ANALYTICAL BIOCHEMISTRY 214, 397–408 (1993)

Peptide Mass Maps: A Highly Informative Approach to Protein Identification

John R. Yates III,[1] Stephen Speicher,[2] Patrick R. Griffin,[3] and Tim Hunkapiller
Mail Code FJ-20, Department of Molecular Biotechnology, School of Medicine, University of Washington, Seattle, Washington 98195

Received March 26, 1993

A computer searching algorithm has been used to identify protein sequences in the Protein Information Resource (PIR) database with peptide mass information (mass map) obtained from proteolytic digests of proteins analyzed by microcapillary high-performance liquid chromatography electrospray ionization mass spectrometry. A theoretical analysis of the cytochrome c family demonstrates the ability to identify protein sequences in the PIR database with a high degree of accuracy using a set of six predicted tryptic peptide masses. This method was also applied to experimentally determined peptide masses for a small GTP-binding protein, a protein from pig uterus, the human sex steroid binding protein, and a thermostable DNA polymerase. The results demonstrate that a set of observed masses which is less than 50% of the total number of predicted masses can be used to identify a protein sequence in the database. For the analysis presented in this paper, a mass matching tolerance of 1 amu is used. Under these conditions, mass maps created by fast atom bombardment mass spectrometry and matrix-assisted laser desorption time-of-flight would also be applicable. In cases where multiple matches are observed or verification of the protein identification is needed, tandem mass spectrometry sequencing can be used to establish sequence similarity.

As the genomes of various organisms are sequenced the availability of protein sequence information will be an invaluable aid to define the interrelationships of proteins involved in regulatory pathways. The availability of the complete protein complement will minimize the amount of structural characterization which will be required to identify a protein and thus allow rapid delineation of protein cascades. A key component to strategies for defining molecular pathways will be developing rapid and sensitive methods for generating an "address" with which a protein or gene sequence can be found in a database.

Traditionally, short stretches of amino acid sequences have been used to search databases to locate protein or gene sequences, or to identify proteins of similar sequence. This constitutes a fairly unique address for a protein and generally leads to successful searches. The amino acid sequence information is usually obtained by N-terminal analysis using Edman degradation (1). Approaches have been developed to characterize proteins separated by gel electrophoresis and transferred to membranes by electroblotting with Edman degradation (2–4). In many cases the N-terminus of the protein may be modified, either naturally or artefactually, blocking the initial coupling reaction of the Edman degradation chemistry. To circumvent this problem Aebersold and co-workers and Plaxton and Moorhead have developed *in situ* methods for proteolytic cleavage of proteins on membranes and in polyacrylamide gels, respectively, to obtain internal amino acid sequence information (5,6). The compatibility of *in situ* cleavage methods with microcapillary HPLC ESI–MS[4] for the structural analysis of proteins has been demonstrated by Griffin and co-workers (7). As genomes are sequenced coupling these methods with 2-di-

[1] To whom correspondence should be addressed. Fax: (206) 685-7344.
[2] Department of Biology, Mail Code 139-74, California Institute of Technology, Pasadena, CA 91125.
[3] Analytical Biochemistry, Merck and Company, P.O. Box 2000, R80-A23, Rahway, NJ 07065-0900.

[4] Abbreviations used: ESI–MS, electrospray ionization mass spectrometry; FAB–MS, fast atom bombardment mass spectrometry; PIR, protein information resource; TFA, trifluoroacetic acid; MS/MS, tandem mass spectrometry; MALD-TOF, matrix-assisted laser desorption time-of-flight.

0003-2697/93 $5.00

TABLE 1

Amino Acid Sequences from Cytochrome c Proteins

Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Arabian camel	G	D	V	E	K	G	K	K	I	F	V	Q	K	C	A	Q	C	H	T	V	E	K	G	G	K	H
Chimpanzee	G	D	V	E	K	G	K	K	I	F	I	M	K	C	S	Q	C	H	T	V	E	K	G	G	K	H
Dog	G	D	V	E	K	G	K	K	I	F	V	Q	K	C	A	Q	C	H	T	V	E	K	G	G	K	H
Grey Whale	G	D	V	E	K	G	K	K	I	F	V	Q	K	C	A	Q	C	H	T	V	E	K	G	G	K	H
Hippopotamus	G	D	V	E	K	G	K	K	I	F	V	Q	K	C	A	Q	C	H	T	V	E	K	G	G	K	H
Honeybee	G	I	P	A	G	D	P	E	K	G	K	K	C	A	Q	H	T	I	E	S	G	G	K	H	K	V
Lamprey	G	D	V	E	K	G	K	K	V	F	V	Q	K	C	S	Q	C	H	T	V	E	K	A	G	K	H
Mouse	G	D	V	E	K	G	K	K	I	F	V	Q	K	C	A	Q	C	H	T	V	E	K	G	G	K	H
Ostrich	G	D	I	E	K	G	K	K	I	F	V	Q	K	C	S	Q	C	H	T	V	E	K	G	G	K	H
Rabbit	G	D	V	E	K	G	K	K	I	F	V	Q	K	C	A	Q	C	H	T	V	E	K	G	G	K	H
Spider Monkey	G	D	V	F	K	G	K	R	I	F	I	M	K	C	S	Q	C	H	T	V	E	K	G	G	K	H
Tuna	G	D	V	A	K	G	K	K	T	F	V	Q	K	C	A	Q	C	H	T	V	E	N	G	G	K	H

Species	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
Arabian camel	K	T	G	P	N	L	H	G	L	F	G	R	K	T	G	Q	A	V	G	F	S	Y	T	D	A	N
Chimpanzee	K	T	G	P	N	L	H	G	L	F	G	R	K	T	G	Q	A	P	G	Y	S	Y	T	A	A	N
Dog	K	T	G	P	N	L	H	G	L	F	G	R	K	T	G	Q	A	P	G	F	S	Y	T	D	A	N
Grey Whale	K	T	G	P	N	L	H	G	L	F	G	R	K	T	G	Q	A	V	G	F	S	Y	T	D	A	N
Hippopotamus	K	T	G	P	N	L	H	G	L	F	G	R	K	T	G	Q	S	P	G	F	S	Y	T	D	A	N
Honeybee			G	P	N	L	Y	G	V	Y	G	R	K	T	G	Q	A	P	G	Y	S	Y	T	D	A	N
Lamprey	K	T	G	P	N	L	S	G	L	F	G	R	K	T	G	Q	A	P	G	F	S	Y	T	D	A	N
Mouse	K	T	G	P	N	L	H	G	L	F	G	R	K	T	G	Q	A	A	G	F	S	Y	T	D	A	N
Ostrich	K	T	G	P	N	L	D	G	L	F	G	R	K	T	G	Q	A	E	G	F	S	Y	T	D	A	N
Rabbit	K	T	G	P	N	L	H	G	L	F	G	R	K	T	G	Q	A	V	G	F	S	Y	T	D	A	N
Spider Monkey	K	T	G	P	N	L	H	G	L	F	G	R	K	T	G	Q	A	S	G	F	T	Y	T	E	A	N
Tuna	K	V	G	P	N	L	W	G	L	F	G	R	K	T	G	Q	A	E	G	Y	S	Y	T	D	A	N

Species	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78
Arabian camel	K	N	K	G	I	T	W	G	E	E	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Chimpanzee	K	N	K	G	I	I	W	G	E	D	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Dog	K	N	K	G	I	T	W	G	E	E	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Grey Whale	K	N	K	G	I	T	W	G	E	E	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Hippopotamus	K	N	K	G	I	T	W	G	E	E	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Honeybee	K	G	K	G	I	T	W	N	K	E	T	L	F	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Lamprey	S		K	G	I	V	W	N		E	T	L	F	V	Y	L	E	N	P	K	K	Y	I	P	G	T
Mouse	K	N	K	G	I	T	W	G	E	D	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Ostrich	K	N	K	G	I	T	W	G	E	D	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Rabbit	K	N	K	G	I	T	W	G	E	D	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Spider Monkey	K	N	K	G	I	I	W	G	E	D	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T
Tuna	K	S	K	G	I	V	W	N	E	N	T	L	M	E	Y	L	E	N	P	K	K	Y	I	P	G	T

Species	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
Arabian camel	K	M	I	F	A	G	I	K	K	K	G	E	R	A	D	L	I	A	Y	L	K	K	A	T	N	E	
Chimpanzee	K	M	I	F	V	G	I	K	K	K	E	E	R	A	D	L	I	A	Y	L	K	K	A	T	N	E	
Dog	K	M	I	F	A	G	I	K	K	T	G	E	R	A	D	L	I	A	Y	L	K	A	T	K	E		
Grey Whale	K	M	I	F	A	G	I	K	K	K	G	E	R	A	D	L	I	A	Y	L	K	K	A	T	N	E	
Hippopotamus	K	M	I	F	A	G	I	K	K	K	G	E	R	A	D	L	I	A	Y	L	K	K	A	T	N	E	
Honeybee	K	M	V	F	A	G	L	K	K	P	Q	E	R	A	D	L	I	A	Y	I	E	Q	A	S	K		
Lamprey	K	M	I	F	A	G	I	K	K	E	G	E	R	K	D	L	I	A	Y	L	K	K	S	T	E	S	E
Mouse	K	M	I	F	A	G	I	K	K	K	G	E	R	A	D	L	I	A	Y	L	K	K	A	T	N	E	
Ostrich	K	M	I	F	A	G	I	K	K	K	S	E	R	A	D	L	I	A	Y	L	K	D	A	T	S	K	
Rabbit	K	M	I	F	A	G	I	K	K	K	D	E	R	A	D	L	I	A	Y	L	K	K	A	T	N	E	
Spider Monkey	K	M	I	F	V	G	I	K	K	K	G	E	E	R	A	D	L	I	A	Y	L	K	K	A	T	N	E
Tuna	K	M	I	F	A	G	I	K	K	K	G	E	R	Q	D	L	V	A	Y	L	K	S	A	T	S		

TABLE 2

Summary of (M+H)+ Values Used in Cytochrome c Search

Species	9–13	14–22	28–38	40–53	56–72	92–99
Arabian camel	634	1019	1169	1459	2011	907
Chimpanzee	651	1035	1169	1429	2009	808
Dog	634	1019	1169	1457	2011	907
Gray whale	634	1019	1169	1459	2011	907
Hippopotamus	634	1019	1169	1473	2011	907
Honeybee	634	1234	1195	1473	1383	1322
Pacific lamprey	620	1035	1119	1457	2051	836
House mouse	634	1019	1169	1431	1997	907
Ostrich	634	1035	1147	1489	1997	907
Domestic rabbit	634	1019	1169	1459	1997	907
Spider monkey	651	1035	1169	1475	2009	907
Skipjack tuna	622	1247	1216	1505	2051	950

mensional gel electrophoresis, a method capable of separating over 2000 different proteins in a single analysis, will be a powerful approach for the analysis of complex cellular systems (8). In an effort to create a strategy that would enable the rapid analysis of large numbers of proteins, we have examined alternate strategies for obtaining addresses for proteins that may exist in the database.

Peptide mapping techniques, combining fast atom bombardment and mass spectrometry (FAB–MS), were utilized by Morris and co-workers and Gibson and Biemann to determine the masses of peptides resulting from proteolytic or chemical cleavage of a protein in efforts to verify and correct gene sequences (9,10). This technique has been useful for identifying frame shift, deletion, or addition errors in gene sequences since an amino acid addition or deletion will create a variation in the mass predicted by the gene sequence. The difference in observed and predicted mass can often be indicative of the type of error in the gene sequence. Strategies for peptide mapping have improved with the introduction of electrospray ionization allowing microscale separation techniques such as capillary chromatography and capillary electrophoresis to be interfaced to mass spectrometers (11–17). This approach has led to the rapid and sensitive analysis of complex mixtures of peptides, and thus simplified peptide mapping studies. In conjunction with tandem mass spectrometry, amino acid sequences and post-translational modifications can also be determined (18–25).

Mass spectrometry is emerging as a rapid and sensitive technique for the determination of the molecular weights of peptides and proteins. Combining molecular weight information obtained by mass spectrometry with information represented within a database could form a powerful approach for protein identification. Although an accurate molecular weight of a protein could constitute a unique marker, the presence of post-translational modifications, which are frequently not known or represented in the database, would alter the mass and lead to incorrect matches. A set of peptides generated by specific proteolytic cleavage could create a unique fingerprint for a protein sequence which would not be unilaterally affected by the presence of post-translational modifications. On the supposition that a collection of masses for peptides derived from proteolytic cleavage of a protein creates a highly informative fingerprint for a protein sequence, we explored the potential of using mass data from peptide maps obtained by microcapillary HPLC ESI–MS to search the database to obtain identities of proteins. To establish the feasibility of this approach, a theoretical analysis using cytochrome c from 12 different species was performed. Encouraged by the results of this analysis, experimentally generated peptide maps were used to search the PIR database with subsets of the masses produced in the mapping experiments. The result is that a peptide map produces a highly informative fingerprint which can be used to identify a protein sequence in a protein database. A preliminary account of this work was presented at the 5th Annual Meeting of the Protein Society in Baltimore, Maryland, June 1991.

MATERIALS AND METHODS

Computation. All computation was performed on a Hewlett Packard HP-9000 minicomputer. The computer algorithms were written in FORTRAN. An algorithm was written to convert the protein sequences in the PIR database to the predicted $(M + H)^+$ values for the fragments produced by the enzymes trypsin, *Staphylococcus aureus* V8 protease, endoproteinase Lys-C, and cyanogen bromide. For this work two separate cleavage databases were produced, PIR ver. 28 (36,150 sequences) and 33 (42,215 sequences). The masses for each predicted fragment produced in the cleavage program were stored as protonated molecular weight ((M +

TABLE 3

Summary of Results from the Database Search with (M+H)+ Values from Cytochrome c Proteins

Sequence	6 mass matches	5 mass matches
Arabian camel	CCCM Arabian camel	CCDG dog
	CCGW guanaco	CCHP hippopotamus
		CCRB rabbit
		CCSLE so. elephant seal
Chimpanzee	CCCZ chimpanzee	CCMKP spider monkey
	CCHU human	CCMQR rhesus macque
Dog	CCDG dog	CCBTS long-fingered bat
	CCSLE south. elephant seal	CCCM Arabian camel
		CCGW guanaco
		CCHP hippopotamus
		CCPG pig
		CCWHC gray whale
Gray whale	CCCM Arabian camel	CCDG dog
	CCGW guanaco	CCHP hippopotamus
	CCWHC grey whale	CCRB rabbit
		CSLE so. elephant seal
Hippopotamus	CCHP hippopotamus	CCCM Arabian camel
		CCDG dog
		CCGW guanaco
		CCSLE so. elephant seal
		CCWHC gray whale
Honeybee	CCHB honeybee	None
Pacific lamprey	CCLM pacific lamprey	None
Mouse	CCMS house mouse	CCRB rabbit
	CCRT Norway rat	
	A23057 house mouse gene sequence	
Ostrich	CCEU emu	CCDK domestic duck
	CCOS ostrich	CCPN king penguin
		CCPY domestic pigeon
Rabbit	CCRB domestic rabbit	CCCM Arabian camel
		CCGW guanaco
		CCMS house mouse
		CCRT Norway rat
		CCWHC gray whale
		A23057 house mouse
Spider monkey	CCMKP spider monkey	CCCZ chimpanzee
		CCHU human
		A31764 human
Skipjack tuna	CCBN skipjack tuna	None

$H)^+$) values as the program was originally written for peptide maps produced by FAB–MS. Exhaustive digestion was assumed and no allowances were made for post translational modifications. The $(M + H)^+$ values derived from predicted digestion of each protein sequence were stored as records and links were maintained to each protein's annotation and amino acid sequence. The disk space required for storing the mass data was approximately 30 Mb. The cleavage database can be updated by rerunning the cleavage program as new versions of the PIR become available and requires approximately 25–30 min of CPU time to generate a new mass database.

The program developed to the search the cleavage database provides the user with several options. First is the choice of cleavage database to be used in the search. Four cleavage databases are available: trypsin, *S. aureus* V8 protease, endoproteinase Lys-C, and cyanogen bromide. Next, the user must set the mass tolerance for the matching algorithm and this ranges from 1 to 10 amu. The user then provides a list of the protonated molecular weights for the peptide map and a molecular weight value for the protein.

The search algorithm queries each protein fragment record with the set of input $(M + H)^+$ values. A match occurs when masses are within the selected mass tolerance of each other and within the molecular weight window input by the user. This window is a consecutive amino acid sequence that sums to the molecular weight value and only those mass matches that occur within this window are recorded. For example, if a set of $(M + H)^+$ values are used from a protein presumed to be 10

TABLE 4

Summary of Database Searching Results

(a) Cytochrome c: Dog

Masses (a.a. residues)	4 mass matches	3 mass matches
1170 (28–38)	CCBTS long fingered bat[a,b]	CCCM Arabian camel[b]
1457 (40–53)	CCDG dog[a,b]	CCDK domestic duck[b]
634 (9–13)	CCSLE southern elephant seal[a,b]	CCGW guanaco[b]
906 (92–99)		CCHOD donkey[b]
(13,000)		CCHP hippopotamus[b]
		CCLQ desert locust[b]
		CCMS house mouse[b]

(b) Cytochrome c: Pigeon

Masses (a.a. residues)	4 mass matches	3 mass matches
1170 (28–38)	CCDK domestic duck[c,d]	CCBTS long fingered bat[d]
1489 (40–53)	CCPN king penguin[c,d]	CCCH chicken[d]
678 (74–79)	CCPY domestic pigeon[c,d]	CCCM Arabian camel[d]
906 (92–99)		CCCZ chimpanzee[d]
(13,000)		CCDG dog[d]
		CCEU emu[d]
		CCGW guanaco[d]

(c) GTP-binding protein Rab3: Human

Masses (a.a. residues)	7 mass matches	6 mass matches	4 mass matches
1102 (25–35)	C34323 GTP-binding protein: human	SO1765 gene for rab3 Norway rat	A25970 cholera transcriptional activator
990 (86–93)	A29224 GTP-binding protein: cattle		
1593 (185–200)			
1317 (73–83)			
1512 (13–24)			
1811 (151–166)			
2081 (42–60)			
(26,000)			

[a] No amino acid differences in the peptides used in the search.
[b] Cytochrome c.
[c] No amino acid differences in the peptides used in the search.
[d] Cytochrome c.

kDa in size, but is actually a fragment of a much larger protein, the program will only count the matches that occur within a consecutive 10-kDa region of the larger protein's sequence. If a molecular weight cutoff was used instead of a sliding window, this protein sequence would not be identified. This also allows a margin of error in molecular weight estimates used in the search. The number of mass matches for all the PIR entries is stored and at the end of the search, the list is sorted and the PIR entries with the most matches are displayed on the computer screen. For the data produced in this paper the search program required less than a minute to search 42,215 sequences. The impact on search times of using more peptide $(M + H)^+$ in the search versus less is imperceptible.

Proteolytic digestion. Cytochrome c (dog and domestic pigeon) were obtained from Sigma Chemical Co. Rab 3 was obtained from Dr. Thomas Sudhoff, University of Texas, Dallas, human sex steroid binding protein from Dr. Philip Petra, University of Washington, and uteroferrin from Dr. R. M. Roberts, University of Missouri. The proteins were digested with trypsin by dissolving the protein in 50–200 μl of 100 mM ammonium bicarbonate, pH 8.5, or 100 mM Tris–HCl, pH 8.5 and then exposing the protein to protease (1–2% w/w (enzyme/substrate)) for 12 h at 37°C. The digestion mixture was concentrated on a Speed Vac concentrator and brought to a final concentration of ~1 pmol/μl with the addition of 0.1% trifluoroacetic acid (TFA).

High-performance liquid chromatography. Peptides were fractionated by reverse-phase microcapillary high-performance liquid chromatography with a gradient of 0 to 100% acetonitrile (0.085% TFA) in 0.1% aqueous TFA using an Applied Biosystems 140A solvent delivery

TABLE 5

Summary of Database Searching Results

Uteroferrin: Domestic pig		
Masses (a.a. residues)	5 mass matches	3 mass matches
868 (1–8) 1163 (205–214) 1229 (118–126) 1314 (276–287) 1816 (67–81) (38,000)	A27035 uteroferrin pig A33318 uteroferrin precursor pig	LKRT2 Link 2 protein Norway rat QRECMB chemotaxis mot b protein
Sex steroid binding protein: Human		
Masses (a.a. residues)	8 mass matches	7 mass matches
576 (95–99) 1840 (48–63) 2373 (257–280) 1331 (174–186) 912 (40–47) 925 (100–106) 1153 (64–72) 916 (126–134) (45,000)	S00077 sex steroid binding protein BOHUS sex steroid binding protein	A26339 sex steroid binding protein precursor
Thermostable DNA polymerase: *Thermus flabus*		
Masses	4 mass matches	2 mass matches
1393 1791 3110 697 (100,000)	A28784 SEC7 gene sequence A33530 DNA-directed DNA polymerase I A31068 SEC7 protein sequence	A23584 coagulation factor VIII precursor

system. The HPLC pump was operated at 100 μl/min and the flow was split prior to the sample injection loop (5 μl) to produce a flow through the column of 2 to 4 μl/min. Column eluent was then directed to the electrospray ionization source. Capillary columns were 320-μm i.d. × 15 cm and were packed with 5-μm C_{18} particles (LC Packings, San Francisco, CA).

Mass spectrometry. Mass spectra were recorded with a Finnigan MAT (San Jose, CA) TSQ-700 triple quadrupole mass spectrometer equipped with a 20 keV conversion dynode, DECstation 2100, and an electrospray ion source as previously described (7). The mass-to-charge ratios of peptides were recorded by scanning Q_3 at a rate of ~400 amu/s over a mass range of 400 to 1800 throughout the HPLC gradient. Sequence analysis of peptides was performed during a second HPLC analysis by selecting the parent ion with a 3–4 amu (peak width at half height) wide window in Q1 and passing the ions into the collision cell which was filled with argon to a pressure of 5 mTorr. Collision energies were on the order of 20 to 50 eV. The fragment ions produced in Q2 were transmitted to Q3, which was scanned at 500 amu/s over a mass range from 50 amu to the parent mass to record the fragment ion mass-to-charge ratios.

RESULTS AND DISCUSSION

Methods for the rapid identification of protein sequences will become increasingly important as various genome projects proceed. Peptide mapping techniques based on mass spectrometry have been widely used to correct and verify gene sequences and identify post-translational modifications. These same approaches could have applications for the rapid identification of protein sequences represented in a protein database. In this work we explored the use of peptide maps as a highly informative approach to protein identification.

This study had several objectives. The first was to examine the information content of peptide mass maps and their potential for searching a database. The second was to determine the efficacy of protein database searches among closely related protein sequences, and the third was to apply this searching technique to experimentally derived peptide maps from microcapillary

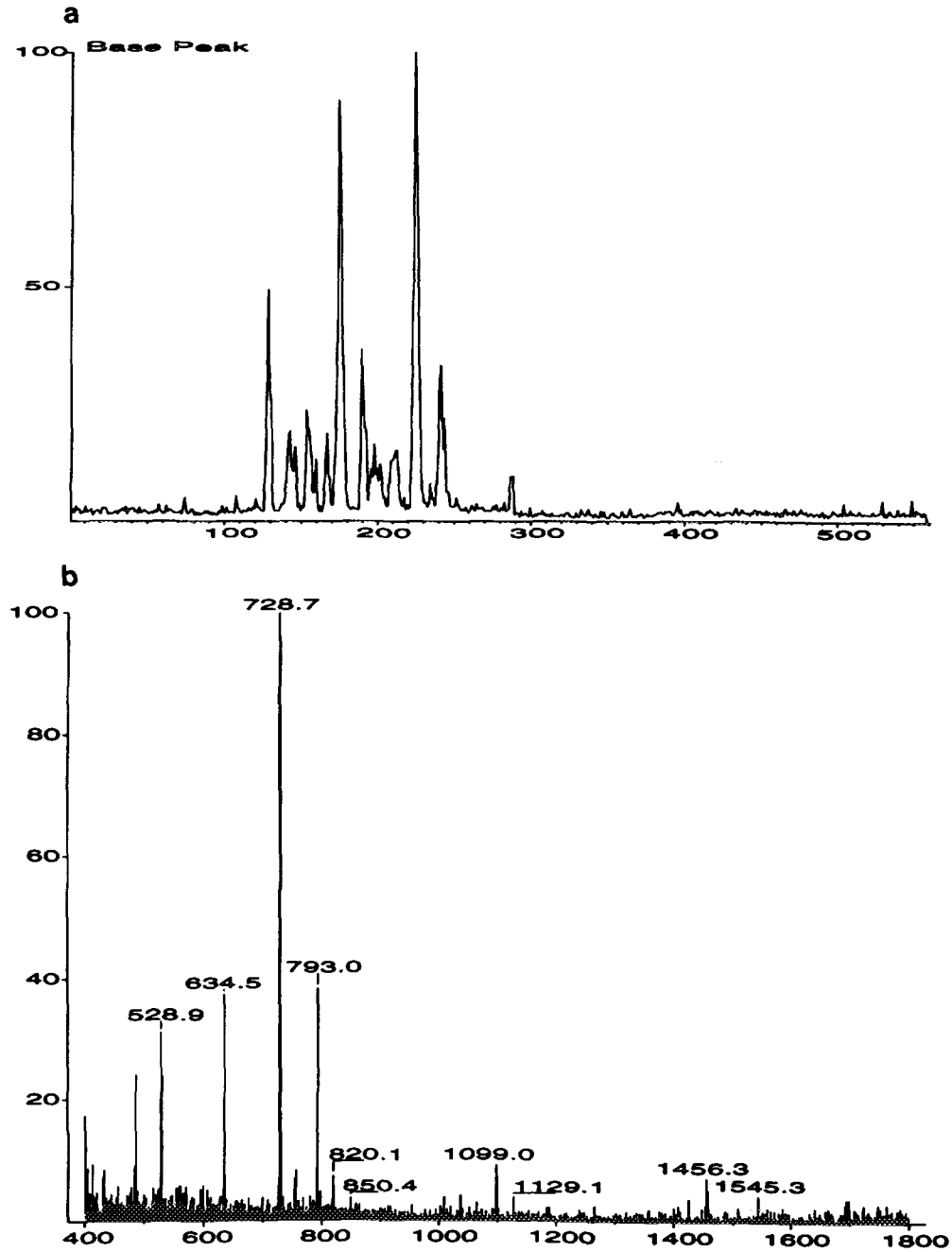

FIG. 1. (a) A plot of the base peak ion intensity verse scan number of the tryptic digest from dog cytochrome c. (b) A mass spectrum produced by summing scans 163–183 from the tryptic digest of dog cytochrome c.

HPLC ESI–MS analysis to judge the reliability and limitations of using mass data as a highly informative approach for searching a protein database. Protonated molecular weights were used that cover a wide range of values, e.g., from small to large. They were chosen without regard to sequence. Ideally, this process would be computer controlled and all masses observed in an ESI–MS analysis would be used in the search.

The search algorithm requires two types of data input. The first is the set of $(M + H)^+$ values observed for the peptides produced from proteolytic cleavage and the second is an approximate molecular weight for the protein. The molecular weight of the protein could have been determined from gel electrophoresis or matrix-assisted laser desorption mass spectrometry experiments. The molecular weight does not need to be accurate and

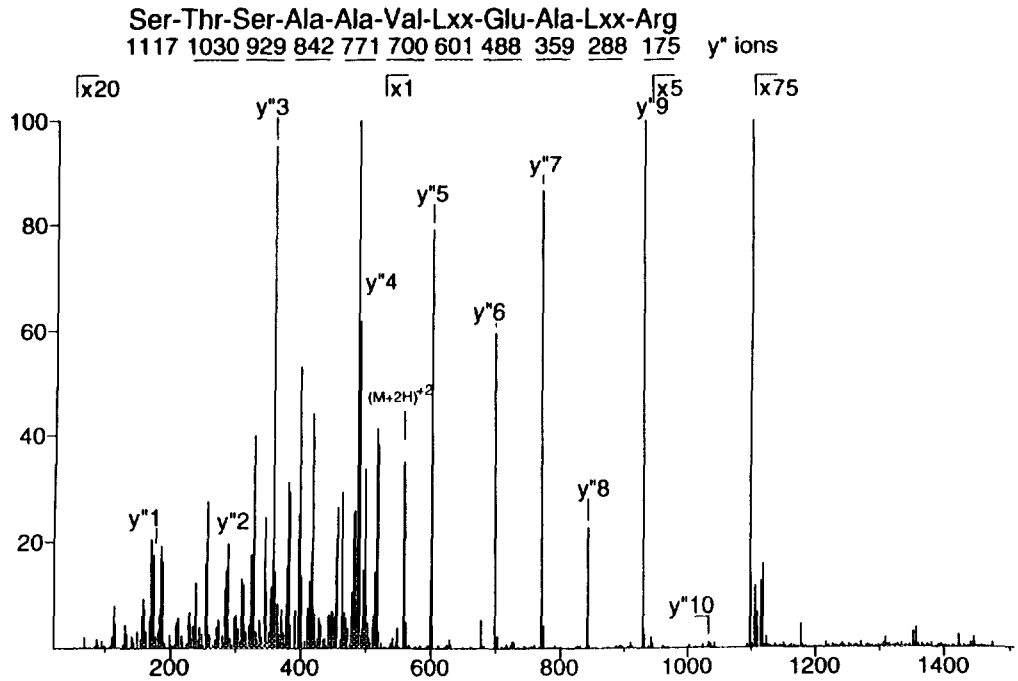

FIG. 2. A capillary HPLC-MS/MS collision-activated dissociation mass spectrum recorded on $(M + 2H)^{2+}$ ions of an 11-residue peptide, STSAAVXEAXR, generated from a 20-pmol injection of a tryptic digest of the thermostable DNA polymerase. The amino acid located at position X could be leucine or isoleucine. Ions of type y" are labeled. The predicted fragment ions for the 11-residue tryptic peptide are given above the spectrum and those type y" ions appearing in the mass spectrum are underlined. All of the type y" ions fragment ions are observed in the mass spectrum.

is used to limit the search primarily to a molecular weight window within a protein sequence. Finally, the cleavage database to be used in the search must be chosen, e.g., trypsin. In general, search times were uneffected by increasing the number of $(M + H)^+$ values used in the search or increasing the mass matching tolerance. All searches required less than a minute of computer time.

Analysis of the Cytochrome c Family

Sequences from the cytochrome c family were used to test the hypothesis that a set of protonated masses obtained from a proteolytic digestion of a protein is sufficient to search a protein database and obtain an identification of the protein. The initial analysis was performed with members of this protein family because it is a large family of proteins with very similar amino acid sequences. Table 1 lists the amino acid sequences for the different species included in the analysis. A protease such as trypsin is a useful enzyme since it cleaves principally on the C-terminal side of Arg and Lys with the exception of Arg–Pro and Lys–Pro. Peptides on the order of 8–30 residues would be expected on the frequency of occurrence of lysine and arginine in the protein database. On the average, any amino acid change that occurs more frequently than 1 in 8 residues would be readily reflected in a change of mass. However, when protein sequences are identical, or when the sequence differences do not occur in the regions of the protein sequence reflected in the values used in the search, no differentiation could be expected.

Table 2 lists the $(M + H)^+$ values for a set of six peptides that would be produced by treatment of each of the proteins with trypsin. Many of the peptides have the same $(M + H)^+$ value and identical sequences at the selected positions. However, the use of a set of values greatly limits the number of protein sequences which would match the entire set. For example, the sequences for the Arabian camel and California gray whale are identical and therefore share the same set of masses. However, there are no $(M + H)^+$ values in common between the lamprey sequence and the Arabian camel sequence as these sequences differ by 18 amino acids (1 in 5.7). Two more closely related sequences are those of the rabbit and mouse which differ by two amino acids (1 in 52), but show differences in the set of peptides selected for the search. In general, this set of six $(M + H)^+$ values represents a sequence coverage of ~61%.

For each species the set of protonated masses representing the same sequence regions was used to search the protein database. The results of the searches are

TABLE 6

Summary of the Effect of Increasing Mass Tolerance on Database Searches

Mass tolerance 3 amu	Mass tolerance 5 amu	Mass tolerance 7 amu
	Cytochrome c, dog	
Long-fingered bat	Long-fingered bat	Long-fingered bat
Arabian camel	Arabian camel	Arabian camel
Dog	Dog	Dog
Southern elephant seal	Southern elephant seal	Southern elephant seal
Gray whale	Gray whale	Gray whale
Domestic rabbit	Domestic rabbit	Domestic rabbit
	Cytochrome c, pigeon	
Domestic duck	Domestic duck	Domestic duck
King penguin	King penguin	King penguin
Domestic pigeon	Domestic pigeon	Domestic pigeon
	Cytochrome P450 hydroxylase homolog	Cytochrome P450 hydroxylase homolog
		Secreted 45 K protein precursor
	GTP-binding protein	
GTP-binding protein human	GTP-binding protein human	GTP-binding protein human
GTP-binding protein cattle	GTP-binding protein cattle	GTP-binding protein cattle
	Uteroferrin	
Uteroferrin pig	Uteroferrin pig	Uteroferrin pig
Uteroferrin pig precursor	Uteroferrin pig precursor	Uteroferrin pig precursor
	Hypothetical 176 K protein	Hypothetical 176 K protein
		Cytochrome P450 52A1
		Regulatory protein degS
	Human sex steroid binding protein	
Sex steroid binding protein	Sex steroid binding protein	Sex steroid binding protein
Sex steroid binding protein	Sex steroid binding protein	Sex steroid binding protein
Sex-hormone binding globulin	Sex-hormone binding globulin	Sex-hormone binding globulin
	DNA polymerase	
SEC7 protein	SEC7 protein	SEC7 protein
DNA-directed DNA polymerase I	DNA-directed DNA polymerase I	DNA-directed DNA polymerase I
S-locus-specific glycoprotein S29-2-precursor	S-locus-specific glycoprotein S29-2-precursor	S-locus-specific glycoprotein S29-2-precursor
	Genome polyprotein	Genome polyprotein
	Protein P3-6b	Protein P3-6b
	virC-region hypothetical protein yscC precursor	virC-region hypothetical protein yscC precursor
		DNA-directed DNA polymerase

Note. Only the protein sequences matching the complete map are shown.

displayed in Table 3. As would be expected, the correct protein sequence matched the complete set of values in all cases. In half of the examples more than one protein sequence matched all six values. In all of these cases the sequences were identical in all of the peptide $(M + H)^+$ values used for the search. All the proteins which matched five of the values were members of the cytochrome c family. The more complete the set of $(M + H)^+$ values used in the search the greater the level of discrimination for this closely related family of proteins.

Experimentally Derived Peptide Maps

To test the sensitivity of the database search with experimentally derived data, proteolytic digestion of six proteins was performed followed by separation and mass analysis of the peptide mixtures by microcapillary

TABLE 7

Summary of Database Search with an Artificial Peptide Map

Masses	Protein sequences (3 mass matches)
634	S15518 Type III restriction endonuclease
1035	S20687 DNA ligase
1119	BVBYD9 RAD9 protein
1489	CCCH chicken cytochrome c
2051	CCDK domestic duck cytochrome c
950	CCEU emu cytochrome c
(13,000)	CCLM pacific lamprey cytochrome c
	CCOS ostrich cytochrome c
	CCPN king penguin cytochrome c

Note. One peptide (M+H)+ value was chosen from each of six cytochrome c species to create a set of values for the search. No protein sequence matched the complete set of (M+H)+ values.

HPLC ESI–MS. This set of proteins consisted of two members of the cytochrome c family (dog and pigeon), a small GTP-binding protein from humans (~24.6 kDa), uteroferrin from pig uterus (~34.2 kDa), the human sex steroid binding protein (~40.4 kDa), and a thermostable DNA polymerase (~100 kDa). All of these proteins were presumed to exist in the PIR database. The HPLC trace of the tryptic digest of the DNA polymerase has been previously published (26). The results of the database searches are summarized in Tables 4 and 5.

Cytochrome c proteins from the dog and domestic pigeon were treated with the protease trypsin to produce a mixture of peptides. The resulting mixture was separated by reverse-phase micro-capillary HPLC and mass analyzed by ESI–MS. Shown in Fig. 1a is the base peak plot for the analysis. This is a plot of the intensity of the base peak ion in each scan as a function of scan number. The summed mass spectra of scans 163–183 for dog cytochrome c is displayed in Fig. 1b. The mass spectra produced by electrospray ionization contain multiply charged ions of the peptides. Ions of peptides generated by tryptic digestion generally contain doubly and triply charged ions. Ions representing incompletely digested fragments can often be identified by their higher charge states (more basic sites for protonation). Deconvolution of the mass spectra can be used to identify the charge state of the ions to locate peptide fragments resulting from incomplete digestion, since these fragments are not represented in the cleavage database (27). In practice these fragments should not interfere with the search provided a sufficient number of $(M + H)^+$ values representing the fragments predicted from exhaustive digestion are present. A set of four $(M + H)^+$ values from these analyses were chosen for the search and the results are displayed in Table 4a. Only the top 10 results were saved and all the matches were cytochrome c proteins. The top three answers were all indistinguishable based on the $(M + H)^+$ values used in the search since there were no differences in the amino acid sequences among those species in the regions chosen, although there are sequence differences between dog and southern elephant seal at position 100 and the long-fingered bat at positions 62 and 88.

The top three results of the search with the peptide $(M + H)^+$ values from pigeon cytochrome c matched all four values and seven results matched three of the peptide $(M + H)^+$ values. Interestingly, all of the top matches were members of the Aves class. No sequence differences existed in the regions used in the search, but the king penguin sequence differs in four positions (35, 69, 80, and 103). The change at position 35 is leucine to isoleucine, a difference that would not have been reflected in a mass change as these two amino acids are isobaric. Changes between glutamine and lysine (same nominal mass) would be detected if trypsin is used to produce the map unless the change occurs in the sequence Lys–Pro. The domestic duck sequence differs at positions 3, 69, and 80.

The peptide map of the small GTP-binding protein matched seven fragments of the human and bovine sequence. These two sequences differ by one amino acid at position 7 in the sequence, a region not represented by any of the $(M + H)^+$ values used in the search. Six matches were found to the gene sequence of Rab 3 from the Norway rat. Again there is a single amino acid difference between this gene sequence and the human sequence at position 196, and this was represented by one of the fragments used in the search allowing differentiation of the two sequences. A total of four $(M + H)^+$ values matched the cholera transcriptional activator protein and these were 991, 991, 1103, and 1511. These are within the 1 amu of the $(M + H)^+$ values used in the search and within the matching tolerances of the search algorithm. There is no apparent sequence similarity between the two proteins.

Mass maps were produced for several additional proteins, uteroferrin, an acid phosphatase isolated from pig uterus (~34.2 kDa), and the human sex steroid binding protein (40.4 kDa), to determine if the specificity of the search decreases with increasing mass. Five $(M + H)^+$ values produced in the map of uteroferrin were chosen representing ~18% of the molecular mass of the protein. The search identified two proteins with five matches, uteroferrin and uteroferrin precursor, and two unrelated proteins with matches to three of the masses. A second larger glycoprotein, human sex steroid binding protein (44.4 kDa), was mass mapped and eight of the values were used in the search. Approximately 13% of the molecular mass of the protein consists of carbohydrate (28). Two entries of the human sex steroid binding protein sequence were found with eight matches, and the precursor protein was found with seven matches.

Two additional questions are the efficacy of the search when a small number of peptide masses are used and the problem of a protein sequence that does not exist in the database. A set of four $(M + H)^+$ values was used in the search of a mass map from a thermostable DNA polymerase (90 kDa), representing ~7% of the molecular mass of the protein. The molecular weight value for the search was set at 100 kDa. Four sequences were identified in the search: a SEC7 gene and protein sequence, a DNA-directed DNA polymerase I sequence, and the coagulation factor VIII precursor. A range of $(M + H)^+$ values from 697 to 3110 was used. A sequence analysis using tandem mass spectrometry of selected regions of the protein revealed those regions to have an identical amino acid sequence to the DNA-directed DNA polymerase I. Two peptides, mass 1116 and 2191, were analyzed by MS/MS and their sequences determined to be STSAAVLELAR (513–523) and LSSSDPNLQNIPVRTPLGQR (574–593). The MS/MS spectra for the peptide of mass 1116 ($(M + 2H)^{+2}$, 559) is shown in Fig. 2. These sequences correspond to regions of the DNA-directed DNA polymerase, strongly suggesting a similarity to this protein sequence.

Limitations

The data presented in Tables 3, 4, and 5 were acquired using a mass matching tolerance of 1 amu. Table 6 summarizes results obtained when the mass matching tolerance is increased to 3, 5, and 7 amu. For the cytochrome c proteins (dog and pigeon) the specificity of the search decreases with an increase in the mass tolerance. Although the search still identifies the peptide maps as belonging to a cytochrome c protein, the species specificity decreases. In general, the sensitivity of the search decreases with an increase in the mass tolerance. However, the GTP-binding protein and the sex steroid binding protein were still the primary sequences identified in the search even though a mass tolerance of 7 amu was used. To maximize specificity for database searches with a peptide map produced with the protease trypsin, the mass matching tolerance should be kept to 3 amu or below.

The use of peptide mass maps for database searching can be an important adjunct to protein studies, although there are some limitations to this approach. A search of the database with any set of masses will produce a match of at least some masses to a protein sequence. Only in cases of a complete match is there a high probability of protein identification. If less than a complete match is observed then some additional form of sequence verification should be attempted, such as Edman degradation or MS/MS analysis. For example, Table 7 summarizes the results of a database search using a nonexistent peptide map. The set of $(M + H)^+$ values were chosen from among the peptides produced by tryptic digestion of the cytochrome c family. One peptide $(M + H)^+$ value was taken from each of six species. A complete match of the peptide map to a protein sequence is not observed so the identified proteins should be considered unrelated. Any similarity of relationships to protein sequences indentified by partial matches should be discovered through amino acid sequence analysis.

The use of a database which represents only complete proteolysis will be inherently limited in relation to real experimental data. This could create a problem with proteins that are difficult to digest such as hydrophobic proteins, or heavily glycosylated proteins. An advantage to using electrospray ionization and tryptic digestion is the ability to identify peptides created by complete digestion on the basis of the charge states observed in the mass spectrum. Peptides resulting from incomplete digestion are more likely to be of the +3, +4, and +5 charge state. An exception will be peptides containing other basic amino acids such as His. The charge states of the peptides can serve as a guide to choosing appropriate peptide masses for the search. Protein identification is possible when less than a complete set of peptides is used in the search. However, at least a portion of the peptides produced from proteolysis must result from complete digestion. It would be entirely possible to create and use a cleavage database representing incomplete proteolysis. This would more than double the disk space needed to store the database and result in longer search times.

As the number of proteins in the database grows the impact on the efficacy of the search should be minimal except in cases where the number of members of a highly similar family of proteins increases. The analysis of the cytochrome c sequences demonstrates that even proteins with small changes in their sequences can be differentiated if there is a large enough set of peptide masses for the search. The more likely outcome of an increase in the size of the database is on search speed. If a linear relationship between search speed and database size is assumed then a factor of 10 increase would result in roughly 10-min searches on the same computer equipment. However, the impact should be minimal since the size of databases is increasing at about the same rate as the computational speed.

CONCLUSIONS

This analysis has demonstrated the feasibility of using mass data derived from proteolytic digestion of proteins to search a protein database. Mass data experimentally derived from microcapillary HPLC ESI–MS experiments can be used to perform the search and it should be possible to perform the same mass mapping

technique with FAB–MS and MALD–TOF. In combination with 1- and 2-dimensional gel electrophoresis techniques this approach could be an invaluable method for limiting amino acid sequencing efforts. This will also add an extra dimension to the information obtained from two-dimensional electrophoresis of complex systems (29). Microcapillary HPLC ESI–MS and *in situ* digestion of proteins in polyacrylamide gels and on membranes have been shown to be compatible making this approach feasible (7). Combining this approach with tandem mass spectrometry experiments allows verification of the protein sequence identified in the search.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, Science and Technology Center Cooperative Agreement 8809710, and Finnigan MAT Corp. through the Science and Technology Center's Industrial Affiliates Program.

REFERENCES

1. Hewick, R. M., Hunkapiller, M. W., Hood, L. E., and Dreyer, W. J. (1981) *J. Biol. Chem.* **256,** 7990–7997.
2. Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J., and Van Montagu, M. (1985) *Eur. J. Biochem.* **152,** 9–19.
3. Aebersold, R. H., Teplow, D., Hood, L. E., and Kent, S. B. H. (1986) *J. Biol. Chem.* **261,** 4229–4238.
4. Matsudaira, P. (1987) *J. Biol. Chem.* **262,** 10035–10038.
5. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. H. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6970–6974.
6. Plaxton, W. C., and Moorhead, G. B. G. (1989) *Anal. Biochem.* **178,** 391–393.
7. Griffin, P. R., Coffman, J. A., Hood, L. E., and Yates, III, J. R. (1991) *Int. J. Mass Spectrom. Ion Proc.* **111,** 131–149.
8. Aebersold, R. H., Leavitt, J., Hood, L. E., and Kent, S. B. H. (1987) Methods in Protein Sequence Analysis, (Walsh, K., Ed.), pp. 277–294, Humana Press, Clifton, NJ.
9. Morris, H. R., Panico, M., and Taylor, G. W. (1983) *Biochem. Biophys. Res. Commun.* **117,** 299–305.
10. Gibson, B. W., and Biemann, K. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1956–1960.
11. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1990) *Mass Spectrom. Rev.* **9,** 37–70.
12. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) *Science* **246,** 64–71.
13. Huang, E. C., and Henion, J. D. (1990) *J. Am. Soc. Mass Spectrom.* **1,** 158–165.
14. Loo, J. A., Udseth, H. R., and Smith, R. D. (1988) *Biomed. Environ. Mass Spectrom.* **17,** 411–414.
15. Loo, J. A., Udseth, H. R., and Smith, R. D. (1989) *Anal. Biochem.* **176,** 404–412.
16. Hail, M., Lewis, S., Jardine, I., Liu, J., and Novotny, M. (1990) *J. Microcol. Sep.* **2,** 285–292.
17. Moseley, M. A., Jorgenson, J. W., Shabanowitz, J., Hunt, D. F., and Tomer, K. B. (1992) *J. Am. Soc. Mass Spectrom.* **3,** 289–300.
18. Hunt, D. F., Bone, W. M., Shabanowitz, J., Rhodes, J., and Ballard, J. (1981) *Anal. Chem.* **54,** 1704.
19. Hunt, D. F., Yates, III, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 620–623.
20. Hunt, D. F., Yates, III, J. R., Shabanowitz, J., Zhu, N.-Z., Zirino, T., Averill, B. A., Larroque, S. T., Shewale, J. G., Roberts, R. M., and Brew, K. (1987) *Biochem. Biophys. Res. Commun.* **144,** 1154.
21. Johnson, R. S., and Biemann, K. (1987) *Biochemistry* **26,** 1209–1214.
22. Griffin, P. R., Kumar, S., Shabanowitz, J., Charbonneau, H., Namkung, P. C., Walsh, K. A., Hunt, D. F., and Petra, P. H. (1989) *J. Biol. Chem.* **264,** 19066–19075.
23. Biemann, K., and Scoble, H. A. (1987) *Science* **238,** 992–998.
24. Michel, H., Hunt, D. F., Shabanowitz, J., and Bennett, J. (1988) *J. Biol. Chem.* **263,** 1123–1130.
25. Hunt, D. F., Michel, H., Dickinson, T. A., Shabanowitz, J., and Cox, A. L. (1992) *Science* **256,** 1817–1820.
26. Griffin, P. R., Furer-Jonshur, K., Hood, L. E., and Yates, III, J. R. (1992) Techniques in Protein Chemistry II, (Hogue-Angeletti, R., Ed.), pp. 467–476, Academic Press, New York.
27. Mann, M., Meng, C. K., and Fenn, J. B. (1989) *Anal. Chem.* **61,** 1702.
28. Petra, P. H., Griffin, P. R., Yates, J. R., Moore, K., and Zhang, W. (1992) *Protein Sci.* **1,** 902–909.
29. Eckershorn, C., Strupat, K., Karas, M., Hillenkamp, F., and Lottspeich, F. (1992) *Electrophoresis* **13,** 664–665.