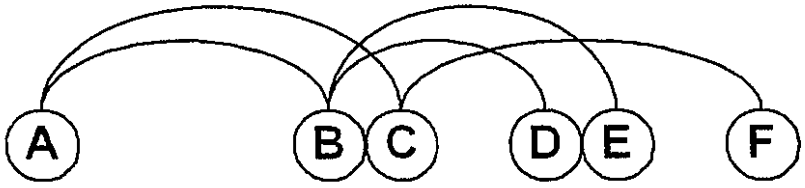



Figure 6.2: The peaks of each spectrum constitute nodes in a directed graph. Only those peaks separated by an amino acid's mass are connected by edges. By recursively traversing these edges from low to high m/z, sequence tags can be inferred from the spectrum.

is generated of appropriate length, the sequence is passed to the scorer for examination. If a tag is the maximum sequence length, the recursion is allowed to go no further. Sequence tags are only compared to other sequence tags of the same length. A configurable number of tags of each length are kept; the thousands of possible tags which can be inferred from a single spectrum would otherwise swamp memory.

*Model construction*

After a sequence tag has been constructed, GutenTag models the intensity and m/z of each peak expected to appear in the tag. The peaks used in the recursion are assumed to be *y* ions, and their complementary peaks are handled as *b* ions. Intensities and m/z for each expected fragment ion are estimated as described below.

GutenTag can estimate a *y* or *b* ion's intensity based on the fragment ion's mass divided by the intact peptide's mass. While *b* ions are fairly uniform in intensity throughout the spectrum, *y* ions crest prominently two-thirds of the way through. GutenTag uses a fourth degree polynomial to model each of these distributions. If the equation returns a value less than zero intensity, the peak's intensity is modeled as zero. See Figure 6.3 for a comparison of the model to the observed median intensities for both series.

Individual fragment ions may vary in intensity due to the amino acid residues flanking them in the intact peptide. Proline, for example, has long been known to augment fragmentation to its N-terminal side while significantly diminishing C-terminal fragmentation. GutenTag incorporates the influences of the two neighboring amino acid residues for estimating each fragment ion's intensity.

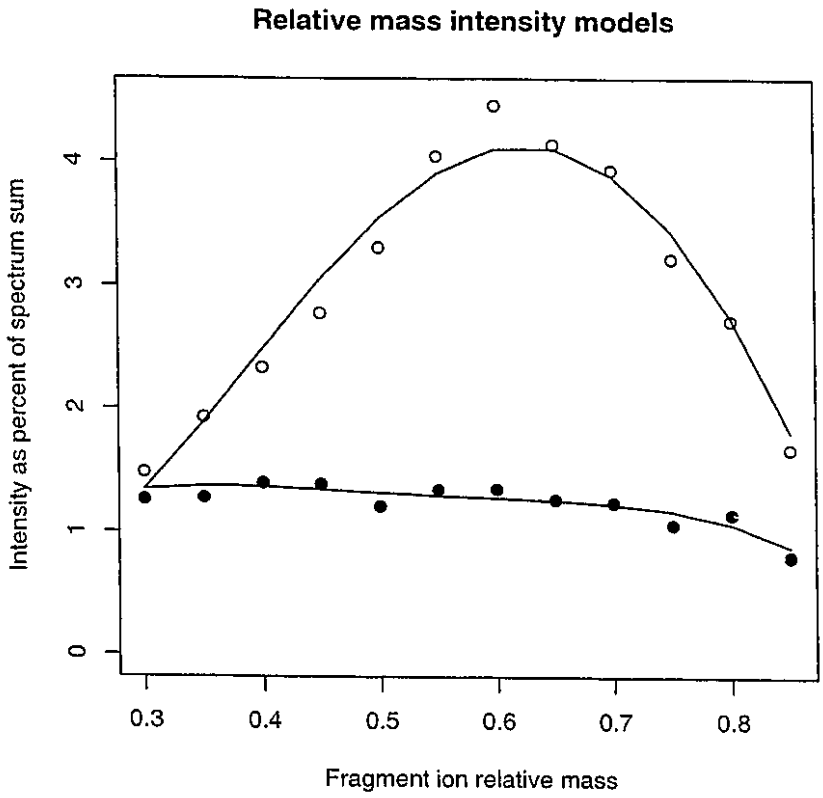



Figure 6.3: The ratio of a fragment ion's mass to the intact peptide's mass can be used to estimate the intensity of the fragment ion. The empty circles above show the mean intensity of $y$ ions within a range of relative mass. The filled circles are mean intensities of $b$ ions. GutenTag uses quartic equations to produce its estimates of peak intensities, shown here by lines traversing the graph.

The influence of each amino acid is read as an "N-bias" value from a configuration file [18]. Each N-bias reports the ratio of intensities expected for the two peaks bordering a specific amino acid. GutenTag divides the ratio of the C-terminal residue by the N-terminal N-bias ratio. In effect, a high N-bias amino acid to the N-terminal side diminishes expected intensity, while a high C-terminal N-bias increases it.

GutenTag adjusts the modeled m/z values of the peaks in each sequence tag. It may initially seem unnecessary to estimate the true m/z values for the fragment ions in the sequence tags; the sequences can be generated only if the peaks are separated by the masses of amino acids. It is more accurate to state that the peaks are separated by the masses of amino acids within a set tolerance. Successive amino acid gaps may be slightly wider or narrower than predicted, causing sequence tags to align better with observed peaks in some positions than in others. GutenTag finds the optimal m/z position for the sequence tag to minimize the overall distance between observed and expected peak locations. By performing this alignment, GutenTag can use deviations between expected and observed m/z values in its scoring of individual tags.

*Sequence tag evaluation*

In some ways, evaluating a sequence tag is more challenging than evaluating a full peptide sequence. If the complete peptide sequence is being predicted, each observed peak may be explained by the correct sequence. A sequence tag, however, can account for only a portion of the spectrum. As a result, the information available to score a predicted tag is less than is available to score a full peptide sequence. To compensate for this difficulty, GutenTag makes use of a two-dimensional evaluation scheme. Both *b* and *y* ions are included in intensity scoring, but only *y* ions are included in m/z scoring.

Scoring intensities is complex because fragment ion intensities do not form Gaussian distributions. Instead, a very small proportion of peaks are intense, while most peaks are quite small. GutenTag transforms peak intensities to a probability domain in order to analyze them; instead of noting that a peak comprises approximately 1% of the spectrum's intensity, GutenTag computes that the peak is taller than 50% of *y* ions. To handle this transformation, GutenTag uses equations of the

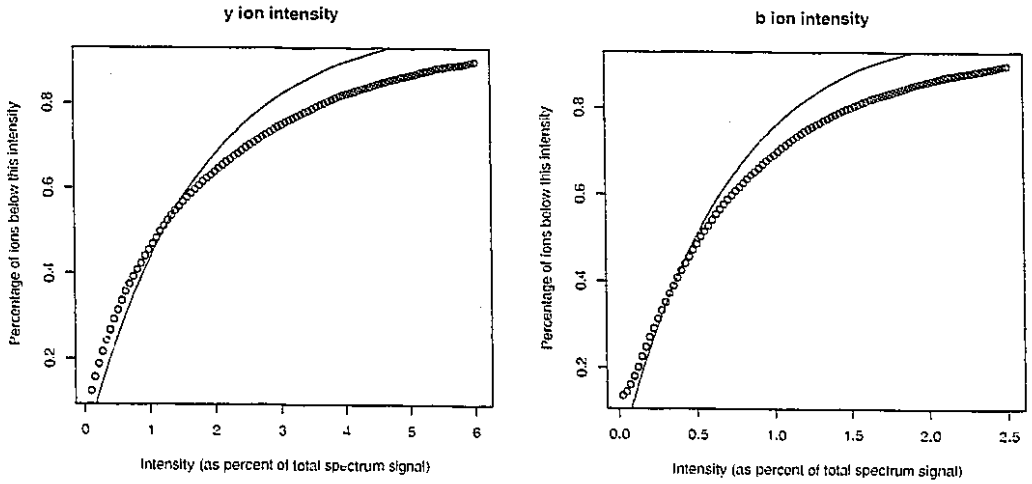



Figure 6.4: The observed distributions of $b$ and $y$ ions are shown by the series of circles in each of the two cumulative density functions above. The equation used by GutenTag to model these functions is superimposed as a line on each graph. The equations differ from the observed values in part because the function passes through (0,0) and (1,1).

form

$$P = \frac{1 - e^{-CI}}{1 - e^{-C}} \tag{6.1}$$

to model the cumulative density functions of these distributions. Given the intensity ($I$) of a peak and the constant for the peak's ion series ($C$), the function returns the percentage of fragment ions which are less intense ($P$). A missing peak has zero intensity and is modelled in the zeroth percentile, while a peak in a single-peak spectrum is modelled as the 100th percentile. The constant is chosen to minimize Kolmogorov-Smirnov [19] deviation from the observed intensity distribution from the appropriate series of ions. See Figure 6.4 for a graphical representation of this model.

The intensity score for each peak is found by determining the modelled percentage of $b$ or $y$ ions falling between the observed and calculated intensities. If the observed $y$ ion peak, for example, is more intense than 70% of $y$ ions and the calculated peak is more intense than 50% of these ions, the absolute difference between these measures is 20%. The difference between modeled and observed intensities is subtracted from one to yield an intensity score of 0.8. If a peak is missing, the score will be lower if a more intense peak is modeled in its location than if a minor peak is modeled there.

These intensity scores are multiplied together for the *b* and *y* ions.

The degree to which observed *y* ion m/z values match those calculated for the sequence tag is evaluated in a similar way. The same equation used to model the intensity cumulative density functions is applied to model the differences between observed and expected m/z values. Given the absolute m/z difference, the equation returns the modelled percentage of fragment ions which would be found nearer to the expected location. These values are subtracted from one and multiplied together. Ions from the *b* series are not included because they may be absent; *y* ions are guaranteed to be found because they were used to generate the sequence being scored.

The final score for a sequence tag, then, has three components: the *y* ion intensity score, the *b* ion intensity score, and the *y* ion m/z score. These three scores are multiplied together to create the tag score. A perfect tag includes only peaks that are exactly as intense as modelled and which are found at the precise m/z values predicted and scores 1.0. Tags that include larger numbers of peaks have more opportunities for error and thus can be expected to score lower than shorter tags.

*Database search and sequence evaluation*

Once the sequence tags for a spectrum have been infered and ranked by their scores, a sequence database can be searched for these short sequences. Rather than search the database once for each top-ranking tag, GutenTag employs an technique to search for all tags simultaneously [20]. Briefly, this algorithm assembles the tag sequences into a search tree, and the tree is traversed as the program moves through the sequence database (see Figure 6.5 for an example). The database is loaded into memory at the start of processing for the full set of spectra in order to speed this process.

If a sequence tag is matched to the database sequence, the program next checks whether or not the flanking sequences match the known masses to the N-terminal and C-terminal sides of the tag. If both masses are matched within 1.5 Da, the full spectrum may be explained by the peptide sequence. If only one of the two masses match, one end of the database peptide matches the spectrum while the other does not. In this way, mis-sequenced or post-translationally modified peptides may be identified.

Typically, GutenTag will identify several candidate sequences to explain each spectrum. To evaluate these candidates, the program will compute the m/z ratios and intensities expected for

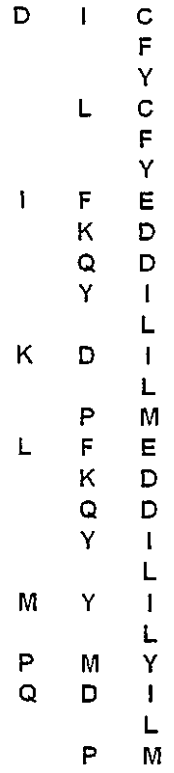



Figure 6.5: Once a set of sequence tags has been infered from the spectrum, GutenTag assembles the tag sequences into a tree to find more rapidly each occurence of each tag in the sequence database. In this example tree, 25 tag sequences are stored. The 25 tags all begin with one of seven letters: D, I, K, L, M, P, or Q. Note that many sequence tags are present twice, once with I and once with L or once with K and once with Q. These amino acid pairs are isobaric, and so the differences between peak m/z values cannot be used to differentiate them. If this sequence tree is used to examine a sequence containing LELQDLYVLG, the tags matched include LQD, QDL, and DLY.

each fragment ion, using the same intensity estimation system employed for sequence tag ions. The program attempts to match observed peaks to the expected ones, and a normalized dot-product score for the sequence is generated:

$$D = \frac{\sum eo}{\sqrt{\sum e^2 \sum o^2}}, \tag{6.2}$$

where $D$ is the normalized dot-product, $e$ is the expected intensity of each peak, and $o$ is the observed intensity of each peak.

Candidate sequences, whether matching only one mass or both masses, are sorted by the dot-product multiplied by the number of ions matched to the sequence. Because scores reflect the numbers of matched ions, partial sequences are at a disadvantage to full sequences. The top five sequences for the spectrum are then recorded to a file in the unified SEQUEST result format.

## *6.3 Results and discussion*

### *6.3.1 Validation on defined protein mix*

GutenTag's ability to identify peptides was validated against Normalized SEQUEST in its analysis of a defined five-protein mix including bovine serum albumin, horse apomyoglobin, rabbit cytochrome C, rabbit phosphorylase A, and bovine beta-casein [21]. Counting only spectra from doubly-charged precursor ions, a tryptic digest produced 6170 spectra, while a proteinase K digest yielded 7972 spectra. Only these spectra were included in this analysis because GutenTag's intensity model is specific to this type. SEQUEST and GutenTag were run on these collections, using a sequence database including the above proteins, common contaminants, and the *S. cerevisae* ORFs. Identifications matching only yeast proteins were labeled as false, and the others were all assessed as true identifications.

The numbers of true and false identifications from each algorithm in each set of spectra were recorded in Table 6.1. GutenTag assigned fewer correct, complete sequences to the spectra than did Normalized SEQUEST. When GutenTag gave an incorrect sequence, however, it generally gave a partial sequence rather than a complete sequence; 70% of tryptic digest spectra that were assigned complete sequences were assigned true identifications. SEQUEST produced a larger number of true sequences than GutenTag, but its limitation to yield only complete sequences meant that the false positives greatly outnumbered the true positives among its identifications.

Table 6.1: Comparison of SEQUEST and GutenTag results. 1) The Code column gives the set of spectra and algorithm. "T" indicates tryptic spectra, while "K" indicates proteinase K spectra. "GT" specifies GutenTag was used, and "SQ" notes that Normalized SEQUEST identified the peptides. 2) The "Full" columns indicate the numbers of true and false positives among complete-sequence identifications. 3) The "Part" columns indicate the numbers of true and false positives among partial-sequence identifications. 4) The rightmost section of the table gives the score which removes 95% of the false positive identifications, the percentage of the identifications retained that were true, and the percentage of all true identifications which are retained by the cutoff. Both algorithms perform better on tryptic digests than proteinase K digests.

| Code | Full T IDs | Full F IDs | Part T IDs | Part F IDs | 95%ile of F scores | %T over cutoff | % of T retained |
|---|---|---|---|---|---|---|---|
| T-GT | 1328 | 558 | 766 | 3515 | 10.4 | 97.7% | 89.2% |
| K-GT | 1039 | 667 | 621 | 5645 | 10.7 | 95.6% | 71.5% |
| T-SQ | 1987 | 4183 | N/A | N/A | 0.348 | 85.1% | 60.2% |
| K-SQ | 2230 | 5742 | N/A | N/A | 0.319 | 80.0% | 51.7% |

A standard approach for finding the best identifications in proteomics data is to accept all identifications above a threshold score. The scores which would reject 95% of false positive complete sequences are reported in Table 6.1. The proportion of identifications above this cutoff that were true positives was substantially higher for GutenTag than for SEQUEST, even in proteinase K spectra. The percentages of all true identifications retained by this cutoff was also higher for GutenTag than SEQUEST. Even though SEQUEST yields a larger number of true positives, GutenTag's true, complete-sequence identifications are more readily separated from false ones on the basis of score (see Figure 6.6).

The numbers of true and false complete-sequence IDs in narrow windows of score were calculated. The top-scoring ten percent of false positive identifications were split into five bins. The top bin, for example, ranged in score from that of the best-scoring false identification to the score of the false identification at the 98th percentile. The numbers of true and false positives in each bin were calculated, and Table 6.2 shows the resulting percentages. The results for complete GutenTag identifications indicate that less separation is achieved between true and false positives in proteinase K spectra than in tryptic spectra, probably resulting from GutenTag's training on tryptic spectra.

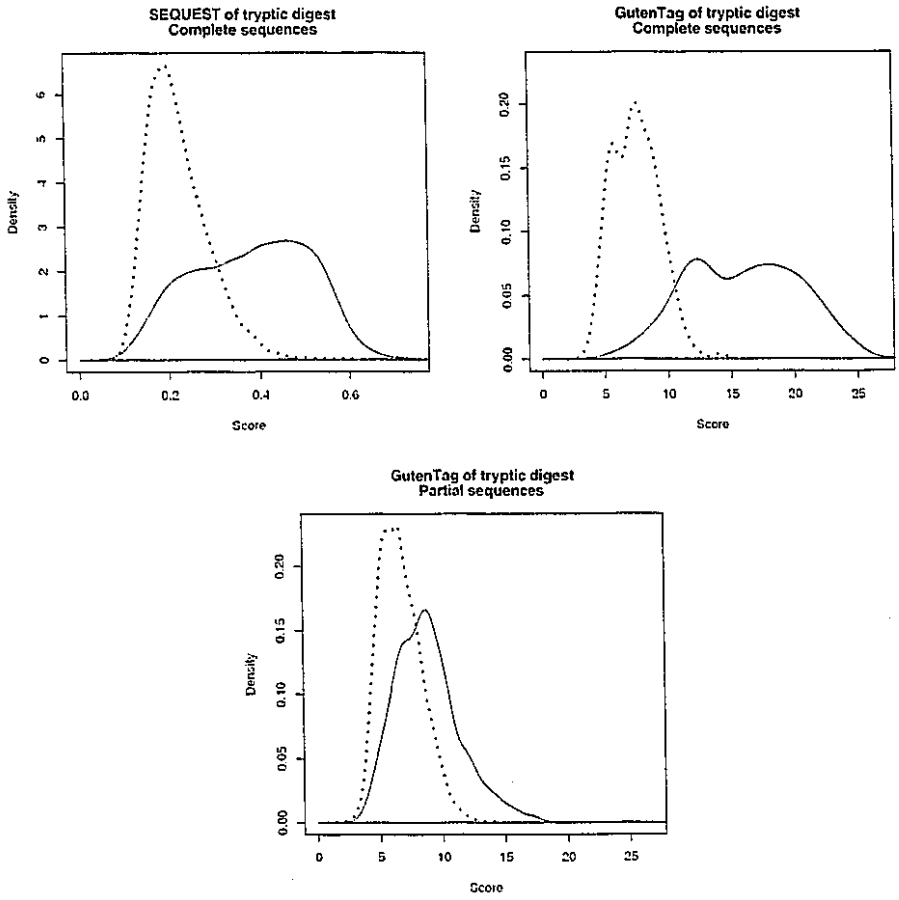



Figure 6.6: Score distributions for identification of tryptic peptides. Both SEQUEST and GutenTag score true identifications (solid line) more highly than false identifications (dotted line), but the degree of score overlap for complete sequences is reduced in GutenTag. Partial sequences are not as well separated, indicating the importance of inferring complete sequences (including necessary post-translational modifications) before acceptance of identifications.

Table 6.2: Percentage of true identifications by score range. The top-scoring ten percent of false identification scores were divided into five bins. The percentage of true identifications in each bin is shown above.

| Code | 100-98%ile | 98-96%ile | 96-94%ile | 94-92%ile | 92-90%ile |
|---|---|---|---|---|---|
| T-GT | 96.5% | 84.1% | 69.4% | 54.2% | 31.3% |
| K-GT | 91.5% | 84.1% | 53.6% | 65.0% | 64.9% |
| T-SQ | 92.3% | 62.5% | 53.4% | 47.1% | 36.9% |
| K-SQ | 87.2% | 70.9% | 51.5% | 47.7% | 34.3% |

The top 100 identifications were examined for each algorithm on both sets of spectra. For GutenTag, only complete sequences were included in the top hundred. Each identification was checked against the result for the other algorithm. GutenTag's identifications matched SEQUEST's for all 100 of the tryptic spectra and 95 of the proteinase K spectra. Of SEQUEST's top 100, 88 matched GutenTag's identification among the tryptics, and 82 matched among the proteinase K spectra. The relatively low percentage of SEQUEST identifications validated by GutenTag identifications reflected that SEQUEST allows more variability in spectra than does GutenTag. GutenTag's scoring model favored peptides terminating with Arg or Lys more highly than SEQUEST; its proteinase K identifications included 18 peptides terminating in these residues, while SEQUEST's proteinase K identifications included 12. GutenTag's spectrum model is more specific to tryptic peptides than is SEQUEST's.

### 6.3.2 *Identification of modified peptides*

Several post-translational modifications were identified among the proteins of the defined mixture. GutenTag's partial sequence identifications were analyzed to find spectra corresponding to these modifications. An acetylated peptide at the amino terminus of rabbit cytochrome C [22] was identified as -EKGKKIF, matching the final seven residues of the peptide. When the sequence of the protein's N-terminus was added to GutenTag's identification, the remaining mass of 42 Da indicated a likely acetylation. Likewise, an acetylation on rabbit phosphorylase A's N-terminus [23] matched to six spectra as -PLSDQEK and three spectra as -RPLSDQEK. A phosphorylation on bovine beta-casein's

serine 50 [24] matched to seven spectra as -QQQTEDELQDK and one spectrum as -EQQQTEDELQDK, with the 80 Da remainder indicating a phosphorylation. The algorithm correctly identified spectra when modifications prevented complete sequence identification.

This capability was put to the test in an assessment of post-translational modifications in human lens tissue from a 4-year-old congenital cataract patient [25]. No modifications were specified in the configuration of GutenTag. Partial sequences matching to the A and B chains of crystallin alpha were analyzed to determine the amount of mass differentiating the observed spectrum from the database sequence.

Of GutenTag's partial sequence identifications to alpha crystallin chain A or B, 134 corresponded to modifications of 69 Da. In chain A, 74 spectra for a peptide of 2553.1 Da were all matched to similar sequences, with the N-terminal portions identified to sequences ranging from LFDQFFGEGLFEY- to LFDQFFGEGLFEYDLLP-. By subtracting the flanking database sequence from the remaining mass, the peptide sequence was found to be LFDQFFGEGLFEYDLLPFLSSS, where threonine 43 had become a serine and one of the final three serine residues had beta eliminated to become dehydroalanine. A SEQUEST search was configured to search for a shift of -32 on threonine residues (-14 for the mutation to serine and -18 for the beta elimination), and 137 spectra were identified to the modified sequence, with a best (non-normalized) SEQUEST XCorr of 5.514. The spectra did not reveal which of the three serines had beta eliminated. In this example GutenTag identified the peptide despite a change in protein sequence from a putative single nucleotide polymorphism and an unanticipated modification.

Chain B showed similar two similar shifts of 69 Da. Five spectra matched to the sequence LFDQFFGEHLLESDLFPTSS, where threonine 42 had become a serine and a beta elimination had taken place. The above SEQUEST seach was able to match 33 spectra to this sequence once this modification was specified, with a best XCorr of 5.651. GutenTag matched 48 spectra to LFDQFFGEHLLESDLFPTSTSS with a beta elimination, which suggested that leucine 44 may also be mutated to a serine in this sample. A total of 308 SEQUEST identifications, scoring as high as 5.7161, were found when the program was configured for this possibility. If these modifications are not in the SEQUEST configuration, these spectra are incorrectly identified.

Other modifications were also identified. Acetylations on the N-termini of both chains were observable from large numbers of spectra. Spectra featuring the characteristic +80 Da mass difference

of phosphorylation agreed with previously documented modifications on these proteins. In addition, many spectra resulted from peptide ions that had already lost ammonia or water. In some cases, the masses of modifications were more apparent than their chemical identities.

### *6.4 Conclusion*

GutenTag incorporates several advantages over existing proteomic algorithms. Its modeling of fragment ion intensity based on sequence and mass enables it to be more accurate in sequence inference than existing algorithms. Its ability to search a database with multiple tags in a single pass allows users to increase the number of tags retained for each spectrum, resulting in increased performance and accuracy. GutenTag's ability to score multiple sequences resulting from database searches sets it apart from existing sequence tag search algorithms. By uniting in a single program the capability to infer tag sequences and search databases, GutenTag outperforms existing techniques and is usable on larger datasets than previously possible.

GutenTag embodies a new capability for large-scale proteomics. The automated and accurate interpretation of partial sequences from tandem mass spectra will play an important role in the identification of protein sequence features not predicted through bioinformatics. We've shown the algorithm accuracy compares quite well with database searching algorithms for peptides created by site-specific and non-specific proteases. The algorithm can also help identify unanticipated modifications, sequence variations, and possibly alternate splice sites in proteins. By combining the best elements of *de novo* and database search algorithms, sequence tagging increases the scope of biological discovery possible using shotgun proteomic strategies.

# Chapter 7

# CONCLUSION

## *7.1 Effects of research*

Although the presented research may initially seem disjoint, the findings in each study drew attention to the next. An attempt to create software to interpret mass spectra led to an attempt to model fragment ion intensity. To understand how spectra should be modeled, it was necessary to characterize intensities in identified spectra. To generate statistics for this exploration, software to select spectral identifications was required. To write such software required a codification of how such spectra were selected manually. These linkages from project to project have underscored that research begets research; proteomics is still a young field, and many opportunities for exploration remain. A brief examination of the accomplishments in this body of work underscores this message.

### *7.1.1 DTASelect and associated advances*

What began as a simple program for selection of identifications became a much more significant contribution to the field. The replacement of SEQUEST Summary and Combine has accelerated the processing of proteomic samples in the Yates Lab and many others. DTASelect has standardized rules for selecting identifications, making it easier to communicate the rules used for selecting identifications. Because it was designed for scalability, DTASelect has greatly increased the scope of proteomics experiments. Even though this software started with more modest aims, it has had the most immediate impact of any of these projects.

If one is to compare an apple to an orange, a standard for comparison must be defined. By defining the rules by which proteomic samples could be cataloged, DTASelect made it possible for multiple samples to be compared meaningfully via Contrast. This development has made it possible to evaluate the reproducibility of MudPIT experiments. In addition, Contrast's scalability enables the comparison of supersamples, where multiple MudPIT runs have been conglomerated together

with similar samples. Literally millions of identifications can be compared to other millions of identifications, a capability that was not imagined while the software was created.

DTASelect's predecessors had not been altered for a few years when DTASelect was initiated. This stagnation had prevented updates to the file formats by which SEQUEST communicated identifications to the summary software. Likewise, the CGI utility programs had remained unaltered for years. Replacement of the summarization system made it possible to design new, more scalable file formats to accomodate the greater volume of data produced by MudPIT experiments. New tools to evaluate spectra and sequence coverage have helped to highlight new areas of research.

DTASelect's identification of the redundancy of spectra in proteomic collections led to an examination of this phenomenon. The extent of spectral duplication in these data sets was a surprise, given that instrument control software ostensibly prevents duplication. By increasing our understanding of the structure of proteomic spectral collections, this research provided insights that may lead to improved accuracy and speed in analyzing proteomic data.

### *7.1.2 Intensity modeling and de novo sequence tagging*

The initial design of this research, of course, had less to do with the creation of summarization software and more to do with improving models of CID fragmentation. Two challenges were encompassed: first, identified peptide spectra were to be statistically analyzed, and second, the observed trends were to be implemented in a model of fragmentation. Intensity of fragment ions was observed to depend upon neighboring peptide residues as well as the location of these ions within the spectrum. While other features undoubtedly influence fragment ion formation, these two principles gave a basis for initial modeling attempts.

This work describes a model produced from this system. The intensity of each fragment ion is first estimated on the basis of its mass relative to that of the full peptide and the series from which it comes. This intensity is then modified by a ratio incorporating information from the two amino acid residues flanking the fragment ion. If only one residue is known, the other residue is given no effect.

While this model was shown to be effective (see Chapter 6), it has important limitations. First, only doubly-charged precursor spectra can be evaluated. In addition, amino acid residues are pre-

sumed to have no effect on nonadjacent fragmentation, probably to the detriment of accuracy. Other fragment ions, such as water and ammonia losses from fragments, are not predicted by this model even though they were included in the statistical analysis. Finally, amino acids may have different results in combination than they do singly; a glycine residue may influence fragmentation differently if next to an alanine than it does when it is adjacent to a proline. These and other influences are not accounted for in the model as it currently exists.

The implementation of this scoring model in GutenTag has produced a tool with interesting properties. By automating partial sequence inference, this software makes sequence tag identification a usable strategy for high-throughput proteomics. The fragmentation model above is combined with a probability-based scorer to find the best tags for each spectrum. While this system is insufficiently powerful to rank correct sequence tags in the first position for each spectrum, the scorer is capable of producing at least one correct tag in the top 25 for the great majority of spectra. This capability was the direct result of the improved fragment ion intensity modeling.

The same fragmentation model is used to predict complete and partial spectra for sequences resulting from the database search. The scorer, this time a normalized dot-product algorithm, makes use of the intensities to determine the fit between the predicted and observed fragment ion intensities. Because a larger number of peaks is included in this round of scoring, the scoring is more effective than in the tagging phase of the software. These successes underscore the importance of modeling spectra accurately in proteomic software.

From a practical rather than a theoretical perspective, however, GutenTag's achievement is the identification of peptides with sequence modifications and / or post-translational modifications. The significance of single-nucleotide polymorphisms (SNPs) in real-world genomic specimens may correspond to single-amino-acyl polymorphisms (SAAPs, to coin an abbreviation) in proteomic specimens. By enabling peptide identification with these variances, the sequence tagging approach may grow to rival standard database search algorithms.

## *7.2 Future work*

In many respects, this research has only scratched the surface of an important and complex problem. As proteomic tandem mass spectrometry becomes more widely used, the algorithms developed to

process it will rely upon improved information characterizing these spectra. Several projects suggest themselves from this body of work.

### *7.2.1 Spectrum modeling*

As Chapter 4 indicates, further work remains for understanding the content of doubly-charged precursor CID spectra. The improved spectral viewing tools described in Chapter 3 revealed that doubly-charged fragment ions occur unexpectedly in these spectra. Understanding the causes for a fragment ion's retention of both protons will require further analysis. The impact of the C-terminal residue of a peptide on its fragmentation is also poorly understood. Do peptides terminating in arginine fragment differently than those ending in lysine? Do peptides ending in non-basic residues have more prominent *b* series ions than normal trypic peptides? Relatively small differences in peptide sequences may lead to significant changes in the appearances of the spectra.

Of course, proteomic spectra may come from other peptides than those with two additional protons. The fragmentation of singly-charged peptides may be the most inherently understandable; as the number of surplus protons decreases, the randomness of fragmentation may diminish as well. Initial examinations have suggested that other residues' chemistries may take the fore if protons are immobilized at arginine residues [1]. This and other explorations will be required for the fragmentation of these spectra to be modeled adequately.

By the same token, spectra from peptides with more than two added protons may exhibit a higher degree of randomness. Doubly-charged fragment ions may be comingled with singly-charged fragments, complicating analysis. Because these ions are separated by m/z rather than mass, the detailed analysis of doubly-charged fragment ions may require mass analyzers of higher mass accuracy than ion traps. The influences causing some of these fragments to be prominent and others to be missing remain nebulous.

### *7.2.2 Peptide identification improvements*

This research may pay other dividends for database identification algorithms. As shown in Chapter 5, spectral similarity can be detected without prior sequence identification. The prediction that this similarity can be leveraged for improved accuracy and performance, however, remains unverified

experimentally. If a database identification algorithm were designed such that spectra were grouped by similarity even before having charge states assigned, the full utility of these interrelationships could be tested.

Existing tools for assigning precursor ion charge states for proteomic spectra need to be redesigned. As described in section 1.4.2, singly-charged precursor spectra are separated from the others, and then the others are divided into doubly- and triply-charged precursor spectra. During development of GutenTag, however, it became apparent that these charge assignment algorithms may make errors, assigning singly-charged precursor spectra to doubly-charged precursors instead. Furthermore, particularly large peptides may be charged more highly than +3. As research characterizes spectra of each state more descriptively, charge discrimination software should evolve to incorporate the best possible models for differentiating these spectra by precursor charge.

### *7.2.3 Modification analysis*

When GutenTag identifies a spectrum with a partial sequence, it leaves the completion of that identification to the user. Hundreds of spectra may be matched to a particular protein in this way, and identifying which of these spectra shows a peptide with a significant, well-identified modification take considerable effort. The protein's full sequence must be sought in the sequence database, and then the difference between the peptide's observed mass and the sequence must be reduced by subtracting off the adjoining amino acid masses until the mass of the modification is reached. Simply determining how many amino acids make up the unidentified portion of the spectrum can be ambiguous. Once the full sequence and the modification mass are known, the task turns to finding which of the added amino acids plays host to the modification. GutenTag can be used to find a particular protein's modifications manually, but using it in complex mixtures to find all possible modifications on multiple proteins can pose a difficult challenge.

Software to automate the process of identifying modified peptides from partial sequence identifications would ease this process. Modifications which are evidenced by multiple spectra should take precendence over singletons. Likewise, peptides with overlapping sequences which cover the same modified residue can help to increase the confidence that the modification is legitimate. A scoring algorithm like that used in GutenTag could ascertain positions of modifications and rate the full-

sequence identification score. This software would make it much simpler to identify modifications for a complex mixture comprehensively and automatically.

### *7.3 Final thoughts*

The existing algorithms for processing proteomic tandem mass spectra set the stage for this research. SEQUEST made it possible to identify a set of spectra sufficient for characterization of these spectra. In many ways, this and other early tools bootstrapped proteomics into existence. It is because of the insights embodied in these programs that more ambitious studies can now be conducted on this class of spectra. It is the author's hope that this body of research will be instrumental in bringing about a new generation of informatic tools for proteomic data.

## END NOTES


1.1 Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nat. Biotechnol.* 2001, 19: 242-247.

1.2 Mann, M. ; Jensen, O. N. *Nat. Biotechnol.* 2003, 21: 255-261.

1.3 Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H. *Nat. Biotechnol.* 2000, 17: 994-999 and Ranish, J. A.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R. *Nat. Genet.* 2003, 33: 349-355.

1.4 Klose, J.; Kobalz, U. *Electrophoresis* 16: 1034-1059.

1.5 McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R. III. *Anal. Chem.* 1997, 69: 767-776 and Gatlin, C. L.; Kleeman, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R. III. *Anal. Biochem.* 1998, 263: 93-101.

1.6 Moyer, S. C.; Cotter, R. J. *Analytical Chemistry* 2002, 74: 468A-476A.

1.7 Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* 1989, 246: 64-71.

1.8 Moyer, S. C.; Cotter, R. J. *Analytical Chemistry* 2002, 74: 468A-476A.

1.9 Miller, P. E.; Denton, M. B. *J. Chem. Educ.* 1986, 63: 617-622 and Jonscher, K. R.; Yates, J. R. III *Anal. Biochem.* 1997, 244: 1-15.

1.10 Link, A. J. *Trends Biotechnol.* 2002, 20: S8-13.

1.11 Lawrence, J. B.; Oxvig, C.; Overgaard, M. T.; Sottrup-Jensen, L.; Gleich, G. J.; Hays, L. G.; Yates, J. R. 3rd; Conover, C. A. *Proc. Natl. Acad. Sci. USA* 1999, 96: 3149-3153.

1.12 Klose, J.; Kobalz, U. *Electrophoresis* 16: 1034-1059.

**1.13** McDonough, J. L.; Neverova, I.; Van Eyk, J. E: *Proteomics* 2002, 2: 978-987.

**1.14** Barrett, A. J; Rawlings, N. D.; Woessner, J. F. *Handbook of Proteolytic Enzymes* Elsevier Academic Press, San Diego, 2003, 1896 pp.

**1.15** McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R. III. *Anal. Chem.* 1997, 69: 767-776 and Gatlin, C. L.; Kleeman, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R. III. *Anal. Biochem.* 1998, 263: 93-101.

**1.16** Link, A. J.; Eng, J. K.; Schieltz D. M.; Carmack, E.; Mize G. J.; Morris D. R. et al. *Nat. Biotechnol.* 1999, 17: 676-682 and Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nat. Biotechnol.* 2001, 19: 242-247.

**1.17** Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* 1989, 246: 64-71.

**1.18** Miller, P. E.; Denton, M. B. *J. Chem. Educ.* 1986, 63: 617-622.

**1.19** Jonscher, K. R.; Yates, J. R. III *Anal. Biochem.* 1997, 244: 1-15.

**1.20** Afonso, C.; Modeste, F.; Breton, P.; Fournier, F; Tabet, J.-C. *Eur. J. Mass Spectrom.* 2000, 6: 443-449.

**1.21** Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* 2000, 35: 1399-1406.

**1.22** Huang, Y.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. III. *Intl. J. Mass Spectrom.* 2002, 219: 233-244.

**1.23** Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Anal. Chem.* 1993, 65: 425-438 and Breci, L. A.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. III. *Anal. Chem* 2003, 75: 1963-1971.

**1.24** Hunt, D. F.; Yates, J. R. III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. USA* 1986, 83: 6233-6237.

**1.25** Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* 1994, 5: 976-989.

**1.26** Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R. 3rd *J. Proteome Res.* 2002, 1: 211-215.

**1.27** Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* 1994, 5: 976-989.

**1.28** Perkins, D. N.; Pappin, J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* 1999, 20: 3551-3567.

**2.1** Tabb, D. L.; Eng, J. K.; Yates, J. R. III. Protein Identification by SEQUEST. In *Proteome Research: Mass Spectrometry*; James, P., Ed.; Springer: New York, 2001; Vol. 1, 125-142.

**2.2** Verma, R.; Chen, S.; Feldman, R.; Schieltz, D.; Yates, J. R. III; Dohmen, J.; Deshaies, R. J. *Mol. Biol. Cell* 2000, 11: 3425-3439.

**2.3** Verma, R.; McDonald, W. H.; Yates, J. R. III; Deshaies, R. J. *Molecular Cell.* 2001, 8: 439-48.

**2.4** Gatlin, C. L.; Kleeman, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R. III. *Anal. Biochem.* 1998, 263: 93-101 and Link, A. J.; Eng, J. K.; Schieltz D. M.; Carmack, E.; Mize G. J.; Morris D. R. et al. *Nat. Biotechnol.* 1999, 17: 676-682 and Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nat. Biotechnol.* 2001, 19: 242-247.

**2.5** Verma, R.; Chen, S.; Feldman, R.; Schieltz, D.; Yates, J. R. III; Dohmen, J.; Deshaies, R. J. *Mol. Biol. Cell* 2000, 11: 3425-3439.

**2.6** Verma, R.; Chen, S.; Feldman, R.; Schieltz, D.; Yates, J. R. III; Dohmen, J.; Deshaies, R. J. *Mol. Biol. Cell* 2000, 11: 3425-3439.

**2.7** Link, A. J.; Eng, J. K.; Schieltz D. M.; Carmack, E.; Mize G. J.; Morris D. R. et al. *Nat. Biotechnol.* 1999, 17: 676-682.

**2.8** Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nat. Biotechnol.* 2001, 19: 242-247.

**3.1** Verma, R.; McDonald, W. H.; Yates, J. R. III; Deshaies, R. J. *Molecular Cell.* 2001, 8: 439-48.

**3.2** SEQUEST Unified File Format. http://fields.scripps.edu/sequest/SQTFormat.html.

**3.3** ExtractMS. http://fields.scripps.edu/sequest/extractms.html.

**3.4** Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. *Nature* 2002, 419: 520-526 and Carlton, J. M.; Angiuoli, S. V.; Suh, B. B.; Kooij, T. W.; Pertea, M.; Silva, J. C.; Ermolaeva, M. D.; Allen, J. E.; Selengut, J. D.; Koo, H. L.; Peterson, J. D.; Pop, M.; Kosack, D. S.; Shumway, M. F.; Bidwell, S. L.; Shallom, S. J.; van Aken, S. E.; Riedmuller, S. B.; Feldblyum, T. V.; Cho, J. K.; Quackenbush, J.; Sedegah, M.; Shoaibi, A; Cummings, L. M.; Florens, L.; Yates, J. R.; Raine, J. D.; Sinden, R. E.; Harris, M. A.; Cunningham, D. A.; Preiser, P. R.; Bergman, L. W.; Vaidya, A. B.; van Lin, L. H.; Janse, C. J.; Waters, A. P.; Smith, H. O.; White, O. R.; Salzberg, S. L.; Venter, J. C.; Fraser, C. M.; Hoffman, S. L.; Gardner, M. J.; Carucci, D. J. *Nature* 2002, 419: 512-9.

**4.1** van Dongen, W. D.; Ruijters, H. F. M. ; Luinge, H.-J.; Heerma, W.; Haverkamp, J. *J. Mass Spectrom.* 1996, 31: 1156-1162.

**4.2** Dančík, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Computation. Bio.* 1999, 6: 327-342.

**4.3** Zhang, Z. *Proc. 50th Amer. Soc. Mass Spectrom.* 2002, Orlando, FL, TPE-126.

**4.4** Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nat. Biotechnol.* 2001, 19: 242-247.

**4.5** Link, A. J.; Eng, J. K.; Schieltz D. M.; Carmack, E.; Mize G. J.; Morris D. R. et al. *Nat. Biotechnol.* 1999, 17: 676-682.

**4.6** Lin, D.; Alpert, A. J.; Yates, J. R. III. *Amer. Genomic / Proteomic Tech.* 2001, 1: 38-46.

**4.7** MudPIT Homepage. http://fields.scripps.edu/mudpit/

**4.8** Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* 1994, 5: 976-989.

**4.9** Saccharomyces Genome Database http://genome-www.stanford.edu/Saccharomyces/

**4.10** Tabb, D. L.; McDonald, W. H.; Yates, J. R. 3rd. *J. Proteome Res.* 2002, 1: 21-26.

**4.11** DaughterDB home page. http://fields.scripps.edu/DaughterDB/

**4.12** Hornik, K. The R FAQ. http://www.ci.tuwien.ac.at/~hornik/R/ ISBN:3-901167-51-X

**4.13** Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Anal. Chem.* 1993, 65: 425-438 and Breci, L. A.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. III. *Anal. Chem* 2003, 75: 1963-1971.

**4.14** Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Anal. Chem.* 1993, 65: 425-438 and Breci, L. A.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. III. *Anal. Chem* 2003, 75: 1963-1971.

**4.15** Arnott, D.; Kottmeier, D.; Yates, N.; Shabanowitz, J.; Hunt, D. F. *Proc. 42nd ASMS Conf. Mass Spectrom. Allied Topics* 1994, Chicago, IL, 470 and Nold, M. J.; Wesdemiotis, C.; Yalcin, T.; Harrison, A. G. *Int. J. Mass Spectrom. Ion Process.* 1997, 164: 137-153 and Yalcin, T.; Khouw, C.; Csizmadia, I. G.; Peterson, M. R.; Harrison, A. G. *J. Am. Soc. Mass Spectrom.* 1995, 6: 1164-1174.

**4.16** Schlosser, A.; Wolf, D. L. *J. Mass Spectrom.* 2000, 35: 1382-1390 and Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* 2000, 35: 1399-1406 and Vaisar, T.; Urban, J. *J. Mass Spectrom.* 1996, 31: 1185-1187.

**4.17** Paizs, B.; Lendvay, G.; Vékey, K; Suhai, S. *Rapid Comm. Mass Spectrom.* 1999, 13: 525-533.

**4.18** Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* 2000, 35: 1399-1406.

**4.19** Farrugia, J. M.; Taverner, T.; O'Hair, R. A. J. *Int. J. Mass Spectrom.* 2001, 209: 99-112.

**4.20** Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* 2000, 35: 1399-1406 and Farrugia, J. M.; O'Hair, R. A. J.; Reid, G. E. *Int. J. Mass Spectrom.* 2001, 210: 71-87.

**4.21** Vachet, R. W.; Asam, M. R.; Glish, G. L. *J. Am. Chem. Soc.* 1996, 118: 6252-6256.

**4.22** Farrugia, J. M.; O'Hair, R. A. J.; Reid, G. E. *Int. J. Mass Spectrom.* 2001, 210: 71-87.

**4.23** Ballard, K. D.; Gaskell, S. J. *J. Amer. Soc. Mass Spectrom.* 1993, 4: 477-481.

**4.24** van Dongen, W. D.; de Koster, C. G.; Heerma, W.; Haverkamp, J. *Rapid Commun. Mass Spectrom.* 1995, 9: 845-850.

**4.25** Noren, C. J.; Wang, J. M.; Perler, F. B. *Angewandte Chemie Int. Ed.* 2000, 39: 450-466.

**4.26** Farrugia, J. M.; Taverner, T.; O'Hair, R. A. J. *Int. J. Mass Spectrom.* 2001, 209: 99-112.

**5.1** Lipton, M. S.; Paša-Tolić, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolić; Udseth, H. R.; Venkateswaran, A.; Wong, K.-K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci.* 2002, 99: 11049-11054 and Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hayes, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolters, D.; Yates, J. R. 3rd. *Proc. Natl. Acad. Sci.* 2002, 99: 11969-11974 and Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nat. Biotechnol.* 2001, 19: 242-247 and Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. *Nature* 2002, 419: 520-526 and VerBerkmoes, N. C.; Bundy, J. L.; Hauser, L.; Asano, K. G.; Razumovskaya, J.; Larimer, F.; Hettich, R. L.; Stephenson, J. L. Jr. *J. Proteome Res.* 2002, 1: 239-252.

**5.2** Thermo Finnigan Product Support Bulletin 105. http://www.thermofinnigan.com

**5.3** Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* 1994, 5: 976-989.

**5.4** ExtractMS home page. http://fields.scripps.edu/sequest/extractms.html.

**5.5** Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R. 3rd *J. Proteome Res.* 2002, 1: 211-215.

**5.6** Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* 1994, 5: 859-866 and Gan, F.; Yang, J.-H.; Liang, Y.-Z. *Anal. Sci.* 2001, 17: 635-638 and Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* 2002, 13: 85-88 and Yates, J. R. III. *J. Mass Spectrom.* 1998, 33: 1-19 and Hansen, B. T.; Jones, J. A.; Mason, D. E.; Liebler, D. C. *Anal. Chem.* 2001, 73: 1676-83.

**5.7** Yates, J. R. III. *J. Mass Spectrom.* 1998, 33: 1-19.

**5.8** Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* 1994, 5: 859-866.

**5.9** Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* 2002, 13: 85-88.

**5.10** Canettieri, G.; Morantte, I.; Guzmán, E.; Asahara, H.; Herzig, S.; Anderson, S. D.; Yates, J. R. 3rd.; Montminy, M. *Nat. Struct. Biol.* 2003, 10: 175-81.

**5.11** Mitchison Lab Protocols. http://mitchison.med.harvard.edu/protocols/tubprep.html.

**5.12** Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R. 3rd *Nat. Biotech.* 2003, 21: 532-8.

**5.13** McDonald, W. H.; Ohi, R.; Miyamoto, D. T.; Mitchison, T. J.; Yates, J. R. 3rd. *Intl. J. Mass Spectrom.* 2002, 219: 245-251.

**5.14** Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R. 3rd *J. Proteome Res.* 2002, 1: 211-215.

**5.15** MacCoss, M. J.; Wu, C. C.; Yates, J. R. 3rd. *Anal. Chem.* 2002, 74: 5593-9.

**5.16** National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/.

**5.17** Tabb, D. L.; McDonald, W. H.; Yates, J. R. 3rd. *J. Proteome Res.* 2002, 1: 21-26.

**5.18** SEQUEST Unified File Format. http://fields.scripps.edu/sequest/SQTFormat.html.

**5.19** Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* 1994, 5: 859-866.

**5.20** Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* 2003, 2: 43-50.

**5.21** Wolters, D. A.; Washburn, M. P.; Yates, J. R. III. *Anal. Chem.* 2001, 73: 5683-5690.

**5.22** Pevzner, P. A.; Dancik, V.; Tang, C. L. *J. Comput. Biol.* 2000, 7: 777-87.

**6.1** Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* 1995, 67: 1426-1436.

**6.2** Perkins, D. N.; Pappin, J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* 1999, 20: 3551-3567.

**6.3** Sakurai, T.; Matsuo, T.; Matsuda, H. *Biomed. Mass Spectrom.* 1984, 11, 396-399.

**6.4** Mann, M.; Wilm, M. *Anal. Chem.* 1994, 66: 4390-4399.

**6.5** Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. III. *Anal. Chem.* 2003, 75: 1155-1163.

**6.6** Bartels, C. *Biomed. Enviro. Mass Spectrom.* 1990, 19: 363-368.

**6.7** Dančík, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Computation. Bio.* 1999, 6: 327-342.

**6.8** Taylor, J. A.; Johnson, R. S. *Rapid Comm. Mass Spectrom.* 1997: 1067-1075.

**6.9** Fernandez-de-Cossio, J.; Gonzalez, J.; Satomi, Y.; Shima, T.; Okumura, N.; Besada, V.; Betancourt, L.; Padron, G.; Shimonishi, Y.; Takao, T. *Electrophoresis* 2000, 21: 1694-1699.

**6.10** Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nat. Biotechnol.* 2001, 19: 242-247.

**6.11** MacCoss, M. J.; Wu, C. C.; Yates, J. R. 3rd. *Anal. Chem.* 2002, 74: 5593-9.

**6.12** Tabb, D. L.; McDonald, W. H.; Yates, J. R. 3rd. *J. Proteome Res.* 2002, 1: 21-26.

**6.13** MacCoss, M. J.; Wu, C. C.; Yates, J. R. 3rd. *Anal. Chem.* 2002, 74: 5593-9.

**6.14** MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. III. *Proc. Natl. Acad. Sci. USA* 2002, 99: 7900-7905.

**6.15** Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. III. *Anal. Chem.* 2003, 75: 1155-1163.

**6.16** Kubinyi, H. *Analytica Chimica Acta* 1991, 247: 107-119.

**6.17** Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. III. *Anal. Chem.* 2003, 75: 1155-1163.

**6.18** Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. III. *Anal. Chem.* 2003, 75: 1155-1163.

**6.19** Chakravarti; Laha; Roy. *Handbook of Methods of Applied Statistics. Volume I* John Wiley and Sons, 1967: 392-394.

**6.20** Aho, A. V.; Corasick, M. J. *Comm. ACM* 1975, 18: 333-340.

**6.21** MacCoss, M. J.; Wu, C. C.; Yates, J. R. 3rd. *Anal. Chem.* 2002, 74: 5593-9.

**6.22** Krishna, R. G.; Chin, C. C.; Wold, F. *Anal. Biochem.* 1991, 199: 45-50.

**6.23** Nakano, K.; Hwang, P. K.; Fletterick, R. J. *FEBS Lett.* 1986, 204: 283-287.

**6.24** Fiat, A. M.; Jolles P. *Mol. Cell. Biochem.* 1989, 87: 5-30.

**6.25** MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. III. *Proc. Natl. Acad. Sci. USA* 2002, 99: 7900-7905.

**7.1** Huang, Y.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. III. *Intl. J. Mass Spectrom.* 2002, 219: 233-244.

## BIBLIOGRAPHY

[1] Afonso, C.; Modeste, F.; Breton, P.; Fournier, F; Tabet, J.-C. *Eur. J. Mass Spectrom.* 2000, 6: 443-449.

[2] Aho, A. V.; Corasick, M. J. *Comm. ACM* 1975, 18: 333-340.

[3] Arnott, D.; Kottmeier, D.; Yates, N.; Shabanowitz, J.; Hunt, D. F. *Proc. 42nd ASMS Conf. Mass Spectrom. Allied Topics* 1994, Chicago, IL, 470.

[4] Ballard, K. D.; Gaskell, S. J. *J. Amer. Soc. Mass Spectrom.* 1993, 4: 477-481.

[5] Barrett, A. J; Rawlings, N. D.; Woessner, J. F. *Handbook of Proteolytic Enzymes* Elsevier Academic Press, San Diego, 2003, 1896 pp.

[6] Bartels, C. *Biomed. Enviro. Mass Spectrom.* 1990, 19: 363-368.

[7] Breci, L. A.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. III. *Anal. Chem* 2003, 75: 1963-1971.

[8] Canettieri, G.; Morantte, I.; Guzmán, E.; Asahara, H.; Herzig, S.; Anderson, S. D.; Yates, J. R. 3rd.; Montminy, M. *Nat. Struct. Biol.* 2003, 10: 175-81.

[9] Carlton, J. M.; Angiuoli, S. V.; Suh, B. B.; Kooij, T. W.; Pertea, M.; Silva, J. C.; Ermolaeva, M. D.; Allen, J. E.; Selengut, J. D.; Koo, H. L.; Peterson, J. D.; Pop, M.; Kosack, D. S.; Shumway, M. F.; Bidwell, S. L.; Shallom, S. J.; van Aken, S. E.; Riedmuller, S. B.; Feldblyum, T. V.; Cho, J. K.; Quackenbush, J.; Sedegah, M.; Shoaibi, A; Cummings, L. M.; Florens, L.; Yates, J. R.; Raine, J. D.; Sinden, R. E.; Harris, M. A.; Cunningham, D. A.; Preiser, P. R.; Bergman, L. W.; Vaidya, A. B.; van Lin, L. H.; Janse, C. J.; Waters, A. P.; Smith, H. O.; White, O. R.; Salzberg, S. L.; Venter, J. C.; Fraser, C. M.; Hoffman, S. L.; Gardner, M. J.; Carucci, D. J. *Nature* 2002, 419: 512-9.

[10] Chakravarti; Laha; Roy. *Handbook of Methods of Applied Statistics. Volume I* John Wiley and Sons, 1967: 392-394.

[11] Dančík, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Computation. Bio.* 1999, 6: 327-342.

[12] Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* 1994, 5: 976-989.

[13] Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* 1995, 67: 1426-1436.

[14] Farrugia, J. M.; Taverner, T.; O'Hair, R. A. J. *Int. J. Mass Spectrom.* 2001, 209: 99-112.

[15] Farrugia, J. M.; O'Hair, R. A. J.; Reid, G. E. *Int. J. Mass Spectrom.* 2001, 210: 71-87.

[16] Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* 1989, 246: 64-71.

[17] Fernandez-de-Cossio, J.; Gonzalez, J.; Satomi, Y.; Shima, T.; Okumura, N.; Besada, V.; Betancourt, L.; Padron, G.; Shimonishi, Y.; Takao, T. *Electrophoresis* 2000, 21: 1694-1699.

[18] Fiat, A. M.; Jolles P. *Mol. Cell. Biochem.* 1989, 87: 5-30.

[19] Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. *Nature* 2002, 419: 520-526.

[20] Gan, F.; Yang, J.-H.; Liang, Y.-Z. *Anal. Sci.* 2001, 17: 635-638.

[21] Gatlin, C. L.; Kleeman, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R. III. *Anal. Biochem.* 1998, 263: 93-101.

[22] Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H. *Nat. Biotechnol.* 2000, 17: 994-999.

[23] Hansen, B. T.; Jones, J. A.; Mason, D. E.; Liebler, D. C. *Anal. Chem.* 2001, 73: 1676-83.

[24] Hornik, K. The R FAQ. http://www.ci.tuwien.ac.at/~hornik/R/ ISBN:3-901167-51-X

[25] Huang, Y.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. III. *Intl. J. Mass Spectrom.* 2002, 219: 233-244.

[26] Hunt, D. F.; Yates, J. R. III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. USA* 1986, 83: 6233-6237.

[27] Jonscher, K. R.; Yates, J. R. III *Anal. Biochem.* 1997, 244: 1-15.

[28] Klose, J.; Kobalz, U. *Electrophoresis* 16: 1034-1059.

[29] Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hayes, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolters, D.; Yates, J. R. 3rd. *Proc. Natl. Acad. Sci.* 2002, 99: 11969-11974.

[30] Krishna, R. G.; Chin, C. C.; Wold, F. *Anal. Biochem.* 1991, 199: 45-50.

[31] Kubinyi, H. *Analytica Chimica Acta* 1991, 247: 107-119.

[32] Lawrence, J. B.; Oxvig, C.; Overgaard, M. T.; Sottrup-Jensen, L.; Gleich, G. J.; Hays, L. G.; Yates, J. R. 3rd; Conover, C. A. *Proc. Natl. Acad. Sci. USA* 1999, 96: 3149-3153.

[33] Lin, D.; Alpert, A. J.; Yates, J. R. III. *Amer. Genomic / Proteomic Tech.* 2001, 1: 38-46.

[34] Link, A. J.; Eng, J. K.; Schieltz D. M.; Carmack, E.; Mize G. J.; Morris D. R. et al. *Nat. Biotechnol.* 1999, 17: 676-682.

[35] Link, A. J. *Trends Biotechnol.* 2002, 20: S8-13.

[36] Lipton, M. S.; Paša-Tolić, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolić; Udseth, H. R.; Venkateswaran, A.; Wong, K.-K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci.* 2002, 99: 11049-11054.

[37] Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Anal. Chem.* 1993, 65: 425-438.

[38] MacCoss, M. J.; Wu, C. C.; Yates, J. R. 3rd. *Anal. Chem.* 2002, 74: 5593-9.

[39] MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. III. *Proc. Natl. Acad. Sci. USA* 2002, 99: 7900-7905.

[40] Mann, M.; Wilm, M. *Anal. Chem.* 1994, 66: 4390-4399.

[41] Mann, M. ; Jensen, O. N. *Nat. Biotechnol.* 2003, 21: 255-261.

[42] McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R. III. *Anal. Chem.* 1997, 69: 767-776.

[43] McDonald, W. H.; Ohi, R.; Miyamoto, D. T.; Mitchison, T. J.; Yates, J. R. 3rd. *Intl. J. Mass Spectrom.* 2002, 219: 245-251.

[44] McDonough, J. L.; Neverova, I.; Van Eyk, J. E. *Proteomics* 2002, 2: 978-987.

[45] Miller, P. E.; Denton, M. B. *J. Chem. Educ.* 1986, 63: 617-622.

[46] Moyer, S. C.; Cotter, R. J. *Analytical Chemistry* 2002, 74: 468A-476A.

[47] Nakano, K.; Hwang, P. K.; Fletterick, R. J. *FEBS Lett.* 1986, 204: 283-287.

[48] Nold, M. J.; Wesdemiotis, C.; Yalcin, T.; Harrison, A. G. *Int. J. Mass Spectrom. Ion Process.* 1997, 164: 137-153.

[49] Noren, C. J.; Wang, J. M.; Perler, F. B. *Angewandte Chemie Int. Ed.* 2000, 39: 450-466.

[50] Paizs, B.; Lendvay, G.; Vékey, K; Suhai, S. *Rapid Comm. Mass Spectrom.* 1999, 13: 525-533.

[51] Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* 2003, 2: 43-50.

[52] Perkins, D. N.; Pappin, J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* 1999, 20: 3551-3567.

[53] Pevzner, P. A.; Dancik, V.; Tang, C. L. *J. Comput. Biol.* 2000, 7: 777-87.

[54] Ranish, J. A.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R. *Nat. Genet.* 2003, 33: 349-355.

[55] Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R. 3rd *J. Proteome Res.* 2002, 1: 211-215.

[56] Sakurai, T.; Matsuo, T.; Matsuda, H. *Biomed. Mass Spectrom.* 1984, 11, 396-399.

[57] Schlosser, A.; Wolf, D. L. *J. Mass Spectrom.* 2000, 35: 1382-1390.

[58] Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* 1994, 5: 859-866.

[59] Tabb, D. L.; Eng, J. K.; Yates, J. R. III. Protein Identification by SEQUEST. In *Proteome Research: Mass Spectrometry*; James, P., Ed.; Springer: New York, 2001; Vol. 1, 125-142.

[60] Tabb, D. L.; McDonald, W. H.; Yates, J. R. 3rd. *J. Proteome Res.* 2002, 1: 21-26.

[61] Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. III. *Anal. Chem.* 2003, 75: 1155-1163.

[62] Taylor, J. A.; Johnson, R. S. *Rapid Comm. Mass Spectrom.* 1997: 1067-1075.

[63] Vachet, R. W.; Asam, M. R.; Glish, G. L. *J. Am. Chem. Soc.* 1996, 118: 6252-6256.

[64] Vaisar, T.; Urban, J. *J. Mass Spectrom.* 1996, 31: 1185-1187.

[65] van Dongen, W. D.; de Koster, C. G.; Heerma, W.; Haverkamp, J. *Rapid Commun. Mass Spectrom.* 1995, 9: 845-850.

[66] van Dongen, W. D.; Ruijters, H. F. M. ; Luinge, H.-J.; Heerma, W.; Haverkamp, J. *J. Mass Spectrom.* 1996, 31: 1156-1162.

[67] VerBerkmoes, N. C.; Bundy, J. L.; Hauser, L.; Asano, K. G.; Razumovskaya, J.; Larimer, F.; Hettich, R. L.; Stephenson, J. L. Jr. *J. Proteome Res.* 2002, 1: 239-252.

[68] Verma, R.; Chen, S.; Feldman, R.; Schieltz, D.; Yates, J. R. III; Dohmen, J.; Deshaies, R. J. *Mol. Biol. Cell* 2000, 11: 3425-3439.

[69] Verma, R.; McDonald, W. H.; Yates, J. R. III; Deshaies, R. J. *Molecular Cell.* 2001, 8: 439-48.

[70] Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* 2002, 13: 85-88.

[71] Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nat. Biotechnol.* 2001, 19: 242-247.

[72] Wolters, D. A.; Washburn, M. P.; Yates, J. R. III. *Anal. Chem.* 2001, 73: 5683-5690.

[73] Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R. 3rd *Nat. Biotech.* 2003, 21: 532-8.

[74] Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* 2000, 35: 1399-1406.

[75] Yalcin, T.; Khouw, C.; Csizmadia, I. G.; Peterson, M. R.; Harrison, A. G. *J. Am. Soc. Mass Spectrom.* 1995, 6: 1164-1174.

[76] Yates, J. R. III. *J. Mass Spectrom.* 1998, 33: 1-19.

[77] Zhang, Z. *Proc. 50th Amer. Soc. Mass Spectrom.* 2002, Orlando, FL, TPE-126.

# VITA

David L. Tabb

*Education*

- University of Washington, Dept. of Molecular Biotechnology / The Scripps Research Institute, Dept. of Cell Biology. Doctor of Philosophy: June 2003

- University of Arkansas, Major in Biology, Minor in Computer Science, Summa cum laude, Bachelor of Science: May 1996

---

*Professional Experience*

- Non-TSRI Graduate Student, TSRI Department of Cell Biology

  Developed DTASelect proteomic data mining tool, Contrast proteomic differentiation software, and DaughterDB peptide identification analyzer. Constructed and maintained two Linux file servers totalling more than 2 TB of data. 2000 - present.

- Graduate Student, UW Department of Molecular Biotechnology

  Developed El CID *de novo* peptide sequence inference software. Provided expertise for assembling and maintaining Windows PCs. 1996 - 2000.

- Data Processing Software Director for KAW Trucking

  Maintained and repaired extensive UNIX FoxPro database systems for national transportation company. 1996.

- Scientific Applications Programmer for G.D. Searle and Company

Developed FoxPro databases in Windows and Macintosh environments for document storage and hardware / software cataloguing. Managed team of 10 temporary workers in sitewide computer installation. 1995.

---

*Grants and Awards*

- 1996 National Science Foundation Graduate Research Fellowship
- 1994 Research Opportunity with University of Lyon, France
- 1992 National Science Foundation National Science Scholar
- 1992 White House Presidential Scholar
- 1992 Sturgis Fellow
- 1992 Tandy Technology Scholar

---

*Publications*

- Breci, LA; Wysocki, VH; Tabb, DL; Yates, JR III. Analysis of cleavage N-terminal to proline using a database of peptide tandem mass spectra. *Anal. Chem.* 2003 75: 1963-1971.
- Tabb, DL; MacCoss, MJ; Wu, CC; Anderson, SD; Yates, JR III. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.* 2003 75: 2470-2477.
- Tabb, DL; Smith, LL; Breci, LA; Wysocki, VH; Lin, D; Yates, JR III. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* 2003 75: 1155-1163.
- Lin, D; Tabb, DL; Yates, JR III. Large-scale protein identification using mass spectrometry. *Biochimica et Biophysica Acta* 2003 1646: 1-10.

- Florens, L; Washburn, MP; Raine, JD; Anthony, RM; Grainger, M; Haynes, JD; Moch, JK; Muster, N; Sacci, JB; Tabb, DL; Witney, AA; Wolters, D; Wu, Y; Gardner, MJ; Holder, AA; Sinden, RE; Yates, JR; Carucci, DJ. A proteomic view of the Plasmodium falciparum life cycle. *Nature* 2002 419(6906): 520-6.

- Tabb, DL; McDonald, WH; Yates, JR 3rd. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 2002 1:21-26.

- Huang, Y; Wysocki, VH; Tabb, DL; Yates, JR 3rd. The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *Intl. J Mass Spectrom.* 2002, 219:233-244.

- Tabb, DL; Eng, JK; Yates, JR 3rd. Protein Identification by SEQUEST. *Proteome Research: Mass Spectrometry.* James, P ed. Springer, New York, 2001 1:125-142.

- Krahmer. MT; Walters, JJ; Fox, KF; Fox, A; Creek, KE; Pirisi, L; Wunschel, DS; Smith, RD; Tabb, DL; Yates, JR 3rd. MS for identification of single nucleotide polymorphisms and MS/MS for discrimination of isomeric PCR products. *Anal. Chem.* 2000 72:4033-40.


---

*Presentations*

- Oak Ridge National Lab, Life Sciences Division, 2003
- University of Washington, Dept. of Genome Sciences, 2003
- Merck Research Labs, 2003
- University of Southern California, School of Pharmacy, 2002
- Oak Ridge National Lab, Life Sciences Division, 2002
- California Institute of Technology, Dept. of Biology, 2002
- University of Arkansas, Dept. of Biology, 2002

- Association for Laboratory Automation, 2002
- University of Arizona, Dept. of Chemistry, 2001
- Proteome Society, San Diego, 2001
- University of Toronto, Dept. of Medical Genetics, 2001
- Salk Institute, Peptide Biology Laboratories, 2001
- LinuxFest, 2000
- University of Washington Interdisciplinary Graduate Student Symposium, 1999 (poster)
- University of Washington, Dept. of Biochemistry, 1998