J. theor. Biol. (1997) **186,** 7–15

Cross-Species Protein Identification using Amino Acid Composition, Peptide Mass Fingerprinting, Isoelectric Point and Molecular Mass: A Theoretical Evaluation

MARC R. WILKINS AND KEITH L. WILLIAMS*

Macquarie University Centre for Analytical Biotechnology, Macquarie University, Sydney, NSW 2109, *Australia*

(*Received on* 28 *September* 1995, *Accepted in revised form on* 11 *June* 1996)

Proteins can be identified by rapid techniques that do not involve Edman degradation sequencing. These approaches entail the matching of amino acid compositions or tryptic peptide masses of proteins against databases, often in conjunction with estimated protein molecular weight and isoelectric point. As genome sequencing projects progress, proteins from poorly molecularly defined organisms will increasingly be identified by cross-species comparison to proteins from well-defined organisms. To investigate the application of rapid techniques for cross-species protein identification, a total of 65 theoretical cross-species comparisons involving 21 proteins (nine human and 12 *E. coli*) were undertaken. The degree of conservation of amino acid composition, tryptic peptides, protein isoelectric point and mass was established. Protein amino acid composition was well conserved across species boundaries, whilst tryptic peptides were poorly conserved. The molecular weight of proteins was generally well conserved, but protein isoelectric point was not. These results suggest that cross-species protein identification by rapid techniques will be done best by protein amino acid composition and protein molecular weight.

Introduction

The identification of proteins separated by two-dimensional electrophoresis links proteomes (the PROTEin complement expressed by a genOME or tissue) to genomes. It forms the basis of protein-based gene expression analysis [for review see Wilkins *et al.* (1995)]. Two techniques have recently emerged as rapid means of identification in this regard. Protein identification by peptide mass fingerprinting involves the creation of peptides from proteins by endoproteinase digestion, the determination of the masses of these peptides by mass spectrometry, and the matching of masses against a library of theoretical peptide masses generated from protein databases (for reviews see Patterson, 1994; Cottrell, 1994). As this technique relies on the specificity of endoproteinases for certain amino acids, it is a sequence specific method of identification. Use of amino acid (AA) composition for protein identification involves the determination of the AA composition of a protein by chromatographic or radiolabelling techniques, followed by the matching of this data against the theoretical AA compositions of proteins in databases (Garrels *et al.*, 1994; Hobohm *et al.*, 1994; Wilkins *et al.*, 1996; Yan *et al.*, 1996). This means of identification does not directly rely on protein sequence, yet can identify proteins accurately and with confidence. For example, one recent study correctly identified 75% of a group of 59 *E. coli* proteins prepared by two-dimensional electrophoresis (Wilkins *et al.*, 1996). The isoelectric point (pI) of proteins and their molecular weight (MW), estimated

* Author to whom correspondence should be addressed.

0022–5193/97/090007 + 09 $25.00/0/jt960346

from two-dimensional gels, are parameters also commonly used to aid identification.

Rapid protein identification techniques have in the most part been designed for, and applied to, the identification of proteins from one organism by matching against database entries for the species in question. However, it was recently shown that these methods can be applied to ''cross-species matching'', where proteins from poorly molecularly defined species are identified by matching against database entries for all other species (Cordwell *et al*., 1995; Wasinger *et al*., 1995). This potentially allows rapid and detailed protein-based definition of organisms without the need for DNA sequencing initiatives. However, the degree of conservation of AA composition, tryptic peptide maps, pI and MW for protein types from species to species, and how this affects protein identification by cross-species matching, is unexplored. Here we present a theoretical evaluation of these issues.

Methods

Test proteins from *Escherichia coli* and human sera were from the SWISS-PROT database (Bairoch & Boeckmann, 1994). For amino acid (AA) composition cross-species matching, the theoretical percentage composition of proteins was determined for the 16 AAs usually measured by AA analysis after acid hydrolysis of proteins (Asx, Glx, Ser, His, Gly, Thr, Ala, Pro, Tyr, Arg, Val, Met, Ile, Leu, Phe, Lys, where Asx = Asn plus Asp, and Glx = Gln plus Glu). Compositions of test proteins were then matched against all entries of the SWISS-PROT database without the use of pI or MW windows, using the AACompIdent program from the ExPASy world wide web server (Appel *et al*., 1994a; Wilkins *et al*., 1996). This provided a list of best-matching proteins, ranked by a score which represents the degree of difference between the composition of the query protein and a protein in the database. Lists were examined for proteins that had the same identity as the test protein, but from a different species. The scores of such proteins were noted. A score of zero is a perfect match, and larger scores represent increasing compositional differences.

For peptide cross-species matching, theoretical peptide maps were created by ''cleaving'' test proteins with trypsin using Peptidesort in the GCG suite of programs (Devereux *et al*., 1984). Peptides in the mass range 700–3000 were then matched against the MOWSE/OWL database with peptide mass error tolerance of zero, and a MW window of 10% around the known MW value of the entire protein (Pappin *et al*., 1993). This yielded a list of best-matching proteins, ranked according to the percent of peptides from the query protein in common with a database entry. Lists were examined for proteins that had the same identity as the test protein, but from a different species, and their percent peptide identities were noted. Where proteins were identified across species boundaries by AA composition (above) but were not found in lists from tryptic peptide cross-species matching, percent peptide conservation was calculated manually.

The GCG Bestfit program (Devereux *et al*., 1984) was used to establish percent sequence identity of pairs of cross-species matched proteins examined by the above techniques. The pI and MW of cross-species matched proteins were determined using the Compute pI/MW or AACompIdent programs from the ExPASy world wide web server (Appel *et al*., 1994a, b), and the absolute values of differences between pairs calculated. All pI differences were calculated in units and MW differences in percent to correspond with the manner used in most AA composition and peptide mass fingerprinting database matching programs.

Results

AN EXAMPLE OF CROSS-SPECIES MATCHING

Cross-species matching of theoretical AA composition and tryptic peptides of *E. coli* DNA-binding protein H-NSA showed the potential of both techniques to identify proteins across species boundaries. The best matching proteins against *E. coli* DNA-binding protein H-NSA were the same protein from *Shigella flexneri*, *Salmonella typhimurium*, *Serratia marcescens*, and *Proteus vulgaris* (Tables 1 and 2). The degree of conservation of AA composition and tryptic peptides was generally reflected in the degree of protein sequence identity (Table 3). The proteins matched cross-species against *E. coli* showed a maximum pI difference of 0.96 units, and a maximum MW difference of 1.8%.

THEORETICAL CROSS-SPECIES MATCHING OF *E. COLI* AND HUMAN SERA PROTEINS

To better define the degree of protein conservation in cross-species matching, a total of 65 cross-species comparisons of 21 proteins were undertaken (Table 4). Proteins were chosen from *E. coli* and humans as these organisms are the best characterised bacteria and vertebrates, respectively, and will be reference organisms for much cross-species matching in the future. Each cross-species protein comparison

TABLE 1
Ten best matches in the SWISS-PROT database against the amino composition of HNSA_ECOLI

		Protein HNSA_ECOLI			
The closest SWISS-PROT entries for the species ALL:					
Rank	Score	Protein	(pI	MW)	Description
1	0	HNSA_ECOLI	5.44	15408	DNA-BINDING PROTEIN H-NS
2	0	HNSA_SHIFL	5.44	15408	DNA-BINDING PROTEIN H-NS
3	9	HNSA_SALTY	5.32	15411	DNA-BINDING PROTEIN H-NS
4	15	HNSA_PROVU	5.03	15134	DNA-BINDING PROTEIN H-NS
5	18	HNSA_SERMA	6.40	15472	DNA-BINDING PROTEIN H-NS
6	21	MYSN_DROME	5.54	232016	MYOSIN HEAVY CHAIN, NON-MUSCLE
7	24	MYSC_CAEEL	5.99	223008	MYOSIN HEAVY CHAIN C (MHC C)
8	27	HNSB_ECOLI	7.95	15347	DNA-BINDING PROTEIN H-NSB
9	27	MYSA_CAEEL	5.51	225508	MYOSIN HEAVY CHAIN A (MHC A)
10	31	MYSB_HUMAN	5.63	223112	MYOCIN HEAVY CHAIN, CARDIAC MUSCLE

Matching was done using the AACompID program from the ExPASy server (Wilkins *et al*. 1995) without the use of pI or MW windows.

examined differences in AA composition, tryptic peptide maps, pI and MW as for the example of *E. coli* DNA-binding protein H-NSA above. The percent identity in AA sequence was also determined for all comparisons as a benchmark. This revealed some clear trends.

CROSS-SPECIES CONSERVATION OF AMINO ACID COMPOSITION

Amino acid compositions of test proteins were well conserved across species, with pairs of proteins showing a gradual increase in AA difference scores as their percent sequence identity decreased (Fig. 1). With five exceptions in 56, protein pairs with sequence identities of 60% or more had AA composition difference scores of less than 20. The AA difference score was above 25 in only nine of 65 matches. Amino acid composition was extremely well conserved in two cases, where pairs of proteins with sequence identities of 44% and 46% had AA composition difference scores of less than 25. There was a correlation between AA composition difference scores of protein pairs and their percent sequence identity ($R^2 = 0.67$).

TABLE 2
Five best matches in the MOWSE/OWL database (*Pappin et al*., 1993) *against the theoretical tryptic digest of HNSA_ECOLI, showing peptide conservation with E. coli*

Escherichia coli	*Shigella flexneri*	*Salmonella typhimurium*	*Serratia marcescens*	*Proteus vulgaris*
Rank from MOWSE matching	2	3	4	5
2227	2227	1568	1568	957
1568	1568	1419	957	805
1419	1419	1304	805	742
1304	1304	957	742	707
957	957	805	707	
950	950	742		
805	805	707		
778	778			
742	742			
707	707			
Percent common with *E. coli*	100	70	50	40

Only peptides in mass range 700–3000 were used in matching. Peptide error mass tolerance was 0, and MW window for the entire protein was 15000 ± 10%.

CROSS-SPECIES CONSERVATION OF TRYPTIC PEPTIDES

Tryptic peptides of MW 700–3000 were not well conserved across species, with pairs of proteins showing a rapid decrease in percent tryptic peptide identity with decreases in protein sequence identity (Fig. 2). Nine of 19 protein pairs with 95% or more sequence identity showed above 80% conservation of peptides, but the remaining ten showed lower peptide identity, with the five lowest having identities between 52% and 60%. When sequence identity of protein pairs fell below 70% virtually no peptides of MW 700–3000 were conserved. In fact, of 14 protein pairs with less than 70% sequence identity, 12 had no peptides conserved, and the remaining two showed 8% and 11% peptide conservation. There was a correlation between percent tryptic peptide identity of protein pairs and their percent sequence identity ($R^2 = 0.66$).

COMPARISON OF AMINO ACID COMPOSITION CONSERVATION AND TRYPTIC PEPTIDE CONSERVATION

The percent conservation of tryptic peptides from protein pairs was plotted against their AA composition difference scores (Fig. 3). This revealed that over the range of 100% to 5% tryptic peptide identity, there were only small changes in protein AA compositions. All protein pairs with more than 50%

TABLE 3
Sequnce identity of E. coli protein HNSA with the same protein from four other species

HNSA_ECOLI matched against:	HNSA_SHIFL	HNSA_SALTY	HNSA_SERMA	HNSA_PROVU
% sequence identity	100	95	86	78

Protein sequences were from the SWISS-PROT database, and sequence identities were determined using the Bestfit program from GCG (Devereaux *et al*., 1984). Bestfit matching was done using default parameters.

tryptic peptide identity showed AA composition difference scores of less than ten, and all protein pairs with more than 5% tryptic peptide identity showed AA composition difference scores of less than 20. Only when 0% of peptides were conserved between protein pairs were there AA composition difference scores greater than 20. There was a correlation between percent tryptic peptide identity and AA composition difference scores for all 65 protein comparisons ($R^2 = 0.40$), and a stronger correlation observed when protein pairs with 0% peptide identity were not considered ($R^2 = 0.52$).

CROSS-SPECIES CONSERVATION OF PROTEIN MOLECULAR WEIGHT AND ISOELECTRIC POINT

The MW of proteins was highly conserved across species boundaries, with a mean absolute difference of 1.9% and standard deviation of 3.3% (Table 4;

TABLE 4
Proteins from E. coli and humans used in this study, and the mean pI and MW differences found when compared with proteins of the same function from other species

Proteins cross-matched:	Number of matches undertaken	Mean pI difference in units	Mean MW difference in percent
Human sera			
Alpha-1-antitrypsin	4	0.3	1.5
Apolipoprotein A1	5	0.1	1.2
Apolipoprotein A2	3	0.1	0.9
Apolipoprotein C2	1	0.0	1.3
Apolipoprotein H	3	0.2	5.7
Fatty acid binding protein	2	1.6	0.7
Hemoglobin alpha	5	0.3	0.2
Hemoglobin beta	5	0.0	0.3
Serum albumin	5	0.1	0.5
E. coli			
10 kD chaperonin	5	0.1	1.7
Alkyl hydroperoxide reductase C22	1	0.0	0.1
Aspartate carbamyltransferase (catalytic chain)	2	0.3	1.6
Aspartate carbamyltransferase (regulatory chain)	1	0.8	0.2
Cysteine synthase	2	0.1	2.5
DNA-binding protein HNS-A	4	0.4	0.6
Glyceraldehyde 3-phosphate dehydrogenase	5	0.5	10.6
Outer membrane protein	4	0.6	1.8
Periplasmic oligopeptide-binding precursor	1	0.0	0.8
Phosphotransferase factor III	1	0.0	0.0
Serine hydroxymethyl transferase	2	0.1	0.6
Superoxide dismutase	4	0.9	1.1
Mean		0.30	1.9
standard deviation		0.42	3.3

A maximum of 5 cross-species matches were examined per protein. All means were calculated from the absolute value of all differences.

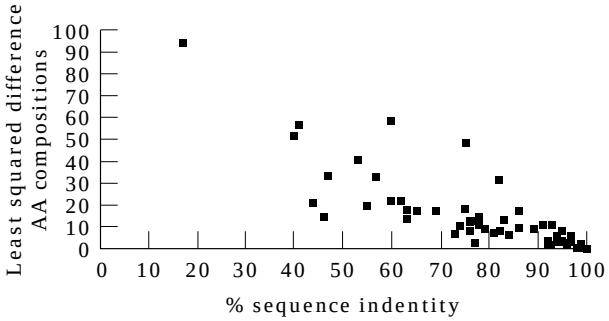

FIG. 1. Cross-species conservation of protein amino acid composition, as compared with sequence identity. Each point represents the comparison of two proteins which are functionally identical, but from different species. Note, however, that one of the proteins in each comparison is an *E. coli* or human protein as shown in Table 4. The AA composition difference score calculated from the two proteins is plotted against their percent sequence identity. Amino acid compositions considered only the 16 amino acids generated from proteins after acid hydrolysis.

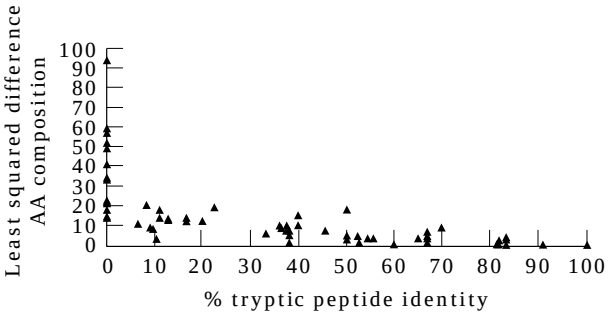

FIG. 3. Cross-species conservation of protein amino acid composition, compared with conservation of protein tryptic peptides. Each point represents the comparison of two proteins which are functionally identical, but from different species. Note, however, that one of the proteins in each comparison is an *E. coli* or human protein as shown in Table 4. The percent conservation of tryptic peptides between the two proteins is plotted against their AA composition difference score. Amino acid compositions considered only the 16 amino acids generated from proteins after acid hydrolysis, and only tryptic peptides in the range 700–3000 MW were examined.

Fig. 4). The high standard deviation was due to the MW differences between glyceraldehyde 3-phosphate dehydrogenase from *E. coli* and five other species (8.7–11.3% difference), and to a MW difference of 15.4% between apolipoprotein H from human and rat. Isoelectric point of proteins was less well conserved (Table 4; Fig. 5), with test proteins not showing consistently large or small pI differences across species boundaries. There was a mean absolute difference of 0.3 units between test proteins and others, with a standard deviation of 0.4 units. An extreme example of pI differences was seen in fatty acid binding protein, where there was a pI difference of 2.1 units between the protein from human and chicken. There were no correlations observed between pI differences and MW differences ($R^2 = 0.02$), between pI differences and percent sequence identity ($R^2 = 0.06$), pI differences and AA difference scores ($R^2 = 0.03$), pI differences and percent tryptic peptide identity ($R^2 = 0.05$), MW differences and percent sequence identity ($R^2 = 0.00$), MW differences and AA difference scores ($R^2 = 0.00$), or MW differences and percent tryptic peptide identity ($R^2 = 0.00$).

Discussion

This study has compared proteins of identical function across species boundaries. To serve as a benchmark, AA sequence identity of all cross-species

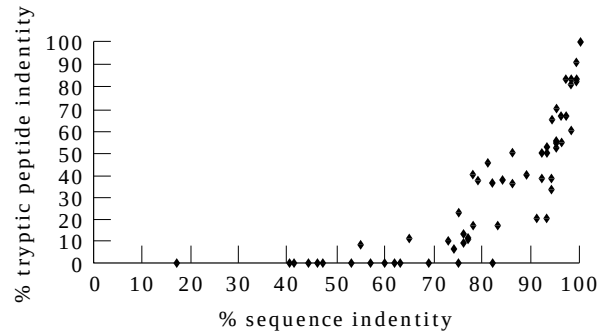

FIG. 2. Cross-species conservation of protein tryptic peptides, as compared with sequence identity. Each point represents the comparison of two proteins which are functionally identical, but from different species. Note, however, that one of the proteins in each comparison is an *E. coli* or human protein as shown in Table 4. The percent conservation of tryptic peptides between the two proteins is plotted against their percent sequence identity. Only tryptic peptides in the range 700–3000 MW were considered here.

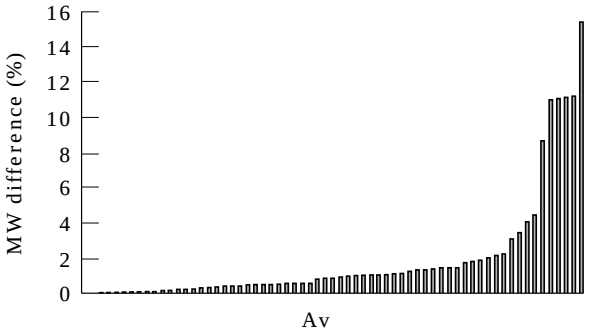

FIG. 4. Cross-species differences in protein molecular weight for 65 proteins. Each column represents the comparison of two proteins which are functionally identical, but from different species. Note, however, that one of the proteins in each comparison is an *E. coli* or human protein as shown in Table 4. The absolute value of percent differences in molecular weight are shown. The mean difference is 1.9%, with standard deviation of 3.3%. The high standard deviation is mostly due to the six highest differences.

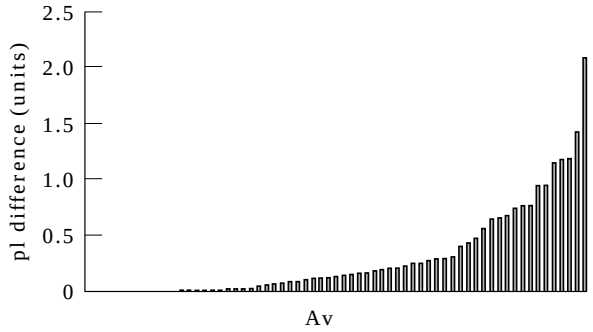

FIG. 5. Cross-species differences in protein isoelectric point for 65 proteins. Each column represents the comparison of two proteins which have the same function, but are from different species. Note, however, that one of the proteins in each comparison is an *E. coli* or human protein as shown in Table 4. The absolute value of differences in units are shown. Twelve comparisons showed no pI difference. The mean difference is 0.3 units, with a standard deviation of 0.4 units.

matched proteins was established. Proteins then had AA compositions, tryptic peptide maps, pI and MW compared. As this was undertaken to determine the usefulness of these parameters for cross-species protein identification from two-dimensional gels, AA compositions were compared using the 16 AAs that are produced after acid protein hydrolysis (Wilkins *et al*., 1996), and only peptide masses of range 700–3000 were used in tryptic peptide mapping (Pappin *et al*., 1993). No consideration was given to the error that is inherent to the techniques of AA analysis or peptide mass fingerprinting. However, the reader is referred to a recent review for discussion of the technicalities of these methods (Wilkins *et al*., 1995).

One of the key findings of this study was that differences in protein sequence resulted in small changes in AA composition, but comparatively large changes in tryptic peptides. In fact, virtually no peptides were conserved from protein to protein once sequence identity fell below 70%. This is essentially what would be expected. Single AA changes in a protein would not result in substantial alterations to the AA composition, provided the protein is of moderate size. However, of all 380 possible single AA changes that can occur, only two will not alter the mass of a peptide (Table 5). If a mass error tolerance of ± 3 Da is allowed, only 28 changes will not affect the peptide mass. Even more drastic alterations in peptide masses, and indeed peptide maps, can occur if single amino acids which define the endoproteinase cleavage sites in a protein are changed. For example if Lys or Arg residues in a protein are changed, and the protein is mapped using trypsin, sites of digestion and thus peptides will be lost. Conversely, peptides will be created if other amino acids are changed to Lys or Arg.

The cross-species differences in protein MW and pI were also investigated in this study. The MW of most proteins was well conserved across species bound-

TABLE 5

Peptide masses are altered when amino acids in any peptide are changed

	Peptide mass difference if one residue changed from:																			
to:	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
A		−85	−43	−44	−32	−58	−57	14	−66	−42	−42	−57	−60	−76	−26	−16	−30	−115	−92	−28
R	85		42	41	53	27	28	99	19	43	43	28	25	9	59	69	55	−30	−7	57
N	43	−42		−1	11	−15	−14	57	−23	1	1	−14	−17	−33	17	27	13	−72	−49	15
D	44	−41	1		12	−14	−13	58	−22	2	2	−13	−16	−32	18	28	14	−71	−48	16
C	32	−53	−11	−12		−26	−25	46	−34	−10	−10	−25	−28	−44	6	16	2	−83	−60	4
E	58	−27	15	14	26		1	72	−8	16	16	1	−2	−18	32	42	28	−57	−34	30
Q	57	−28	14	13	25	−1		71	−9	15	15	−0.2	−3	−19	31	41	27	−58	−35	29
G	−14	−99	−57	−58	−46	−72	−71		−80	−56	−56	−71	−74	−90	−40	−30	−44	−129	−106	−42
H	66	−19	23	22	34	8	9	80		24	24	9	6	−10	40	50	36	−49	−26	38
I	42	−43	−1	−2	10	−16	−15	56	−24		0	−15	−18	−34	16	26	12	−73	−50	14
L	42	−43	−1	−2	10	−16	−15	56	−24	0		−15	−18	−34	16	26	12	−73	−50	14
K	57	−28	14	13	25	−1	0.2	71	−9	15	15		−3	−19	31	41	27	−58	−35	29
M	60	−25	17	16	28	2	3	74	−6	18	18	3		−16	34	44	30	−55	−32	32
F	76	−9	33	32	44	18	19	90	10	34	34	19	16		50	60	46	−39	−16	48
P	26	−59	−17	−18	−6	−32	−31	40	−40	−16	−16	−31	−34	−50		10	−4	−89	−66	−2
S	16	−69	−27	−28	−16	−42	−41	30	−50	−26	−26	−41	−44	−60	−10		−14	−99	−76	−12
T	30	−55	−13	−14	−2	−28	−27	44	−36	−12	−12	−27	−30	−46	4	14		−85	−62	2
W	115	30	72	71	83	57	58	129	49	73	73	58	55	39	89	99	85		23	87
Y	92	7	49	48	60	34	35	106	26	50	50	35	32	16	66	76	62	−23		64
V	28	−57	−15	−16	−4	−30	−29	42	−38	−14	−14	−29	−32	−48	2	12	−2	−87	−64	

This table shows the effect of changing one amino acid in an arbitrary peptide to another amino acid. Only two amino acid changes in the possible 380 do not alter the mass of an arbitrary peptide. If a peptide mass error of ± 3 Da is tolerated, 28 changes do not alter peptide mass. Note that some amino acid changes can also result in the creation of new peptide cleavage sites, which will change the peptide map of the protein. Single amino acid code is used. Masses are in Daltons.

aries, presumably due to the presence of shared domains which classify the proteins as functionally identical. However, protein pI was frequently not well conserved. It is startling that proteins from different species known to have the same function can have isolelectric points which are greatly different.

The purpose of comparing the pI, MW, AA composition and tryptic peptide maps of proteins from different species has been to determine the usefulness of these parameters for cross-species protein identification from two-dimensional gels. The pI of unmodified proteins can be determined accurately from two-dimensional gels, and can be a useful parameter for protein identification (Bjellqvist *et al*., 1993; Hobohm *et al*., 1994; Wilkins *et al*., 1996). For example, one study showed the pI of 26 out of 27 *E. coli* proteins estimated from a two-dimensional gel to be within 0.25 units of that calculated from the SWISS-PROT database, allowing a pI search window of gel estimate ± 0.25 units to be used for matching against *E. coli* database entries (Wilkins *et al*., 1996). However, if pI is to be used in cross-species matching, a much larger window will need to be employed. Using the data set described here as a guide, a pI window of ± 1.1 units [i.e. mean + (2 × S.D.)] will account for most sequence-derived differences in protein pI. A further enlargement of windows by 0.25 units, giving a final pI window size of gel estimate ± 1.35 units should also account for pI estimation error from the two-dimensional gel. This window is large compared to that used in single species matching, but will still greatly reduce the number of proteins from the database to be considered during searches.

Protein MW determined from two-dimensional gels is often used to aid protein identification (Henzel *et al*., 1993; Hobohm *et al*., 1994; Pappin *et al*., 1993; Wilkins *et al*., 1996). However, accurate determination of protein MW from polyacrylamide gels can be difficult. For example, error margins of $\pm 20\%$ around MW estimates from an *E. coli* gel were necessary to create windows which included most database MW values (Wilkins *et al*., 1996). As the mean MW difference observed in cross-species matching was 1.9%, it is likely that a MW window of $\pm 20\%$ around estimates from the gel will be adequate to account for both error from the gel and cross-species MW differences in most cases. However, it may be beneficial to increase window size when dealing with small proteins (less than 10000 MW), where inaccurate MW estimations have a greater effect in percent terms than in larger proteins. It is pertinent to consider that MALDI-TOF mass spectrometry has recently been applied to the direct mass determination of PVDF-bound proteins from two-dimensional gels (Eckerskorn *et al*., 1992; Strupat *et al*., 1994). This allows the determination of protein masses to within 0.1% of predicted. This method of MW estimation would be immensely useful in cross-species protein identification, as it avoids the error inherent in gel-derived MW estimation. In this case, a MW search window of $\pm 8.5\%$ [i.e. mean + (2 × S.D.)] should be sufficient to account for most cross-species differences in protein MW, except when proteins are post-translationally modified.

It has been shown in this study that peptide mass fingerprinting can be useful in cross-species matching. However, there are some clear limitations in this application. The rapid decrease in peptide identity with decreases in sequence identity will make it difficult for proteins to be confidently identified cross-species when the best matching protein is of less than 80% sequence identity. Database error and polymorphisms in populations will also affect peptide-based identification more than other identification techniques. So whilst it is possible to assign an identity to proteins with three or four peptides (Henzel *et al*., 1993; Pappin *et al*., 1993), it is only when all peptide masses match those from a database entry that there is high confidence in the protein identification (Henzel *et al*., 1993; Mortz *et al*., 1994; Yates *et al*., 1993). Therefore it is likely that cross-species peptide mass fingerprinting will nearly always require corroborative evidence to give high confidence in protein identities.

Amino acid composition of proteins has been shown to be well conserved across species boundaries in the present study. As AA composition is known to be an accurate means of protein identification in single species matching (Hobohm *et al*., 1994; Garrels *et al*., 1994; Wilkins *et al*., 1996; Yan *et al*., 1996) it should thus also be a useful cross-species protein identification tool, even when protein sequence identities are below 80%. In single species matching, score patterns are known to provide a clear indication of the confidence of protein identification by considering the goodness of fit of the data as well as the uniqueness of the match (Wilkins *et al*., 1996). Specifically, correctly identified proteins always have a first ranked protein with score below 30 and a large difference between the score of the first and second ranked proteins. The cross-species matching undertaken here has not shown score patterns as seen in single species matching (data not shown), but this is because the matching was done against all species in the database, and without pI or MW windows. This approach is in fact the least powerful form of cross-species matching. When complete nucleotide

sequences of organisms like *E. coli* and *S. cerevisiae* become available, a more directed form of cross-species matching will be feasible. Protein AA composition, in conjunction with MW and pI windows as defined in the present study, should be matched cross-species against the database entries for a single, well-defined organism. For example, proteins from *Candida albicans* could be matched against *S. cerevisiae*, or proteins from *S. typhimurium* matched against *E. coli*. This approach should produce either "correct" score patterns, providing confidence in protein identities, or score patterns which suggest proteins identities should be checked by other means. We are currently using the guidelines outlined in this theoretical study to define score patterns when empirical AA composition, pI and MW data is used for cross-species protein identification (Cordwell *et al*., in prep.).

It is of interest to relate the findings of this study to those in Cordwell *et al*. (1995) and Wasinger *et al*. (1995), in which cross-species protein identification was achieved by comparing results from AA composition matching with those from tryptic peptide mass fingerprinting. The efficiency of protein identification by the two techniques reflected the predictions of the theoretical study undertaken here. Specifically, a total of 13 out of 17 proteins subjected to the AA analysis identification procedure showed the correct protein as rank no. 1, but only two of the same 17 proteins analysed by tryptic peptide mass fingerprinting showed the correct protein as rank no. 1 (Cordwell *et al*., 1995; Wasinger *et al*., 1995). The accuracy of AA identification was highlighted by the fact that AA analysis correctly identified one protein across species boundaries where the percent sequence homology of the proteins was only 43% (Cordwell *et al*., 1995). The remaining proteins were identified by comparing lists of best-matching proteins generated by both techniques, even though neither technique had ranked the correct identification as no. 1. This combined approach to cross-species matching is clearly a rapid, powerful and effective identification approach when protein identities cannot be confidently established by one technique alone.

This manuscript has theoretically investigated the use of rapid techniques for the cross-species matching of proteins from two-dimensional gels. There are some clear conclusions that can be drawn. (1) Protein pI differences across species boundaries outweigh the experimental error in pI estimation from two-dimensional gels. The use of pI windows greater than those used in single species protein identification is necessary. (2) The experimental error in MW determination from two-dimensional gels is generally far greater than MW differences observed between proteins across species boundaries. The use of MW windows similar to those in single species matching should be appropriate. (3) Cross-species protein identification by peptide mass fingerprinting will be useful only where phylogenetic distances between the species under study and those described in the database are small. Identity confidence will usually need to be increased through other techniques. (4) Protein AA composition is well conserved across species boundaries. When matched cross-species against only the database entries for a single, well-defined organism, it should offer the best rapid means of cross-species protein identification. Its use in conjunction with protein pI and MW estimations will be desirable. Nevertheless, there will be proteins whose identities require other means of confirmation.

The above conclusions should aid in the rapid identification of many proteins across species boundaries. This will assist in the detailed description of proteomes of organisms which are poorly defined at the molecular level.

MRW was the recipient of an Australian Postgraduate Research Award. KLW acknowledges funding from the Australian Research Council.

REFERENCES

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

Appel, R. D., Sanchez, J.-C., Bairoch, A., Golaz, O., Ravier, F., Pasquali, C., *et al*. (1994b). The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis. *Nucl. Acids Res*. **22,** 3581–3582.

Bairoch, A. & Boeckmann, B. (1994). The SWISS-PROT protein sequence bank: current status. *Nucl. Acids Res*. **22,** 3578–3580.

Bjellqvist, B., Hughes, G., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., *et al*. (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14,** 1023–103l.

Cordwell, S., Wilkins, M. R., Cerpa-Poljak, A., Gooley, A. A., Duncan, M., Williams, K. L. & Humphery-Smith, I. (1995). Cross-species identification of proteins separated by two-dimensional electrophoresis using MALDI-TOF and amino acid composition. *Electrophoresis* **16,** 438–443.

Cottrell, J. S. (1994). Protein identification by peptide mass fingerprinting. *Peptide Research* **7,** 115–124.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res*. **12,** 7463–7476.

Eckerskorn, C., Strupat, K., Karas, M., Hillenkamp, F. & Lottspeich, F. (1992). Mass spectrometric analysis of blotted proteins after gel electrophoretic separation by matrix-assisted laser desorption/ionization. *Electrophoresis* **13,** 664–665.

Garrels, J. I., Futcher, B., Koboyashi, R., Latter, G. I., Schwender, B., Volpe, T., *et al*. (1994). Protein identification for a *Saccharomyces cerevisiae* protein database. *Electrophoresis* **15,** 1466–1486.

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. U.S.A.* **90,** 5011–5015.

HOBOHM, U., HOUTHAEVE, T. & SANDER, C. (1994). Amino acid analysis and protein database compositional search as a rapid and inexpensive method to identify proteins. *Anal. Biochem.* **222,** 202–209.

MORTZ, E., VORM, O., MANN, M. & ROEPSTORFF, P. (1994). Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biol. Mass Spect.* **23,** 249–261.

PAPPIN, D. J. C., HOJRUP, P. & BLEASBY, A. J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biology* **3,** 327–332.

PATTERSON. D. (1994). From electrophoretically separated proteins to identification: strategies for sequence and mass analysis. *Anal. Biochem.* **221,** 1–15.

STRUPAT, K., KARAS, M., HILLENKAMP, F., ECKERSKORN, C. & LOTTSPEICH, F. (1994). Matrix-assisted laser desorption ionization mass spectrometry of proteins electroblotted after polyacrylamide gel electrophoresis. *Anal. Chem.* **66,** 464–470.

WASINGER, V. C., CORDWELL, S. J., POLJAK, A., YAN, J. X., GOOLEY, A. A., WILKINS, M. R., *et al.* (1995). Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* **16,** 1090–1094.

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. *Biotechnology and Genetic Engineering Reviews* **13,** 19–50.

WILKINS, M. R., PASQUALI, C., APPEL, R. D., OU, K., GOLAZ, O., SANCHEZ, J.-C., *et al.* (1996). From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology* **14,** 61–65.

YAN, J. X., WILKINS, M. R., OU, K., GOOLEY, A. A., WILLIAMS, K. L., SANCHEZ, J.-C., *et al.* (1996). Large scale amino acid analysis for proteome studies. *J. Chromatogr. A.* **736,** 291–302..

YATES, J. R. III, SPEICHER, S., GRIFFIN, P. R. & HUNKAPILLER, T. (1993). Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214,** 397–408.