PROTEOMICS METHODS FOR DETECTION OF MODIFIED PEPTIDES

by

Beau Tanana Hansen

A Dissertation Submitted to the Faculty of the

COMMITTEE ON PHARMACOLOGY AND TOXICOLOGY (GRADUATE)

In Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2 0 0 5

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Beau Tanana Hansen

entitled Proteomics Methods for Detecting Protein Modifications

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: 4/18/05
Daniel C. Liebler

_____ Date: 4/18/05
A. Jay Gandolfi

_____ Date: 4/18/05
John Regan

_____ Date: 4/18/05
Barbara Timmermann

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 4/18/05
Dissertation Directors: Daniel C. Liebler / A. Jay Gandolfi

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirement for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: _____

ACKNOWLEDGEMENTS

The author would like to acknowledge the contribution of several individuals who contributed to the research described in this dissertation. Dr. Liebler came up the original concept of studying electrophile-adducted peptide libraries by mass spectrometry, with the objective of identifying diagnostic fragmentation patterns in the spectra of adducted peptides. It was his ultimate goal to use these fragmentation patterns as the basis for the development of a software program that could be used to screen the outputs from LC-MS-MS experiments for evidence of adducted peptides. With his guidance I developed the prototypes for both the SALSA and P-Mod algorithms. Sean Davey was instrumental in converting these prototypes to stand-alone Windows applications. Other members of the Liebler Laboratory contributed data that was important for validating the algorithm prototypes. Specifically the author acknowledges the contributions of Laura Tiscareno, Juliet Jones, and Daniel Mason.

Instrumentation for this research was made available by the Southwest Environmental Health Sciences Center Analytical Core and ThermoFinnigan. The author thanks Dr George Tsaprailis and Sherry Daugherty of the analytical core for their continuous assistance with the equipment and for their positive and professional demeanor.

The author is also very grateful for the support and encouragement from Dr. Zeynep Hansen, the author's wife.

Funding for this research was provided by the following grants: NIH ES06694, NIH ES10056.

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACN	acetonitrile
amu	atomic mass unit
APAP	acetaminophen
Ambic	ammonium bicarbonate
BSA	bovine serum albumin
BQ	benzoquinone
CASH	cortical androgen stimulating hormone
CID	collision induced dissociation
CL	charged loss
Da	Dalton
DHP	dehydropyrole
DTT	dithiothreitol
ESI	electrospray
ESI-tandem-MS	electrospray tandem mass spectrometry
Frog24	atrial natriuretic peptide, frog
GSH	glutathione
HPLC	high performance liquid chromatography
IAA	iodoacetic acid
ICAT	isotope coded affinity tags
LCQ	quadrupole ion trap manufactured by ThermoFinnigan
LC-MS	liquid chromatography mass spectrometry

MALDI	matrix-assisted laser desorption ionization
MS	mass spectrometry
MS-MS	tandem mass spectrometry
MS^3	MS-MS-MS
m/z	mass to charge ratio
NL	neutral loss
2D-PAGE	two dimensional polyacrylamide gel electrophoresis
PepC	peptide, sequence AGAGCAGAG
ppm	parts per million
SALSA	Scoring ALgorithm for Spectral Analysis
SDS-PAGE	sodium dodecyl sulfate-PAGE
TCEP	*tris*-(2-carboxyethyl)phosphine hydrochloride
TFA	trifluoroacetic acid
TIC	total ion current
%TIC	percent total ion current
TpepC	peptide, sequence AVAGCAGAR
TOF	time-of-flight

ABSTRACT

The recent emergence of the field of proteomics has been driven by advances in mass spectrometry methods and instrumentation. Due to the large amount of data generated, success at peptide and protein identification is contingent on reliable software algorithms. The software programs in use at the time the work in this dissertation was carried out were well suited to the task of identifying unmodified peptides and proteins in complex mixtures. However, the existing programs were not able to reliably identify protein modifications, especially unpredicted modifications. This dissertation describes the development of two novel software algorithms that can be used to screen LC-MS-MS data files, and identify MS-MS spectra that correspond to peptides with either predicted or unpredicted modifications. The first program, SALSA, is highly flexible and uses user defined search criteria to screen data files for spectra the exhibit fragmentation patterns diagnostic of specific modifications or peptide sequences. SALSA facilitates exhaustive searches, but requires user expertise to both generate search criteria and to validate matched spectra. The second program, P-Mod, provides automated searches for spectra corresponding to peptides in a search list. P-Mod is able to identify spectra derived from either modified or unmodified peptides. All sequence-to-spectrum matches reported in the P-Mod output are assigned statistical confidence levels derived using extreme value statistics.

CHAPTER ONE – Introduction

MS Instrumentation for Proteomics

Advances in mass spectrometry (MS) methods and instrumentation have driven the emergence of proteomics over the last decade (1, 2). Much of the initial research in proteomics has been focused on the identification and quantification of proteins in complex mixtures. However, the methods used in proteomics research can potentially provide highly detailed information about protein modifications caused by either endogenous mechanisms or reactions with reactive xenobiotics. This dissertation describes in detail the development of two novel data analysis algorithms that can be integrated with existing proteomics methods in order to detect and characterize protein modifications derived from a variety of sources. In order to appreciate how these algorithms operate and the impetus behind their development it is essential that the reader have a basic understanding of proteomics methods and instrumentation.

There are multiple configurations of mass spectrometers available for research in proteomics, each with different strengths and weaknesses. All mass spectrometers are composed of two main parts; an ionization source and a mass analyzer. In the source, analytes are freed from the sample matrix and energized to produce charged gas phase ions. Mass spectrometers come equipped with many different sources. Of the various source options, Matrix Assisted Laser Desorption Ionization (MALDI), and Electrospray Ionization (ESI) have proven to be the most useful in analyses of large non-volatile biomolecules such as peptides or proteins (3). Analyte ions generated in the source are

directed into a mass analyzer which separates ions on the basis of their mass-to-charge ratio (*m/z*), after which a detector records the abundance of ions in each *m/z* category.

Both MALDI and ESI are widely used in proteomics, although the two sources are best suited for different applications. Prior to analysis on an instrument equipped with a MALDI source, samples are mixed with an UV-absorbing matrix and spotted on a target. When bombarded by laser light in the source, the matrix assists in the transfer of light energy to molecules in the sample and the generation of gas phase ions. Predominately singly charged ions are produced by MALDI and molecules with molecular weights in excess of 200,000 daltons can be ionized. When a MALDI source is coupled to a time of flight mass analyzer, it is possible to acquire highly accurate mass estimates for very large proteins. Because peptide ions are predominately singly charged, and because tandem MS (MS-MS) methods are less developed for MALDI than ESI, MALDI equipped mass spectrometers are best suited for the identification of proteins by peptide mass fingerprinting which will be discussed in greater detail below.

By comparison, ESI can be accomplished on molecules in solution without the use of a matrix (4). Samples do not have to be prepared and spotted on a target prior to analysis. Ionization in ESI occurs in solution, and gas phase ions are produced by desolvation. The solvent solution entering the mass spectrometer is nebulized to produce a spray of very fine droplets containing the analyte ions. The droplets pass through a heated tube where droplet size is decreased through evaporation until such time that the concentration of charge in the droplets causes them to explode due to colombic interactions into even smaller droplets. Continuation of this process leads to ever finer

droplets until eventually all of the solvent has been stripped away and all that is left are gas phase analyte ions. In contrast to MALDI, multiply charged ions are often produced by ESI, and large molecules in excess of 100,000 daltons are less efficiently ionized. The production of multicharged ions in ESI reduces the *m/z* of large molecules such as proteins. Consequently, ESI is often coupled with quadrupole or ion trap mass analyzers which typically have a limited mass range of 2,000 daltons or less. Mass spectrometers equipped with an ESI source can be linked directly to a High Performance Liquid Chromatography (HPLC) system in order to analyze molecules as they elute from an HPLC column. Such online coupling is precluded on instruments equipped with a MALDI source because of the need for matrix and laser excitation. Online separation is a critical component in the analysis of complex mixtures such as peptide digests of unpurified protein samples. Applications of MALDI to protein and peptide analyses are limited to relatively simple mixtures, and require considerable prefractionation and/or purification of samples prior to being introduced to the instrument. In general, the greater the capacity for online separation, the greater the complexity of mixtures that can be analyzed and the greater the sensitivity for detecting low abundance proteins or peptides.

Multiply charged peptide ions produced by ESI have more stored energy and have greater potential for fragmentation in MS-MS experiments than the singly charged ions produced by MALDI. MS-MS fragmentation of peptides is accomplished by first selecting and isolating peptide ions of a particular mass, then colliding these ions into either an inert gas or solid surface in order to induce dissociation into a population of fragment ions. The resulting fragment ions are then passed through a second mass

analyzer, or are reanalyzed by the original mass analyzer, to generate a MS-MS spectrum. MS-MS spectra represent the relative abundance and *m/z* of the fragment ions produced by the dissociation the selected peptide ions. As such MS-MS spectra are rich in structural information pertinent to the determination of peptide sequence and potentially the mass and sequence location of peptide modifications.

Mass analyzers fall into one of three categories: time of flight, quadrupole, or quadrupole-ion trap. Time of flight (TOF) mass analyzers are more or less long tubes with the source at one end and the detector at the other. Ions leaving the source are accelerated into the TOF tube with a constant kinetic energy; however, ions with different *m/z* are imparted with different velocities and therefore travel the length of the tube at different rates. Low *m/z* ions have greater velocity and reach the detector ahead of their high *m/z* counterparts (5). The resolving power, and therefore the mass accuracy, of early TOF mass analyzers was limited due to broadening of ion populations in the source. Incorporation of delayed extraction and reflectron technologies has greatly increased their mass accuracy (6). TOF mass analyzers are able to measure *m/z* over an extended range and are therefore frequently coupled with a MALDI source to take advantage of the production of singly charged high molecular weight ions.

Quadrupole mass analyzers employ a combination of radio frequency (RF) and direct current (DC) fields to act as a mass filter. At a particular ratio of RF amplitude to DC voltage an electric field is established which only allows ions of a particular *m/z* to pass through the length of the quadrupole rods; all other ions have unstable trajectories and end up crashing into the rods prior to reaching the outlet of the mass analyzer. In

order to scan a *m/z* range the RF amplitude and DC voltages are ramped at a fixed ratio. Quadrupoles have lower sensitivity and resolving power compared to TOF mass analyzers. Quadrupoles also have a significantly reduced *m/z* range and are only able to analyze ions with a *m/z* less than 4000. Nonetheless, quadropoles are frequently used because they are relatively cheep and robust; and they are available in configurations which facilitate MS-MS experiments. Because multiply charged ions produced by ESI have lower *m/z* ratios, these mass analyzers are usually coupled with an ESI source to compensate for the restricted *m/z* range (7).

Ion trap mass analyzers operate by similar principles as quadrupoles, with the exception that they use alternating RF and DC currents to trap ions in three dimensional space (8). Newer linear ion trap analyzers trap ions in two dimensional space (9). Ion traps are the most versatile mass analyzers from the point of view of proteomics research. Unlike TOF and quadrupole mass analyzers, ion traps do not require any special tandem configurations to carry out true MS-MS experiments. Ion traps are also exceptionally sensitive, with newer instruments being capable of acquiring quality MS-MS spectra for peptide ions with low femtomolar concentrations when configured with a nanoelectrospray source. Ion traps are used primarily for automated "data-dependent" acquisition of peptide MS-MS spectra (10). In "data-dependent" scan mode, the ion trap alternates between MS and MS-MS modes. In the MS mode the instrument determines the highest intensity ion in the sample and selects that ion for MS-MS analysis. After acquiring a MS-MS spectrum of the selected ion, the instrument software places the selected *m/z* on an exclusion list. The instrument then carries out another MS scan and

selects the highest intensity ion with a *m/z* not on the exclusion list for MS-MS. "Data-dependent" scanning is exceptionally useful for LC-MS-MS analyses of complex mixtures. Because of this, and because ion traps like quadrupoles have a restricted *m/z* range, these mass analyzers are almost exclusively coupled with ESI sources. Ion traps typically have unit resolution, meaning that *m/z* estimates are accurate to ± 0.5 *m/z*.

An additional feature of ion trap mass analyzers is that they are capable of MS^n experiments. That is, fragment ions in MS-MS spectra can be selected for fragmentation leading to the generation of MS-MS-MS spectra. Fragment ions in these spectra can also be selected for still higher order fragmentation. This feature can assist in the verification of assigned fragment ions in MS-MS spectra. For some peptide modifications, such as pyrollizidine alkaloid adducts, which were researched as part of this dissertation, adduction leads to the disruption of normal peptide CID patterns. For adducted peptides of this type, MS-MS-MS provides the only means to determine peptide sequence.

These basic mass analyzers are packaged in different configurations which expand their capabilities. For example, quadrupole mass analyzers are often the first component in tandem mass analyzers. Triple quadrupole mass analyzers are very common and facilitate MS-MS analyses. In instruments with this configuration, the first quadrupole serves as a mass filter to select an ion with a particular *m/z*, the second quadrupole provides an environment for collision induced dissociation of the selected ion, and the third quadrupole operates as a mass analyzer to acquire a MS-MS spectrum of the fragment ions. Because these instruments do not have the "data-dependent" scanning capabilities of ion traps and have lower sensitivity they are not nearly as useful

for acquiring peptide spectra from complex mixtures. However, they still have utility quantifying ions by selected ion monitoring, and can be used to detect certain classes of molecules or peptide modifications by constant neutral loss or product ion scans.

In a similar fashion, quadrupoles can be coupled with TOF and ion trap mass analyzers to produce quadrupole-TOF and quadrupole-ion trap instruments. TOF mass analyzers can also be combined to produce a TOF-TOF mass analyzer. MALDI-TOF-TOF instruments are capable of MS-MS experiments, and are capable of high energy collision induced dissociation which can be useful for distinguishing between isobaric amino acids such as leucine and isoleucine or lysine and glutamine. These instruments also have exceptional sensitivity and resolution. However, singly charged precursors produced by the MALDI source undergo less extensive fragmentation than the multi-charged precursors generated by instruments with an ESI source and therefore result in less informative MS-MS spectra.

Mass Spectrometry Approaches to Protein Analysis

A modern proteomics laboratory equipped with appropriate mass spectrometry instrumentation is capable of high throughput identification and quantification of proteins, as well as detailed structural analysis that can lead to the elucidation of protein modifications. There are two main approaches for protein identification used in proteomics research: peptide mass fingerprinting and peptide sequencing from MS-MS spectra. Both approaches involve the analysis of peptide sequences derived from either the enzymatic or chemical digestion of cellular proteins. However, there are significant differences between the two approaches, some of which have important implications for the analysis of protein modifications.

Even though the term "proteomics" implies that it is possible to carry out high throughput analyses of cellular proteomes, this is a bit of a misnomer. Considerable sample preparation is still required prior to introducing proteins into the mass spectrometer. In any given eukaryotic cell there are over 20,000 expressed proteins. To complicate matters, there can be over 6 orders of magnitude difference in the expression levels between various proteins. Thus, at least some degree of protein purification is usually essential; the extent of purification required depends on the MS approach to be used and objectives of the research. In order for peptide mass fingerprinting to be successful it is necessary to purify proteins to the point that the sample only contains 1-3 proteins of any significant concentration. Because of the increased information contained in MS-MS spectra and because the peptide sequencing approach to protein analysis usually involves coupling with online liquid chromatography (LC), much more complex

samples can be analyzed. However, even when LC-MS-MS methods are used, some degree of fractionation of the proteome is required if one hopes to identify low abundance proteins or characterize protein modifications.

Analyses can be simplified by prefractionating the proteome by one of several methods. Various sub-cellular fractionation kits are available for isolating cytosolic, nuclear, mitochondrial, or membrane bound proteins. The proteome may also be fractionated by preparative isoelectric focusing or liquid chromatography. Either way, the goal is to reduce the complexity of the protein sample to be analyzed and possibly target proteins of interest. For more targeted studies, imunoprecipitation or affinity chromatography may be used to purify specific proteins. Alternatively, proteins may be targeted by cutting bands out of either 1-D or 2-D sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gels. 1-D SDS-PAGE gels separate proteins on the basis of size, while 2-D gels separate proteins both on the basis of size and isoelectric point. On some large format 2-D gels it is possible to resolve as many as 10,000 separate protein spots. However, it should be noted that even at this level of resolution, most spots contain multiple proteins, and most spots on SDS-PAGE gels of whole cell lysates correspond to high abundance proteins (11). Even using the most sensitive florescent dyes or silver staining to visualize protein spots, low abundance proteins are usually not observed unless previously enriched or samples are loaded onto the gels in high milligram quantities, which can lead to decreased spot resolution.

Intact proteins purified or otherwise can be analyzed directly by MALDI-TOF to determine protein mass with a high degree of accuracy and achieve tentative

identification. Because protein masses are not unique, with several proteins sharing the same nominal mass, conclusive identification of unknown whole proteins is not possible using this technique; moreover, the presence of protein modifications further complicate the situation. Likewise the mass of purified proteins can be determined by ESI on a quadrupole or ion trap instrument; however, mass estimates are not as accurate on these instruments further decreasing the reliability of protein identification. Much more precise protein identification can be achieved through evaluation of the peptides produced by digestion of sample proteins. Consequently, protein samples that have been purified or enriched using one of the above techniques are subjected to enzymatic or chemical digestion prior to MS analysis.

Before digestion, proteins are denatured by heat and incubation in either urea or ammonium bicarbonate so as to disrupt tight protein folding that may restrict access to proteolytic sites. Recently, the use of urea has become less common due to artifactual carbamylation of lysine residues and protein and peptide N-termini. Protein disulfide bonds are reduced with either dithiothreitol (DTT) or *tris*-(2-carboxyethyl)phosphine hydrochloride (TCEP) to assist in denaturation and to ensure that proteins do not refold when buffer levels are decreased. A number of proteolytic enzymes are available for protein digestion, each of which with slightly different characteristics and amino acid specificity (Table 1-1). Of these, the most commonly used is trypsin. Digestion with trypsin cleaves proteins on the C-terminal side of lysine and arginine residues except in cases where these residues are followed by the amino acid proline. Typically tryptic digestion results in peptides with an average length of 5 to 25 amino acids, ensuring a

Table 1-1. Amino acid specificity of commonly used proteases

Enzyme	Cleavage Specificity
Trypsin	C-terminal side of K and R not preceding P
Chymotrypsin	C-terminal side of Y, W, L, M, A, D, E
Endoproteinase Glu-C	C-terminal side of E, D
Endoproteinase Lys-C	C-terminal side of K
Endoproteinase Asp-N	N-terminal side of D, cysteic acid
Endoproteinase Arg-C	C-terminal side of R
Carboxypeptidase Y	non-specific cleavage from carboxylic end of peptide

rich collection of peptides from all but the most insoluble membrane bound proteins. Peptides in this mass range are ideally sized for MS-MS peptide sequencing. Moreover, tryptic peptides are charged exceptionally well in ESI due to the presence of c-terminal basic amino acids, increasing the sensitivity of detection, and facilitating MS-MS fragmentation. Trypsin is also a very robust enzyme able to tolerate a variety of buffers and up to 2 M urea.

Even though trypsin produces many useful peptides for MS and MS-MS analyses, use of trypsin alone is not likely to produce useful peptides over the entire length of a protein. This is especially the case for large proteins or membrane bound proteins that have extended hydrophobic domains devoid of basic amino acid residues. The opposite problem occurs with basic regions, which yield peptides too small for satisfaactory MS-MS analysis. This is not necessarily a hindrance for protein identification, but it can result in decreased sequence coverage significantly reducing the opportunity to detect protein modification sites. A greater number of useful peptides can be produced, and sequence coverage can be enhanced if several different proteases are used during digestion or if a protease is used in conjunction with a chemical cleavage agent such as cyanogen bromide.

Relatively simple peptide mixtures produced from the digestion of purified proteins can be analyzed directly on a MALDI-TOF instrument by a technique known as peptide mass fingerprinting (12). Just as the mass of an intact protein provides the basis for tentative protein identification, the mass of an intact peptide facilitates tentative peptide identification; although it should be noted that many more peptides share the

same nominal mass. The ability to identify proteins by peptide mass fingerprinting is based on the idea that it is possible to obtain mass estimates for many peptides from a given protein. Each individual peptide mass is compared to a protein or DNA sequence database to generate a list of potential peptide and protein matches. Inspection of the potential match lists for multiple peptides from a purified protein allows the identification of a consensus protein assignment. Another way to look at this is that many proteins may contain peptides of a particular mass; however, fewer proteins contain two peptides of particular masses, and still fewer contain three or more peptides of particular masses. If the protein sample only contains a single protein prior to digestion, the protein can usually be reliably identified by matching the masses of three to six peptides. The number of required matches is influenced by the mass accuracy of the data (3) and by the degree of redundancy of particular peptide masses in the available databases. More complex protein samples may require more peptide matches for conclusive identifications. Crude protein samples are not amendable to this technique.

If one is interested in investigating protein modifications, peptide mass fingerprinting is of limited value. Peptides with predicted modifications of known mass and sequence specificity may be identified using this technique, but less well characterized modifications can not be matched correctly. Comparison of peptide digests from both modified and unmodified versions of the same purified protein may lead to the observation of new or shifted peaks in the spectrum of the modified sample. However, conclusive identification of the modified peptides and determination of the sequence

specificity of observed modifications requires more detailed structural information that can only be acquired by MS-MS.

The alternative approach to peptide mass fingerprinting is peptide sequencing from peptide MS-MS data. Peptide sequencing is possible because of the predictable way in which peptides undergo collision induced dissociation (CID) (13). During CID peptides predominantly fragment along the peptide backbone at the peptide bonds, producing both b- and y- series ions, depending on which fragment retains a charge (Figure 1-1). Fragments that contain the N-terminus of the peptide are known as b- ions, while those that contain the C-terminus of the peptide are known as y- ions. Fragmentation at different points along the peptide backbone produces different sized b- and y- ions, with successive ions separated by a mass equal to the mass of extra amino acid included in the larger fragment. Quality MS-MS spectra contain a nearly complete b- and y- ion series which can be interpreted to decipher the amino acid sequence of the selected peptide. Other types of fragment ions are also observed in MS-MS spectra. A-, c-, x-, and z- series ions are also produced from fragmentation at different points along the peptide backbone; although these ion types are not observed as consistently as b- and y- ions. Fragmentation can also occur at amino acid side chains resulting in numerous other fragment types which can assist in the assignment of specific amino acids. Fragmentation can even occur at certain peptide modifications, potentially providing insight into the structure of modifications beyond what can be ascertained from simple peptide mass estimates.

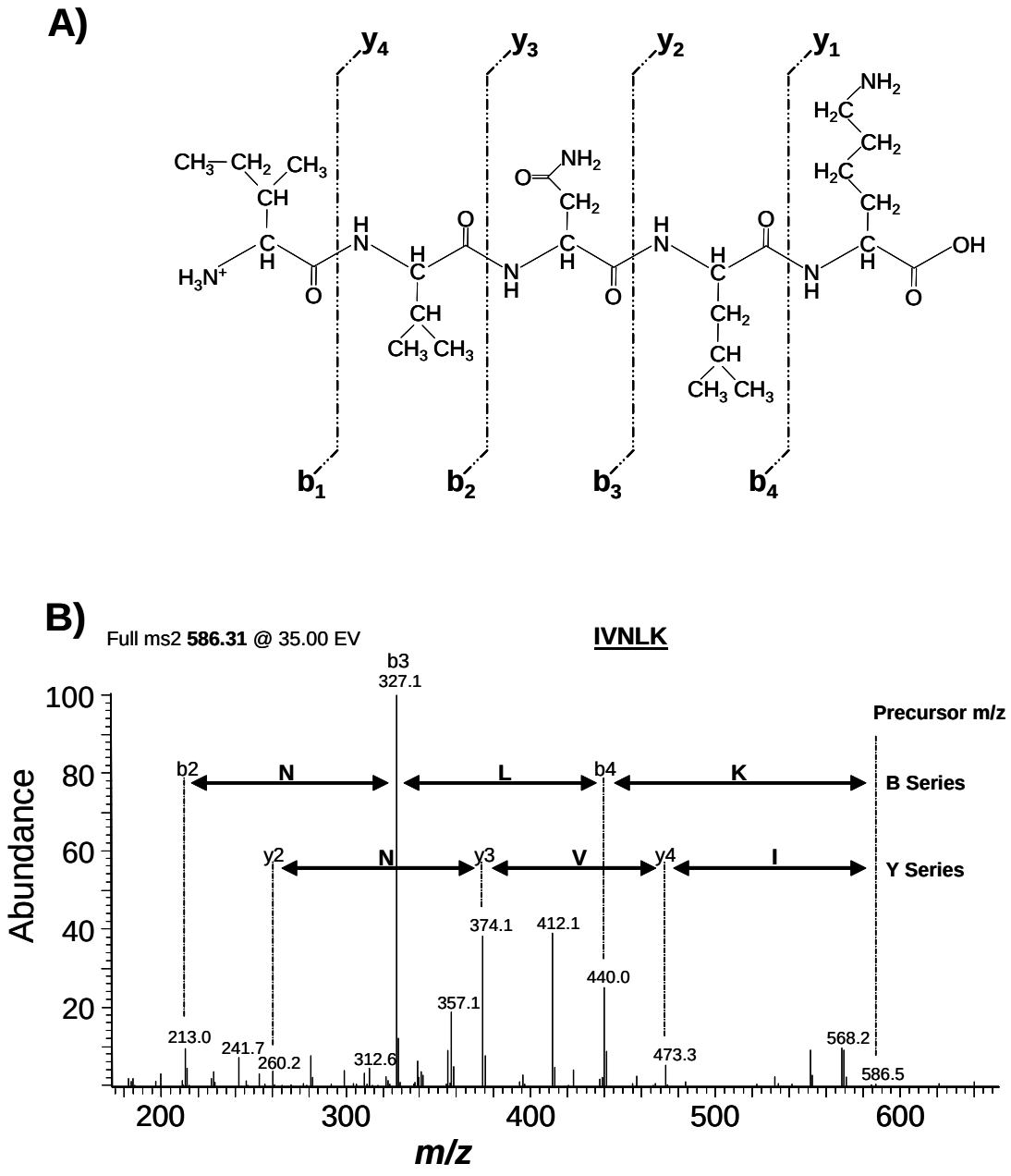


Figure 1-1. CID peptide fragmentation. A) Roepstorff nomenclature for classifying N-terminal and C-terminal peptide fragment ions. B) MS-MS spectrum of the gamma fibrogen peptide IVNLK illustrating peptide sequencing.

The wealth of information contained in peptide MS-MS spectra makes the peptide sequencing approach ideally suited for identification and mapping of protein modifications. Identification of peptides based on derived sequences from interpretation of MS-MS spectra is much more reliable than identification based on mass alone. Consequently, proteins can be identified on the basis of fewer peptides by peptide sequencing than by peptide mass fingerprinting. Additionally, the coupling of instruments with an ESI source to online HPLC allow for the identification of proteins in much more complex mixtures. In theory, peptide MS-MS spectra can be interpreted manually to infer peptide sequence, however, this process is very time consuming and error prone. Consequently a number of software algorithms have been developed which can match uninterpreted MS-MS spectra to theoretical spectra produced from the expected fragmentation of peptides from either protein or DNA sequence databases. Figure 1-2 summarizes the process by which peptides and proteins are identified using LC-MS-MS peptide sequencing methods.

A critical component to maximizing the number of peptides that can be analyzed in an individual analysis is the online separation of peptides prior to entering the mass spectrometer. Reverse phase liquid chromatography facilitates the separation and concentration of peptides and assists in the elimination of salts, which can suppress peptide ionization and damage the mass spectrometer. The temporal separation of peptides along with the data dependent scanning capabilities of ESI-ion traps allows hundreds to thousands of peptide MS-MS spectra to be acquired in a single 1 hour LC-MS-MS experiment. Greater degrees of online separation increase the time the mass

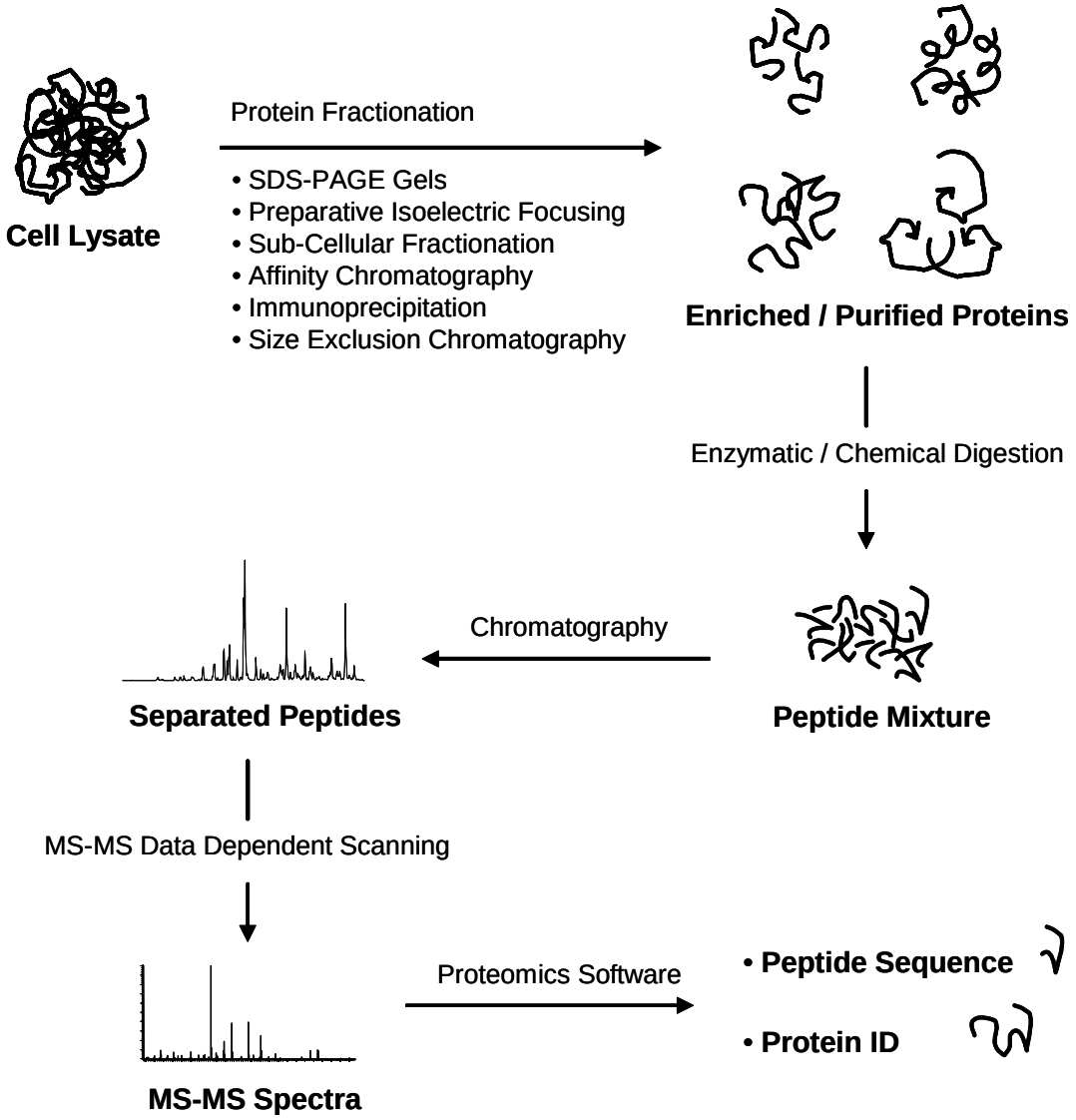


Figure 1-2. Schematic depicting the LC-MS-MS approach to peptide sequencing and protein identification

spectrometer can spend probing individual peptide peaks, increasing both the number of acquired spectra and the opportunity to acquire spectra from low abundance peptides. Sensitivity can be enhanced by reducing LC flow rates to nanoliters per minute and using small capillary columns for what is known as nanospray ESI (14).

Online separation can also be improved using multidimensional chromatography techniques (15, 16). Typically, a cation-exchange HPLC column is placed in front of and in line with a reverse phase C18 column. This system separates peptides on the basis of both ionic strength and hydropobicity and provides for a much finer degree of peptide separation than can be achieved by reverse phase LC alone. Complex peptide mixtures from the digests of hundreds of proteins can be analyzed by this type of system, and the sequence coverage of proteins in simpler mixtures can be significantly improved. The down side to multi-dimensional chromatography is that run times can be very long, some times in excess of 24 hours. The system is also not very robust and repeated experiments take a long time to carry out. Similar results can be achieved by simply carrying out the multi-dimensional separation in two discrete steps, as opposed to the online configuration (17). This can be done by first separating sample peptides on a cation-exchange column and collecting fractions. Individual fractions can then be subjected to reverse phase LC-MS-MS. The collected fractions are likely to be more course than can be achieved if the two separation systems are linked in line. But this approach is more robust, uses less instrument time, and generates smaller, more easily analyzed data files.

Several methods have been developed for quantifying peptides by LC-MS-MS. Because of different peptides have significantly different ionization potentials, absolute

quantification is not possible without generating a standard curve for each peptide of interest. However, relative quantification is possible. The standard approach developed by Aebersold and co-workers (18) utilizes isotope coded affinity tags (ICAT). The ICAT reagent is designed to react quantitatively with protein cysteine sulfhydrals. The reagent has been prepared in two forms; an isotopically light form and an isotopically heavy form for which hydrogen atoms on the reagent molecule have been replaced by deuterium atoms. A biotin moiety has also been incorporated into the reagent so as to facilitate purification of labeled peptides by affinity chromatography on an avidin column following tryptic digestion of sample proteins. Comparison of the protein expression levels between two samples with equal quantities of total protein is a carried out as follows. Treated separately, each sample is treated with a reducing agent such as DTT or TCEP to cleave protein disulfide bonds and generate free cysteine sulhydral groups to react with the ICAT reagent. One sample is then treated with the isotopically light ICAT reagent and the other is treated with the isotopically heavy form of the reagent. The two samples are then combined and digested with trypsin. The resulting peptide digest is purified by avidin affinity chromatography, enriching the sample in labeled peptides. This enriched mixture is then analyzed by LC-MS-MS. Individual peptides are identified on the basis of their MS-MS spectra; however, inspection of the full scan MS spectra acquired during the same time frame in data dependent scan mode should reveal two peptides separated by 8 a.mu. (the mass difference between the light and heavy isotope tags). Integration of these two peaks over the course of several MS scans allows for the determination of the relative quantities of specified peptides in each of the two samples.

While a powerful tool for characterizing protein expression levels, the ICAT method is not well suited for quantifying peptide modifications. First, free cysteine sulfydryls are prime targets for oxidation or reaction with electrophilic metabolites. Modified cysteine residues are not available for reaction with the ICAT reagent. Second, modifications at other amino acids may not be co-localized on peptides with cysteines and these peptides would not be captured by the affinity purification of ICAT labeled peptides. Fortunately, several other isotope labeling methods have been developed for labeling either the N- terminal amino (19, 20) or the C- terminal carboxylic acid (21) of digested peptides. The advantage of these methods is that they facilitate quantitative comparisons of all peptide including those with peptide modifications or those without cysteine residues. The disadvantage of these methods is that they do not utilize an affinity purification step to simplify the analyzed peptide mixtures. In order to be successful, these approaches are best applied to purified protein samples.

Data Analysis Algorithms and Software

The volume and complexity of data generated by mass spectrometry based proteomics methods renders routine manual interpretation of the data impossible. As a consequence, numerous algorithms and software programs have been developed for identification of peptides and proteins from either MS or MS-MS data. Programs developed for interpretation of full scan MS or MALDI-TOF MS data support the peptide mass fingerprinting method of protein identification. Programs developed to operate on MS-MS data support peptide sequencing methods.

There are a number of available programs that support peptide mass fingerprinting. The most basic of these programs which include Peptide Search (12), PeptIdent (22), and MS-Fit (23), rank database proteins by the number of matched peptide masses. Rankings based simply on the number of peptide matches have been shown to be biased toward the detection of large proteins which have a large number of constituent peptides. The more peptides a protein has, the greater the likelihood that both random and non-random matches will be observed. The somewhat more sophisticated MOWSE algorithm, which is incorporated into the widely used Mascot program, factors protein size into its rankings (24). MOWSE provides variable weights to assigned peptides based on an observed relationship between peptide and protein sizes, compensating for the non-random distribution of peptide molecular weights.

Mascot uses an probability based scoring model to adjust protein rankings generated by the underlying MOWSE algorithm (25). The Mascot generated probabilities provide an estimate of the likelihood that a reported match is a false positive

or random event. Statistical based scoring of this sort is essential to proteomics research due to the large amount of data generated and consequent impossibility of manual verification of most reported matches. Statistical scoring models facilitate automated analyses and can be used to establish criteria for establishing proteomics databases (26). The ProFound program also incorporates probabilistic scoring, although it uses a different algorithm from that used in MASCOT (27). ProFound statistics are based on a Bayesian model that allows the user to include additional information such as the source organism or the estimated isoelectric point or molecular weight of isolated proteins. The added information increases the sensitivity and selectivity of the algorithm. Profound also uses a two step scoring function that facilitates the analysis of simple protein mixtures.

The peptide mass fingerprinting algorithms are generally intolerant of peptide modifications. The exception is the program FindMod (28). FindMod has been designed to carry out exhaustive searches for peptides from a particular protein incorporating more than 20 endogenous protein modifications into its search parameters. The program also allows the user to customize searches for other modifications providing that the mass and amino acid specificity of the modifications are known beforehand. However, it should be noted that identification of peptide modifications based solely on full scan data is problematic for several reasons. First of all, unless further MS-MS experiments are carried out it is impossible to verify the peptide identity or the amino acid location of observed modifications. In addition, it is impossible to assign probability estimates to

matches to individual peptides based solely on a single mass estimate, so the likelihood that observed matches are not merely random events cannot be estimated.

There are also a number of programs that have been developed for assigning peptide MS-MS spectra (2, 26, 27, 29). These programs all take advantage of the predictable MS-MS fragmentation pattern of peptides. An early program developed for assigning peptide MS-MS spectra was the program MS-Tag (30). MS-Tag combines short sections of peptide sequence information, called sequence tags, along with the peptide mass to carry out database sequence searches. Incorporating MS-MS data into the search parameters increases the reliability of individual peptide matches relative to peptide mass fingerprinting methods. However, the sequence tags used by MS-Tag must be manually determined by the user by partial *de novo* sequencing of peptide spectra, greatly limiting throughput. Recently, a similar program, called Guten-Tag, was developed which is capable of automated generation of sequence tags (31). This program is capable of high throughput analysis and is able to detect peptide spectra derived from peptides with unanticipated modifications. The principal limitations to this program are that scoring is not statistically based and peptide modifications that significantly disrupt normal peptide fragmentation cannot be detected.

Of the programs designed to work with MS-MS data, SEQUEST (29, 32) is the most commonly used. SEQUEST first matches estimated precursor ion masses to database peptide sequences with the same mass. The program then derives theoretical MS-MS spectra from each of the database peptides and compares these derived spectra to the observed spectra using a cross-correlational analysis. The peptide sequence receiving

the best cross-correlation score is reported as a match. SEQUEST spectrum assignments are generally thought to be quite reliable, providing that the quality of the assigned spectrum is good and that the correct peptide sequence is included in the reference database. However, SEQUEST cross-correlation scores do not allow the user to estimate statistical confidence levels for assigned matches. Consequently, other programs have been developed to further process SEQUEST outputs in order to assign statistical significance to peptide and protein assignments (33, 34). SEQUEST allows the user to search for select protein modifications with known mass and sequence specificity, such as phophorylation, but the program is unable to detect unanticipated modifications or modifications that disrupt normal peptide fragmentation.

MS-MS spectra provide useful fingerprints of modified peptides, as they encode not only peptide sequences, but also the masses and sequence positions of modifications. Adaptation of existing database search algorithms to identify modified peptide sequences offers one approach to mapping protein modifications. For example, prior specification of modifications of known mass and sequence specificity (e.g., phosphorylation of Ser, Thr or Tyr) allows these modified peptides to be correctly identified in database searches with their MS-MS spectra (35). An alternative approach was employed by Gatlin et al., who used a modified version of the widely used program SEQUEST to detect protein sequence polymorphisms (36). The program utilized a virtual database approach, in which each sequence variant or potential modification was predicted in advance. Similarly, Creasy and Cottrell have reported an extension of the Mascot database search algorithm, in which a comprehensive list of known protein modifications and a residue

substitution matrix are used to perform error-tolerant searches of peptide MS-MS spectra (37).

These approaches share two common problems. First, searches with even a small set of modifications present a substantial combinatorial challenge, due to the large number of possible variants against which spectra are searched. Search times can become impractically long for searches involving a large number of proteins and error rates increase with the expanded pool of potential sequence matches. Second, not all relevant modifications can necessarily be predicted prior to analysis. Protein sequence polymorphisms or posttranslational modifications not found in the available databases are not included as search criteria. Protein modifications resulting from exposure to many endogenous electrophiles, xenobiotics and xenobiotic metabolites are difficult to predict and many have not yet been discovered, nor is their amino acid targeting specificity known (38). Thus, MS-MS spectra of peptides containing unanticipated sequence polymorphisms or modifications cannot be detected using the above approaches.

Pevzner et al. reported a mutation- and modification-tolerant database search algorithm (SHERENGA) based on spectral convolution, spectral alignment and branch-and-bound approaches, which provides a means of detecting modified peptides, even when such modifications are unknown at the outset of the experiment (39). Detection of unanticipated modifications employs a spectral alignment approach, which considers nearly all spectra as potential matches to database sequences, regardless of precursor mass. A difference in the precursor mass of the fragmented peptide and the expected mass of a database sequence is assumed to represent one or more modifications to the

original sequence. The theoretical fragmentation pattern of the target sequence is adjusted accordingly and then scored. Like the virtual database approach, spectral alignment results in a greatly expanded pool of potential sequence matches for each MS-MS spectrum. Consequently, run times are expected to be quite long for searches involving a large number of proteins and the error rates are likely to be higher for matches to modified sequences.

Unique Challenges to Characterization of Protein Modifications

While methods for identification and quantification of cellular proteins are well established, it remains a daunting challenge to characterize protein modifications, especially modifications resulting from adduction from electrophilic xenobiotics. In most cases, protein modification is expected to be less than quantitative. In other words, modified protein isoforms are likely present in significantly lower concentrations than their unmodified counterparts. Particular modifications may also occur at different sites on the same protein, with different sites having different implications for protein function. This diffuse distribution creates a situation where the concentrations of constituent modified peptides are even lower than the pooled concentration of a particular modified protein isoform. Traditional methods that rely on the use of either antibodies or radiolabeled compounds to visualize spots on SDS-PAGE gels may have insufficient sensitivity to detect modified proteins, especially when modifications occur on low abundance proteins. In many cases, antibodies or radiolabeled compounds may simply not be available for studies involving xenobiotic adducts.

Even if adducted proteins can be effectively targeted for analysis, it can be difficult to acquire MS-MS spectra of modified peptides. Enzymatic or chemical digestion of most proteins does not lead to a uniform distribution of peptides covering the entire protein sequence. Because of the irregular spacing of cleavage sites, digested peptides may be either too long or too short for effective MS-MS analysis. Even after denaturation protein folding may restrict access to some of the available cleavage sites resulting in either missed cleavages or decreased concentrations of peptides cleaved at the

restricted sites. Under ideal circumstances involving a high concentration of purified protein and using multidimensional LC-MS-MS, it is difficult to achieve 100% sequence coverage. While only a few peptides are sufficient for identification, gaps in sequence coverage may occur at modification sites; meaning that MS-MS spectra may not be acquired for certain modified peptides.

While acquisition of modified peptide spectra remains a challenge, the greatest barrier to identification of modified peptides at the time the research in this dissertation was carried out was that available software algorithms were unable to characterize many types of modifications. The widely used programs such as SEQUEST and Mascot were and are able to assign spectra belonging to peptides with anticipated modifications with known mass and sequence specificity. However, only one modification could be searched for at a time and search parameters had to be set for each modification prior to analysis. In the last few years, improvements to the SEQUEST algorithm now facilitate searches for a small set of known modifications. The disadvantage of this approach is that even a small set of modifications presents a substantial combinatorial challenge due to the large number candidate peptide sequences that have to be considered for each precursor mass. Search times can become impractically long for searches involving a large number of proteins and error rates increase with the expanded pool of potential sequence matches.

These and all other proteomics algorithms at the time were unable to assign spectra of peptides with unanticipated modifications. This was especially problematic for studies aimed at characterizing the distribution of protein targets impacted by reactive

xenobiotics. There are no consensus sequences that allow one to anticipate the location of such adducts and in some cases it is possible that adduction occurs at several different amino acids. Also, the mass of important adducts may not always be known, as several metabolites may be involved, some of which may remain to be characterized. Complex modifications such as ubiquitination may sufficiently perturb normal peptide MS-MS fragmentation so as to confound existing algorithms even when the mass and sequence specificity of the modification are known. Similarly, bifunctional alkylating agents (i.e., molecules with two reactive sites) may cause protein cross-links, thus greatly complicating MS-MS fragmentation and confusing existing programs.

During the time frame that this dissertation research was carried out several new proteomics algorithms were introduced. Of these, the SHERENGA algorithm was designed to detect unanticipated peptide modifications. Unfortunately, the sensitivity and specificity of this algorithm have not been well established. This is particularly relevant when attempting to identify unanticipated modifications because large numbers of peptides sequences are considered to be potential matches to each spectrum and peptide mass ceases to be defining characteristic. The SHERENGA algorithm and all other proteomics algorithms developed for the purpose of assigning peptide MS-MS spectra work by matching CID fragment ions to theoretical spectra for database peptide sequences. Modifications that significantly disrupt normal peptide MS-MS fragmentation patterns can not be detected by use of these algorithms.

Protein Modifications in Toxicity

Metabolic conversion of chemicals to reactive intermediates has long been recognized as a critical factor in the toxicity of a diverse array of environmental and therapeutic agents (40-42). However, the exact mechanism or mechanisms by which reactive intermediates produce toxicity remains poorly defined. Many reactive intermediates have the potential to form covalent bonds with cellular macromolecules, and toxicity is often correlated with DNA and protein binding (43). Covalent modification of DNA has been shown to be associated with mutagenesis and carcinogenesis. However, in tissues exposed to reactive metabolites the majority of affected cells do not become cancerous, but rather simply die either through apoptotic or necrotic mechanisms. The cell signaling pathways involved in cell death have been worked out in considerable detail, but the molecular mechanisms by which reactive metabolites initiate these pathways remains unclear. Although extensive DNA damage caused by reactive metabolites can potentially trigger apoptotic pathways, it is apparent that cellular stress and ultimately cellular death can be initiated by covalent modification of proteins (44).

Over the last several decades a body of evidence has emerged supporting the general hypothesis that covalent modification of specific proteins may be an initiating event in target organ toxicity and carcinogenesis (45, 46). Proteins contain a diverse collection of nucleophilic amino acids capable of reacting with reactive eletrophiles. Of the potential amino acid targets, cysteine residues are the strongest nucleophiles at physiological pH. Following exposure to electrophilic xenobiotics including

iodoacetamide and metabolites of acetaminophen (APAP), diclofenac, and halothane researchers have noted that the distribution of adducts is not uniform across the proteome, but rather select proteins appear to be modified (47-50). Until recently, attempts to identify protein targets of reactive electrophiles have relied on the use of radio-labeled compounds and adduct specific antibodies. These methods lack the sensitivity to detect low abundance adducts, thus most of the early adducts described have been modifications on high abundance proteins. Newer methods for identifying proteins using mass spectrometry have considerably more sensitivity and have the potential to detect protein modifications on low abundance proteins.

Studies of protein modifications produced by reactions with reactive xenobiotics are complicated by the fact that the identities of reactive metabolites are not always known. Because of this it can be difficult to predict the mass and amino acid specificity of xenobiotic adducts. Without this information, commonly used proteomics algorithms are likely to be unable to detect many important xenobiotic adducts.

Endogenous Protein Modifications

Modified and variant protein forms are abundant in living systems and are known to have profound biological implications. Table 1-2 provides a short list of some of the better known endogenous post-translational protein modifications and illustrates the diverse processes that they help regulate. In addition to the modifications listed in this table, there are many more protein modifications that have been reported such as: amidation, C-mannosylation, flavinylation, formylation, geranylation, hydroxylation, methylation, palmitoylation, and many others. Providing that the masses and sequence specificities of each of these endogenous medications are known, it is possible to identify and sequence peptides with these modifications using proteomics methods that were available at the start of this dissertation research. However, the proteomics algorithms that were available at the time were only able to identify expected modifications that were entered as search criteria by the user, but were unable to conduct exhaustive searches for all peptide modifications. If a researcher did not know the mass of a particular modification, or did not anticipate certain modifications to be present on the proteins that he/she was studying, potentially important modifications could go undetected. Complex modifications such as ubiquitination may also sufficiently perturb normal peptide MS-MS fragmentation so as to confound commonly used algorithms. Along the same line as post-translational modifications, it is also possible that proteins contain undocumented single amino acid polymorphisms or splice variants. Because virtually all proteomics database search algorithms match peptide mass spectrometry data to database sequences, undocumented modifications of this sort are rarely detected.

Table 1-2. Representative endogenous protein modifications

Modification	Biological Significance	References
Acetylation	Chromatin remodeling Transcription factor modulation Membrane targeting of some receptors	(51) (52) (53)
Biotinylation	Chromatin remodeling Activation of quanylate cyclase Regulation of biotin carboxylases	(54) (55)
Deamidation	Protein degradation and turnover	(56)
Farnesylation	Membrane targeting of signal transduction proteins	(57)
Myristoylation	Membrane targeting of signal transduction proteins Regulation of FAS mediated apoptosis	(57) (58)
O-linked N-acetylglucosamine	Regulation of signal transduction and gene expression	(59)
Phosphorylation	Regulation of G protein-coupled receptors Signal transduction Transcription factor modulation	(60) (61) (62)
Sulphation	Modulation of membrane protein interactions	(63)
Sumoylation	Intracellular protein localization	(64)
Ubiquitination	Target proteins for degradation Activation of protein kinases	(65) (66)

Proteins can also be modified by a variety of endogenous reactive oxygen species (ROS) (67). Superoxide anion (O_2^-) and hydrogen peroxide (H_2O_2) are regularly produced by macrophages and neutrophiles as well as by mitochondria and peroxisomes in all cells. Disease states or redox cycling of xenobiotics can dramatically increase the production of these oxidants (68-71). O_2^- can react with nitric oxide to produce to the potent oxidant peroxynitrite ($ONOO^-$). Similarly H_2O_2 can react with ferric iron to produce hydroxide radicals ($HO\cdot$). Both $ONOO^-$ and $HO\cdot$ are capable of directly oxidizing proteins at cysteine, tyrosine and methionine residues. These compounds also react with unsaturated lipids to form a variety of potent electrophiles including 4-hydroxynonenal (4HNE), 4-oxononenal (4ONE), and malondialdehyde (MDA). 4HNE and 4ONE have been shown to react with cysteine, histidine, and lysine amino acids at physiological pH (72, 73). MDA reacts preferentially with protein lysine residues (74). Surprisingly, these compounds appear to exhibit selectivity in the types of proteins and amino acid residues modified (67).

Elucidation of protein modifications caused by reactions with endogenous oxidants and electrophiles is essential to understanding the molecular processes involved in oxidative stress. Once again, it is possible to use previously existing proteomics methods to study individual predicted modifications. However, an exhaustive study of these modifications would be incredibly laborious.

Research Focus

The work presented in this dissertation was inspired by research that I had been conducting on pyrrolizidine alkaloids. Pyrrolizidine alkaloids are plant secondary metabolites found in species of the Senecio and Crotolaria genera. 1,2 -Unsaturated alkaloids are hepato- and pneumo-toxic in animals due to metabolism by cytochrome P-450 liver enzymes to reactive pyrroles. The structure and metabolic pathways of the alkaloid monocrotaline are depicted in Figure 1-3. Pyrrolizidine pyrrole metabolites are potent electrophiles and have been shown react with glutathione as well as cellular proteins and DNA. The metabolites are bi-functional alkylating agents with the potential to form a variety of protein, DNA, and glutathione crosslinks. The amino acid target specificity of protein pyrrole adduction was unknown at the start of this dissertation research, as was the identities of affected proteins. Believing that the identification of pyrrole protein targets could shed additional light onto the molecular mechanisms of pyrrolizidine toxicity, I collaborated with Dr. Liebler and those in his laboratory to develop new techniques to characterize protein modification sites.

At the beginning of this research, myself and others in Dr. Liebler's laboratory carried out experiments with various electrophiles and a model peptide library. The purpose of these experiments was to determine the amino acid specificity of the electrophiles that we were working with, and to characterize the impact of adduction by these electrophiles on peptide MS-MS fragmentation patterns. These initial experiments demonstrated that pyrrolizidine alkaloid pyrroles are capable of reacting with any nucleophilic amino acid even though there is likely a preference for reactions with

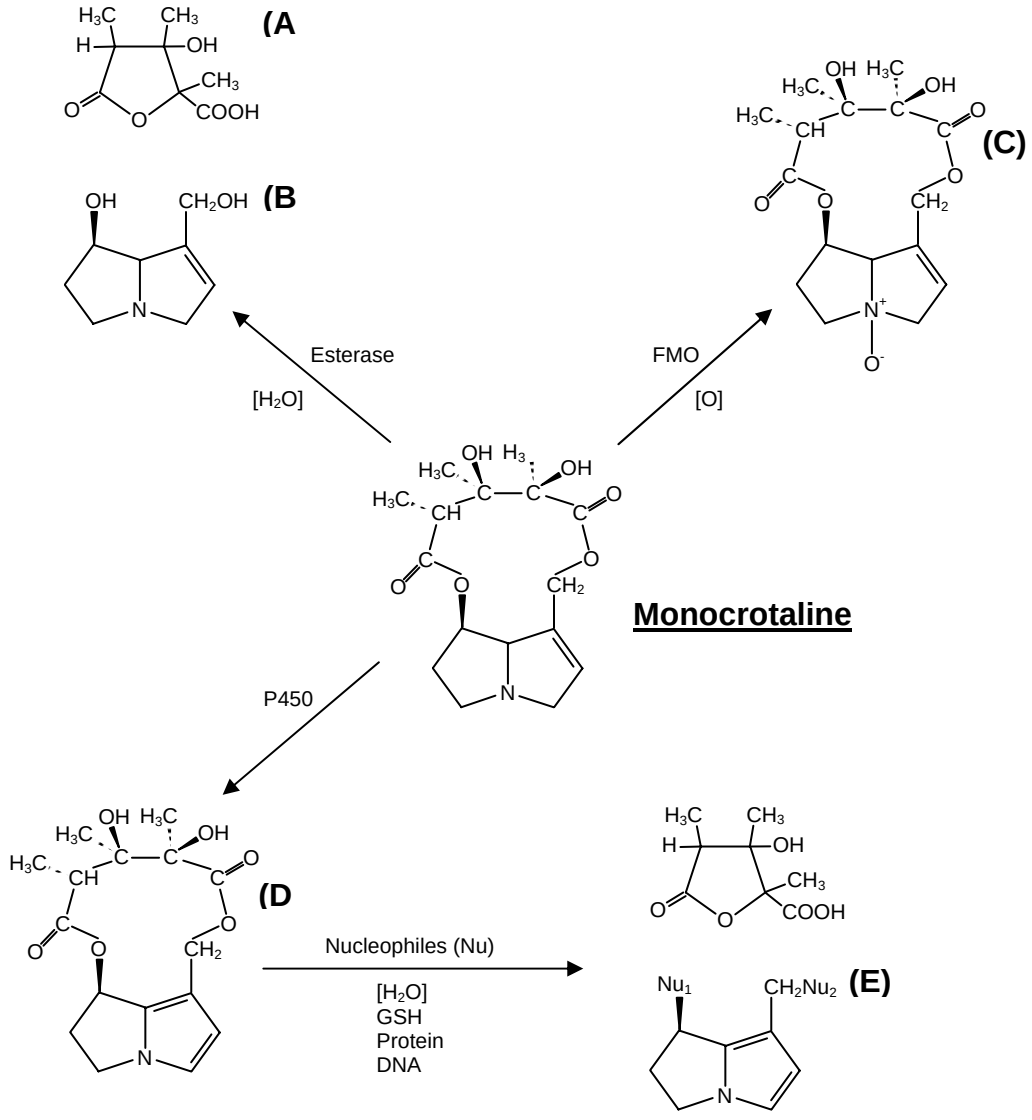


Figure 1-3. Monocrotaline metabolism. In humans, hydrolysis by esterases is a very minor metabolic pathway resulting in formation of A) monocrotalic acid and B) retronecine. Oxidation by flavin-containing monooxygenases is another minor pathway producing the water soluble non-toxic metabolite C) monocrotaline N-Oxide.

Cytochrome P-450 dehrogenation produces the toxic electrophile D) moncrotaline pyrrole, which reacts with cellular nucleophiles to produce E) dehydro-pyrrole adducts. cysteine thiols *in vivo*. They also demonstrated that pyrrole adduction significantly disrupts normal peptide MS-MS fragmentation patterns, to the point that meaningful sequence information can no longer be obtained from pyrrole modified peptide spectra. The lack of sequence specificity and the disruption of peptide MS-MS fragmentation meant that existing proteomics algorithms would be unable to detect peptide pryrrole adducts. However, the spectra of pyrrole adducted peptides exhibit a number of highly diagnostic characteristics which can be used to differentiate these spectra from the spectra of typical unmodified peptides. Furthermore, once the MS-MS spectra of theses adducts are identified it is possible to acquire peptide sequence information from the MS-MS-MS fragmentation of fragment ions produced from the neutral loss of the pyrrole adduct which have a mass equal to that of the unadducted peptide.

Initially I attempted to use the MS-MS fragmentation characteristics of pyrrole adducted peptides to design methods for detecting these peptides using standard tandem mass spectrometry approaches. Precursor ion scans and constant neutral loss scans are two techniques that can be used on triple quadrupole mass spectrometers that had potential for detecting pyrrole adducts (75). For precursor ion scans, the third quadrupole is set at a fixed *m/z* and the first quadrupole is scanned, allowing for the detection of precursor ions that fragment to produce a specific product ion. For constant neutral loss scans both quadrupoles are scanned with the third quadrupole set at a defined *m/z* below the first quadrupole, facilitating detection of the loss of neutral fragments. Because

pyrrole adducted peptide spectra exhibit characteristic product ions and neutral losses, precursor ion scans and constant neutral loss scans can both be used to detect adducted peptides. However, it turns out that detection of modified peptides was very much concentration dependent. At low adduct concentrations the background signal produced by low frequency fragmentations of high abundance peptides swamps the signal from the adducted peptides. An additional limitation of the tandem MS approaches is that peptide sequence information is not obtained during these experiments; in the case of pyrrole adducted peptides it is impossible to obtain peptide sequence information from the MS-MS experiments that can be carried out on a triple quadrupole mass spectrometer.

To overcome the above limitations and to improve detection capabilities for modified peptides in general, those of us in the Liebler laboratory saw a need to develop new data analysis algorithms to assist in identifying modified peptide MS-MS spectra. This dissertation provides a detailed description of the development and performance characteristics of two novel algorithms which have been tailored to facilitate detection of modified peptide spectra.

The first program developed was the SALSA (Scoring ALgorithm for Spectral Analysis) algorithm. This algorithm was inspired by the studies of pyrrolizidine pyrole adducted peptide MS-MS spectra, which had very distinct pyrrole “fingerprints” but lacked standard peptide fragment ions. Improvements to SALSA have made the program capable of studying all types of protein modifications, and have even made the program applicable to studies involving non-peptide molecules. SALSA provides the user tremendous flexibility in constructing customized search criteria for evaluating MS-MS

spectra. These customized criteria can be tailored to different adducts or specific peptides allowing for very focused searches. SALSA is unique in its approach, and has been licensed to ThermoFinnigan for packaging with their BioWorks platform.

While highly flexible, SALSA performs best with peptide modifications that have very pronounced MS-MS fragmentation patterns, and is most applicable to searches for modifications such as the pyrrole adducts that disrupt normal peptide MS-MS fragmentation. For other adducts and peptide modifications the second algorithm developed in our laboratory, called P-Mod, is a superior algorithm. P-Mod enables discovery and sequence mapping of modifications to target proteins known to be represented in the analysis or previously identified by another proteomics program such as SEQUEST. Similar to previous proteomics algorithms, P-Mod matches MS-MS spectra to peptide sequences in a search list. For spectra of modified peptides, P-Mod calculates mass differences between search peptide sequences and MS-MS precursors and localizes the mass shift to a sequence position in the peptide. Because modifications are detected as mass shifts, P-Mod does not require the user to guess at masses or sequence locations of modifications. Standardization of P-Mod searches and scoring made it possible to assign statistical confidence levels to assigned matches, reducing output volume and decreasing the time spent on manually confirming the authenticity putative peptide modifications.

CHAPTER TWO - SALSA: A Pattern Recognition Algorithm to Detect Electrophile-Adducted Peptides by Automated Evaluation of CID Spectra in LC-MS-MS Analyses

Introduction

New MS techniques for protein and peptide analysis make it possible to characterize the proteomes of living organisms (1, 76-78). Although MS is widely used for protein identification, new extensions of MS methods offer opportunities to detect and map modifications of the proteome. Precise characterization of protein modifications by both endogenous and exogenous chemicals is essential to understanding mechanisms of chemical toxicity and disease. Alkylation of cellular proteins may have numerous consequences ranging from altered gene transcription, cell signaling, or enzymatic activity to altered protein turnover or disruption of cytoskeletal integrity (42, 43, 45, 79-81). Protein adduction triggers cellular events that lead to acute toxicity and may serve as an initiator of apoptosis, or programmed cell death (79, 82). Identification of protein targets adducted by reactive xenobiotics may provide a better understanding of how cells respond to environmental insults.

Identification of electrophile-adducted proteins is a significant analytical challenge, which has been attempted previously with the aid of either radiolabeled substrates or antibodies. A common approach involves separating proteins by 1-D or 2-D gel electrophoresis, identifying spots of interest on audioradiograms or Western blots, and sequencing proteins in these spots by either Edman degradation or mass spectrometry

(48, 83-92). Unfortunately, there is often little direct evidence that proteins identified from gel spots are indeed the targets, as covalently modified peptides are rarely identified. Thus, protein alkylation sites and effects on protein function are highly speculative.

To assist in the identification of electrophile-adducted peptides, we have developed a data mining program, SALSA (scoring algorithm for spectral analysis), which can screen LC-MS-MS data-dependent scan files for precisely defined fragmentation patterns in collision-induced dissociation (CID) spectra. SALSA can simultaneously score multiple user-specified search criteria, including product ions, neutral losses, charged losses, and ion pairs. Search criteria may be combined to reflect overall fragmentation patterns that are diagnostic of specific peptide modifications. CID spectra scored by the algorithm can be recovered for visual inspection or evaluation with other software tools. This program represents an important step forward in the field of proteomics and provides a means of rapidly screening thousands of CID spectra for evidence of peptide adducts.

Here we demonstrate the use of SALSA to detect peptides alkylated by the reactive electrophiles dehydromonocrotaline (DHM), benzoquinone (BQ), and iodoacetic acid (IAA). CID spectra of model peptide adducts were initially evaluated to identify adduct-related fragmentations, which provide search parameters for the SALSA algorithm. The algorithm was then evaluated for its ability to detect modified peptides in a mixed protein digest. These experiments demonstrated the potential for sensitive and specific detection of peptide adducts in complex mixtures. The use of the SALSA

algorithm to simultaneously evaluate multiple characteristics of MS-MS spectra is superior to conventional tandem MS approaches (e.g., precursor or constant neutral loss scanning) to detect chemically diverse peptide modifications.

Methods

Algorithm

The SALSA algorithm has been developed to screen CID spectra obtained in MS-MS experiments on ThermoFinnigan mass spectrometers for user-defined fragmentation patterns. Averaged CID spectra are preprocessed to subtract nonfragment ions, estimate precursor charge, and normalize ion intensities as a percent of the total ion current (% TIC). Processed spectra then are evaluated for user-defined search criteria and scores are computed by taking into account the % TIC values of detected ions along with the assigned hierarchy for entered criteria. The SALSA prototype was initially encoded as a series of linked macros in Microsoft Excel. The algorithm has subsequently been encoded in Visual Basic, Java, and C++. The program has been licensed to ThermoFinnigan, and integrated into that company's BioWorks platform.

Spectra Preprocessing

Prior to scoring, CID spectra with at least 25 fragment ions are preprocessed by a data workup subroutine in which precursor charge is estimated and fragment ions are normalized to % TIC. Initially, the precursor ion and ions within ±0.4% of the precursor m/z are subtracted from each spectrum, along with ions with m/z greater than twice that of the precursor. This assures that any residual precursor ion is not calculated as part of %TIC and won't be identified as a spectral characteristic. The precursor charge is then estimated by calculating the ratio of the summed ion current for ions with m/z greater than the precursor to the total ion current for the remaining ions (Equation 2-1). Spectra

with a ratio greater than 0.1 are assumed to arise from doubly charged precursors, conversely, spectra with a ratio less than or equal to 0.1 are assumed to arise from singly charged precursors.

Equation 2-1. Charge estimation ratio = $$\frac{\sum(\text{ions} > \text{precursor } m/z)}{\sum(\text{all ions in spectrum})}$$

For spectra assigned single charge status, there is an additional subtraction of all ions with m/z greater than the precursor. The remaining fragment ions then are normalized to % TIC, where each ion has a value equal to 100 × (ion intensity/summed ion intensity of the remaining ions). Finally, ions with a % TIC value less than 0.2 are subtracted from the spectrum and the remaining ions are again normalized. These subtractions maximize the % TIC values for fragment ions detected by the algorithm and decrease background noise for ion pair detection (see below).

Scoring Product Ions, Neutral Losses, and Charged Losses

SALSA scores specific product ions by identifying the most abundant ion within a window centered at the designated m/z value ± 0.5 m/z unit and recording the % TIC value for the selected ion. Neutral losses and charged losses from doubly charged precursors are scored in an analogous manner (Figure 2-1). Neutral losses result in product ions that have the same charge as the precursor ion, and may be observed in both singly and doubly charged spectra. The window for neutral loss detection is centered at the precursor m/z minus the user-specified neutral mass/precursor charge. (Note that the actual m/z value for a neutral loss from a doubly charged precursor is half that of the

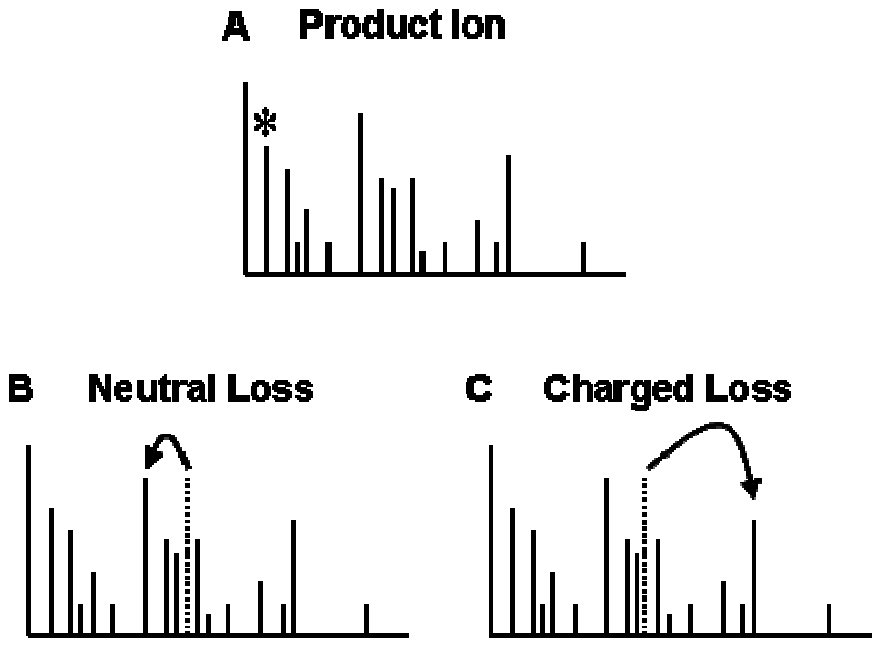


Figure 2-1. SALSA scoring of individual ions. A) Product Ion; B) Neutral loss; C) Charged loss. The * indicates an individual product ion. Dashed lines indicate the m/z of a precursor ion, and arrows point to ions scored by hypothetical losses.

same mass loss from a singly charged precursor.) Charged losses generate product ions that have a charge one unit less than that of the precursor, consequently, charge losses are only observed in spectra arising from doubly charged precursors. Charged losses are calculated by subtracting the specified m/z from the predicted singly charged m/z value for the precursor instead of the actual precursor m/z (i.e., 2 × precursor m/z - 1). When a particular loss is entered as a search criterion, the precursor charge and the charge of the product ion produced by the loss are included in the loss description, allowing the user to define the loss as neutral or charged and to adjust the magnitude of a neutral loss to account for the precursor charge state.

Scoring Ion Pairs

The algorithm can also detect ion pairs, such as sequential b- or y-series ions in peptide CID spectra (13). Ion pairs are defined as two fragment ions that are a specified distance apart on the m/z axis (Figure 2-2). This distance may reflect the residue mass of one or more amino acids or the elimination of a specific adduct, adduct fragment, or other modification. To detect ion pairs, the SALSA algorithm first generates a hypothetical list of fragment ions shifted the designated number of m/z units above the actual fragment ions in the spectrum, then rounds fragment m/z values in both lists to the nearest integer, and finally matches ions from the real list to the hypothetical list. If a match exists, the ion pair is scored as the geometric mean of the % TIC values for the largest fragment ion from each of the rounded integer windows.

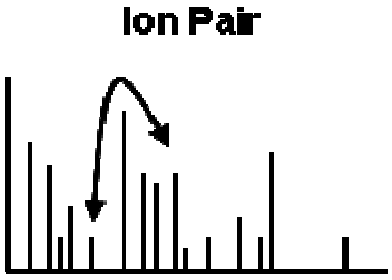


Figure 2-2. SALSA scoring of ion pairs. The figure depicts an ion pair in a hypothetical spectrum separated by a defined mass shift.

Primary and Secondary Search Criteria

Since SALSA scores are contingent on the summed % TIC of detected ions, inclusion of additional search criteria frequently results in higher raw scores. However, non-specific scoring may also be increased. SALSA allows the user to partially compensate for non-specific scoring by differentiating between primary and secondary search criteria. Primary search criteria are considered to be highly diagnostic and are expected to have large % TIC values in target spectra. In contrast, secondary search criteria may have less significant % TIC values in target spectra, and may be commonly observed in non-target spectra.

Secondary search criteria are entered in the same way as primary criteria, except that they are each linked to a specific primary criterion in the search list. While primary criteria are automatically scored when detected, a secondary criterion is only scored when the linked primary criterion is detected in the same MS-MS spectrum. Thus, the scoring of secondary criteria is contingent on the presence of other primary indicators. Scores for secondary characteristics are adjusted to ensure that final SALSA scores are most heavily influenced by primary criteria. The initial calculated % TIC score of a secondary criterion is adjusted by taking the geometric mean of this score and the % TIC score of the primary criterion to which it is linked. Each secondary characteristic is scored only once and is allowed a maximum score equal to the score of the linked primary characteristic. The final spectrum score is calculated as the sum of % TIC values of detected primary characteristics plus the sum of adjusted secondary characteristic scores.

SALSA Output

The output of the SALSA algorithm is in the form of a list of MS-MS scans ranked in order of decreasing score. SALSA provides the precursor m/z, retention time, and scan numbers for each spectrum. All product ions, losses, or ion pairs, scored by the algorithm are reported alongside the spectrum identifiers. It is often possible to estimate spectrum quality directly from this information, prior to recovering the complete CID spectrum for visual inspection. However, newer versions of the SALSA program automatically display the CID spectra for any SALSA outputs selected by the user.

Enzymatic Digestion of Proteins

A mixture of 400 μg of bovine serum albumin (BSA), 300 μg of horse skeletal apomyoglobin (MYO), and 100 μg of rat liver glutathione S-transferase (all from Sigma) was lyophilized and dissolved in 200 μL of 8 M urea/2 M NH4HCO3. Tris(2-carboxyethyl)phosphine (TCEP; Pierce) (125 μL, 30 mM) was added, and the solution was heated at 95 °C for 15 min. (TCEP provides efficient reduction of protein disulfide bonds, essential for enzymatic digestion, yet it produces much less background noise compared to the commonly used reducing agent dithiothreitol.) The solution was diluted to a final volume of 800 μL, and sequencing grade modified trypsin (Promega) was added in a ratio of approximately 1:25 w/w enzyme/protein. The solution was incubated for 24 h at 37 °C and stored at -20 °C.

Preparation of Dehydromoncrotaline-Adducted Peptides

DHM was synthesized by o-brominil oxidation of monocrotaline in chloroform as previously described (93). The product was lyophilized, dissolved in dry acetone (15 μg/μL), and stored at -20 °C. Peptide adducts of DHM were prepared as follows: peptides were dissolved in H2O to 1-3 μg/μL, made alkaline with the addition of 5× molar excess of diisopropylethyamine, and treated with a 2× molar excess of DHM, added dropwise with mixing. Peptides containing disulfide bonds were reduced prior to alkylation reactions with a 5× molar excess of TCEP. Reaction mixtures were left at room temperature for 30 min and stored at -20 C.

Model peptides alkylated by DHM included the biological peptides glutathione (γECG), urotensin (AGTADCFWKYCV), antiflammin-2 (HDMNKVLDL), leucine enkephalin (YAGFLR), anaphylatoxin C3a fragment 70-77 (ASHLGLAR), frog atrial natriuretic peptide-24 (SSDCFGSRIDRIGAQSGMGCGRRF), amyloid β-protein 22-35 (EDVGSNKGAIIGLM), and β-neurokinin (DMHDFFVGLM-NH2) as well as the synthetic peptides GRGDSPC, AGAGCAGAG, AGAGKAGAG, AVAGCAGAR, AVAGKAGAR, and LVACGAK. The frog atrial natriuretic peptide (FANP) and the GRGDSPC peptide were obtained from Peninsula. The other biological peptides were from Sigma, whereas the remaining synthetic peptides were synthesized by Sigma Genosys. The FANP peptide was subjected to enzymatic digestion with trypsin (as above) both before and after alkylation by DHM, producing two cysteine-containing peptide fragments SSDCFGSR and IGAQSGMGCGR. DHP adducts were observed on

both peptide fragments when reactions were carried out either before or after digestion, indicating that DHP adducts are stable to digestion conditions.

Two peptide-DHP adducts were purified on a Hamilton 5-μm PRP-1 analytical column (2.1 × 150 mm) for use in spiked protein digest experiments. A γECG-DHP adduct was purified with a H2O/CH3CN/0.01% TFA gradient, which started at 1% CH3CN from 0 to 1 min, then was programmed to 9% CH3CN by 27 min, and then to 90% CH3CN at a flow rate of 0.5 mL/min with UV detection at 230 nm. A peak eluting between 26 and 27 min was collected and analyzed by MS and MS-MS on the TSQ. Full-scan MS revealed a base peak ion at m/z 425 and no detectable ion at m/z 443. Tandem MS of the m/z 425 precursor confirmed the identity of a singly charged γECG-DHP M+117 adduct. A LVACGAK-DHP adduct was also purified using the same solvent system with a gradient starting at 7% ACN from 0 to 1 min, then to 25% ACN at 17 min, and then to 98% CH3CN at a flow rate of 1.0 mL/min. The synthetic peptide-DHP adduct eluted between 15.5 and 17 min. Tandem MS analysis of m/z 778 confirmed the identity of a singly charged LVACGAK-DHP M+117 adduct. The collected HPLC fractions from these two preparations were pooled and lyophilized. The adducts were dissolved in methanol and quantified with a modified Ehrlich's reagent using DHM as a reference standard (94). The two HPLC-purified DHP adducts were then spiked into the above protein digest at concentrations of 1, 5, and 20 pmol /μg of protein.

Preparation of Benzoquinone-Adducted Peptides

The synthetic peptide AVAGCAGAR was alkylated with BQ as described earlier (95). The adduct was HPLC purified on a 5-μm Vydac Protein and Peptide C18 column (2.1 × 250 mm) eluted with a H2O/CH3CN/0.01% TFA gradient, starting at 1% CH3CN for the first 3 min and then programmed to 85% CH3CN over 27 min. The flow rate was set to 0.2 mL/min, and UV absorbance was monitored at 220 and 280 nm. The peptide-hydroquinone (HQ) adduct eluted between 18 and 19 min. Collected fractions were dried on a Savant SpeedVac concentrator and then dissolved in H2O at a concentration of 0.5 mg/mL. Tandem MS analysis of m/z 883 confirmed the identity of the singly charged AVAGCAGAR-HQ adduct, which was added to an aliquot of the protein digest at a final concentration of 57 pmol/μg of protein. An additional HQ adduct (m/z 713) observed as a contaminant in the above preparation was tentatively identified as AGCAGAR-HQ by tandem MS. The precise concentration of this contaminant in the spiked protein digest was not determined. However, the concentration was assumed to be less than 20% of the full length peptide adduct since the purity of the synthetic peptide was greater than 80%.

Preparation of Iodoacetic Acid-Adducted Peptides

Recrystallized IAA (1.86 g) was dissolved in 100 mL of distilled water to produce a 0.1 M solution and was added to 1 mg of AVAGCAGAR in a ~25× molar excess. The solution was incubated in the dark at 37 °C for 3 h, then removed from the incubator, and left under ambient light for 1 h to decompose unreacted IAA. Tandem MS of m/z 417 confirmed the identity of the doubly charged S-carboxymethylated peptide

(AVAGCAGAR-CM). LC-MS analysis of the reaction mixture indicated a ~50% yield for the reaction. The product was added without further purification to an aliquot of the protein digest at a concentration of ~12 pmol/μg of protein. The CM adduct of the AGCAGAR peptide was not observed in this preparation.

Mass Spectrometry

Model peptides and peptide adducts were analyzed by +ESI-MS and MS-MS experiments on a Finnigan MAT TSQ 7000 triple quadrupole instrument and a Finnigan LCQ ion trap instrument to identify diagnostic fragmentation characteristics of adducted peptides. Full-scan experiments were initially carried out to identify precursor ions corresponding to peptide adducts. Adducted and nonadducted peptide precursors were then subjected to CID in MS-MS experiments at -30 to -35 eV on the TSQ or at an activation amplitude of 40% on the LCQ.

Spiked protein digests were analyzed by +ESI LC-MS-MS with data-dependent scanning on the LCQ. Chromatography of DHM-adducted peptides was carried out on a 5-μm Vydac 259VHP column (1.0 × 150 mm) with a H2O/CH3CN/0.01% TFA gradient at a flow rate of 30 μL/min. The mobile phase contained 3% CH3CN from 0 to 4 min, then increased to 12% CH3CN at 13 min, 21% at 35 min, 27% at 75 min, 45% at 105 min, 70% at 120 min, and then to 95% CH3CN from 125 to 150 min. The LCQ was set to acquire MS and MS-MS data between 20 and 150 min. Chromatography of the BQ and IAA adducted peptides was achieved with a 5-μm Vydac 218TP column (1.0 × 250 mm) using H2O/CH3CN/0.01% TFA gradient elution at a flow rate of 30 μL/min. The

gradient was initiated at 3% CH3CN for the first 3 min, increasing to 5% CH3CN at 10 min and held for 10 min, and then increasing to 30% CH3CN at 115 min, 75% at 125 min, 85% at 130 min, and then to 95% CH3CN at 135 min. In data-dependent scan experiments, the instrument was set to conduct repeated cycles of a single MS scan followed by three successive MS-MS scans of a selected precursor ion. The dynamic exclusion mass limit was set to ±1.5 m/z, with ions residing on the exclusion list for 5 min.

Many of the DHP adducts were analyzed by flow injection tandem MS precursor ion and constant neutral loss scans on the TSQ. Precursor ion scans of m/z 118 provided the most consistent and sensitive detection of DHP adducts at a collision energy of -35 eV. LC-MS-MS precursor ion scans (m/z 118) of the peptide-DHP spiked protein digest were carried out using the same chromatographic conditions described for the LCQ data-dependent scan experiments. Data analysis of the tandem MS experiments was performed with Thermoquest Xcalibur software.

Results

DHP Adducts

DHM alkylated model peptides were analyzed by flow injection +ESI-MS-MS experiments to elucidate fragmentation characteristics of adducted peptides for use as search parameters in the SALSA algorithm. Previous reports have indicated that DHM reacts preferentially with cysteine thiols but may also alkylate nitrogen nucleophiles on proteins (96, 97). The majority of peptides we studied contained either cysteine or histidine. However, other peptides contained lysine or arginine as the only nucleophilic amino acid residues other than the N-terminal amines. Under the mildly basic conditions used, DHM reacted with all of the model peptides, although the highest apparent yields were with peptides containing cysteine thiols. Three chromatographically separable alkylation products were observed in peptide reactions with DHM, which is a bifunctional alkylating agent (98). In accord with previous observations, DHM produced dehydropyrrole (DHP) alkylation products having masses of M+117, M+135, or 2M+117, where M is the mass of the target peptide (99-101). While the cross-linked 2M+117 adducts may be biologically important products, our experiments to date have focused on the M+117 and M+135 peptide-DHP adducts (Figure 2-3).

+ESI-MS-MS spectra of peptide-DHP adducts obtained on the TSQ triple-quadrapole demonstrate several diagnostic fragmentation characteristics (Table 2-1), including DHP-derived product ions at m/z 106, 118, 120, and 136, as well as several losses from the precursor ion that are specific for each DHP-adduct isomer and precursor

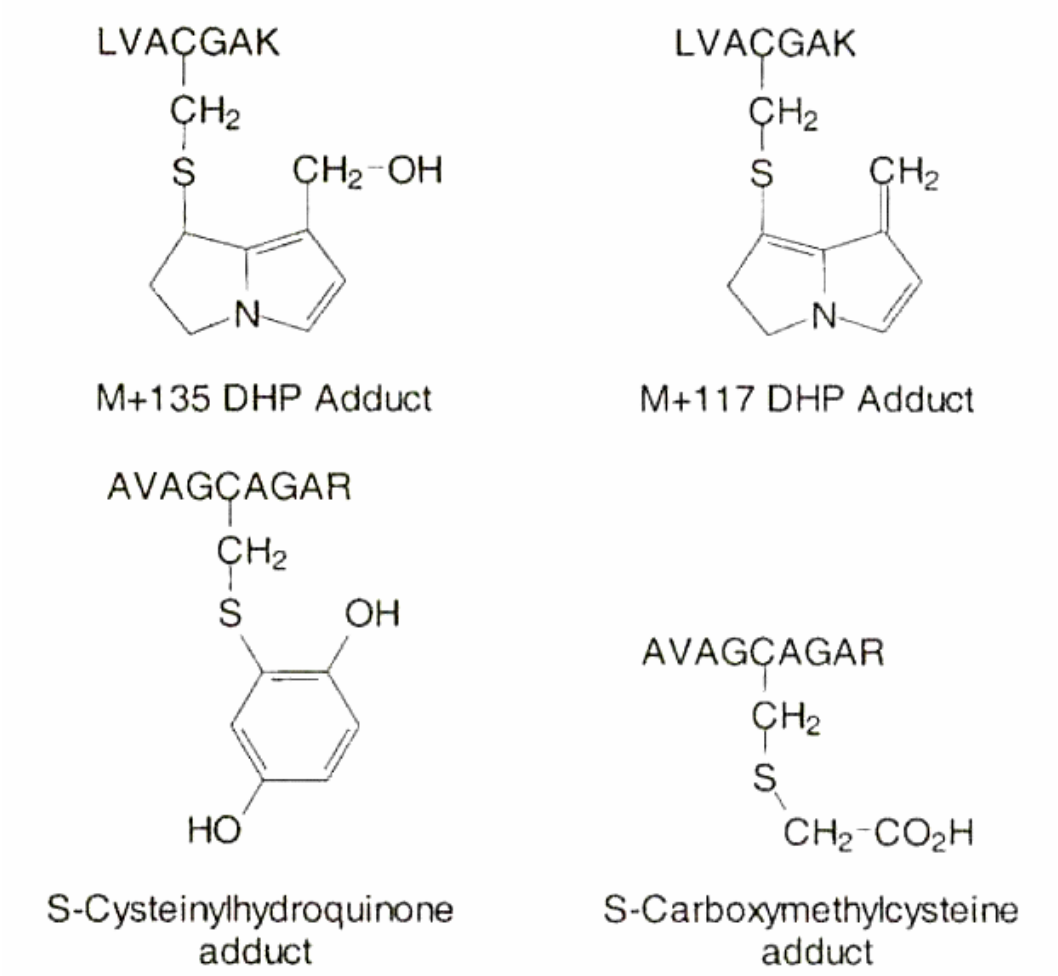
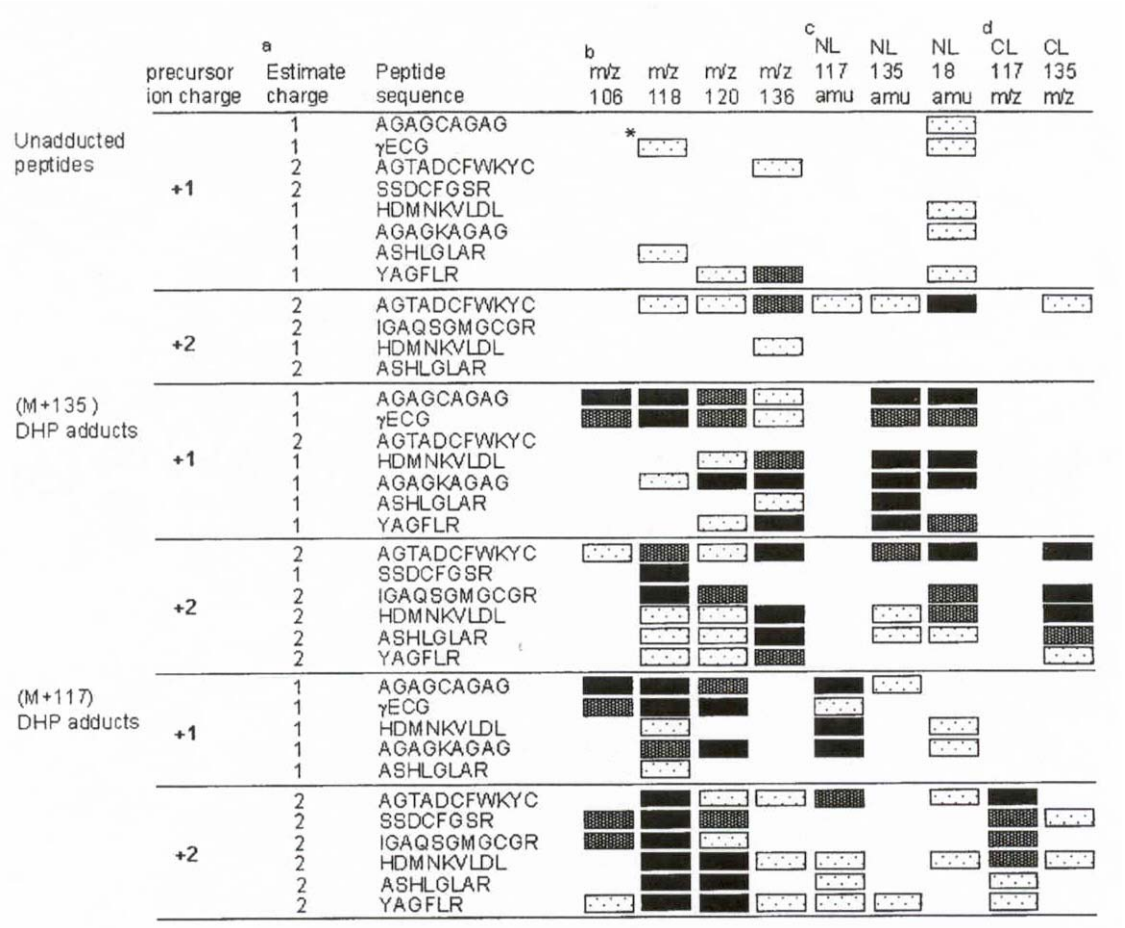


Figure 2-3. Representative structures of peptide adducts.

Table 2-1. Fragmentation Pattern of DHP-Adducted Peptides Observed in TSQ CID Spectra.



*The Shaded boxes indicate the %TIC values for SALSA detected fragment ions: ▒ = 0.2 – 2.0 %TIC, ▓ = 2.1 – 6.0 %TIC, and ■ > 6.0 %TIC. [a] SALSA estimate of precursor ion charge; [b] Product ion at designated m/z; [c] Neutral loss of specified amu / the estimated charge from the precursor m/z; [d] Charged loss from doubly charged precursror.

charge state. Singly charged precursors exhibit a neutral loss of the DHP adduct, either 117 or 135 amu depending on the DHP isomer collided. Doubly charged precursors produce corresponding neutral losses of 58.5 and 67.5 m/z units. However, CID spectra of doubly charged precursors exhibit more prominent charged losses of the DHP adduct, resulting in product ions corresponding to the singly charged unadducted peptides. M+135 adducts also exhibit a prominent neutral loss of water (18 or 9 m/z units from singly or doubly charged precursors, respectively). The formation of specific DHP-derived fragments is variable between different peptide-DHP adducts, depending on the amino acid adducted and the surrounding peptide sequence. The product ions at m/z 106 and 118 are observed more frequently and at higher abundance in cysteine adducted peptides than in peptides adducted at other amino acids. The m/z 136 product ion is specific to M+135 adducts and is not observed in the majority of CID spectra of dehydrated M+117 isomers.

Several of the fragmentation characteristics of peptide-DHP adducts are occasionally observed in CID spectra of unadducted peptides (Table 2-1). TSQ MS-MS spectra of peptides containing tyrosine or phenylalanine exhibit immonium ions at m/z 136 and 120, respectively. The TSQ CID spectrum of the urotensin peptide exhibits a product ion at m/z 118 and a charged loss of 117 amu from the double-charged precursor due to the C-terminal valine. However, background signals from these sources do not appear to cause significant interference for the SALSA algorithm, which scores spectra on both the presence of specific ions and their relative intensity in % TIC. The fragment

ions observed in unadducted peptides, while real, have much lower % TIC values compared to analogous ions in peptide-DHP spectra.

CID spectra of peptide-DHP adducts were also evaluated on the LCQ ion trap. In contrast to the TSQ, the trap exhibits a low mass cutoff and cannot detect product ions with m/z less than ~25% of the precursor m/z. Thus, the pyrrole fragment ions at m/z 106, 118, 120, and 136 are not observed in the LCQ MS-MS spectra of most peptide-DHP adducts. However, neutral and charged losses are much more prominent than in TSQ spectra. DHP adduction results in the disruption of normal peptide fragmentation in CID spectra acquired on either instrument. Peptide specific b- and y-series ions are infrequent and appear at low abundance in adducted peptide spectra, complicating attempts to obtain sequence information directly from the MS-MS spectra of adducted peptides (97). However, adducted peptides may be effectively sequenced in follow up MS-MS-MS experiments on the LCQ. In these experiments the product ions resulting from the neutral or charged loss of the DHP adduct are selected and activated for another CID cycle. These loss ions correspond to, and exhibit exactly the same fragmentation pattern as the original unadducted peptides.

The model peptide experiments provided the basis for the development of a general set of SALSA search parameters for the detection of peptide-DHP adducts (Table 2-2). The criteria used for TSQ spectra were as follows: The product ions at m/z 118 and 120 were entered as primary criteria, as well as the neutral and charged losses (from single- and double-charged precursors, respectively) of 117 and 135 amu. The product ion at m/z 106 was entered as a secondary criterion linked to primary detection of the m/z

Table 2-2. SALSA Search Criteria for Peptide-DHP CID Spectra Obtained on Either the TSQ or LCQ.

TSQ Spectra		LCQ Spectra	
Primary	Secondary[a]	Primary	Secondary
PI 118[b]	PI 106	NL 117 (SCP) or CL 117 (DCP)	NL 117 (DCP)
PI 120		NL 135 (SCP) or CL 135 (DCP)	NL 18 NL 135 (DCP)
NL 117[c] (SCP[e]) or CL 117[d] (DCP[f])	NL 117 (DCP)		
NL 135 (SCP) or CL 135 (DCP)	NL 18 PI 136 NL 135 (DCP)		

[a] Scoring of Secondary Criteria is contingent on the simultaneous detection of the adjacent primary criterion. [b] Product ion. [c] Neutral loss. [d] Charged loss. [e] Singly charged precursor. [f] Doubly charged precursor.

118 product ion. Likewise, the product ion at m/z 136 and the neutral loss of H2O from either single- or double-charged precursors were together entered as secondary to the neutral or charged loss of 135 amu. Inclusion of the m/z 136 product ion as a secondary criterion reduces the potential for tyrosine immonium ion interference. Neutral losses from doubly charged precursors of 117 and 135 amu (58.5 and 67.5 m/z) were entered as secondary criteria linked to the corresponding charged losses. Search criteria for LCQ spectra were identical except for the exclusion of the low m/z product ions.

Using these search criteria, SALSA gave substantially higher scores to CID spectra of DHP adducts compared to unadducted peptide spectra (Table 2-3). Expressing the SALSA scores on a log scale indicates that spectra scoring greater than 0.9 are likely to be derived from DHP adducts. None of the unadducted peptide CID spectra evaluated thus far has ever scored this high. However, 21 of 24 TSQ spectra and all 17 LCQ spectra of model peptide-DHP adducts scored above the 0.9 threshold. Both hydrated and dehydrated DHP adducts were detected equally well, as were both singly and doubly charged precursor ions. Two of the three DHP adduct CID spectra that received scores below threshold can be attributed to incorrect charge estimation by SALSA during initial MS-MS scan processing, which caused the algorithm to miscalculate neutral and charged losses. The precursor ion charge was assigned correctly for 39 of the 41 peptide-DHP CID spectra evaluated.

Experiments with spiked protein digests demonstrate improved sensitivity and specificity of DHP adduct detection using the SALSA algorithm to evaluate LCQ data-dependent scan CID spectra compared to conventional tandem MS detection strategies.

Table 2-3. Comparison of SALSA Scores for CID Spectra of Unadducted Peptides and Peptide-DHP Adducts Using the Search Criteria Listed in Table 2-2.

	TSQ Spectra[a]		LCQ Spectra[b]	
Log (SALSA Score + 1)	Unadducted[c] Peptides	Peptide-DHP[d] Adducts	Unadducted Peptides	Peptide-DHP Adducts
0	7 (63.5)[e]	1 (4.2)	508 (97.5)	
0.01 – 0.3	2 (18.2)		11 (2.1)	
0.31 – 0.6	1 (9.1)	1 (4.2)	1 (0.2)	
0.61 – 0.9	1 (9.1)	1 (4.2)	1 (0.2)	
0.91 – 1.2		2 (8.3)		2 (11.8)
1.21 – 1.5		9 (37.5)		4 (23.5)
1.51 – 1.8		9 (37.5)		7 (41.2)
1.81 - 2.1		1 (4.2)		4 (23.5)

[a] Flow injection CID spectra of peptides listed in Table 2-1. [b] LC-MS-MS data-dependent scan spectra. [c] Spectra from a tryptic digest of BSA. [d] DHP adducts of AVAGCAGAR, LVACGAK, AVAGKAGAR, GRGDSPC, EDVGSNKGAIIGLM, and DMHDFFVGLM. [e] Number of spectra (relative frequency, %).

The intense pyrrole-specific fragmentation pattern in the MS-MS spectra of DHP adducts facilitates detection in simple mixtures by tandem MS precursor ion or constant neutral loss scans on the TSQ. As the low m/z product ions are often the most abundant fragments in TSQ spectra, precursor ion scans provide the most sensitive tandem MS detection of peptide-DHP adducts. However, there is considerable variability in the relative production of diagnostic product ions for DHP adducts of different peptides (Table 2-1), making no single precursor ion scan universally successful. Figure 2-4 depicts the LC-MS-MS analysis of a BSA/MYO/GST tryptic digest spiked with DHP adducts of γECG and LVACGAK. Precursor ion scanning for m/z 118 detected the LVACGAK-DHP adduct ($[M + H]^+$ m/z 778) in the 20 pmol/μg spiked protein digest. Panel A shows the reconstructed total ion current (RTIC) of a m/z 118 precursor ion scan (m/z 400-2000) with an apparent peak between 67 and 70 min. This peak coincides with the base peak trace of m/z 778 shown in panel B. The γECG-DHP adduct ($[M + H]^+$ m/z 425) eluted between 26 and 28 min, as indicated by the weak base peak signal for m/z 425 in panel C. The ion signal was very low for this adduct, apparently due to a lower ionization potential for the γECG peptide compared to LVACGAK and to possible dilution of the signal from peak splitting during chromatography. No peak corresponding to γECG-DHP was observed in the RTIC trace. When both peptide adducts were present in concentrations of or below 5 pmol/μg of protein, the precursor ion scan failed to clearly differentiate between either adduct and the background produced by other peptides in the mixture (panel D). The specificity of precursor ion scans for DHP adducts may be improved by decreasing the energy of collision. However, this improvement

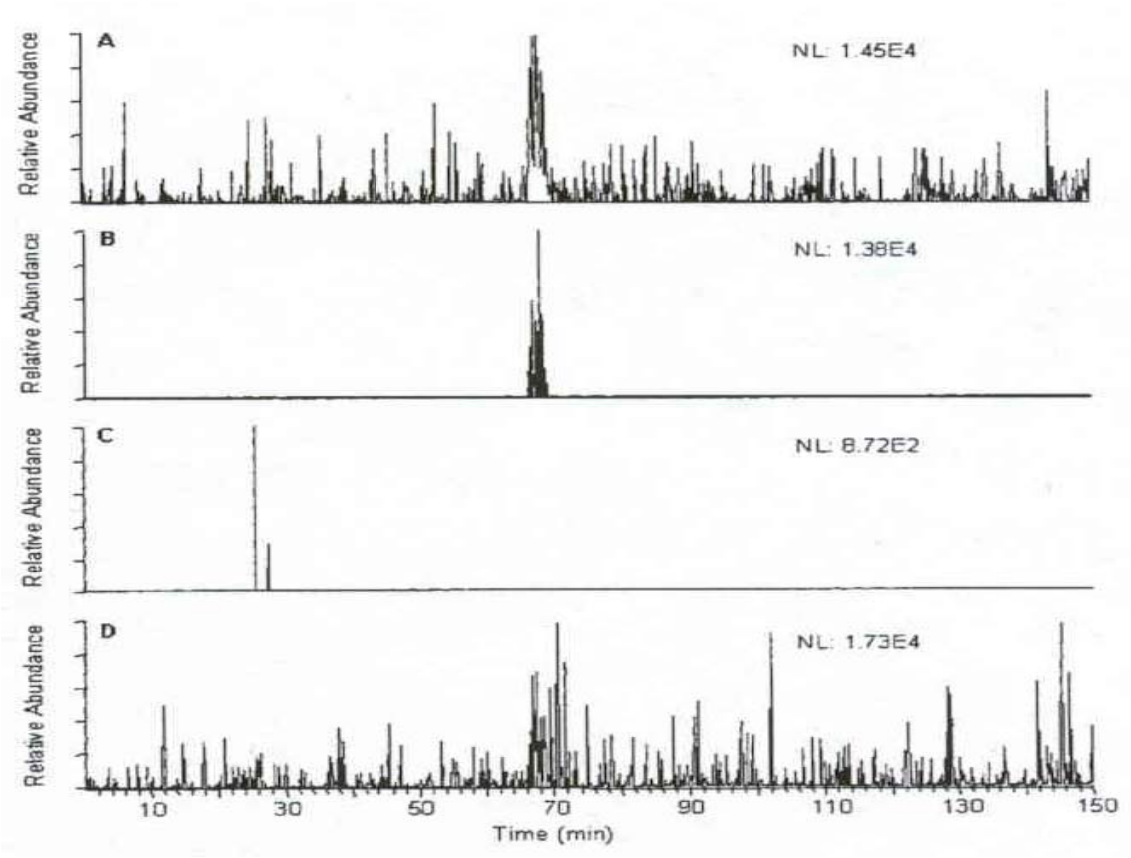


Figure 2-4. LC-MS-MS precursor ion scan of m/z 118 for protein digest spiked with γECG-DHP (m/z 425) and LVACGAK-DHP (m/z 778). (A) RTIC (m/z 400-2000) for 20 pmol/μg spiked digest; (B) precursor ion scan base peak trace of m/z 777.5-778.5 for 20 pmol/μg spiked digest; (C) precursor ion scan base peak trace of m/z 424.5-425.5 for 20 pmol/μg spiked digest; (D) RTIC (m/z 400-2000) for 5 pmol/μg spiked digest.

would come at the expense of sensitivity. The adduct levels in these experiments are most likely higher than can be expected in real samples. Thus, it appears that precursor ion scans may have insufficient sensitivity to detect unknown DHP adducts in protein digests.

In comparison, when the spiked digests were analyzed by LC-MS-MS with data-dependent scanning on the LCQ, SALSA was able to clearly differentiate both adducts down to a concentration of 1 pmol/μg of protein (Table 2-4). The algorithm scores for LVACGAK-DHP spectra were consistent over the three concentrations studied, with log scores ranging from 1.69 to 1.74. All of these scores are over the log score threshold value of 0.9 established in Table 2-3. The reproducibility of the scores demonstrates the relative concentration independence of SALSA scoring. The γECG-DHP adduct (m/z 425.14 or 425.16 eluting at 20 min) was detected in the 1 and 5 pmol/μg spiked digests with log scores of 0.95 and 1.23, respectively. However, this adduct was not detected in the 20 pmol/μg sample due to an anomalous CID spectrum that contained numerous high-abundance fragments with m/z greater than the precursor m/z. This caused the algorithm to classify the singly charged precursor as doubly charged and to miscalculate the expected loss. It is interesting to note from Table 2-4 that the two spectra with log scores between 0.61 and 0.9 in the 20 pmol/μg sample were verified as M+2 isotopomers of the two peptide-DHP adducts.

Table 2-4. Distribution of SALSA Scores for LC-MS-MS Spectra from a Tryptic Digest of BSA, GST, and MYO Spiked with HPLC Purified M+117 Isomers of γECG-DHP and LVACGAK-DHP at Concentrations of 1, 5, and 20 pmol/μg Protein.

Log(SALSA Score + 1)	1 pmol/μg	5 pmol/μg	20 pmol/μg
0	1000	939	963
0.01 – 0.3	21	12	7
0.31 – 0.6	3	6	2
0.61 – 0.9			2[b]
0.91 – 1.2	1 (0.95)[a]		
1.21 – 1.5		1 (1.23)	
1.51 – 1.8	1 (1.73)	1 (1.73)	3 (1.69, 1.73,1.69

[a] Numbers in parentheses indicate Log(SALSA score + 1) for spectra of confirmed DHP adducts. [b] CID spectra for M+2 isotopomers of the two adducted peptides.

BQ and IAA Adducts

To evaluate the ability of SALSA to detect adducts displaying different CID fragmentation patterns, two other adducts were studied. The first, S-carboxymethylcysteine (Figure 2-1), is a modification commonly encountered in peptide analysis. CID spectra of peptides containing S-carboxymethylcysteine exhibit neither adduct-derived ions nor losses. However, an ion pair separated by 161 m/z often is observed in the b- and/or y-ion series due to the S-carboxymethylcysteine residue (102). In experiments with these adducts, the 161 m/z ion pair is used as the only primary SALSA search criterion. Table 2-5 shows the distribution of SALSA scores for the protein digest spiked with AVAGCAGAR-CM. SALSA scored the AVAGCAGAR-CM spectrum (m/z 417.46 eluting at 19.4 min) in the top 9 scores out of a file containing 937 CID spectra. MS-MS of the $[M + 2H]^{2+}$ ion produced nearly complete sets of b- and y-series ions. Signals at m/z 299.1 and 460.0 corresponded to the b-series ion pair representing the cleavage of the carboxymethylated cysteine while signals at m/z 374.3 and 535.3 corresponded to the y-series ion pair. While SALSA cannot differentiate CM adducts from unadducted peptides to the same degree as with DHP adducts, the algorithm can identify a small subset of spectra that are most likely to be derived from CM adducts.

S-Cysteinylhydroquinone (Figure 2-1) adducts eliminate a benzoquinol-SH neutral fragment of 142 amu from single-charged precursors (95). This gives rise to ion pairs separated by 142 and 211 m/z due to the presence of the adducted cysteine in the b- or y-series and to cleavage of the hydroquinone-SH moiety from the b- or y-ion containing the adducted cysteine residue. SALSA search criteria for peptide-HQ adducts

include the neutral loss of 142 amu as a primary criterion, with the neutral loss of 160 amu and ion pairs of 142 and 211 m/z entered as secondary criteria linked to the primary loss. With these criteria, SALSA generated the score distribution of the peptide-HQ spiked digest shown in Table 2-5. The top score from the algorithm (log score 1.19) was the singly charged ion of the AVAGCAGAR-HQ adduct (m/z 883.59 eluting at 41 min). The next highest scoring ion in the output (log score 1.16) was the truncated AGCAGAR-HQ adduct (m/z 713.64 eluting at 42.3 min) discussed earlier (95).

Table 2-5. Distribution of SALSA Scores for CID Spectra from a Tryptic Digest of BSA, GST, and MYO Spiked with HPLC-Purified AVAGCAGAR-HQ (57 pmol/μg of Protein) and AVAGCAGAR-CM (12 pmol//μg of Protein).

Log(SALSA Score + 1)	Peptide-HQ[a] Spiked Digest	Peptide-CM[b] Spiked Digest
0	662	431
0.01 – 0.2		
0.21 – 0.4	3	
0.41 – 0.6	5	96
0.61 – 0.8		233
0.81 – 1.0		79
1.01 -1.2	4 (1.19, 1.16)[c]	8 (1.02)
1.21 -1.4		1

SALSA search criteria for HQ adducts included the neutral loss of 142 amu from singly charged precursors as a primary criterion, with the neutral loss of 160 amu from singly charged precursors and ion pairs of 142 and 211 m/z as secondary criteria. CM adducts were detected with only the ion pair of 161 m/z as a primary criterion. Numbers in parentheses indicate Log(SALSA score + 1) for spectra of confirmed adducts.

Discussion

SALSA is a data mining algorithm developed to detect user-specified fragmentation patterns in CID spectra obtained in LC-MS-MS data-dependent scan experiments. Detected fragmentation patterns may consist of a combination of product ions and/or neutral losses, as well as charged losses from double-charged precursors and ion pairs located anywhere in the spectrum. Multiple criteria may be scored simultaneously either as independent events or linked together in a hierarchical fashion. Normalization of spectra prior to analysis renders SALSA scores insensitive to concentration, facilitating detection of low-abundance analytes in a complex matrix. Search criteria may be optimized for a specific chemical class by conducting preliminary experiments with model compounds to elucidate diagnostic MS-MS fragmentation characteristics. However, appropriate search parameters may be derived in many instances by referring to the scientific literature, or simply making an educated guess about the expected fragmentation of particular adducts.

The analyses of peptide-DHP adducts, -HQ adducts, and -CM adducts demonstrate the potential of SALSA for searching the proteome for protein modifications by reactive electrophiles. Using the search criteria developed specifically for peptide-DHP adducts, SALSA was able to differentiate DHP adducts from unadducted peptides by comparing spectra scores to a threshold value established in preliminary experiments. DHP adduct spectra were consistently scored above a log score threshold value of 0.9, while none of the spectra from unadducted peptides received a score above this value. Selectivity and sensitivity of the algorithm was demonstrated by identification of γECG-

DHP and LVACGAK-DHP adducts spiked into a protein digest containing BSA, MYO, and GST. Even at a concentration of 1 pmol/μg protein, the adducted peptides consistently received the highest scores out of LC-MS-MS data-dependent scan experiments with approximately 1000 CID spectra.

Analyses of peptide-HQ and CM adducts provide further evidence of the applicability of SALSA to detecting adducts with different MS-MS fragmentation characteristics. SALSA gave the AVAGCAGAR-HQ and AGCAGAR-HQ adducts the highest two scores out of 674 CID spectra obtained in a LC-MS-MS analysis of a spiked protein digest. The detection of CM adducts on the basis of a single ion pair represents the most challenging situation to which SALSA can be applied. Ion pair detection is intrinsically noisy, with any given pair being detected in the majority of spectra. Nevertheless, the algorithm scored the AVAGCAGAR-CM adduct in the top 1% of all spectra scored in the spiked digest experiments. This suggests that detection of specific species on the basis of ion pair scoring is feasible, although not as effective as detection based on losses, product ions, or combinations of these parameters. While it is possible to carry out certain analyses using only ion pair detection, specificity is greatly increased by including ion pairs as secondary search criteria. Included as secondary criteria, ion pairs do not add appreciably to background noise and, instead, provide enhancement to scores for spectra that exhibit other primary criteria as well as information useful in spectra interpretation.

Post-analysis data reduction provides several distinct advantages compared to traditional tandem MS detection strategies. SALSA provides greater flexibility in the

types of fragmentation events that can be detected, such as charged losses and ion pairs; increasing the range of adducts that can be detected. Simultaneous evaluation of multiple characteristics by SALSA provides synergistic scoring of spectra that exhibit a diagnostic fragmentation pattern. Although CID spectra of peptide adducts may display multiple adduct-related characteristics, some characteristics may be weak or missing in the spectra of individual adducts as illustrated by the study of DHP-peptides (see Table 1). No single precursor ion or constant neutral loss scan provides universal detection of DHP adducts. However, since multiple characteristics are combined in the SALSA analysis, the different DHP adduct types can be detected reliably. Moreover, the SALSA method potentially provides lower limits of detection compared to tandem MS experiments. Sensitivity in precursor ion scan and neutral loss experiments is directly proportional to a given analyte's concentration, ionization potential, and relative abundance of the fragment ion being scanned for. However, SALSA scoring is largely independent of concentration and ionization potential, providing that the analyte is present at sufficient concentration to produce a CID spectrum in a data dependent scan experiment. While SALSA scores are proportional to the relative abundance of designated fragment ions, scores are not influenced by the concentration of other anaytes in a mixture, allowing for the detection of specific but weak fragmentation patterns.

Even in cases where tandem MS experiments successfully detect adducts, subsequent MS-MS experiments are required to verify the identity of putative adduct ions. In contrast, the SALSA algorithm identifies adducts from their MS-MS spectra, so there is no need to perform additional analyses (except in the case of DHP-adducts,

which require MS-MS-MS experiments to obtain reliable sequence information). An additional advantage of the data reduction approach is that after a single LC-MS-MS experiment the resulting data file can be analyzed numerous times using different search criteria. Several sets of search criteria may be developed for a single adduct type, or completely different criteria can be used to search for multiple adduct types within the sample. This presents a distinct advantage over precursor or constant neutral loss scanning experiments, where each change in the parameters monitored requires a new analysis.

Although the applications we present here are focused on identification of modified peptides, SALSA can be adapted to the detection of any species whose MS-MS spectra exhibit specific characteristics. In addition to peptides the program can be used to detect small molecule natural products, synthetic compounds, or metabolites. The program can be used to confirm the presence of known compounds in complex mixtures, or it may be used to search for unknowns that belong to a class of molecules that share sufficient structural similarities as to have predictable MS-MS fragmentation patterns.

CHAPTER THREE - Peptide Sequence Motif Analysis of Tandem MS Data with the SALSA Algorithm

Introduction

The heterogeneity of proteins in living systems presents a major analytical challenge in field of proteomics. Genetic polymorphisms and splice variants produce altered protein sequences which may or may not be accounted for in available sequence databases. Proteins also frequently contain posttranslational modifications, and proteins with such modifications vary in the number and types of modifications they possess. Because research in posttranslational modification of proteins is ongoing, protein sequence databases represent an incomplete record of the types and locations of these modifications. Even if a complete record were available, standard proteomics software packages such as Sequest do not routinely search for these modifications. In addition to naturally occurring posttranslational modifications, proteins may be modified by chemically reactive xenobiotics, which are most certainly not accounted for in sequence databases. A modified version of Sequest has been reported which is capable of detecting amino acid sequence variants for some proteins (36). Nevertheless, the general problem of detecting and characterizing unanticipated peptide variants remains a significant barrier to comprehensive characterization of proteomes.

As part of an ongoing effort in our laboratory to investigate new methods for the identification of modified peptides, we have developed an algorithm named SALSA (Scoring Algorithm for Spectral Analysis). The SALSA algorithm, as described

previously (103), scores spectra for product ions, neutral losses, charged losses, and ion pairs. Search criteria may be combined in a user-specified hierarchy; allowing for the inclusion of multiple characteristics to enhance sensitivity while simultaneously minimizing background interference. When applied in a proteomics context, search criteria may be tailored to specific amino acid modifications in peptide ions subjected to CID. The ranking of CID spectra according to their SALSA scores allows the user to quickly identify those scans corresponding to the peptide(s) with specific amino acid modifications.

Here we report an extension of the SALSA algorithm to facilitate the scoring of ion series. The earlier version of the program, while well suited for the detection of specific amino acid modifications with highly diagnostic CID fragmentation patterns, struggled to detect modifications that resulted in more subtle CID patterns. As the previously described experiments with Iodoacetic acid-adducted peptides demonstrated, select peptide modifications do not result in any diagnostic product ions or losses, leaving ion pairs as the only search criteria. The background signal for individual ion pairs was shown to be quite high, limiting the programs potential for the sensitive and specific detection of peptide modifications of this sort. Moreover, the earlier version of the program required advance knowledge of the peptide modification being sought after, as well as the predicted fragmentation pattern of said modification. By including ion series as potential SALSA search criteria it is now possible to search for specific peptides irregardless of modifications, overcoming the above limitations. We demonstrate that

SALSA can be used to selectively mine LC-MS-MS data not only for specific peptides but also for their variant and modified forms.

Experimental Procedures

Algorithm

The SALSA algorithm has been described in detail previously (103). SALSA was written to analyze MS-MS data from ThermoFinnigan Xcalibur .raw files. All spectra were recorded by data-dependent scanning on a ThermoFinnigan LCQ instrument (San Jose, CA). Full-scan analysis was used to identify candidate precursor ions, which then were automatically subjected to CID, and three MS-MS scans were recorded. Each set of three spectra is then averaged prior to further analysis. (Hereafter, the term "MS-MS spectrum" refers to the averaged spectrum derived from the three individual scans.) Prior to SALSA analysis, the data are preprocessed to eliminate nonfragment ions, estimate precursor charge, and normalize product ion intensities as a percent of the total ion current (% TIC). Ions with % TIC values below a user-specified threshold value (typically 0.2% TIC) are subtracted from the spectrum. Processed spectra then are evaluated for specific criteria and scores are computed from the % TIC values of the detected ions together with user-specified scoring heirarchy for the scoring criteria.

To detect ion series, we have extended the method we described for the detection of ion pairs in MS-MS spectra (103). Ion series are defined as a group of ions (i_1, i_2, i_3,...i_n) separated by specific m/z values (m_1, m_2, m_3,...m_n), where $m_n = i_n - i_{n+1}$ (Figure 3-1A). It should be noted that lower subscripts in an ion series denote higher m/z values. In the case of peptide sequence motifs, the distances between ions in the series correspond to the average residue masses of the amino acids in their sequence in the peptide. To detect ion series, SALSA first generates a hypothetical list of fragment ions

separated by the average residue mass differences for amino acid series. This hypothetical ion series list corresponds to a "virtual ruler" in which the relative distances between ions in the series are fixed along the *m*/*z* axis. The first ion in this hypothetical series (i_1) is then aligned with the highest *m*/*z* fragment ion in the actual MS-MS spectrum being evaluated (Figure 3-1A). SALSA then detects the actual ions that align with the hypothetical ions within a user-specified mass tolerance (typically ±0.5 *m*/*z* unit). The ions detected by alignment with the hypothetical ion series are scored as described below. The hypothetical ion series is then aligned beginning with the next lower *m*/*z* ion in the MS-MS spectrum and the matches again are recorded and scored (Figure 3-1B). The user may specify a minimum number of ions *x* to be detected in order for the series to be scored. In the example depicted in Figure 3-1B, only two matches are detected and the spectrum would not be scored if $x > 2$. Because some MS-MS spectra may be missing the highest mass ion in a particular series, the hypothetical series also is matched to the spectrum beginning with the second hypothetical ion (i_2) and matches between real ions and hypothetical ions i_2 - i_n then are recorded and scored (Figure 3-1C). Alignments of the hypothetical ion series with MS-MS data are continued through ions i_{n-x}, where *x* is the user-specified minimum number of matches required for scoring.

Spectra scores are calculated from the % TIC values of detected ions corresponding to the hypothetical ion series i_1-i_n (Figure 3-1D). Scores are calculated as shown in Equation 3-1, where *N* is the number of detected ions that correspond to hypothetical ions i_1-i_n in the series, n is the number of ions in the hypothetical series, and I_1, I_2, I_3...I_n represent the % TIC values of the ions in the series. For spectra in which one

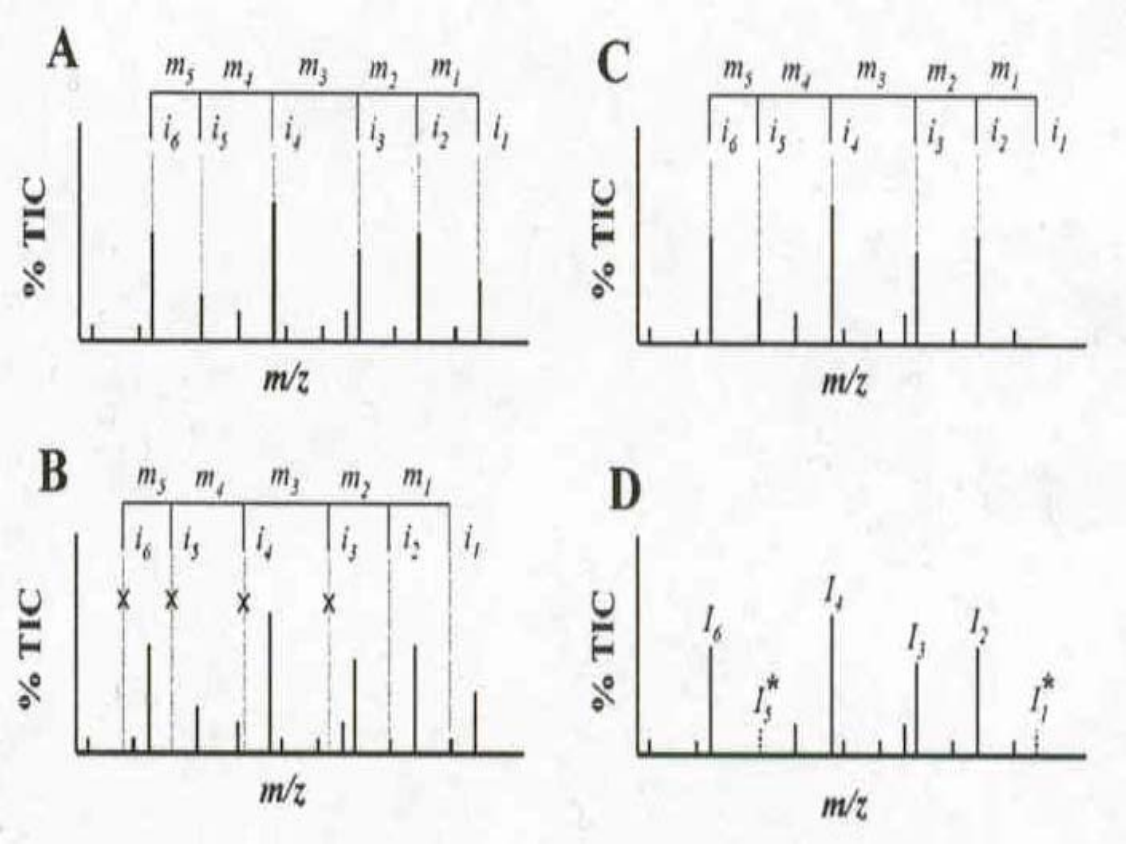


Figure 3-1. SALSA detection of ion series using hypothetical ruler. See text for discussion.

or more of the ions in the series are missing, the algorithm inserts a value I_n equal to the threshold value for ion detection discussed above (typically 0.2% TIC).

Equation 3-1. $\text{Score} = N(I_1 \bullet I_2 \bullet I_3 \ldots \bullet I_n)^{1/n}$

In the example depicted in Figure 3-1D, the score would be calculated as shown in Equation 3-2, where only four of the six ions in the series (i.e., I_2, I_3, I_4, and I_6) were actually detected in the spectrum and threshold % TIC values are used for I_1 and I_5, which were not detected.

Equation 3-2. $\text{Score} = 4(0.2 \bullet I_2 \bullet I_3 \bullet I_4 \bullet 0.2 \bullet I_6)^{1/6}$

As noted above, if $N < x$ (the user-specified minimum number of detected ions), then a score of zero would be assigned to the spectrum. The SALSA algorithm reports the scores for all sets of three averaged MS-MS scans receiving nonzero scores. The algorithm reports the total score, the scan number, LC retention time, the precursor *m*/*z*, and the ions detected in the MS-MS spectrum that matched the hypothetical series.

Tryptic Digestion

Bovine serum albumin (BSA) and human serum albumin (HSA) were purchased from Sigma. For tryptic digestion, 100 μg of BSA was dissolved in 0.02 mL of 8 M urea/2 M NH_4HCO_3. Tris(carboxyethyl)phosphine (Pierce) was added to a concentration

of 30 mM, and the solution was heated at 50 °C for 45 min. The solution was diluted to a final volume of 0.8 mL with 0.1 mL of water, and sequencing grade modified porcine trypsin (Promega) was added in a ratio of ~1:50 (w/w) enzyme/protein. The solution was incubated for 24 h at 37 °C. The solution was then subjected to LC-MS-MS analysis. For mixed digestions of BSA and HSA, the proteins were combined in a 1:1 ratio and the same digestion protocol was used.

LC-MS-MS Analysis

Peptide digests were analyzed by positive LC-ESI-MS-MS on a ThermoFinnigan LCQ instrument equipped with a standard ThermoFinnigan ESI source. Chromatography was done on a 5-μm Vydac 218TP column (1.0 × 250 mm) using $H_2O/CH_3CN/0.01\%$ CF_3COOH gradient elution at a flow rate of 0.03 mL/min. The gradient was initiated at 3% CH_3CN for the first 3 min, increased to 5% CH_3CN at 10 min and held for 10 min, then increased to 30% CH_3CN at 115 min, 75% at 125 min, 85% at 130 min, and finally to 95% CH_3CN at 35 min, and then held for 5 min before ramping back down to 3% CH_3CN at 145 min. In data-dependent scan experiments, the instrument was set to conduct repeated cycles of a single MS scan followed by three successive MS-MS scans of a selected precursor ion. The dynamic exclusion mass limit was set to ±1.5 *m*/*z*, with ions residing on the exclusion list for 5 min. Data analysis of the tandem MS experiments was performed with ThermoFinnigan Xcalibur software. Sequest analyses of the data were done using TurboSequest version 27 (revision 11) run under the Sequest Browser (ThermoFinnigan). Relevant settings for Sequest searches included enzyme =

trypsin, fragment ion tolerance = 0.00, peptide mass tolerance = 1.50, and maximum internal cleavage sites = 2. All Sequest searches were done with the NCBI nonredundant FASTA database at ftp://ftp.ncifcrf.gov/pub/nonredun/protein.nrdb.Z.

Results

SALSA Scoring of Ion Series

The SALSA algorithm permits the user to specify which features of MS-MS spectra to detect and score. For scoring of ion series, one may vary the number of ions in the series or the direction of the ion series search (i.e., the b-ion series or the y-ion series). One also may link the occurrence of product ions at specific *m*/*z* values (e.g., y- or b-series ions) as secondary to the ion series and score these product ions only in spectra that contain the ion series. The advantages of using these different search criteria alone or in combination are described below.

A screen capture image of the SALSA user interface is shown in Figure 3-2. An amino acid sequence motif is entered as an "ion series" and any linked secondary search criteria (e.g., product ions) are entered below the ion series search string as indicated. In Figure 3-2, the entered series with linked product ions will perform search 8 in Table 3-1 (see below). After search criteria are entered and an LC-MS-MS data file is selected, SALSA analysis is initiated by clicking the "score" button. Figure 3-3 depicts a screen capture image of the SALSA output screen. The output includes the SALSA score, HPLC retention time, and precursor *m*/*z* for each set of three averaged MS-MS scans. The right-hand column indicates ions in the MS-MS spectra that were detected and scored by SALSA based on user-specified search criteria. In Figure 3-3, the MS-MS scans are ranked in order of decreasing SALSA score. The search output depicted in Figure 3-3 is for search 8 from Table 3-1 (see below).

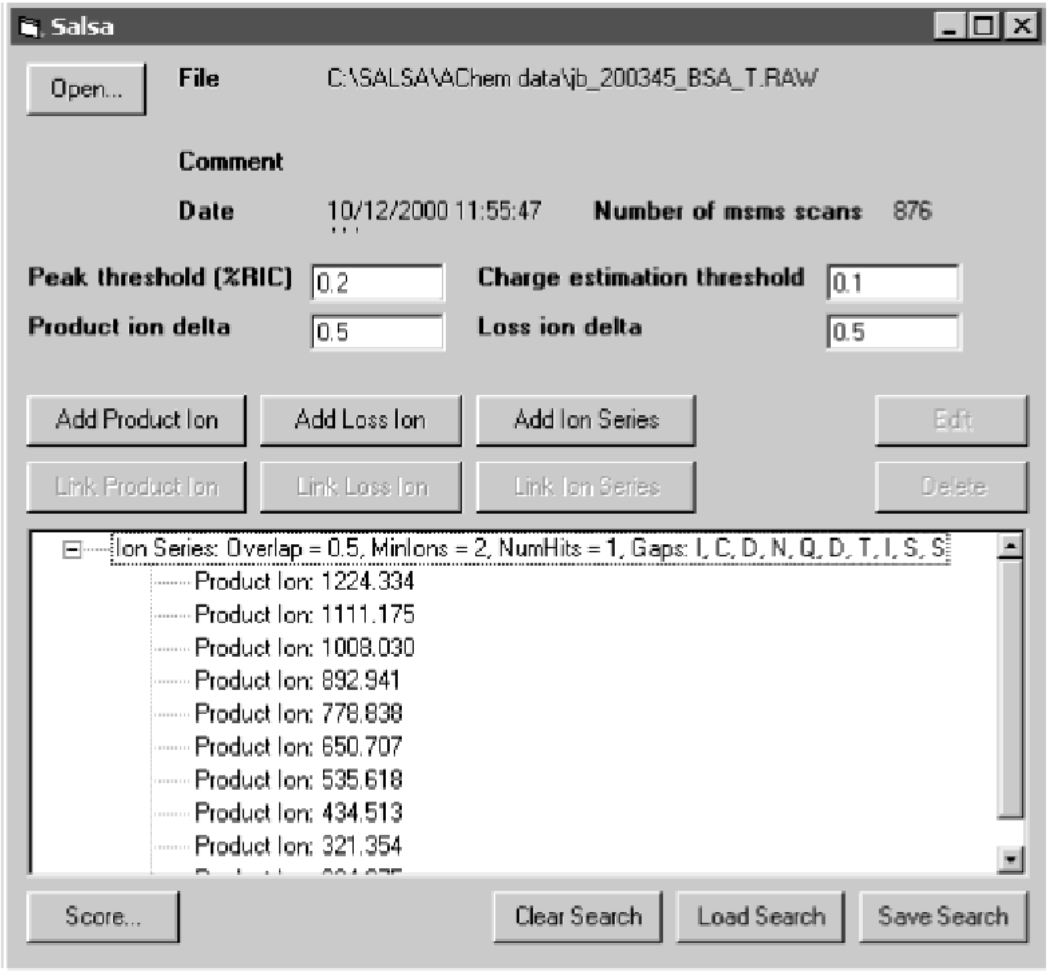


Figure 3-2. SALSA User interface. The entered ion series and linked product ions correspond to search 8 in table 3-1.

Table 3-1. Effects of Ion Series Search Parameters on the SALSA Scores and Rankings for MS-MS Spectrum of the BSA Peptide YICDNQDTISSK[a]

Search	Motif Searched	Search Method [b]	SALSA Score (Rank)
1	QD	Y1	4.58 (102)
2	DNQDTI	Y1	12.81 (1)
3	ICDNQDTISS	Y1	11.97 (1)
4	SSITDQNDCI	B1	2.02 (4)
5	ICDNQDTISS / SSITDQNDCI	Y1 / B1	13.99 (1)
6	ICDNQDTISS / SSITDQNDCI	Y1 / B2	16.89 (1)
7	SSITDQNDCI / ICDNQDTISS	B1 / Y2	4.03 (2)
8	ICDNQDTISS	Y1 / y"-ions2	59.15 (1)
9	SSITDQNDCI	B1 / b-ions2	10.01 (2)
10	ICDNQDTISS / SSITDQNDCI	Y1 / y"-ions2 // B1 / b-ions2	69.34 (1)

[a] All SALSA sequence motif searches were done on the same LC-MS-MS data file, in which the peptide YICDNQDTISSK was detected in scans 1613-1615 by CID of the doubly charged ion at m/z 693.92. Search methods are denoted by a letter or phrase followed by a number. The letter Y indicates that the search sequence motif was entered in the N- to C- terminal direction and corresponds to a search of the y-ion series. The letter B indicates that the search sequence was entered in the opposite direction and corresponds to a search of the b-ion series. The number following each letter indicates

whether the motif was specified as a primary or secondary search characteristic. Combined searches involving multiple search parameters are indicated by entries separated by a slash. Where indicated, b-ions or y”-ions were entered as secondary product ion search characteristics linked to the indicated primary motif.

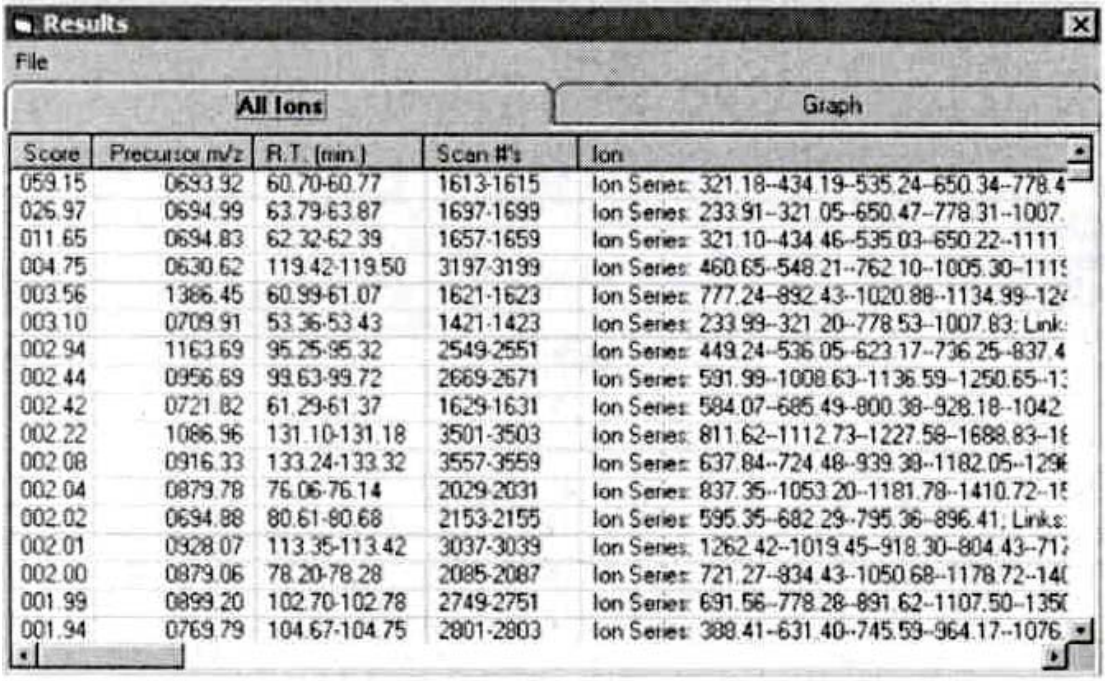


Figure 3-3. SALSA search output. The search output depicted is for search 8 in Table 3-1.

A sample of BSA was digested with trypsin and analyzed by LC-MS-MS with data-dependent scanning. The data file then was searched with the SALSA algorithm for MS-MS scans that corresponded to the BSA tryptic peptide YICDNQDTISSK, which was detected in scans 1613-1615 of the analysis (MS-MS of the doubly charged ion at *m*/*z* 693.92). The effects of manipulating ion series search parameters are summarized in Table 3-1. The first three entries in Table 3-1 illustrate the effect of the number of ions used in the search motif on the SALSA score for the target MS-MS scans 1613-1615 and the ranking of scans relative to other MS-MS scans in the datafile. The data file was searched with segments of the motif "ICDNQDTISS" (i.e., 10 linked ion pairs separated by 113, 103, 115, 114, 128, 115, 101, 113, 87, and 87 amu). Entry of these parameters in this order creates a "virtual ruler" that matches the y-series ions in an MS-MS spectrum (Figure 1). A minimum of two ions matching the search series was required for a "hit" (i.e., for an MS-MS spectrum to receive a score). Search 1 used the motif "QD", which corresponds to a three-ion series. This search gave a score of 4.58 to the target scans (1613-1615), which was ranked 102nd of the 876 sets of three averaged MS-MS scans in the data file. The highest score (6.77) was given to scans 2705-2707, which were not from MS-MS of the desired peptide. In searches 2 and 3, use of the longer search motifs DNQDTI and ICDNQDTISS allowed SALSA to assign the highest ranked scores (12.81 and 11.97, respectively) to the target scans.

The next five entries in Table 3-1 illustrate the effect of search series direction on SALSA scoring. In search 4, entry of the ICDNQDTISS sequence in reverse order (i.e., SSITDQNDCI) constitutes a b-series search and yielded a lower SALSA score of 2.02,

which was the fourth ranked score. This reflects the considerably lower intensity of the b-series ions relative to the y-series ions in the target MS-MS spectrum (not shown). The SALSA algorithm also permits search queries to be combined and prioritized. Thus, the y-series (ICDNQDTISS) and b-series (SSITDQNDCI) motifs can both be searched with equal priority and their scores combined (search 5). This search awarded the highest ranked score (13.99) for the correct MS-MS scan.

Search parameters in SALSA can also be ranked in priority as primary or secondary. A primary characteristic is scored whenever it is detected, whereas a secondary characteristic is scored only when the primary characteristic to which it is linked is also detected (15). A SALSA search with the y-series (ICDNQDTISS) denoted as a primary characteristic and the b-series (SSITDQNDCI) as a secondary characteristic (search 6) awarded the highest ranked score (16.89) to the correct MS-MS scan.

MS-MS of peptides yields both y- and b-series product ions at specific *m*/*z* values. The last three entries in Table 3-1 indicate the utility of using expected y- or b-series product ions as secondary search characteristics, which are scored only if the ion series corresponding to the target peptide is also detected in the same MS-MS scan. Figure 3-2 depicts a search of the y-series motif ICDNQDTISS together with the expected y-series ions (*m*/*z* 147.2, 234.3, 321.4, 434.5, 535.6, 650.7, 778.8, 892.9, 1008.0, 1111.2, 1224.3) as linked secondary characteristics. This search (search 8) awarded the highest SALSA score (59.15) to the target MS-MS scans. Comparison of this score to that from search 3 (11.97 for the same ion series searched without linked product ions) illustrates the ability of secondary scoring of product ions to amplify the SALSA score. A similar search

(search 9) of the b-series motif (SSITDQNDCI) with the b-series product ions linked as secondary search characteristics yielded a considerably lower score of 9.59, which was nevertheless the second ranked score in the datafile. A combined search of the y-series motif and the b-series motif each linked to its expected product ions awarded the highest SALSA score (69.34) to the correct MS-MS scans (search 10).

The results presented here are exemplary of SALSA analyses of MS-MS data for peptide sequence motifs. Several different search strategies (e.g., y-ion series, b-ion series, linked product ions, etc.) can identify MS-MS scans corresponding to a target peptide sequence. The most thorough strategy would employ both b- and y-ion series motifs with b- and y-series product ions linked as secondary search characteristics. Nevertheless, y-series motifs with linked y-product ions usually are sufficient to detect MS-MS scans of doubly charged ions of tryptic peptides, in which y-ions often are more abundant than b-ions. As shown above (Table 3-1, searches 1-3), search series containing more hypothetical ions will match a greater number of actual ions in the MS-MS scans corresponding to the target peptide sequence. Search motifs corresponding to the central motif in a peptide typically yield the highest scores, as these correspond to the most intense ions in the MS-MS spectrum, and low mass fragment ions at the end of the sequence are often omitted in LCQ CID spectra. Spectra with missing ions can receive somewhat lower scores because the scoring formula for ion series inserts a TIC value of 0.2% (the user-defined threshold) for ions that are not detected.

Significance and Use of SALSA Scores

SALSA scores are determined by several factors including (1) the search strategy used, (2) the length of the search motif, (3) the number of ions that match the search series, and (4) the intensities of the scored ions. SALSA scores do not provide an absolute measure of spectral quality or the fidelity of the match between the search motif and the MS-MS spectrum. Thus, the absolute values of SALSA scores are less important than the relative values for different MS-MS scans in a data set. A ranking of the MS-MS scans by SALSA score quickly identifies those MS-MS scans originating from the target peptide or its modified or variant forms.

Figure 3-3 depicts the output of a typical SALSA search. The three highest ranked MS-MS scans received scores of 59.15, 26.97, and 11.65, respectively. Inspection of the precursor *m*/*z* values indicates that all correspond to the doubly charged *m*/*z* expected for YICDNQDTISSK (*m*/*z* 694.259). The "ions" column on the right of the output display indicates that similar *m*/*z* ions were scored for these three sets of MS-MS scans. Thus, a quick inspection of the SALSA output provides evidence (via the precursor *m*/*z* and the occurrence of the same product ions) that the three highest scoring scans are all "hits" for the desired peptide sequence. The averaged MS-MS scans 1613-1615 displayed a complete y-ion series and a nearly complete b-ion series, whereas MS-MS scans 1697-1699 and 1657-1659 had several missing b- or y-ions, thus accounting for their lower SALSA scores. Other scored scans display considerably lower SALSA scores and bear no apparent characteristics of the target peptide. In our experience, SALSA analyses typically rank MS-MS scans for target peptides among the top five

scans and most often as the highest scoring scan. On the other hand, in SALSA analyses where a search motif fails to identify higher scoring scans, inspection of LC-MS-MS data files usually reveals that no MS-MS spectrum was recorded for the desired precursor ion.

We have not attempted to statistically analyze the relationship between SALSA scores, MS-MS spectral characteristics, and peptide sequences. As noted above, SALSA scores are highly dependent on search strategy and spectral characteristics and are useful only as relative measures of concordance between spectral features and search criteria. In its current form, SALSA is a tool for mining MS-MS data sets to quickly identify those spectra that display characteristics of the target peptide sequence motif. However, depending on the search strategy used and features of individual spectra, SALSA may assign relatively high scores to spectra that do not contain the motif of interest. Thus, it is important that the user validate SALSA "hits" by inspection of the indicated spectra.

LC-MS-MS and SALSA Analysis of a BSA Tryptic Digest

The Sequest algorithm has become a widely used tool for identifying proteins from MS-MS data. Sequest correlates MS-MS spectra with virtual spectra derived from database sequences consistent with the measured *m*/*z* of the peptide precursor ion (32). SALSA directly scores MS-MS spectra for user-defined ion series that match specific peptide sequences, regardless of the *m*/*z* value of the peptide precursor ion. To compare these complementary approaches to MS-MS data evaluation, we used Sequest and SALSA to analyze a datafile generated by LC-MS-MS of a tryptic digest of BSA (Figure 3-4). The LC-MS-MS datafile contained 876 sets of three MS-MS scans. Sequest

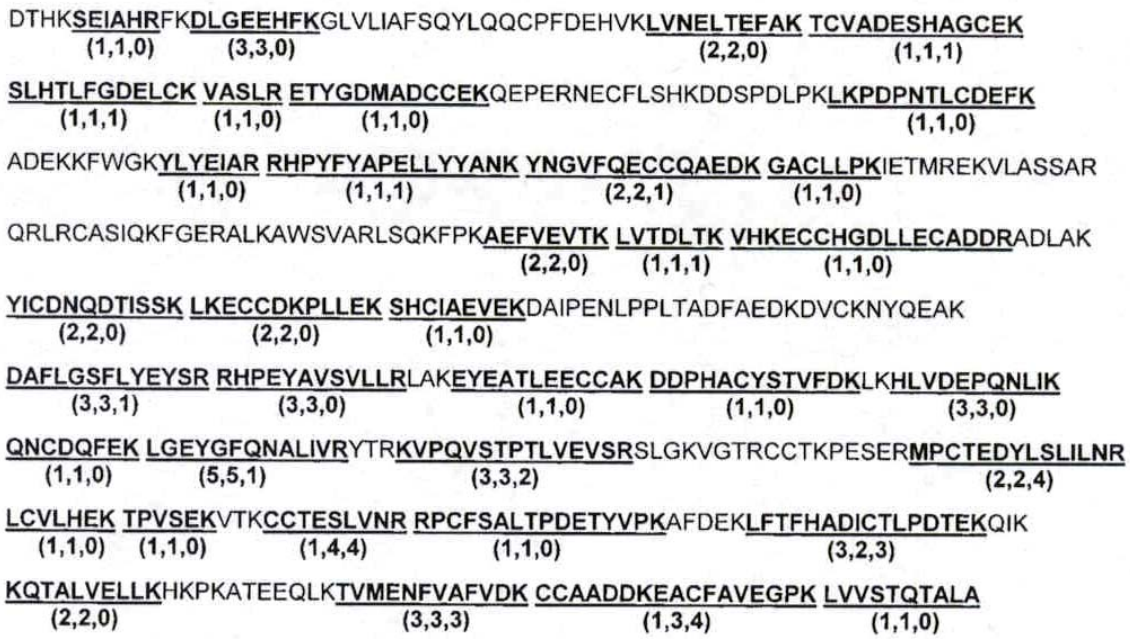


Figure 3-4. BSA tryptic peptides for which MS-MS scans were detected by Sequest and SALSA. Detected peptides are indicated by highlighted sequences. Numbers in parentheses in the format (*a*,*b*,*c*) beneath each highlighted sequence indicate (*a*) the number of MS-MS scans for the indicated sequence detected by Sequest, (*b*) the number of MS-MS scans for the indicated sequence detected by SALSA, and (*c*) the number of MS-MS scans for variants of the indicated sequence detected by SALSA.

assigned MS-MS spectra to 37 BSA tryptic peptides with correlation scores of greater than 1.50, corresponding to 66.2% coverage by amino acid sequence. A SALSA analysis of the same data file was performed with ion series searches corresponding to the central sequence of the peptide and entered in an N- to C-terminal direction (i.e., a y-ion series search (see above)). For peptides containing even numbers of amino acids, a search motif of up to eight amino acids in length was entered; for peptides containing odd numbers of amino acids, a search motif of up to nine amino acids was entered. N- and C-terminal amino acids were not included in search motifs. The SALSA analysis assigned significant scores to the same MS-MS spectra assigned to the peptides by Sequest. In virtually all cases, the highest SALSA score was assigned to the same MS-MS spectrum assigned to the sequence by Sequest. SALSA scores of MS-MS spectra assigned by both Sequest and SALSA ranged from 9.69 to 112.36 with a mean value of 48.20 ± 25.50. The mean raw SALSA score for all spectra in these analyses was 1.82 ± 0.38. Thus, SALSA scores for MS-MS scans corresponding to target peptide sequences were well differentiated from scores for other MS-MS scans, and there was a high overall concordance between Sequest and SALSA in the assigment of MS-MS scans to specific BSA peptide sequences.

SALSA detected MS-MS spectra of several variant peptides not assigned by Sequest. These additional MS-MS spectra resulted from CID of precursor ions that differed by more than ±1.5 m/z unit from the target precursor peptide ion of the same charge state. All detected MS-MS scans assigned as variants of a target sequence displayed very strong y-ion series identity or homology (i.e., the series was displaced

along the *m*/*z* axis) with the unmodified peptides (data not shown). As indicated in Figure 3-4, one MS-MS scan corresponding to a variant peptide was found for eight of the peptides detected. Two variants were found for another peptide (KVPQVSTPTLVEVSR), three variants were found for two other peptides (LFTFHADICTLPDTEK, TVMENFVAFVDK), and four variants each were found for three other peptides (MPCTEDYLSLILNR, CCTESLVNR, CCAADDKEACFAVEGPK).

A more detailed summary of the data for MS-MS scans corresponding to variant forms of the peptides MPCTEDYLSLILNR and CCAADDKEACFAVEGPK is presented in Table 3-2. Of the MS-MS spectra for the five MPCTEDYLSLILNR peptides reported, only that for the doubly charged ion of the unmodified peptide (*m*/*z* 834.90) was detected by Sequest. The variants were detected by SALSA on the basis of their intact y-ion series. The M + 16 and M + 32 variants reflect probable oxidative modification at the cysteine and cysteine/methionine, respectively, whereas the M - 32 variant indicated loss of sulfur from the cysteine residue, as verified by inspection of the MS-MS spectra. The CCAADDKEACFAVEGPK peptides represent incomplete tryptic cleavage products found by motif searching with the ACFAVEGPK sequence. Modifications of M - 48, M + 26, and M + 48 were all assigned to the N-terminal CC residues by inspection of the MS-MS spectra.

Finally, SALSA detected what appear to be N-terminal carbamylation adducts of the peptides SLHTLFGDELCK, RRHPYFYAPELLYYANK, HLVDEPQNLIK, and KVPQVSTPTLVEVSR (data not shown). These are typically detected as M + 43

Table 3-2. Detection of MS-MS Scans for Variant Forms of MPCTEDYLSLILNR and CCAADDKEACFAVEGPK in a Tryptic Digest of BSA

Peptide	Precursor m/z	Scan No.	SALSA Score	Sequest Xcorr [a]	Identity / Comment
MPCTEDYLSLILNR [b]	834.90	3565-3567	88.14	4.68	MPCTEDYLSLILNR 2+
	851.14	3545-3547	83.23	NS [c]	MPCTEDYLSLILNR 2+ (M+32); [d] modification on MPC
	843.15	3433-3435	65.38	NS	MPCTEDYLSLILNR 2+ (M+16); modification on M
	819.12	3553-3555	56.02	NS	MPCTEDYLSLILNR 2+ (M-32); modification on C
	836.64	3573-3575	46.35	NS	MPCTEDYLSLILNR 2+ (M+4)
CCAADDKEACFAVEGPK	586.51	2033-2035	45.38	3.94	CCAADDKEACFAVEGPK 3+
	855.65	2109-2111	36.68	NS	CCAADDKEACFAVEGPK 2+ (M-48); modification on CCAA
	595.37	2189-2191	34.14	NS	CCAADDKEACFAVEGPK 3+ (M+26); modification on CC
	903.29	2057-2059	28.29	NS	CCAADDKEACFAVEGPK 2+ (M+48); modification on CC
	892.89	2185-2187	24.58	NS	CCAADDKEACFAVEGPK 2+ (M+26); modification on CC
	878.00	2041-2043	24.06	NS	CCAADDKEACFAVEGPK 2+

[a] Xcorr score for match of MS-MS scan to the listed peptide sequence. [b] SALSA analysis of MS-MS data were done as described above using the highlighted search motifs entered as an ion series with linked y"-ions as secondary search characteristics. [c] NS, not scored. [d] Denotes the calculated difference in amu between the detected peptide and the theoretical mass for the unmodified peptide sequence.

adducts of doubly charged peptide ions. These adducts display HPLC retention times different from that of the unmodified peptides. MS-MS spectra of the adducts appear similar to those of the unmodified peptides. The y-ion series generally is unchanged, but some b-series ions appear at 43 *m*/*z* units above the expected values. The formation of N-terminally carbamylated peptides during digestions or storage in urea has been reported previously (104). Our experience is that storage of digests in urea results in a progressive accumulation of carbamylated peptides, which are not detected by Sequest unless N-terminal carbamylation is specified as an expected modification. SALSA assigned comparable scores to both carbamylated and unmodified peptides.

SALSA Analyses of Combined BSA/HSA Tryptic Digests

The ability of SALSA to detect MS-MS spectra based on ion series searching raises the possibility of using SALSA to identify variant peptide sequences in protein digests. Variant peptide sequences would correspond to protein products of mutant or polymorphic genes or junction points of splice variants and are of considerable biomedical significance (105). To model a situation in which a protein sample contains highly similar variant peptides, we mixed HSA and BSA at a ratio of 1:1 (mol/mol), digested the mixed samples with trypsin, and analyzed by LC-MS-MS with data-dependent scanning. HSA contains 14 tryptic peptides that differ from corresponding BSA peptides by a single amino acid substitution, 5 others that differ by two amino acids, and approximately another 10 that display greater than 50% sequence identity. We analyzed the data file for these mixtures with the SALSA algorithm to detect four

different pairs of peptides that differed by one or two amino acids (Table 3-3). All searches were done with the indicated sequence motifs as written (i.e., in the y-series direction).

The first entry in each group indicates the SALSA score assigned to the MS-MS scan for the BSA peptide when searched with the BSA sequence motif. Each SALSA search gave the MS-MS spectrum from the target BSA peptide the highest ranked score in the data file. The second entry in each peptide group indicates the score calculated by SALSA for the MS-MS scans corresponding to the HSA peptide variants. These MS-MS scans received lower SALSA scores than those given to the MS-MS scans for the corresponding BSA peptides. Moreover, their rankings of the HSA peptide MS-MS scans among all the MS-MS scans in each datafile varied from 3rd (for LVNEVTEFAK) to 100th (for DLGEENFK). A search of the data with the search motif CDNQDTIS from the BSA peptide YICDNQDTISSK assigned a score of 3.08 (31st) to the MS-MS scans for the HSA peptide YICENQDSISSK, which differs at two amino acids from the BSA peptide. Thus, although SALSA can detect MS-MS spectra for sequence variant peptides, the scores and rankings given to the detected variants may vary considerably due to the effects of the modifications on the spectral characteristics evaluated by SALSA.

These searches all used an ion series as a primary search characteristic with specific product ions linked to the ion series as secondary search characteristics. MS-MS spectra for peptides that match the search motif displayed both the expected ion series and the expected product ions and received the highest SALSA scores and ranking.

Table 3-3. SALSA analysis of LC-MS-MS data from BSA-HSA peptide mixtures

Target Sequence	**Search Motif**	**SALSA Score (Rank)**
DLGEE**H**FK (b[a])	LGEE**H**F (b)	50.63 (1)
DLGEE**N**FK (h[b])	LGEE**H**F (b)	2.49 (100) / 15.77 (4)[C]
DLGEE**N**FK (h)	LGEE**N**F (h)	36.67 (1)
LVNE**L**TEFAK (b)	VNE**L**TEFA (b)	74.34 (1)
LVNE**V**TEFAK (h)	VNE**L**TEFA (b)	7.09 (3)
LVNE**V**TEFAK (h)	VNE**V**TEFA (h)	64.58 (1)
YIC**D**NQD**T**ISSK (b)	C**D**NQD**T**IS[d] (b)	48.97 (1)
YIC**E**NQD**S**ISSK (h)	C**D**NQD**T**IS[d] (b)	3.08 (31)
YIC**E**NQD**S**ISSK (h)	C**E**NQD**S**IS[e] (h)	36.35 (1)

[a] Bovine. [b] Human. [c] Score for FKDLGEENFK 2+. [d] Used y''-ions predicted for BSA peptide YIC**D**NQD**T**ISSK as secondary search characteristics. [e] Used y''-ions predicted for HAS peptide YIC**E**NQD**S**ISSK as secondary search characteristic.

MS-MS spectra of peptides with substitutions or modifications may contain much of the ion series, but shifts in the series along the *m*/*z* axis may cause the actual product ion values to differ from those used as the secondary search criteria. This is exemplified by searches for MS-MS spectra of the HSA peptide DLGEE**N**FK using the BSA search motif LGEEHF. The substitution of N for H in the HSA sequence displaces the entire y-ion series from the values expected for the BSA peptide DLGEE**H**FK. Thus, much of the series (i.e., DLGEE) was preserved, but there was little correspondence between the BSA product ions used as secondary search parameters and the actual y-ions from the HSA peptide. This greatly lowers both the SALSA score and the ranking.

The third entry in each group in Table 3-3 indicates the SALSA scores and rankings for the MS-MS spectra of human sequence variant peptides when human sequence search motifs were used. All of the HSA peptide MS-MS spectra were ranked first in analyses of their respective data files. In the case of the target HSA sequence DLGEENFK, MS-MS scans were detected instead for the incomplete digestion product FKDLGEENFK. These data indicate that SALSA can discriminate between MS-MS spectra of closely related sequence variant peptides. However, optimal discrimination between variant forms is achieved when the SALSA search criteria are optimized to the target peptide. Application of SALSA to sequence motif searching may prove valuable for investigation of genetic polymorphisms at the level of the proteome.

SALSA Versus Other Algorithms for Analysis of Tandem MS Data

At the time that the information presented in this chapter was first published (106), several other algorithms and associated software tools were already are available for the analysis of tandem MS data including Sequest (107), Mascot (25), Pep-Frag (108), MS-Tag (23), and Pep-Sea (30). In the years following the original SALSA manuscripts, other algorithms have been introduced which will be discussed in greater detail in following chapters.

There are two fundamental ways in which SALSA differs from these tools. First, SALSA finds MS-MS spectra that display characteristics of a sequence motif rather than database sequences that match spectra. This is the reverse of the approach taken with Sequest and related tools. All of these tools identify proteins from tandem MS data by comparing features of the spectra with those of theoretical spectra of protein sequences in databases. In short, the other proteomics tools identify sequences from MS-MS scans; SALSA identifies MS-MS scans from sequences. This can be both an advantage and a disadvantage. The advantage is that SALSA is much more flexible that other sequence driven proteomics tools. SALSA may be the only available tool for detecting spectra of peptides containing certain modifications which disrupt the predicted CID fragmentation pattern of peptide ions, such as dehydromonocrotaline adducts (103), or peptides containing multiple unpredicted modifications. Because SALSA searches are focused on a particular peptide sequence motif or fragmentation pattern, the output contain a more exhaustive list of potential matches ensuring that no potentially useful data is excluded. However, these advantages can sometimes be disadvantages. The flexibility of SALSA

requires that the user develop custom search criteria for each analysis, whereas other proteomics tools automate the generation of search criteria directly from the sequence databases. The exhaustive output list resulting from SALSA searches results in large output files for each target sequence, and can require significant man-hours to confirm potential hits.

The second important difference is that the SALSA algorithm searches for spectral features without regard either to a peptide ion precursor *m*/*z* or to specific *m*/*z* values of product ions. This allows SALSA to identify MS-MS scans that contain *m*/*z* signals in some specified relation to each other, regardless of their absolute positions on the *m*/*z* axis. SALSA then can identify MS-MS scans that correspond to the target peptide and any variants, whether anticipated or not. The ability to identify unanticipated variants is the key advantage of SALSA over existing tools. Sequest and related tools can correlate MS-MS data of modified peptides with database sequences if the user specifies the nature of the modification and the amino acid modified (32). Similarly, one may specify either "missed cleavages" or "no enzyme" to increase the ability of Sequest to accurately match MS-MS spectra to database sequences. One also may search databases containing variant sequences to enable the successful correlation of MS-MS spectra of sequence variant peptides (36). Nevertheless, unanticipated modifications preclude correct correlation of precursor *m*/*z* or the MS-MS data with database protein sequences. For example, Steen and Mann recently reported that thioether oxidation in peptides containing alkylated cysteine residues shifted both the precursor *m*/*z* and the product ion *m*/*z* values in CID of the modified peptides (109). This is important because

thiol alkylation (and oxidation of the resulting thioethers) is a near-universal element of proteomic analysis. Moreover, protein differential expression analysis with isotope-coded affinity tags (ICAT) (18) also is subject to this complication. Thus, SALSA may be particularly useful for the evaluation of LC-MS-MS data from ICAT analyses.

Analysis of primary sequence variation in proteins is a challenging problem to which SALSA may be applied. Although the error-tolerant sequence tag approach described by Mann and Wilm (30) is applicable to this problem, the sequence tag approach requires partial de novo interpretation of all spectra to be evaluated. This is impractical with data files containing hundreds or thousands of spectra. However, initial screening of data files with SALSA for sequence motifs of interest could identify candidate MS-MS scans worthy of further analysis with sequence tags or related approaches.

Conclusion

The SALSA algorithm has been modified to include ion series as one of the search parameters available to users. The ion series search parameter greatly increases the algorithm's effectiveness at detecting MS-MS spectra of unmodified peptide ions, as well as peptide ions containing unanticipated modifications or modifications with weakly diagnostic CID fragmentation patterns. In contrast to other proteomics tools, SALSA makes no assumptions about the expected mass of peptide precursor or fragment ions. Instead, the algorithm tries to find evidence of targeted fragmentation patterns in each CID spectrum in a data dependent scan LC-MS-MS data file. Because of the exhaustive nature of SALSA searches, the program is able to identify nearly all of the variants of a target peptide sequence in a particular data file.

The principal caveat to ion series searches is that they are most effective at detecting either unmodified peptides, peptides with anticipated modifications, or those with modifications located near either the N- or C-terminus. Any modification to the target peptide sequence, if unanticipated, produces a shift in the sequence of fragment ions. If this shift occurs near the ends of the peptide, the series of b- or y- fragment ions remains mostly intact, ensuring that the spectra with this type of shift receive reasonably high SALSA scores. However, unanticipated modifications located at amino acids near the center of the peptide sequence create a shift near the middle of the anticipated ion series, resulting in approximately half of the series being scored. Since ion series scores are proportional to the number scored ions, spectra with centrally located modifications may receive scores that are not ranked near the top of the output list.

SALSA is an important addition to the proteomics toolbox. It will prove to be the most useful for researchers interested in mining the proteome for protein modifications. The greatest utility of SALSA will be in combination with Sequest and similar tools. An initial analysis of MS-MS data files with Sequest identifies proteins that are represented in the analyzed sample. The user then may use SALSA motif searches to mine the data for MS-MS scans corresponding to modified and variant peptide forms. Inspection of these MS-MS scans allows confirmation of the sites and masses of modifications.

CHAPTER FOUR – P-Mod: A Statistically Based Algorithm For Mapping Peptide Modifications Using Tandem MS Data

Introduction

Here we describe a new algorithm and software program called P-Mod, which detects MS-MS spectra corresponding to target peptide sequences along with variant and modified forms. In contrast to SALSA, which is used to conduct targeted searches for individual peptides or modifications, P-Mod can simultaneously screen for hundreds of peptides and can detect numerous even unanticipated modifications. P-Mod makes use of extreme value statistics to assign p value estimates to sequence-to-spectrum matches. The reported p values are scaled to account for the number of comparisons, so that error rates do not increase with the expanded search lists that result from incorporating potential peptide modifications. P-Mod enables the rapid discovery of protein modifications from MS-MS data and thus complements established tools for protein identification.

P-Mod screens data files for MS-MS spectra corresponding to peptide sequences in a search list. Modification of the primary peptide sequence will result in a shift in the peptide mass. This shift in the peptide mass may be experimentally observed as a difference between the measured mass of the modified peptide precursor ion (adjusted for charge state) and the predicted mass of the unmodified peptide. If a modification is located at a particular amino acid residue in the sequence, the mass shift also will be observed in the *m/z* values of some of the fragment ions.

P-Mod search lists typically consist of all potential peptides of between 5 and 30 amino acids in length expected from an enzymatic digest of a protein. P-Mod is applied in situations where the identity and sequence of the protein of interest is known or is established by prior data analysis (e.g., Sequest). Searches based on peptide lists from larger numbers of proteins would create a significant combinatorial problem and greatly extend analysis time. Moreover, the reported p values of hypothetical matches take into consideration the number of sequence comparisons made to each spectrum. Larger search lists result in matches with lower statistical significance and ultimately decrease the sensitivity of the analysis.

P-Mod considers each MS-MS spectrum in a data file as a potential match to each peptide search sequence. In cases where the precursor ion mass differs from the expected peptide mass, the program assumes the spectrum corresponds to a modified version of the original sequence. To score sequence-to-spectrum comparisons, the program generates an array of search criteria corresponding to the expected MS-MS fragment ions and incorporating the observed mass shift at all possible sequence positions. For each spectrum in a data file, P-Mod determines which sequence in the search list provides the best match. This match then localizes the modification to a specific amino acid position in the matched sequence. Ultimately, all putative matches are assigned a p value, which reflects the probability of a false positive sequence-to-spectrum match.

P-Mod Algorithm

Preliminary Workup of MS-MS Spectra

P-Mod operates on Thermo Xcalibur datafiles (v. 1.2 or higher) and .dta files. All MS-MS spectra with 50 or more fragment ions are eligible for P-Mod scoring and are subjected to preliminary workup. The fragment ions of each spectrum are indexed in 110 *m/z* increments or bins beginning with the lowest *m/z* ion recorded. Indexing effectively divides each spectrum into compartments and serves two purposes. First, compartmentalization improves the efficiency of fragment ion assignments and decreases analysis times. Second, indexing makes it possible to estimate the local background signal from unassigned ions in multiple regions of the MS-MS spectrum (see below). During workup the charge state of the precursor ion is estimated, either from the output accompanying the .dta files, or directly from the MS-MS spectrum as previously described for the SALSA algorithm (103). Spectra with less than 12% ion current above the precursor *m/z* are considered singly charged, whereas those with a higher percentage of ion current above the precursor *m/z* are assumed to arise from doubly charged precursors. Charge estimates are used to estimate a neutral precursor mass for each precursor ion and to determine whether multi-charged ions are assigned when calculating background (see below).

Mass Shift Estimation

Before scoring, the mass difference between the estimated neutral mass of the precursor ion and the theoretical monoisotopic mass of each peptide sequence is

calculated (Equation 4-1). All comparisons with a positive mass shift or a negative mass shift that is no greater than an amino acid side chain mass are considered viable comparisons and are scored.

Equation 4-1. Mass shift = neutral precursor mass - sequence mass

To enable detection of triply charged peptide ions and to account for occasional errors in charge estimation, negative mass shifts with a magnitude greater than ¾ of precursor *m/z* are adjusted by recalculating the theoretical mass of the precursor assuming an additional +1 to the charge state.

Generation of Search Criteria

An array of customized search criteria is generated for every sequence-to-spectrum comparison, taking into consideration the primary peptide sequence, the observed mass shift, the precursor *m/z* and instrumental limitations of ion trap mass spectrometers. The first element in each search array is a list of all of the expected b- or y- series fragment ions for the unmodified peptide sequence. Succeeding elements in the search array consist of these same fragment ions, tailored to reflect the mass shift localized at different amino acid residues in the sequence. For example, if the mass shift was located on the N-terminal residue of a given peptide, one would predict that all of the b- series fragment ions and none of the y- series fragment ions would be mass shifted. If the mass shift was located on a residue near the middle of the sequence, the higher *m/z*

fragment ions of both series which contained the modified residue would all be mass shifted, whereas the low *m/z* fragment ions that did not contain the modified residue would not be shifted. P-Mod makes no assumptions about the potential locations of peptide modifications or about which amino acids are even subject to modification. Thus, every location in the sequence is used to generate a separate element in the search array. Positive mass shifts are localized to each possible sequence location. Negative mass shifts whose absolute value is greater than the amino acid side chain mass at a specific sequence location are not scored.

Before being applied as search criteria, the fragment ions in the search array are trimmed so as not to include those fragment ions that are likely to be missing from MS-MS spectra acquired on an ion trap MS instrument. Such trimming is necessary because missing ions negatively impact the scoring of sequence-to-spectrum comparisons (see below). The b_n ion and fragment ions within 2 units of the precursor *m/z* or those with a *m/z* less than 25% of the precursor *m/z* are not typically observed in ion trap data and are therefore not included as applied search criteria. Likewise, fragment ions with a *m/z* greater than twice the precursor *m/z* or greater than 2000 are not applied as search criteria.

Scoring of MS-MS Spectra

Each element in the search array with more than 6 applied search criteria is given a raw score according to Equation 4-2, where **n** = the number of applied search criteria, $\mathbf{i_n}$ = intensity of the largest ion within 1.25 *m/z* of the expected location for the n^{th} search

criterion, $\mathbf{b_{ci}}$ = background intensity in the index compartment which contains the scored ion, and **d** = distance in *m/z* between the scored ion and its expected location.

Equation 4-2. $\text{Score} = 1/n * \sum (\ln (1 + i_n / (b_{ci} * (1 + 3d^2))))$

Raw scores depend on the ratio of the number of detected vs. applied search criteria, the mass-accuracy of assigned ions, as well as the ratio of ion intensities of assigned ions compared to background intensity in the spectrum compartments containing those ions. The background intensity in Equation 4-2 is defined as the largest unassigned ion in each spectrum compartment. P-Mod attempts to assign every fragment ion in each compartment, either as one of the applied search criteria ions or as an expected water loss or multi-charged ion (if the spectrum is assigned multi-charged status during work-up). All ions within 1.5 amu of an expected ion are considered assignable to that species. The intensity of the largest remaining un-assigned ion in each compartment is used to establish the background for that compartment and serves as the standard to which assigned b-/y- series ions from the same compartment are compared.

The b-/y- series fragment ions in peptide spectra are expected to have higher intensities than remaining un-assignable ions. If the background signal in a compartment is much larger than the scored ions, it is an indication that other assignments may be possible and the scores for ions in that compartment are decreased proportionately. The background is determined separately for each compartment to adjust for the fact that fragment ion intensity in typical CID spectra varies depending on where the fragment

ions are located on the *m/z* axis. Fragment ions near the middle of the b- or y- series tend to be much more intense than fragment ions near the low or high *m/z* extremes of spectra. By scoring the ratio of detected fragment ions to regional background signals, P-Mod gives approximately equal weight to all detected ions.

The mass accuracy of detected fragment ions is modeled through incorporation of term d in Equation 4-2. Term d represents the difference between the observed and expected *m/z* of scored ions. Scores for individual fragment ions are maximized when d=0. For this reason, spectra obtained on high mass accuracy instruments are expected to produce higher sequence-to-spectrum comparison scores for correct matches. The increased scores should increase the statistical significance of correct matches, improving both the sensitivity and specificity of P-Mod assignments.

Although each element in each peptide search array is scored separately, only the highest scoring element is recorded as a potential match. As described previously, each element in a given peptide search array represents the mass shift located at a different position in the peptide sequence. The element receiving the highest score provides the best fit to the data and reflects the position in the sequence where the observed mass shift is most likely located. Final determination of which sequence produces the best match to a given spectrum is accomplished by comparison of the putative match scores of all the sequences in the search list. Only one sequence from the search list is ultimately considered to be a potential match for each spectrum and is selected because it produces the best match score.

P Value Estimation

The fact that only the highest scoring element from the search array is recorded means that the raw scores assigned to individual sequence-to-spectrum comparisons are extreme values (110). Moreover, the sequence from the search list that provides the best match to a given spectrum is the sequence that produces the most extreme score. The scores assigned to putative matches are outliers and are much higher than the mean or median scores for random sequence-to-spectrum comparisons. To model the distribution of scores assigned to sequence matches, raw scores from equation 2 were fit to a series of extreme value distributions (Equations 4-3 and 4-4), where Y = the extreme value reduced variate, S = raw score, μ = a conditional location parameter, α = a conditional scale parameter, k = the number of comparisons and p = the estimated p value (110).

Equation 4-3. $Y = (S - \mu) / \alpha - \ln(k/100)$

Equation 4-4. $p = 1 - \exp(-\exp^{(-Y)})$

Initial experimentation with P-Mod indicated that raw scores were contingent upon the peptide sequence length, or more precisely, the number of applied search criteria. Shorter peptide sequences generate fewer search criteria, making it easier to achieve a higher ratio of detected ions relative to comparisons with longer sequences. It was not possible to correct for these observed scoring differences though simple linear regression. This is because linear regression methods compare differences between the means of different categories or groups. Although there are differences in the mean

scores for sequences of different lengths, these differences are much greater when one compares the extreme scores of matched sequences. Thus, instead of calibrating the raw scores and fitting them to a single extreme value distribution, scores were fit to multiple extreme value distributions conditioned on the number of applied search criteria.

Extreme value distributions, like many other statistical distributions, are defined by both location (μ) and scale parameters (α), which are similar to the familiar mean and standard deviation parameters from the normal distribution. For P-Mod, these parameters have been estimated conditionally based on the number of applied search criteria. For each potential number of applied search criteria, we have generated unique estimates for μ and α. Following each sequence-to-spectrum comparison, raw scores are converted to an extreme value reduced variate by Equation 4-3 using the conditional parameter estimates. This reduced variate, which is analogous to the z-statistic from the normal distribution, is a measure of how extreme a score is. For the initial estimates of the reduced variate, it is assumed that each sequence comparison score is the extreme score derived from 100 comparisons to the spectrum. After all sequence comparisons have been completed, the reduced variate estimates are sorted and the sequence with the largest reduced variate is identified. The sequence producing the largest reduced variate may not have the largest raw score, but it is the greatest statistical outlier, thus indicating that it is the least likely to be a product of a randomly matched sequence-to-spectrum comparison. The reduced variate of the matched sequence is rescaled to reflect the actual number of comparisons made to the given spectrum, and then inserted into Equation 4-4 to generate an estimated p value. This p value indicates the probability that a spectrum

matched to a random sequence of a given length would produce a score as large as the score observed.

It is important to emphasize that the p value estimates generated from Equations 4-3 and 4-4 are scaled to the number of comparisons. Greater numbers of random sequence to spectrum comparisons increase the resulting raw match score. Unless scores are adjusted for the number of comparisons, random matches derived from large peptide sequence search lists would receive higher p values than matches derived from smaller search lists. Indeed, if the search list were large enough, nearly every spectrum would be matched to one of the search sequences at a statistically significant p value. For this reason, the reduced variate is scaled to account for the number of measurements (110). This is critical to the accuracy of P-Mod, because the number of sequences in the search list can vary from experiment to experiment and because nearly every sequence in the search list is considered as a potential match.

P-Mod Program and Graphical User Interface

P-Mod is written in C++ for Microsoft Windows. C++ was chosen because the algorithm is computationally intensive. The user interface for P-Mod consists of a search generation window, a results window and spectrum display window. The search generation window aids the user in creating a list of peptides to search by providing *in silico* digestion by various enzymes with or without missed cleavages and by allowing the user to specify fixed amino acid modifications (e.g., carboxamidomethylation of Cys residues), which then will not be mapped as mass shifts by P-Mod. The results window

shows search results with color-coded highlighting of mass shifts on the indicated amino acid residue. The results window also allows for sorting by the various columns and exporting of the results to external .CSV files. The spectrum display window shows the selected MS-MS spectrum, highlights the matched b and y ions in color and displays a list of all b- and y-ions matched by P-Mod. For detailed inspection of the spectrum, the user can expand selected sections of the *m/z* axis.

Experimental Procedures

Generation of Reference Spectra

Extreme value distribution parameters were estimated through a series of simulations in which MS-MS spectra were scored against random peptide sequences. The MS-MS spectra were obtained by LC-MS-MS analyses of tryptic digests from eight reference protein samples. The reference samples consisted of bovine proteins (muscle actin, serum albumin, histones, ubiquitin, and α-casein) as well as human hemoglobin and prothrombin, and E-coli β-galactosidase, all of which were obtained from Sigma. The proteins were each dissolved in 8 M urea, 400 mM NH_4HCO_3, 4 mM tris(carboxyethyl)phosphine at a final protein concentration of 5 μg/μl. Samples were heated at 50°C for 45 min, and then diluted 1:4 with deionized water prior to digestion. Sequencing grade modified porcine trypsin (Promega) was then added to each sample at an enzyme:protein ratio of 1:50 (w/w). Digestion was carried out overnight at 37°C.

Peptide digests were analyzed by LC-MS-MS on a ThermoFinnigan LCQ instrument equipped with a standard ThermoFinnigan ESI source (Thermo Electron, San Jose, CA). Chromatography was carried out on a 5-μm Vydac 218TP column (1.0 x 250 mm) using a $H_2O/CH_3CN/0.01\%CF_3COOH$ gradient elution at a flow rate of 0.03 ml/min. Data-dependent scanning used repeated cycles of a single MS scan followed by three successive MS-MS scans of selected precursor ions with a *m/z* greater than 300. MS-MS spectra were derived from averaging the three MS-MS scans. MS-MS spectra with fewer than 50 fragment ions were not assigned P-Mod scores or considered in the statistical analyses. A total of 2404 MS-MS spectra were obtained from the peptide

digests. These spectra served as the reference dataset in subsequent simulations that were used to establish the statistical parameters of the P-Mod algorithm.

Simulated Comparisons With Random Peptide Sequences

All spectra in the reference dataset were scored against random peptide sequences between 6 and 25 amino acids in length. Random sequences were generated using the natural occurrence rates of the 20 common amino acids in mammalian proteins (111) with the exception that seven out of eight generated sequences were tryptic peptides ending in either lysine (52.5%) or arginine (35%). Each reference spectrum was scored by two distinct and unique random sequence lists. The first list was composed of 5000 random sequences with expected masses within ± 1 amu of the precursor mass. The second list consisted of 10^5 random sequences with expected masses that differed from the precursor mass by more than 2 amu. New random sequence lists were generated for each spectrum to account for differences in precursor mass and to minimize serial correlation of results.

All sequence-to-spectrum comparisons were assigned raw scores according to Equation 4-2. The resulting scores were divided into either mass-shifted or non mass-shifted groups and further categorized according to the number search criteria applied in each case. Scores were recorded for comparisons utilizing between 7 and 36 applied search criteria. Wherever possible, exactly 100 scores were recorded for all applicable search criteria categories. Due to the restriction against large negative mass shifts (see above), comparisons between spectra derived from low mass precursor ions and high

mass sequences were not considered. Thus, scores for higher applied search criteria categories were not recorded for all spectra in the reference data base. In cases were more than 100 comparison scores were generated in a particular category, the list of scores was truncated to include only the first 100 comparison scores.

From these preliminary sets of 100 scores, the most extreme scores were identified and saved for further analysis. At this stage, a population of extreme scores for each applied search criteria category had been collected for both mass shifted and non-shifted sequence-to-spectrum comparisons. Each extreme score was taken from 100 sequence-to-spectrum comparisons with the same number of applied search criteria. The extreme scores in each category were fit to a series of extreme value distributions by maximum-likelihood estimation in SAS. The parameter estimates from this analysis are portrayed in Figure 4-1. These parameters, in turn, provide the basis for estimating p values for each comparison score based on the mass shift and the number of applied search criteria.

Analysis of BSA Peptides

BSA (20 μg)was dissolved in 100 μL 0.1 M ammonium bicarbonate containing 4 mM tris(carboxyethyl)phosphine and 10 mM DTT and incubated at 50ºC for 15 min. Iodoacetamide was added to a final concentration of 20 mM for 15 min to convert thiols to carboxamidomethyl derivatives. Modified porcine sequencing grade trypsin then was added in a 1:50 protein:trypsin ratio and the samples were incubated at 37ºC for 18-24 h.

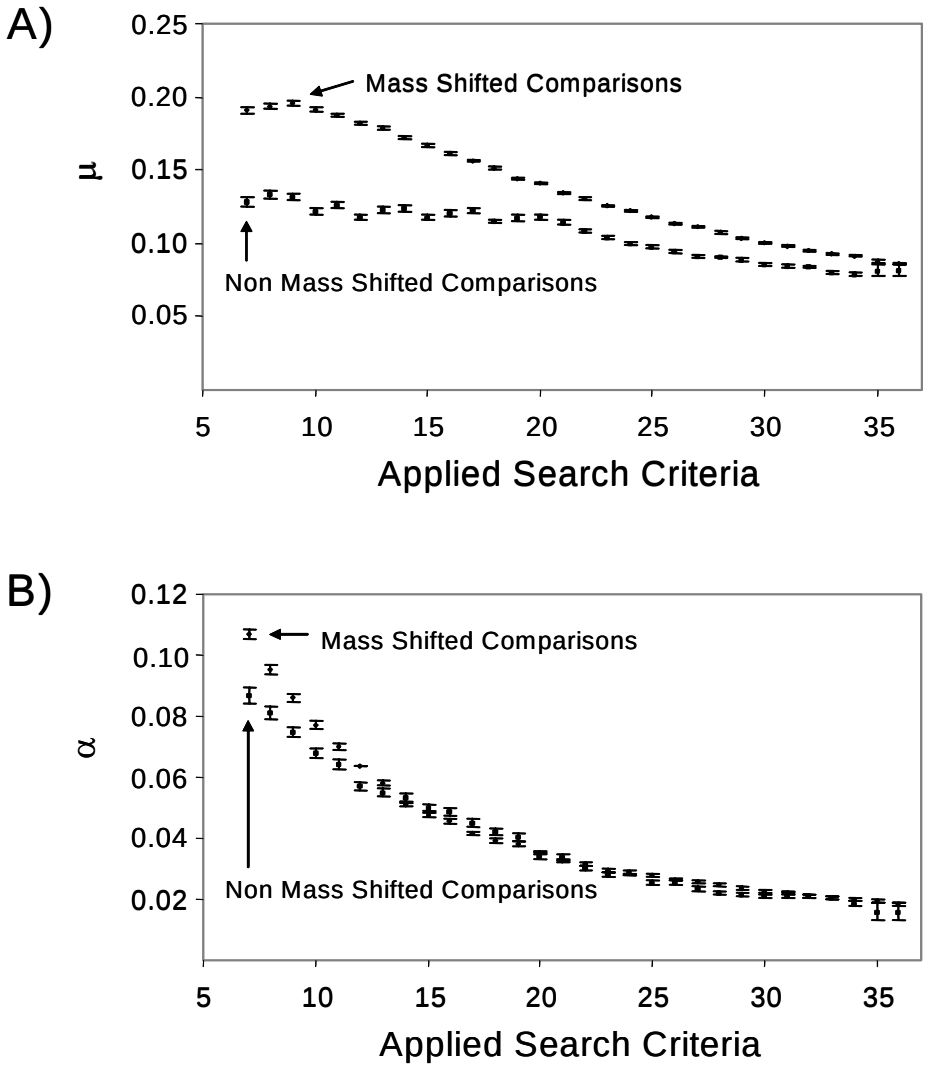


Figure 4-1. Conditional extreme value parameter estimates for μ (A) and α (B). 2404 CID spectra obtained in LC-MS-MS experiments with peptide digests of 8 test proteins were scored by lists of 100 random peptide sequences, each of which were either mass-shifted or not, and each with the same number of applied search criteria. For each applied search criteria category and mass-shift state, the extreme scores observed from the 100 random sequence comparisons to each of the test spectra were fit to extreme value distributions by maximum-likelihood estimation in SAS.

A the end of digestion, the digestion mixture was acidified with formic acid and analyzed by LC-MS-MS on a ThermoFinnigan LTQ linear ion trap MS instrument equipped with a ThermoFinnigan Surveyor LC system and microelectrospray source and operated with Xcalibur 1.4 and Bioworks 3.1 software (Thermo Electron, San Jose, CA). LC-MS-MS analyses were done by reverse phase chromatography on a 11 cm fused silica capillary column (100 μm ID) packed with Monitor C-18 (5 μm) (Column Engineering, Ontario, CA) and eluted first with water:acetonitrile:formic acid (98:2:0.1, v/v/v) for 5 min. A linear gradient then increased acetonitrile to 60% by 45 min, to 80% by 47 min and then held at this solvent composition for another 18 min. MS-MS spectra were acquired in data-dependent scanning mode with one full scan followed by one MS-MS scan on the most intense precursor with dynamic exclusion of previously selected precursors for a period of 3 min.

Results

Validation of P-Mod Statistical Estimates

The statistical estimates of P-Mod were validated through a series of additional simulations comparing lists of peptide sequences to the MS-MS spectra in the reference dataset. In the first set of simulations, the reference spectra were scored against lists of either 50, 100, or 200 random sequences of between 6 and 25 amino acid residues in length. The random peptide sequences for these simulations were generated in a similar manner to that described above; however, no consideration was made about the mass shift or the number of applied search criteria. The mixed composition of peptide sequences in these lists was analogous to what one could expect from a tryptic digest of a hypothetical protein. Any matches to sequences from these lists are completely artifactual and for the most part are expected to have high p values. However, since the p values are an estimate of the false positive rate, one should expect that approximately 10% of the random matches should have p values less than 0.1 and that 1% of the matches should have p values less than 0.01.

The MS-MS spectra in the reference database were scored one at a time by the random sequences in the three lists. For each list, the individual sequences were scored and assigned p values using Equations 4-1 – 4-4. The sequences then were ranked by p value, and the sequence with the smallest (i.e., most significant) p value was retained as a hypothetical match. The estimated p values for the resulting random matches then were plotted vs. their observed frequency (Figure 4-2). The data for all three sets of

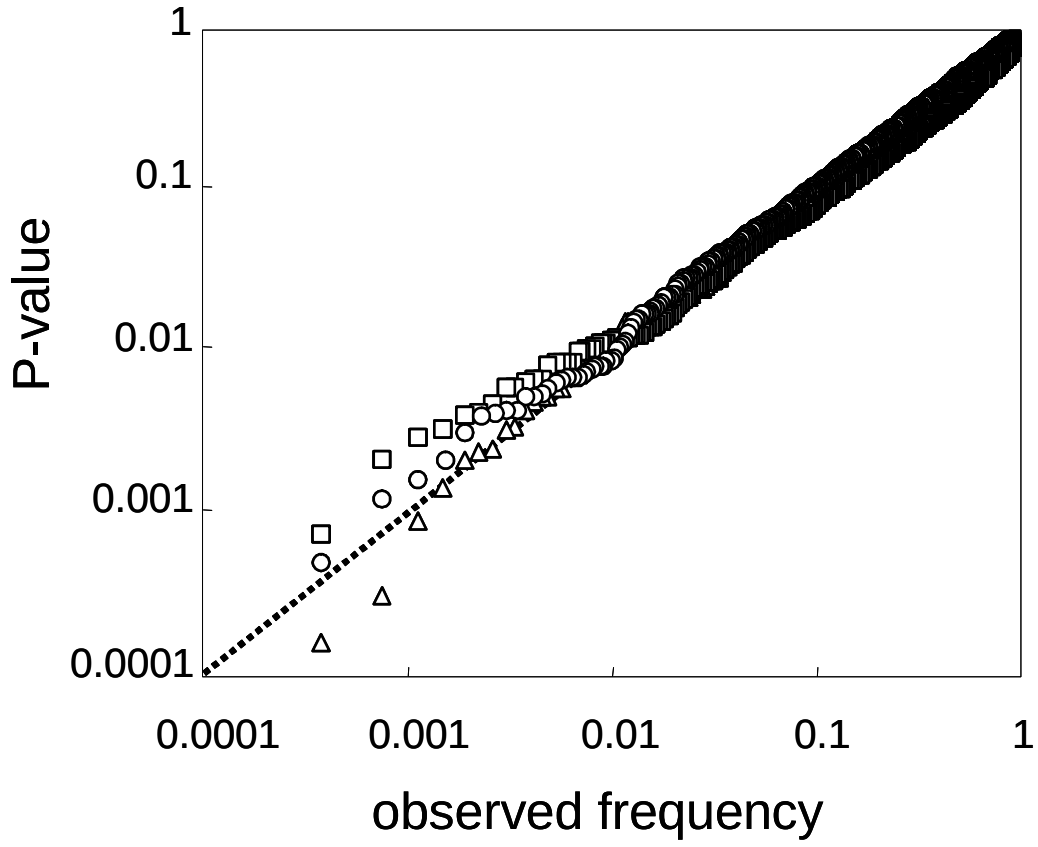


Figure 4-2. P-value estimates for extreme scores, or matches, resulting from comparisons to search lists containing different numbers of random peptide sequences. The test spectra were rescored by lists of either 50 ∆, 100 ○, or 200 □ peptide sequences of random length and composition. For each comparison, raw scores were converted to p-values using equations 4-3 and 4-4 and the conditional extreme value parameters illustrated in Figure 4-1. For each spectrum, the comparison with the lowest p-value was considered as a prospective match. The resulting p-values were plotted versus their observed frequency. The dashed line indicates the expected p-value distribution for such random comparisons.

comparisons cluster around the dashed diagonal line, indicating that calculated p values are distributed in close agreement with expected false positive rates.

Figure 4-2 indicates that when comparisons are carried out with lists containing approximately 100 peptide sequences the p value estimates of assigned matches follow an expected distribution and are highly accurate. Search lists containing more than 100 peptide sequences result in matches with p value estimates that are slightly, but consistently, above the dashed line, indicating that the reported p values are somewhat conservative. Comparisons with smaller sequence lists result in less conservative estimates. However, when scored by a search list of 50 peptide sequences, only 2 out of the 2404 MS-MS spectra in the reference database were given p value estimates more significant than would be expected by chance. Thus, p value estimates for the P-Mod algorithm are highly accurate and to scale appropriately with the number of comparisons made to each spectrum.

Sensitivity of P-Mod Algorithm

Additional simulations were carried out to establish the sensitivity at which P-Mod is able to detect MS-MS spectra corresponding to either modified or unmodified peptide sequences. Sensitivity is a measure of the percentage of spectra that can be detected at different p value thresholds. Ideally, all spectra that should be matched to one of the sequences from the search list would be detected at highly significant p values. However, this is not always the case. A major impediment to detecting spectra of unmodified peptide sequences is noise in the MS-MS spectra. Spectra with weak or

incomplete diagnostic fragment ion series or those that contain a significant number of large unassignable fragment ions may result in matches that have numerically high p values. Likewise, spectra in which fragment ions are recorded with poor mass accuracy are likely to receive lower raw scores and numerically higher p values. Detection of modified peptide spectra is further complicated by two additional constraints. First, because random matches to mass shifted sequences have higher extreme scores (Figure 4-1), true matches to mass shifted sequences are expected to have less significant p values compared with matches to native sequences. Furthermore, because the raw scores in Equation 4-2 are dependent on the mass accuracy of the fragment ions, modified peptide spectra can receive reduced scores if there is considerable error in the precursor mass estimate. Error in the precursor mass estimate leads to an error in the mass shift estimate in Equation 4-1, which, in turn leads to errors in the assumed *m/z* of mass shifted fragment ions.

The sensitivity of the P-Mod algorithm was modeled through simulated analyses of validated MS-MS spectra from the reference dataset generated as described above. The reference dataset was screened to identify spectra that were positive matches to unmodified peptides from the known digested proteins. Criteria for this screening were that spectra 1) had calculated masses within 2 amu of the expected masses of matched sequences, 2) had at least 50 fragment ions, and 3) had p values less than 0.05 for P-Mod sequence to spectrum matches. The matched peptide sequences then were systematically mutated to introduce sequence polymorphisms. Two different sets of mutated sequences were generated. The first was composed of peptide sequences with a single

polymorphism, whereas the second contained sequences with two polymorphisms. Each peptide sequence was systematically mutated at every amino acid residue. Mutated amino acids were selected at random; the only requirement was that the mass difference between original and mutated residues was greater than 2 amu.

Figure 4-3A shows the results of comparisons between the mutated sequences and validated MS-MS spectra that had been corrected for errors in the precursor mass estimates. For mass accurate precursor ions, spectra of peptides containing a single modification were detected with sensitivity comparable to that of unmodified peptides. However, spectra matched to sequences with two nonconsecutive modifications received p values approximately three orders of magnitude less significant than did matches to either unmodified sequences or to sequences containing only a single modification. It is important to note that this loss of sensitivity was only for spectra containing multiple unpredicted modifications. If modifications are predicted at the outset of the experiment and incorporated into the peptide sequence search list, modified peptides can be detected with the same sensitivity as for unmodified peptides. The P-Mod search generation window allows the user to specify fixed modifications.

Figure 4-3B illustrates how error in the precursor mass estimate affects the ability of P-Mod to detect modified peptide spectra. The simulation was repeated with the validated peptide MS-MS spectra and the list of peptide sequences containing single polymorphisms. After correcting for errors in the precursor mass estimates, error was systematically reintroduced. The precursor mass error was set at either 0.0, 0.5, 1.0, or 2.0 amu. These experiments demonstrated that for every 0.5 amu error in the precursor

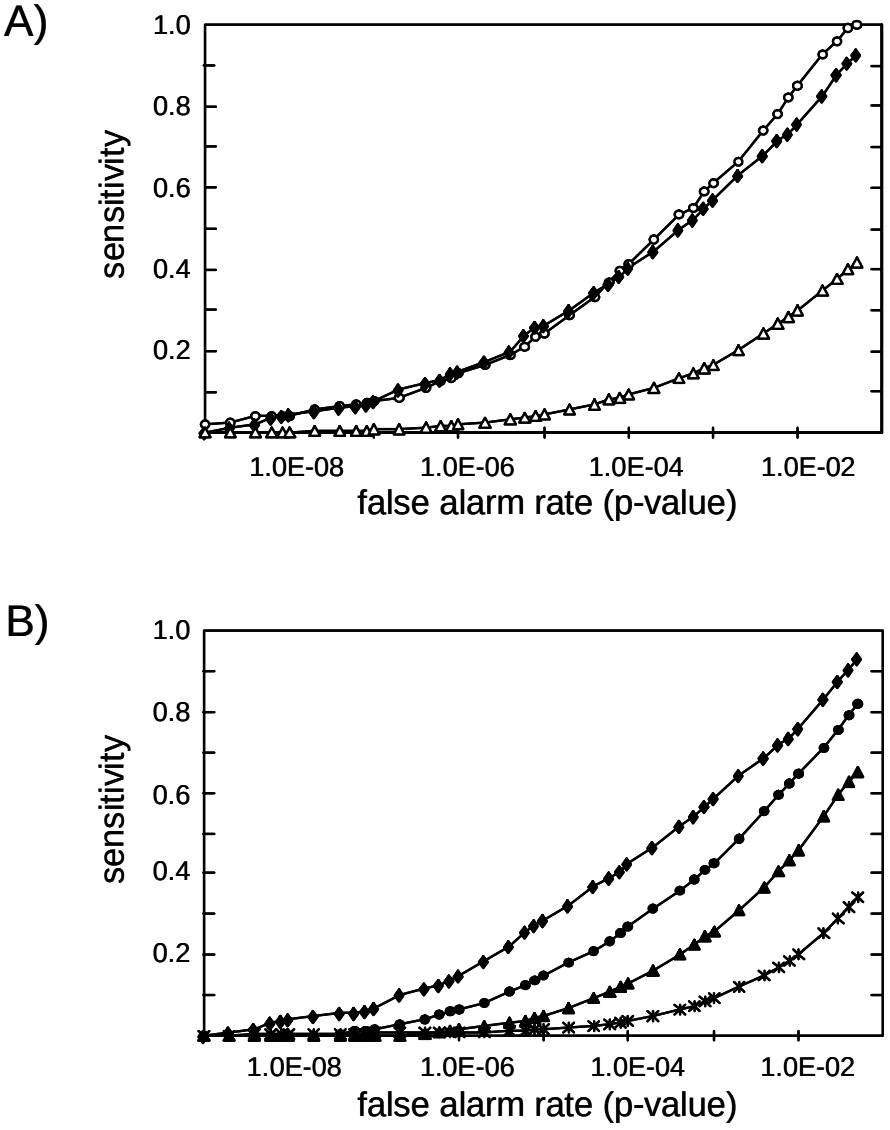


Figure 4-3. Receiver operator curves for the P-Mod algorithm. (A) Sensitivity versus false alarm rate for spectra with precise precursor mass estimates. Detection of unmodified peptide sequences ○ is compared to detection of sequences with single amino acid polymorphisms ♦ and sequences with two nonconsecutive polymorphisms Δ. (B) Effect of inaccurate precursor mass estimates on the detection of single amino acid polymorphisms. Shown is the comparable sensitivity when the error in precursor mass estimation is 0.0 amu ♦, 0.5 amu ●, 1.0 amu ▲, or 2.0 amu ×.

mass estimate there is approximately a one order of magnitude decrease in p value significance for matches to modified sequences.

Accuracy of P-Mod Localization of Modifications

In a separate but related issue, we wanted to document the dependability of P-Mod at localizing mass shifts to the correct position in the peptide sequence when matches to modified sequences are detected. The validated peptide spectra were once again analyzed against a list of peptide sequences containing single amino acid polymorphisms. As described above, the list of mutated peptide sequences contained multiple sequence variants for every primary sequence that had been matched to one of the validated spectra. Since each primary sequence was systematically mutated once at every amino acid residue, each sequence had a number of variants equal to the number of amino acids that it contained. Each validated spectrum was scored against all of the variant sequences corresponding to the matched primary sequence and the percentage of variant sequences for which the mass shift position was correctly assigned was recorded. For the initial set of comparisons, the precursor ions of the validated peptide spectra were mass corrected so that the error in the precursor ion mass estimate was 0.0 amu. Comparisons with the variant peptide sequences were then repeated at different error levels for the precursor mass estimate: 0.5, 1.0, 1.5, 2.0, and 2.5 amu.

The results from these simulations are summarized in the box plots presented in Figure 4-4. The box plots represent distributions of percentage scores for the validated peptide spectra under different precursor mass estimate error levels. In turn, the

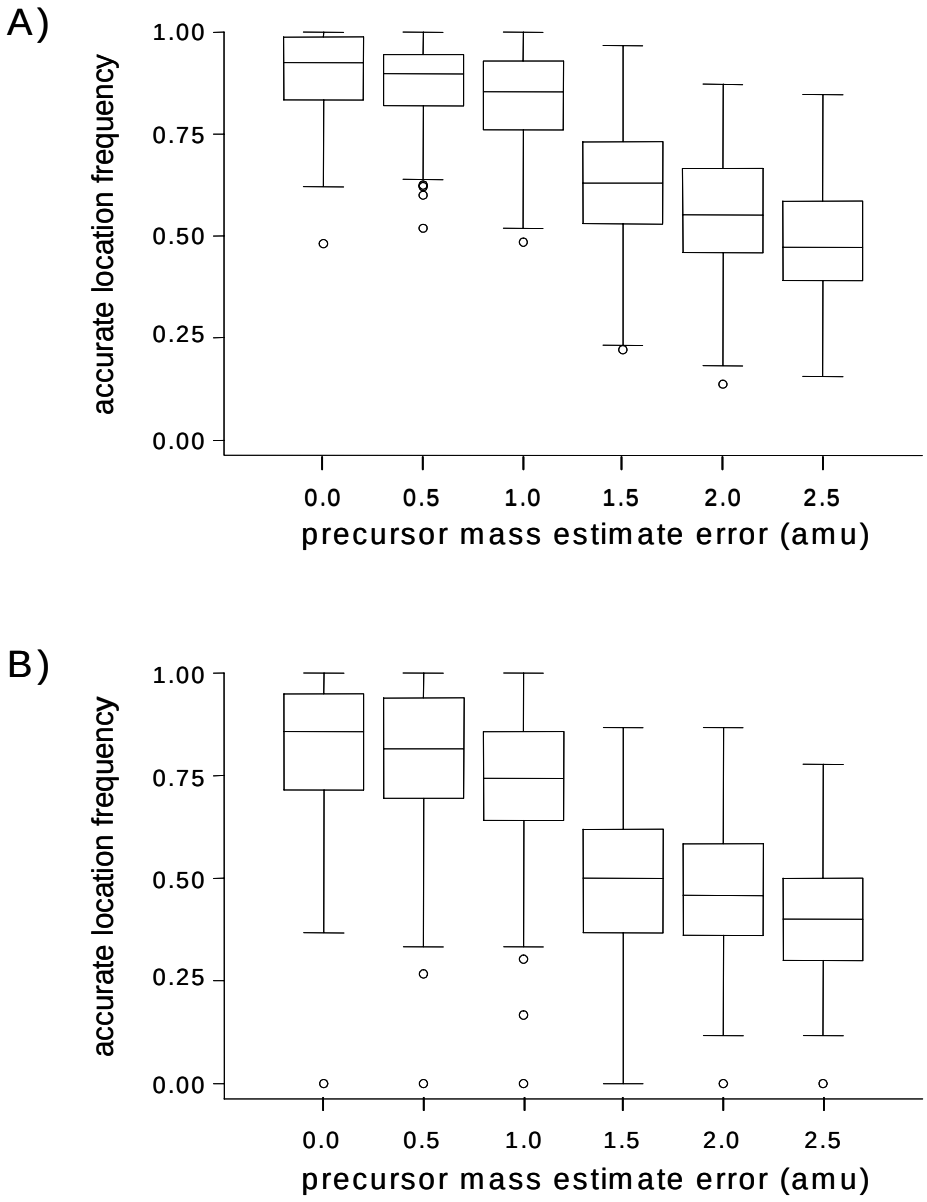


Figure 4-4. The frequency at which P-Mod assigns the correct sequence position to observed mass-shifts is a function of both precursor mass accuracy and spectrum quality. (A) Location accuracy for high quality spectra with p-values less than 0.001. (B) Location accuracy for low quality spectra with p-values greater than 0.001.

percentage scores reflect the proportion of variant sequences that, when compared to the appropriate validated spectrum, were assigned the correct mass shift position. The line in the center of each box plot indicates the median percentage score, while the upper and lower box edges indicate the quartile scores. The extended lines protruding from the boxes encompass the outlying percentage scores within twice the box height and the open circles beyond these lines are the point estimates of extreme percentage score outliers. Figure 4-4A summarizes the results from sequence comparisons to high quality peptide spectra which, when scored by matching unmutated sequences, were assigned p values less than 0.001. Figure 4-4B summarizes the results from lower quality peptide spectra for which the matches to unmutated sequences were assigned p values greater than 0.001. High and low quality spectra were evaluated separately because low quality spectra that receive p values greater than 0.001 have fewer and less intense b-/y- series fragment ions than do high quality spectra. The missing fragment ions in these spectra were expected to hinder accurate localization of peptide modifications.

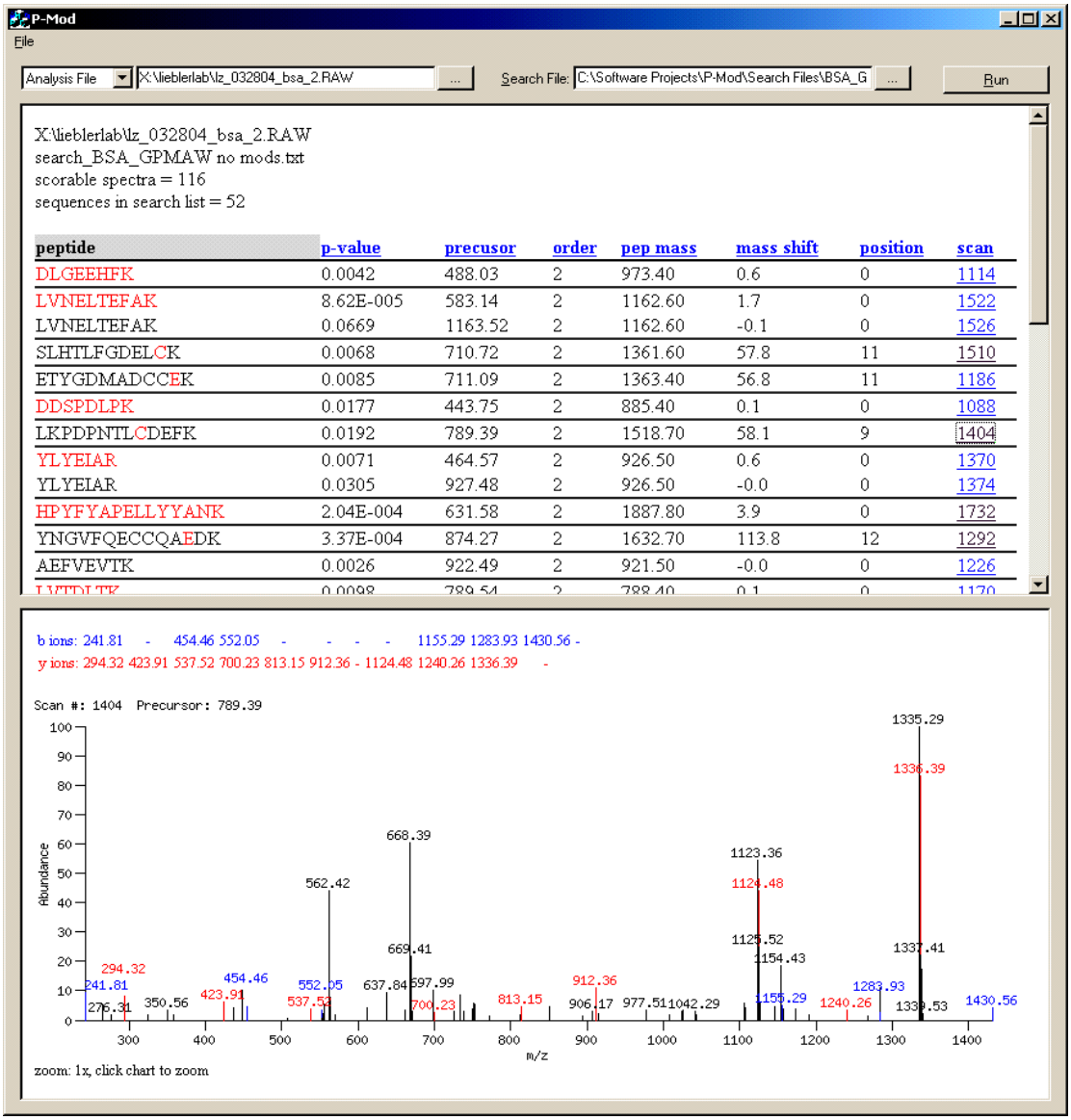
These simulations demonstrated that while modifications are localized with a high degree of accuracy in high quality spectra with match p values less than 0.001, precision is decreased for spectra with less significant match p values. At every precursor mass estimate error level, the median percentage of sequence variants for which the mass shift position was correctly assigned is approximately 10 percent higher for high quality spectra. Low quality spectra exhibit a greater degree of variation in the percentage of correct position assignments; the box plots for these spectra are broader and there are many more low percentage extremes. Error in the precursor mass estimate also decreases

the ability to precisely localize peptide modifications. Any error in the precursor mass estimate is carried over as an error in the estimated mass shift, which in turn interferes with the detection of b-/y- series fragment ions. P-Mod is relatively tolerant of precursor mass estimate errors less than 1.0 amu, however, larger errors substantially reduce the percentage of modifications assigned to correct positions.

What Figure 4-4 does not illustrate is that even when P-Mod assigns a mass shift to the wrong amino acid residue in a peptide sequence, the true mass shift is frequently at an adjacent residue. If the chemical nature of the modification can be reasonably inferred from the mass shift, the precise position can often be deduced from the known chemistries of the adjacent amino acids in the sequence. Localization errors are more frequent for modifications at amino acids near either end of sequences because the b-/y-series fragment ions for cleavages at these sites are often missing in MS-MS spectra. Modifications located near the center of peptide sequences tend to be assigned more reliably, since such assignments are corroborated by overlapping b-/y- series fragment ions.

Discovery of BSA Peptide Variants

We used P-Mod to analyze datafiles from 12 LC-MS-MS analyses of trypsin digests of BSA. After spectral workup as described above, each datafile yielded between 2,000 and 5,000 scorable spectra. A P-Mod search file containing tryptic peptides between 5-30 amino acids in length and with up to one missed cleavage was used to search all the datafiles. A typical search output screen is shown in Figure 4-5. P-Mod



peptide	p-value	precusor	order	pep mass	mass shift	position	scan
DLGEEHFK	0.0042	488.03	2	973.40	0.6	0	1114
LVNELTEFAK	8.62E-005	583.14	2	1162.60	1.7	0	1522
LVNELTEFAK	0.0669	1163.52	2	1162.60	-0.1	0	1526
SLHTLFGDELCK	0.0068	710.72	2	1361.60	57.8	11	1510
ETYGDMADCCEK	0.0085	711.09	2	1363.40	56.8	11	1186
DDSPDLPK	0.0177	443.75	2	885.40	0.1	0	1088
LKPDPNTLCDEFK	0.0192	789.39	2	1518.70	58.1	9	1404
YLYEIAR	0.0071	464.57	2	926.50	0.6	0	1370
YLYEIAR	0.0305	927.48	2	926.50	-0.0	0	1374
HPYFYAPELLYYANK	2.04E-004	631.58	2	1887.80	3.9	0	1732
YNGVFQECCQAEDK	3.37E-004	874.27	2	1632.70	113.8	12	1292
AEFVEVTK	0.0026	922.49	2	921.50	-0.0	0	1226

Figure 4-5. P-Mod search output screen. The upper panel lists the datafile searched, the search file, number of scorable spectra after preliminary workup and search sequences. The upper panel also lists the peptide search sequences, p-values for matches, precursor *m/z*, scan order, search peptide mass, mass shift, modification sequence position and scan number. The lower panel displays the selected scan, the b- and y-ions assigned by

P-Mod and the corresponding signals are coded blue, red and black for b-ions, y-ions and unassigned ions, respectively.

lists the search peptide, the p-value for the match, the precursor mass, the order of the scan (e.g., 2 for MS-MS, 3 for MS-MS-MS, etc.), the mass of the peptide search sequence, the calculated mass shift, sequence position and scan number. The selected scan in Figure 4-5 is scan 1404, which corresponds to the peptide LKPDPNTLCDEFK with a carboxamidomethylation mapped to the Cys at position 9. Results can be sorted by any of these output parameters by clicking on the column heading. The spectrum window displays the selected spectrum with a list of b- and y-ions that match the mass-shifted peptide sequence. Dashes indicate the positions of missing b- and y-ions.

The combined search outputs for LC-MS-MS of the 12 BSA digests yielded 1457 "hits" with p values <0.05 and mass shifts ranging from -85.5 to 878.1 Da. Sorting of the data by mass shift indicated groups of peptides containing characteristic, consistent mass shifts at specific amino acids. A number of such groupings are evident in the dataset; the most prominent being +57 shifts at Cys residues due to iodoacetamide-induced carboxamidomethylation during sample preparation. We also noticed a consistent -58 Da mass shift at position 11 of the peptide TVMENFVAFVDK, which corresponds formally to a loss of the carboxymethylene ($-CH_2CO_2H$) group from the Asp residue (Figure 4-6). Because this change does not appear at other Asp residues in any of the other analyzed samples, we conclude that the shift is due to a D → G substitution at position 579 in the BSA sequence. An examination of the UniProt website (http://www.pir.uniprot.org/) indicates that this BSA variant has not been reported previously.

Peptide	p-value	prec.m/z	pep.Mass	mass.Shift	position	scan
QEPER	0.0415	593.13	657.3	-65.2	4	5714
GVFRR	0.0395	576.27	633.4	-58.1	4	10448
TVMENFVAFVDK	4.64E-07	671.31	1398.7	-58.1	11	8263
TVMENFVAFVDK	3.26E-04	671.53	1398.7	-57.7	11	8922
TVMENFVAFVDK	2.03E-04	671.53	1398.7	-57.7	11	9684
TVMENFVAFVDK	8.04E-04	671.58	1398.7	-57.6	11	8074
TVMENFVAFVDK	0.0071	671.58	1398.7	-57.6	11	1788
TVMENFVAFVDK	0.0053	671.71	1398.7	-57.3	11	9074
TVMENFVAFVDK	0.0038	671.86	1398.7	-57	12	4124
QEPER	0.0212	609.26	657.3	-49	4	1329
QEPER	0.0358	609.31	657.3	-49	4	1457
CCTKPESER	0.0465	502.43	1051.4	-48.5	4	6904

Figure 4-6. Section of combined P-Mod search output for datafiles from 12 LC-MS-MS analyses of BSA tryptic digests. The combined outputs were sorted by mass shift and the highlighted section shows entries for the peptide TVMENFVAFVDK with a mass shift of -58 Da located at position 11, which corresponds to a D→G substitution at position 579 of BSA.

Discussion

The discovery of protein modifications, particularly unanticipated modifications is one of the biggest challenges in proteome analysis. Although Sequest and Mascot searches of MS-MS data with allowances for variable modifications can detect modified peptides (35-37), the mass and amino acid specificity must be known beforehand. Our introduction of the SALSA algorithm enabled the identification of MS-MS spectra corresponding to modified peptides through searches based on fragmentation characteristics of the modifying moieties or on sequence motifs within a target protein (103, 106). Nevertheless, this powerful approach to discovery of modifications and variants required time-consuming manual interpretation of MS-MS spectra detected by SALSA. The P-Mod algorithm and software we describe here automate the detection and sequence mapping of modifications from MS-MS spectra.

In comparing P-Mod to widely used protein identifications tools such as Sequest and Mascot, it is important to understand the differences and relationship between these tools and P-Mod. Sequest, Mascot and other similar tools search database sequences with uninterpreted MS-MS spectra to identify the corresponding proteins. P-Mod uses sequences to search datafiles of MS-MS spectra rather than databases. P-Mod is applied in situations where the identity and sequence of the target protein is known or has been established through prior data analysis (e.g., with Sequest). In contrast to Sequest and Mascot, P-Mod can identify unknown and unanticipated modifications by identifying MS-MS spectra corresponding to mass-shifted variants of search sequences. Thus, P-Mod is a unique resource that complements existing proteomics database search tools.

In contrast to SALSA and database search tools, P-Mod assigns p value estimates for all matched peptide sequences. The p values reflect the likelihood that the reported match for a given spectrum is the product of a random or otherwise erroneous comparison. Conducting analyses with a p value threshold of 0.01 provides sensitive detection of modified peptides and ensures that false positive matches are expected for no more than 1% of the analyzed spectra. This greatly reduces the effort required to confirm reported matches. The p values calculated by P-Mod are scaled to account for the number of sequences compared to each spectrum prior to assignment of a match. Larger numbers of sequence comparisons are likely to result in random matches with higher raw scores. However, because the p values are scaled to the number of comparisons, p value estimates remain valid for search lists containing different numbers of peptide sequences.

P-Mod is able to detect unexpected peptide modifications by considering nearly every peptide sequence in a restricted search list as a potential match to any given MS-MS spectrum. The program estimates a mass shift for every sequence-to-spectrum comparison for which there is a difference between the peptide search sequence mass and the observed precursor ion mass. The expected MS-MS fragmentation pattern of the peptide is then adjusted to account for the magnitude of the mass shift as well as potential locations for the mass shift within the peptide sequence. In this regard, the P-Mod mass shift is analogous to the spectral alignment approach described for SHERENGA (39). The spectral alignment function described for SHERENGA is more complex than the P-Mod mass shift and is able to adjust for the possibility that the observed mass difference for a sequence-to-spectrum comparison is due to multiple modifications. However,

dividing the mass difference into multiple mass shifts greatly expands the number of possible permutations, increases analysis times and decreases sensitivity. Because multiple unanticipated modifications to a single peptide are probably uncommon, it is usually sufficient that P-Mod is able to identify peptide spectra that have single unanticipated modifications. In cases where multiple modifications may occur, the anticipated modifications can be incorporated into the sequences in the search list.

Recent developments in mass analyzer technology, including Q-q-TOF, TOF-TOF, FT-ICR and new triple quadrupole designs provide higher resolution and mass accuracy for both precursor and product ions in MS-MS spectra. More accurate mass measurements are particularly useful for discriminating between modifications or combinations of modifications. Most widely used proteomics programs were originally developed for use with lower resolution ion trap and triple quadrupole instrument data and do not take advantage of higher mass accuracy. Algorithms such as Sequest, Mascot and SALSA detect spectral features that lie within user-designated tolerances of expected values. Even though P-Mod was developed and validated on a lower resolution ion trap instrument, P-Mod scores incorporate the mass accuracy of MS-MS fragment ions. The data in Figures 4-3 and 4-4 indicate that increasingly accurate measurement of precursor *m/z* improves the sensitivity and positional accuracy for detection of peptide modifications by P-Mod.

P-Mod is, to our knowledge, the first proteomics software program to utilize extreme value theory to assign p values to sequence-to-spectrum matches. However, there is a precedent for the use of extreme value theory in the field of genomics where it

provides the basis for p value estimations in the widely used BLAST and FASTA algorithms (112, 113). Sequence matches in these algorithms are extreme values because they are best fits resulting from multiple sequence alignments. The same logic applies to P-Mod raw scores. At the level of individual sequence-to-spectrum comparisons, the preliminary match score is the best fit resulting from multiple configurations of the observed mass shift. Furthermore, the preliminary match scores for all of the sequences in the search list are ranked by significance to determine which sequence is ultimately the best match. Assigned sequence match scores are therefore extremes taken from a set of extreme values. This same line of reasoning could be carried over to other proteomics programs. As discussed above, spectral alignment in the SHERENGA algorithm is similar to the P-Mod mass shift. Thus, individual sequence-to-spectrum comparisons in SHERENGA are extreme values. Moreover, nearly all proteomics programs assign ultimate sequence matches by ranking preliminary match scores of multiple potential sequence matches. In this context, all assigned sequence matches are extreme values.

Finally, it is important to emphasize that the p values for assigned matches reflect the probability that by chance an individual MS-MS spectrum would receive a sequence match score of the observed magnitude. These p values are estimates for the false positive rates for individual spectra matched to a single sequence from a list of potential sequences. However, LC-MS-MS data files typically contain hundreds to thousands of MS-MS spectra. The likelihood of observing a false positive over the course of analyzing all of these spectra is approximately equal to the number of spectra times the p value threshold. For example, if the p value threshold is set at 0.01, then 1% of the

reported matches are potentially false positives. For a data file containing 500 MS-MS spectra one should expect that approximately 5 spectra may be matched to one of the search list sequences just by chance. At a p value threshold of 0.001 the number of expected false positives in this same data file is 0.5. That is, there is only a 50% likelihood that a single random match would be reported out of the 500 analyzed spectra. Similarly, if the p value threshold is decreased to 0.0001, the likelihood of even a single random match is only 5%.

We have found that detection of variant and modified peptides generally requires multiple LC-MS-MS analyses of the same or replicate samples. This reflects the low abundance of many modified peptide forms. Identification of all but the most abundant modifications with P-Mod is best done by combining the results of P-Mod analysis of multiple datasets, as we did with the 12 LC-MS-MS datasets the generated for BSA digests. Sorting the combined P-Mod outputs by mass shift allows infrequently detected modifications to be identified and this approach identified a previously unreported BSA D579→G variant (Figure 4-6). The substitution arises from a GAC to GGC substitution at position 1768 in the BSA coding sequence. Inspection of the data in Supplemental Table 4-1 reveals several other groupings of mass shifts that repeatedly appear in specific peptides or at certain residues, yet do not correspond to any known posttranslational modifications. These include an apparent M + 40 modification at Cys, which may correspond to an elimination of 17 Da from the carboxamidomethylated derivative. The nature of this modification remains to be established, as do those of other mass shifts

detected in these analyses. Nevertheless, this underscores the utility of P-Mod for discovery of unexpected protein posttranslational modifications.

In addition to endogenous regulatory modifications, there are many reactive chemical species formed from xenobiotics and endogenous oxidative processes that contribute to protein modifications (38, 67, 114, 115). Typical of these are reactive α,β-unsaturated carbonyl products of lipid oxidation, which react with Cys, Lys, His, Arg residues and N-terminal amines to form multiple products (67, 116, 117). The diversity of potential targets and adduct masses would make it nearly impossible to detect the majority of these adducts without employing an open-ended discovery strategy. Nevertheless, some of these modified proteins may serve as biomarkers for disease processes or chemical exposures. It is for this purpose that we developed P-Mod and we have begun to apply LC-MS-MS and P-Mod to the discovery of modified protein forms as biomarkers of disease.

CHAPTER FIVE – Suggested Improvements for SALSA and P-Mod and the Future of Protein Modification Analysis

Past research on protein modifications has relied on the use of gel electrophoresis and specific antibodies. This approach has provided important insights, but is limited in important respects. Such research is inherently biased towards the analysis of high abundance proteins and reliant on the availability of highly specific antibodies. For many types of protein modifications, such as those caused by adduction from electrophilic xenobiotics, suitable antibodies do not exist and can be difficult and expensive to generate. Even when bands on a gel are visualized by the binding of an antibody, extensive additional research is required to conclusively verify the identity of the protein bound by the antibody. Using conventional methods, it remains extremely difficult to conclusively verify the location, and for some adducts the chemical composition, of protein modifications. Without this information it is impossible to predict changes in protein structure and activity.

The emerging field of proteomics, driven by advances in mass spectrometry, promises to provide a more detailed look at the proteome than ever before. Through the LC-MS-MS analysis of peptides generated by enzymatic protein digestion, it is possible to conclusively identify large numbers of proteins in complex mixtures using a limited number of experiments. The wealth of information contained in peptide CID spectra produced in these experiments can potentially be used to both establish the chemical composition and location of protein modifications. However, the complexity and sheer

volume of data produced renders manual interpretation impossible. The use of computer algorithms to match peptide CID spectra to database sequences is absolutely required and already well established. Unfortunately, previous proteomics algorithms have limited utility when applied to the study of protein modifications.

The algorithms that are presented in this dissertation significantly enhance our ability to study protein modifications. Both programs are substantially different from previously developed proteomics algorithms. Table 5-1 compares the two programs to each other as well as to the commonly used Sequest algorithm. SALSA is unique in that it is the only program that allows users to screen MS-MS spectra datasets for specific fragmentation patterns. All of the other programs used in proteomics operate in exactly the opposite fashion by attempting to match individual spectra to database sequences. While other programs are capable of matching unmodified peptides or peptides with a limited number of expected modifications, SALSA excels at identifying unexpected modifications. In cases where modifications disrupt the normal pattern of peptide CID, as we observed with the pyrrole adducts, database matching programs fail to match the resulting peptide spectra. SALSA is the only program capable of detecting such modifications. Likewise, P-Mod is also unique and well suited to the study of protein modifications. Not only can P-Mod detect most modifications, even unexpected ones, it can also assign sequence location to reported modifications. The assigned sequence locations assist the user in identifying and determining the authenticity of reported modifications. The statistical model used in P-Mod represents the first example of the application of extreme value statistics to proteomics. The majority of previous

Table 5-1. Comparison of proteomics algorithms

	Sequest	**P-Mod**	**SALSA**
Application	Screen LC-MS-MS data from protein digests to identify protein components	Re-screen data files to confirm presence of peptides from specific protein(s), and identify spectra corresponding to modified variants of targeted sequences	Exhaustive searches of data files, looking for specific peptides and/or peptide modifications
Strengths	♦ Easy to use ♦ Sensitive identification of protein components in complex mixtures ♦ Searches involve all potential peptide sequences in corresponding databases ♦ Able to identify limited number and type of specified peptide modifications	♦ Easy to use ♦ Rapid search for up to several hundred peptide sequences ♦ Accurate p-values for sequence-to-spectrum matches ♦ Small and highly meaningful output files ♦ Able to identify most peptide modifications including those that are 'unpredicted' ♦ Sequence interpretation of identified modifications	♦ Highly flexible ♦ Exhaustive searches for specified targets ♦ Only program able to identify spectra corresponding to peptide modifications that disrupt normal MS-MS fragmentation patterns
Weaknesses	♦ Unable to identify spectra from peptides with 'unexpected' modifications or modifications that disrupt normal MS-MS fragmentation patterns ♦ Program is error prone and the statistical reliability of individual sequence-to-spectrum matches is unknown ♦ Large output files	♦ Unable to identify spectra from peptides with modifications that disrupt normal MS-MS fragmentation patterns ♦ The identities of protein components must be known or guessed prior to analysis	♦ Requires skilled users ♦ Searches can be cumbersome and time consuming to set up ♦ Large output files ♦ All reported matches require manual verification ♦ Scoring is not statistically based

proteomics algorithms were not statistically based. Those that were relied upon the raw score distributions of sequence-to-spectrum matches. Such models do not take into consideration the number of comparisons made. Large numbers of comparisons resulting from database searches, especially when such searches include possible protein modifications, are likely to result in random matches with large enough raw scores so as to be considered to be statistical outliers. P-Mod is the only proteomics algorithm that accurately scales statistical estimates to the number of computed sequence-to-spectrum comparisons. The accurate statistical estimates produced by P-Mod are essential to automating the search for protein modifications. Despite the strengths of these two programs, as is the case with most dissertation research, there remains room for improvement.

SALSA Improvements

While the SALSA algorithm is highly flexible and adaptable to other research contexts outside of the field of proteomics, it is handicapped by its dependence on user experience and expertise. For the effective use of SALSA, the user must have detailed expectations for the MS-MS fragmentation characteristics of target peptides and must be able determine which of these characteristics to use in constructing a SALSA search protocol. In addition to the skill requirement, SALSA is limited by the amount of time it takes to develop and enter search criteria. Currently, in order to conduct a search for a specific peptide along with all of its variants, the user has to enter several ion series and product ions into the search parameters, which can take a minute or two. If the user

wishes to conduct searches for a large number of individual peptides, the time it takes to construct search criteria can become prohibitive.

SALSA would benefit from further research to develop pre-constructed search modules for peptides and peptide modifications, as well as certain classes of small molecules. It would be a relatively simple matter to construct a front end program that would automate the generation of search criteria for peptides and common peptide modifications such as phosphorylation, acetylation, or glycosylation. By ensuring that the search criteria are accessible for editing, the user would be able to more rapidly generate custom search criteria for other, less common, modifications.

The productivity of SALSA would also be enhanced by making it possible to set up multiple searches at the outset of the analysis. For example, one could search for all of the peptides of a specific protein, along with potential modifications, simply by sending the program a list of expected peptide sequences. Or once a user had developed several custom search modules, it would be a simple matter to apply all of the search modules sequentially to each data file. In addition to expanding the capacity to conduct multiple searches for every data file, it would be beneficial to allow the user to queue up multiple data files. These program modifications would make it possible for a user to carry out a large number of SALSA searches in a day without constantly monitoring and interacting with the program.

However, if the above suggestions are incorporated into SALSA, many users will soon find themselves swamped with SALSA outputs files. Simply conducting a SALSA search does not ensure that a user will find an MS-MS spectrum of the target peptide or

peptide modification. In fact, the program will provide a list of potential matches for every attempted search, regardless of whether or not a correct match exists in the LC-MS-MS data file. Because SALSA scores are not statistically validated and because matched spectra are not interpreted by the program, it is up to the user to manually validate all potential matches. Fortunately, the user does not have to wade though the entire SALSA output file to ensure that all potential matches are evaluated. SALSA ranks outputs by raw scores and, even though the scores are not statistically based, correctly-matched spectra usually have higher scores compared to randomly matched spectra.

The observation that correctly matched spectra usually have raw scores that rank near the top of SALSA output files suggests several strategies for decreasing the size of SALSA outputs. The simplest means for decreasing the size of SALSA outputs is to add a parameter to the graphical user interface that lets the user decide on the maximum number of reported matches for each search. For example, if the user decided to limit the output to 5 potential matches, the SALSA output would only contain 5 matched spectra for each search. Such a truncated list would obviously contain fewer false positive matches, but there is a danger if the list were made too small that legitimate matches could be inadvertently eliminated.

A less arbitrary approach would be to fit all of the raw scores in the SALSA output to a statistical distribution, such as the lognormal distribution, and only report spectra with scores that are more than two standard deviations from the mean. It is important to note that such an evaluation will not provide a statistical measure of the

quality of putative matches; it will only help to reduce the size of the SALSA output files. It will still be up to the user to evaluate the nature and quality of all reported matches. Ideally, an evaluation of the distribution of SALSA raw scores would lead to statistical measures that would accurately reflect the likelihood that reported matches are false positives. This does not appear to be possible given the current configuration of SALSA. SALSA searches using different sets of search criteria have wildly different raw score distributions. In order to generate statistics that measure false positive rates for SALSA searches a database of raw SALSA scores would need to be generated for each set of SALSA search criteria.

P-Mod Improvements

Like SALSA, there is also considerable room for improving upon and enhancing P-Mod. The most obvious and most rapidly implemental P-Mod improvements have to do with the algorithm's scoring and statistics. It should be possible to enhance the sensitivity and specificity of the algorithm by incorporating more parameters as applied search criteria. Currently, only singly charged b-/y- ions are scored by the algorithm. While multiply charged b-/y- ions and water loss ions are assigned for the purpose of background determination, these ions are not scored. Scoring these ions and assigning other common peptide CID fragment ions, such as x-/z-/a-/c- ions, would likely enhance the observable difference in raw scores when comparing correct and false positive matches. The greater the magnitude of difference in raw scores, the greater the statistical significance for correct matches. By assigning greater statistical significance to correct

matches, the sensitivity of P-Mod would be improved; that is, fewer correct matches would be rejected at a given p-value threshold. Specificity would also be improved as the improved sensitivity would allow the user to apply more a more stringent p-value cutoff thereby decreasing the number of false positive matches. In addition, applying more search criteria would likely decrease the variability of false positive match scores (see Figure 4-1B) and consequently decrease the occurrence of sporadic scores with greater than expected statistical significance (Figure 4-2).

P-Mod is a computationally intensive program and fairly slow. Changing the scoring routine so that more ions are assigned and scored will likely slow the program even more. It may be possible to speed the program up by converting some of the code relating to the scoring routine to another computer language such as Array Basic. The reason that the current calculations are so slow is due to the time it takes for the program to work through a series of do-loops in order to assign all targeted ions in a spectrum. Such do-loops are necessary in programming languages such as C++ or java. However, the language Array Basic allows for the direct look up of array elements, such as fragment ions in a peptide CID spectrum, without the use of do-loops. If it turns out to be impractical to work Array Basic code into the P-Mod program, it may still be possible to enhance computational speed by further dividing peptide spectra into compartments. The program already divides spectra into 100 amu compartments. In addition to being necessary for background calculations, these compartments help decrease the time it takes to look up assigned ions by decreasing the size of the do-loops. Dividing peptide

spectra into smaller compartments, such as 50 amu, would provide a significant boost to look up efficiency.

The most important P-Mod improvement, at least from the standpoint of making the program available to the greater community of researchers in the field of proteomics, has to do with calibrating the statistics to individual mass spectrometers. Currently, the program statistics are only truly valid for the instrument and settings used during development. The distribution of peptide fragment ions would be different on other mass spectrometers, and could even be different on the same instrument if the instrument settings were altered. It is difficult to predict how altered peptide fragmentation would affect P-Mod statistics. In fact, the scores assigned to false positive matches are likely to be quite similar from one instrument to another, as there is no reason to believe that random matches would score any better with an altered fragmentation profile. However, it is obvious that the sensitivity, that is the scoring of correct matches, could be vary considerably with altered peptide fragmentation. The methods used to establish P-Mod statistics can be automated so that the statistical estimates can be calibrated for data acquired on different instruments. The automated calibration subroutine would simply score large batches of random peptide sequences versus a test database of peptide MS-MS spectra, group scores by the number of applied search criteria, take the extreme scores from groups of 100 comparisons for each category, and fit the extreme scores to conditional extreme value distributions using maximum likelihood estimation.

The main obstacle to applying P-Mod to data acquired on different instruments has to do with interfacing with the different output formats. Each instrument formats

outputs a little differently, interfering with P-Mod's ability to read in the distribution of MS-MS fragment ions, as well as generating a final report which displays spectra corresponding to putative matches. In order to continue interfacing in the manner that P-Mod does currently, one would have to acquire specific operation and control codes for each instrument. Unfortunately, these codes are not publicly available and would have to be provided by the instrument manufacturers. Thus, application of P-Mod to multiple types of instrument data is dependent on the ability to get manufacturers to agree to provide access to these codes. This problem is not unique to P-Mod. The lack of uniformity in MS instrument outputs limits other proteomics software as well. Consequently there has recently been an emergence of a new open-source data representation format called mzXML which would standardize the outputs from different mass spectrometers (118). Work is underway in Dr. Liebler's laboratory to adapt P-Mod to this new format.

In addition to these improvements, the utility of P-Mod would be enhanced by setting up a database environment in order to store and manage P-Mod outputs. Such a database would allow users to store the results from a large number or LC-MS-MS experiments and P-Mod analyses. Ideally, the database could then be sortable by protein, peptide, or protein modification mass-shift. The sequence coverage of a particular protein may vary experiment to experiment. Combining the results of multiple experiments and sorting the results by protein provides the best chance of maximizing sequence coverage and thereby maximizing the opportunity to observe modifications to the selected protein. Sorting results to focus on a particularly reactive peptide could

provide a quick summary of different modifications on a specific protein. Similarly, sorting results by the mass shift caused by protein modifications would provide a summary picture of the range of peptide modifications present in a batch of samples. This summary picture can provide insight into previously unknown modifications, or point to chemical artifacts resulting from sample preparation. Clustering modifications could also lead to more accurate mass estimates for observed mass-shifts. Re-evaluation of the database using corrected mass-shifts would enhance sensitivity and accuracy of mass shift localization.

The database could also play a role in decreasing output volume. If multiple analyses were conducted on the same data set using different peptide lists, the database could provide a means of eliminating conflicting assignments to a certain spectra. If multiple assignments were made to the same spectrum, the match scores would be ranked by p-value and matches with less significant p-values would be dropped from the database. In a similar fashion, duplicate spectra assigned to the same peptide and peptide modification could be dropped from the database. In a single experiment it is not uncommon to see a single peptide sequence matched to half a dozen spectra. If a large number of experiments were pooled, a single sequence could be matched to a substantial number of spectra. It would be helpful to limit the output such that only the two or three spectra with the most significant p-values are stored for each peptide or peptide modification. Inclusion of additional duplicates does not help the user and unnecessarily adds to the output volume.

Data Limitations

This dissertation has focused on new analytical tools that assist in the detection of protein modifications, however, the most significant limitations to this type of research are instrumental in nature. The sensitivity of tandem mass spectrometry is vastly superior to that of 2-D gels, however, it is still difficult to acquire peptide spectra from low abundance proteins when analyzing protein mixtures. In some cases this difficulty may be due to instrument sensitivity and true limits of detection, more often, the difficulty lays with the sample heterogeneity and the data dependent scan protocols that facilitate data acquisition.

Data dependent scanning provides an automated protocol for generating numerous MS-MS spectra from a mixture of components with unknown masses. When operating in data dependent scan mode the mass spectrometer alternates between full scan and tandem MS modes as it samples compounds eluting from an HPLC column. Following each full scan the instrument software determines the *m/z* with the highest ion intensity, and then selects that *m/z* for MS-MS analysis. The selected m/z is then placed on a dynamic exclusion list for a period of several minutes and the process is repeated. For every full scan after the first, the dynamic exclusion list is referenced; if the m/z with the highest ion intensity is on the exclusion list, the instrument selects the next most intense ion that is not on the list. In principle data dependent scanning allows the mass spectrometer to acquire MS-MS spectra for multiple components in each HPLC peak, providing an opportunity to sample low abundance peptides in the sample. In practice, low abundance peptides are still often passed over either because they elute prior to being

selected by the instrument due to prioritization of high abundance ions, or because mass spectrometers do not seem to apply the dynamic exclusion list uniformly and often acquire multiple MS-MS spectra for high abundance ions.

Bias against low intensity ions makes it difficult-to-imposable to achieve 100% sequence coverage of even high abundance proteins. Such coverage is required if we are to reliably map protein modifications. When analyzed as part of a mixture, peptide spectra from low abundance proteins are poorly represented, meaning that the sequence coverage for these proteins is exceptionally poor. It is much easier to achieve a high degree of sequence coverage for purified proteins; however, dependence on purification makes the study of modifications at the level of the proteome impossible.

Representation from low abundance peptide spectra can be enhanced by using HPLC protocols that increase the separation of peptides in complex digests. However, even under the most favorable conditions it is usually only possible to generate several thousand unique peptide spectra during a single LC-MS-MS run. When we consider that there are several million peptides in a proteome we begin to see the scope of the problem. Ultimately, new methods will have to be developed so that more comprehensive datasets can be acquired.

APPENDIX

How to Acquire SALSA and P-Mod

SALSA has been licensed by the University of Arizona to Thermo Corporation for commercial distribution and is included as a component of the Bioworks software suite. Inquiries regarding availability of SALSA should be directed to Thermo.

The compiled P-Mod program and associated source code will soon be available for free download from Dr. Liebler's laboratory web site http://www.mc.vanderbilt.edu/lieblerlab/ .

REFERENCE LIST

1. Yates, J. R., III (1998) *J. Mass Spectrom.* **33**, 1-19.

2. Aebersold, R. & Mann, M. (2003) *Nature* **422**, 198-207.

3. Loo, J. A. (2003) in *Proteome Characterization and Proteomics*, eds. Smith, R. D. & Veenstra, T. A. (Academic Press, San Diego), pp. 25-56.

4. Dalluge, J. J. & Reddy, P. (2000) *Biotechniques* **28**, 156-160.

5. Cotter, R. J. (1997) *Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research* (Oxford University Press, Oxford).

6. Cornish, T. J. & Cotter, R. J. (1993) *Rapid Commun. Mass Spectrom.* **7**, 1037-1040.

7. Yost, R. A. & Boyd, R. K. (1990) *Methods Enzymol.* **193**, 154-200.

8. Stafford, G. C., Kelley, P. E., Syka, J. P. P., Reynolds, W. E. & Todd, J. F. J. (1984) *Int. J. Mass Spectrom. Ion Proc.* **60**, 85-98.

9. Hager, J. W. (2002) *Rapid Commun. Mass Spectrom.* **16**, 512-526.

10. Wenner, B. R. & Lynn, B. C. (2004) *J. Am. Soc. Mass Spectrom.* **15**, 150-157.

11. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. & Aebersold, R. (2000) *Proc. Natl. Acad. Sci. U. S. A* **97**, 9390-9395.

12. Pappin, D. J., Hojrup, P. & Bleasby, A. J. (1993) *Curr. Biol.* **3**, 327-332.

13. Roepstorff, P. & Fohlman, J. (1984) *Biomed. Mass Spectrom.* **11**, 601.

14. Yates, J. R., III, McCormack, A. L., Link, A. J., Schieltz, D., Eng, J. & Hays, L. (1996) *Analyst* **121**, 65R-76R.

15. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. & Yates, J. R., III (1999) *Nat. Biotechnol.* **17**, 676-682.

16. Washburn, M. P., Wolters, D. & Yates, J. R., III (2001) *Nat. Biotechnol.* **19**, 242-247.

17. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. (2003) *J. Proteome. Res.* **2**, 43-50.

18. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H. & Aebersold, R. (1999) *Nat. Biotechnol.* **17**, 994-999.

19. Zhang, X., Jin, Q. K., Carr, S. A. & Annan, R. S. (2002) *Rapid Commun. Mass Spectrom.* **16**, 2325-2332.

20. Mason, D. E. & Liebler, D. C. (2003) *J. Proteome. Res.* **2**, 265-272.

21. Liu, P. & Regnier, F. E. (2002) *J. Proteome. Res.* **1**, 443-450.

22. Wilkins, M. R., Gasteiger, E., Wheeler, C. H., Lindskog, I., Sanchez, J. C., Bairoch, A., Appel, R. D., Dunn, M. J. & Hochstrasser, D. F. (1998) *Electrophoresis* **19**, 3199-3206.

23. Clauser, K. R., Baker, P. & Burlingame, A. L. (1999) *Anal. Chem.* **71**, 2871-2882.

24. Mann, M., Hojrup, P. & Roepstorff, P. (1993) *Biol. Mass Spectrom.* **22**, 338-345.

25. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. (1999) *Electrophoresis* **20**, 3551-3567.

26. Fenyo, D. (2000) *Curr. Opin. Biotechnol.* **11**, 391-395.

27. Zhang, W. & Chait, B. T. (2000) *Anal. Chem.* **72**, 2482-2489.

28. Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R., Molloy, M. P., Binz, P. A., Ou, K., Sanchez, J. C., Bairoch, A., Williams, K. L. *et al.* (1999) *J. Mol. Biol.* **289**, 645-657.

29. Eng, J. K., McCormack, A. L. & Yates, J. R., III (1995) *J. Am. Soc. Mass Spectrom.* **5**, 976-989.

30. Mann, M. & Wilm, M. (1994) *Anal. Chem.* **66**, 4390-4399.

31. Tabb, D. L., Saraf, A. & Yates, J. R., III (2003) *Anal. Chem.* **75**, 6415-6421.

32. Yates, J. R., III, Eng, J. K., McCormack, A. L. & Schieltz, D. (1995) *Anal. Chem.* **67**, 1426-1436.

33. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. (2002) *Anal. Chem.* **74**, 5383-5392.

34. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. (2003) *Anal. Chem.* **75**, 4646-4658.

35. 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. *et al.* (2002) *Proc. Natl. Acad. Sci. U. S. A* **99**, 7900-7905.

36. Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C. & Yates, J. R., III (2000) *Anal. Chem.* **72**, 757-763.

37. Creasy, D. M. & Cottrell, J. S. (2002) *Proteomics.* **2**, 1426-1434.

38. Liebler, D. C. (2002) *Environ. Health Perspect.* **110 Suppl 1**, 3-9.

39. Pevzner, P. A., Mulyukov, Z., Dancik, V. & Tang, C. L. (2001) *Genome Res.* **11**, 290-299.

40. Miller, E. C. & Miller, J. A. (1981) *Cancer* **47**, 2327-2345.

41. Guengerich, F. P. & Liebler, D. C. (1985) *Crit Rev. Toxicol.* **14**, 259-307.

42. Hinson, J. A., Pumford, N. R. & Nelson, S. D. (1994) *Drug Metab Rev.* **26**, 395-412.

43. Nelson, S. D. & Pearson, P. G. (1990) *Annu. Rev. Pharmacol. Toxicol.* **30**, 169-195.

44. Zhan, Y., van de, W. B., Wang, Y. & Stevens, J. L. (1999) *Oncogene* **18**, 6505-6512.

45. Cohen, S. D., Pumford, N. R., Khairallah, E. A., Boekelheide, K., Pohl, L. R., Amouzadeh, H. R. & Hinson, J. A. (1997) *Toxicol. Appl. Pharmacol.* **143**, 1-12.

46. Pumford, N. R. & Halmes, N. C. (1997) *Annu. Rev. Pharmacol. Toxicol.* **37**, 91-117.

47. van de, W. B., Wang, Y., Asmellash, S., Liu, H., Zhan, Y., Miller, E. & Stevens, J. L. (1999) *Chem. Res. Toxicol.* **12**, 943-951.

48. Qiu, Y., Benet, L. Z. & Burlingame, A. L. (1998) *J. Biol. Chem.* **273**, 17940-17953.

49. Hargus, S. J., Amouzedeh, H. R., Pumford, N. R., Myers, T. G., McCoy, S. C. & Pohl, L. R. (1994) *Chem. Res. Toxicol.* **7**, 575-582.

50. Pumford, N. R., Martin, B. M., Thomassen, D., Burris, J. A., Kenna, J. G., Martin, J. L. & Pohl, L. R. (1993) *Chem. Res. Toxicol.* **6**, 609-615.

51. Marks, P. A., Richon, V. M., Miller, T. & Kelly, W. K. (2004) *Adv. Cancer Res.* **91**, 137-168.

52. Schmitz, M. L., Mattioli, I., Buss, H. & Kracht, M. (2004) *Chembiochem.* **5**, 1348-1358.

53. Graham, T. R. (2004) *Curr. Biol.* **14**, R483-R485.

54. Rodriguez-Melendez, R. & Zempleni, J. (2003) *J. Nutr. Biochem.* **14**, 680-690.

55. Chapman-Smith, A. & Cronan, J. E., Jr. (1999) *Biomol. Eng* **16**, 119-125.

56. Reissner, K. J. & Aswad, D. W. (2003) *Cell Mol. Life Sci.* **60**, 1281-1295.

57. Resh, M. D. (2004) *Subcell. Biochem.* **37**, 217-232.

58. De Jonge, H. R., Hogema, B. & Tilly, B. C. (2000) *Sci. STKE.* **2000**, E1.

59. Comer, F. I. & Hart, G. W. (1999) *Biochim. Biophys. Acta* **1473**, 161-171.

60. Hata, J. A. & Koch, W. J. (2003) *Mol. Interv.* **3**, 264-272.

61. Whitmarsh, A. J. & Davis, R. J. (1999) *Sci. STKE.* **1999**, E1.

62. Holmberg, C. I., Tran, S. E., Eriksson, J. E. & Sistonen, L. (2002) *Trends Biochem. Sci.* **27**, 619-627.

63. Kehoe, J. W. & Bertozzi, C. R. (2000) *Chem. Biol.* **7**, R57-R61.

64. Wilson, V. G. & Rangasamy, D. (2001) *Exp. Cell Res.* **271**, 57-65.

65. Wilkinson, K. D. (2000) *Semin. Cell Dev. Biol.* **11**, 141-148.

66. Sun, L. & Chen, Z. J. (2004) *Curr. Opin. Cell Biol.* **16**, 119-126.

67. Marnett, L. J., Riggins, J. N. & West, J. D. (2003) *J. Clin. Invest* **111**, 583-593.

68. Butterfield, D. A. & Boyd-Kimball, D. (2004) *Brain Pathol.* **14**, 426-432.

69. Molavi, B. & Mehta, J. L. (2004) *Curr. Opin. Cardiol.* **19**, 488-493.

70. Catella-Lawson, F. & FitzGerald, G. A. (1996) *Diabetes Res. Clin. Pract.* **30 Suppl**, 13-18.

71. Boelsterli, U. A. (2003) *Toxicol. Appl. Pharmacol.* **192**, 307-322.

72. Doorn, J. A. & Petersen, D. R. (2003) *Chem. Biol. Interact.* **143-144**, 93-100.

73. Sayre, L. M., Smith, M. A. & Perry, G. (2001) *Curr. Med. Chem.* **8**, 721-738.

74. Uchida, K. (2000) *Free Radic. Biol. Med.* **28**, 1685-1696.

75. Baillie, T. A. (1992) *Int. J. Mass Spectrom. Ion Proc.* **118-119**, 289-314.

76. Pandey, A. & Mann, M. (2000) *Nature* **405**, 837-846.

77. Gygi, S. P., Han, D. K., Gingras, A. C., Sonenberg, N. & Aebersold, R. (1999) *Electrophoresis* **20**, 310-319.

78. Yates, J. R., III, McCormack, A. L. & Eng, J. (1996) *Anal. Chem.* **68**, 534A-540A.

79. Stevens, J. L., Liu, H., Halleck, M., Bowes, R. C., Chen, Q. M. & van de, W. B. (2000) *Toxicol. Lett.* **112-113**, 479-486.

80. Nelson, S. D. (1995) *Drug Metab Rev.* **27**, 147-177.

81. Nicotera, P., Bellomo, G. & Orrenius, S. (1990) *Chem. Res. Toxicol.* **3**, 484-494.

82. Chen, Q., Yu, K. & Stevens, J. L. (1992) *J. Biol. Chem.* **267**, 24322-24327.

83. Lame, M. W., Jones, A. D., Wilson, D. W., Dunston, S. K. & Segall, H. J. (2000) *J. Biol. Chem.* **275**, 29091-29099.

84. Kleiner, H. E., Rivera, M. I., Pumford, N. R., Monks, T. J. & Lau, S. S. (1998) *Chem. Res. Toxicol.* **11**, 1283-1290.

85. Pumford, N. R., Halmes, N. C. & Hinson, J. A. (1997) *Drug Metab Rev.* **29**, 39-57.

86. Rombach, E. M. & Hanzlik, R. P. (1997) *Chem. Res. Toxicol.* **10**, 1407-1411.

87. Bulera, S. J., Cohen, S. D. & Khairallah, E. A. (1996) *Toxicology* **109**, 85-99.

88. Hargus, S. J., Martin, B. M., George, J. W. & Pohl, L. R. (1995) *Chem. Res. Toxicol.* **8**, 993-996.

89. Bulera, S. J., Birge, R. B., Cohen, S. D. & Khairallah, E. A. (1995) *Toxicol. Appl. Pharmacol.* **134**, 313-320.

90. Bruschi, S. A., West, K. A., Crabb, J. W., Gupta, R. S. & Stevens, J. L. (1993) *J. Biol. Chem.* **268**, 23157-23161.

91. Bartolone, J. B., Birge, R. B., Bulera, S. J., Bruno, M. K., Nishanian, E. V., Cohen, S. D. & Khairallah, E. A. (1992) *Toxicol. Appl. Pharmacol.* **113**, 19-29.

92. Hayden, P. J., Ichimura, T., McCann, D. J., Pohl, L. R. & Stevens, J. L. (1991) *J. Biol. Chem.* **266**, 18415-18418.

93. Mattocks, A. R., Jukes, R. & Brown, J. (1989) *Toxicon* **27**, 561-567.

94. Huxtable, R., Ciaramitaro, D. & Eisenstein, D. (1978) *Mol. Pharmacol.* **14**, 1189-1203.

95. Mason, D. E. & Liebler, D. C. (2000) *Chem. Res. Toxicol.* **13**, 976-982.

96. Robertson, K. A., Seymour, J. L., Hsia, M. T. & Allen, J. R. (1977) *Cancer Res.* **37**, 3141-3144.

97. Lame, M. W., Jones, A. D., Morin, D., Wilson, D. W. & Segall, H. J. (1997) *Chem. Res. Toxicol.* **10**, 694-701.

98. Huxtable, R. J. (1979) *Gen. Pharmacol.* **10**, 159-167.

99. Mattocks, A. R., Croswell, S., Jukes, R. & Huxtable, R. J. (1991) *Toxicon* **29**, 409-415.

100. Reed, R. L., Miranda, C. L., Kedzierski, B., Henderson, M. C. & Buhler, D. R. (1992) *Xenobiotica* **22**, 1321-1327.

101. Lame, M. W., Jones, A. D., Morin, D., Segall, H. J. & Wilson, D. W. (1995) *Drug Metab Dispos.* **23**, 422-429.

102. Jones, J. A. & Liebler, D. C. (2000) *Chem. Res. Toxicol.* **13**, 1302-1312.

103. Hansen, B. T., Jones, J. A., Mason, D. E. & Liebler, D. C. (2001) *Anal. Chem.* **73**, 1676-1683.

104. Stark, G. R. (1965) *Biochemistry* **4**, 1030-1036.

105. Yan, H., Kinzler, K. W. & Vogelstein, B. (2000) *Science* **289**, 1890-1892.

106. Liebler, D. C., Hansen, B. T., Davey, S. W., Tiscareno, L. & Mason, D. E. (2002) *Anal. Chem.* **74**, 203-210.

107. Yates, J. R., III, Eng, J. K. & McCormack, A. L. (1995) *Anal. Chem.* **67**, 3202-3210.

108. Fenyo, D., Qin, J. & Chait, B. T. (1998) *Electrophoresis* **19**, 998-1005.

109. Steen, H. & Mann, M. (2001) *J. Am. Soc. Mass Spectrom.* **12**, 228-232.

110. Kinnison, R. R. (1985) *Applied Extreme Value Statistics* (Battelle Press, Columbus, OH).

111. Lihninger, A. L., Nelson, D. L. & Cox, M. M. (1993) *Principles of Biochemistry* (Worth Publishers, Inc., New York, NY).

112. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119-129.

113. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. U. S. A* **85**, 2444-2448.

114. Baynes, J. W. & Thorpe, S. R. (1999) *Diabetes* **48**, 1-9.

115. Baynes, J. W. (2003) *Clin. Chem. Lab Med.* **41**, 1159-1165.

116. Isom, A. L., Barnes, S., Wilson, L., Kirk, M., Coward, L. & rley-Usmar, V. (2004) *J. Am. Soc. Mass Spectrom.* **15**, 1136-1147.

117. Doorn, J. A. & Petersen, D. R. (2002) *Chem. Res. Toxicol.* **15**, 1445-1450.

118. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R. *et al.* (2004) *Nat. Biotechnol.* **22**, 1459-1466.