Scoring Methods in MALDI Peptide Mass Fingerprinting: ChemScore, and the ChemApplex Program

Kenneth C. Parker
Applied Biosystems, Framingham, Massachusetts, USA

No universally accepted score is currently available to determine when a matrix-assisted laser desorption ionization (MALDI) peptide mass fingerprint (PMF) experiment has been successfully carried out. We describe a software program (ChemApplex) based on a calculated parameter (Combined Protein Score) that takes into account (1) peak intensity, (2) the mass accuracy of the match, and (3) ChemScore, a theoretical intensity factor that estimates the probability of observing a particular peptide based on a combination of chemical considerations, in particular the amino acid composition of the peptide and the amino acid sequence of the amino acids that span the cleavage site. When these three factors are taken into account both at the level of individual peptides and at the protein level, protein components in mixtures whose peptides contribute less than 1% of the total intensity can often be correctly identified, as is demonstrated for mixtures of standard proteins. Moreover, it is possible to make robust database identifications that are nearly independent of the number of masses submitted and the mass error threshold used for matching. Protein scoring based on Combined Protein Score is orthogonal to many of the commonly used probability-based scoring schemes, and makes it possible to archive a more complete set of parameters that more thoroughly characterize the validity of the database match, which increases the confidence in the identifications. (J Am Soc Mass Spectrom 2002, 13, 22–39) © 2002 American Society for Mass Spectrometry

Currently, a crucial stage in the field of proteomics involves identification of proteins from 2-D gels by PMF [1–6]. When this step is successful, the protein may be deemed identified, and attention is transferred to the next sample in line. When it is unsuccessful, and especially when the mass spectrum is of a sufficient quality, tandem MS methods may be applied to obtain peptide sequence information that can be correlated to a known protein sequence. Ideally, this decision as to what constitutes successful identification should be simple and unambiguous. In reality, different scientists have different criteria as to what constitutes success at this stage, and use different scoring strategies to determine what is acceptable and how to archive the matches. The first PMF papers demonstrated that when single components were present, simple matching of theoretical peptides versus observed masses was often sufficient to correctly identify a protein [1–6]. Later developments include the use of statistical information, as used by the MASCOT program [7] and the Protein Prospector MS-FIT program [8], which in addition calculate the probability that the highest scoring protein could have been chosen at random, given the starting assumptions. However, none of the previously published methods take into account what is known about the following three factors: (1) The fine specificity of trypsin action, that is, taking account of which amino acids precede and follow the arginine and lysine. (2) The different sensitivity of detection for individual peptides by MALDI. (3) The observed degree of methionine oxidation, pyroglutamic acid formation, or degree of cysteine alkylation in the experiment in hand. All of this information can be deduced by the most compelling identifications to abundant proteins when many protein mixtures are submitted to PMF in parallel.

Currently, the most fundamental parameter that is used in PMF is the number of peptides matched. The second most important parameter is typically a threshold for mass error, often expressed in parts per million (ppm error). The third most important parameter is the percentage coverage of the primary sequence by the matched peptides (% seq). A final parameter is the number of missed cleavage sites (p for partial). A striking aspect of the current publicly available PMF

Published online November 19, 2001

Address reprint requests to Dr. Kenneth C. Parker, Applied Biosystems, 500 Old Connecticut Path, Framingham, MA 01701, USA. E-mail: parkerkc@appliedbiosystems.com

1044-0305/02/$20.00
PII S1044-0305(01)00320-8

Received June 12, 2001
Revised August 15, 2001
Accepted August 15, 2001

programs is that the intensity of the peaks is not used directly in the scoring. Instead, it is tacitly assumed that the user knows which masses have meaningful intensities, and that this determination is straightforward. Most programs perform at their best if the user chooses a small number of masses, say between 5 and 50. A second omission is that the only factor that is used in PMF which deals with the chemistry of the peptides is in the parameter p, the number of missed cleavages. No consideration is taken for which cleavage sites are missed, which peptides are found, or which peptides are not detected that ought to be found. Typically, the user can specify whether to check for matches with peptides containing oxidized methionines or alkylated cysteine residues, but cannot in any way quantify expectations. In this article, we describe a software program called ChemApplex that takes into account intensity and chemical expectations (quantified as ChemScore). We believe that the result is more secure database identifications than would otherwise be possible. The higher the confidence in identifying the primary component in a mixture, the more practical it becomes to identify additional components. To this end, ChemApplex attempts to allocate the remaining unexplained masses to additional components in the protein database, as many times as desired. This feature makes ChemApplex useful in identifying and quantifying proteins in mixtures. In addition, the attention paid to intensity makes ChemApplex less sensitive to large peaks lists, which usually are a problem for other PMF programs. ChemApplex also takes into account the degree of mass error, making ChemApplex relatively insensitive to the mass tolerance threshold. Because of the quantitative information that is calculated, ChemApplex can be used to determine which of several spectra have the highest information content. The ideal peptide mass fingerprinting program might well combine the features included in ChemApplex with information about probabilities of correct identification as calculated by other PMF programs [7, 8] as well as other database and spectrum parameters [9].

Terms and Considerations for Database Matching

The following discussion describes the input database searching parameters that can be set by the operator, and the calculated parameters that have been found effective in identifying the best protein matches in the database. The concept of ChemScore is central to this approach. A special set of input parameters, listed in Table 1, can be used to calculate the ChemScore for any peptide given its primary sequence and its location within a protein sequence. The 23 other input parameters specific to database searching are listed in Table 2, along with a brief description of the parameters. First, a predigested peptide database for the organism(s) of interest and encompassing the desired peptide *m/z* range is generated, which requires the ChemScore input parameters in Table 1, and parameters 9, 16, and 17 of Table 2. Second, decisions have to be made regarding peak detection at the level of the mass spectrum, which is controlled by parameters 18 and 19. Third, one must decide how to conduct the search process; that is, how many masses are to be required for a protein to be considered, the mass tolerances to be employed, and which proteins from the database are to be considered. These functions are controlled by parameters 1–15. Finally, there is the problem of how to perform protein sorting; in particular, how to deal with the problem of masses that can be accounted for by more than one protein, which is controlled by parameters 20–23.

Internal Calibration

When MALDI mass spectra are collected, there are usually significant, consistent differences between the observed masses and the theoretical masses (when the theoretical masses are known). The accuracy of MALDI mass spectrometers is such that the mass error is commonly reported in parts per million (ppm). Because the offset in ppm is usually a linear function of mass, and many masses are usually tentatively identified, the validity of a tentative identification can most easily be tested by performing internal calibration on the theoretical masses even in the case of an unknown. Ideally, this internal calibration procedure can also take into account any other known molecular masses, like those corresponding to doped-in calibrants or trypsin autodigestion peaks. When the evidence is strong for the identification of a major component in a mixture, recalibration makes it possible to identify minor components because the ppm tolerance can be set lower at the level of database searching. After internal calibration, we commonly use the values of parameters 1–15 listed in Table 2 column "Used" to perform mass searches in two stages (see below). The high mass accuracy obtained by internal calibration restricts the number of source proteins to a manageable number even when as few as two peptides are required for matching.

Two Stage Searches

First, for a protein to be considered further, at least one peptide (Table 2, parameter 1) must match within 15 ppm (parameter 2), and it must have a ChemScore of at least 9 (parameter 3), and must also be one of the 100 most intense masses (parameter 4) that were detected. At the second stage, at least one additional peptide must match (parameter 5) within 25 ppm (parameter 7), and must be one of the 200 most intense masses (parameter 8). Because of this feature, very minor protein components can be tentatively identified, so long as one expected peptide is matched to a relatively intense mass with high mass accuracy. This is especially important when PMF is applied to small proteins in complex mixtures; for example, proteins that have been

Table 1. ChemScore rules

Rule	Peptide[a]	ChemScore[b]	Used[c]
	Initial settings		
	To whichever of these that first applies:		
1	Arg containing	100	yes
2	VP-Cys containing	100	no
3	Lys-containing	10	yes
4	Biotinylated-Cys containing	10	no
5	None of the above	1	yes
	Detrimental cysteine-containing peptides		
	Whenever this applies:		
6	Cys free SH	÷ 10	
7	Cys acrylamide adduct	÷ 10	yes
	Methionine-containing peptides		
8	Methionine Oxidation Factor (MetOxF)[d]	0.2	
9	Count No. reduced Met = R		
10	Count No. oxidized Met = X		
11	If MetOxF >1	÷ $MetOxF^R$	yes
12	If MetOxF <1	× $MetOxF^X$	yes
13	If MetOxF = 1	÷ 2	yes
	Proline at beginning of peptide[e]		
14	P^	÷ 100	yes
	Missed cleavage rules		
15	1 missed cleavage	÷ 100	yes
	To whichever of these that apply:		
16	^[KR]P^[e]	× 100	yes
17	[KR]*	× 30	yes
18	^[DE][KR]*	× 20	yes
19	^[KR][DE]^	× 20	yes
20	^[KR][ILV]^	× 5	yes
21	^[KR][KRX]	× 3	yes
22	^[DE] X[KR]*	× 2	yes
23	^[KR] X[DE]^	× 2	yes
24	X[KR]*	× 2	yes
25	^[KR]X[KRX]	× 1.5	yes
	Terminal peptide adjustments		
26	[DE]^	× 0.9	no
27	[ILV]^	× 0.95	no
28	^[DE][KR]	× 0.9	no

[a]^ means 0 or more aa of any kind; * means at least one aa of any kind; X means any one aa; aa in brackets[ ] are alternatives. In rules 21 and 25, [KRX] indicates any aa, which will usually be K or R, but may be X in the case of a C-terminal peptide.
[b]The ChemScore is assigned by rules 1–5, and further manipulated by the remaining rules.
[c]Some of these rules have not yet been implemented, and are listed as possibilities.
[d]The MetOxF is assigned based on whether Met in peptides in the same experiment is largely oxidized or reduced. The range is between 0.2 and 5.0. 5.0 means strongly oxidized, 0.2 means strongly reduced. 1.0 means half oxidized, half reduced.
[e]Note cleavage of [KR]P is considered 1 missed cleavage; intact [KR]P is considered 0 missed cleavage, thus [KR]P does not impact the Partial Cleavage Factor.

isolated from sodium dodecylsulfate (SDS) gels with an apparent molecular weight of less than 15K.

Peak Detection and Intensity Matched

One of the most valuable pieces of information in a mass spectrum is the intensity of each of the observed masses. It is always desirable in performing PMF to account for as many of the masses as possible, starting with the most intense masses. To accomplish this, one needs to have a reliable peak detection algorithm, followed by appropriate smoothing, and de-isotoping, in which the masses and intensities that correspond to the different isotopic forms of each peptide are reduced to two numbers: the mass of the monoisotopic form of the molecular ion, and the sum of the intensities of all of the isotopic forms of the molecule. Once this has been performed, an important parameter that characterizes a given mass spectrum is the sum of the intensity of all of the detected isotope clusters. Ideally, the peak detection process ought to be dependent on statistical factors that exclude electronic noise and peaks for which the signal-to-noise ratio is below a given threshold, for example, a ratio of 4.0. This works fine in many instances, but not so well when the low mass region of the spectrum is dominated by real peaks that derive from matrix-related clusters. To get around this problem, one can perform two stages of peak detection. In the first stage, prior to de-isotoping, peak detection is set wide open, and a large number of peaks are detected (often many thousand). The peak list is then reduced to a much smaller one by retaining only the most intense 30 (see

Table 2. ChemApplex input parameters

	Parameter name	Values			Description
		Used[a]	Minimum[b]	Maximum[c]	
	Top peptide parameters				
1	Minimum number peptides to match	1	1	4	Number of intense peptides that must match for protein to be considered.
2	Maximum ppm error	15	5	50	Ppm error between theoretical and experimental.
3	ChemScore minimum	9	1	20	ChemScore required.
4	Lowest Intensity Rank	100	30	100	Rank required.
	All peptide parameters				
5	Minimum number of peptides to match	2	2	5	Total number of peptides that must match.
6	Minimum ppm error	2	1	50	Minimum ppm error to use in calculations. Peptides that match better than this are considered to match perfectly.
7	Maximum ppm error	25	10	200	Maximum ppm error for peptide to be reported in table.
8	Maximum number of peptides	200	50	500	Maximum number of peptides from mass list (sorted by intensity) to be considered. If more peptides than this are detected, only the 200 most intense are used.
	Database				
9	Organisms	*E. coli*, chicken, cow			Organisms from SwissProt release 39.2 to be searched.
10	Number of Proteins searched	7261			Number of proteins in database.
11	Number of Peptides searched	81487			Number of peptides in database.
	Protein parameters				
12	Lowest	1 K	0	100	Lower MW limit for proteins.
13	Highest	3000 K	50	9000	Upper protein MW limit for proteins.
14	Fragment	1000 K	5	200	MW of the fragment of the protein that must contain all of the peptides that fulfill matching criteria. Turned off in these experiments.
15	Minimum % ChemScore Matched	20	10	40	Minimum ChemScore required for protein to be considered further.
	Peptide parameters				
16	Lower *m/z* limit	800	500	1200	Lower *m/z* limit for peptides to be searched, and for calculation of Protein ChemScore.
17	Upper *m/z* limit	3600	3000	6000	Upper *m/z* limit for peptides to be searched, and for calculation of Protein ChemScore.
18	Masses per 100 u (detect)	30	20	80	Maximum number of peptides that can be detected every 100 u before de-isotoping.
19	Masses per 100 u (deisotope)	8	3	15	Maximum number of peptides that can be detected every 100 u after de-isotoping.
	Sorting				
20	No. masses to truncate in sort	1	0	3	Number of Peptide TriScores to truncate to calculate Protein TriScore.
21	Min ChemScore SortOut	5	1	20	Minimum ChemScore of peptides from a higher protein on the sort list required before the corresponding mass will be removed from the MALDI mass list.
22	Max ppm SortOut	25	10	50	Maximum ppm required to match the peptide of a higher ranking protein before the mass is removed.
22	Loss Factor SortOut	2500	1	10,000	The attenuation factor for previously matched intensities and ChemScores.
23	Iterations to Sort	50	5	All found	Number of iterations of removal of peptides from the mass list. When set to 50, the program removes peptides from the top 50 proteins from the mass list, and then stops.

[a]The value of the parameter that was used in the other tables.
[b]A reasonable lower limit value for the parameter.
[c]A reasonable upper limit value for the parameter.

Table 2 parameter 18) masses every 100 aμ, followed by subtracting the intensity of the 31st peak from the top 30 peaks, thus eliminating chemical noise in a mass region-dependent fashion. Parameter 18 can be set to a number higher than 30 for particularly rich samples. The reduced mass list is then submitted to a de-isotoping tool, which functions much better after most of the chemical noise has been eliminated. After de-isotoping, the peak list is reduced again by retaining the most intense 8 (parameter 19) masses every 100 u. To eliminate meaningless or insignificant mass signals from chemical noise, the intensity of the ninth most intense isotope cluster can then be subtracted by each of the remaining 8 peaks in the 100 u mass window. In this fashion one obtains a spectrum intensity value that is only slightly dependent on the peak detection criteria.

For every possible database match, one can now sum the intensities of the peaks that are accounted for, and calculate the parameter % Intensity Matched (see Table 3 column 10 "% Intensity Matched").

ChemScore

Definition of ChemScore. To take advantage of what is known about the chemistry of peptide ionization, the fine specificity of polypeptide cleavage by trypsin [10] and other chemical aspects of the experimental system [11], we introduce the concept of ChemScore, which is designed to be a score that reflects the theoretical detection probability for individual peptides under a specific set of experimental conditions, in this case, MALDI PMF of trypsin-digested proteins. This score is calculated from the peptide's sequence and the neighboring amino acids, and a series of user-adjustable parameters that reflect the experimental conditions (Table 1). We have found that the parameters listed in Table 1 are useful in predicting the intensities of peptides that are observed in many trypsin digests. The parameters are based on four separate considerations: (1) Some peptides are detected more efficiently by MALDI than others; most importantly, arginine-containing peptides are detected with greatest sensitivity [12], followed by lysine-containing peptides. Chemical modifications to natural peptides often have a profound impact on detection, and need also to be taken into account. For example, pyridylethylation by vinyl pyridine [13] promotes detection, whereas introduction of sulfonic acid groups onto peptides significantly diminishes detection [14], at least in positive-ion mode. (2) Some peptides are generated more efficiently than others by trypsin digestion. Depending on the amount of trypsin added, the digestion may proceed to differing degrees of completion. (3) Some peptides become modified during the experiment, in particular, methionines sometimes become oxidized, some N-terminal glutamine-containing peptides are converted partially to the pyroglutamic acid form, and cysteine-containing peptides may be differentially recovered based on whether they were alkylated, how efficiently they were alkylated, and what alkylating reagent was used. (4) Finally, in most cases, the spectrum usually has a lower *m/z* limit and upper *m/z* limit; peptides whose masses fall outside of that range have a ChemScore of 0. These rules have been incorporated into the ChemApplex program so that the ChemScores for all tryptic peptides can be automatically calculated (see e.g., Table 4 column 3 "Ch").

Terminal tryptic peptides versus missed cleavage containing peptides. With the ChemApplex program, peptides are first categorized according to whether they have an arginine residue or not, and if not, whether they have a lysine residue. Thus, in Table 4, five peptides are assigned a ChemScore of 100 because they are terminal arginine-containing peptides. In addition, peptides are tagged as to whether or not they have a missed tryptic cleavage site. (Because KP and RP bonds are rarely cleaved by trypsin, KP and RP sequences do not count as missed cleavages, as is conventional. We have adopted the additional convention that peptides which are nonetheless generated by cleavage at KP or RP bonds should be considered to have a missed cleavage because such cleavages are so infrequent). A series of string searches are then carried out on the sequence of each peptide to detect the presence of certain motifs that may affect cleavage by trypsin, as listed in Table 1. Furthermore, if more than one of these missed cleavage motifs are present in the same peptide, then the ChemScore for the peptide reflects all such motifs. At the same time, the ChemScore of a peptide containing a missed cleavage never exceeds that of a peptide with the same composition but no missed cleavage.

The ChemScore Partial Factor. This is accomplished mathematically by the parameter ChemScore Partial Factor (ChPF). Terminal trypsin cleavage products are assigned the Arginine ChemScore (e.g., 100, Table 1, rule 1) if they contain at least one arginine, and the Lysine ChemScore (e.g., 10, rule 3) if they contain no arginine but contain at least one lysine, and the Basal ChemScore (e.g., 1, rule 5) if they contain neither arginine nor lysine. The ChemScore of a peptide with missed cleavages is then divided by ChPF, which is a function of the Basal Missed Cleavage Factor (BMCF), typically 100 (rule 15), and the product of the Missed Cleavage Factors (MCF) that apply to the peptide in question (rules 16–25), according to the equation

$$\mathrm{ChPF} = [\mathrm{BMCF} + \Pi(\mathrm{MCF})]/\Pi(\mathrm{MCF}). \quad (1)$$

For example, in Table 4, the second peptide has sequence RHGLDNYR, to which rule 17 applies, because of the N-terminal Arg residue. Because no other missed cleavage rule applies to this peptide, the ChPF is therefore (100 + 30)/30, which causes the ChemScore for the peptide to equal about 23. As a second example,

Table 3. Mixing tryptic digests of lysozyme, ovalbumin and BSA

Number of identifications[a]	Number of masses submitted	Protein name	Protein molecular weight	Unique peptide matches[b]	All peptide matches	% ChemScore Matched[c]	Protein-Based Protein TriScore[c]	Combined Protein Score[c]	% Intensity Matched[c]	PPW[c]	Arg % Intensity[d]	Lys % Intensity[e]
1	2	3	4	5	6	7	8	9	10	11	12	13
	Experiment 1: 5 pmole lysozyme : 5 pmole ovalbumin : 5 pmole bsa											
1	200	bsa	67162	7	11	54.8	911	191052	24.9	3.0	60.3	1.6
1	200	ovalbumin	42723	11	15	65.6	602	132758	22.9	5.0	78.9	0.1
1	200	lysozyme	16229	5	7	79.7	314	90836	11.8	6.0	99.6	0.0
1	200	trypsin	23291	3	4	45.1	15	19518	0.5	3.0		
1	200	hypothetical 43.6 kda protein	43557	4	6	18.7	7	951	1.4	7.0	30.3	0.0
	Experiment 2: 2 pmole lysozyme : 2 pmole ovalbumin : 2 pmole bsa											
1	200	lysozyme	16229	6	8	89.3	694	1383619	11.7	3.0	98.0	0.0
1	200	bsa	67162	7	13	55.0	1424	775400	25.9	2.0	48.3	4.3
1	200	ovalbumin	42723	11	14	65.6	818	724761	24.9	4.0	80.6	0.3
1	200	trypsin	23291	3	4	46.7	10	12040	0.5	5.0		
1	200	chloramphenicol acetyltransferase	25647	2	4	38.7	4	201	1.0	22.0	51.1	13.6
	Experiment 3: 0.5 pmole lysozyme : 5 pmole ovalbumin : 5 pmole bsa											
11	198	bsa	69249	7.7	15.5	51.4	1190	490561	27.4	2.4	61.3	2.1
11	198	ovalbumin	42723	10.4	13.8	65.5	939	404130	25.4	3.5	77.2	0.2
9	198	trypsin	23291	2.3	3.9	43.4	9	6519	0.6	5.9		
8	198	lysozyme	14305	2.9	3.1	47.7	13	4350	0.7	5.0	99.6	0.0
1	183	hemolysin-activating lysine-acyltransfer	19758	3	3	30.3	10	2361	0.3	2.0	5.4	0.0
	Experiment 4: 0.2 pmole lysozyme : 2 pmole ovalbumin : 2 pmole bsa											
9	196	ovalbumin	42723	10.7	13.4	64.5	924	746996	26.3	3.7	79.9	1.7
9	196	bsa	67162	8.0	16.7	55.4	1194	572559	27.6	2.6	54.4	4.3
4	192	lysozyme	14305	2.8	3.0	45.4	15	16038	0.6	3.8	100.0	0.0
8	196	trypsin	23291	3.4	4.9	46.8	13	7804	1.0	7.6		
1	168	arylamine n-acetyltransferase	32895	2	2	33.4	7	2411	0.5	5.0	100.0	0.0
	Experiment 5: 5 pmole lysozyme : 0.5 pmole ovalbumin : 5 pmole bsa											
11	152	lysozyme	16229	5.6	8.5	82.5	1664	456743	22.9	2.3	99.4	0.0
11	152	bsa	69249	8.6	17.5	60.5	2144	380566	51.5	2.9	56.5	2.8
11	152	ovalbumin	42723	5.4	6.1	37.5	21	3513	1.4	5.2	89.5	0.0
10	153	trypsin	23291	4.6	4.9	53.3	9	2233	0.7	8.6		
1	200	hypothetical acetyltransferase	19237	3	3	32.1	7	487	0.6	6.0	0.0	0.0
	Experiment 6: 2 pmole lysozyme : 0.2 pmole ovalbumin : 2 pmole bsa											
1	200	lysozyme	16229	6	9	89.3	873	1146155	14.7	3.0	98.3	0.0
1	200	bsa	67162	7	17	55.8	2281	1139575	40.9	2.0	49.7	7.4
1	200	trypsin	23291	4	5	49.1	16	7261	2.3	14.0		
1	200	ovalbumin	42723	5	7	28.9	16	6499	1.9	7.0	84.9	0.0
1	200	hypothetical 23.1 kda protein in glne-cc	23063	2	3	23.4	2	1261	0.2	4.3	50.0	0.0
	Experiment 7: 5 pmole lysozyme : 5 pmole ovalbumin : 0.5 pmole bsa											
1	200	lysozyme	16229	7	11	93.0	1942	2326762	20.9	2.0	97.9	0.0
1	200	ovalbumin	42723	12	16	69.5	1854	1162386	40.0	3.0	74.7	0.0
1	200	bsa	69249	7	10	38.4	27	8700	1.8	5.0	47.1	1.5
1	200	trypsin	23291	3	4	45.1	13	7687	1.0	7.0		
1	200	4-hydroxy-2-oxovalerate aldolase	36448	3	4	25.4	25	764	7.3	15.0	6.1	0.0
	Experiment 8: 2 pmole lysozyme : 2 pmole ovalbumin : 0.2 pmole bsa											
12	177	ovalbumin	42723	12	15	68.2	1308	636823	38.4	4.0	85.4	1.7
12	177	lysozyme	16229	5	7	74.0	417	500896	8.0	2.8	95.3	0.0
10	182	trypsin	23291	3	4	47.1	10	6948	1.0	9.6	0.0	0.0
6	162	bsa	67162	4	5	34.6	17	2432	1.7	7.0	63.2	0.4
2	186	4-hydroxy-2-oxovalerate aldolase	36448	3	4	25.4	18	2284	5.0	14.5	20.6	0.0

[a]The number of times that the protein was identified. For each experiment, the largest number equals the number of spectra that were analyzed.
[b]The number of masses that matched the protein that were not accounted for by higher ranking proteins and that fulfilled the ChemScore and mass accuracy requirements of Table 2 parameters 21 and 22.
[c]calculated using the peptides from column "unique peptide matches."
[d]The Percentage of the Matched Intensity that was attributable to terminal, arginine-containing peptides.
[e]The Percentage of the Matched Intensity that was attributable to terminal lysine-containing but not arginine-containing peptides.

Table 4. Lysozyme peptides

Theo[a]	No. obs[b]	Ch[c]		Sequence		Mod[f]
874.42	6	100.0	R[d]	HGLDNYR	G[e]	
1030.52	3	23.3	K	RHGLDNYR	G	
1041.44	7	5.0	R	WWCNDGR	T	VP
1045.54	31	100.0	K	GTDVQAWIR	G	
1373.67	13	0.5	R	GYSLGNWVCAAK	F	VP
1428.65	28	100.0	K	FESNFNTQATNR	N	
1675.80	11	100.0	K	IVSDGNGMNAWVAWR	N	
1753.84	25	100.0	R	NTDGSTDYGILQINSR	W	
1803.90	8	30.1	K	KIVSDGNGMNAWVAWR	N	

[a]The calculated (or theoretical) *m/z* ratio of the peptide.
[b]The number of times the peptide was detected among the 47 spectra.
[c]The ChemScore for the peptide, calculated using Table 1 rules 1–25.
[d]The single letter amino acid code for the amino acid preceding the peptide.
[e]The single letter amino acid code for the amino acid following the peptide.
[f]The chemical modification status of the peptide. VP indicates vinylpyridine-modified cysteine.

for a peptide whose sequence is DKLDAAAK (not in Table 4), the peptide would get an initial ChemScore of 10 for having at least 1 lysine residue but no arginine. According to eq 1, the ChPF would equal $[100 + (20 \times 5 \times 2 \times 2)]/[(20 \times 5 \times 2 \times 2)] = 1.25$, because it contains the motifs from lines 18, 20, 23, and 24 in Table 1. Hence the overall ChemScore for this peptide would be 8. If a peptide contains more than 1 missed cleavage, then the initial ChemScore for the peptide is divided by the product of the ChPFs for each missed cleavage calculated separately.

N-terminal depletion factor and other terminal peptide adjustment factors. Similarly, one can adjust downward the ChemScores of terminal peptides by some factor if the N-terminal or C-terminal cleavage sites are known to be unfavorable. This is of greatest importance when a terminal trypsin digestion peptide is preceded by two

Table 5. Peptides matched to lysozyme from Experiment 4

No.[a]	Maldi[b]	Theo[c]	Ppm[d]	Intensity[e]	Rank[f]
1	1045.53	1045.54	−12.4	104	83
1	1428.65	1428.65	−2.1	241	44
1	1753.83	1753.84	−6.3	513	27
2	1753.85	1753.84	3.4	260	57
3	1041.48	1041.44	38.4	121	143
3	1753.83	1753.84	−6.3	620	58
4	1041.44	1041.44	3.8	347	97
4	1045.53	1045.54	−9.6	116	158
4	1753.84	1753.84	−3.4	888	56
5	1428.63	1428.65	−11.9	423	79
5	1753.84	1753.84	−2.9	1115	43
6	1041.45	1041.44	14.4	594	110
6	1045.54	1045.54	−3.8	1431	66
6	1428.66	1428.65	3.5	4667	42
6	1753.84	1753.84	−1.7	4860	41
7	1041.43	1041.44	−7.7	92	111
7	1045.51	1045.54	−33.5	221	64
7	1428.67	1428.65	11.2	137	92
8	1041.45	1041.44	8.6	421	84
8	1045.57	1045.54	25.1	199	121
8	1753.84	1753.84	0.0	1241	43
9	1041.44	1041.44	3.8	887	67
9	1045.57	1045.54	24.9	409	107
9	1428.65	1428.65	−0.7	1186	55
9	1753.83	1753.84	−6.8	1045	63

[a]The spectrum ID number, from Table 3 Experiment 4.
[b]The measured *m/z* ratio after internal calibration.
[c]The calculated *m/z* ratio for the peptide.
[d]The mass error in parts per million.
[e]The measured intensity for the isotope cluster in question.
[f]The rank of the peptide upon sorting by intensity.

basic residues. In this case, depending on the amount of trypsin added and the incubation time, the terminal peptide may not be generated very efficiently, and may instead remain with an N-terminal basic residue, as has been noted previously many times. In cases where cleavage is clearly incomplete in this regard, taking this factor into consideration aids in the interpretation of the data. We have not found it advantageous for database searching purposes to decrease the ChemScores of other terminal peptides that are the counterparts of the other special cases of missed cleavages, though such rules ChemScore (rules 26–28) are useful for studying the extent of tryptic digestion of known proteins, because the expected trypsin digestion products of a protein can be sorted by ChemScore, and these rules can be used to subcategorize terminal tryptic peptides with special characteristics.

Special treatment of methionine: the Methionine Oxidation Factor. Methionine-containing peptides present a special case, because they are often recovered in two forms: unaltered and methionine-sulfoxide containing. The degree of oxidation is dependent on the experimental conditions; for example, we and others have observed that when proteins are isolated from gels, the intensities of methionine sulfoxide–containing peptides are usually more intense than the native peptide, whereas in solution digests, the opposite is usually the case. When many methionine-containing peptides are studied, the ratio of the intensity of the methionine sulfoxide-containing peptide to the reduced form is roughly constant. The peptide ChemScore can be appropriately adjusted by means of the Methionine Oxidation Factor (MetOxF). If 80% of the peptide is found to be oxidized, then the ChemScore of the reduced form of the peptide ought to be divided by the MetOxF, which ought to equal about 5.0. If there are two methionines in the peptide, then the ChemScore of the theoretical peptide containing one reduced methionine would be reduced by 5-fold, whereas the theoretical peptide containing both reduced methionines ought to be reduced by 25-fold. The same logic applies in reverse if the peptide digest is mostly reduced to begin with: in that case the ChemScore of the theoretical peptide containing two oxidized methionines would be reduced by the MetOxF squared. We have adopted the convention that a MetOxF >1 decreases the ChemScore for peptides containing reduced methionine, whereas a MetOxF <1 decreases the ChemScore for peptides containing oxidized methionine. Note that if both forms of a methionine-containing peptide are matched, the current version of the program counts them separately, and does not give any particular bonus to having matched two masses that differ by the mass of an oxygen atom to a peptide that contains methionine, though one could argue that this might be desirable.

Cysteine alkylation effects. A similar logic applies to cysteine residues. Some alkylating agents like vinyl pyridine cause peptides to ionize as well as arginine, thus the ChemScore of a peptide containing pyridylethylated cysteine ought to be the same as arginine. Biotinylated iodoacetamides to a lesser degree promote ionization, whereas iodoacetamide and iodoacetic acid alkylation are neutral with respect to ionization but promote peptide recovery. If a protein is not alkylated at all, then peptides derived from it are usually recovered from gels as the acrylamide adduct in low yield, or in the reduced form in low yield, whereas in solution protein digests these peptides are often recovered efficiently as reduced cysteine. These expectations are implemented by a series of user-adjustable Cysteine Factors. Vinylpyridine labeled and biotinylated-cysteine must be dealt with at the top level (Table 1 rules 2 and 4), whereas the extent of recovery issue is dealt with separately (rules 6 and 7). In this fashion, one can deal with a situation in which vinyl pyridine alkylation was intended, but was not very efficient, as in the experiments described in Tables 3, 4, 5, 6, and 7 (see below).

Table 6. Summary of lysozyme data from Table 5

No. obs[a]	*m/z*[b]	Rank (ave)[c]
6	1041.45	102.0
6	1045.54	99.8
5	1428.65	62.4
8	1753.84	48.5

[a]The number of times the peptide (from lysozyme) was detected in Table 3, Experiment 4.
[b]The theoretical *m/z* ratio of the lysozyme peptide.
[c]The average intensity rank for the peptide, from those spectra in which it was detected.

Pyroglutamic acid. The degree of pyroglutamic acid formation appears to depend on how the digest is treated, and the ChemScores of peptides containing pyroglutamic acid can therefore be adjusted by the Pyroglutamic Acid Factor. Apparently, in some cases, digests contain little pyroglutamic acid, whereas in other cases, the peptides that contain an N-terminal glutamine are split about 50:50 between the pyroglutamic acid form and the unreacted glutamine form.

Other considerations. In principle, similar factors could be developed for other chemical modifications. For example, tryptophan residues are sometimes oxidized (either singly or doubly), and peptides that contain adjacent asparagine–glycine residues are often dehydrated to aspartimide, and hydrolyzed back to aspartic acid (either to the natural aspartic acid or to form an isopeptide bond). Certain glutamines and asparagines get partially deamidated. Conversion of lysines to guanidyllysines promotes detection of lysine-containing peptides [15, 16]. In all of these cases, it should be possible to add new ChemScore rules to reflect the observed apportioning of intensity between the various chemically modified forms. Correctly accounting for all

Table 7. Protein rank versus sorting parameter[a]

SpecID[b]	Rank[c]	M[d]	%Ch[e]	PBPT[f]	%I[g]	PPW[h]	No. Proteins[i]
Data for lysozyme							
1	3	5-11	3	3	15	8	21
2	not found						20
3	not found						22
4	16	9-18	12	16	31	5	35
5	9	9-16	5	9	22	7	32
6	3	5-8	5	3	16	3	24
7	not found						28
8	20	10-13	10	20	27	12	31
9	4	6-9	4	4	16	7	21
average	9.2	10.8	6.5	9.2	21.2	7.0	27.3
Data for ovalbumin							
1	1	2	1	2	2	4	21
2	1	2	3	1	2	3	20
3	2	2	1	2	2	6	22
4	1	2	2	2	2	6	35
5	2	2	1	2	2	3	32
6	1	2	3	2	2	5	24
7	2	2	2	2	1	5	28
8	1	2	3	2	2	3	31
9	1	2	1	2	2	4	21
average	1.3	2.0	1.9	1.9	1.9	4.3	26.0
Data for BSA							
1	2	1	2	1	1	3	21
2	2	1	4	2	1	1	20
3	1	1	2	1	1	2	22
4	2	1	4	1	1	4	35
5	1	1	4	1	1	1	32
6	2	1	4	1	1	4	24
7	1	1	3	1	2	3	28
8	2	1	4	1	1	1	31
9	2	1	2	1	1	1	21
average	1.7	1.0	3.2	1.1	1.1	2.2	26.0
Data for trypsin							
1	6	3	8	17	10	22	21
2	3	4	5	3	8	18	20
3	3	3	5	3	10	12	22
4	3	6	5	3	16	19	35
5	3	3	6	4	13	20	32
6	4	5-8	6	7	12	21	24
7	3	5-8	5	3	11	22	28
8	3	3-4	5	5	19	22	31
9	4	3	3	5	10	21	21
average	3.6	5.5	5.3	5.6	12.1	19.7	26.0

[a]The sorting parameter refers to one of the parameters in columns 3–7, which was used to determine the ranking of the proteins.
[b]The spectrum ID number, from Table 3, Experiment 4, which corresponds to an individual spectrum.
[c]The rank of the proteins upon sorting by Combined Protein Score.
[d]The rank of the protein upon sorting. When more than one number is listed, more than one protein had the same rank. Thus for lysozyme in SpecID 1, seven different proteins had the same rank of between 5 and 11 when the proteins were sorted solely by number of peptides matched.
[e]The rank of the protein when sorted by % ChemScore Matched only.
[f]The rank of the protein when sorted by Protein Based Protein TriScore only.
[g]The rank of the protein when sorted by % Intensity Matched only.
[h]The rank of the protein when sorted by PPW (intensity-weighted average ppm error).
[i]The number of proteins that passed the minimal protein matching requirements.

of the chemically modified peptides is expected to promote database identification, especially from complex mixtures (see below). Failure to do so leaves peptides in the mass list after identification of the primary component(s) which the program will attempt to match to additional proteins. If too many chemical modifications are considered, then masses will be removed from the mass list that are due to other proteins, diminishing the likelihood that these other proteins will be identified.

Protein ChemScore. Another concept that follows from the Peptide ChemScore concept is the Protein ChemScore. This is defined as the sum of the ChemScores of

the peptides being considered. Peptides that would be impossible to detect are defined as having a ChemScore of zero; thus if one is collecting data between *m/z* 800 and 3600, the only peptides that contribute to the ChemScore of a protein have masses within this range. Obviously, Protein ChemScore is dependent on all of the ChemScore input parameters. It can be thought of as similar to an extinction coefficient for MALDI analysis. We are investigating how useful Protein ChemScore is in quantification of protein mixtures.

Percent ChemScore Matched. It is useful to track the validity of a PMF database identification by calculating % ChemScore Matched (see Table 3 column 7). It is similar to the % primary sequence covered parameter that is calculated by other PMF programs, and both of these parameters have independent advantages. A protein with a high % ChemScore Matched is weighted higher (see Combined Protein Score below) than another protein with a larger number of matching peptides with a lower % ChemScore Matched, making it easier to identify smaller proteins. In addition, a protein that has few arginine residues is more easily detected using % ChemScore Matched, because a higher value is placed on the identification of the few arginine containing peptides that are invariably detected whenever the protein is present. This parameter is so powerful that we have invoked it as an additional filter at the level of protein matching: if desired one can demand that proteins be further considered only if they have a minimal % ChemScore Matched of >20%.

ChemScore conclusions. To reiterate, the ChemScore value of each peptide represents a combination of factors that have a chemical basis, including ionization efficiency, degree of trypsin digestion, and degree of chemical modification. Luckily, for digests that are processed in parallel, we have found that most peptides appear to be modified to a similar degree, making it useful to adjust some of the ChemScore Factors according to experimental observations. This would be a circular argument if the ChemApplex program was needed to identify the highest scoring components. In practice, it is not a problem because the primary components of even complex mixtures can often be identified with high confidence with other software.

We expect that in many cases, it may be found that the use of ChemScore is not important for making an initial identification of a protein. But because ChemScores allow the users to sort the expected peptides by chemical properties, one can easily identify peptides which ought to be detected that are missing, which might therefore be chemically modified. In cases of borderline identifications, the choice of input ChemScore parameters can make a lot of difference in suggesting the presence of minor components. Until more MS-MS data from these digests is analyzed, we cannot be sure to what degree these suggestions are correct; however, in many cases the same peptides that are detected in digests in which the protein is a minor component are the most prominent peptides in digests in which the protein is a major component. In the case of major components, ChemScore is useful in deciding which of several candidate peptides is most likely to explain the mass observed.

Peptide TriScore and Minimum Ppm Error

After internal calibration is performed, Peptide TriScore can be defined as follows, taking into account the three most relevant parameters for confidence in individual peak assignment:

$$\text{Peptide TriScore} = \frac{\text{Intensity} \times \text{ChemScore}}{\text{Abs(Ppm Error)}} \tag{2}$$

where Abs( ) indicates the absolute value of the Ppm Error is to be used. To prevent division by zero, a minimum value for Ppm Error needs to be defined. This can be accomplished either by postulating that Ppm Error must always be greater than a Minimum Ppm Error, e.g., 2 ppm (as in Table 2 parameter 6), or one can add a Minimum Ppm Error term, e.g., 2 ppm to the ppm difference between the theoretical and experimental masses so that:

$$\text{Peptide TriScore} = \frac{\text{Intensity} \times \text{ChemScore}}{(\text{Abs(Ppm Error)} + \text{Minimum Ppm Error})} \tag{3}$$

This distinction has minimal impact on the results. The choice of the value for this Minimum Ppm Error is dependent on instrument parameters and on whether or not internal calibration was performed.

Peptide-Based Protein TriScore

The same logic can be extended to an intact protein. The sum of the Peptide TriScores for all peptides that match to a given protein is defined as the Peptide-Based Protein TriScore. Note that individual peptides often make very different contributions to the Peptide-Based Protein TriScore. In contrast, in scoring schemes other than with ChemApplex software, each matching peptide is accorded equal weight, or perhaps some lower weight is assigned to peptides with missed cleavages. The ChemApplex program is designed to accord each peptide a weight that is roughly proportional to the incremental credibility by which that peptide strengthens the database identification. Thus, little weight should be given to peptides that match poorly, that have relatively low intensity, or that are found by experimentation to be difficult to detect.

Protein-Based Protein TriScore

A second way to calculate Protein TriScore is at the whole protein level, using the same logic as above. For a whole protein, one can in addition normalize the Intensity and the ChemScore terms. One way to accomplish this is to convert the intensity to % Intensity Matched, and to convert ChemScore to % ChemScore Matched. As was the case for Peptide TriScore, it is also desirable to define Protein Error so that it cannot be less than the Minimum Ppm Error. A simple way to accomplish this is to add the Minimum Ppm Error to the Average (Absolute) Ppm Error for all the peptides that match, and then normalize:

$$\text{Protein-Based Protein TriScore} = (\%\ \text{Intensity}) \times (\%\ \text{ChemScore})/\text{Protein Error} \quad (4)$$

where Protein Error = (Average Ppm Error + Minimum Ppm Error)/Minimum Ppm Error.

In this case, a protein whose peptides all matched perfectly would be assessed at 100%. If the average mass error was 4 ppm and Minimum Ppm Error was 2 ppm, then the Protein-Based Protein TriScore would be reduced two-fold based on the Protein Error term alone. Using eq 4, the Protein-Based Protein TriScore becomes dimensionless with a maximal possible value of 10,000, which would take place if all of the intensity was accounted for by the protein in question, if all the expected peptides were detected, and if all of these masses matched perfectly (see Table 3 column 8 "Protein-Based Protein TriScore").

When a sufficient number of peptides are required to be matched and the mass spectrum is complex (that is, it consists of more than one or two significant peaks), the absolute value of the Protein-Based Protein TriScore can be used to assess the validity of a match. Thus, an excellent score is >1000, an almost always credible score is >50, and some scores as low as 5 appear to be meaningful, provided that other proteins have also been identified. If no protein has a score higher than 50, then very likely a major component in the sample is not in the database, or there is such a large number of components that none of them can be identified with confidence. Surprisingly, this last circumstance is difficult to achieve when dealing with organisms with relatively small defined genomes like E. coli. Thus, typically one or several components can be identified with reasonable confidence from digests of chromatographic fractions of proteins, or even from digests of whole organisms. Nothing is identified with confidence, of course, if peptides are separated and individual fractions are analyzed. One must also be careful with the number of peptides that are required for matching (parameter 5 Table 2). If parameter 5 is set to too low a number (e.g., 3 or less), then the value of Protein-Based Protein TriScore may be misleading by itself.

Intensity-Weighted Parts Per Million Error (PPW)

Because as the Maximum Ppm Error (Table 2 parameter 7) increases, an increasing number of peptides will match randomly, the Protein-Based Protein TriScore described above is strongly dependent on the Maximum Ppm Error. Usually, the higher the Maximum Ppm Error, the higher the Average PpmError term, and the lower the Protein-Based Protein TriScore. To counteract this problem, one can calculate instead an intensity-weighted Ppm Error term:

$$\text{PPW} = \Sigma\ (\text{Intensity} \times \text{Ppm error})/\Sigma\ \text{Intensity} \quad (5)$$

Now spurious matches significantly increase PPW (thereby decreasing Protein-Based Protein TriScore) only if they happen to also be intense peaks, which is much less likely to be the case. In addition, the peaks with lower intensity that are correctly assigned often have a larger mass error than the more intense peaks. For these reasons, the PPW for correctly identified proteins is usually significantly less than the Protein Ppm (see Table 3 column 8 "PPW"). Alternatively, PPW could be redefined so that it is calculated based on the average Ppm Error and average intensity for those peptides that fulfill the minimum peptide matching requirements (Table 2 parameter 5), but is intensity-weighted for all additional matches, so that its value is not dominated by single peptides, yet is not compromised by borderline signals. In cases where two different peptides from the same protein can account for the same mass, some decision must be made to determine which peptide should be used to calculate PPW. It could be either the peptide with the lowest Ppm Error, or the peptide with the highest Peptide TriScore.

The Dominant Peptide Limitation

The major drawback to giving peptides quantitative values for matching to the database is the danger that a single peptide or a small number of peptides may so dominate the score that arbitrary matches will be found. Thus, if one submits a mass spectrum with a single peak at *m/z* 1000.5, and 20 small peaks with 100-fold less intensity, the ChemApplex program will automatically identify the smallest protein in the database that has a peptide with a relatively high ChemScore that has an *m/z* ratio as close as possible to 1000.5, regardless of the Peptide ChemScores of the 20 small peaks, because their overall contribution would be so small, so long as the protein fulfills the requirements for minimal number of peptides to be matched. This protein may well be the right answer if the intense peak does in fact correspond to a tryptic peptide derived from a protein in the database, but very likely a wrong answer because of the possibility that the peak in question is due to a contaminant or calibrant. There are several ways to overcome this limitation, the simplest of which is to sort proteins using Protein TriScore where all matching masses are

used to fulfill the minimal peptide matching requirement, but the (usually one or two) peptides with the highest Peptide TriScores are set to the same value as the quantitative contribution of the second or third highest peptide (TriScoreMinus). In this fashion, every peptide contributes to the identification, but no one peptide dominates. The Protein TriScore including all peptides is more suitable for archiving purposes, or for the purpose of comparing spectra.

Combined Protein Score

One way to combine the parameters into a simple score is:

$$\text{Combined Protein Score} = (\Sigma\ \text{PepT} - \text{PepT}_1 + \text{PepT}_2) \times \%\ \text{ChemScore} \quad \text{Matched/PPW} \qquad (6)$$

where PepT = Peptide TriScore.

This formula uses ChemScore and the mass error term in two places, both at the level of individual peptides and at the level of the whole protein (see Table 3 column 9 "Combined Protein Score"). In most cases this results in appropriate protein sorting. The exact formula used has little impact on the sort order when the identifications are robust.

Multiple Protein Matching Considerations

As a further level of sophistication, in database matching it is reasonable to consider removing (or marking) previously matched masses from the MALDI mass list used to calculate the Protein TriScore of secondary components. We have tried to be careful to ensure that the highest scoring component is correctly identified prior to any subtraction step. This highest scoring component could be trypsin, or human keratins, depending on the Peptide TriScores for the matched peptides, and therefore masses that match to human keratins would be subtracted only if there was a substantial amount of evidence for the presence of human keratin. Because the primary criteria for sorting is not the sheer number of matched peptides, however, it is possible to mark masses that have previously been matched by diminishing either that mass's intensity or to diminish the ChemScore of any peptide that it is found to match that peptide later by some factor (say 100). In this fashion, the peptide will still be listed as matched to the protein, but the protein will get minimal credit for the match (see below). This is carried out automatically by the ChemApplex program, which recalculates Protein TriScore and resorts the proteins using the Combined Protein Score by adjusting the intensities and ChemScores of the matched peptides starting from the highest scoring protein on the list and working down. The number of times the automatic resorting process takes place is controlled by parameter 23 in Table 2.

The Optimal Protein Fragment MW

A second level of sophistication allows the user to postulate a maximal MW (FragMW) of protein from which the peptides in question must derive for the match to be considered similar to that used in some versions of Mascot software. For a protein of higher MW to be considered, the peptides that match must lie in a fragment of no larger than the FragMW. In this fashion, even the largest proteins can be tested without the % Intensity term dominating. To make it fair for the fragments, it is possible to calculate a Fragment ChemScore, consisting of the sum of ChemScores of the peptides within the fragment, and to use this Fragment ChemScore in calculating the % ChemScore term.

Pseudoproteins and the Supplementary Protein Database

In PMF, many signals are detected from frequently encountered contaminants. These include the trypsin used to perform the digestion, keratins from human skin, calibrant peptides that cross-contaminate at the level of the MALDI plate, masses derived from dyes like Coomassie brilliant blue, and masses derived from polymerized forms of the matrix itself. A key feature of these contaminants is that in most cases, there are several distinguishing masses. It is relatively rare to encounter only one contaminant mass, although often only one mass is prominent. There is an easy solution for dealing with this problem using the ChemApplex software system: one can define a pseudoprotein consisting of any combination of masses that are commonly encountered together, and assign ChemScores to each mass in approximate proportion to the relative intensity of these masses. Experience indicates that this is a powerful way to ensure that the program will identify trypsin correctly even if only two masses from trypsin autolysis are present on the mass list, so long as those two masses correspond to the most frequently detected masses from trypsin. In practice, our pseudoprotein trypsin contains many masses with lower ChemScores that are MALDI artifacts or peak processing artifacts. These artifacts include sodium adducts of the most prominent trypsin masses, masses that are one and two masses higher than the prominent trypsin masses that arise as residuals from incomplete de-isotoping, and doubly-charged forms of prominent masses. Some masses are nearly always found in certain trypsin batches, and there is no reason why these masses cannot also be added to the trypsin pseudoprotein. Moreover, if more than one constellation of trypsin-related peaks is observed, for example, in correlation with different batches of trypsin, more than one trypsin pseudoprotein can be defined with overlapping masses. Typically,

in this case, the program would identify both trypsin pseudoproteins, but the score from the best matched pseudoprotein would be the highest, and the score of the alternative trypsin pseudoprotein would be drastically reduced at the stage of the Multiple Protein Matching step. The program would then report how much of the intensity could be accounted for by the highest ranked pseudoprotein. In the case of human keratins, we have chosen to use the standard ChemScores for the keratin-derived peptides, and find that the ChemApplex program does an excellent job of identifying them when they are present. We have not found it necessary to combine these proteins, even though cytokeratins k1, k2, k9, and k10 are commonly observed together, albeit in varying ratios.

Observed Versus Calculated ChemScores: Learning to Identify Proteins

As a corollary to the idea of pseudoproteins, it is possible to have the program automatically add to the supplementary protein list any protein whose identification exceeds an arbitrary threshold, for example, by adding the % intensity observed for each peptide that was found to the originally calculated ChemScore for that peptide, and then renormalizing the ChemScores of each peptide to the original Protein ChemScore. In this fashion, the ChemScores of a commonly encountered protein would come to be proportional to the observed intensities rather than the ChemScores calculated from Table 1. In effect, the program would automatically learn to identify the protein better. We are confident that this would have the effect of increasing the sensitivity of identification for that protein in any succeeding analysis so long as the experimental design was the same (same alkylating agent, same degree of methionine oxidation, etc.). We have not tested this corollary extensively.

Original Scores Versus Resorted Scores

For evaluating the overall validity of database matches, we have found it useful to retain the original values of any parameters that are updated during the Multiple Protein Matching step. Thus, the number of peptides that match has two values: the original number of matching peptides (column 6 in Table 3 "All peptide matches"), and the number of peptides that match not including those accounted for by higher ranking proteins (column 5 in Table 3 "Unique peptide matches"). Because some peptides are matched randomly to proteins, we have further filtered the peptides tabulated in column 5 to include only those peptides whose ChemScores are higher than 5, and that match within 25 ppm. The output of the program also has two columns for ChemScore, Protein-Based Protein TriScore, Combined Protein Score, % Intensity Matched, and PPW, but only the recalculated values for these parameters are listed in Table 3. Because the data are originally housed in MS-Excel, in which the whole table is easily sorted by any combination of columns, one can easily determine what proteins would have been identified, had certain masses not been accounted for.

Homologous Proteins

Frequently in PMF a protein is identified that has many homologues in the database that is searched. In such cases, even MS-MS analysis cannot identify the correct protein if the peptide that is fragmented is shared between the homologues. In such a case, the ChemApplex program will initially identify all of the homologues that pass the minimal criteria for matching, but the highest ranking homologue will be sorted to the top, and the other homologues will be sorted down. In many cases this resorting process has drastic effects on which proteins are on the top of the list. Because the original scores are retained by the program, one can easily flip back and forth between unsorted and resorted proteins. This sorting feature can also be used to make the program distinguish between different forms of the protein, for example, between proteins whose signal peptides have been cleaved off versus the preprotein, or between substitution variants, or even between chemically modified forms of the protein. For this to take place, one must manually adjust the masses and ChemScores of the protein in a predigested form of the database to reflect the chemical modifications. Alternatively, the program that creates the predigested database could be designed to use intelligently the information contained in the annotation fields in the available protein databases, as is currently performed by some of the Expasy tools [17].

Methods

All reagents were from Sigma-Aldrich (St, Louis, MO), except acetone and acetonitrile, which were from EM Science (Darmstadt, Germany), and SDS, which was from Bio-Rad (Hercules, CA). Lysozyme, ovalbumin, and bovine serum albumin (BSA) were reduced and alkylated in 0.1% SDS using vinylpyridine, acetone precipitated, and then digested with bovine trypsin (Sigma T-1426). The proteins were mixed together in different amounts as described in Table 3, where the amounts listed correspond to the amount of protein loaded onto the MALDI plate in a volume of 0.5 microliter. The matrix was 0.5 microliter of recrystallized α-cyano cinnamic acid dissolved in 55% acetonitrile containing 0.3% TFA.

A Voyager DE-STR Workstation (Applied Biosystems, Framingham, MA, USA) was used to collect 200 shots and the spectra were analyzed using the ChemApplex program. The ChemScore parameters used to generate the data in Tables 3, 4, 5, 6, and 7 are listed in Table 1. The remaining ChemApplex software

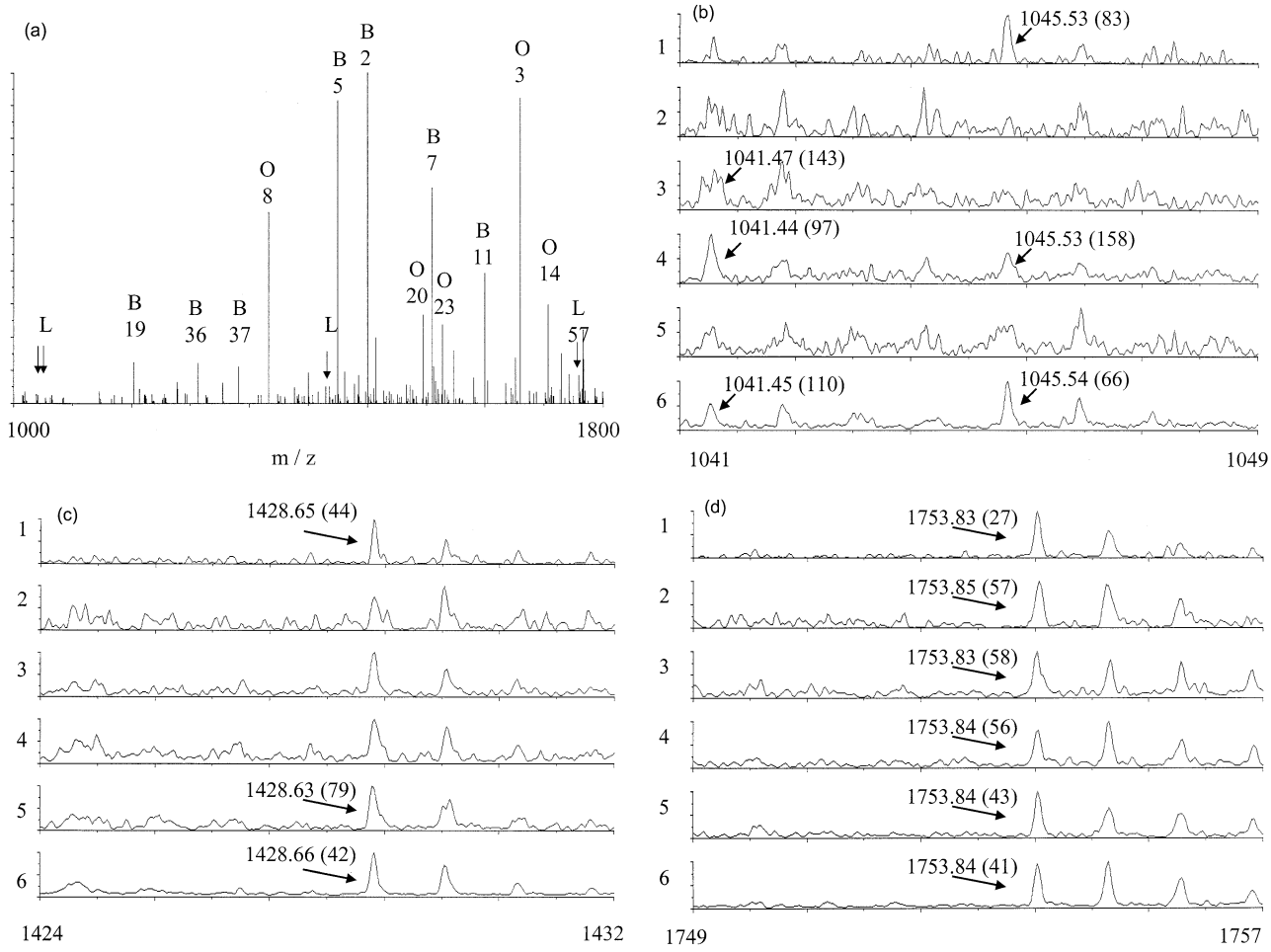

Figure 1. (**a**) Mass Spectrum of mixture containing 2 pmole of ovalbumin, 2 pmole of BSA, and 0.2 pmole of lysozyme, corresponding to Tables 5 and 7, SpecID 2. The spectrum has been de-isotoped and each isotope cluster is plotted at the position of the mono-isotopic mass. Only the region between a *m/z* of 1000 and 1800 is shown. Each substantial peak that matches to BSA or ovalbumin is labeled with a B or an O, respectively. The number beneath indicates the intensity rank of the mass in question. The three regions corresponding to expected lysozyme peptides are marked with an L. From this spectrum, the only peak that was automatically detected and matches lysozyme has an *m/z* value of 1753.83, and was the 57th most intense. (**b**), (**c**), and (**d**) Each panel shows an inset of an 8 u region for each of six spectra, referred to as SpecID 1–6 in Tables 5 and 7. In each case, the mass region displayed corresponds to an expected lysozyme peptide (see Table 6). In each case, the raw spectra are displayed; no smoothing or de-isotoping was performed. Every peak that was automatically detected is labeled with the measured *m/z* value, with the intensity rank in parentheses.

parameters are listed in Table 2 and briefly described in column 4 labeled "Description."

Results: Summary of Data from Mixtures of Three Standard Proteins

To demonstrate the ability of the ChemApplex program to identify minor components in protein mixtures, three proteins, lysozyme, ovalbumin, and BSA, were digested separately, and then mixed together in different proportions, as listed in the headings of Table 3. For each sample, 162 to 200 masses were submitted (see column 2), and searched against a database containing all of the proteins in SwissProt release 39.2 from *E. coli*, chicken and cow (7261 proteins). A region of one representative spectrum out of the 47 analyzed, extending from *m/z* 1000 to 1800 is shown in Figure 1a. In all cases, the most prominent peaks matched to expected peptides from BSA, ovalbumin, lysozyme or trypsin as expected. Table 3 lists the five highest ranking proteins identified by the ChemApplex program from these samples. In some cases (Experiments 3–5 and 8), multiple spectra were collected and analyzed separately. The highest number listed in the first column for each experiment lists the number of spectra that were analyzed in each experiment, and the numbers in columns 5–13 represent an average for the parameters listed. In some of these cases the proteins at the lower concentration were identified from some of the spectra only, as listed in column 1. For example, for the third experiment, which

contained 0.5 pmole lysozyme, 5 pmole ovalbumin, and 5 pmole BSA, BSA and ovalbumin were identified from all 11 spectra, but trypsin was detected nine times, and lysozyme was detected 8 times (see Tables 4, 5, and 6 for details). The sixth column ("Unique peptide matches") indicates that on average when lysozyme was detected, 3.1 peptides were detected. On average, 2.9 of these peptides could not also be assigned to higher proteins on the list, and also had ChemScores of higher than 5.0, and matched within 25 ppm. The seventh column indicates that the % ChemScore Matched for lysozyme was on average 47.7%, which is not much lower than BSA, which had a % ChemScore Matched of 51.4% based on an average of 15.5 peptides. Thus when lysozyme was sorted to a high-ranking position in the table, it was because an average of three matches out of 162–300 masses submitted corresponded to lysozyme peptides with relatively high ChemScores. The MW of each protein is listed in column 4 to emphasize the finding that under these search conditions the highest scoring false positive identification is usually not a high MW protein, which is in contrast to most other PMF search programs. Incorrectly identified low MW proteins can be excluded by increasing the number of peptides required for matching (Table 2 parameter 5).

Column 8 lists the Protein-Based Protein TriScore, which is on the order of 1000 for the top two components for all eight experiments, indicating that the top two proteins match with high confidence and account for a substantial percentage of the total intensity. Column 9 lists the Combined Protein Score, which was used to sort the proteins. In all 8 experiments, the average Combined Protein Scores for the four proteins known to be present (average calculated only from those experiments in which the protein was detected at all) was always higher than the highest Combined Protein Score for any other protein (though just barely in Experiment 8). However, in 14 of the 45 times that one of the three standard proteins was present at 10-fold lower concentration than the other two proteins, that protein was not identified by the ChemApplex program. In comparison, trypsin was detected all but 6 out of 47 times. Identification may have been more successful for trypsin because it was present in the database as a pseudoprotein, and therefore the most easily detected peptides from trypsin had higher ChemScores, making it easier to detect. The discussion below addresses why the minor protein was not always identified.

The % Intensity Matched for each protein is listed in column 10, while PPW is listed in column 11. Note that the minor standard protein never had a higher % Intensity Matched than 1.9%, even though by design the minor component comprised ~5% of the mixture at a molar level (1 part in 21). This is unexpected, but because the minor component contributes such a small percentage of the intensity, it is a worthy challenge for the ChemApplex program. Note also that the highest PPW observed for any of the standard proteins was 7.0, even though peptides were considered to a maximum ppm difference of 25 ppm. Hence, most masses that match are not random. The low total % Intensity Matched value observed for all matched proteins is at least partially explained by a technical problem that took place with these digests: There must have been a problem at the alkylation stage, because the vinyl pyridine-modified cysteine containing peptides were poorly recovered from all three of the standard proteins. The ChemScore for these peptides was adjusted downwards on that account, so the database identification did not suffer very much. It remains possible that many of the masses that did not match any protein correspond to some unrecognizable form of the cysteine containing peptides, thus explaining why the total % Intensity Matched for the known components varies between 49% to 77%. In other experiments, we have often obtained values of higher than 90% Intensity Matched even from 200 masses submitted. It is of course possible to easily obtain 100% Intensity Matched by limiting the number of masses submitted to 10–20 masses. The highest Protein-Based Protein TriScore that has been obtained for any protein using the parameters listed in Table 1 is 6840, which occurred from an E. coli 2-D gel sample containing EF-TS in which 21 out of 121 masses were matched, accounting for 89% of the intensity, with a ChemScore Matched of 80.7% and a PPW of 2.1 (data not shown). The maximum possible Protein-Based Protein TriScore is 10,000, which would occur if every possible peptide that was predicted was identified, accounting for 100% of the intensity observed, with a PPW of less than 2.0. It is also possible to get higher scores by using different ChemScore values, for example, by setting the ChemScore of all peptides other than arginine-containing to zero. Although this would cause the highest scoring protein to get a higher score, minor protein components will be detected with much greater difficulty because lysine-containing peptides from the top-ranked peptide would not be accounted for. If a consistent set of input parameters is used with the ChemApplex program, then the Protein Based Protein TriScore becomes an excellent measure of data quality because it is normalized. When more than one protein component is present, then the highest quality mass spectrum has the highest Protein Based Protein TriScore for each component. This should be useful in optimizing spot deposition on MALDI plates, as well as at the level of spectrum acquisition.

The remaining two columns in Table 3 list the % of the matched intensity that was attributable to arginine-containing complete digestion products ("Arg % Intensity"), or non arginine-containing lysine-containing peptides ("Lys % Intensity"). Note that the highest percentage intensity of lysine-containing terminal digestion products from the standard proteins was 7.4%, whereas the lowest % of intensity attributable to arginine-containing peptides was 47.1%. Most of the remaining Intensity Matched was due to arginine-containing peptides that had missed cleavages. The values of these

parameters are not reported for trypsin because it is a special case: the autodigestion peptides that are derived from trypsin are cleaved from a native protein, with intact disulfide bonds. As it turns out, there are no arginine-containing terminal peptides that would be predicted from the sequence of bovine trypsin with an *m/z* ratio higher than SAASLNSR at MW 805.4167. This peptide is sometimes observed at relatively low intensity, perhaps because it is inefficiently cleaved from intact trypsin compared to the more N-terminal lysine-containing fragments of trypsin, which are usually more intense.

It is fair to ask why the minor component was not detected correctly in every case. To examine this question, the data for lysozyme from Table 3 are considered in detail, with special attention paid to Experiment 4. Of the 33 peptides from lysozyme that are in the appropriate *m/z* range for MALDI analysis, only 14 were ever observed in any of the eight experiments in Table 3. Of these, 9 were observed more than twice (Table 4). Only six of the masses corresponding to these peptides were ever observed in Experiment 4, whether or not lysozyme was identified. Two of these six peptides ranked in intensity >150th and were observed only once and therefore are in all likelihood accidental hits, and were not considered further. The data for the remaining four peptides are summarized in Table 6. The most intense of these peptides was on average about the 48th most intense peak in the spectrum, whereas the weakest peak ranked on average in 102nd place. Eight u regions of the first six spectra from Experiment 4 are shown in Figures 1b, c, and d. The region corresponding to the first two lysozyme peptides are shown in Figure 1b, whereas the third and fourth peptides are shown in Figures 1c and d, respectively. Whenever a peak that matched a lysozyme peptide was detected, it is labeled in the figure, regardless of whether there was sufficient information in the rest of the spectrum to warrant identification of lysozyme.

The details describing the matches of those four masses from lysozyme in nine different spectra obtained from the same MALDI spot are listed in Tables 5 and 7. For the nine spectra, in six cases, lysozyme was detected as a protein to be further considered (between 20 and 35 proteins, see Table 7 column 8 "No. proteins"). In two cases (Nos. 3 and 7), the reason that lysozyme was not chosen for further consideration was that it did not pass the minimum % ChemScore Matched criterion of 20% (data not shown); in the remaining case (No. 2), only one of the six expected masses that derive from lysozyme was detected (Table 5). Figures 1b, c, and d show that although there are peaks corresponding to lysozyme in spectra 2 and 3, they were below the threshold for automatic peak detection. In these three cases, the program should not have been able to identify lysozyme due to lack of evidence.

To determine why the ChemApplex program had trouble identifying lysozyme in three of the six cases in which there was evidence for lysozyme, the proteins that passed the minimal detection criteria were sorted successively by the following parameters: No. of peptides matched (Table 7 column 3 "All peptide matches"), % ChemScore Matched (column 4), Protein Based Protein TriScore (column 5), % Intensity Matched (column 6), and PPW (column 7). In each case, the proteins which fulfilled the matching requirements listed in Table 2 (lines 1–18) were sorted only by the parameter indicated in Table 7; no masses were subtracted for higher ranking proteins. For comparison, the rank of the protein that was obtained upon sorting by Protein Based Protein TriScore when masses were subtracted for higher ranking proteins is listed in column 2 "Rank". Strikingly, BSA and ovalbumin always ranked either 1 or 2 when sorted by M, and % I. Thus % I, % Ch, and M, the first two of which are independent parameters, each by itself promotes the correct protein identification. In several cases, there were higher ranking proteins upon % Ch sorting; in these cases the proteins that ranked higher than BSA or ovalbumin had two peptides matched (all cases but one) or three peptides matched, and had a molecular weight of <20K. Not surprisingly, % Ch favors the identification of smaller proteins, like lysozyme. To reduce the danger of false positives, one must make M larger, at the price of eliminating proteins from consideration that are correct but produce a small number of detectable peptides. This has been frequently observed for proteins isolated from the lower MW regions of both 1 and 2-D gels (data not shown).

Because one peptide that matches fortuitously can dominate PPW, it is not surprising that some random proteins ranked higher than BSA or ovalbumin when the proteins were sorted solely by PPW. Thus PPW is not useful on its own for identification unless it is combined with a high number of required peptide matches. However, if for some reason PPW is aberrantly high, there needs to be some explanation (often imperfect calibration or intense masses that match but are spurious). In these experiments, trypsin illustrates this situation. Although trypsin was identified in third or fourth place in all cases except for in spectrum No. 1, in all cases, when the proteins were sorted using PPW, trypsin was ranked significantly lower than the other correct proteins. This was largely due to a fortuitous match of a rather intense mass at *m/z* 1725.8 (~rank 20) with very low ChemScore (0.0038) to trypsin that matched with poor mass accuracy (~18 ppm). Because it elevated PPW, trypsin would have been ranked higher had this mass not matched to it! This same mass could also be explained by an ovalbumin peptide that was itself dubious because the peptide was found only in the methionine sulfoxide form, which would not be expected in solution-digested ovalbumin and was not found for other methionine-containing ovalbumin-derived peptides. Thus, one of the factors that complicates protein identification in complex mixtures is accidental matches. Because relatively few peptides have a high

ChemScore, individual peptides that match with low ChemScore must be viewed with skepticism, unless other peptides with higher ChemScore are detected from the same protein.

In the six cases where lysozyme was detected, in two cases, (Nos. 4 and 8), lysozyme was not ranked highly by the program. In both cases, this was because of poor % I and low % Ch, rather than high PPW. In both spectra Nos. 4 and 8, lysozyme accounted for at most 0.3% of the total intensity (data not shown), which caused lysozyme to rank 31st out of 35 for spectrum No. 4 (Table 7 column 6 "% I"), and 27th out of 31 in spectrum No. 8. Thus, the program did an excellent job in identifying lysozyme whenever the lysozyme peaks were reasonably intense.

Conclusions

The ChemApplex program is often able to identify correctly minor components in predefined mixtures, even when the evidence for the presence of the minor component is borderline. In the case of the simple protein mixtures studied here, the minor component was often identified if three plausible peptides were detected, even when those peptides were so weak that about one hundred masses needed to be considered for them to be detected. However, when the evidence for the presence of a protein is below a certain threshold, automated identification fails. It is likely that further improvements to the algorithm, perhaps based on a more sophisticated understanding of ChemScore, could lower the threshold somewhat. In many of these borderline cases, automated MS-MS analyses would also fail to identify the minor component because none of the signals from the minor component would be selected for fragmentation. Of course, given sufficient time and sufficient peptide separation, MS-MS analyses can identify large numbers of minor components in complex mixtures [18]. When real biological systems are investigated, the ChemApplex program correctly identifies minor components as long as a sufficiently complete database is available and a high quality mass spectrum is obtained, for example, in digests of extracts from slices of SDS gels, from digests of protein chromatographic fractions, or even in digests of whole cell lysates (data not shown). In real systems, the major confounding factor is expected to be the presence of large numbers of unaccountable peaks originating from peptides from fragments of proteins, or unexpected chemical modifications. The strategy of using chemical knowledge to predict which peptides should be detectable by mass spectrometry enables PMF to identify a larger number of proteins in complex mixtures, which is especially valuable when a limited amount of MS-MS time is available. To this end, the ChemApplex program lists which masses are indicative of each protein in the mixture, as well as which masses cannot be explained by PMF.

How does the ability of the ChemApplex program to identify proteins compare with the existing alternatives? When the same mass lists are submitted, the same parameters are used for searching, and the same databases are searched, then the same proteins are found by most PMF programs at the initial pass prior to the sorting step. However, depending on exactly which parameters are used, and which mass list is submitted, traditional PMF programs typically determine different proteins to be the primary component, and in most cases make no attempt to determine whether additional protein components are also present, or whether they represent the most plausible alternatives. The proteins that ChemApplex identifies are also dependent on the parameters used, but to a much lesser degree, because the quality of the match for each mass is considered individually. In addition, ChemApplex does distinguish between alternative proteins and proteins that may simultaneously be present. ChemApplex is especially useful in extending the gray zone in two situations. (1) When there is sufficient information in the spectrum to warrant the identification of multiple proteins. In this case, we believe ChemApplex typically identifies more proteins with greater confidence than alternative programs, and certainly accumulates more information about the matches, which is valuable for considering borderline cases. (2) When the statistics that support the identification of the highest scoring protein are not compelling, the additional information returned by ChemApplex often makes it easier to judge the validity of the match. The large amount of information returned by ChemApplex is also useful in a third area of research- in determining which spectrum among many is the best, which is important in optimizing experimental protocols. Experiments are in progress to document the use of ChemApplex in identifying proteins in complex mixtures from a variety of biological materials.

References

1. Mann, M.; Hojrup, P.; Roepstorff, P. Use of Mass Spectrometric Molecular Weight Information to Identify Proteins in Sequence Databases. *Biol. Mass Spectrom.* **1993,** *22,* 338–345.
2. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. Identifying Proteins from Two-Dimensional Gels by Molecular Mass Searching of Peptide Fragments in Protein Sequence Databases. *Proc. Natl. Acad. Sci. U.S.A.* **1993,** *90,* 5011–5015.
3. Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. Rapid Identification of Proteins by Peptide-Mass Finger Printing. *Current Biology* **1993,** *3,* 327–332.
4. James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. Protein Identification by Mass Profile Fingerprinting. *Biochem. Biophys. Res. Commun.* **1993,** *195,* 58–64.
5. Yates, J. R.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. Peptide Mass Maps: A Highly Informative Approach to Protein Identification. *Anal. Biochem.* **1993,** *214,* 397–408.
6. Cottrell, J. S. Protein Identification by Peptide Mass Fingerprinting. *Peptide Research* **1994,** *7,* 115–124.
7. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999,** *20,* 3551–3567.

8. Clauser, K. R.; Baker, P.; Burlingame, A. L. Role of Accurate Mass Measurement (+/− 10 ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching. *Anal. Chem.* **1999,** *71,* 2871–2882.
9. Gras, R.; Mueller, M.; Gasteiger, E.; Gay, S.; Binz, P.–A.; Bienvenut, W.; Hoogland, C.; Sanchez, J.-C.; Bairoch, A.; Hochstrasser, D. F.; Appel, R. D. Improving Protein Identification from Peptide Mass Fingerprinting Through a Parameterized Multi-Level Scoring Algorithm and an Optimized Peak Detection. *Electrophoresis* **1999,** *20,* 3535–3550.
10. Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Rudel, T.; Jungblut, P. R. Analysis of Missed Cleavage Sites, Tryptophan Oxidation and N-Terminal Pyroglutamylation After In-Gel Tryptic Digestion. *Rapid Commun. Mass Spectrom.* **2000,** *14,* 496–502.
11. Parker, K. C.; Garrels, J. I.; Hines, W.; Butler, E. M.; McKee, A. H.; Patterson, D.; Martin, S. Identification of Yeast Proteins from Two-Dimensional Gels: Working Out Spot Cross-Contamination. *Electrophoresis* **1998,** *19,* 1920–1932.
12. Krause, E.; Wenschuh, H.; Jungblut, P. R. The Dominance of Arginine-Containing Peptides in MALDI-Derived Tryptic Mass Fingerprints of Proteins. *Anal. Chem.* **1999,** *71,* 4160–4165.
13. Moritz, R. L.; Eddes, J. S.; Reid, G. E.; Simpson, R. J. S-Pyridylethylation of Intact Polyacrylamide Gels and in situ Digestion of Electrophoretically Separated Proteins: A Rapid Mass Spectrometric Method for Identifying Cysteine-Containing Peptides. *Electrophoresis* **1996,** *17,* 907–917.
14. Keough, T.; Youngquist, R. S.; Lacey, M. P. A Method for High-Sensitivity Peptide Sequencing Using Postsource Decay Matrix-Assisted Laser Desorption Ionization Mass Spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **1999,** *96,* 7131–7136.
15. Beardsley, R. L.; Karty, J. A.; Reilly, J. P. Enhancing the Intensities of Lysine-Terminated Tryptic Peptide Ions in Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **2000,** *14,* 2147–2153.
16. Brancia, F. L.; Oliver, S. G.; Gaskell, S. J. Improved Matrix-Assisted Laser Desorption/Ionization Mass Spectrometric Analysis of Tryptic Hydrolysates of Proteins Following Guanidination of Lysine-Containing Peptides. *Rapid Commun. Mass Spectrom.* **2000,** *14,* 2070–2073.
17. Wilkins, M. R.; Gasteiger, E.; Bairoch, A.; Sanchez, J.-C.; Williams, K. L.; Appel, R. D.; Hochstrasser, D. F. *In 2-D Proteome Analysis Protocols;* Link, A. J., Ed.; Humana Press: New Jersey, 1998, pp 531–552.
18. Washburn, M. P.; Wolters, D.; Yates, J. R., III. Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. *Nat. Biotechnol.* **2001,** *19,* pp 242–247.