J. Mol. Biol. (1998) **278**, 599–608

JMB

Protein Identification with N and C-Terminal Sequence Tags in Proteome Projects

Marc R. Wilkins[1,2]*, Elisabeth Gasteiger[2], Luisa Tonella[1], Keli Ou[3] Margaret Tyler[3], Jean-Charles Sanchez[1], Andrew A. Gooley[3] Bradley J. Walsh[3], Amos Bairoch[2], Ron D. Appel[4], Keith L. Williams[3] and Denis F. Hochstrasser[1,2]

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

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

[3]*Australian Proteome Analysis Facility, Macquarie University Sydney NSW 2109, Australia*

[4]*Medical Informatics Division Geneva University Hospital 24 Rue Micheli-du-Crest 1211 Geneva 14, Switzerland*

**Corresponding author*

Genome sequences are available for increasing numbers of organisms. The proteomes (protein complement expressed by the genome) of many such organisms are being studied with two-dimensional (2D) gel electrophoresis. Here we have investigated the application of short N-terminal and C-terminal sequence tags to the identification of proteins separated on 2D gels. The theoretical N and C termini of 15,519 proteins, representing all SWISS-PROT entries for the organisms *Mycoplasma genitalium*, *Bacillus subtilis*, *Escherichia coli*, *Saccharomyces cerevisiae* and human, were analysed. Sequence tags were found to be surprisingly specific, with N-terminal tags of four amino acid residues found to be unique for between 43% and 83% of proteins, and C-terminal tags of four amino acid residues unique for between 74% and 97% of proteins, depending on the species studied. Sequence tags of five amino acid residues were found to be even more specific. To utilise this specificity of sequence tags for protein identification, we created a world-wide web-accessible protein identification program, TagIdent (http://www.expasy.ch/www/tools.html), which matches sequence tags of up to six amino acid residues as well as estimated protein pI and mass against proteins in the SWISS-PROT database. We demonstrate the utility of this identification approach with sequence tags generated from 91 different *E. coli* proteins purified by 2D gel electrophoresis. Fifty-one proteins were unambiguously identified by virtue of their sequence tags and estimated pI and mass, and a further 11 proteins identified when sequence tags were combined with protein amino acid composition data. We conlcude that the TagIdent identification approach is best suited to the identification of proteins from prokaryotes whose complete genome sequences are available. The approach is less well suited to proteins from eukaryotes, as many eukaryotic proteins are not amenable to sequencing *via* Edman degradation, and tag protein identification cannot be unambiguous unless an organism's complete sequence is available.

© 1998 Academic Press Limited

Keywords: protein identification; sequence tag; proteome; two-dimensional gel electrophoresis

Introduction

The identification of proteins separated by two-dimensional (2D) gel electrophoresis is central to proteome projects, which aim to characterise all proteins expressed by a genome or tissue (Wilkins *et al.*, 1995). Accordingly, it has been the focus of much recent research. A wide range of identification strategies has been described, most of which

Present address: M. R. Wilkins, Macquarie University Centre for Analytical Biotechnology, Macquarie University, Sydney NSW 2109, Australia.

Abbreviations used: 2D, two-dimensional; MS-MS, tandem mass spectrometry; IPG, immobilised pH gradient; PVDF, polyvinylidene difluoride; MALDI-TOF, matrix-assisted laser desorption ionisation time of flight.

0022–2836/98/180599–10 $25.00/0/mb981726

© 1998 Academic Press Limited

couple analytical techniques with database searching tools. For example, proteins from 2D gels have been identified by their amino acid composition, by peptide mass fingerprinting, through the generation of protein sequence by tandem mass spectrometry (MS-MS) or post-source decay techniques, and through combinations of these approaches (for reviews, see James, 1997; Wilkins & Gooley, 1997). A feature of many of these approaches is their sensitivity, whereby nanogram quantities (1 to 5 pmol) of starting material can be used for analysis.

Concomitant with the above-mentioned advances in protein identification technology but perhaps less appreciated are recent advances in micropreparative 2D gel electrophoresis. The loading of samples onto immobilised pH gradient (IPG) strips by in-gel rehydration (Rabilloud *et al*., 1994; Sanchez *et al*., 1997) has eliminated precipitation problems experienced with gel loading, shortened focusing time, and dramatically increased the resolution of micropreparative separations. Perhaps most importantly, this technique allows up to 15 mg of a sample to be loaded onto a single gel. When blotted to a PVDF membrane and stained with amido black, gels prepared in this manner can yield more than 1000 proteins in high nanogram to low microgram quantities (Sanchez *et al*., 1997). Hence many low abundance proteins are amenable to analysis not only with high-sensitivity MS techniques but also the robust and straightforward technique of Edman degradation. Some of the high abundance spots, for which as much as 10 to 20 μg of protein may be present (e.g. see Packer *et al*., 1996), may also be amenable to chemical C-terminal sequence analysis.

Given the availability of complete genomes in databases, advances in micropreparative 2D PAGE, and advances in protein identification technologies, the question that arises is not only how proteins can be identified, but what the simplest and most efficient way of doing this actually is. Mass spectrometry is clearly useful in this regard, although it can be quite ''information intensive'', especially data from tandem mass spectrometry or MALDI-TOF post-source decay, and results can be challenging to interpret. The generation of *de novo* sequence data remains challenging. Proteins generally require digestion into peptides before analysis, which can be time-consuming. Clearly, any approach that uses whole proteins, generates easily interpreted data and requires a minimum amount of information for identification would also be useful.

Here we examine the applicability of N and C-terminal protein ''sequence tags'' (Mann & Wilm, 1994; Wilkins *et al*., 1996a) to large-scale protein identification. This is first done at a theoretical level, and then verified with empirical data generated through chemical protein sequencing techniques and a new protein identification program TagIdent. This program allows proteins from 2D gels to be identified with data including sequence tag, estimated pI and molecular mass, keywords and species (or a group of species) of interest. It is available on the world-wide web at: http://www.expasy.ch/www/tools.html

Results

Large scale analysis of N and C-terminal sequence tags

The availability of genome sequences in databases makes possible the large-scale analysis of protein N and C termini. To investigate the uniqueness and thus specificity of protein termini, and to determine the length of sequence tag that will lead to an unambiguous identification, three to five amino acid residues of sequence at the predicted N and C termini of all proteins from a number of molecularly well defined species were studied. These included all proteins in the SWISS-PROT database for *Mycoplasma genitalium*, *Bacillus subtilis*, *Escherichia coli* and *Saccharomyces cerevisiae*. Proteins in SWISS-PROT for humans, which represent only a small percentage of the proteome for this species, were also examined. Proteins were processed to their mature forms according to their SWISS-PROT annotations before the above analysis was undertaken. Protein fragments were ignored, and all plasmid-borne *E. coli* proteins were excluded. A total of 15,519 proteins was considered.

Protein N and C-terminal sequence tags were found to have a wide range of uniqueness, according to the species studied. N-terminal sequence tags of three amino acid residues were found to be unique for 23% of proteins in the small-genome *M. genitalium*, but unique for only 8% of proteins in *S. cerevisiae* (Table 1). N-terminal tags of four amino acid residues were more specific, being unique for 83% (*M. genitalium*) to 41% of proteins (*S. cerevisiae*). N-terminal sequence tags of five amino acid residues showed remarkable specificity, being unique for 97% (*M. genitalium* and *B. subtilis*) to 78% of proteins (human). Compared to N-terminal sequence tags, C-terminal sequence tags were found to be more frequently unique. Sequence tags at the C terminus of three amino acid residues were unique identifiers for 82% of proteins in *M. genitalium*, to 36% of human proteins. The C-terminal tags of four amino acid residues were unique for an impressive 97%, 93% and 92% of proteins from *M. genitalium*, *B. subtilis* and *E. coli*, respectively, and unique for 86% of all proteins in SWISS-PROT for *S. cerevisiae*. Almost all proteins from the prokaryotes studied had unique five residue sequence tags at their C termini, having 99.6% (*M. genitalium*) to 98% (*B. subtilis* and *E. coli*) of unique tags. Five residue C-terminal sequence tags were also very frequently unique in proteins from *S. cerevisiae* (94%) and those from human (81%).

Where N and C-terminal tags were not unique identifiers of proteins, the number of times that

Table 1. Frequency of unique sequence tags at protein N and C termini

Species (total proteome size)	Total proteins examined	Number of unique terminal tags (% of all proteins examined) 3AA N-term	4AA N-term	5AA N-term	3AA C-term	4AA C-term	5AA C-term
Mycoplasma genitalium	469	108	387	456	385	455	467
(469 proteins[a])		(23)	(83)	(97)	(82)	(97)	(99.6)
Bacillus subtilis	1889	230	1178	1769	1105	1775	1842
(approx. 4000 proteins)		(12)	(62)	(97)	(59)	(93)	(98)
Escherichia coli	3456	395	1960	3221	1531	3173	3390
(4285 proteins)		(11)	(57)	(93)	(44)	(92)	(98)
Saccharomyces cerevisiae	4770	360	1963	4022	1788	4084	4459
(5885 proteins)		(8)	(41)	(84)	(37)	(86)	(94)
Human	4935	1137	2779	3858	1762	3656	3998
(approx. 100,000 proteins)		(23)	(56)	(78)	(36)	(74)	(81)

AA, amino acid residues; N-term, N-termimal; C-term, C-terminal.
[a] As revised in SWISS-PROT release 34.

such tags were found in all proteins of a species was tallied. The five most common N and C tags of three to five amino acid residues are shown in Table 2. In the prokaryotes, it was found that whilst some N tags of three amino acid residues were common to large numbers of proteins (e.g. MKK is common to about 4% of all proteins in *B. subtilis* and *M. genitalium*), the most frequent four residues N tag of MKTL was shared between only 12 proteins (representing 0.4% of those from *E. coli*), and the most frequent five residue N tag shared between only four proteins. The three and four amino acid residue tags at prokaryote protein C termini were shared between less proteins than corresponding N-terminal tags, with the most frequent tags of length three residues found to be common to 11 or less proteins, and the most frequent C tags of length four residues common to a maximum of six proteins. Interestingly, C tags of five residues were not found to be less frequent than those of four residues. Where N-terminal or C-terminal tags were common to a group of pro-

Table 2. The five commonest three, four and five amino acid residue sequence tags at protein N and C termini in five organisms, and their frequency of occurrence

Species (number considered)	3AA N-term Tag	Freq.	4AA N-term Tag	Freq.	5AA N-term Tag	Freq.	3AA C-term Tag	Freq.	4AA C-term Tag	Freq.	5AA C-term Tag	Freq.
Mycoplasma genitalium	MKK	21	MKKI	4	MDLKK	3	KKS	4	FFTN	2	VKKRN	2
(469)	MAK	10	MAKK	3	MKKAI	2	KNF	4	KHLK	2	AEFKQ	1
	MIK	8	MDLK	3	MKKIN	2	KNS	4	KKRN	2	AEGKK	1
	MKL	8	MELN	3	MKKVI	2	KRN	4	LSWI	2	AHMRW	1
	MDK	7	MINA	3	MLIAI	2	ELN	3	QKNS	2	AKKGV	1
Bacillus subtilis	MKK	57	MLKK	10	AGTKT	3	AAA	6	AAAA	6	TAAAA	6
(1889)	MKI	30	MKKL	9	MKKIL	3	AKK	6	TAQA	4	STAQA	4
	MKL	29	MKIK	7	MKKLL	3	LGE	6	AVSV	3	GTEPN	3
	MKT	27	MKKA	7	MRLSE	3	LKK	6	IQKG	3	EIERT	2
	MSK	25	MKTK	7	LTAPS	2	LTK	6	KRKF	3	EKLGE	2
Escherichia coli	MSK	42	MKTL	12	MELKK	4	AKK	11	AKKK	4	GSGLS	3
(3456)	MKK	40	MKKI	10	MKILV	4	KKK	9	EAAQ	3	QRTIA	3
	MSE	40	MKKL	9	MKTLI	4	AAQ	7	EAKK	3	AKRGK	2
	MKT	36	MALL	7	RIGAP	4	EEA	7	FGSN	3	ANTAA	2
	MSQ	36	MAKN	6	AEIYN	3	EEE	7	GEKL	3	AYYGQ	2
Saccharomyces	MSS	159	MSSS	31	MVKLT	15	SKK	15	TIAN	10	YTIAN	10
cerevisiae	MSE	96	MVKL	17	MESQQ	9	SKL	14	EVGE	9	REVGE	9
(4770)	MSD	79	MSSE	16	MKVSD	8	AKK	12	HDEL	9	GQPMY	7
	MST	77	MSSL	15	MKENE	7	KKK	12	KWIH	7	NKWIH	7
	MSL	71	MSEE	14	MSSSK	6	DEL	11	QPMY	7	FGLFD	5
Human	MAA	102	GSHS	81[a]	GSHSM	79[a]	LTA	54[a]	SLTA	53[a]	VSLTA	53[a]
(4935)	GSH	81[a]	MAAA	29	DIQMT	20[b]	IKR	29[b]	ACKV	21[a]	TACKV	21[a]
	MAS	75	DIQM	20[b]	CSHSM	12	VSS	25[c]	EIKR	20[b]	VTVSS	17[c]
	MAE	50	IVGG	20	CDLPQ	11	VLG	24[d]	TVLG	20[d]	VEIKR	15[b]
	MAG	47	YGGF	14	MDPNC	11	CKV	21[a]	TVSS	17[c]	LTVLG	14[d]

Tags that are underlined represent the alphabetical top of the list of those that are present the same number of times in the organism, meaning that there can be many other tags of the same frequency that are not shown here.

The list of human tags is dominated by different forms of the [a] HLA class I histocompatibility antigen alpha chain, [b] Ig kappa chain, [c] Ig heavy chain and [d] Ig lambda chain.

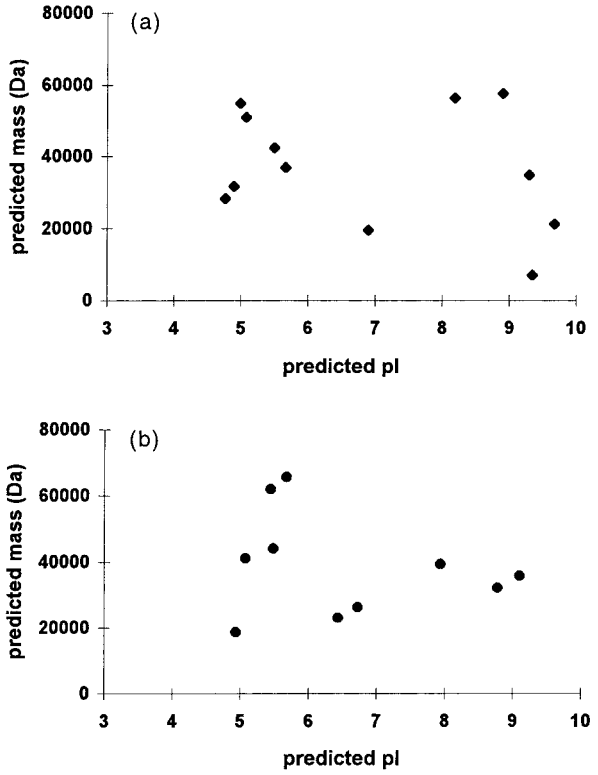

Figure 1. Predicted pI and molecular mass of proteins from *E. coli* that share sequence tags. Note that many of these proteins have pI and mass considerably different from those of other proteins in this group, highlighting the usefulness of these two parameters for distinguishing between proteins that share terminal sequence tags. (a) Proteins from *E. coli* with N tag MKTL, and their predicted pI and mass. (b) Proteins from *E. coli* with the C tag AKK, and their predicted pI and mass. Note that one protein is not on this graph, as it had a mass of 141,000 Da.

teins, it was noted that the differences in pI and molecular mass for each protein may allow the correct protein identity to be selected. For example, Figure 1 shows the predicted pI and molecular mass for *E. coli* proteins sharing the N tag MKTL and the three residue C tag AKK. As many proteins do not have a pI or molecular mass common to other proteins with the same tag, these parameters should be useful in identity searches.

For the eukaryotes *S. cerevisiae* and human, the commonest N and C-terminal tags were tallied (Table 2). In *S. cerevisiae*, the most frequent N tag of three residues (MSS) was shared between 159 proteins, which at 3.3% of the total was similar in occurrence to the most common N tags in prokaryotes. The commonest tag of four residues (MSSS) was present in 31 proteins, which at 0.6% of proteins studied was similar in occurrence to the most frequent prokaryotic tag (MKTL) in *E. coli*. As seen in prokaryotes, the most frequent *S. cerevisiae* C tags were common to relatively small numbers of proteins. Only 15 proteins shared the most frequent C tag of three residues (SKK), and ten proteins shared the commonest C tag of four (TIAN) or five residues (YTIAN). In humans, the list of most frequent tags was found to be dominated by proteins from gene families, notably those for the histocompatibility antigens, and immunoglobulin lambda, kappa and heavy chains. Apart from these, the number of proteins sharing N tags of three amino acid residues was a little lower than seen in yeast, with the most frequent tag (MAA) being common to 1.8% of proteins, and the most frequent N tag of four residues not from a gene family (MAAA) shared between 0.5% of proteins. The most frequent C-terminal tags for proteins not from a gene family (EKP, DEL, KKK, RRH, GEKP, EEVD and SCCA; data not shown) were found to be common to small numbers of proteins, with the most frequent C tag of EKP shared between 17 proteins (0.4% of total) and the tag of GEKP shared between 12 proteins (0.3% of total). C-terminal tags of five amino acid residues did not allow many members of gene families to be distinguished from one another.

Finally, the occurrence of amino acids in the most frequent terminal sequence tags was examined. Interestingly, after the initialising amino acid residue methionine (M), the N termini of proteins from the three prokaryotes showed a strong bias for the charged amino acid lysine (K; Table 2). The N-terminal tag MKK was very frequent in all prokaryotes, and other frequent prokaryote tags often contained one, two or more lysine residues. By contrast, *S. cerevisiae* showed a strong bias for serine (S) after the initialising methionine residue in three to five amino acid residue tags, and human proteins showed some bias for alanine (A). In the most frequent prokaryote protein C termini shown in Table 2, lysine was also found to be often present, although to a lesser degree than at protein N termini. There was some bias for lysine found in the C-terminal tags of *S. cerevisiae*, however this was not seen in proteins from human.

The TagIdent tool

Having found short N and C-terminal sequence tags to be frequently unique and thus useful for the identification of proteins, we developed the tool TagIdent. This matches sequence tags of up to six amino acid residues in length against proteins from a specified species, genus, family, phylum or kingdom of interest in the SWISS-PROT database. Estimated protein pI and molecular mass from 2D gels can also be used as part of the identification procedure if desired, as can any relevant keyword. A sample output from TagIdent is shown in Figure 2.

Identification of proteins from 2D gels using TagIdent

The TagIdent identification approach was tested with 91 different proteins from PVDF blots of *E. coli* 2D gels. Samples were analysed once, except for 16

```
Search performed with following values:
      Query name: 2D-KVF

      pI =        5.35
      Mw =        46086
      delta-pI =  1.00
      delta-Mw =  23043
      OS or OC = \ESCHERICHIA
      KW keyword =     ALL
      Sequence Tag =   SKIV
      Display the N-terminal sequence.
      Print only the sequences matching your tag SKIV.
----------------------
1229 proteins found in pI and MW range
---
Proteins with sequence tag: 2
The number before the sequence indicates the position in the mature
protein where your tag SKIF has been found.
The sequence tag itself is printed in lower case.
---
      ENO_ECOLI   (P08324)
            ENOLASE (EC 4.2.1.11) (2-PHOSPHOGLYCERATE DEHYDRATASE) (2-
            PHOSPHO-D-GLYCERATE HYDRO-LYASE).
            pI: 5.32, MW: 45523.75
      1     skivKIIGREIIDSRGNPTVEAEVHLEGGFVGMAAAPSGA...

      MT57_ECOLI  (P25240)
            MODIFICATION METHYLASE ECO57I (EC 2.1.1.72) (ADENINE-SPECIFIC
            METHYLTRANSFERASE ECO57I) (M.ECO57I).
            pI: 5.82, MW: 62012.96
      194   MKFKADQTSQKLRGGYYTPQNLADYVTKWVLSKNPKTILE...

--------------------
```

Figure 2. Sample output from the TagIdent program for protein 11 from an *E. coli* 2D reference gel (see Table 3). Of the 1229 proteins from *E. coli* within the specified pI and mass range, only ENO_ECOLI and MT57_ECOLI carry the sequence tag SKIV anywhere in their sequence. As ENO_ECOLI carries the tag at the amino terminus, and the pI and mass of the whole protein match well with that of the query protein, it is the correct identification. By comparison, MT57_ECOLI is unlikely to be the protein identification because the tag at position 194 in the protein is approximately one-third of the way through the protein.

proteins that were analysed twice. Proteins were N-terminally sequenced for four cycles, and tag data matched against all SWISS-PROT entries for *E. coli* within the window of pI $\pm$ 1.0 and molecular mass $\pm$ 50%. This allowed 51 proteins to be unambiguously identified (Table 3), their sequence tag matching the amino terminus of only one protein of appropriate pI and molecular mass (e.g. see Figure 2). For these proteins, no further biochemical analysis was undertaken. Proteins that were not unambiguously identified because their sequence tag was not unique and/or all amino acid residues were not assigned in the tag, were further analysed by subjecting the sample of PVDF-bound protein used for microsequencing to amino acid analysis. The protein data of amino acid composition, sequence tag, pI and molecular mass were then matched against *E. coli* proteins in SWISS-PROT using the AACompIdent tool (Wilkins *et al.*, 1996a,b). A further 11 proteins were identified by this approach (Table 4; e.g. see Figure 3). All identified proteins can be seen at the SWISS-2DPAGE *E. coli* database, at URL address: http://www.expasy.ch/ch2d/ch2d-top.html

For all remaining unidentified proteins, the picomolar yields from protein sequencing and amino acid analysis were compared. Four proteins showed discrepancies between picomolar yields of sequence and amino acid analysis that suggested that the proteins were blocked, and a further two proteins were confirmed to be present in low quantities where sequencing gives unreliable results (Table 5). The remaining 23 proteins had sequence tags that did not match against the N terminus of any *E. coli* protein in the SWISS-PROT database, even when matched within the windows of pI $\pm$ 2.0 and a molecular mass range of 0 to 100,000 (Table 5). This could be either because the relevant protein was not yet in the SWISS-PROT database, or because the sequence tag was derived from two comigrating proteins that were not separated by the 2D procedure.

Discussion

We have explored the utility of protein N and C-terminal sequence tags for the identification of proteins. The examination of all available proteins in SWISS-PROT for *M. genitalium*, *B. subtilis*, *E. coli*, *S. cerevisiae* and human revealed that a high number of proteins in each organism had unique sequence tags at their termini, even when only three, four or five amino acid residues of sequence were considered. Where terminal sequence tags of proteins were not unique, it was shown that remarkably few proteins shared any particular three to five residue sequence tag, especially at protein C termini in the prokaryotes and in *S. cerevisiae*.

Whilst we have not considered the full proteomes for all species here, and indeed have considered only about 5% of proteins for humans, some conclusions can be drawn. Firstly, it is clear that sequence tags of three and four amino acid residues at protein C termini are more valuable for protein identification than those at protein N termini. This is mostly because of the high frequency of methionine as the first amino acid residue at protein N termini, which increases the likelihood that a sequence tag will be shared. Secondly, terminal sequence tags are highly suited to protein identification in organisms such as prokaryotes and single-celled eukaryotes with small proteomes (6000 or so proteins), but less suited to large, poorly defined organisms. Thus, whilst there are 160,000 combinations of sequence tag of four amino acid residues, there is considerable bias in the use of amino acids at protein N and C termini, and many tags are found in more than one protein. So, if 0.85% of proteins share a tag in the 469 proteins of *M. genitalium*, this equates to four proteins, and it is likely that these proteins will show differences in pI and molecular mass, allowing their identification. However if 0.85% of proteins in humans share a terminal tag, the proteome size of approximately 100,000 will give a set of 850 proteins. The existence of large protein families in higher eukaryotes that share sequence tags further limits the utility of sequence tag identification in such organisms. It should be noted that accurate TagIdent protein identification in small genome organisms is best applied where genomes are competely known (e.g. *M. genitalium*, *S. cerevisiae*, *E. coli*, *Haemophilus influenzae*, *Methanococcus jannaschii*, *Streptococcus pneumoniae*, *Staphylococcus aur-*

eus, *Archeobacter globus*, *Treponema pallidium* and *Helicobacter pylori*). In this manner, if only one protein within a given pI and molecular mass range is found with a certain terminal tag, one can be confident that there is no other, as yet undescribed, protein that could otherwise match the tag. In fully sequenced organisms the procedure is thus self-checking.

The database searching program TagIdent has a number of distinct features that make it useful in the identification of proteins from 2D gels. It can accept small sequence tags of one to six amino acid residues, and can match against one species, or any group of species defined by a classification term (e.g. prokaryota or mammalia). In this manner, searches are highly directed and more useful for tag protein identification than BLAST or FASTA (Altschul *et al.*, 1990; Pearson & Lipman, 1988), which are global alignment tools that either cannot search with small sequences or return lists containing many irrelevant proteins. TagIdent accepts estimated pI and molecular mass values in searches, which can greatly increase matching power. Thus, if the mass of an unknown protein has been accurately determined with mass spectrometry, TagIdent searches can be done with pre-

Table 3. Identification of 51 *E. coli* proteins with N-terminal sequence tag, estimated pI and mass (*M*)

No.	Estimated *M* (Da)	*M* from database (Da)	Estimated pI	pI from database	Sequence tag	Protein identification	
						Name	SWISS-PROT AC
1	65,203	64,422	5.81	5.85	MKLP	dhsa_ecoli	P10444
2	57,746	58,360	5.75	5.85	ADVP	oppa_ecoli	P23843
3[a]	57,621	52,022	5.83	6.02	MLRI	imdh_ecoli	P06981
4	55,175	58,360	5.58	5.85	ADVP	oppa_ecoli	P23843
5	53,063	51,884	5.54	5.54	SQNV	gltd_ecoli	P09832
6	52,948	51,884	5.61	5.54	SQNV	gltd_ecoli	P09832
7	52,605	57,407	5.78	5.75	KTLV	dppa_ecoli	P23847
8	51,812	49,779	5.83	5.90	AKTL	leu2_ecoli	P30127
9	48,654	47,114	5.37	5.24	MKLY	thrc_ecoli	P00934
10[a]	48,654	46,829	7.59	7.87	AETS	htra_ecoli	P09376
11	46,086	45,524	5.35	5.32	SKIV	eno_ecoli	P08324
12[a]	44,323	45,078	5.83	6.12	APQV	sura_ecoli	P21202
13	43,000	49,779	5.56	5.90	XKTLY	leu2_ecoli	P30127
14	40,419	40,987	5.10	5.08	SVIK	pgk_ecoli	P11665
15	39,779	40,707	5.16	5.22	KIEE	male_ecoli	P02928
16	34,894	32,337	5.70	5.61	MKVA	mdh_ecoli	P06994
17[a]	34,623	32,337	5.57	5.61	MKVA	mdh_ecoli	P06994
18	34,175	35,576	5.39	5.16	AMYQ	ydaa_ecoli	P03807
19	33,734	30,292	5.14	5.22	AEIT	efts_ecoli	P02997
20	33,255	31,270	5.96	5.98	MFTG	dapa_ecoli	P05640
21[a]	33,212	31,138	5.55	5.44	AVVA	yebl_ecoli	P39172
22	32,655	31,138	5.53	5.44	AVVA	yebl_ecoli	P39172
23	32,486	31,138	5.66	5.44	AVVA	yebl_ecoli	P39172
24	30,307	26,233	5.13	5.17	AIPQ	hisj_ecoli	P39182
25	30,182	23,586	5.54	5.55	MRII	kad_ecoli	P05082
26	30,059	28,425	5.89	5.86	AVTK	pmg1_ecoli	P31217
27	29,935	26,972	5.63	5.64	MRHP	tpis_ecoli	P04790
28	28,789	25,042	5.28	5.32	AETI	arti_ecoli	P30859
29	28,203	22,860	5.14	5.20	MTQD	rpia_ecoli	P27252
30	27,972	24,908	5.89	5.97	AEKI	artj_ecoli	P30860
31	25,291	19,965	6.06	6.19	S(TGL)EK	ylad_ecoli	P77791
32	25,239	21,132	5.09	5.42	AQYE	dsba_ecoli	P24991
33	25,084	22,487	5.24	5.16	AEKF	leud_ecoli	P30126
34	24,624	21,135	5.68	5.58	SFEL	sodf_ecoli	P09157
35	24,675	20,532	5.95	6.34	SEAP	nusg_ecoli	P16921
36[a]	23,975	20,630	5.08	5.03	SLIN	ahpc_ecoli	P26427
37	23,584	20,630	5.02	5.03	XLIN	ahpc_ecoli	P26427
38	22,265	18,120	4.8	4.73	GLFD	ptga_ecoli	P08837
39	20,953	18,161	5.41	5.42	ENNA	osmy_ecoli	P27291
40[a]	20,122	17,528	5.00	5.06	MQEG	dksa_ecoli	P18274
41	18,625	19,407	5.37	5.26	AEKR	arok_ecoli	P24167
42	18,420	17,528	5.08	5.06	MQEG	dksa_ecoli	P18274
43	18,218	18,153	5.66	5.51	MVTF	cypb_ecoli	P23869
44	13,854	10,387	5.29	5.15	MNIR	ch10_ecoli	P05380
45	14,132	15,769	6.38	6.17	MQVIL	rl9_ecoli	P02418
46	12,804	15,935	6.14	6.03	MYKT	up12_ecoli	P39177
47	12,562	14,284	5.09	5.09	MITG	yfid_ecoli	P33633
48[a]	11,373	12,164	4.79	4.60	SITK	rl7_ecoli	P02392
49	11,332	14,284	5.08	5.09	MITG	yfid_ecoli	P33633
50	10,989	14,284	5.19	5.09	MITG	yfid_ecoli	P33633
51	9344	11,532	5.71	5.79	MLTV	ygin_ecoli	P40718

X, unknown amino acid residue.
[a] Protein analysed twice.

Table 4. Identification of 11 *E. coli* proteins with N-terminal sequence tag, amino acid composition, estimated pI and mass (*M*)

No.	Estimated *M* (Da)	*M* from database (Da)	Estimated pI	pI from database	Sequence tag	Protein identification Name	SWISS-PROT AC
52[a]	66,200	61,158	4.99	4.89	MESF	rs1_ecoli	P02349
53	57,621	57,138	4.98	4.85	MLRG	ch60_ecoli	P06139
54	56,141	57,138	4.95	4.85	MLRE	ch60_ecoli	P06139
55	55,898	57,138	4.93	4.85	MLXG	ch60_ecoli	P06139
56	51,365	48,193	4.97	4.83	KVRI	tig_ecoli	P22257
57[a]	39,305	39,016	5.70	5.52	SKIF	alf_ecoli	P11604
58	38,452	39,016	5.57	5.52	SKIF	alf_ecoli	P11604
59	33,169	51,164	4.44	4.50	AQVI	flic_ecoli	P04949
60	24,523	16,687	4.81	4.66	MDIR	bccp_ecoli	P02905
61	10,086	15,704	5.76	4.93	MRHY	rs6_ecoli	P02358
62	9781	15,408	5.11	5.44	SEAL	hns_ecoli	P08936

X, unknown amino acid residue.
[a] Protein analysed twice.

cise mass windows (e.g. mass ± 1%). Even if a tag is not available, the use of TagIdent with pI and mass values alone can create a ''shortlist'' of proteins that includes the protein identity (as also in the program PeptideSearch (Mann, 1994)). To assist in the identification of proteins that are known to be post-translationally processed, TagIdent uses annotation in SWISS-PROT to cleave entries to their mature forms before calculating protein pI and molecular mass, and displaying protein N or C termini.

The utility of terminal tag protein identification has been demonstrated here, whereby 56% of 91 different proteins from *E. coli* were identified by N-terminal sequence tags of four amino acid residues. Use of amino acid analysis data in conjunction with a sequence tag raised the total identified to 68%. This is impressive considering the ease of analysis, the minimum of data handling and interpretation required, and that most samples were analysed once only. We believe these data give a good indication of what can be typically achieved when a TagIdent approach is applied to the study of small, well-defined proteomes. However, this raises some questions as to the best way that such tags can be generated analytically. Edman degradation, as used here, is robust and produces easily interpreted N-terminal sequence data. If rapid cycles or parallel processing are used in a 16 cartridge sequenator (Wilkins *et al.*, 1996a; Gooley *et al.*, 1997) it can offer sample throughput of more than 50 samples a week per sequenator, with minimal operator invention. As very large numbers of proteins can be prepared micropreparatively on 2D gels (Sanchez *et al.*, 1997), the pmol to high fmol sensitivity of Edman degradation is acceptable. However, there are organisms such as *S. cerevisiae* where up to 50% of proteins are N-terminally blocked, thus reducing the efficiency of identification with N tags.

Sequence tags at protein C termini are more specific than those at protein N termini, but C-terminal sequencing methods are not well established. Chemical sequencing at protein C termini can be undertaken on proteins prepared by electrophoresis and blotting to Teflon or Gore-Tex membranes (Burkhart *et al.*, 1996), but more than 50 pmol of protein is usually needed per sample, which restricts the analysis to relatively few proteins on any 2D gel. By comparison with chemical sequencing techniques, the majority of mass spectrometric protein identification techniques involve the study of enzymatically generated peptides, from which sequence tags can be generated by techniques including MS-MS or MALDI-TOF post-source decay fragmentation, carboxypeptidase digestion, or chemical ladder sequencing techniques (Mann & Wilm, 1994; Griffin *et al.*, 1995;

```
**************************************************************************
SEARCH VALUES:
Calibration protein: ALBU_BOVIN ( P02769 )
Species searched: ESCHERICHIA
Keyword searched: ALL
Name given to unknown protein: ALBU_BOVIN
Tagging: SKIF
The N-terminal sequence of the protein will be printed.
**************************************************************************

pI:   5.57  Range: (  5.07,   6.07)
Mw:  38452  Range: ( 30762,  46142)

The SWISS-PROT entries having pI and Mw values in the specified range
for the specified species and keyword:

Rank Score   Protein    (pI      Mw)  N-terminal Sequence
=========================================================
*  1    15 ALF_ECOLI    5.52    39016 skifDFVKPGVITGDDVQKVFQVAKENNFALPAVNCVGTD
   2    16 YHHX_ECOLI   6.07    38765 MVINCAFIGFGKSTTRYHLPYVLNRKDSWHVAHIFRRHAK
   3    23 YEBL_ECOLI   5.27    31135 AVVASLKPVGFIASAIADGVTETEVLLFDGASEHDYSLRP
   4    24 AGP_ECOLI    5.38    43560 QTVPEGYQLQQVLMMSRHNLRAPLANNGSVLEQSTPNKWP
   5    25 F16P_ECOLI   5.67    36834 MKTLGEFIVEKQHEFSHATGELTALLSAIKLGAKIIHRDI
   6    26 RFFG_ECOLI   5.63    39684 MRKILITGGAGFIGSALVRYIINETSDAVVVVDKLTYAGN
   7    27 RBB2_ECOLI   5.09    40641 MKILVTGGAGFIGSAVVRHIINNTQDSVVNVDKLTYAGNL
   8    28 YCL2_ECOLI   6.05    37602 MKYLVTGAAGFIGFHVSKRLLEAGHQVVGIDNLNDYYDVS
   9    30 ACKA_ECOLI   5.85    43290 MSSKLVLVLNCGSSSLKFAIIDAVNGEEYLSGLAECFHLP
  10    31 YAIM_ECOLI   5.54    31324 MELIEKHVSFGGWQNMYRHYSQSLKCEMNVGVYLPPKAAN
**************************************************************************
```

Figure 3. AACompIdent can be used to clarify ambiguous TagIdent searches. Protein 58 from *E. coli* had an N sequence tag SKIF, which was common to the N termini of proteins ALF_ECOLI and CYSK_ECOLI. The PVDF-bound protein used for sequence analysis was subsequently used for amino acid analysis, and the resulting data, in conjunction with protein pI, mass and sequence tag, was matched against SWISS-PROT using the AACompIdent tool (Wilkins *et al.*, 1996a). The identity of the protein was confirmed to be ALF_ECOLI, as it was the first-ranked protein in the AACompIdent output, and carried the sequence tag. Note that the asterisk beside the protein rank indicates if the sequence tag has been found anywhere in the sequence of the corresponding protein.

Table 5. Proteins not identified with their sequence tags (numbers 63 to 85) and proteins where sequence tags were not able to be determined (numbers 86 to 91)

No.	Estimated *M* (Da)	Estimated pI	Sequence tag
63[a]	66,200	4.96	R(GF)QDE[b]
64	57,247	6.41	S(PV)IN
65	55,415	5.60	ANVP
66	40,581	5.15	AVAA
67	36,873	5.60	MIKF
68	36,726	5.38	AEGF
69[a]	35,305	4.84	(DS)GES
70	35,076	4.94	XVAEF
71	35,030	4.92	KVAE
72	33,515	5.74	ANLX
73[a]	33,039	5.77	(TSA)NLK
74	30,307	4.74	NYGA
75	28,907	5.74	RFIQ
76	26,245	5.08	SVMD
77[a]	21,992	6.41	SEMA
78	16,863	6.86	KLLD
79	15,381	4.91	(MRSH)(TL)EL
80[a]	13,874	5.52	KLNF
81	12,911	5.75	MNKF
82	12,646	5.05	FAV
83	12,133	6.36	QNI
84	11,670	5.85	SGKK
85	11,542	5.46	KLSG
86[a]	42,658	5.94	—
87	38,683	5.84	—
88	24,981	5.46	—
89	52,037	5.96	—
90	47,714	5.00	—
91	43,437	5.43	—

X, unknown amino acid residue.
[a] Protein analysed twice.
[b] Residues in parentheses were found simultaneously in one cycle of sequencing.

Patterson *et al.*, 1995; Thiede *et al.*, 1995; Knierman *et al.*, 1994; Bartlet-Jones *et al.*, 1994). However, none of these techniques is currently being used routinely to generate N or C-terminal sequence data with whole proteins. This is due to either the difficulty of accurately measuring mass differences between a protein and the same protein minus one or more amino acid residues, or the inefficiencies of the chemical or enzymatic agents in removing terminal amino acid residues from large polypeptides. Interestingly, as lysine is very common in protein N and C-terminal tags and the enzymes trypsin and Lys-C are almost exclusively used for the cleavage of proteins in preparation for mass fingerprinting, peptides at protein N and C termini will frequently be two or three amino acid residues in length and probably not analysed in mass spectrometry procedures.

The recommended strategy for use of TagIdent can be summarised as follows. (1) Generate terminal tag data for protein from 2D gel. (2) Match tag against SWISS-PROT using TagIdent for the species of interest within a suitable pI and molecular mass range. (3) If results are unambiguous (as in Figure 2) and the organism molecularly well-defined, the protein identification can be accepted. If the sequence tag is not found at any protein terminus, check that the protein identity is not outside the chosen pI and mass range by rematching with larger windows. Alternatively, rematch using the amino acid residue X for any ambiguous residue in the tag. (4) If the sequence tag is found to be common to many protein termini, further analytical data for the same protein are needed. The same or a duplicate sample can be used for amino acid analysis and the compositional data and sequence tag matched against SWISS-PROT using the AACompIdent tool (Wilkins *et al.*, 1996a,b). The picomolar yields from the different approaches can be compared to determine if proteins are blocked. Alternatively, a duplicate sample can be used for peptide mass fingerprinting, and the peptide mass data and sequence tag matched against databases using the MS-Edman program (Clauser & Baker, at http://falcon.ludwig.ucl.ac.uk/msedman.htm). Provided the protein is in sequence databases, these approaches should resolve its identity.

In conclusion, here we have described a method and a program, TagIdent, that exploits the specificity of N or C-terminal sequence tags for protein identification. This is suited for the screening of proteins separated by 2D PAGE from small, molecularly defined organisms. We expect the utility of this approach to grow with the increasing availability of genomes and proteomes in databases.

Materials and Methods

Protein databases, sequences and processing

All proteins for theoretical evaluation were from SWISS-PROT release 35.0 and updates to December 1997 (Bairoch & Apweiler, 1998), and all were processed to mature forms before being used for analysis. Any database entry known to represent a protein fragment was not used in this study.

TagIdent protein identification program

The program TagIdent compares a user-specified tag of six amino acid residues or less, including X for any unknown amino acid residues in that sequence, against protein sequences of interest in the SWISS-PROT database (Bairoch & Apweiler, 1998). Sequences of interest are selected by specifying windows of protein pI (estimated from 2D gels), protein molecular mass (estimated from a gel or measured using mass spectrometry), and by defining a species or group of species of interest (by entering a word from the organism species (OS) or organism classification (OC) line in the SWISS-PROT database). SWISS-PROT keywords can be used as part of the search strategy. Before user sequence tags are compared with database sequences, SWISS-PROT entries are first processed to their mature forms, through the removal of known signal sequences, transit peptides and propeptides. Processing of database entries to multiple mature chains is also automatically undertaken where appropriate. Users can choose the way that results should be displayed, thus accommodating tag data derived from protein N termini, C termini, or from internal peptides. TagIdent is available on the ExPASy World-Wide Web server (Appel *et al.*, 1994) at URL address http://www.expasy.ch/www/tools.html, and

search results are sent to the user by e-mail. Further details of program code are available from the authors.

Two-dimensional gel electrophoresis, blotting, and gel image analysis

Whole-cell lysates of *E. coli* were prepared according to Pasquali *et al*. (1996), and 4 mg was separated using micropreparative 2D PAGE. Briefly, commercial sigmoidal immobilised gradient strips of pH range 3.5 to 10 (Pharmacia) were loaded using in-gel rehydration and focused for 100 kV hours according to Sanchez *et al*. (1997). Strips were then equilibrated, treated with iodoacetamide, and run on a standard second-dimension polyacrylamide gel as described (Hochstrasser *et al*., 1988a,b). Gels were blotted to PVDF membranes in a solution of 10 mM Caps (pH 11.0) in 10% (v/v) methanol and, after extensive rinsing in water, blots were stained with amido black and dried. High-resolution scans of these blots were matched against the *E. coli* SWISS-2DPAGE reference map (Pasquali *et al*., 1996) using the Melanie II software (BioRad; Wilkins *et al*., 1996c) to provide estimated pI and molecular mass values for proteins of interest.

Protein sequencing and amino acid analysis

PVDF spots were excised and subjected to four cycles of Edman degradation on a Beckman LF3000 protein sequenator equipped with a prototype 16 cartridge carousel (Gooley *et al*., 1997) or an ABD Procise 494 sequenator equipped with four cartridges. Where necessary, spots used for sequencing were subsequently used for automated amino acid analysis as described (Yan *et al*., 1996; Ou *et al*., 1996; Wilkins *et al*., 1996a).

Acknowledgements

M.R.W. and E.G. 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

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Appel, R. D., Bairoch, A. & Hochstrasser, D. F. (1994). A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trend Biochem. Sci.* **19**, 258–260.

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

Bartlet-Jones, M., Jeffery, W. A., Hansen, H. F. & Pappin, D. J. (1994). Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Commun. Mass Spectrom.* **8**, 737–742.

Burkhurt, W. A., Moyer, M. A., Bailey, J. M. & Miller, C. G. (1996). Electroblotting of proteins to Teflon tape and membranes for N- and C-terminal sequence analysis. *Anal. Biochem.* **236**, 364–367.

Gooley, A. A., Ou, K., Russell, J., Wilkins, M. R., Sanchez, J. C., Hochstrasser, D. F. & Williams, K. L. (1997). A role for Edman degradation in proteome studies. *Electrophoresis,* **18**, in the press.

Griffin, P. R., MacCoss, M. J., Eng, J. K., Blevins, R. A., Aaronson, J. S. & Yates, J. R., III (1995). Direct database searching with MALDI-PSD spectra of peptides. *Rapid Commun. Mass. Spectrom.* **9**, 1546–1551.

Hochstrasser, D. F., Harrington, M., Hochstrasser, A.-C., Miller, M. J. & Merril, C. R. (1988a). Methods for increasing the resolution of two-dimensional protein electrophoresis. *Anal. Biochem.* **173**, 424–435.

Hochstrasser, D. F., Patchornik, A. & Merril, C. R. (1988b). Development of polyacrylamide gels that improve the separation of proteins and their detection by silver staining. *Anal. Biochem.* **173**, 412–423.

James, P. (1997). Of genomes and proteomes. *Biochem. Biophys. Res. Commun.* **231**, 1–6.

Knierman, M. D., Coligan, J. E. & Parker, K. C. (1994). Peptide fingerprints after partial acid hydrolysis: analysis by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **8**, 1007–1010.

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. & Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399.

Ou, K., Wilkins, M. R., Jan, J. X., Gooley, A. A., Fung, Y., Scheumack, D. & Williams, K. L. (1996). Improved high-performance liquid chromatography of amino acids derivatised with 9-fluorenylmethyl chloroformate. *J. Chromatog. sect. A,* **723**, 219–225.

Packer, N. H., Wilkins, M. R., Golaz, O., Lawson, M., Gooley, A. A., Hochstrasser, D. F., Redmond, J. W. & Williams, K. L. (1996). Characterisation of human plasma glycoproteins separated by 2-D gel electrophoresis. *Bio/Technology,* **14**, 66–70.

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 *E. coli* SWISS-2DPAGE database. *Electrophoresis,* **17**, 547–555.

Patterson, D. H., Tarr, G. E., Regnier, F. E. & Martin, S. A. (1995). C-terminal ladder sequencing via matrix assisted laser desorption mass spectrometry coupled with carboxypeptidase Y time-dependent and concentration-dependent digestions. *Anal. Chem.* **67**, 3971–3978.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444–2448.

Rabilloud, T., Valette, C. & Lawrence, J. J. (1994). Sample application by in-gel rehydration improves the resolution of two-dimensional electrophoresis with immobilized pH gradients in the first dimension. *Electrophoresis,* **15**, 1552–1558.

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

Thiede, B., Wittmann-Liebold, B., Bienert, M. & Krause, E. (1995). MALDI-MS for C-terminal sequence determination of peptides and proteins degraded by carboxypeptidase Y and P. *FEBS Letters,* **357**, 65–69.

Wilkins, M. R. & Gooley, A. A. (1997). Protein identification in proteome projects. In *Proteome Research: New Frontiers in Functional Genomics* (Wilkins, M. R., Williams, K. L., Appel, R. D. & Hochstrasser, D. F., eds), pp. 35–64, Springer-Verlag, Berlin, Heidelberg.

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., Ou, K., Appel, R. D., Sanchez, J.-C., Yan, J. X., Golaz, O., Farnsworth, V., Cartier, P., Hochstrasser, D. F., Williams, K. L. & Gooley, A. A. (1996a). Rapid protein identification using N-terminal ''sequence tag'' and amino acid analysis. *Biochem. Biophys. Res. Commun.* **221**, 609–613.

Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L. & Hochstrasser, D. F. (1996b). From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology,* **14**, 61–65.

Wilkins, M. R., Sanchez, J.-C., Bairoch, A., Hochstrasser, D. F. & Appel, R. D. (1996c). Integrating 2D gel databases using the Melanie II software. *Trends Biochem.* **21**, 496–497.

Yan, J. X., Wilkins, M. R., Ou, K., Gooley, A. A., Williams, K. L., Sanchez, J-C., Golaz, O., Pasquali, C. & Hochstrasser, D. F. (1996). Large scale amino acid analysis for proteome studies. *J. Chromatogr. sect. A,* **736**, 291–302.

Edited by R. Huber

(Received 14 August 1997; received in revised form 18 February 1998; accepted 24 February 1998)