An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database

Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III
Department of Molecular Biotechnology, University of Washington, Seattle, Washington, USA

A method to correlate the uninterpreted tandem mass spectra of peptides produced under low energy (10–50 eV) collision conditions with amino acid sequences in the Genpept database has been developed. In this method the protein database is searched to identify linear amino acid sequences within a mass tolerance of ±1 u of the precursor ion molecular weight. A cross-correlation function is then used to provide a measurement of similarity between the mass-to-charge ratios for the fragment ions predicted from amino acid sequences obtained from the database and the fragment ions observed in the tandem mass spectrum. In general, a difference greater than 0.1 between the normalized cross-correlation functions of the first- and second-ranked search results indicates a successful match between sequence and spectrum. Searches of species-specific protein databases with tandem mass spectra acquired from peptides obtained from the enzymatically digested total proteins of *E. coli* and *S. cerevisiae* cells allowed matching of the spectra to amino acid sequences within proteins of these organisms. The approach described in this manuscript provides a convenient method to interpret tandem mass spectra with known sequences in a protein database. *(J Am Soc Mass Spectrom 1994, 5, 976–989)*

Amino acid sequence analysis is often the initial step in characterizing a newly isolated protein. Conventional sequencing strategies employ chemical reagents to remove one amino acid at a time from the amino terminus followed by isolation and analysis of the released amino acid derivative [1, 2]. Limitations in the chemical efficiency of the process prevents determination of the complete sequence of a protein from small quantities of sample. Partial sequence information, however, can be used to search a protein or nucleotide database to discover relationships to previously identified proteins or to determine if the protein sequence is novel [3, 4]. Although sequence information may have been determined previously, the context in which the protein is identified may be relevant to the biological process under study [5]. Another method to identify known protein sequences employs site-specific proteolysis followed by measurement of the mass-to-charge ratios of the peptides by mass spectrometry. The set of observed peptide mass-to-charge ratios is then used to search a protein database to find a set of peptide masses predicted from enzymatic digestion of each protein in the database [6–10]. Both chemical degradation and peptide mapping approaches require the use of fairly homogeneous samples to avoid ambiguity in assigning the amino acid sequence and establishing the origin of the set of measured peptide ions, respectively. These approaches are not well suited to the study of biological problems where complex mixtures of peptides may be isolated.

Innovations such as electrospray ionization have led to improved methods for interfacing liquid separation techniques to mass spectrometers [11]. Combining microcolumn liquid chromatography (LC) with tandem mass spectrometry improves the ability to manipulate, for sequence analysis, small quantities of peptides contained in complex mixtures [12]. Ions of a single mass-to-charge ratio can be selected for transmission through the first mass analyzer and into a cell for collision-induced dissociation (CID) even though that peptide may be introduced into the mass spectrometer as part of a mixture of peptides [12, 13]. Under CID conditions, peptides fragment to create patterns characteristic of a specific amino acid sequence. These patterns are reproducible and, in general, predictable. Experimental conditions that employ multiple low energy collisions (10–50 eV) cause peptides to fragment primarily at the amide bonds to produce a ladder of fragment ions [14]. Depending on the gas-phase basicity of the amino acids within the sequence, the charge can be retained upon fragmentation on the amino terminus of the ion to form an acylium ion (type-*b* ion, $NH_2-CHR_1-CO\cdots NHCHR_nCO^+$) or, by H rearrangement, on the carboxy terminus (type-*y* ion, $NH_2-CHR_n-CO\cdots NHCHR_1-CO_2H + H^+$) of

Address reprint requests to John R. Yates, III, FJ-20, Department of Molecular Biotechnology, University of Washington, Seattle, WA 98185.

1044-0305/94/$7.00

Received June 2, 1994
Revised June 28, 1994
Accepted June 29, 1994

the ion. The value of R depends on the amino acid and ranges from 1 to 131 u for Gly and Trp, respectively. A series of one type of fragment ion allows the amino acid sequence to be determined by the differences in the masses of adjacent sequence ions.

Interpretation of a CID spectrum for an unknown sequence proceeds by identifying a consecutive series of fragment ions whose differences correspond to residue masses for amino acids [15, 16]. This ion series can correspond to either the type-*b* or -*y* ion series or a combination of both. The putative amino acid sequence can be confirmed by identifying fragment ions from the other ion series if they are present in the spectrum. Ambiguities in the sequence can be resolved by obtaining a CID spectrum of the peptide modified by reaction with acetic anhydride or methanolic HCl [15, 16]. The tandem mass spectrometer sequencing approach has been successfully applied to proteins and peptides, and more recently to peptides in complex mixtures such as antigens released from major histocompatibility (MHC) molecules [17, 18].

Computer programs designed to aid in the interpretation of tandem mass spectra from peptides of unknown sequence have been developed [19–21], but have had limited general utility because there are an enormous number of sequence combinations that must be considered in the interpretation. More classical approaches to the identification of mass spectra have involved creation of mass spectral libraries and library searching routines [22]. The development of mass spectral libraries for peptides, however, is impractical because of the enormous number of spectra required to create a usable library. In many cases, tandem mass spectrometry is used to confirm a peptide sequence. This requires matching the predicted sequence ions of the peptide to the fragment ions contained in the spectrum. Computer programs have been developed to calculate the predicted fragment ions for a given amino acid sequence to facilitate the process [23–25].

In the present study, we demonstrate a computer algorithm that converts the character-based representations of amino acid sequences in a protein database to a fragmentation pattern that can be used to match fragment ions in a tandem mass spectrum. The algorithm initially identifies amino acid sequences in the database that match the measured mass of the peptide ion and predicts the fragment ions expected for each sequence. A score is calculated for each amino acid sequence by matching the predicted ions to the ions observed in the tandem mass spectrum. The highest scoring amino acid sequences are then reported. This approach is demonstrated by interpreting tandem mass spectrometer data from peptides of known amino acid sequence, peptides generated by proteolytic digestion of proteins from whole cell lysates, antigenic peptides released from class II MHC molecules, and peptides obtained from the heavy chain of glycoasparaginase. The potential use of this method for protein identification also is discussed.

Experimental

Peptide and Protein Sources

Peptides and proteins of known sequence were obtained from the following commercial sources. Myoglobin, *Physeter catadon* (sperm whale) (cat. no. 19290, lot no. 24040), U.S. Biochemical Corp. (Cleveland, OH). α-Casein, *Bos taurus* (cat. no. C-6780, lot no. 78F-9555), Angiotensin I, *Homo sapiens* (cat. no. A9650, lot no. 59F-58301), cytochrome *c*, *Columba livia* (pigeon) heart (cat. no. C-9261, lot no. 79F-7221), cytochrome *c*, *Saccharomyces cerevisiae* (cat. no. C-6913, lot no. 40H-7280), β-lactoglobulin A, *Bos taurus* (cat. no. L-7880, lot no. 88F-8095), and [Glu^1]-fibrinopeptide B, *Homo sapiens* (cat. no. F-3261) were obtained from Sigma Chemical Co. (St. Louis, MO). Sequencing grade trypsin (cat. no. 1047-841, lot no. 12676220-10) and chymotrypsin (cat. no. 1418 467, lot no. 13115720-01) were obtained from Boehringer Mannheim (Indianapolis, IN). Peptides were generated from intact proteins by digestion with the enzyme trypsin in 50-mM Tris-HCl, pH 8.6, for 4–8 h at 37 °C. Tandem mass spectra were acquired during liquid chromatography–tandem mass spectrometry (LC-MS/MS) analysis of the digestion products as described in succeeding text.

*Isolation of MHC Class II Peptides from HLA-DRB*0401 Cells*

Major histocompatibility molecules were isolated from Epstein Barr-virus-transformed B-cells (10^{10} cells) homozygous for HLA-DRB*0401 by using antibody affinity chromatography. Bound peptides were released by the method of Demotz et al. [26] and isolated by filtration through a Centricon 10 spin column. The peptide filtrate was fractionated by reverse-phase high performance liquid chromatography (HPLC) with a Vydac (Hesperia, CA) 2.1- × 250-mm C_{18} column. Solvent A was 0.07% trifluoroacetic acid (TFA) in water and solvent B was 0.06% TFA in 80/20 acetonitrile to water. The gradients were 0–45 min, 2–40% B; 45–60 min, 40–75% B; 60–65 min, 75–98% B. Separation was monitored at 220 nm. Flow rate was 200 μL/min, and 200-μL fractions were collected. HPLC fractions were concentrated to a final volume of 20–30 μL. One-microliter aliquots were removed for analysis by mass spectrometry and represented ~ 0.5–5 pmol of peptide. This mixture of peptides was provided by Professor Michael Davey, Oregon Health Sciences University.

Digestion of Glycoasparaginase

The heavy chain of glycoasparaginase [N^4-(β-N-acetylglucosaminyl)-*L*-asparaginase, EC 3.5.1.26] was isolated from human leukocytes as previously described and was provided by Dr. Ilkka Mononen, Kupio University Hospital, Finland [27]. The protein was reduced and alkylated prior to digestion [27].

Chymotrypsin was added in a ratio of 100/1 protein/enzyme and the mixture was incubated at 37 °C for 8 h. One-microliter aliquots were removed for analysis by mass spectrometry.

E. coli *Whole Cell Lysate Digestion*

An acetone precipitation of an *E. coli* cellular lysate was obtained from Sigma Chemical Co. (St. Louis, MO; cat. no. E-0125, lot no. 42C6900, strain ATCC 11246). Proteins from the *E. coli* lysate (1 mg) were digested in 500 μL of 50-mM ammonium bicarbonate that contained 10-mM Ca^{++}, pH 8.6. The peptides produced from this digest were analyzed directly by liquid chromatography–mass spectroscopy (LC-MS) and liquid chromatography–tandem mass spectrometry (LC-MS/MS) without further purification. Approximately 1/500 of the mixture was used for each analysis by mass spectrometry.

S. cerevisiae *Whole Cell Lysate Digestion*

S. cerevisiae cells (α $Trp_{1\text{-}ochre}$ ρ^0) were grown to a density of $\approx 10^7$ cells/mL in 500 mL of Yeast extract-Peptone-Dextrose media. The cells were collected by centrifugation and washed. The ≈ 4.7 mg (wet weight) of cells were lysed by using the procedure of Kolodziej and Young [28]. The lysis buffer contained 50-mM Tris-HCl, pH 8.0, 0.1-M NaCl, 1-mM ethylenediamine tetracetic acid (EDTA), 1-mM NaF, 1-mM NaN_3, 1% dimethylsulfoxide (DMSO) (v/v), 30-μg/mL phenyl methyl sulfonyl fluoride, 1.5-μg/mL leupeptin, and 3-μg/mL pepstatin. After lysis the protease inhibitors were removed by batch concentration of the solution on a Centriprep 10. The retentate was washed with 5 mL of 50-mM Tris-HCl, pH 8.0, 0.1-M NaCl, 1-mM EDTA, 1-mM NaF, 1-mM NaN_3, and 1% DMSO (v/v) and reconcentrated. The final volume of solution after removal from the Centriprep 10 was 9 mL. An aliquot of 200 μL was removed, 10 μg of trypsin was added, and the mixture was incubated for 41 h at 37 °C. Aliquots of 1–3 μL of the digested material were injected onto the microcapillary column by pneumatic injection for analysis by mass spectrometry.

Electrospray and Tandem Mass Spectrometry

Analysis of the resulting peptide mixtures was performed by LC-MS and LC-MS/MS on a Finnigan MAT (San Jose, CA) TSQ70 equipped with atmospheric pressure ionization as previously described [29]. Briefly, 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 2–6 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 10–50 eV (E_{lab}). The second mass analyzer was scanned at 500 u/s to record the mass-to-charge ratio of the fragment ions. Peak widths in the second mass analyzer ranged from 1.5 to 2.5 u.

Computer Analysis of Tandem Mass Spectrometry Data

All computer algorithms were written in the C programming language under the UNIX operating system. The Genpept database was obtained as an ASCII text file in the FASTA format from the National Center for Biotechnology Information by anonymous ftp (NCBI-Genbank Release 77.0; release date 06/15/93; 74,938 proteins containing 29,127,050 amino acids). This database contains protein sequences translated from nucleotide sequences. A search of the entire Genpept database required approximately 10–15 min on a HP9000 minicomputer (Hewlett-Packard, Palo Alto, CA). The amino acid masses used to calculate the mass of peptides were based on average chemical masses. To account for changes in mass created by chemical modification, the mass of the appropriate amino acid can be changed. Species-specific databases were created by combining protein sequences derived from *Homo sapiens* (16,340 sequences), *E. coli* (6637 sequences), and *S. cerevisiae* (4285 sequences) from the Genpept and PIR database (National Biomedical Research Foundation, Washington, DC). Sequences of [Glu^1]–fibrinopeptide and the peptide CRGDSY were added to the *Homo sapiens* database. To allow a more representative assessment of searches with peptides derived from β-lactoglobulin (*Bos taurus*) and α-casein (*Bos taurus*), these protein sequences were added to the human database. The sequence of trypsin (*Bos taurus*) was added to the *E. coli* database to screen for autolysis products. A search of a species-specific database requires approximately 3–5 min on the HP9000 minicomputer.

Results and Discussion

The objectives of the research reported here are twofold. The first is to develop a computer algorithm to search a protein database by using uninterpreted tandem mass spectra from a single peptide. No additional information about the peptide beyond the mass of the peptide is used. In this manner unrelated peptides generated with proteases of unknown specificity can be analyzed, and if the amino acid sequence is in the protein database, an interpretation of the tandem mass spectrum can be performed. The second objective is to test the method on representative peptides of known and unknown sequence (unknown to the researcher) to determine the effectiveness of tandem mass spectrometry data to identify the origin of the sequence.

Database Searching with Tandem Mass Spectrometry Data

Figure 1 illustrates the approach. The analysis strategy begins with computer reduction of the tandem mass spectrometry data (Step 1). Amino acid sequences are then identified in a protein database for comparison to the processed tandem mass spectrometry data by matching the molecular weight of the peptide to a linear sequence (Step 2). The predicted fragment ions of the sequences derived from the database are compared to the mass spectral information to produce a ranked list of the top 500 best fit sequences (Step 3). These 500 sequences are then subjected to a correlation-based analysis to generate a final score and ranking of the sequences (Step 4).

Step 1: Tandem mass spectrometry data reduction. The first element of the program involves pre-search analysis and reduction of the tandem mass spectrometry data. Fragment ion mass-to-charge ratios are converted to the rounded nominal values (nearest integer). We realize an increase in computational speed by a factor of ≈ 2 compared to the use of fractional masses. A 10-u window around the precursor ion is removed to eliminate possible matches of a predicted fragment ion to the mass-to-charge ratio of the precursor ion. An erroneous match to the precursor ion can result in an artificially inflated score, which yields poor sequence correspondence. To eliminate noise from the spectrum and to reduce the number of ions to be considered, all but the 200 most abundant ions are removed and the remaining ions are renormalized to 100. The abundances of fragment ions within ± 1 u of each other are equalized to the higher value. We evaluated the effect on search accuracy by using the most abundant 200, 300, 400, and 500 ions and found no significant changes in search accuracy. Low mass-to-charge ratio ions that correspond to the structure $+ NH_2{=}CH(R)$ (immonium ions) are diagnostic features present in the tandem mass spectra of peptides and convey information about the amino acid content of the peptide [14]. Immonium ions are frequently observed when the amino acids His, Met, Trp, Tyr, or Phe are present in the spectrum. The presence of these ions in the mass spectrum is noted during this preprocessing step. The graphically displayed output of processed tandem mass spectrometry data for the peptide DLRSWTAADTAAQISQ is shown in Figure 2a.

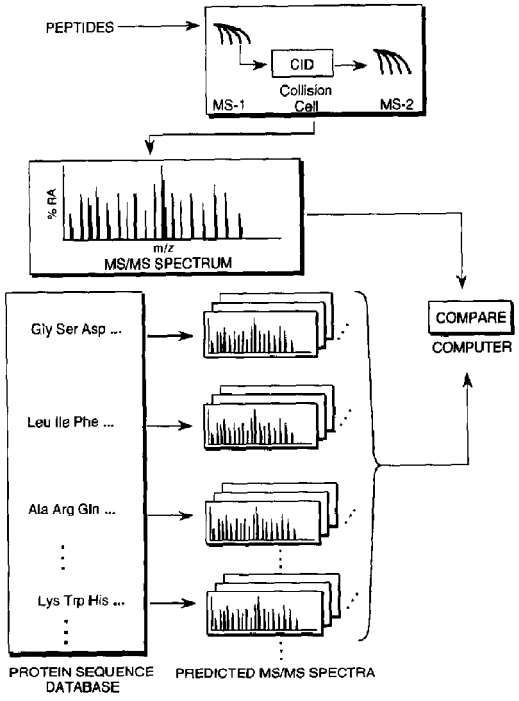

Figure 1. Flow chart that depicts the alogrithm for searching protein databases with tandem mass spectrometry data.

Step 2: Search method. To match a tandem mass spectrum to a sequence in the database, protein sequences are retrieved and scanned to find linear combinations of amino acids, proceeding from the N to the C terminus, that match the mass of the peptide. The identity of the terminating amino acid, which would be indicative of protease specificity, is not considered in the selection of sequences. Masses for the amino acid sequences are summed until the mass of the peptide falls within a specified mass tolerance. In general, this tolerance is set at $\pm 0.05\%$ or a minimum of 1 u, although the value can be adjusted. As the mass tolerance increases, the number of matched sequences increases. We have found in some cases the mass tolerance can be increased to ± 3 u without a significant decrease in accuracy. In a search of both the *S. cerevisiae* and the Genpept databases, the amino acid sequences listed in Table 1 were identified by using a mass tolerance of ± 3 u. Narrowing the mass window to ± 1 u resulted in identification of the same top ranking amino acid sequence. The mass-to-charge ratio values for the fragment ions predicted for each amino acid sequence are calculated in the manner

$$b_n = \sum a_n + 1$$
$$y_n = \mathrm{MW} - \sum a_n \qquad (1)$$

where a_n is the mass of the amino acid, b_n is a type-*b* ion, and y_n is a type-*y* ion. These values are used in the third step to compare sequences to the mass-to-charge ratio list derived from the tandem mass spectrum.

Chemical modifications to amino acids can be considered in this search by changing the amino acid masses used to calculate the masses of the peptides. The modified amino acid is then considered at every occurrence in the sequence. Modifications such as gly-

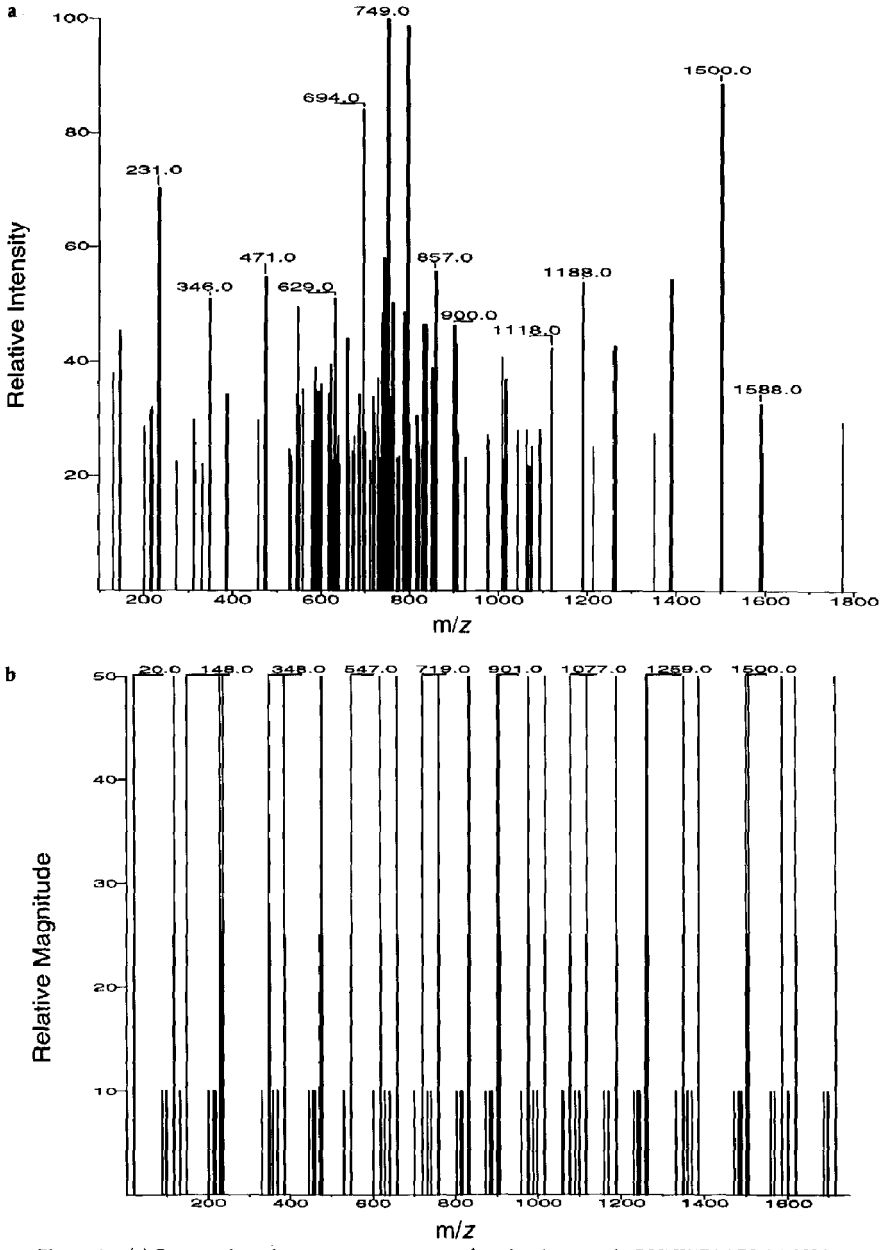

Figure 2. (a) Processed tandem mass spectrometry data for the peptide DLRSWTAADTAAQISQ. The tandem mass spectrum was obtained from the doubly charged ion $[M + 2H]^{+2}$ at m/z 868. The 200 most abundant ions are shown in the graphical display. A 10-u window around the precursor ion at m/z 868 has been removed. The abundances of fragment ions within 1 u of each other are equalized to the higher value. (b) A graphical display of the reconstructed data used for the correlation analysis for the amino acid sequence DLRSWTAADTAAQISQ. A magnitude of 50.0 is assigned to the predicted mass-to-charge ratio values for the type-*b* and -*y* ions and ions with mass-to-charge ratios ± 1 u from the type-*b* and -*y* ions are assigned a value of 25.0. The neutral losses of water and ammonia and the *a*-type ions are assigned values of 10.0. (c) A graphical display of the processed experimental tandem mass spectrum used in the correlation analysis. A 10-u region around the precursor ion is removed. The spectrum is then divided into 10 equal sections and the ion abundances in each section are normalized to 50.0. (d) A graphical display of the result of the cross-correlation function for the spectrum displayed in Figure 2b and c. The final score attributed to the analysis is the value at $\tau \approx 0$ minus the mean of the cross-correlation function over the range $-75 < \tau < 75$.

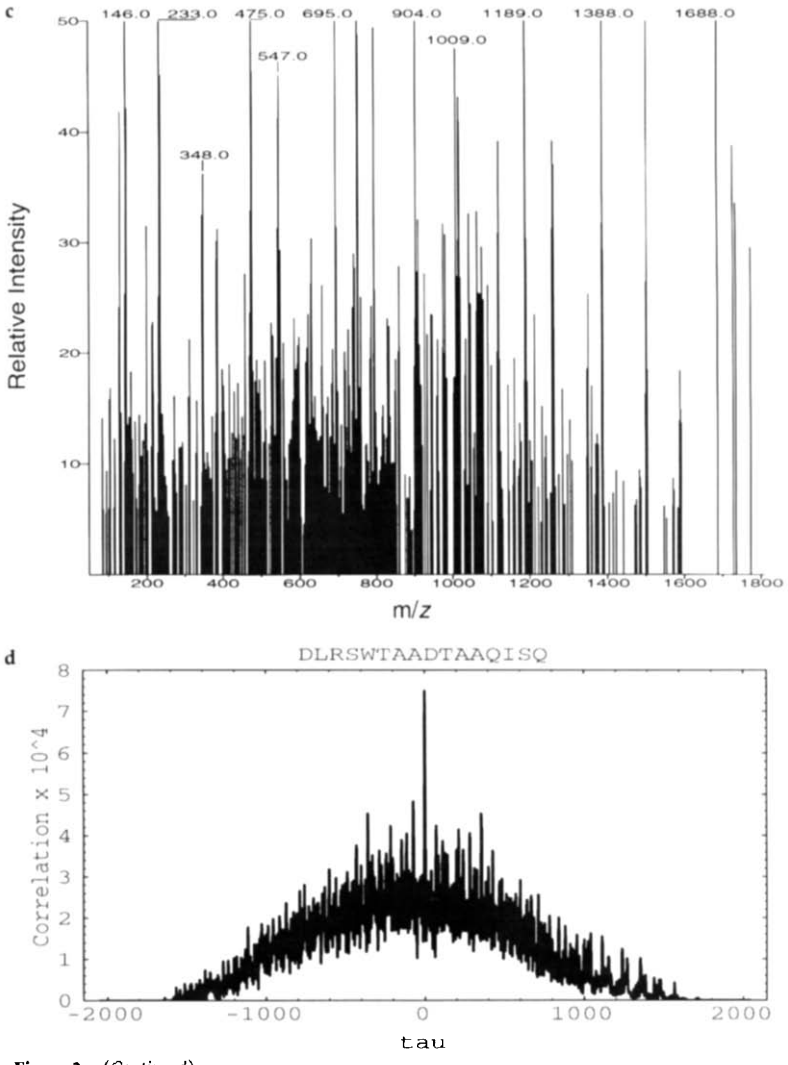

Figure 2. *(Continued)*

cosylation cannot be considered as yet because they frequently affect the fragmentation patterns of the peptide in the CID process. A modification such as phosphorylation may not be present at every occurrence of a Thr, Ser, or Tyr, and one or more of these amino acids may exist in the sequence. Each possibility would need to be considered, and this is computationally more difficult. At present the search algorithm only considers mass changes that would be present at every occurrence of the modification site.

Step 3: Scoring method. Once an amino acid sequence fits the defined mass tolerance, the sequence is evaluated by using several different criteria. First, the number (n_i) of predicted fragment ions that match ions observed in the spectrum within ± 1 u and their abundances (i_m) are summed. The continuity of an ion series is considered by incrementing a component of the score (β) for each consecutive fragment ion matched. If an immonium ion for the amino acids His, Tyr, Trp, Met, and Phe is present in the spectrum along with the associated amino acid, then an additional component of the score (ρ) is incremented. If the amino acid is not present in the sequence, then ρ is decreased. The values used for β and ρ are 0.075 and 0.15, respectively. The total number of predicted sequence ions is also noted (n_t). A score is calculated for each amino acid sequence by using the relationship

$$S_p = \left(\sum i_m\right) n_i (1 + \beta)(1 + \rho)/n_t \qquad (2)$$

A total of 57 tandem mass spectra from peptides of different sequences were analyzed with this program

Table 1. Results of searches of protein databases by using the tandem mass spectra of peptides obtained by trypsin digestion of an *S. cerevisiae* lysate with mass tolerances of ±1 and ±3 u[a]

		C_n			
		±3 u		±1 u	
Source of peptide	Amino acid sequence	S	G	S	G
Enolase gene	EEALDLIVDAIK	1	1	1	1
Enolase gene	NPTVEVELTTEK	1	1	1	1
Enolase gene	DPFAEDDWEAWSH	1	1	1	1
Pyruvate kinase	LPGTDVDLPALSEK	1	1	1	1
Hexokinase PI gene	IEDDPFVFLEDTDDIFQK	1	1	1	1
Hypusine cont protein HP2	APEGELGDSLQTAFDEGK	1	1	1	1
Chromosome III complete DNA seq.	IPAGWQGLDNGPESR	1	1	1	1
Chromosome III complete DNA seq.	TGGGASLELLEGK	1	1	1	1
BMH1 gene product	QAFDDAIAELDTLSEESYK	1	1	1	1

[a] All tandem mass spectra were acquired from the doubly charged ions of peptides created by electrospray ionization. The columns designated C_n refer to the ranking of the identified sequences by the correlation parameter. These rankings are provided for searches of yeast sequence database (S) and the Genpept database (G).

(Tables 2 and 3). The rankings of the spectra for peptides of known sequence via formula 2 are given in Table 2 under the heading S_p. The foregoing scoring method provides a reasonably accurate method to assess sequence matches. Reproducibility of the searches for data acquired at different times is good and this is illustrated in Table 2 with several spectra acquired for the same peptide (numbers 18–22, 26, and 27). In some examples, the tandem mass spectra obtained from a different charge state are illustrated. In one case the tandem mass spectrum of the +3 charge state of the peptide VYVEELKPTPEGDLEILLQK resulted in a poor identification in the Genpept search (rank 3). In this particular tandem mass spectrum, some fragment ions exist as doubly charged ions, which are not considered in the scoring routine, and poor identification results. The accuracy of the search for small peptides is generally poor and is more pronounced in a search of a large database. For example, the tandem mass spectra for peptides in rows 35–37 of Table 2 failed to rank in the top 500 sequences in a search of the Genpept database. To improve the scoring of small peptides, at least those appearing in the top 500 sequences, and the assessment of false positives, an additional independent scoring procedure was used to evaluate the top 500 sequences generated in the search.

Step 4: Cross-correlation analysis. A cross-correlation analysis is used to compare the top 500 identified amino acid sequences obtained from the search with the experimental data. This method has been employed to provide a measurement of spectral similarity for infrared library searches [30] and has been reviewed at length by Owen [31] for use in mass spectral analysis. To compare the amino acid sequences obtained from the database search to the acquired tandem mass spectrum by using a correlation analysis, a "spectrum" is reconstructed from the character-based amino acid sequences. The "spectrum" contains values for the predicted mass-to-charge ratio of fragment ions of the given amino acid sequence as well as a magnitude component. All the factors in the CID process that contribute to peptide fragmentation are not fully understood; consequently a priori prediction of fragment ion abundances for a given amino acid sequence is impossible at the present time with any degree of accuracy. A magnitude component is assigned to the predicted mass-to-charge ratio values of the fragment ions by using an empirical knowledge of the appearance of tandem mass spectra for peptides and the constraints of correlation analysis. The relative abundances of type-*b* and -*y* ions cannot be predicted, but we know from experience that sequential losses from these ions are generally less facile at the collision energies used to generate low energy tandem mass spectra of peptides. All values that represent the mass-to-charge ratios of fragment ions of type-*b* and -*y* ions are assigned a magnitude of 50.0. A magnitude of 25.0 is assigned to mass-to-charge ratios within ±1 of the *b* or *y* ion values. The neutral losses of ammonia, water, and carbon monoxide (type-*a* ions) and the mass-to-charge ratios ±1 are assigned a magnitude of 10.0. By increasing the peak width of the reconstructed tandem mass spectrum, we increase its similarity to the experimental tandem mass spectrum, thereby improving the coherence measured by the cross-correlation function. Figure 2b is a graphically displayed "spectrum" for the amino acid sequence DLRSWTAADTAAQISQ.

To compare the reconstructed "spectrum" to the original tandem spectrum with a cross-correlation function, the original spectrum is processed by first removing the mass-to-charge ratio for the precursor ion and then dividing the spectrum into 10 equal regions and normalizing the ions in each region to a value of 50.0 (Figure 2c). By eliminating the precursor ion, a major feature of the tandem mass spectrum is

Table 2. Results of searches of the protein database by using the tandem mass spectra from peptides of known sequence[a]

No.	$n+$	Db	Source of peptide	Sequence of peptide	S		G	
					C_n	S_p	C_n	S_p
1	+2	H	β-lactoglobulin, bovine, 92–101	VLVLDTDYKK	2	8	—	—
2	+2	H	β-lactoglobulin, bovine, 125–138	TPEVDDEALEKFDK	1	1	1	1
3	+3	H	β-lactoglobulin, bovine, 41–60	VYVEELKPTPEGDLEILLQK	1	1	13	3
4	+2	H	α-s1-casein, bovine, 91–100	YLGYLEQLLR	1	1	1	1
5	+2	H	α-s1-casein, bovine, 23–34	FFVAPFPQVFGK	2	2	2	9
6	+2	H	α-s1-casein, bovine, 8–22	HQGLPQEVLNENLLR	1	1	1	1
7	+1	H	aspartylglucosaminidase, 1–6	SSPLPL	2	50	1	135
8	+1	H	aspartylglucosaminidase, 102–112	TLLVGESATTF	1	1	1	1
9	+2	H	aspartylglucosaminidase, 146–155	RNVIPDPSKY	1	1	1	1
10	+2	H	aspartylglucosaminidase, 136–145	LARNCQPNYW*	2	2	19	11
11	+2	H	aspartylglucosaminidase, 113–135	AQSMGFINEDLSTSASQALHSDW	1	1	1	1
12	+2	H	aspartylglucosaminidase, 207–223	KIHGRVGDSPIPGAYGAY	1	1	1	1
13	+1	H	aspartyglucosaminidase, 75–82	DVGAVGDL	1	2	1	11
14	+2	H	aspartylglucosaminidase, 129–135	INEDLSTSASQASQALHSDW	1	1	1	1
15	+2	H	aspartylglucosaminidase, 113–131	AQSMGFINEDLSTSASQAL	1	1	1	1
16	+2	H	aspartylglucosaminidase, 53–74	GGSPDELGETTLDAMIMGTTM	1	1	1	1
17	+2	H	aspartylglucosaminidase, 309–323	NSEKNQPTEEKVDCI*	1	1	1	5
18	+3	H	angiotensin I	DRVYIHPFHL	1	1	1	1
19	+3	H	angiotensin I	DRVYIHPFHL	1	1	1	1
20	+3	H	angiotensin I	DRVYIHPFHL	1	1	1	1
21	+2	H	angiotensin I	DRVYIHPFHL	1	11	3	111
22	+2	H	angiotensin I	DRVYIHPFHL	1	10	1	60
23	+1	H	angiotensin I	DRVYIHPFHL	2	9	4	43
24	+2	H	myoglobin, sperm whale, 17–31	VEADVAGHGQDILIR	1	1	1	1
25	+2	H	myoglobin, sperm whale, 64–77	HGVTVLTALGAILK	1	1	1	1
26	+2	H	[Glu^1]-fibrinopeptide	EDVNDNEEGFFSAR	1	2	1	2
27	+2	H	[Glu^1]-fibrinopeptide	EGVNDNEEGFFSAR	1	2	1	2
28	+2	H	synthetic	CRGDSY	2	29	5	204
29	+1	H	synthetic	CRGDSY	3	53	10	238
30	+2	H	synthetic	IPTSLALLCCVRSANA	1	1	2	1
31	+2	H	synthetic	YPHFMPTNL	1	2	5	3
32	+2	H	cytochrome *c*, pigeon, 28–38	TGPNLHGLFGR	1	1	—	—
33	+2	H	cytochrome *c*, pigeon, 40–53	TGQAEGFSYTDANK	1	1	1	4
34	+2	H	cytochrome *c*, pigeon, 56–72	GITWGEDTLMEYLENPK	1	1	1	1
35	+1	Y	cytochrome *c*, *S. cerevisiae*, 2–5	TEFK	6	93	—	—
36	+1	Y	cytochrome *c*, *S. cerevisiae*, 86–93	MAFGGLK	1	95	—	—
37	+2	Y	cytochrome *c*, *S. cerevisiae*, 12–17	GATLFK	1	55	—	—
38	+1	Y	cytochrome *c*, *S. cerevisiae*, 11–17	KGATLFK	1	1	1	2
39	+2	Y	cytochrome *c*, *S. cerevisiae*, 96–105	DRNDLITYLK	1	4	2	42
40	+1	Y	cytochrome *c*, *S. cerevisiae*, 45–52	HSGQAEGY	1	1	1	1
41	+2	Y	cytochrome *c*, *S. cerevisiae*, 45–60	HSGQAEGYSYTDANIK	1	1	1	1

[a] Peptides derived by proteolytic digestion of proteins with sequences known to the investigators are listed with the residue numbers of the protein sequence. All spectra were acquired under electrospray ionization conditions. Each entry into the table represents a separate acquisition of the data. The column labeled "No." references the position of each peptide to Figure 3. Charge states of the ion used in the tandem mass spectrometry experiment are listed under the column n^+. The column designation Db refers to the specific species database used in the search: H = human; Y = yeast. Under the heading "Source of peptide," the residue numbers of the protein sequence from which the peptides were derived is listed. The column labeled "Sequence of peptide" is the known amino acid sequence of the peptide used to acquire the tandem mass spectrum. The designations S_p and C_n refers to rankings of the correct amino acid sequences by score and correlation parameter, respectively. These rankings are provided for searches of species-specific databases (S) and the Genpept database (G). The mass tolerance used in the searches was ±0.05% or a minimum of ±1 u. Fragment ion mass tolerance was ±1 u. The amino acids marked with an asterisk used a mass of 161 u for carboxymethylated Cys.

Table 3. Results of searches of protein databases by using the tandem mass spectra of peptides obtained from whole cell lysates of *E. coli* and *S. cerevisiae*, and from peptides bound to class II MHC molecules[a]

No.	Db	Protein identified	Sequence identified	S ΔC_n	S C_n	G C_n
42	E	trypsin, bovine	SSGTSYPDVLK	0.035	1	1
43	E	trypsin, bovine	TLNNDIMLIK	0.407	1	1
44	E	trypsin, bovine	SIVHPSYNSNTLNNDIMLIK	0.012	4	17
45	E	trypsin, bovine	VASISLPTSCASAGTQCLISGWGNTK	0.012	1	7
46	E	trypsin, bovine	VASISLPTSCASAGTQCLISGWGNTK	0.367	1	1
47	E	tufA gene product, *E. coli*	VGEEVEIVGIK	0.278	1	1
48	E	tufA gene product, *E. coli*	VTLIHPIAMDDGLR	0.037	1	2
49	E	ORF-D, *E. coli*	GGDTVTLNETDLTQIPK	0.281	1	1
50	E	GAD α protein, *E. coli*	GWQVPAFTLGGEATDIVVMR	0.494	1	1
51	H	mRNA for HLA-DR antigens	MATPLLMQALP	0.017	5	—
52	H	mRNA for HLA-DR antigens	MATPLLMQALPM	0.098	2	9
53	H	mRNA for HLA-DR antigens	KPPKPVSKMR*	0.015	7	26
54	H	mRNA for HLA-DR antigens	PKPPKPVSKMR*	0.163	1	2
55	H	lymphocyte antigen	DLRSWTAADTAAQISQ	0.347	1	1
56	H	MHC Class I HLA	DLRSWTAADTAAQITQ	0.152	2	2
57	Y	enolase gene, *Sc*	EEALDLIVDAIK	0.305	1	1
58	Y	enolase gene, *Sc*	NPTVEVELTTEK	0.292	1	1
59	Y	enolase gene, *Sc*	DPFAEDDWEAWSH	0.416	1	1
60	Y	pyruvate kinase, *Sc*	LPGTDVDLPALSEK	0.325	1	1
61	Y	hexokinase PI gene, *Sc*	IEDDPFVFLEDTDDIFQK	0.312	1	1
62	Y	hypusine protein HP2, *Sc*	APEGELGDSLQTAFDEGK	0.476	1	1
63	Y	chromosome III compl. DNA, *Sc*	IPAGWQGLDNGPESR	0.314	1	1
64	Y	chromosome III compl. DNA, *Sc*	TGGGASLELLEGK	0.310	1	1
65	Y	BMH1 gene product, *Sc*	QAFDDAIAELDTLSEESYK	0.321	1	1

[a] All spectra were acquired using electrospray ionization conditions. Each entry in the table represents a separate acquisition of the data. The column labeled "No." references the position of each peptide to Figure 3. The column designation Db refers to the specific species database used in the search: E-*E. coli*; H = human; Y = yeast. The column labeled "Protein identified" lists the protein from which the peptide was derived. Under the heading "Sequence identified," the amino acid sequence represented by the tandem mass spectrum is listed. The column designation ΔC_n refers to difference of the correlation parameters for the first- and second-ranked sequences from the species-specific database search. The rankings of the correct sequence following a search of the species-specific and Genpept databases, as determined by inspection of the tandem mass spectrum, are listed in the columns designated by the letters S and G, respectively. The mass tolerance used in the search was ±0.05% or a minimum of ±1 u. Fragment ion mass tolerance was ±1 u. Sequences marked with an asterisk used a mass for Lys increased by 42 u. All tandem mass spectra were acquired on doubly charged ions with the exception of number 49, which was triply charged. The abbreviation *Sc* refers to *S. cerevisiae*.

removed, and the similarity between the fragment ion patterns is the major contributor to the τ value calculated by the cross-correlation function.

A cross-correlation between two continuous signals $x(t)$ and $y(t)$ can be calculated by using the formula

$$C_{xy} = \int_{-\infty}^{+\infty} x(t)\, y(t + \tau)\, dt \qquad (3)$$

where τ is a displacement value between the two signals. The correlation function measures the coherence of two signals by, in effect, translating one signal across the other. The displacement value τ is the amount by which the signal is offset during the translation and is varied over a range of values. If two signals are the same, the correlation function should maximize at $\tau = 0$, where there is no offset between the signals. The reconstructed "spectrum" x_i and the experimental "spectrum" y_i represent discrete input signals and use the form of the cross-correlation

$$R_\tau = \sum_{i=0}^{n-1} x[i]\, y[i + \tau] \qquad (4)$$

where τ is again a displacement value. A discrete correlation of two real functions x and y is one member of the discrete Fourier transform pair

$$R_\tau \leftrightarrow X_\tau Y_\tau^* \qquad (5)$$

where X and Y are the discrete Fourier transforms of x and y and the Y* denotes complex conjugation. The cross-correlations are computed by fast Fourier transformation of the two 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. A graphical output of the cross-correlation of the two "spectra," Figure 2b and c, is shown in Figure 2d. 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 $-75 < \tau < 75$ [30]. The scores are normalized to 1.0 and termed C_n.

The differences between the normalized cross-correlation parameter of the first- and second-ranked amino acid sequences have shown a trend useful for distinguishing correct identifications from false positives. Figure 3 shows a plot of the difference between the normalized correlation parameter of the first and second answer for searches of species-specific databases. When a difference greater than 0.1 is observed between the normalized correlation parameter of the first- and the second-ranked sequences, then the first answer is usually correct. Figure 3 shows a plot of the cross-correlation differences for every tandem mass spectrum analyzed against the species database. The same trend is observed with the Genpept database (data not shown). A related trend was observed when correlation analysis of infrared library spectra [30] was used. Rankings by both scoring procedures are included in Table 2. The addition of the cross-correlation parameter improves the scoring for many of the small peptides. At present, searches of species-specific databases are more accurate than searches of the whole Genpept database. If no sequence match is identified for a tandem mass spectrum in the species database, then a search of the entire database would find a match only if there is a sequence and mass-conserved domain in a protein from another species.

The success of a search is also sensitive to the collision energy conditions used to obtain the tandem mass spectrum. Searches that use spectra from [Glu^1]-fibrinopeptide (EGVNDNEEGFFSAR) obtained under energy conditions varying from 10 to 20 eV are shown in Table 4. At laboratory collision energies of 10–12.5 eV, the correct sequence is considerably lower in the rankings. At energies closer to those typically used for a peptide of this size, the rankings are closer to the first position (Table 4). In several cases (rows 2, 3, and 5 of Table 4) the number one ranked sequence was QGVNDNEEGFFSAR, a sequence from the fibrinogen β-chain precursor. This sequence is identified because the mass of this peptide sequence is within the mass tolerance used for the search and the predominant set of sequence ions observed for [Glu^1]-fibrinopeptide is the y-type ion. Consequently, the single amino acid difference at the N terminus of the fibrinogen β-chain peptide does not change the mass-to-charge ratio values for the y ions of this sequence and this results in a high score.

Application of the Search Program to Identify Class II MHC Peptides

This data analysis approach should be useful in a variety of different sequencing applications. One specific application is the analysis of peptides presented by class II major histocompatibility molecules. Many of

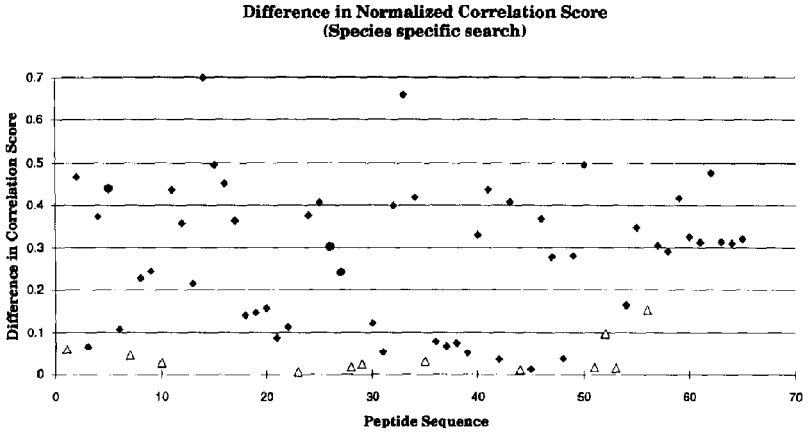

Figure 3. A plot of the differences calculated for the cross-correlation parameters of the first and second answers. The squares indicate the differences when the correct answer is ranked number one. The triangles indicate the differences for the first- and second-ranked sequences when the correct answer is not in the number one ranked position. The numbers on the x-axis reference the results from Tables 2 and 3.

Table 4. Results of database searches by using tandem mass spectra of [Glu^1]-fibrinopeptide acquired at different collision energies[a]

Description	Collision energy (E_{lab}) (eV)	Sequence	S		G	
			C_n	S_p	C_n	S_p
[Glu^1]-fibrinopeptide	10.0	EGVNDNEEGFFSAR	2	21	10	154
[Glu^1]-fibrinopeptide	12.5	EGVNDNEEGFFSAR	2	2	6	2
[Glu^1]-fibrinopeptide	15.0	EGVNDNEEGFFSAR	2	2	2	4
[Glu^1]-fibrinopeptide	17.5	EGVNDNEEGFFSAR	1	2	1	2
[Glu^1]-fibrinopeptide	20.0	EGVNDNEEGFFSAR	2	2	2	1

[a] All tandem mass spectra were acquired under electrospray ionization conditions by using the doubly charged ion. The column designations C_n and S_p refer to rankings of the identified sequences by the correlation parameter and score, respectively. These rankings are provided for both a human sequence database (S) and the Genpept database (G). The mass tolerance used in the searches was ±0.05% or a minimum of ±1 u. Fragment ion mass tolerance was ±1 u.

the peptides presented by class II MHC proteins are derived from exogenous and endogenous serum proteins [32] or from proteins known to be present in the cellular processing compartment where complexation occurs [33]. Many of these peptides are derived from proteins or genes of known sequence, and rapid prescreening of the experimental data should speed up the identification process of those antigens of known sequence. A representative example for the interpretation of a peptide released from MHC molecules is given.

An analytical HPLC separation of the peptides released from HLA DRB*0401 class II MHC molecules produced 50 fractions that were collected at a rate of one per minute. Approximately 30 of the fractions showed UV absorbance at 220 nm, which indicates the presence of peptides. An aliquot representing 1/30 of the material was removed from one of the collected fractions and analyzed by microcolumn LC-MS. Peptide ion signals with signal-to-noise ratio greater than 3/1 were observed for at least 30 peptides from this one analytical HPLC fraction. A tandem mass spectrum of a doubly charged peptide ion at m/z 868, $[M + 2H]^{+2}$, was obtained and used in the database

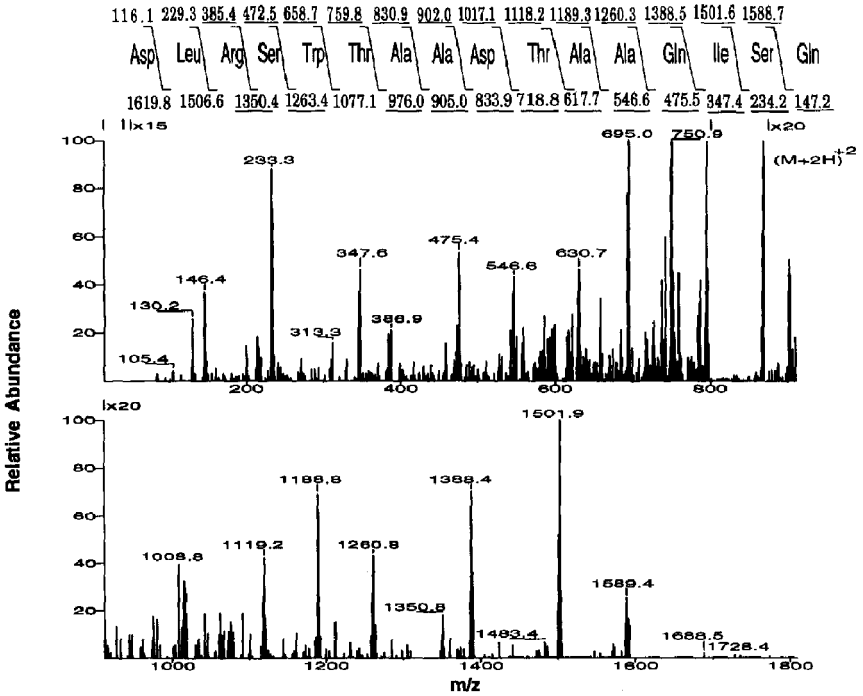

Figure 4. A tandem mass spectrum of an $[M + 2H]^{+2}$ ion at m/z 868. The fragment ions are labeled with their computer assigned mass-to-charge ratios. The molecular weight of the peptide is 1734 u. The calculated sequence ions are shown on the amino acid sequence. Ions above the sequence are type-*b* ions and those below are type-*y* ions. The underlined values are the sequence ions identified during the search of the Genpept database with the processed data in Figure 2a.

searching algorithm. The tandem mass spectrum is shown in Figure 4. A search of the Genpept database identified a total of 384,398 different amino acid sequences of a molecular weight of 1734 ±1 u.

Two amino acid sequences were identified in the search as best fitting the spectrum: DLRSWTAADTAAQISQ and DLRSWTAADTAAQISK. Both sequences produced the same cross-correlation parameter and score. The sequence DLRSWTAADTAAQISQ originates in the *Homo sapiens* class I histocompatibility antigen (residues 145–160) and the class I histocompatibility antigen B-48 B*4801 α-chain precursor (residues 153–168). The sequence DLRSWTAADTAAQISK appears in nine immunologically related proteins from *Homo sapiens:*

HLA G locus gene product residues 30–45
HLA G mRNA for nonclassical class I transplantation antigen residues 129–144
In residues 153–168 of the following HLA class I histocompatibility antigens:
G α-chain precursor
class I histocompatibility antigen HLA G α-chain precursor
MHC class I histocompatibility antigen HLA-6.0 precursor
HLA g mRNA for nonclassical class I transplantation antigen
β2 microglobulin
MHC class I HLA-6.09 gene
The lymphocyte antigen (residues 153–168)

Fragmentation information in the low energy tandem mass spectra does not allow differentiation of Gln or Lys. Consequently the correct sequence was determined to be DLRSWTAADTAAQISQ by acetylation of the peptide and measurement of the change in mass. The mass increased by 42 u, which indicates the presence of only one free amino group. Class II MHC molecules are known to present peptides from regions of proteins with highly conserved sequences [32]; thus the protein origin of the peptides cannot be identified with certainty. By combining LC-MS/MS with database searching, however, a reduction in data analysis time can be realized, which results in more efficient analysis of the antigens presented by class II MHC molecules.

Application of the Search Program to Complex Mixtures of Peptides Obtained from Single Cell Organisms

An additional application would involve the correlation of experimental tandem mass spectrometry data to protein sequence information. This would be useful for identification of protein sequences translated from genomic or cDNA sequences. To examine the ability to identify protein sequences by analysis of a corresponding peptide, the total proteins obtained from whole cell lysates of *E. coli* and *S. cerevisiae* cells were digested with the enzyme trypsin. For each organism, tandem mass spectra were obtained from peptides contained in the complex mixtures. The spectra were then used to search the respective species databases and the Genpept database. There was no prior knowledge of the amino acid sequences or the protein identities before searching the database. Results of the computer search with the eight tandem mass spectra derived from the trypsin digest of the *E. coli* proteins (four from *E. coli* proteins and four trypsin autolysis products) and nine spectra acquired from peptides derived from the *S. cerevisiae* lysate are shown in Table 3. These spectra were obtained from the most abundant peptide ions present in the mixture. All the tandem mass spectra were interpreted by using the sequences identified in the computer search and were found to be in good agreement with the sequences proposed by the computer search. Tandem mass spectra obtained from peptides at least 10 amino acids in length should provide a high probability identification of the protein. The search program also indicates the number of identified proteins that contain the amino acid sequence, which is useful to determine if the sequence is from a conserved domain or is unique. Once the complete nucleotide sequence for the *E. coli* and *S. cerevisiae* genomes have been determined, the complete genetic information that defines prokaryotic and eukaryotic organisms will be available [34]. The ability to identify sequences quickly, even if the protein is present in a mixture, by correlation of tandem mass spectral data to the translated gene sequences in the database will greatly facilitate biochemical studies.

Application of the Search Program to Single Protein Sequences

This data analysis approach also can be used to identify the peptides generated by enzymatic digestion of a single protein. Seven additional spectra were acquired from a tandem mass analysis of the peptides generated by chymotryptic digestion of the S-carboxymethylated heavy chain of glycoasparaginase. Proteolysis with the more specific enzyme trypsin resulted in sporadic and incomplete digestion, which required the use of the less specific chymotrypsin. The spectra were used to search the Genpept database as well as the glycoasparaginase sequence. A search of the single protein sequence required less than 5 s of computer time. The differences in the cross-correlation parameter for the first and second answers are quite large, which indicates excellent matches between sequences and spectra (Table 5). Search results for the tandem mass spectra through the Genpept database also are shown in Table 5. Although these peptides were produced by a chymotrypsin digest of the protein, this information was not included in the computer analysis. In this process 50% of the sequence was confirmed against the nucleotide-derived protein sequence with less than 1 min

Table 5. Results of searches by using the tandem mass spectra of peptides obtained from a homogenous preparation of glycoasparaginase digested with chymotrypsin[a]

Mass	$n+$	Sequence identified	Glycoasparaginase protein		S		G	
			C_n	ΔC_n	C_n	S_p	C_n	S_p
613.7	+1	SSPLPL	1	0.683	2	50	1	135
1139.3	+1	TLLVGESATTF	1	0.816	1	1	1	1
1189.4	+2	RNVIPDPSKY	1	0.871	1	1	1	1
1323.5	+2	LARNCQPNYW	1	0.462	2	2	19	11
2496.7	+2	AQSMGFINEDLSTSASQALHSDW	1	0.904	1	1	1	1
745.0	+2	DVGAVGDL	1	0.967	1	2	1	11
1875.0	+2	INEDLSTSASQASQAL	1	0.827	1	1	1	1
1971.1	+2	AQSMGFINEDLSTSASQAL	1	0.862	1	1	1	1
2243.5	+2	GGSPDELGETTLDAMIMGTTM	1	0.628	1	1	1	1

[a] Results are given for searches of the glycoasparaginase protein, the human database (S), and the Genpept database (G). All spectra were acquired by using electrospray ionization conditions. Charge states of the ions used in the tandem mass spectrometry experiment are listed under the column n^+. The column designations C_n and S_p refer to rankings of the identified sequences by the correlation parameter and score, respectively. The value ΔC_n indicates the difference between the first and second correlation parameter for the search of the glycoasparaginase sequence. The mass tolerance used in the search was ±0.05% or a minimum of ±1 u. Fragment ion mass tolerance was ±1 u. A mass of 161 u was used for carboxymethylated Cys.

of data analysis time [35]. By combining this data analysis approach with automated LC-MS/MS techniques [36], highly specific amino acid sequence information can be obtained and analyzed concurrent with the analysis. Thus, reasonably homogenous proteins could be analyzed in a highly specific manner even when small amounts of contaminating proteins or isoforms were present or when proteolytic processing had occurred.

Conclusions

We have found the use of fragmentation patterns in tandem mass spectra to be effective to identify amino acid sequences from a protein database. This is made possible by the ability to predict the fragment ions for an amino acid sequence and then to correlate the spectrum with the experimental data. Considering the large number of sequences that fall within the mass tolerance of the search, the fragmentation pattern acts as a good discriminator to identify the correct amino acid sequence. No other information, such as the enzyme specificity used to create the peptide, was used in the search. This method allows experimental data to be used directly for database searches, instead of first performing a manual interpretation of the data and then searching a database with the character-based representations of the sequence. Tandem mass spectra produced by other types of mass spectrometers, for example, high resolution tandem mass spectrometers [37], Fourier transform mass spectrometers [38], quadrupole ion traps [39], and reflectron time-of-flight mass spectrometers [40], should be amenable to this data interpretation approach.

This approach offers several advantages. Foremost is the ability to correlate directly uninterpreted tandem mass spectra to sequences in the database. This provides the ability to pre-screen data through the collection of all known sequences, to correlate amino acid sequences with cDNA sequencing, and to help assign protein functional roles to genome sequence data. The enormous efforts to sequence the human genome and the genomes of model organisms (*C. elegans, nematode, S. cerevisiae, D. melanogaster, E. coli,* etc.) will provide complete protein complements that will facilitate the comparison and interpretation of biological studies of these organisms. In addition, the ability to use information obtained from one peptide to search the protein database without regard to enzymatic specificity will allow the analysis of data obtained from collections of peptides derived from mixtures of unrelated proteins. In light of the exponential increase in the number of nucleotide and protein sequences, both at present and in the future, these databases and the techniques to access the data will become important resources for biological research.

Acknowledgments

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 Corporation. The authors thank Dr. Alan Blanchard for helpful discussions.

References

1. Edman, P.; Begg, G., *Eur. J. Biochem.* **1967,** *1,* 80–91.
2. Hewick, R. M.; Hunkapiller, M. W.; Hood, L. E.; Dryer, W. J. *J. Biol. Chem.* **1981,** *256,* 7990–7997.
3. Pearson, W. R. In *Methods in Enzymology,* Vol. 183; Doolittle, R. F., Ed.; Academic: San Diego, 1990; pp 63–98.
4. Doolittle, R. F. In *Methods in Enzymology,* Vol. 183; Doolittle, R. F., Ed.; Academic: San Diego, 1990; pp 99–110.
5. Cox, A. L.; Skipper, J.; Chen, Y.; Henderson, R. A.; Darrow, T. L.; Shabanowitz, J.; Engelhard, V.; Hunt, D. F.; Slinghuff, C. L., Jr. *Science* **1994,** *216,* 716–719.

6. Henzel, W.; Billeci, T.; Stults, J.; Wond, S.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A.* **1993,** *90,* 5011–5015.
7. Yates, J. R.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993,** *214,* 397–408.
8. Pappin, D.; Hojrup, P.; Bleasby, A. *Curr. Biol.* **1993,** *3,* 327–332.
9. James, P.; Qaudroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993,** *195,* 58–64.
10. Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993,** *22,* 338–345.
11. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989,** *246,* 64–71.
12. Arnott, D.; Shabanowitz, J.; Hunt, D. F. *Clin. Chem.* **1993,** *39,* 2005–2010.
13. Biemann, K. *Ann. Rev. Biochem.* **1992,** *61,* 977–1010.
14. 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.
15. Hunt, D. F.; Griffin, P. R.; Yates, J. R., III; Shabanowitz, J.; Fox, J. W.; Beggerly, L. K. In *Techniques in Protein Chemistry;* Hugli, T. E., Ed.; Academic: San Diego, 1989; pp 580–588.
16. Hunt, D. F.; Alexander, J. E.; McCormack, A. L.; Martino, P. A.; Michel, H.; Shabanowitz, J. In *Techniques in Protein Chemistry II;* Villafranca, J. J., Ed.; Academic: San Diego, 1991; pp 455–465.
17. Hunt, D. F.; Henderson, R. A.; Shabanowitz, J.; Sakaguchi, K.; Michel, H.; Sevilir, N.; Cox, A. L.; Apella, E.; Engelhard, V. N. *Science* **1992,** *255,* 1261–1263.
18. Henderson, R. A.; Michel, H.; Sakaguchi, K.; Shabanowitz, J.; Appella, E.; Hunt, D. F.; Engelhard, V. H. *Science* **1992,** *255,* 1264–1266.
19. Johnson, R. S.; Biemann, K. *Biomed. Env. Mass Spectrom.* **1989,** *18,* 945–957.
20. Hines, W. M.; Falick, A. M.; Burlingame, A. L.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1992,** *3,* 326–336.
21. Yates, J. R., III; Zhou, J.; Griffin, P. R.; Hood, L. E. In *Techniques in Protein Chemistry II;* Villafranca, J. J., Ed.; Academic: San Diego, 1990; pp 477–485.
22. Martinsen, D. P.; Song, B-H. *Mass Spectrom. Rev.* **1985,** *4,* 461–490.
23. Lee, T. D.; Vemuri, S. *Biomed. Environ. Mass Spectrom.* **1990,** *19,* 639–645.
24. Papayannopoulos, I. A.; Biemann, K. *J. Am. Soc. Mass Spectrom.* **1991,** *2,* 174–177.
25. Watkins, P. J.; Jardine, I.; Zhou, J. X. *Biochem. Soc. Trans.* **1991,** *19,* 957–962.
26. Demotz, S.; Grey, H.; Appella, E.; Sette, A. *Nature (London)* **1989,** *342,* 682–684.
27. Kaartinen, V.; Williams, J. C.; Tomich, J.; Yates, J. R., III; Hood, L. E.; Mononen, I. *J. Biol. Chem.* **1991,** *266,* 5860–5869.
28. Kolodziej, P.; Young, R. In *Guide to Yeast Genetics and Molecular Biology;* Guthrie, C.; Fink, G., Eds.; Academic: San Diego, 1991; pp 508–519.
29. Griffin, P. R.; Coffman, J. A.; Hood, L. E.; Yates, J. R., III. *Int. J. Mass Spectrom. Ion Processes* **1991,** *111,* 131–149.
30. Powell, L. A.; Heiftje, G. M. *Anal. Chim. Acta* **1978,** *100,* 313–327.
31. Owens, K. *Appl. Spectrosc. Rev.* **1992,** *27,* 1–49.
32. Chicz, R. M.; Urban, R. G.; Gorga, J. C.; Vignali, A. A.; Lane, W. S.; Strominger, J. L. *J. Exp. Med.* **1993,** *178,* 27–47.
33. Rudensky, A. Y.; Preston-Hurlburt, P.; Hong, S-C.; Barlow, A.; Janeway, C. A., Jr. *Nature* **1991,** *353,* 622–627.
34. Daniels, D. L.; Plunkett, G., III; Burland, V.; Blattner, F. R. *Science* **1992,** *257,* 771–778.
35. Mononen, I.; Heisterkamp, N.; Kaartinen, V.; Williams, J. C.; Yates, J. R., III; Griffin, P. R.; Hood, L. E.; Groffen, J. *Proc. Natl. Acad. Sci. U.S.A* **1991,** *88,* 2941–2945.
36. Stahl, D. C.; Martino, P. A.; Swiderek, K. M.; Davis, M. T.; Lee, T. D. *Proceedings of the 40th ASMS Conference on Mass Spectrometry and Allied Topics;* Washington DC, 1992; pp 1801–1802.
37. Johnson, R. S.; Biemann, K. *Biochemistry* **1987,** *26,* 1209–1214.
38. Hunt, D. F.; Shabanowitz, J.; Yates, J. R. *J. Chem. Soc. Chem. Commun.* **1987,** *8,* 548–550.
39. Kasier, R. E., Jr.; Cooks, R. G.; Syka, J. E. P.; Stafford, G. C., Jr. *Rapid Mass Spectrom.* **1990,** *4,* 30–33.
40. Kaufmann, R.; Spengler, B.; Lutzenkirchen, F. *Rapid Commun. Mass Spectrom.* **1993,** *7,* 902–910.