Anal. Chem. **1995,** *67,* 1426–1436

Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database

John R. Yates, III,* Jimmy K. Eng, Ashley L. McCormack, and David Schieltz

Department of Molecular Biotechnology, FJ-20, University of Washington, Seattle, Washington 98195

A method to correlate uninterpreted tandem mass spectra of modified peptides, produced under low-energy (10–50 eV) collision conditions, with amino acid sequences in a protein database has been developed. The fragmentation patterns observed in the tandem mass spectra of peptides containing covalent modifications is used to directly search and fit linear amino acid sequences in the database. Specific information relevant to sites of modification is not contained in the character-based sequence information of the databases. The search method considers each putative modification site as both modified and unmodified in one pass through the database and simultaneously considers up to three different sites of modification. The search method will identify the correct sequence if the tandem mass spectrum did not represent a modified peptide. This approach is demonstrated with peptides containing modifications such as S-carboxymethylated cysteine, oxidized methionine, phosphoserine, phosphothreonine, or phosphotyrosine. In addition, a scanning approach is used in which neutral loss scans are used to initiate the acquisition of product ion MS/MS spectra of doubly charged phosphorylated peptides during a single chromatographic run for data analysis with the database-searching algorithm. The approach described in this paper provides a convenient method to match the nascent tandem mass spectra of modified peptides to sequences in a protein database and thereby identify previously unknown sites of modification.

The synthesis of proteins in biological organisms proceeds by transcription of deoxyribonucleotide sequences to messenger RNA (mRNA). In eukaryotic organisms, mRNA is processed to remove introns and is then translated on the ribosomal complex to synthesize the protein. Almost all protein sequences are post-translationally modified in processes that may range from simple proteolytic cleavage to covalent modification of specific amino acid residues.[1] As many as 200 types of covalent modifications may exist, and their biological functions are poorly understood.[1] Organisms can increase chemical diversity, control enzymatic activity, or transmit signals through posttranslational modification of protein structure. The initiatives to completely sequence the genomes of the human and numerous model organisms will provide complete sequence information for all gene products.[2,3] Currently, information relevant to posttranslational modifications cannot be inferred from nucleotide sequence information. Thus, key elements of the control and regulation of biological processes will be lacking. For example, it is estimated that at least 1000 kinases may exist in the human genome, indicating phosphorylation is a common mechanism for the transmission of signals and control of enzyme activation.[4,5] The identification of the covalent modifications in their biological context will lead to a more detailed understanding of their regulatory roles.

Recent innovations in mass spectrometry have permitted the facile analysis of mixtures of peptides and concomitant modifications.[6] Interfacing reversed-phase high-performance liquid chromatography to electrospray ionization has led to improved approaches for the analysis of amino acid sequences and their covalent modifications.[7–9] One type of instrument capable of sequence analysis of modified peptides is the tandem mass spectrometer, in particular triple–quadrupole mass spectrometer.[10] Sequence information is obtained by selecting a precursor ion with the first mass analyzer and passing the ion into a collision cell where it is collisionally induced to dissociate. The resulting fragment ions are analyzed in the second mass analyzer.[10] Many different types of modifications have been analyzed by using tandem mass spectrometry.[11,12] Of particular interest in biology is the identification of the sites and stoichiometry of phosphorylation, a reversible modification frequently used in enzyme activation and signal transduction.[5] For example, the phosphorylation sites of the mitogen-activated protein kinase were identified by using tandem mass spectrometry.[13] A method recently proposed by Huddleston et al. for the analysis of phosphopeptides uses collision-induced dissociation (CID) in the skimmer region of an electrospray ionization source to generate phosphopeptide-specific marker ions.[14] This method can identify peptide ions containing modifications, but cannot always identify the specific

* Tel, (206)685-7388; fax, (206)685-7344; e-mail, jyates@u.washington.edu.

(1) Krishna, R. G.; Wold, F. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1993**, *67*, 265–298.
(2) Olson, M. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 4338–4344.
(3) Olson, M. *Curr. Opin. Biol.* **1992**, *2*, 221–223.
(4) Hunter, T. *Cell* **1987**, *50*, 823–829.
(5) Errede, B.; Levin, D. E. *Curr. Opin. Biol.* **1993**, *5*, 254–260.
(6) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64–71.
(7) Loo, J. A.; Udseth, H. R.; Smith, R. D. *Anal. Biochem.* **1989**, *179*, 404–412.
(8) Hail, M.; Lewis, S.; Jardine, I.; Liu, J.; Novotny, M. *J. Microcolumn. Sep.* **1990**, *2*, 285–292.
(9) Covey, T.; Huang, E.; Henion, J. *Anal. Chem.* **1991**, *63*, 1193–1200.
(10) Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *84*, 620–623.
(11) Johnson, R. S.; Ohguro, H.; Palczewski, K.; Hurley, J. B.; Walsh, K. A.; Neubert, T. A. *J. Biol. Chem.* **1994**, *269*, 21067–21071.
(12) Hackett, M.; Guo, L.; Shabanowitz, J.; Hunt, D. F.; Hewlett, E. L. *Science* **1994**, *266*, 433–435.
(13) Payne, D. M.; Rossomando, A. J.; Martino, P.; Erickson, A. K.; Her, J. H.; Shabanowitz, J.; Hunt, D. F.; Weber, M. J.; Sturgill, T. W. *EMBO J.* **1991**, *10*, 885–892.

0003-2700/95/0367-1426$9.00/0

modified residue. Phosphorylation may exist on Ser, Thr, and Tyr, and one or more of these amino acid residues may exist in the peptide. In contrast, product ion MS/MS scans provide information specific to the site of modification. Interpretation of the spectra, however, can be tedious and time-consuming.

We have demonstrated an approach to correlate tandem mass spectrometry data to the character-based amino acid sequences contained in databases.[15] This procedure takes the form of a reverse pseudospectral library search and employs a cross-correlation function to compare a "predicted" tandem mass spectrum to the experimentally derived spectrum. This approach is effective for matching sequences contained in the protein database to their corresponding tandem mass spectra. An exact match to an amino acid sequence can serve to both interpret the spectrum and identify the protein sequence from which it was derived. This approach utilizes complete uninterpreted tandem mass spectra for the search, rather than partial information such as mass or sequence, or combinations of both, for several reasons: (1) a complete spectrum allows verification of the results of the search, (2) this approach is amenable to batch processing of large numbers of spectra without the necessity of reviewing the data prior to analysis, and (3) if the sequence is not contained in the database, there is generally sufficient information to interpret the spectrum and obtain the amino acid sequence of the peptide.

The capability to match tandem mass spectra of posttranslationally modified peptides to sequences contained in the database would facilitate the process of identifying sites of modification. Additionally, an approach for identifying and including modifications within the information generated by the genome projects will add information about protein regulation and function. In this report, we describe a method to match uninterpreted tandem mass spectra of modified peptides to their corresponding sequences in the database and in the process identify the sites of modification. In the course of the analysis, specific amino acid residues are considered as both modified and unmodified. This approach, as well as an experimental method to target the acquisition of modified peptides for data analysis, is illustrated.

EXPERIMENTAL SECTION

Peptide and Protein Sources. Peptides and proteins of known sequence were obtained from the following commercial sources. α-Casein (bovine; Catalog No. C-6780, lot no. 78F-9555), α-lactalbumin (bovine; Catalog No. L-4379 , lot no. 75C-8110), cytochrome *c*, (chicken; Catalog No. C-0761, lot no. 11H-7030), amino acylase (porcine; Catalog No. A-7264, lot no. 26F-9705), and ribonuclease A (bovine; Catalog No. R-5503, lot no. 49F-8000) were obtained from Sigma Chemical Co. (St. Louis, MO). The peptides, RRLIEDNEY(PO_3H_2)TARG and DGVY(PO_3H_2)QPLRDRDDAQY-(PO_3H_2)SHLGG, were obtained from Quality Controlled Biochemicals, Inc. (Hopkinton, MA). A protein mixture containing six proteins, phosphorylase *b* (rabbit muscle), serum albumin (bovine), ovalbumin (hen), carbonic anhydrase (bovine), trypsin inhibitor (soybean), and lysozyme (hen) (SDS–PAGE low molecular weight protein standards; lot no. 27074) was obtained from BioRad (Richmond, CA). Sequencing grade trypsin (Catalog No. 1047-841, lot no. 12676220-10) was obtained from Boehringer Mannheim (Indianapolis, IN). The Sendai virus protein was generously provided by K. Gupta (Rush Presbyterian–St. Luke's Medical Center, Chicago, IL).

Peptides were generated from intact proteins by digestion with the enzyme trypsin in 50 mM Tris-HCl, pH 8.6, or 50 mM ammonium bicarbonate, pH 8.6, for 4–8 h at 37 °C. To generate the test mixture of phosphorylated and nonphosphorylated peptides, 100 pmol of the peptide RRLIEDNEY(PO_3H_2)TARG was combined with 5 μg each of phosphorylase *b* (rabbit muscle), serum albumin (bovine), ovalbumin (hen), carbonic anhydrase (bovine), trypsin inhibitor (soybean), and lysozyme (hen) and digested with the enzyme trypsin in 50 μL of 50 mM ammonium bicarbonate, pH 8.6, for 2–3 h at 37 °C. A 0.5 μL aliquot of the resulting peptide mixture was used for the neutral loss analysis.

Microcolumn High-Performance Liquid Chromatography. The model peptides were analyzed as single components or as mixtures using microcolumn high-performance liquid chromatography. Microcolumns were made by using the method of Kennedy and Jorgenson employing 98 μm i.d. fused-silica capillary tubing.[16] The columns were packed with Perseptive Biosystems (Boston, MA) POROS 10 R2, a 10 μm reversed-phase packing material, to a length of 15–20 cm. Samples were injected onto the column as previously described.[17] During injection, the effluent from the end of the column was collected with a 1–5 μL graduated glass capillary to measure the amount of liquid displaced from the column. Once a sufficient volume had been injected, the column was connected to the HPLC pumps. HPLC was performed using Applied Biosystems (Foster City, CA) 140B microgradient pump with dual-syringe pumps and a 250 μL dynamic mixer. The flow from the pump was reduced from 100 to 1 μL/min using a splitting tee and a length of restriction tubing made from fused silica. The mobile phase used for gradient elution consisted of (A) 0.5% acetic acid and (B) acetonitrile/water 80:20 (v/v) containing 0.5% acetic acid. The gradient was linear from 0 to 80% B over 30 min. The fritted end of the column was inserted directly into the electrospray needle. A sheath liquid flowed concentrically around the end of the column at a flow rate of 1.5 μL/min and was a methanol/water (70:30) mixture containing 0.1% acetic acid.

Electrospray Ionization and Tandem Mass Spectrometry. Mass spectra were recorded on a Finnigan MAT (San Jose, CA) TSQ700 equipped with an electrospray ionization source as previously described.[18] Electrospray ionization was performed under the following conditions. The needle voltage was set at 4.6 kV. The sheath and auxiliary gases consisted of nitrogen gas (99.999%) and were set at 20 psi and 5 units, respectively. The heated capillary temperature was set at 150 °C, and a potential of 70 V was placed on the capillary. The molecular weights of peptides were recorded by scanning the mass analyzer at a rate of 500 u/s over a range of 400–1400 u throughout the HPLC gradient. Sequence analysis of peptides was performed during a second HPLC analysis by selecting the precursor ion with a 2–3 u (full width at half-height) wide window in the first mass analyzer and passing the ions into a collision cell filled with argon to a pressure of 3–5 mTorr. Collision energies were on the order of

(14) Huddleston, M. J.; Annan, R. S.; Bean, M. F.; Carr, S. A. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 710–717.
(15) Eng, J.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(16) Kennedy, R. T.; Jorgenson, J. W. *Anal. Chem.* **1989**, *56*, 1128–1135.
(17) Yates, J. R., III; McCormack, A. L.; Hayden, J. B.; Davey, M. P. *Cell Biology: A Laboratory Handbook*; Celis, J. E., Ed.; Academic Press: San Diego, CA, 1994; pp 380–388.
(18) Griffin, P. R.; Coffman, J. A.; Hood, L. E.; Yates, J. R., III *Int. J. Mass Spectrom. Ion Processes* **1991**, *111*, 131–149.

10−50 eV (Elab). The second mass analyzer was scanned at 500 u/s to record the *m/z* of the fragment ions. Peak widths in the second mass analyzer ranged from 1.5 to 2.0 u.

To selectively obtain product ion spectra of phosphopeptides, the scan modes of the mass spectrometer were controlled through a computer program written in the TSQ 700's instrument control language (ICL). At the start of the program, the instrument was configured to scan in a neutral loss mode to detect loss of 49 u from doubly charged phosphopeptide ions (98/2 = 49) over a mass range of 400−1400 with a scan time of 1.5 s. When the ion current for a particular *m/z* value was above a preset threshold value of 50 000 counts, the program switched the instrument to scan in the product ion MS/MS configuration. The *m/z* value measured in the neutral loss mode was used to set the precursor *m/z* value in the first mass analyzer. The mass range was set at 50 u to 2 times the precursor ion *m/z* value with a scan time of 1.5 s during acquisition of product ion spectra. The instrument acquired five scans and then returned to the neutral loss mode of scanning. The analysis was performed with the collision cell filled with argon to a pressure of 3.5 mTorr. The collision offset during the neutral loss scans was 20 eV. The collision energy for product ion scans was set by dividing the *m/z* value of the precursor ion by a constant (−35).

Computer Analysis of MS/MS Data. All computer algorithms were written in the C-programming language under the UNIX operating system. The OWL database version 24.0 (88 823 entries) was obtained as an ASCII text file in the FASTA format from the National Center for Biotechnology Information (Washington, DC). OWL is a nonredundant database comprising protein sequences from the GenBank (Release 84.0, National Center for Biotechnology Information, Washington, DC), SWISS-PROT (Release 29, Swiss Institute of Technology), Protein Information Resource (Release 41, National Biology Resource Foundation, Georgetown, Washington, DC), and National Research Laboratory (Release 14, Brookhaven National Laboratory, Brookhaven, NY) databases. A modified human species-specific database was constructed by selecting protein sequences derived from *Homo sapiens* (14 190 sequences) from the OWL database and adding the peptide and protein sequences of the various standards. The sequence of trypsin (bovine) was added to the modified human database to screen for autolysis products. All searches were performed on a Hewlett Packard HP9000 minicomputer (Palo Alto, CA).

A detailed description of the methods used for data reduction, preliminary scoring, and final scoring used in the computer algorithm have been presented elsewhere.[15] A brief description will be presented here with specific details of the search method for the spectra of modified peptides described in the Results and Discussion section. Each tandem mass spectrum undergoes two types of data preprocessing. In the first method, termed search preprocessing, spectrum data were converted to a list of masses and intensities and a 10 u window around the precursor ion was removed from the file. The ion intensity values were normalized to 100.0. All ions except the 200 most intense ions were removed from the file. A second data preprocessing method, termed correlation preprocessing, was employed to prepare the experimentally obtained spectrum for correlation analysis. The spectrum was processed by removing a 10 u region around the *m/z* value of the precursor ion and dividing the spectrum into 10 equal sections. The ions in each section were then normalized to a maximum value of 50.0.

Protein sequences were analyzed sequentially until the entire database had been searched. The amino acid masses used to calculate the mass of amino acid sequences obtained from the database were based on average chemical masses. Fragment ion values were calculated as previously described.[15] Mass values used to calculate phosphoserine, phosphothreonine, and phosphotyrosine residues were 167.08, 181.11, and 243.18 u, respectively. The mass values used for oxidized methionine and S-carboxymethylated cysteine were 147.19 and 161.18 u, respectively. A preliminary score (S_p) was determined by comparing the information contained in the search preprocessed data to the predicted fragment ion values calculated for each sequence derived from the database. The following formula was used:

$$S_p = (\sum i_m) n_i (1 + \beta)(1 + \varrho)/\eta_\tau \qquad (1)$$

where i_m is the matched intensities (within ±1 u), n_i is the number of matched fragment ions, β is the type b- and y-ion continuity, ϱ is the presence of immonium ions in the spectrum and their respective amino acids in the predicted sequence, and n_τ is the total number of predicted fragment ions.

Cross-Correlation Analysis. A cross-correlation function was used to compare the tandem mass spectrum to a "spectrum" reconstructed using each of the amino acid sequences identified in the search. The appearance of a tandem mass "spectrum" was reconstructed by calculating the *m/z* of the fragment ions and assigning abundance values of 50.0 to the b- and y-type ions and a value of 25.0 for signals within *m/z* ±1 of the b and y ions. The neutral losses of ammonia and water from b- and y-type ions are assigned values of 10.0, and carbon monoxide losses (type a ions) are given values of 10.0. Cross-correlations are computed by fast Fourier transformation of the reconstructed and correlation preprocessed data sets zero padded to 4096 points, multiplication of one transform by the complex conjugate of the other, and inverse transformation of the resulting product. The final score attributed to each candidate peptide sequence is the value of the function when $\tau = 0$ minus the mean of the cross-correlation function over the range $-75 \leq \tau \leq 75$.[15,19] The scores are normalized to 1.0 and termed C_n.

RESULTS AND DISCUSSION

A covalent modification of an amino acid results in a change of the mass predicted by its sequence. The difference in mass is generally sufficient to detect the presence of a modification, but not the specific site within the sequence. This method, however, requires a priori knowledge of the peptide's sequence and cannot be extended to a peptide whose sequence is not known. The accumulation and storage of nucleotide and protein sequence information in databases has facilitated the rapid identification of peptides and proteins by using mass spectrometry and computer data analysis with protein databases.[15,20−24] With the increasing amount of sequence information accumulating in databases, a

(19) Powell, L. A.; Heiftje, G. M. *Anal. Chim. Acta* **1978**, *100*, 313−327.
(20) Henzel, W.; Billeci, T.; Stults, J.; Wond, S.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A* **1993**, *90*, 5011−5015.
(21) Yates, J. R.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397−408.
(22) Pappin, D.; Hojrup, P.; Bleasby, A. *Curr. Biol.* **1993**, *3*, 327−332.
(23) James, P.; Qaudroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58−64.

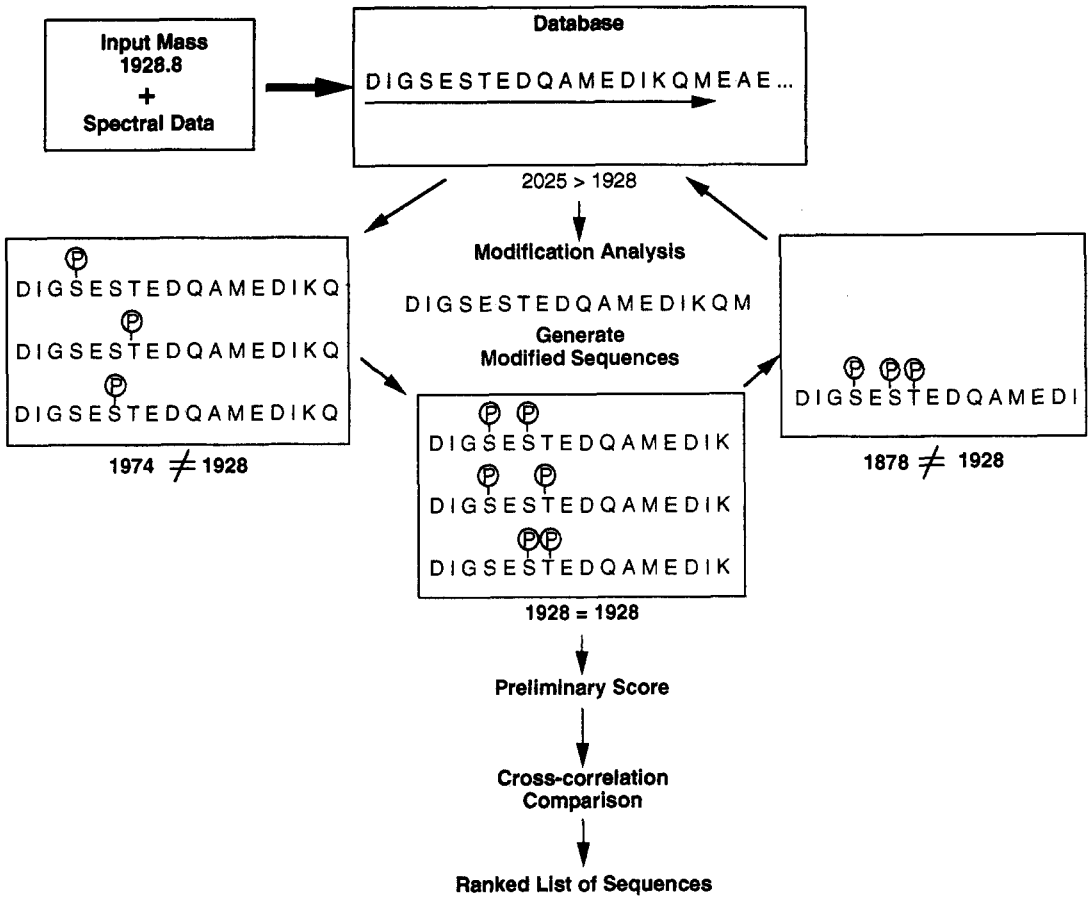

Figure 1. Schematic of the approach used by the computer algorithm to match tandem mass spectra of modified peptides to sequences in the protein database.

peptide or protein's sequence may not be known to the researcher but may exist in a database. The use of computer methods to identify the sequence of a modified peptide in a database, however, requires altering the amino acid mass value used to calculate peptide masses from the database to account for the suspected modification. This would result in the universal application of the mass to all occurrences of the putative modification site. If a modification does not exist at every occurrence of a putative modification site (amino acid), then the search will fail for unmodified sequences as well as for sequences containing two or more amino acid residues with only one modified site. To subject a spectrum to a more general and appropriate search, each putative modification site should be considered as both modified and unmodified. Thus, our criteria for the searching method presented here was to (1) identify modifications without using the mass for the modified amino acid at all occurrences of the amino acid (differential modification), (2) simultaneously consider more than one site of modification, and (3) identify the correct sequence if the tandem mass spectrum did not represent a modified peptide.

To study the analysis of tandem mass spectra obtained from modified peptides and test the efficacy of this approach, we considered three different cases of covalent modification. The simplest case, modification of a single type of amino acid, is illustrated using peptides containing S-carboxymethylated Cys residues or oxidized Met. A second example, simultaneous consideration of modifications to two different types of amino acids, is illustrated with peptides containing both S-carboxymethylated Cys and oxidized Met. The last example is simultaneous consideration of modification to three different amino acids. This example is illustrated with phosphopeptides where all occurrences of Ser, Thr, and Tyr are considered as possible sites of modification.

Algorithm for Modified Searches. To search the database with tandem mass spectra of modified peptides, we have modified our original algorithm to consider the possibility of covalent modification.[15] The analysis strategy begins with computer reduction of the MS/MS data as described in the Experimental Section. Linear amino acid sequences are then identified in a protein database for comparison to the processed MS/MS data when the mass of the linear sequence matches or exceeds the input mass. No assumptions are made about the identity of the terminating amino acid based on the type of proteolytic enzyme used to create the peptides. If the mass of the candidate peptide is equal to the input mass (mass tolerance ±3 u), the candidate peptide undergoes preliminary scoring (S_p) directly (eq 1). By comparing direct matches, a potentially nonmodified sequence is given equal consideration to a modified sequence. Whenever the mass of the candidate peptide sequence is greater than the input mass, the algorithm returns to the first amino acid in this series and recalculates the mass considering differential modification at each putative modification site. For example, if three potential sites of phosphorylation exist in the sequence, all combinations of modification are considered and the mass is calculated for each combination (Figure 1). All candidate peptides that match the input mass undergo preliminary scoring (eq 1). The search

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

Table 1. Results of Searches by Using the Collision-Induced Dissociation Mass Spectra of Peptides Containing Modifications to a Single Type of Amino Acid Obtained by Proteolytic Digestion of Proteins of Known Sequence[a]

description	sequence	OWL C_n	OWL ΔC_n	specie C_n	specie ΔC_n	prot C_n	prot ΔC_n
b. α-lactalbumin	GYGGVSLPEWV***C***TTFH	1(1)	0.409	1	0.474	1	0.808
b. α-lactalbumin	LDQWL***C***EK	2(1)	0.065	1	0.017	1	0.574
b. α-lactalbumin	SSNI***C***NIS***C***DK	1(1)	0.260	1	0.284	1	0.794
b. ribonuclease	SLADVKAV***C***SQK	1(1)	0.050	1	0.436	1	0.653
b. serum albumin	ADI***C***TLPDTEK	1(1)	0.132	1	0.203	1	0.625
b. serum albumin	DDPHA***C***YSTVFDK	1(1)	0.070	1	0.165	1	0.617
b. serum albumin	EA***C***FAVEGPK	1(1)	0.110	2	0.007	1	0.318
b. serum albumin, +3	E***CC***HGDLLE***C***ADDR	1(1)	0.015	1	0.080	1	0.432
b. serum albumin	EYEATLEE***CC***AK	1(1)	0.094	1	0.097	1	0.097
b. serum albumin	HADI***C***TLPDTEK	1(1)	0.250	1	0.267	1	0.623
b. serum albumin	NE***C***FLSHK	2(1)	0.056	2	0.056	1	0.454
b. serum albumin	TLFGDEL***C***K	3(1)	0.031	1	0.028	1	0.603
b. serum albumin	YI***C***DNQDTISSK	1(1)	0.240	1	0.250	1	0.727
b. serum albumin	YNGVFQE***CC***QAEDK	1(1)	0.175	1	0.334	1	0.334
p. amino acylase	LALELEI***C***PASTDAR	1(1)	0.038	1	0.038	1	0.854
h. HLA class I, +3	VDDTQFVRFDSDAASQR***M***E[b,c]	1	0.112	1	0.187	1	0.603
c. cytochrome *c*	GITWGEDTL***M***EYLENPK[c]	1	0.224	1	0.383	1	0.747
c. cytochrome *c*	**GITWGEDTLMEYLENPK**[c]	1	0.278	1	0.424	1	0.766
b. trypsin	**APILSDSSCK**	5(−)	0.021	1	0.011	1	0.434
b. trypsin	**LQGIVSWSGCAQK**	1(−)	0.111	1	0.265	1	0.714
b. trypsin	**VCNYVSWIK**	1(−)	0.021	1	0.021	1	0.495
b. trypsin	**VASISLPTSCASAGTQCLISGWGNTK**	1(−)	0.299	1	0.321	1	0.854

[a] In the course of the search, each putative site of modification is considered as modified and unmodified. The boldface, italic type letters ***C*** and ***M*** designate S-carboxymethylated Cys and oxidized Met, respectively. Unmodified peptide sequences are displayed in boldface, roman type. These spectra were analyzed under the same search conditions as the spectra representing modified sequences. The numbers in parentheses in the OWL database column are the rankings of spectra determined during searches considering universal modification of Cys amino acid residues. All spectra were acquired under electrospray ionization conditions. Charge states of the ion used in the MS/MS experiment are listed along with the protein description if it was different from +2. The designations C_n refers to rankings of the correct amino acid sequences by correlation parameter. These rankings are provided for searches of the OWL database, a human sequence database (specie), and through the individual protein sequence (prot). The column ΔC_n is the difference between the top scoring sequence and the next highest scoring sequence. The mass tolerance used in the searches was ±3 u. Fragment ion mass tolerance was ±1 u. The abbreviations b, c, p, and h indicate bovine, chicken, porcine, and human, respectively. [b] Aspartic acid and glutamic acid were converted to their corresponding methyl esters; consequently, their masses were considered as 129.09 and 143.12 u, respectively, throughout the search. [c] These peptides were analyzed considering oxidized Met as the modification.

algorithm then starts summing from the next position in the protein to determine the next candidate peptide sequence. This process is repeated until the entire database is searched. The preliminary score (S_p) for each sequence is ranked. The top 500 sequences are then subjected to a correlation-based analysis to generate a final score and ranking of the sequences.[15]

Modifications to a Single Type of Amino Acid. Results from the analysis of spectra representing peptides with modifications to a single type of amino acid are displayed in Table 1. Each amino acid that could function as a potential site of modification was considered as modified and unmodified in the course of a single search. In Table 1, sites of modification identified in the amino acid sequences are presented in boldface italic type. The only information input into the program was the spectrum, the mass, and the type of modification to consider. In a majority of the examples, the correct result is obtained regardless of whether the peptide is modified or not. Of the examples of modified peptides in Table 1, three false positives were observed through the OWL database, although no correct answer ranked lower than third. The number of false positives was reduced to two in a search of the modified human sequence database. As controls, spectra from five unmodified peptides, containing Cys or Met, were analyzed under the same search conditions. Four of the controls rank number one in all searches, and the identified amino acid sequences are displayed at the bottom of Table 1 in boldface, roman type. Covalent modification of Cys is a procedure frequently employed in the preparation of proteins for proteolysis. In general, the procedure is quantitative, so no differential modification within a sequence would be expected. Oxidation of Met in the course of sample isolation is more likely to result in differential modification of the Met residues contained in the protein sequence. These modifications served as models to test the method since numerous examples could be analyzed. Peptides modified at specific regulatory sites by enzymatic action are more likely to show differential modification within a sequence.

The intent of this study was to determine the efficacy of a search that attempts to match a tandem mass spectrum to a sequence considering the spectrum to represent a sequence containing a single type of modification. By employing a strategy to consider each potential modification site as modified or unmodified, a process that increases the number of amino acid sequences evaluated during the search, the likelihood of false positives could increase. The observed rate of false positives, however, does not increase significantly over the use of a universal mass change for the potential modification. These data are shown in Table 1 and are given in parentheses under the OWL database search column. A search procedure employing a universal mass change was described previously.[15]

An advantage to using a search method which considers differential modification in a single search is the ability to employ a batch mode for analyzing spectra. Results from the batch mode analysis of a set of 19 spectra considering modifications to Cys are shown in Table 2. All spectra were analyzed under the same search conditions without any presearch analysis of the data, e.g., interpretation. The total time required for data analysis was slightly over 32 min or an average of 1 min and 40 s per spectrum. Through the use of computer-controlled data acquisition, we have acquired as many as 215 tandem mass spectra in a single LC/MS/MS analysis.[25] Thus, the ability to analyze tandem mass

Table 2. Results of a Batch-Mode Search Considering a Single Type of Modification for a Set of Spectra Produced by Using Collision-Induced Dissociation[a]

no.	description	sequence	modified human C_n
1.	b. α-lactalbumin	GYGGVSLPEWV***C***TTFH	1
2.	b. α-lactalbumin	LDQWL***C***EK	1
3.	b. α-lactalbumin	SSNI***C***NIS***C***DK	1
4.	b. ribonuclease	SLADVKAV***C***SQK	1
5.	b. serum albumin	ADI***C***TLPDTEK	1
6.	b. serum albumin	DDPHA***C***YSTVFDK	1
7.	b. serum albumin	EA***C***FAVEGPK	2
8.	b. serum albumin, +3	E***CC***HGDLLE***C***ADDR	1
9.	b. serum albumin	EYEATLEE***CC***AK	1
10.	b. serum albumin	HADI***C***TLPDTEK	1
11.	b. serum albumin	NE***C***FLSHK	2
12.	b. serum albumin	TLFGDEL***C***K	1
13.	b. serum albumin	YI***C***DNQDTISSK	1
14.	b. serum albumin	YNGVFQE***CC***QAEDK	1
15.	p. amino acylase	LALELEI***C***PASTDAR	1
16.	b. trypsin	**APILSDSSCK**	1
17.	b. trypsin	**LQGIVSWSGCAQK**	1
18.	b. trypsin	**VCNYVSWIK**	1
19.	b. trypsin	**VASISLPTSCASAGTQCLISGWGNTK**	1

[a] In the course of the search, each putative site of modification is considered as modified and unmodified. Total time for consideration of all the spectra listed in the table was 32 min. The complete amino acid sequences displayed in boldface, roman type are unmodified. The letter ***C*** in boldface, italic type designates S-carboxymethylated Cys. All spectra were acquired under electrospray ionization conditions. Charge states of the ion used in the MS/MS experiment are listed along with the protein description if it was different than +2. The designations C_n refers to rankings of the correct amino acid sequences by correlation parameter. These rankings are provided for searches of a modified human sequence database. The mass tolerance used in the searches was ±3 u. Fragment ion mass tolerance was ±1 u. The abbreviations b and p indicate bovine and porcine, respectively.

spectra in a batch mode has become important for efficient analysis of the enormous quantity of data that can be produced during LC/MS/MS.

Modification to Two Different Amino Acids: Oxidation of Methionine and S-Carboxymethylation of Cysteine. The CID spectra of two peptides containing both Met and Cys are shown in Figure 2. These spectra were obtained by proteolytic digestion of α-lactalbumin (bovine) by using the enzyme trypsin. Figure 2A contains the tandem mass spectrum of the peptide Phe-Leu-Asp-Asp-Asp-Leu-Thr-Asp-Asp-Ile-Met-Cys-Val-Lys with a modified Cys residue. A search through the OWL database using this spectrum resulted in the identification of the above sequence as containing S-carboxymethylated Cys and unoxidized Met with a ΔC_n score of 0.288. The spectrum shown in Figure 2B was correctly identified as Phe-Leu-Asp-Asp-Asp-Leu-Thr-Asp-Asp-Ile-Met-Cys-Val-Lys containing Met (ox) and S-carboxymethylated Cys at positions 11 and 12, respectively. A ΔC_n score of 0.301 was observed. Both spectra were correctly identified as coming from bovine α-lactalbumin in a search of the OWL database.

An important consideration in determining a site or sites of modification is the presence of sufficient sequence information to identify the specific site. It can be discerned from the measured mass of each peptide that they contain modifications, but to identify the sites of modification, sequence ions must be present in the spectrum which allow the sites to be distinguished. Both spectra in Figure 2 contain strong b- and y-type fragment ions. The b-ion series is the same for both peptides up to the b_{11} ion. The b_{11} ion observed in Figure 2B is shifted in mass by 16 u above the corresponding ion in the spectrum shown in Figure 2A, but no b_{12} is present to indicate the presence of S-carboxymethylated Cys. A strong set of y-type ions is also present in the spectrum. Because the modification occurs in the C-terminal portion of the sequence, all the ions beyond y_4 increase in mass by 16 and 58 u. The presence of these ions in the spectrum increases the score of the sequence containing oxidized Met and results in a correct identification of both the site and type of modification. If sufficient signal to noise (S/N) is not present in the spectrum to discern the site of modification, the algorithm may fail to correctly identify the site but may still correctly identify the sequence if there are other potential sites of modification in the sequence. The mass of the sequence including modifications must correspond to the mass calculated from the spectrum.

By considering modifications to two different amino acids, the number of amino acid sequences that are examined increases in some cases up to 5-fold. This strategy increases the time required to analyze a single spectrum and has an impact on the processing time for batch-mode analysis. A set of 22 spectra representing the amino acid sequences displayed in Table 3 were analyzed in a batch mode. The total time of analysis for a search of the modified human proteins database was ~ 1 h and 6 min or an average of 3 min per spectrum. The analysis time was roughly double over the time required for a single type of modification.

Oxidation of Met is often a byproduct of sample manipulation in the presence of oxygen. The presence of oxidized Met within a sequence could result in an incorrect identification of the sequence during manual interpretation, since its mass is the same as Phe. Even though oxidized Met has the same residue mass as Phe, the algorithm calculates the mass and fragment ions for the peptide sequences contained in the database. Consequently, when Met is encountered, its mass is considered as 131.19 and 147.18 u in two separate cycles through the sequence. The level at which this modification exists in vivo has been difficult to determine since isolation procedures frequently expose proteins to conditions favoring oxidation.[26] Fliss et al. attempted to measure the amount of oxidized Met in neutrophils and found it

(25) John R. Yates, Ashley L. McCormack, Jimmy Eng *Identification of Individual Proteins in Mixtures using Micro-Column HPLC Tandem Mass Spectrometry and Automated Database Analysis*; presented at the Methods in Protein Structure Analysis Meeting, Snowbird, UT, September 9–13, 1994.

(26) Brot, N.; Weissbach, H. *Arch. Biochem. Biophys.* **1983**, *223*, 271–281.

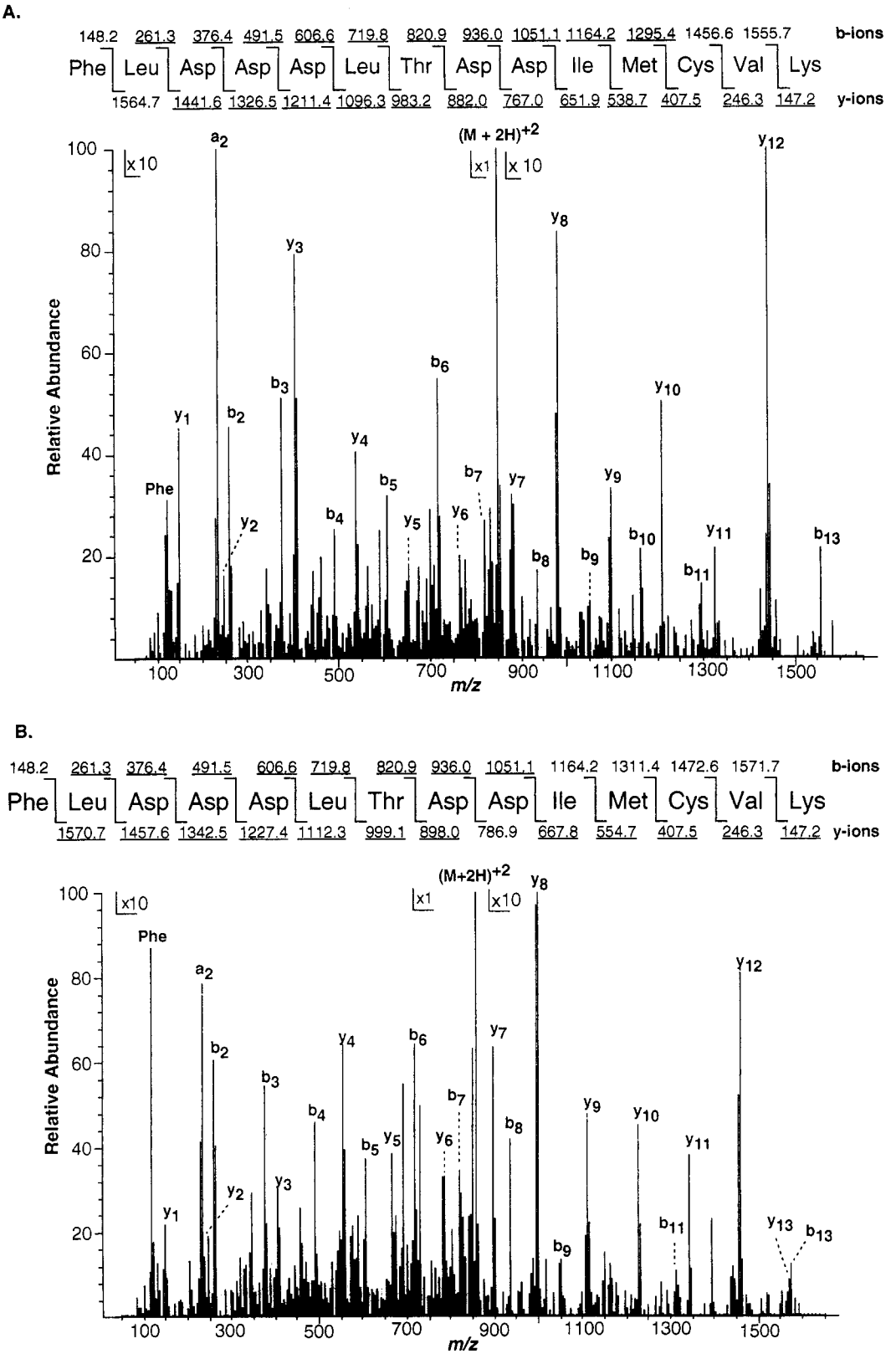

Figure 2. (A) CID mass spectrum recorded on the $(M + 2H)^{2+}$ ions at *m/z* 851 of a peptide derived from trypsin digestion of the protein α-lactalbumin. Fragments of type b and y ions having the general formulas $H(NHCHRCO)_n^+$ and $H_2(NHCHRCO)_nOH^+$, respectively, are shown above and below the amino acid sequence at the top of the figure. Ions observed in the spectrum are underlined. Leu and Ile were assigned by correspondence to the sequence derived from the database. (B) CID mass spectrum recorded on the $(M + 2H)^{2+}$ ions at *m/z* 860 of a peptide derived from trypsin digestion of the protein α-lactalbumin. Fragments of type b and y ions having the general formulas $H(NHCHRCO)_n^+$ and $H_2(NHCHRCO)_nOH^+$, respectively, are shown above and below the amino acid sequence at the top of the figure. Ions observed in the spectrum are underlined. Leu and Ile were assigned by correspondence to the sequence derived from the database.

to be present at levels of 9–22%, depending on the state of cellular activation with phorbol 12-myristate 13-acetate, a potent inducer of a respiratory burst that results in increased levels of oxidizing reagents.[27]

Table 3. Results of a Batch-Mode Search Considering Two Different Types of Modifications for a Set of Spectra Produced by Using the Collision-Induced Dissociation[a]

no.	description	sequence	modified human C_n
1.	b. α-lactalbumin	GYGGVSLPEWV***C***TTFH	1
2.	b. α-lactalbumin	LDQWL***C***EK	1
3.	b. α-lactalbumin	SSNI***C***NIS***C***DK	1
4.	b. ribonuclease	SLADVKAV***C***SQK	1
5.	b. serum albumin	ADI***C***TLPDTEK	1
6.	b. serum albumin	DDPHA***C***YSTVFDK	1
7.	b. serum albumin	EA***C***FAVEGPK	2
8.	b. serum albumin, +3	E***CC***HGDLLE***C***ADDR	1
9.	b. serum albumin	EYEATLEE***CC***AK	1
10.	b. serum albumin	HADI***C***TLPDTEK	1
11.	b. serum albumin	NE***C***FLSHK	2
12.	b. serum albumin	TLFGDEL***C***K	1
13.	b. serum albumin	YI***C***DNQDTISSK	1
14.	b. serum albumin	YNGVFQE***CC***QAEDK	1
15.	p. amino acylase	LALELEI***C***PASTDAR	1
16.	b. trypsin	**APILSDSSCK**	1
17.	b. trypsin	**LQGIVSWSGCAQK**	1
18.	b. trypsin	**VCNYVSWIK**	1
19.	b. trypsin	**VASISLPTSCASAGTQCLISGWGNTK**	1
20.	b. lactalbumin	FLDDDLTDDI***MC***VK	1
21.	b. lactalbumin	FLDDDLTDDIM***C***VK	1
22.	b. serum albumin	MP***C***TEDYLSLILNR	1

[a] In the course of the search, each putative site of modification is considered as modified and unmodified. Total time of analysis for the spectra listed in the table was 1 h and 6 min. The letters ***C*** and ***M*** displayed in boldface, italic type designate S-carboxymethylated Cys and oxidized Met, respectively. The complete amino acid sequences displayed in boldface, roman type are unmodified. All spectra were acquired under electrospray ionization conditions. Charge states of the ion used in the MS/MS experiment are listed along with the protein description if it was different than +2. The designations C_n refers to rankings of the correct amino acid sequences by correlation parameter. These rankings are provided for a search of a modified human sequence database. The mass tolerance used in the searches was ±3 u. Fragment ion mass tolerance was ±1 u. The abbreviations b and p indicate bovine and porcine, respectively.

Table 4. Results for the Analysis of a Phosphopeptide Derived by Using Trypsin Digestion of the Protein α-Casein through the OWL Database[a]

rank (S_p)	$(M + H)^+$	C_n	ΔC_n	ions	amino acid sequence
1 (1)	1928.8	1.0000	0.0000	22/30	DIG***S****E****S****T*EDQAMEDIK
2 (4)	1928.8	0.9275	0.0725	22/30	DIGSE***S****T*EDQAMEDIK
3 (27)	1928.8	0.9070	0.0930	20/30	DIG***S****E****S****T*EDQAMEDIK
4 (51)	1929.2	0.8960	0.1040	20/30	EVEILYTVFKAYPDIQ
5 (5)	1927.0	0.8208	0.1792	20/30	EEIVQEEG***T***VTEEIIQ

[a] The first column signifies the rank of the sequence using a cross-correlation function and the rank, in parentheses, of the sequence from preliminary scoring (S_p). C_n signifies the normalized score from the cross-correlation function. ΔC_n is the difference between the cross-correlation parameter of the top-scoring sequence and the listed sequence. The column designated ions is the number of ions observed in the mass spectrum versus the number of ions predicted. The phosphorylated residues are indicated in boldface, italic type.

Modification to Three Different Amino Acids: Phosphorylation. A representative tandem mass spectrum of the phosphopeptide DIGS(PO_3H_2)ES(PO_3H_2)TEDQAMEDIK, derived from tryptic digestion of α-casein (bovine), is shown in Figure 3. This sequence contains three possible sites of phosphorylation (two Ser and one Thr). A search of the OWL database successfully matched the spectrum to one of seven possible phosphorylated peptides for this sequence. No information relevant to the number of expected phospho amino acids was input into the program. Only the sequences containing two phosphate groups fit the mass, and each of those three possibilities was compared to the spectrum and a preliminary score calculated (eq 1). The scores were then ranked against all the possibilities found in the database, and the top 500 candidates were compared to the experimental spectrum using a cross-correlation function.[15] The results for an analysis against the 88 823 protein sequences contained in the OWL database are displayed in Table 4. Consideration of all possible modification sites at three potential locations is a large combinatorial problem and results in a search time on the order of 1 h. The results obtained for other phosphopeptides are shown in Table 5. At the bottom of Table 5, data from searches conducted using the same search conditions with spectra representing nonphosphorylated sequences are shown as negative controls. All answers for the negative controls are correct. Nine out of 13 of the spectra for phosphopeptides were correctly identified as well as the sites of modification in the OWL database search. Of the four false positives, two were identified as the correct sequence, but the modification site was incorrectly identified. The most critical factor to a successful search is obtaining a spectrum containing the relevant sequence ions with good S/N to determine the sites of modification. This is a factor in the failure of several spectra to correctly identify the modification site or the sequence in a search of the OWL database. In particular, the spectrum of the +2 charge state of the peptide DDAQY(PO_3H_2)SHLGG ranks 18th in the search. The fragmentation of this peptide under CID conditions (data not shown) exhibits poor S/N, possibly due to the lack of a strong basic amino acid residue to direct fragmentation. The tandem mass spectrum for the +1 charge state of the peptide resulted in improved S/N and subsequent score.

Tyrosine residues are also known to be sulfated. The mass difference between sulfated and phosphorylated tyrosine is insufficient to distinguish these two modifications of tyrosine. Tyrosine sulfate occurs in ~1% of proteins and is, apparently, an enzymati-

(27) Fliss, H.; Weissbach, H.; Brot, N. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 7160–7164.

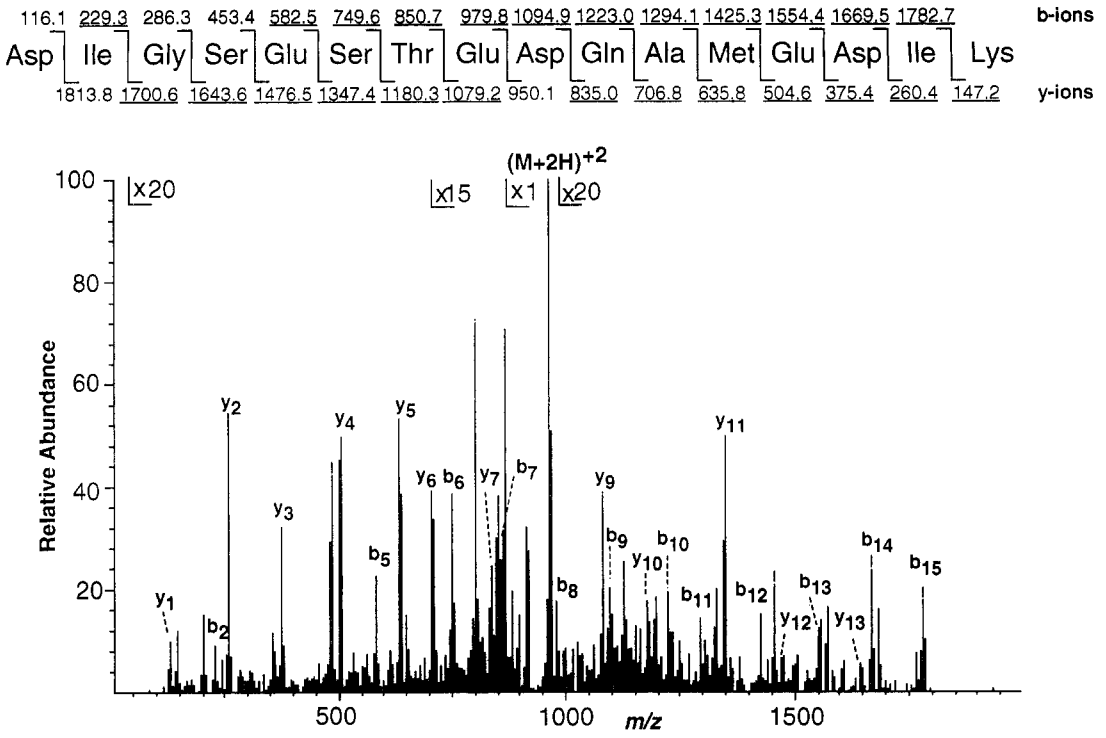

Figure 3. CID mass spectrum recorded on the $(M + 2H)^{2+}$ ions at *m/z* 964 of a peptide derived from trypsin digestion of α-casein. Fragments of type b and y ions having the general formulas $H(NHCHRCO)_n^+$ and $H_2(NHCHRCO)_nOH^+$, respectively, are shown above and below the amino acid sequence at the top of the figure. Ions observed in the spectrum are underlined. Leu and Ile were assigned by correspondence to the sequence derived from the database.

cally irreversible modification. This modification has been found in multicellular eukaryotic organisms and appears to the most abundant modification to tyrosine.[28] Gibson and Cohen have observed this modification to be more acid labile, which may result in greater neutral losses from the precursor ion during mass spectrometry.[29] At present it is impossible to differentiate between these two modifications to tyrosine with our search procedure.

We have combined this computer algorithm for data analysis with a method for the selective generation of product ion spectra of phosphopeptides contained in complex mixtures of peptides. This method utilizes a neutral loss analysis, based in part on previous work of several groups, to automatically select peptides for product ion MS/MS analysis.[14,30,31] The generation of a signal above a preset threshold, due to a neutral loss of 49 u (H_3PO_4, 98/2 from phospho serine or threonine or HPO3 + H2O, (80 + 18)/2 from phosphotyrosine) from the precursor ion, triggers the instrument to convert to a product ion MS/MS scan mode. The *m/z* value obtained from the neutral loss scan is then used as the precursor *m/z*. Figure 4A shows an LC/MS analysis of six proteins digested with the protease trypsin and spiked (prior to digestion) with the phosphotyrosine-containing peptide RRLIEDNEY(PO_3H_2)TARG. Figure 4B displays an LC-neutral loss/product ion MS/MS analysis of the same mixture. Tandem mass spectra for two peptides were generated in the procedure. Both spectra were analyzed by the search algorithm, and the sequences RRLIEDNEY(PO_3H_2)TARG (mass 1674 Da) and RLIEDNEY(PO_3H_2)TARG (mass 1516 Da) were identified as the number one ranked sequences. The doubly charged ions of these two peptides produced sufficient signal to trigger product ion MS/MS scans using the doubly charged *m/z* values as the precursor ions. The advantages of this approach are the selective generation of tandem mass spectra of phosphorylated peptides and the acquisition of data that contain information about sites of modification. Elimination of neutral phosphoric acid from phosphotyrosine-containing peptide ions is thought to be a less favored process than from phosphoserine- or phosphothreonine-containing peptides.[14] Huddleston et al. noted this loss of 98 Da may be due to the combined loss of HPO_3 and H_2O from the precursor ion.[14] In the experiment shown in Figure 4A, sufficient ion current from the neutral loss is generated in this process to trigger the acquisition of product ion MS/MS spectra for phosphotyrosine-containing peptides. By combining selective data acquisition techniques with the above data analysis approach, MS/MS data for phosphopeptides contained in complex mixtures of peptides can be obtained and analyzed.

The use of neutral loss methods to target doubly charged phosphopeptides should function best on proteins that have been digested with trypsin as most of the peptides produced will contain two basic sites within the sequence. Some fraction of the peptides may contain other basic residues such as His, Lys-Pro, and Arg-Pro and appear as triply charged ions. In many instances most of the ion current is carried in the triply charged ion, but some will be present in the doubly charged ion. This ion will be detected in the neutral loss scan and used for tandem mass spectrometry if sufficient ion current is present. Additionally, this

(28) Huttner, W. B. *Trends Biochem. Sci.* **1987**, *12*, 361.
(29) Gibson, B. W.; Cohen, P. In *Methods in Enzymology*; McCloskey, J. A., Ed.; Academic Press: San Diego, CA, 1990; Vol. 193, pp 480–501.
(30) Covey, T.; Shushan, B.; Bonner, R.; Schroder, W.; Hucho, F. In *Methods in Protein Sequence Analysis*; Jornvall, H., Hoog, J. O., Gustavsson, A. M., Eds.; Birkhauser Press: Basel, 1991; pp 249–256.
(31) Ding, J. M.; Burkhart, W.; Kassel, D. B. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 94–98.

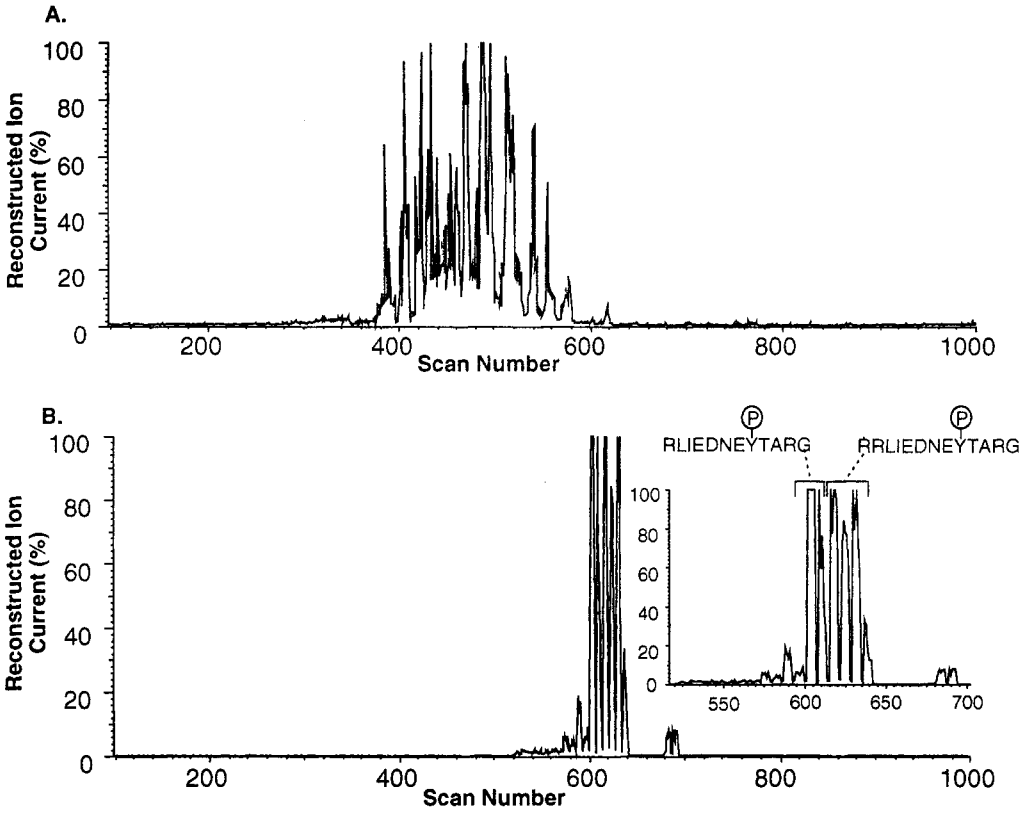

Figure 4. Comparison of the data acquired for two different analyses of a mixture of peptides generated by trypsin digestion of phosphorylase *b*, bovine serum albumin, ovalbumin, carbonic anhydrase, trypsin inhibitor, lysozyme, and RRLIEDNEY(PO_3H_2)TARG. (A) Plot of the reconstructed total ion current for an LC/MS analysis of the digested mixture of proteins. (B) Plot of the reconstructed total ion current for the neutral loss/product ion MS/MS analysis for the digested mixture of proteins. A blowup of the scan region 500–700 is inset into the figure. The peaks represented in this inset correspond to peptides having the sequence RLIEDNEY(PO_3H_2)TARG (mass 1516 Da) and RRLIEDNEY(PO_3H_2)-TARG (mass 1674 Da).

technique may not be completely specific since any peptide ion which produces a fragment ion that has an *m/z* value appearing 49 u below the precursor ion will trigger a product ion MS/MS analysis. By combining this experimental approach with our data analysis software, those peptides that do not contain the modification under study will be identified as such. This method can be extended to other types of modification, such as prenylation, which have been shown to exhibit a modification-specific neutral loss from the precursor ion.[32]

CONCLUSION

The objective of this research was to develop and test methods to match the tandem mass spectra of modified peptides to their corresponding sequence in the database. A long-term goal is to develop an approach for the construction of a general search that will identify a corresponding sequence in the database, whether or not it is modified. If the spectrum does represent a modified peptide, then the site and type of modification would be identified. With over 200 different types of modifications known, this represents an enormous search space over which to consider all possibilities. The likelihood is small that any one of a large majority of the potential modifications exists in a peptide; consequently, only the more common modifications, natural and artifactual, would need to be considered on a routine basis. The preliminary studies presented here demonstrate the efficacy of searches for simultaneous analysis of up to three different modifications.

The three examples of modified peptides discussed illustrates that the tandem mass spectra of modified peptides can be matched to their corresponding sequences in a protein database, even though no information relevant to modifications is contained in the character-based sequences. Any modified peptide that produces a spectrum containing specific information as to the location of the modification will be amenable to the described approach. Extremely labile modifications or bulky modifications such as glycosylation disrupt the fragmentation pattern of the peptide backbone and would most likely not be amenable to analysis with this method. The residual carbohydrate unit left after digestion with endoglycosidases should allow the identification of glycosylation sites by using the above search method.[33] Currently, several different types of modifications can be simultaneously considered in a single search, as was demonstrated in this paper. Extending this approach to simultaneously consider a larger set of the commonly occurring modifications in a single search will be the subject of further work. Combining selective MS/MS scan modes to trigger the acquisition of product ion tandem mass spectra into an analysis routine to target specific types of modifications will allow the use of more directed computer analyses.

The enormous efforts to sequence the human genome and the genomes of model organisms (*M. musculus, C. elegans, S. cerevisiae, D. melanogaster, E. coli,* etc.) will provide complete protein complements for these organisms. This information will

(32) Tuinman, A. A.; Thomas, D. A.; Cook, K. D.; Xue, C. B.; Naider, F.; Becker, J. M. *Anal. Biochem.* **1991**, *193*, 173–177.

(33) Tarantino, A. L.; Maley, F. *J. Biol. Chem.* **1974**, *249*, 811–817.

Table 5. Results of Searches by Using the Collision-Induced Dissociation Mass Spectra of Phosphorylated and Nonphosphorylated Peptides of Known Sequence under Search Conditions To Identify Phosphorylation of Ser, Thr, and Tyr[a]

description	sequence	OWL C_n	OWL ΔC_n	specie C_n	specie ΔC_n	prot C_n	prot ΔC_n
b. α-casein	VPQLEIVPN***S***AEER	1	0.041	1	0.041	1	0.625
b. α-casein	DIG***S******E******S***TEDQAMEDIK	1	0.073	1	0.073	1	0.073
b. opsin	DDEASTTV***S***KTETSQ	4*	0.010	4*	0.010	4*	0.010
b. opsin	DDEAS***T***TV***S***KTETSQ	1	0.001	1	0.001	1	0.001
b. opsin	DDEASTTV***S***KTET***S***QVAPA	*		*		1	0.007
h. phosphatidylcholine 2-acylhydrolase, +3	HIVSNDSSD***S***DDESHEPK	1	0.070	1	0.070	1	0.070
tyrosine phosphopeptide	EDVD***Y***VTLKH[b]	1	0.207	1	0.243	1	0.243
s.v. RNA polymerase α subunit, +3	TPATVPGTR***S***PPLNRY			*		1	0.001
tyrosine kinase substrate	RLIEDNE***Y***TARG[c]	1	0.099	1	0.099	1	0.099
tyrosine kinase substrate	RRLIEDNE***Y***TARG[c]	1	0.123	1	0.084	1	0.084
tyrosine phosphopeptide, +1	DDAQ***Y***SHLGG[d]	1	0.066	1	0.075	1	0.128
tyrosine phosphopeptide	DDAQ***Y***SHLGG[d]	18	0.001	6	0.099	1	0.161
tyrosine phosphopeptide	DRDDAQ***Y***SHLGG[d]	1	0.048	1	0.119	1	0.119
b. serum albumin	**DAIPENLPPLTADFAEDK**	1	0.090	1	0.164	1	0.598
c. cytochrome *c*	**GITWGEDTLMEYLENPK**	1	0.278	1	0.388	1	0.766
c. lysozyme	**FESNFNTQATNR**	1	0.251	1	0.333	1	0.763
b. carbonic anhydrase	**SFNVEYDDSQDK**	1	0.196	1	0.204	1	0.771
s.c. enolase	**SGETEDTFIADLVVGLR**	1	0.268	1	0.420	1	0.735
s.c. triose phosphate	**KPQVTVGAQNAY**	1	0.136	1	0.136	1	0.601
h. Ig heavy chain	**EVQLVESGGGLVQPGR**	1	0.120	1	0.276	1	0.570

[a] In the course of the search, each putative site of modification is considered as modified and unmodified. Modified amino acids are displayed in boldface, italic type. Complete amino acid sequences displayed in boldface, roman type are not phosphorylated, and served as negative controls. All spectra were acquired under electrospray ionization conditions. Charge states of the ion used in the MS/MS experiment are listed along with the protein description if it was different than +2. The column designation C_n refers to the ranking of the correct amino acid sequences by correlation parameter. The column ΔC_n is the difference between the top-scoring sequence and the next highest scoring sequence. These rankings are provided for searches of the OWL database, a modified human sequence database (specie), and the individual protein sequence (prot). An asterisk next to a rank indicates the correct sequence was identified in the top position, but the wrong modification site was found. The mass tolerance used in the searches was ±3 u. Fragment ion mass tolerance was ±1 u. Abbreviations used in the table are the following: b, bovine; c, chicken; p, porcine; h, human; s.v., Sendai virus; s.c., *Saccharomyces. cerevisiae.* [b] Mouse b-cell receptor cd-22 β precursor was used for the single protein search. [c] Added the peptide sequence to bovine casein to perform a single protein search. [d] Human t-cell receptor t3 δ chain was used for the single protein search.

facilitate the comparison and interpretation of biological studies conducted with these organisms. The genome sequencing projects, however, will not provide direct information pertaining to post translational modifications or their roles in the regulation of biological pathways. The information provided by tandem mass spectrometry can provide fragmentation patterns diagnostic of the presence and site(s) of modification. The ability to correlate the fragmentation patterns of modified peptides to sequences obtained through large-scale nucleotide sequencing projects will facilitate biological and functional studies.

ACKNOWLEDGMENT

This work was supported by the University of Washington's Research Royalty Fund. Partial support was derived from the National Science Foundation, Science and Technology Center Cooperative Agreement 8809710, and Digital Equipment Corp.

Received for review November 10, 1994. Accepted February 7, 1995.[⊗]

AC9411029

[⊗] Abstract published in *Advance ACS Abstracts,* March 15, 1995.