Article No. jmbi.1999.2794 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 289, 645–657

JMB

High-throughput Mass Spectrometric Discovery of Protein Post-translational Modifications

**Marc R. Wilkins[1,2,3]*, Elisabeth Gasteiger[3,4], Andrew A. Gooley[1]
Ben R. Herbert[1], Mark P. Molloy[1], Pierre-Alain Binz[2,3], Keli Ou[1]
Jean-Charles Sanchez[2], Amos Bairoch[3,4], Keith L. Williams[1]
and Denis F. Hochstrasser[2,3]**

[1]*Macquarie University Centre for Analytical Biotechnology and Australian Proteome Analysis Facility, Macquarie University, Sydney, NSW 2109 Australia*

[2]*Central Clinical Chemistry Laboratory, Geneva University Hospital, 24 Rue Micheli-du-Crest, 1211, Geneve 14 Switzerland*

[3]*Medical Biochemistry Department, University of Geneva, 1 Rue Michel Servet 1211, Geneve 4, Switzerland*

[4]*Swiss Institute of Bioinformatics, 24 Rue Micheli-du-Crest, 1211, Geneve 14 Switzerland*

**Corresponding author*

The availability of genome sequences, affordable mass spectrometers and high-resolution two-dimensional gels has made possible the identification of hundreds of proteins from many organisms by peptide mass fingerprinting. However, little attention has been paid to how information generated by these means can be utilised for detailed protein characterisation. Here we present an approach for the systematic characterisation of proteins using mass spectrometry and a software tool FindMod. This tool, available on the internet at http://www.expasy.ch/sprot/findmod.html, examines peptide mass fingerprinting data for mass differences between empirical and theoretical peptides. Where mass differences correspond to a post-translational modification, intelligent rules are applied to predict the amino acids in the peptide, if any, that might carry the modification. FindMod rules were constructed by examining 5153 incidences of post-translational modifications documented in the SWISS-PROT database, and for the 22 post-translational modifications currently considered (acetylation, amidation, biotinylation, C-mannosylation, deamidation, flavinylation, farnesylation, formylation, geranyl-geranylation, gamma-carboxyglutamic acids, hydroxylation, lipoylation, methylation, myristoylation, *N*-acyl diglyceride (tripalmitate), O-GlcNAc, palmitoylation, phosphorylation, pyridoxal phosphate, phospho-pantetheine, pyrrolidone carboxylic acid, sulphation) a total of 29 different rules were made. These consider which amino acids can carry a modification, whether the modification occurs on N-terminal, C-terminal or internal amino acids, and the type of organisms on which the modification can be found. We illustrate the utility of the approach with proteins from 2-D gels of *Escherichia coli* and sheep wool, where post-translational modifications predicted by FindMod were confirmed by MALDI post-source decay peptide fragmentation. As the approach is amenable to automation, it presents a potentially large-scale means of protein characterisation in proteome projects.

Keywords: post-translational modification; proteome; two-dimensional gel electrophoresis; peptide mass fingerprinting; mass spectrometry

Introduction

Large-scale projects are making a huge impact on the study of biological systems. Genome sequencing programs have defined the informational content of a range of organisms, from the small-genome *Haemophilus influenzae* to that of the eukaryotic budding yeast *Saccharomyces cerevisiae* and *Caenorhabditis elegans* (Fleischmann *et al*., 1995;

Abbreviations used: 2D, two-dimensional; MALDI-TOF, matrix-assisted laser desorption time of flight; FT, feature; PSD, post-source decay.

E-mail address of the corresponding author: mwilkins@rna.bio.mq.edu.au

Goffeau *et al.*, 1996; The *C. elegans* Sequencing Consortium, 1998). Full genomic sequences for other, more complex model systems such as *Drosophila melanogaster* and *Arabidopsis thaliana* will be available in the near future. This definition of genes, and related large-scale analyses of mRNAs with techniques such as SAGE (Velculescu *et al.*, 1997), is providing insight as to what proteins we expect to find in an organism, to what degree they may be present, and at what time they might be expressed.

Proteome projects aim to identify and characterise all proteins expressed by an organism or tissue (Wilkins *et al.*, 1995), and thus complement genome projects through the study of the functional, rather than informational, molecules of biological processes. Recent advances in protein analysis technology, in particular the availability of exquisitely accurate and sensitive mass spectrometers (Roepstorff, 1997; Mann & Talbo, 1996), are making possible the very large scale analysis of proteins. Proteins are separated in parallel by two-dimensional (2-D) electrophoresis techniques, which array up to thousands of proteins in quantities sufficient for analysis (Sanchez *et al.*, 1997). Robots can then excise protein spots from gels, place them into 96-well plates, subject them to endoproteinase digestion and prepare resulting peptides for automated analysis on mass spectrometers (Traini *et al.*, 1998). Throughput on matrix-assisted laser desorption time of flight (MALDI-TOF) mass spectrometers is now over 100 samples per day.

Progress in proteomic methods and technology so far has largely focused on high-throughput protein identification, especially by the technique of peptide mass fingerprinting. This involves the digestion of a protein with an endoproteinsase of known cleavage specificity, the measurement of the masses of resulting peptides by mass spectrometry, and protein identification by matching the observed peptide masses against databases of proteins whose peptide masses have been generated theoretically (Henzel *et al.*, 1993; James *et al.*, 1993; Mann *et al.*, 1993; Yates *et al.*, 1993). The efficiency of this technique is such that it is becoming commonplace for hundreds of protein identifications to be reported in a single journal article (e.g. Shevchenko *et al.*, 1996; Langen *et al.*, 1997; Traini *et al.*, 1998). However, comparatively little attention has been paid to how researchers can harness peptide mass fingerprinting results, genomic sequences, and protein annotations in databases to systematically analyse proteins for post-translational processing and modifications. These modification events, whilst difficult or impossible to predict from protein sequences alone, are crucial in the structure and function of many proteins and in the control of biochemical pathways.

Here we present an approach for the systematic characterisation of proteins using mass spectrometry and a software tool, FindMod. This tool examines peptide mass fingerprinting data and applies intelligent rules to predict amino acids in peptides that might carry protein post-translational modifications. FindMod can also check for single amino acid substitutions. Finally, FindMod predictions can be tested by mass spectrometry peptide fragmentation techniques.

Results

The FindMod rationale

A series of software tools, such as MOWSE, MS-Fit and PeptIdent are available on the internet to do database matching for peptide mass fingerprinting (Pappin *et al.*, 1993; Mann, 1994; Clauser *et al.*, 1995; M.R.W. *et al.*, unpublished). As part of the matching procedure, these programs can search for a few common post-translational (e.g. phosphorylation) or atifactual modifications (e.g. oxidation of methionine), whereby modifications are ''found'' by nature of the mass change they impart to any particular peptide. Such an approach may reveal modified peptides, but the types of modifications that are considered are generally few. This is because it becomes very computationally intensive to look for modification-derived mass differences between peptides from a query protein and the millions of peptides from the tens to hundreds of thousands of proteins in current databases. In creating FindMod, our rationale has been to dissociate the peptide mass fingerprinting database matching procedure from that of protein characterisation. Thus, the user first identifies the query protein using available peptide mass fingerprinting search engines, assigning a group of peptide masses to a particular protein. FindMod is then used to search for protein post-translational modifications by comparing experimental peptide masses that did not match with the protein against those calculated from the assigned protein sequence, seeking mass differences that may be due to post-translational modifications. When mass differences corresponding to a post-translational modification are found between an experimental peptide and a theoretical peptide for that protein, FindMod applies a set of intelligent rules to make predictions as to which amino acids within the peptide, if any, might carry the modification. A flow diagram for this procedure is shown in Figure 1.

FindMod rules for the prediction of modifications on amino acids

To define FindMod rules for the prediction of modifications on amino acids in a peptide, the feature (FT) lines of all proteins in the SWISS-PROT database (Bairoch & Apweiler, 1998) were checked for post-translational modifications. A total of 5153 incidences of modifications were examined, noting the type of modification, the amino acid on which it was found, the position in the protein where the modification occurred (internal or N or C-terminal), and the phylogenetic classification of the

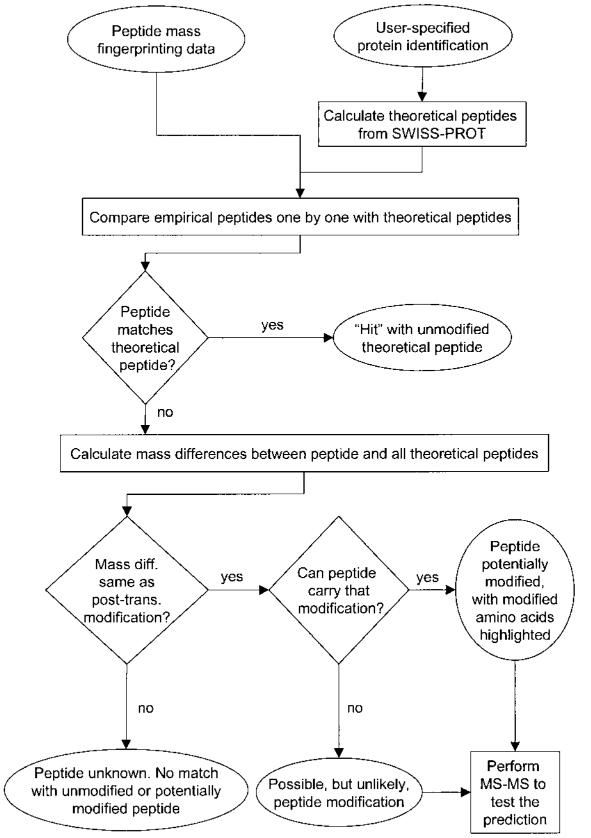

Peptide mass fingerprinting data
User-specified protein identification
Calculate theoretical peptides from SWISS-PROT
Compare empirical peptides one by one with theoretical peptides
Peptide matches theoretical peptide?
yes
"Hit" with unmodified theoretical peptide
no
Calculate mass differences between peptide and all theoretical peptides
Mass diff. same as post-trans. modification?
yes
Can peptide carry that modification?
yes
Peptide potentially modified, with modified amino acids highlighted
no
no
Peptide unknown. No match with unmodified or potentially modified peptide
Possible, but unlikely, peptide modification
Perform MS-MS to test the prediction

Figure 1. Flow chart for the discovery of protein post-translational modifications with mass spectrometry and the FindMod tool.

organism for which the modification was present. All modification events considered in SWISS-PROT had been derived from the literature. The PROSITE database of protein families and domains (Bairoch *et al.*, 1997) was also examined for entries concerning post-translational modifications. Note that *N*-glycosylation, whilst having a well-defined sequence motif, was not considered for FindMod as the mass difference imparted to a peptide is difficult to predict, encompassing a very large family of masses. This modification, however, is being considered elsewhere with a tool which predicts the structures of *N*-linked glycoforms (Packer *et al.*, unpublished results).

We formulated FindMod rules for the prediction of 22 of the more common post-translational modifications to amino acids (Table 1). All of these modifications impart a discrete mass change to a peptide. Rules for some modifications are simple, for example, it is possible for formylation to occur on any N-terminal amino acid on any protein from any species. Other modification rules, such as for methylation, need to be different between prokaryotes and eukaryotes. PROSITE patterns were adopted in whole or part for other FindMod rules, for example, the biotinylation modification. In total, we created 29 separate rules to address the 22 considered modifications. Simple rules, describing which amino acids can be modified, were also implemented for the modifications of glycation, carbamylation, sulphoxide formation and Cys-propionamide formation. These modifications are artifacts that sometimes occur during protein preparation.

Discovery of post-translational modifications with FindMod

To best show the utility of our approach, below we present four examples of the discovery of modified peptides using FindMod. In each case, this involved peptide mass fingerprinting of a protein, identification of the protein using existing peptide mass fingerprinting tools, then use of the FindMod tool to make predictions as to which amino acids in which peptides might be modified. Finally, we used peptide fragmentation to test the FindMod predictions. All proteins were single spots from 2-D gels, subjected to peptide mass fingerprinting using MALDI-TOF MS, and peptide fragmentation was done using MALDI-TOF post-source decay (PSD; see Materials and Methods).

Case 1: Lysine dimethylation in Escherichia coli *elongation factor TU*

After identification of a protein from *E. coli* as elongation factor TU (P02990), it was noted that a peptide of mass 1631.80 did not match any unmodified peptide from this protein. This peptide was predicted by FindMod to have the sequence AFDQIDNAPEEKAR, modified in one of three ways to account for a delta mass of 28.03 (Figure 2). Modifications that conformed with FindMod rules were a dimethylation on one amino acid (e.g. K), or alternatively, for there to be a monomethylation of Lys56 (as documented in SWISS-PROT) as well as a monomethylation of one of the amino acids D, Q, N, E, K or R. FindMod also highlighted the possibility that a formylation event would produce an identical delta mass. However, as the FindMod rules define formylation to occur only at the N termini of proteins, and as the potentially formylated peptide was internal to the protein, FindMod suggested this modification to be unlikely. PSD fragmentation results confirmed the peptide to be of sequence AFDQIDNAPEEKAR, and localised the additional 28.03 mass units to the lysine residue, showing it to be dimethylated (Figure 3 and Table 2). This dimethylation accounts for the missed tryptic cleavage site at the Lys in the peptide.

Case 2: acetylation and Cys-propionamide in wool keratin 1

Following identification of a protein form sheep wool as keratin 1 (P02534), FindMod was used to investigate a peptide of mass 1557.77,

Table 1. Modification masses and rules used in FindMod for the prediction of modified amino acids

Modification	Delta mass (monoisotopic, average)	Kingdom	Amino acids	Position in protein	Rule
Acetylation	42.0106, 42.0373	All	K	Anywhere	-
			All but F, H, K, N, R, W, Y		
Amidation	−0.9840, −0.9847	Eukaryotes	All	C terminus after terminal cleavage event	G must have been C-terminal to site of cleavage
Biotin	226.0776, 226.2934	All	K	Anywhere	PROSITE pattern PDOC00167[a]
*Carbamylation	43.00581, 43.02504	All	All	N terminus	-
			K	Anywhere	-
C-Mannosylation	162.0538, 162.1424	Eukaryotes	W	Anywhere	-
*Cys-propionamide	71.0371, 71.0788	All	C	Anywhere	-
Deamidation	0.9840, 0.9847	All	N	Anywhere	Must be followed by a G
			Q	Anywhere	-
Flavin-adenine dinucleotide	783.1415, 783.542	All	C, H	Anywhere	-
Farnesylation	204.1878,204.3556	Eukaryotes, Viruses	C	Anywhere	-
Formylation	27.9949, 28.0104	All	All	N terminus	-
*Glycation (glucosylation)	162.0528, 162.1424	All	All	N terminus	-
Geranyl-geranylation	272.2504, 272.4741	Eukaryotes, Viruses	C	Anywhere	-
Gamma-carboxyglutamic acid	43.9898, 44.0098	Eukaryotes	E	Anywhere	-
Hydroxylation	15.9949, 15.9994	Eukaryotes	D, N, K, P	Anywhere	-
Lipoyl	188.033, 188.3027	All	K	Anywhere	PROSITE pattern PDC00168[c]
Methylation	14.0157, 15.0269	All	C, D, E, H, K, N, R, Q	Anywhere	-
		Prokaryotes	A	N terminus	-
		Eukaryotes	A, P	N terminus	-
		Prokaryotes	F, I, L, M, Y	N terminus after cleavage of signal peptide	PROSITE pattern PS00409[d]
Myristoylation	210.1984, 210.3598	Eukaryotes	K	Anywhere	-
		Eukaryotes, Viruses	G	N terminus	-
N-acyl diglyceride (tripalmitate)	788.7258, 789.3202	Prokaryotes, Archaebacteria, Phage	C	N terminus after cleavage of signal peptide	PROSITE pattern PDOC00013[b]
O-GlcNac	203.0794, 203.1950	Eukaryotes	S, T	Anywhere	-
Palmitoylation	238.2297, 238.4136	Eukaryotes	C, S, T	Anywhere	-
Phosphorylation	79.9663, 79.9799	Prokaryotes	C, D, H, S, T	Anywhere	-
		Eukaryotes, Viruses	H, D, S, T, Y	Anywhere	-
Pyridoxal phosphate	229.014, 229.129	All	K	Anywhere	D, P not allowed −1, E, P not allowed +1
Phosphopantetheine	339.078, 339.3234	All	S	Anywhere	PROSITE pattern PDC0012[e]
Pyrrolidone carboxylic acid	−17.0266, −17.0306	All	Q	N terminus	-
Sulphation	79.9568, 80.0642	Eukaryotes	Y	Anywhere	PROSITE pattern PDOC0003[f]
*Sulphoxide formation	15.9949, 15.9994	All	M	Anywhere	-

Modifications preceded by an asterisk (*) are generally artifactual and can also be searched for by FindMod.

[a] The regular expression [GN]-[DEQTR]-x-[LIVMFY]-x(2)-[LIVM]-x-[AIV]-M-K-[LMA]-x(3)-[LIVM]-x-[SAV] where K is the biotin attachment site.

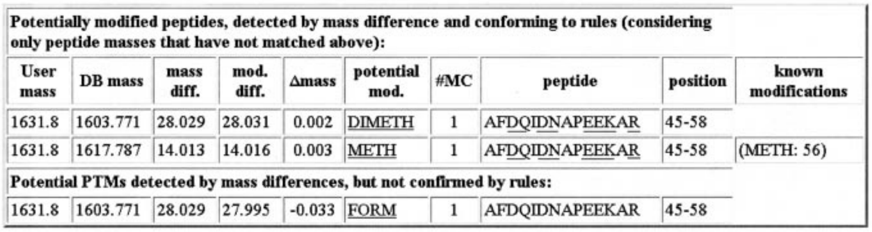
[b] The regular expression {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C where C is the lipid attachment site. Additional rules are that: (1) The cysteine must be between positions 15 and 35 of the sequence in consideration (2) There must be at least one Lys or one Arg in the first seven positions of the sequence.

[c] The regular expression [GN]-x(2)-[LIVF]-x(5)-[LIVFC]-x(2)-[LIVFA]-x(3)-K-[STAIV]-[STAVQDN]-x(2)-[LIVMFS]-x(5)-[GCN]-x-[LIVMFY] where K is the lipoyl-binding site.

[d] The regular expression [KRHEQSTAG]-G-[FYLIVM]-[ST]-[LT]-[LIVP]-E-[LIVMFWSTAG].

[e] The regular expression [DEQGSTALMKRH]-[LIVMFYSTAC]-[GNQ]-[LIVMFYAG]-[DNEKHS]-S-[LIVMST]-{PCFY}-[STAGCPQ-LIVMF]-[LIVMATN]-[DENQGTAKRHLM]-[LIVMWSTA]-[LIVGSTACR]-x(2)-[LIVMFA] where S is the pantetheine attachment site.

[f] See http://www.expasy.ch/sprot/prosite.html.



Potentially modified peptides, detected by mass difference and conforming to rules (considering only peptide masses that have not matched above):

User mass	DB mass	mass diff.	mod. diff.	Δmass	potential mod.	#MC	peptide	position	known modifications
1631.8	1603.771	28.029	28.031	0.002	DIMETH	1	AFDQIDNAPEEKAR	45-58	
1631.8	1617.787	14.013	14.016	0.003	METH	1	AFDQIDNAPEEKAR	45-58	(METH: 56)
Potential PTMs detected by mass differences, but not confirmed by rules:									
1631.8	1603.771	28.029	27.995	-0.033	FORM	1	AFDQIDNAPEEKAR	45-58	

Figure 2. FindMod output for the tryptic peptide of mass 1631.8 from EFTU_ECOLI (P02990), with mass tolerance of 0.1 Da. DB mass is the theoretical peptide mass calculated from the database, #MC is the number of masssed cleavages for the peptide, and the known modifications column shows modifications already documented in the SWISS-PROT database. In the genuine FindMod output, amino acids in the peptide that may carry a modification are shown in blue. In this Figure, these are underlined instead.

which did not match any unmodified peptide from the protein. FindMod predicted the peptide as likely to have sequence SFNFCLPNLSFR or SENARLVVQIDNAK. If of sequence SFNFCLPNSLFR, the peptide was predicted to have mass 1557.77 if it carried either an N-terminal acetylation and artifactual Cys-propionamide, or one trimethylation at one of three sites and an artifactual Cys-propionamide. If the peptide is SENARLVVQIDNAK, a deamidated Gln could also result in a mass of 1557.7 (Figure 4). These three possibilities conformed to FinMod rules. Fragmentation of the peptide by PSD showed that the peptide sequence was SFNFCLPNLSFR, where the N-terminal Ser was acetylated and the Cys was modified by acrylamide to give Cys-propionamide (Figure 5 and Table 3).

Case 3: a wool keratin peptide with acetylation, Cys-propionamide and a Phe to Tyr substitution

In peptide mass fingerprinting this wool protein had homology to sheep keratin 1 (P02534), although we suspected it to be a different gene product from keratin 1 due to its different migration on 2-D gels (data not shown). We used FindMod to establish if a modification in a peptide from sheep keratin 1 (P02534) or a single amino acid substitution could explain the peptide of mass 1573.82. FindMod predicted that the peptide might be of sequence SFNFCLPNLSFR with N-terminal acetylation and artifactual Cys-propionamide, carrying one hydroxylated amino acid (e.g. hydroxyproline). Alternatively, FindMod showed that the single amino acid substitution from Phe to Tyr in the peptide SFNFCLPNSLFR, the most likely substitution by the BLOSUM62 score (Henikoff & Henikoff, 1992), could also account for the same mass change if the peptide also was N-terminally acetylated and contained Cys-propionamide (Figure 6). PSD fragmentation of the peptide showed the latter prediction to be true, with the peptide sequence being S(acet)**Y**NFC (pam)LPNSFR (Figure 7 and Table 4). Particularly important in the interpretation of the PSD

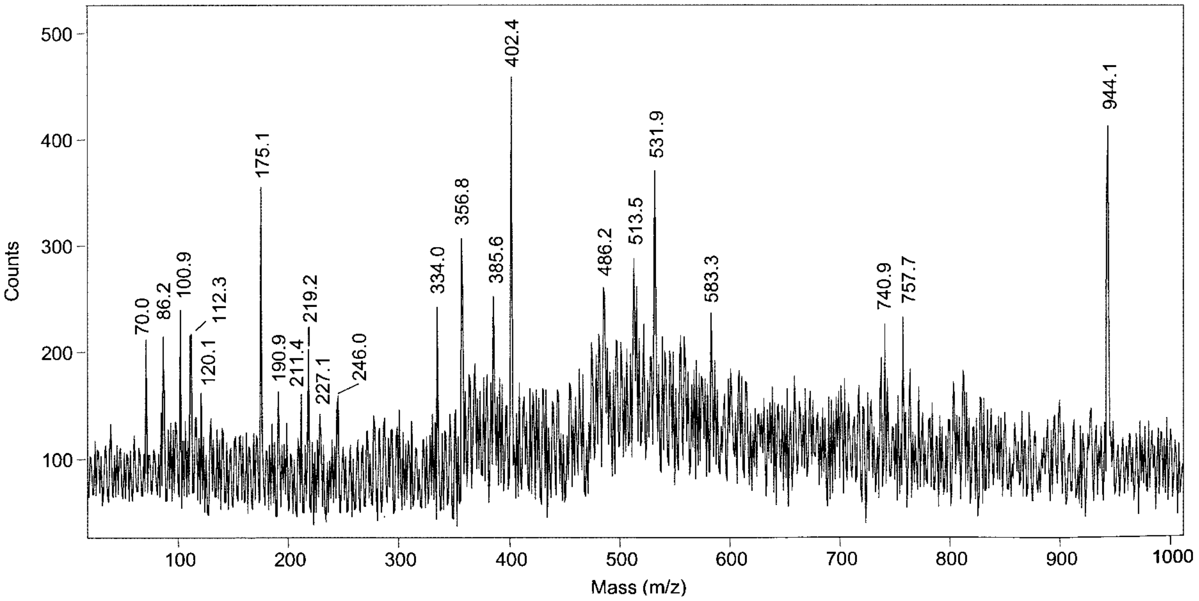


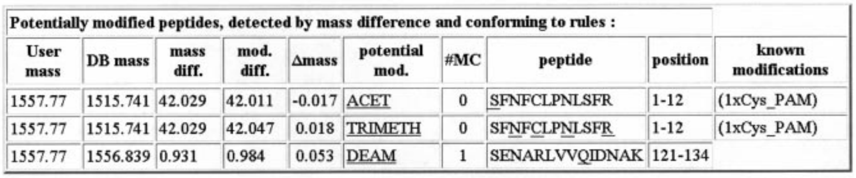
Figure 3. MALDI-TOF PSD spectrum of region *m/z* 50 to 1000, showing fragmentation of peptide mass 1631.8 from EFTU_ECOLI (P02990). See Table 2 for explanation of peptide fragments.

Table 2. Post-source decay fragmentation results for peptide 1631.8 from EFTU_ECOLI (P02990)

Ion type	Amino acid or sequence	Measured mass	Theoretical mass
Immonium	P	70.0	70
	I/L	86.2	86
	Q	100.9	101
	R	112.3	112
	F	120.1	120
a 2	AF-	190.0	191.3
b 2	AF-	219.2	219.3
b 3	AFD-	334.0	334.4
y 1	-R	175.1	175.2
y 2	-AR	246.0	246.3
y 3-NH_3	-K(dimeth)AR	385.6	385.5
y 3	-K(dimeth)AR	402.4	402.5
y 4-NH_3	-EK(dimeth)AR	513.5	514.6
y 4	-EK(dimeth)AR	531.9	531.6
y 6-NH_3	-PEEK(dimeth)AR	740.9	740.8
y 6	-PEEK(dimeth)AR	757.7	757.9
Internal	-PE-	227.1	227.2
Fragments[a]	-PEE-	356.8	356.4
	-EEK(dimeth)A-	486.2	486.5
	-PEEK(dimeth)A-	583.3	583.7
	-DQIDNAPEEK(dimeth)A-	1222.2	1222.3

This confirms the FindMod prediction for the peptide to be of sequence AFDQIDNAPEEKAR and contaiing a dimethylated lysine, K(dimeth). See Figure 3 for the PSD spectrum. Unmatched fragment masses were 105.4, 109.9, 211.4, 944.1, 1202.2, 1447.1.

[a] b ions unless shown otherwise.



Potentially modified peptides, detected by mass difference and conforming to rules :

User mass	DB mass	mass diff.	mod. diff.	Δmass	potential mod.	#MC	peptide	position	known modifications
1557.77	1515.741	42.029	42.011	-0.017	ACET	0	SFNFCLPNLSFR	1-12	(1xCys_PAM)
1557.77	1515.741	42.029	42.047	0.018	TRIMETH	0	SFNFCLPNLSFR	1-12	(1xCys_PAM)
1557.77	1556.839	0.931	0.984	0.053	DEAM	1	SENARLVVQIDNAK	121-134	

Figure 4. FindMod output for peptide of mass 1557.77 from sheep keratin (K1M1_SHEEP; P02534), with mass tolerance 0.1 Da. In the genuine FindMod output, amino acids in the peptide that may carry a modification are shown in blue. In this Figure, these are underlined instead.

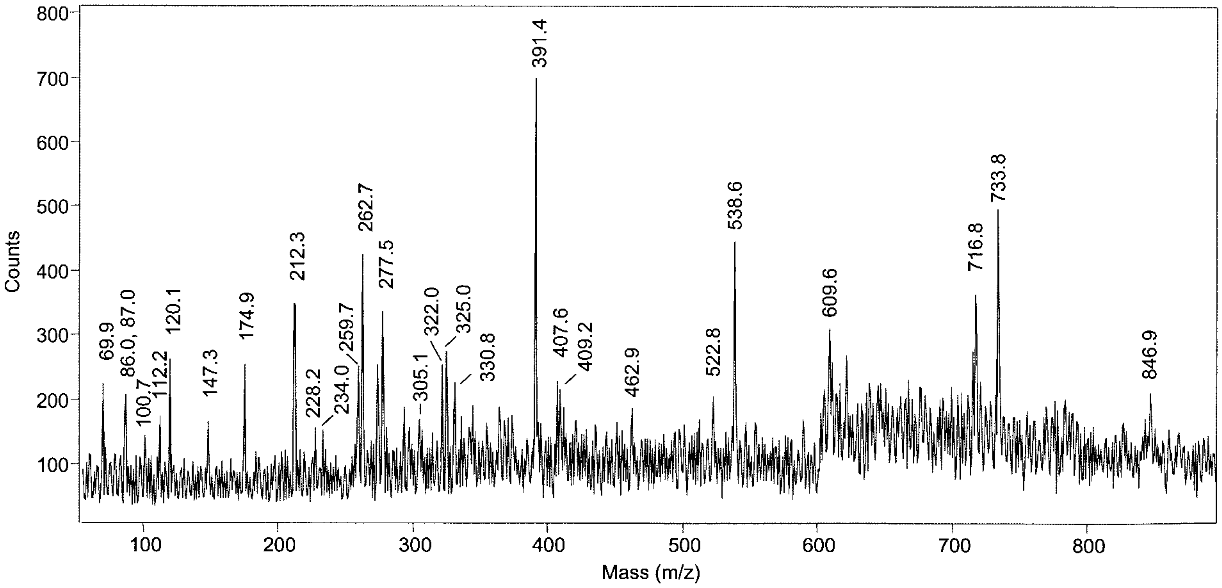


Figure 5. MALDI-TOF PSD spectrum of region *m/z* 50 to 850, showing fragmentation of peptide mass 1557.77 from sheep keratin (K1M1_SHEEP; P02534). See Table 3 for explanation of peptide fragments.

Table 3. Post-source decay fragmentation results for peptide 1557.77 from sheep keratin (K1M1_SHEEP; P02534)

Ion type	Amino acid or sequence	Measured mass	Theoretical mass
Immonium	P	69.9	70
(related ions)	I/L	86	86
	N	87	87
	(R)	100.7	100
	(R)	112.2	112
	F	120.1	120
	C(pam)	147.3	147
b 1-H_2O	Ac-S-	112.2	112.1
b 2-H_2O	Ac-SF-	259.7	259.3
b 2	Ac-SF-	277.5	277.3
b 3	Ac-SFN-	391.4	391.4
b 4	Ac-SFNF-	538.6	538.6
y 1	-R	174.9	175.2
y 2-NH_3	-FR	305.1	305.4
y 2	-FR	322	322.4
y 3	-SFR	409.2	409.5
y 4	-LSFR	522.8	522.6
y 6-NH_3	-PNLSFR	716.8	716.8
y 6	-PNLSFR	733.8	733.8
y 7	-LPNLSFR	846.9	847.0
Internal	-PN-	212.3	212.2
Fragments[a]	-NL-	228.2	228.3
	-NF-/-FN- (a ion)	234	234.3
	-NF-/-FN	262.7	262.3
	-PNL-/-LPN-	325	325.4
	LSF—H2O	330.8	330.4
	-FC(pam)L-	407.6	407.6
	-NLSF-	462.9	462.5

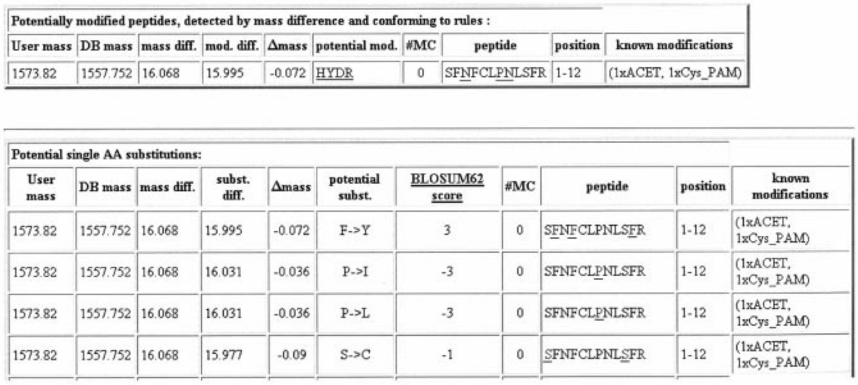
This confirms FindMod predictions for the peptide to be SFNCLPNLSFR with an N-terminal serine acetylation, Ac-S, and artifactual Cys-propionamide, C(pam). See Figure 5 for the PSD spectrum. Unmatched fragment masses were 273.8, 609.6, 1454.9, 1504.1.

[a] b ions unless shown otherwise.

spectra was the discovery of the immonium ion for Tyr (mass 135.9 Da) and the b2, b3 and b4 series of ions.

Case 4: methionine sulphoxide in E. coli *AHPC and species-specific rules*

After identification of a protein from *E. coli* as alkyl hydroperoxide reductase (P26427), a peptide of mass 1381.6 was found not to match any unmodified peptide in the protein. FindMod predicted the peptide to have the sequence YAMIGDPTGALTR containing a methionine sulphoxide residue (Figure 8). Note that artifactual modifications like methionine sulphoxide and Cys-propionamide are classed as ''known'' modifications in FindMod due to their ubiquitous nature in proteins from 2-D gels. FindMod also noted that the 15.9 Da mass



Potentially modified peptides, detected by mass difference and conforming to rules :

User mass	DB mass	mass diff.	mod. diff.	Δmass	potential mod.	#MC	peptide	position	known modifications
1573.82	1557.752	16.068	15.995	-0.072	HYDR	0	SFNFCLPNLSFR	1-12	(1xACET, 1xCys_PAM)

Potential single AA substitutions:

User mass	DB mass	mass diff.	subst. diff.	Δmass	potential subst.	BLOSUM62 score	#MC	peptide	position	known modifications
1573.82	1557.752	16.068	15.995	-0.072	F->Y	3	0	SFNFCLPNLSFR	1-12	(1xACET, 1xCys_PAM)
1573.82	1557.752	16.068	16.031	-0.036	P->I	-3	0	SFNFCLPNLSFR	1-12	(1xACET, 1xCys_PAM)
1573.82	1557.752	16.068	16.031	-0.036	P->L	-3	0	SFNFCLPNLSFR	1-12	(1xACET, 1xCys_PAM)
1573.82	1557.752	16.068	15.977	-0.09	S->C	-1	0	SFNFCLPNLSFR	1-12	(1xACET, 1xCys_PAM)

Figure 6. Truncated FindMod output for peptide of mass 1573.82 from sheep keratin, showing the possible modifications and part of the amino acid substitution table. Mass tolerance was 0.1 Da. In the genuine FindMod output, amino acids in the peptide that may carry a modification are shown in blue, and potentially substituted amino acids shown in red. In this Figure, such amino acids are underlined instead.

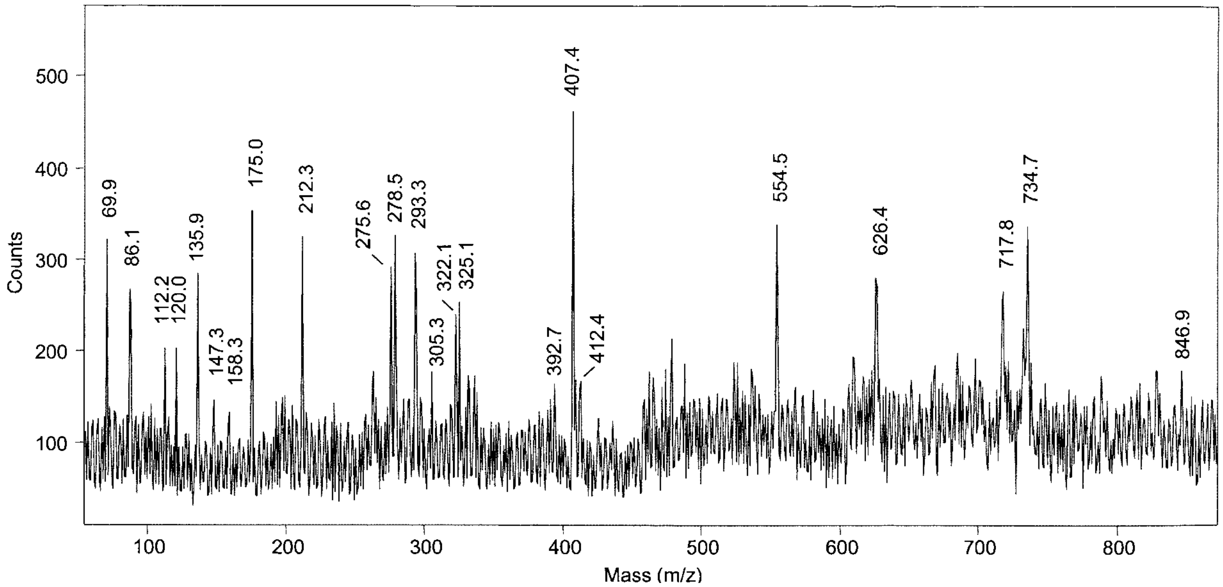


Figure 7. MALDI-TOF PSD spectrum of region *m/z* 50 to 850, showing fragmentation of peptide mass 1573.82 from sheep keratin. See Table 4 for explanation of peptide fragments.

difference could be consistent with an hydroxylation event. However as this modification is yet to be reported in prokaryotes, this was not confirmed by rules and is thus unlikely. The PSD fragmentation results clearly confirmed the peptide sequence as YAMIGDPTGALTR, and localised the extra 15.9 mass units to the Met residue (Figure 9, Table 5).

Discussion

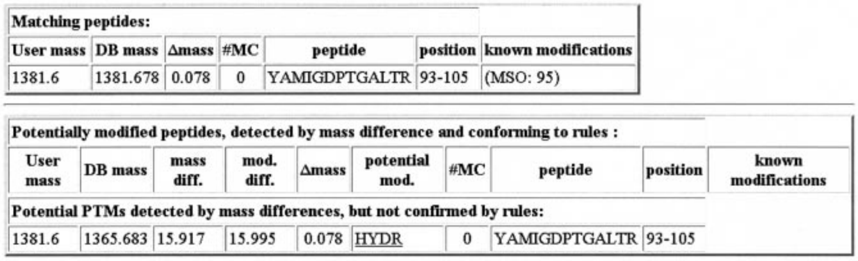
Here we have presented an approach for the systematic analysis of proteins for post-translational modifications and single amino acid substitutions, using mass spectrometry and the tool FindMod. In essence, this approach examines peptide masses

Table 4. Post-source decay fragmentation results for peptide 1573.82 from sheep keratin

Ion type	Amino acid or sequence	Measured mass	Theoretical mass
Immonium	P	69.9	70
(related ions)	I/L	86.1	86
	N	87.1	87
	(R)	112.2	112
	F	120	120
	Y	135.9	136
	C(pam)	147.3	147
b 2-H_2O	Ac-SY-	275.6	275.3
b 2	Ac-SY-	293.3	293.4
b 3	Ac-SYN-	407.4	407.4
b 4	Ac-SYNF-	554.5	554.6
y 1-NH_3	-R	158.3	158.2
y 1	-R	175	175.2
y 2-NH_3	-FR	305.3	305.4
y 2	-FR	322.1	322.4
y 3-NH_3	-SFR	392.7	392.4
y 3	-SFR	409.3	49.5
y 6-NH_3	-PNLSFR	717.8	717.8
y 6	-PNLSFR	734.7	733.8
Internal	-PN-	212.3	212.2
Fragments[a]	-YN-	278.5	278.3
	-FC(pam)-H20	294.4	294.4
	-PNL-/-LPN-	325.1	325.4
	-PNLS-	412.4	412.5
	-FC(pam)LPNLS-/ C(pam)LPNLSF-	846.9	847.0

Fragmentation confirmed a FindMod prediction that the peptide be of sequence SYNFCLPNLSFR with N-terminal acetylation, Ac-S, and artifactual Cys-propionamide, C(pam). See Figure 7 for the PSD spectrum. Unmatched fragment masses were 626.4, 1386.1, 1471.6, 1503.9, 1520.9, 1558.6.

[a] b ions unless shown otherwise.



Matching peptides:						
User mass	DB mass	Δmass	#MC	peptide	position	known modifications
1381.6	1381.678	0.078	0	YAMIGDPTGALTR	93-105	(MSO: 95)

Potentially modified peptides, detected by mass difference and conforming to rules :

User mass	DB mass	mass diff.	mod. diff.	Δmass	potential mod.	#MC	peptide	position	known modifications
Potential PTMs detected by mass differences, but not confirmed by rules:									
1381.6	1365.683	15.917	15.995	0.078	HYDR	0	YAMIGDPTGALTR	93-105	

Figure 8. FindMod output for peptide of mass 1381.6 from *E. coli* alkyl hydroperoxide reductase (P26427), with mass tolerance 0.1 Da. In the genuine FindMod output, amino acids in the peptide that may carry a modification are shown in blue. In this Figure, these are underlined instead.

generated by peptide mass fingerprinting techniques, establishes which peptides match unmodified peptides in the protein, and examines remaining peptides to see if they may carry post-translational modifications or single amino acid substitutions. This is done by examining mass differences between empirical and theoretical peptides and the application of a set of intelligent rules. Where peptides are predicted to be post-translationally modified, predictions can be tested by undertaking fragmentation analysis of the peptide with MS-MS techniques.

We believe that the FindMod approach is a fundamental tool for protein characterisation because of the following: (1) FindMod considerably simplifies the analysis of mass spectrometry data required when searching for post-translational modifications in proteins. This is because FindMod simultaneously searches for 22 post-translational modifications, four artifactual modifications, and, if desired, one or two user-specified modifications. The program suggests which peptides might carry a modification, as well as the amino acids, if any, in the peptide which may be modified. Importantly, FindMod can be requested to look for two post-translational modifications per peptide, and this may be done whilst also considering artifactual modifications (see case (2)). FindMod can also search to see if any single amino acid substitution at any position in any peptide gives a delta mass value that matches with that between an empirical and theoretical peptide. This too can be done at the same time as searching for post-translational modifications (see case (3)). Clearly, this matrix of modifications and substitutions is difficult or impossible to calculate without a dedicated program such as FindMod.

(2) FindMod can be used without preconception as to what modifications one expects to find in a protein. Furthermore, whilst FindMod applies rules to predict which amino acids in a peptide might carry a post-translational modification, pep-

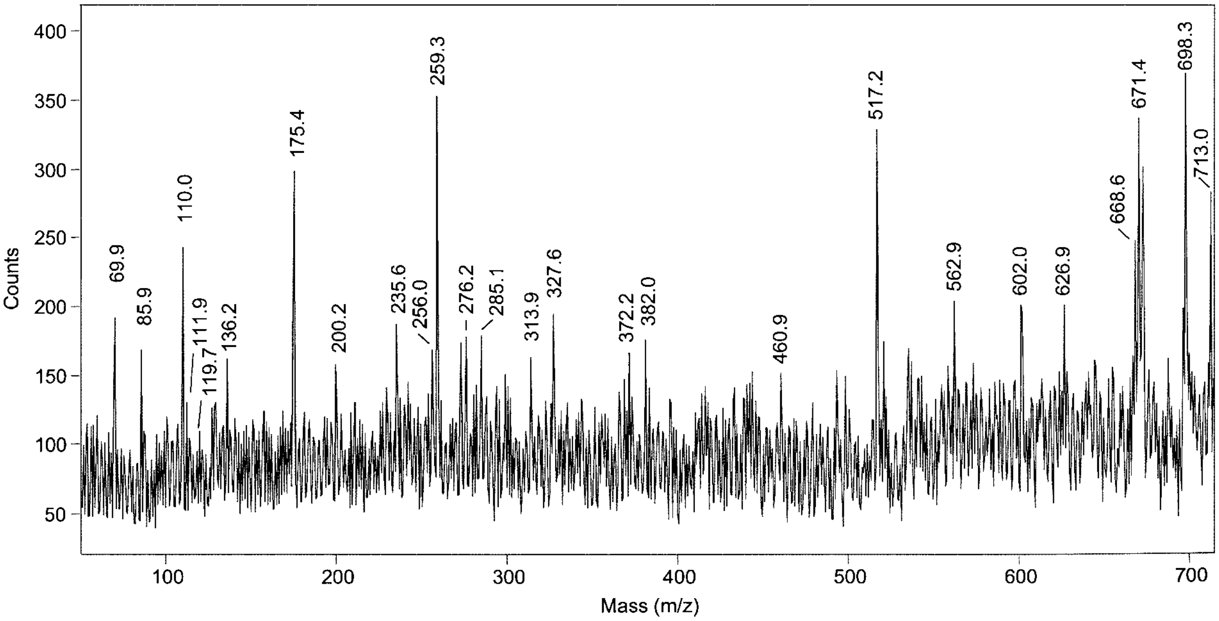


Figure 9. MALDI-TOF PSD spectrum of region *m/z* 50 to 720, showing fragmentation of peptide mass 1381.6 from *E. coli* alkyl hydroperoxide reductase (P26427). See Table 5 for explanation of peptide fragments.

Table 5. Post-source decay fragmentation results for peptide 1381.6 from *E. coli* alkyl hydroperoxide reductase (P26427)

Ion type	Amino acid or sequence	Measured mass	Theoretical mass
Immonium (related ions)	P	69.9	70
	I/L	85.9	86
	R	111.9	112
	M(so)	119.7	120
	Y	136.2	136
a 1	Y-	136.2	136.2
b 2	YA-	235.6	235.3
b 3	YAM(so)-	382	382.5
b 6	YAM(so)IGD-	668.6	667.8
b 10-H_2O	YAM(so)IGDPTGA-	977.3	976.1
a 12	YAM(so)IGDPTGALT-	1179	1180.4
a 3-NH_3	YAM(so)IGDPTGALTR-	1318.6	1319.5
y 1	-R	175.4	175.2
y 2-NH_3	-TR	259.3	259.3
y 2	-TR	276.2	276.3
y 3-NH_3	-LTR	372.2	372.4
y 4	-ALTR	460.9	460.6
y 5	-GALTR	517.2	517.6
y 6-NH_3	-TGALTR	602	601.7
y 7-NH_3	-PTGALTR	698.3	698.8
Internal Fragments[a]	-PT-	200.2	199.2
	-PTG-	256	256.3
	-DPT-	313.9	314.3
	-PTGA-	327.6	327.4
	-M(so)IGDPTG-	671.4	670.8
	-GDPTGALT-	713	713.8

Fragmentation confirmed a FindMod prediction that the peptide is of sequence YAMIGDPTGALTR, where the methionine has been oxidised to a methionine sulphoxide, M(so). See Figure 9 for the PSD spectrum. Unmatched fragment masses were 110, 273.2, 285.1, 562.9, 626.9, 1104.6, and 1366.1.

[a] b ions unless shown otherwise.

tides whose masses suggest they carry a certain modification but which do not conform to rules are also listed to alert the user that a novel form of a modification may be present (see Figure 8). In this manner, FindMod provides the user with a set of hypotheses to test in MS-MS fragmentation analysis of peptides. As MS-MS peptide fragmentation results can be difficult to interpret, it is useful to have a series of predicted peptides against which MS-MS fragmentation results can be checked rather than having to interpret fragmentation results *de novo*.

(3) FindMod considers the annotations concerning protein processing and modifications when proteins of interest are in the SWISS-PROT database. Therefore, proteins are cleaved to mature forms before searching for mass differences between theoretical and empirical peptides. Modified peptides that match with a previously documented modification event, or a modification that SWISS-PROT has deemed as probable due to cross-species similarity or a consensus sequence pattern, are shown separately in the ''known modifications'' column of the output (see Figure 2). This simplifies the user's task of protein characterisation as it can take advantage of approximately 80,500 processing events as well as some 11,000 post-translation modification events documented in the SWISS-PROT database.

To best use FindMod for protein characterisation, a number of technical considerations should be noted. Firstly, proteins should be separated to purity before being submitted to endoproteinase digestion and mass spectrometry. High resolution narrow range 2-D gel electrophoresis (e.g. Sanchez *et al.*, 1997) is ideal in this regard. A pure protein is desirable so that FindMod does not make spurious predictions because of contaminating peptides. Secondly, as FindMod considers a user-specified mass tolerance value in its predictions, the quality of predictions will be increased by high accuracy measurements of peptide masses. ESI-TOF mass spectrometers or MALDI-TOF apparatus equipped with delayed extraction and ion reflectors are ideal for this as most can deliver monoisotopic masses ±40 ppm when two-point internal calibration is used. Less accurate peptide mass data will require a larger mass tolerance in FindMod and will result in larger numbers of possible modifications or amino acid substitutions being suggested. A third technical consideration is that the degree of characterisation of a protein achievable with FindMod will be directly related to the percentage peptide coverage that is obtained during the peptide mass fingerprinting procedure. When working with proteins from 2-D gels, it is therefore advisable that in-gel digestions of reduced and alkylated proteins is done as this will yield higher coverage than some other techniques. However, it must be noted that peptides may not be detected during mass spectrometry if they are not efficiently ionised (Kratzer *et al.*, 1998).

Finally, it is worth exploring the potential that the FindMod approach has for the large-scale discovery of protein post-translational modifications. When subjected to 2-D gel electrophoresis, cellular extracts can be purified into hundreds to thousands of discrete proteins. Peptides from these proteins can be analysed in an essentially automated fashion with use of robotics and automated MALDI mass spectrometry, at the rate of tens to hundreds in one day (Traini *et al.*, 1998). When there is genomic information for the organism under study, the automated analysis of proteins assigns a sequence to protein spots from 2-D gels. At this point FindMod can be given a protein's sequence and its corresponding peptide masses, to give a list of peptides that are potentially modified and the amino acids that might carry the modification. If automated fragmentation of potentially modified peptides can be implemented, for example *via* the use of chip-based automated ESI-MS-MS (Figeys *et al.*, 1997) or through use of automated MALDI-TOF PSD, the potential for screening large numbers of proteins for post-translational modifications can be realised. Clearly this is highly desirable for proteome projects. We are currently extending our previous work (Traini *et al.*, 1998) to provide total integration of such technologies and make this feasible.

Materials and Methods

Two-dimensional gel electrophoresis

Growth of *E. coli* type K-12 strain W3110 and micropreparative 2-D gel electrophoresis of proteins was performed as described by Pasquali *et al.* (1996) and Molloy *et al.* (1998). Preparation of proteins from wool by 2-D gel electrophoresis was as described by Herbert *et al.* (1997).

Protein digestion

Coomassie-blue stained gel spots of interest were cut from 2-D polyacrylamide gels using the ARRM-214 excision robot (Traini *et al.*, 1998) and automatically placed into 96-well plates. The 96-well plates were then placed into a Canberra Packard MultiProbe robot (Downers Grove, Illinois), where they were digested with trypsin (Promega). Peptides were then extracted, and samples spotted onto a 100-sample MALDI-TOF MS sample target with α-cyano-4-hydroxy-cinnamic acid. This digestion and extraction process was done essentially as described by Traini *et al.* (1998), however gel pieces were destained extensively and dried under vacuum before addition of the endoproteinase (Rosenfeld *et al.*, 1992).

Mass spectrometry

Samples were analysed using a PerSeptive Biosystems Voyager-DE STR MALDI mass spectrometer, equipped with a 337 nm N_2 UV laser. Parent ion masses were measured in reflectron/delayed extraction mode, with accelerating voltage of 20 kV, grid voltage of 72.5 % and a 200 ns delay. Between 20 to 50 scans were averaged per sample, and spectra subjected to two-point internal calibration with trypsin autolysis peaks (*m/z* 842.51 and 2211.10) to give a typical mass accuracy of ±30 ppm. Peptides chosen for post-source decay analysis were isolated using the timed ion selector, and a series of spectra collected by first adjusting the reflectron mirror ratio to 0.9, and then stepping in multiples of 0.75 of the previous mirror ratio until a maximum in-focus *m/z* of 50 was achieved. Typically, this required the acquisition of 10 to 12 spectra, where each was the average of 128 scans. Composite PSD spectra were constructed by stitching together the spectra from different *m/z* regions, and calibrating externally with the mass of the parent ion.

FindMod program

FindMod is available on the internet at http://www.expasy.ch/sprot/findmod.html. Users can specify either a SWISS-PROT entry or a protein sequence in single amino acid code for the protein of interest. For SWISS-PROT entries, processing to mature forms is undertaken by reference to annotation for signal sequences, chains and propeptides. If the sequence is entered in amino acids, the user should specify the phylogenetic group the protein sequence is from (e.g. prokaryote, eukaryote, virus, archaebacteria). Cleavage of sequences to yield theoretical peptide masses is according to Wilkins *et al.* (1997), with optional allowance for up to three missed cleavage sites. For SWISS-PROT entries, FindMod will consult feature (FT) tables to yield masses of any peptides carrying known or predicted post-translational modifications. User-entered peptides can be [M] or $[M + H]^+$ and any treatment of cysteine residues can be specified, as can the possibility of oxidation of methionine residues.

For the matching of user-specified peptides to those generated theoretically, FindMod first generates a list of experimental peptides that correspond to unmodified peptides. FindMod then creates a list of experimental peptides that match to peptides which carry modifications to cysteine residues, oxidized methionine residues, or other modifications as documented in SWISS-PROT. All matching peptides must be within a user-specified mass tolerance. To find novel modifications, FindMod then calculates the mass differences between all experimental peptides and all theoretical peptides. If a mass difference corresponds to the mass of any of the modifications in Table 1, peptides are classed as ''potentially modified''. FindMod then examines all potentially modified peptides with a set of intelligent rules (see Table 1) to determine if the protein, and an amino acid within the relevant peptide, can carry the modification in question. Where peptides carry a modification that agrees with the rules, these peptides are classed as ''potentially modified and conforming with rules'' and the amino acid(s) that are potentially modified in the peptide are highlighted in the FindMod output. Importantly, FindMod can be requested to look for two modifications per peptide, in conjunction with artifactual modifications, such that predictions can be made for highly modified peptides. If desired, FindMod can also search for one or two amino acid substitutions in peptides that may account for a mass difference between a query peptide and those from a specified protein sequence. This search can be done at the same time as searches for protein post-translational modifications.

Acknowledgements

M.R.W., E.G. and A.B. were supported by the Helmut Horten Foundation. We also acknowledge the support of the Swiss National Fund for Scientific Research (grant 31-33658-92), the Inter-Maritime and the Montus Foundation. This research has been facilitated by access to the Australian Proteome Analysis Facility established under the Australian Government's Major National Research Facilities Program.

References

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* **26**, 38-42.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217-221.

Clauser, K. R., Hall, S. C., Smith, D. M., Webb, J. W., Andrews, L. E., Tran, H. M., Epstein, L. L. & Burlingame, A. L. (1995). Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE. *Proc. Natl Acad. Sci. USA,* **92**, 5072-5076.

Figeys, D., Ning, Y. & Aebersold, R. (1997). A microfabricated device for rapid protein identification by microelectrospray ion trap mass spectrometry. *Anal. Chem.* **69**, 3153-3160.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C. & Gocayne, J. D. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* **269**, 496-512.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. (1996). Life with 6000 genes. *Science,* **274**, 546-567.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.

Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C. & Watanabe, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl Acad. Sci. USA,* **90**, 5011-5015.

Herbert, B. R., Molloy, M. P., Yan, J. X., Gooley, A. A., Bryson, W. G. & Williams, K. L. (1997). Characterisation of wool intermediate filament proteins separated by micropreparative two-dimensional electrophoresis. *Electrophoresis,* **18**, 568-572.

James, P., Quadroni, M., Carafoli, E. & Gonnet, G. (1993). Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195**, 58-64.

Kratzer, R., Eckerskorn, C., Karas, M. & Lottspeich, F. (1998). Suppression effects in enzymatic peptide ladder sequencing using ultraviolet - matrix assisted laser desorption/ionization - mass spectrometry. *Electrophoresis,* **19**, 1910-1919.

Langen, H., Gray, C., Roder, D., Juranville, J. F., Takacs, B. & Fountoulakis, M. (1997). From genome to proteome: protein map of *Haemophilus influenzae*. *Electrophoresis,* **18**, 1184-1192.

Mann, M. (1994). Sequence database searching by mass spectrometric data. In *Microcharacterisation of Proteins* (Kellner, R., Lottspeich, F. & Meyer, H. E., eds), pp. 223-245, VCH, Weinheim.

Mann, M. & Talbo, G. (1996). Developments in matrix-assisted laser desorption/ionization peptide mass spectrometry. *Curr. Opin. Biotechnol.* **7**, 11-19.

Mann, M., Hojrup, P. & Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22**, 338-345.

Molloy, M. P., Herbert, B. R., Walsh, B. J., Tyler, M. I., Traini, M., Sanchez, J. C., Hochstrasser, D. F., Williams, K. L. & Gooley, A. A. (1998). Extraction of membrane proteins by differential solubilization for separation using two-dimensional gel electrophoresis. *Electrophoresis,* **19**, 837-844.

Pappin, D. J. C., Hojrup, P. & Bleasby, A. J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3**, 327-332.

Pasquali, C., Frutiger, S., Wilkins, M. R., Hughes, G. J., Appel, R. D., Bairoch, A., Schaller, D., Sanchez, J.-C. & Hochstrasser, D. F. (1996). Two-dimensional gel electrophoresis of *Escherichia coli* homogenates: the *Escherichia coli* SWISS-2DPAGE database. *Electrophoresis,* **17**, 547-555.

Roepstorff, P. (1997). Mass spectrometry in protein studies from genome to function. *Curr. Opin. Biotechnol.* **8**, 6-13.

Rosenfeld, J., Capdevielle, J., Guillemot, J. C. & Ferrara, P. (1992). In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Anal. Biochem.* **203**, 173-179.

Sanchez, J.-C., Rouge, V., Pisteur, M., Ravier, F., Tonella, L., Moosmayer, M., Wilkins, M. R. & Hochstrasser, D. F. (1997). Improved and simplified in-gel sample application using reswelling of dry immobilized pH gradients. *Electrophoresis,* **18**, 324-327.

Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H. & Mann, M. (1996). Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl Acad. Sci. USA,* **93**, 14440-14445.

Traini, M., Gooley, A. A., Ou, K., Wilkins, M. R., Tonella, L., Sanchez, J. C., Hochstrasser, D. F. & Williams, K. L. (1998). Towards an automated approach for protein identification in proteome projects. *Electrophoresis,* **19**, 1941-1949.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: platform for investigating biology. *Science,* **282**, 2012-2018.

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science,* **270**, 484-487.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr, Hieter, P., Vogelstein, B. & Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell,* **88**, 243-251.

Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F. & Williams, K. L. (1995). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* **13**, 19-50.

Wilkins, M. R., Lindskog, I., Gasteiger, E., Bairoch, A., Sanchez, J. C., Hochstrasser, D. F. & Appel, R. D. (1997). Detailed peptide characterization using PEPTIDEMASS-a World-Wide-Web-accessible tool. *Electrophoresis,* **18**, 403-408.

Yates, J. R., 3rd, Speicher, S., Griffin, P. R. & Hunkapiller, T. (1993). Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214**, 397-408.

Edited by R. Huber

(Received 26 November 1998; received in revised form 8 April 1999; accepted 12 April 1999)