JOURNAL OF MASS SPECTROMETRY, VOL. 32, 370–378 (1997)

Conserved Motifs as the Basis for Recognition of Homologous Proteins Across Species Boundaries Using Peptide-mass Fingerprinting

Stuart J. Cordwell,[1] Valerie C. Wasinger,[1] Anne Cerpa-Poljak,[2,3] Mark W. Duncan[2] and Ian Humphery-Smith[1]*

[1] Centre for Proteome Research and Gene-Product Mapping, National Innovation Centre, Australian Technology Park, Eveleigh, Australia, 1430

[2] Biomedical Mass Spectrometry Unit, University of New South Wales, Australia, 2052

[3] Cooperative Research Centre for Biopharmaceutical Development, 384 Victoria Street, Darlinghurst, Australia, 2010

Two-dimensional gel electrophoresis of any biological system presently resolves a plethora of highly purified proteins for which no function or identity has been determined. Theoretical and experimental data were used to demonstrate that peptide-mass fingerprinting (PMF) could aid in the recognition of conserved motifs across species boundaries, and thereby assist in attributing putative function to some of these molecules. Amino acids residue substitutions produced by biological diversity and phylogenetic distance combine to highlight regions of functional significance within proteins. Using 10 prokaryotic and two eukaryotic elongation factors (EF), up to 25 peptide fragments (>800 Da) per molecule were compared across species boundaries within a 12 × 12 contingency table (66 cross-species comparisons), based upon the degree of molecular mass and amino acid sequence identity. Total amino acid sequence identity ranged from 29.4–80.9% for these molecules. Peptide fragments with homologous sequence across three or more EF were defined as containing, or being near to, conserved functional motifs. Twelve such fragments (>800 Da) were found in this group of proteins. In addition, an 808.9 Da peptide of unknown functional significance was seen to occur in three of the 12 molecules studied and in another three EF-Tu molecules. At the 83% (five of six residues) identity level, this fragment was found in a further 35 EF-Tu molecules and in 14 unrelated proteins. Further investigation should reveal a role for this fragment (motif) in structural integrity or protein function. A FASTA search conducted on a peptide fragment containing a conserved GTP-binding motif (GHVDHGK) of EF-Tu from *Euglena gracilis* was used as an example to putatively attribute partial function to three hypothetical proteins derived from DNA sequencing initiatives. © 1997 by John Wiley & Sons, Ltd.

J. Mass Spectrom. **32**, 370–378 (1997)
No. of Figures: 2 No. of Tables: 4 No. of Refs: 78

KEYWORDS: MALDI-TOF mass spectrometry; protein characterization; elongation factor; two-dimensional gel electrophoresis; peptide-mass fingerprinting; motifs; profiles; *Spiroplasma melliferum*; *Mycoplasma gallisepticum*; *Euglena gracilis*

INTRODUCTION

Peptide-mass fingerprinting of protein samples digested by endoproteinases has emerged as a powerful tool for the initial characterization of proteins from a variety of host organisms.[1–16] The simplicity of this procedure when combined with two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry has meant that biologists with little or no training in the otherwise demanding field of mass spectrometry can now rapidly access useful information for large numbers of samples. The latter is of paramount importance when one considers the many thousands of proteins which can be efficiently purified from complex mixtures on a single two-dimensional electrophoresis gel.[17]

The rate of expansion of gene and protein databases has increased exponentially in recent years due to the advent of large-scale genome sequencing projects of organisms such as *Mycoplasma pneumoniae*, *Escherichia coli*, *Mycobacterium leprae*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Mus musculus* and the highly publicised Human Genome Project. Two bacterial genomes, *Haemophilus influenzae* and *Mycoplasma genitalium*, have recently been fully sequenced.[18,19] In the immediate future, it is unlikely that large numbers of additional sequencing initiatives will be undertaken. Thus, there is a need to rapidly extract information about other organisms so as to differentiate their unique genes and gene-products from those already identified in other organisms. The ability to cross species boundaries during protein identification can: (1) take advantage of information already acquired for homologous gene-products in other species; (2) maximize the use of limited research resources and (3) rapidly transform our understanding of organisms otherwise poorly defined at the molecular level.

In the past, immunoblotting of whole proteins or Edman degradation chemistry applied to either the N-

* Correspondence to I. Humphrey-Smith.

CCC 1076–5174/97/040370–09 $17.50

Received 23 May 1996
Accepted 12 August 1996

terminus or to internal digestion fragments, has allowed protein identification on a small scale.[20–25] Examination of the large numbers of proteins separated by two-dimensional gel electrophoresis has not previously been feasible without considerable expenditure of both time and financial resources.[26] Cordwell *et al.*[27] and Wasinger *et al.*[28] have recently proposed a novel method combining two rapid techniques for protein characterization: amino acid analysis[29–37] and peptide-mass fingerprinting (as above). These methods have been used to successfully identify proteins present within databases; however, neither method alone has been able to confidently determine the identity of a protein from heterologous species, i.e. cross-species matching, unless also accompanied by the time- and cost-intensive procedure of N-terminal micro-sequencing.[29,38] The 'combined approach' provides unequalled confidence in results obtained for heterologous species matches.[27,28].

The problem of 'uncertain' identification has not been limited to cross-species protein identification. Hobohm *et al.*[32] suggested that the correct identity of a protein was more likely to be amongst the first 20 candidates, rather than simply the candidate ranked first, when analysing by amino acid composition, while Mørtz *et al.*[5] concluded that identification by endoproteinase digestion and subsequent mass profiles needed confirmation by considering other properties of the protein. This problem can be overcome by the use of either the 'combined'[27] or the 'sequence tag'[10] approach. In the latter, identification of the protein is based upon the calculation of molecular mass of a peptide fragment and the sequencing of three to four amino acid residues from within this fragment. This protein microsequence can be acquired by either electrospray ionization or post-source decay MALDI-TOF mass spectrometry (reviewed in Refs 13–15).

It was noted previously that amino acid analysis provided more accurate identities than peptide-mass fingerprinting when cross-species matching.[27] This was thought to be due to the more conserved nature of sequence composition between species. The present study was undertaken to examine the inherent significance of similar peptide masses detected between species and their eventual utility in attributing putative function to novel genes. Elongation factors Tu and 1-α were chosen for study due to the large number of entries within the SWISS-PROT database and because the molecular biology of *E. coli* EF-Tu is well understood (reviewed in Refs 39–41). These EF bind aminoacyl-tRNA to the ribosome during protein biosynthesis. EF-Tu also contains a binding site for elongation factor Ts and is a target for several antibiotics, the most studied of which is kirromycin.[40]

EXPERIMENTAL

Theoretical tryptic digests and peptide-mass fingerprints of EF-Tu and 1-α

Twelve unrelated elongation factor (EF) proteins (EF-Tu and EF1-α) were chosen at random from within the SWISS-PROT database,[42] but were indicative of a range of total protein sequence identities. These proteins are shown in Table 1. Each EF was subjected to theoretical tryptic digestion using the PEPTIDESORT program, accessed via the GCG subdirectory (Program Manual for the Wisconsin Package, Ver. 8, 1994). Only peptides with Mr >800 Da were considered in this study, due to the low specificity of peptides with Mr <800 Da. EF molecules from several species were then compared against each other to determine the degree of fragment identity/similarity. A match was defined as any peptide, ± 6 Da, equal to the peptide under examination. The choice of ± 6 Da was arrived at not as a function of accuracy in the determination of molecular mass via mass spectrometry, but rather as a function of the MOWSE[6] software used to compare entries in protein databases, and the size of sample sets extracted during cross-species searches. Finally, all fragments were tabulated to determine their frequency in each of the EF proteins.

PILEUP and BESTFIT sequence comparisons were performed to determine identity between EF proteins at the primary level. Peptide fragment sequence was compared with other entries in the SWISS-PROT and PIR databases using the FASTA search program.[43] The PROSITE database was used to identify portions of sequence with known functional significance within peptide fragments.[44]

Theoretical MOWSE analysis of *Mycoplasma gallisepticum* EF-Tu

Peptides generated from PEPTIDESORT of *M. gallisepticum* EF-Tu were used to search the MOWSE

Table 1. EF-Tu and EF-1α proteins used in this study and their SWISS-PROT accession codes

Organism	SWISS-PROT accession code	Mr (kDa)	Residues	Number of peptides[a]	Number of peptides (>800 Da)
Cyanophora paradoxa	EFTU_CYAPA	44.6	409	47	21
Mycoplasma gallisepticum	EFTU_MYCGA	44.0	394	50	24
Bacillus subtilis	EFTU_BACSU	43.6	396	43	22
Haloarcula marismortui	EFTU_HALMA	45.6	420	30	19
Thermus aquaticus	EFTU_THEAQ	44.7	405	49	25
Euglena gracilis	EFTU_EUGGR	45.1	409	51	22
Mycobacterium tuberculosis	EFTU_MYCTU	43.6	396	45	21
Escherichia coli	EFTU_ECOLI	43.1	393	47	21
Chlamydia trachomatis	EFTU_CHLTR	43.2	393	47	19
Spirulina platensis	EFTU_SPIPL	44.8	410	47	23
Saccharomyces cerevisiae (1α)	EF1A_YEAST	50.0	458	67	19
Homo sapiens (1α)	EF12_HUMAN	50.5	463	64	20

[a] Number of peptides as derived from PEPTIDESORT analysis following theoretical digestion with trypsin.

database[6] by entering into the program: (1) peptides conserved in *M. gallisepticum* EF-Tu and other EF; and (2) peptides unique to *M. gallisepticum.* This was designed to determine the significance of fragment choice when analysing raw data from unknowns by peptide-mass fingerprinting.

Peptide-mass fingerprinting of *Spiroplasma melliferum* EF-Tu

Whole cell lysates of *S. melliferum* (strain A56) were subjected to IPG-DALT electrophoresis, as described previously.[27,45,46] Proteins were electrotransferred to ProBlott (Applied Biosystems, CA, USA) and visualized with Amido Black staining.[47] The spot previously reported as EF-Tu using the 'combined approach' was excised and treated for mass spectrometry as described previously.[27] Masses of protonated species were calculated using a Finnigan LaserMat 2000 (Hempstead, UK) MALDI-TOF system. Digests were incorporated in a 0.5 μl α-cyano-4-hydroxycinnamic acid matrix [10 mg/ml in 70% acetonitrile (MeCN)]. Internal calibration was performed against substance P (Sigma, St Louis, MO), Mr = 1348.7. Fragment masses generated were used to search the MOWSE database.[6]

RESULTS

Cross-species peptide-mass fingerprinting

The masses of theoretical tryptic digest fragments for 10 EF-Tu and two EF1-α molecules were compared between species (Table 2). All matching peptides were examined via PILEUP and/or BESTFIT to determine whether similarity (±6 Da) in detected mass was due to a 'fluke' (corresponding Mr, but lacking sequence identity) or conserved sequence. The 12 molecules studied showed amino acid sequence identity varying from 29.4–80.9% across species boundaries. The number of peptide fragments showing similarity in molecular mass between species varied between 0 and 9 Table 2). Of these, as many as seven of nine (*C. paradoxa* EF-Tu compared with *S. platensis* EF-Tu) referred to sequence similarity. The numbers shown in parentheses in Table 2 demonstrate the effective range of sequence similarity for which peptide mass can be expected to assist in protein characterization. On 23 of 66 occasions (three above 61.0%; the remainder <37.4% total sequence identity), no fragments matching by both Mr and sequence identity were detected for EF sharing between 29.4 and 68.8% total amino acid sequence identity. However, peptide fragments showing both Mr and sequence identity were found for EF with as little as 29.6% total amino acid sequence identity in 43 of 66 (65.2%) analyses. In the 60–70% amino acid sequence identity range, the technique produced consistently useful data (one to four fragments containing conserved sequence with respect to two to nine fragments showing mass similarity) for 20 of 23 comparisons.

Masses of all peptide fragments (>800 Da) generated by PEPTIDESORT from 12 EF molecules were assessed to detect the number of EF containing a particular peptide fragment. Tryptic peptides with matching Mr and conserved sequence across three or more species were interpreted as containing, or being near to, a conserved motif. Thirteen fragments were present in three or more of 12 EF proteins studied. PROSITE analysis and literature surveys[39–41,48] showed that residues within or surrounding 12 of these 13 fragments were of known functional significance (Table 3). No function had previously been attributed for an 808.9 Da peptide. FASTA sequence analysis of this six residue (PQFYVR) peptide showed six other EF-Tu molecules with 100% identity. A further 35 EF-Tu proteins and 14 unrelated proteins showed identity at the 83% (five of six residues)

Table 2. Comparisons of elongation factors across species boundaries using peptide-mass fingerprinting and FASTA sequence analysis

Organism	1[a]	2	3	4	5	6	7	8	9	10	11	12	Number of peptides (>800 Da)[b]
1. *C. paradoxa*	×	9 (4)	7 (3)	6 (1)	5 (3)	6 (5)	6 (3)	5 (3)	2 (0)	9 (7)	1 (0)	5 (0)	21
2. *M. gallisepticum*	67.2	×	8 (3)	4 (0)	4 (1)	5 (4)	5 (2)	7 (3)	4 (2)	3 (2)	3 (0)	5 (1)	24
3. *B. subtilis*	70.4	72.5	×	2 (0)	7 (4)	5 (3)	6 (5)	8 (6)	5 (1)	5 (3)	3 (1)	4 (0)	22
4. *H. marismortui*	34.0	34.8	36.4	×	4 (0)	3 (0)	2 (0)	1 (0)	2 (0)	3 (0)	1 (1)	2 (1)	19
5. *T. aquaticus*	69.0	69.1	72.6	36.6	×	5 (0)	5 (2)	6 (4)	2 (1)	5 (2)	2 (1)	2 (0)	25
6. *E. gracilis*	78.0	64.3	67.2	33.3	68.8	×	4 (1)	3 (2)	5 (2)	8 (5)	1 (0)	5 (1)	22
7. *M. tuberculosis*	63.8	68.0	71.9	37.4	71.4	62.9	×	5 (3)	5 (1)	4 (3)	5 (0)	4 (0)	21
8. *E. coli*	70.0	70.9	76.8	35.6	71.0	68.6	74.4	×	2 (0)	5 (3)	4 (1)	3 (0)	21
9. *C. trachomatis*	61.4	60.9	62.4	32.9	63.3	60.1	62.9	64.8	×	6 (1)	3 (0)	3 (1)	19
10. *S. platensis*	80.9	65.5	69.3	32.8	70.7	76.5	63.8	69.9	61.3	×	0 (0)	4 (0)	23
11. *S. cerevisiae*	33.7	32.9	33.6	49.9	37.6	33.2	35.8	33.0	29.4	32.0	×	5 (5)	19
12. *H. sapiens*	32.8	32.0	32.8	50.9	35.6	32.1	34.4	32.6	29.6	30.7	80.1	×	20

Bottom left-hand half of table refers to percent identity for protein sequence comparisons.
Top right-hand half of table refers to number of peptide fragments with shared Mr (±6 Da) for cross-species peptide-mass fingerprints (brackets refer to those fragments with amino acid sequence identity).
[a] Molecules from organisms 1–10 refer to EF-Tu and from organisms 11 and 12 to EF-1α.
[b] Peptides as derived from PEPTIDESORT analysis.

Table 3. Conserved peptide fragments of EF-Tu (and EF1-α) molecules from heterologous species

Number of matches	Fragment Mr	Residues in fragment	Function associated with peptide fragment
8 (6)	1711.9	76–90	GTP-binding, Mg^{2+}-guanine co-ordination, structural integrity (interaction with domain III)
8 (6)	1143.3	~250–260	tRNA binding, structural integrity
5 (5)	857.0	~329–334	Kirromycin resistance, structural integrity
4 (4)	2114.0	138–155	GTP-binding, EF-Ts binding
4 (4)	2716.4	91–117*	Structural integrity, tRNA binding
4 (4)	1594.8	11–25** (6–20)	GTP-binding
6 (3)	808.9	340–345	**Unknown**
5 (3)	1685.9	220–234	Aminoacyl-tRNA binding, ribosome binding, kirromycin resistance
4 (3)	851.0	118–124†	tRNA binding, ribosome binding
3 (3)	1566.8	10–24** (6–20)	GTP-binding
3 (3)	2732.0	91–117*	Structural integrity, tRNA binding
3 (3)	1580.8	11–25**	GTP-binding
3 (3)	823.0	118–124†	tRNA binding, ribosome binding

*, ** and † refer to similar peptides from the same conserved region of the protein (as determined by amino acid sequence), differing by one amino acid residue only, i.e. three sets of peptide fragments.

level. Since this region of the molecule has been conserved across phylogenetically distant organisms, our assumption is that further investigation may reveal a role in structural integrity or protein function. Similar levels of similarity were not seen for other six residue sequences selected at random from within the EF molecules, but outside the conserved regions identified in Table 3 (data not shown).

Following detection of conserved peptides across species, the amino acid sequence of the fragment has the potential to attribute putative function to novel genes and/or their gene-products. As an example of this, the peptide fragment (1594.8 Da) containing the conserved GTP-binding motif (GHVDHGK) of EF-Tu from *Euglena gracilis* was compared with other protein entries in the SWISS-PROT and PIR databases. This search revealed numerous other elongation factors: Tu, G, 1α, 2, etc. (data not shown) and apparently unrelated proteins sharing significant similarity (Table 4). It is likely that a significant proportion of these are capable of binding GTP. Furthermore, three 'hypothetical proteins' were putatively attributed partial function.

Thus, in summary, peptide homologies were first detected on the basis of molecular mass conservation. This was followed by amino acid sequence multiple alignment to define the limits of conserved sequence motifs. The information was then used to detect >80% similarity by more traditional approaches for gene-

Table 4. FASTA analysis of the conserved (10 of 12 entries in Table 3) peptide fragment containing the GTP-binding motif of EF-Tu from *Euglena gracilis* (all references to other elongation factors have been deleted)

Accession code	Protein Identity—Species	Sequence position	Sequence*	Identity
EFTU_EUGGR	Elongation factor Tu—*E. gracilis*	11–25	**KPHINIGTIGHVDHGK**	100% in 17 aa
IF2G_HUMAN	Translation initiation factor IF-2-gamma—human	102–115	QAT**INIGTIGHV**A**HGK**	92.3% in 13 aa
S46942	Su (var) 3–9 protein—*D. melanogaster*	41–54	QAT**INIGTIGHV**A**HGK**	92.3% in 13 aa
STTN_RAT	Statin S1—rat	5–21	**KTHINI**VV**IGHVDSGK**	75.0% in 16 aa
S13806	Thesaurin A—African clawed frog	5–21	**KIHINI**VV**IGHVDSGK**	75.0% in 16 aa
GST1_HUMAN	G1 to S phase transition protein—human	72–88	**KEHVNV**VF**IGHVDAGK**	62.5% in 16 aa
YO81_CAEEL	**Hypothetical 72.3 kDa protein 2K1236.1—*C. elegans***	44–56	DKIR**NFGI**VA**HVDHGK**	66.7% in 12 aa
SELB_ECOLI	SelB translation factor—*E. coli*	3–14	...MI**IATAGHVDHGK**	81.8% in 11 aa
S40816	**Hypothetical protein O591—*E. coli***	7–19	EKLR**NIA**II**AHVDHGK**	75.0% in 12 aa
TETM_UREUR	Tetracycline resistance protein—*U. urealyticum*	4–17	MKI**INIG**VLA**HVDAGK**	69.2% in 13 aa
SUP2_PICPI	Omnipotent suppressor protein 2—*P. pinus*	317–333	**KDH**MSII**FMGHVDAGK**	56.2% in 16 aa
S46918	GTP-binding protein—*M. capricolum*	8–20	SKIR**NFSIIAHIDHGK**	58.2% in 12 aa
LEPA_BACSU	GTP-binding protein LepA homologue—*B. subtilis*	16–28	SRIR**NFSIIAHIDHGK**	58.3% in 12 aa
S44254	Alpha-galactosidase—*P. pentosaceus*	35–45	LS**H**LYF**G**..**GHVDH**YH	58.3% in 12 aa
S50374	**Hypothetical protein L8003.7—yeast**	48–60	ENYR**N**FSIVA**HVDHGK**	58.3% in 12 aa
OPAG_NEIGO	Opacity protein OPA52—*N. gonorrhoeae*	139–155	**KPY**IGVRVG**GHVRHG**I	50.0% in 16 aa
VMAT_P14HB	Matrix protein—human parainfluenza 4B virus	42–56	VKQ**IRIRTLGHADH**SN	58.3% in 12 aa
CH13_BRAJA	10 kDa chaperonin (GroES)—*B. japonicum*	41–54	GEV**I**AV**GPGGH**DD**SGK**	53.8% in 13 aa
YN21_CAEEL	Putative ATP-dependent RNA helicase—*C. elegans*	164–179	R**PHI**IVA**T**GRL**VDH**LE	53.3% in 15 aa
S08102	Serine proteinase inhibitor—rat	19–27	DGTLGRD**TLSHEDHGK**	66.7% in 9 aa
S18572	OtrA protein—*S. rimosus*	4–17	MNK**LNLGI**LA**HVDAGK**	53.8% in 13 aa
R5RTLA	Ribosomal protein L27a—rat	14–20	RLRKTRKLR**GHVSHG**H	83.3% in 6 aa

* Conserved amino acid residues appear in bold.
NB The GTP-binding motif is underlined.

```
SCAN of /work6/mowse/owl using reagent Trypsin
USING fragment mws of:
          2732
          2108
          1711
          1594
          1150
          857
No. of hits = 30 (MAX. ALLOWED)
No. of database entries scanned = 72018
1  . EFTU_MYCGA   ELONGATION FACTOR TU (EF-TU). - MYCOPLASMA GALLISEPTICUM.
2  . VIEN_NPVAC   IMMEDIATE-EARLY REGULATORY PROTEIN IE-N. - AUTOGRAPHA CALIFOR
3  . S16893       IS901 protein - Mycobacterium avium
4  . EFTU_MYCHO   ELONGATION FACTOR TU (EF-TU). - MYCOPLASMA HOMINIS.
5  . VE2_RHPV1    E2 PROTEIN. - RHESUS PAPILLOMAVIRUS TYPE 1 (RHPV 1).
6  . TOBTUFB      TOBTUFB LOCUS TOBTUFB 1436 bp ss-mRNA PLN 12-JAN-1993 - Nicot
7  . RATUGT       RATUGT truncated UDP-glucuronosyltransferase; (EC 2.4.1.17) -
8  . ASSY_YEAST   ARGININOSUCCINATE SYNTHASE (EC 6.3.4.5) (CITRULLINE--ASPARTAT
9  . YSCPKIN      YSCPKIN LOCUS YSCPKIN 1836 bp ds-DNA Circular PLN 19-MAR-1993
10 . EFTU_SALTY   ELONGATION FACTOR TU (EF-TU). - SALMONELLA TYPHIMURIUM.
11 . SPEB_STRPY   EXOTOXIN TYPE B PRECURSOR (SPE B). - STREPTOCOCCUS PYOGENES.
12 . EFTU_ECOLI   ELONGATION FACTOR TU (EF-TU). - ESCHERICHIA COLI.
13 . EFTU_CYAPA   ELONGATION FACTOR TU (EF-TU). - CYANOPHORA PARADOXA.
14 . HA12_MOUSE   H-2 CLASS I HISTOCOMPATIBILITY ANTIGEN, D-D ALPHA CHAIN PRECU
15 . ECOGARA3     ECOGARA LOCUS ECOGARA 4649 bp ds-DNA BCT 13-JUL-1993 - Escher
16 . ALFR_DROME   ALDOLASE-RELATED PROTEIN. - DROSOPHILA MELANOGASTER (FRUIT FL
17 . MDSTPKN      MDSTPKN LOCUS MDSTPKN 1696 bp RNA PLN 21-MAY-1993 - Malus dom
18 . COMB_BACSU   B COMPETENCE PROTEIN. - BACILLUS SUBTILIS.
19 . B42214       peroxisome proliferator-activated receptor beta, xPPARbeta=nu
20 . RFAK_SALTY   LIPOPOLYSACCHARIDE 1,2-N-ACETYLGLUCOSAMINETRANSFERASE (EC 2.4
21 . PAPPH633     PAPPH63 putative E2 ORF - Human papillomavirus type 63
22 . VS48_TBRVC   SATELLITE RNA 48 KD PROTEIN. - TOMATO BLACK RING VIRUS (STRAI
23 . CC14_YEAST   PROBABLE PROTEIN-TYROSINE PHOSPHATASE CDC14 (EC 3.1.3.48). -
24 . RCNIF2       RCNIF LOCUS RCNIF 3570 bp DNA BCT 13-MAY-1993 - Rhodobacter c
25 . KEEK_RAT     TYROSINE KINASE EEK RECEPTOR (EC 2.7.1.112) (FRAGMENT). - RAT
26 . 1C03_HUMAN   HLA CLASS I HISTOCOMPATIBILITY ANTIGEN, CW-2 CW*0201 ALPHA CH
27 . S64323       S64323 PGE2 receptor EP3 isoform; For the protein sequence (N
28 . ODB2_PSEPU   LIPOAMIDE ACYLTRANSFERASE COMPONENT (E2) PRECURSOR OF BRANCHE
29 . HIV1ZAM20    HIV1ZAM20 LOCUS HIV1ZAM20 1488 bp ss-RNA VRL 06-OCT-1992 - Hu
30 . REPC_AGRRA   POSSIBLE REPLICATION PROTEIN C. - AGROBACTERIUM RHIZOGENES.
```

Figure 1. MOWSE database search using fragments generated by PEPTIDESORT analysis of *M. gallisepticum* EF-Tu. Conserved fragments were used and cross-species matches appear in bold.

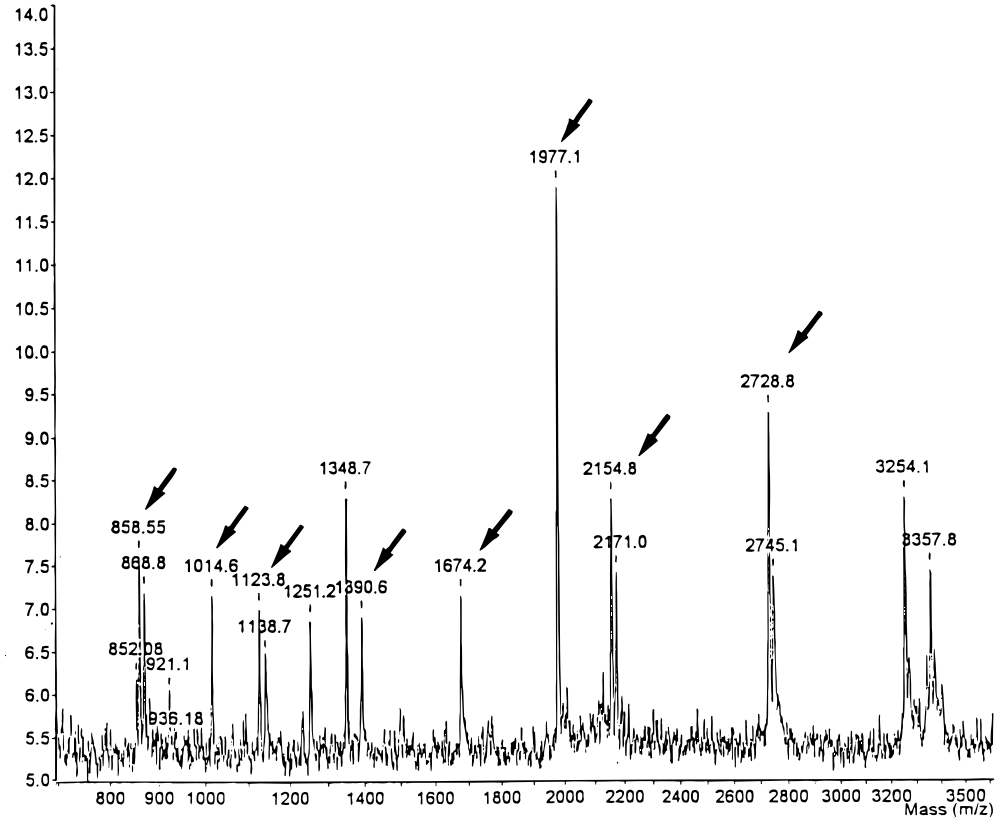

Figure 2. Peptide-mass fingerprint of *S. melliferum* EF-Tu showing peaks used for MOWSE database search.


```
USING fragment mws of:
          2728
          2154
          1977
          1674
          1390
          1123
          1014
          858
No. of hits = 30 (MAX. ALLOWED)
No. of database entries scanned = 72018
1  . STRMLTODX    STRMLTODX homologous to the maltose-binding protein of E. coli
2  . K1CS_HUMAN   KERATIN, TYPE I CYTOSKELETAL 19 (CYTOKERATIN 19) (K19). - HOMO
3  . GBS2_DROME   GUANINE NUCLEOTIDE-BINDING PROTEIN G(S)-S, ALPHA SUBUNIT (ADEN
4  . GBS1_DROME   GUANINE NUCLEOTIDE-BINDING PROTEIN G(S)-L, ALPHA SUBUNIT (ADEN
5  . DROCOL       DROCOL LOCUS DROCOL 1491 bp ds-DNA INV 23-JUN-1992 - Drosophil
6  . BSGENR84     BSGENR LOCUS BSGENR 97015 bp DNA BCT 02-NOV-1993 - Bacillus su
7  . CBG_HUMAN    CORTICOSTEROID-BINDING GLOBULIN PRECURSOR (CBG) (TRANSCORTIN).
8  . S28077       lipA protein - Neisseria meningitidis
9  . CBL_MLVCN    TRANSFORMING PROTEIN CBL. - CAS-NS-1 MURINE LEUKEMIA VIRUS.
10 . EFTU_THEMA   ELONGATION FACTOR TU (EF-TU) (FRAGMENT). - THERMOTOGA MARITIMA
11 . YEFD_ECOLI   HYPOTHETICAL 44.9 KD PROTEIN IN CPSB 5'REGION (ORF 2.4). - ESC
12 . HFLO_ANTMA   FLORICAULA PROTEIN. - ANTIRRHINUM MAJUS (GARDEN SNAPDRAGON).
13 . YO11_MOUSE   HYPOTHETICAL PROTEIN ORF-1137. - MUS MUSCULUS (MOUSE).
14 . PUR2_BACSU   PHOSPHORIBOSYLAMINE--GLYCINE LIGASE (EC 6.3.4.13) (GARS) (GLYC
15 . PGK_TRIVI    PHOSPHOGLYCERATE KINASE (EC 2.7.2.3). - TRICHODERMA VIRIDE.
16 . TN51         TN5 A 'T' at position 1442 causes an ochre stop codon which -
17 . VA23_VACCC   PROTEIN A23. - VACCINIA VIRUS (STRAIN COPENHAGEN).
18 . EFTU_DEISP   ELONGATION FACTOR TU (EF-TU). - DEINONEMA SP.
19 . AATM_HORSE   ASPARTATE AMINOTRANSFERASE, MITOCHONDRIAL (EC 2.6.1.1) (TRANSA
20 . TERA_PSESP   TERPREDOXIN REDUCTASE (EC 1.18.1.-). - PSEUDOMONAS SP.
21 . STAD_RICCO   ACYL-[ACYL-CARRIER PROTEIN] DESATURASE PRECURSOR (EC 1.14.99.6
22 . P53_SALIR    CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSING PROTEIN). - SALM
23 . S28422       1,4-alpha-glucan branching enzyme (EC 2.4.1.18) - cassava (fra
24 . PGK1_RHINI   PHOSPHOGLYCERATE KINASE 1 (EC 2.7.2.3). - RHIZOPUS NIVEUS.
25 . NGFR_CHICK   LOW-AFFINITY NERVE GROWTH FACTOR RECEPTOR PRECURSOR (NGF RECEP
26 . TAU_RAT      MICROTUBULE-ASSOCIATED PROTEIN TAU. - RATTUS NORVEGICUS (RAT).
27 . HRV1APG      HRV1APG LOCUS HRV1APG 1205 bp ss-RNA VRL 11-OCT-1991 - Human r
28 . CPITOXIAB    CPITOXIAB LOCUS CPITOXIAB 5747 bp DNA BCT 10-NOV-1993 - Clostr
29 . YIN2_STRAM   HYPOTHETICAL 44.6 KD PROTEIN IN INSTABLE DNA LOCUS (ORF 2). -
30 . SEN2_YEAST   TRNA-SPLICING ENDONUCLEASE BETA-SUBUNIT. - SACCHAROMYCES CEREV
```

Figure 3. MOWSE database analysis of *S. melliferum* EF-Tu. The closest correct match is highlighted in bold type. Peptide masses were derived from Fig. 2.

product entries in databases, e.g. FASTA.[43] Furthermore, the present approach based upon molecular mass conservation has the potential to achieve near to ideal gap optimization and has the advantage of treating the entire protein rather than small blocks of sequence in association with variable *n*-gram lengths.

Theoretical MOWSE analysis of *M. gallisepticum* EF-Tu

The MOWSE database was searched twice using two separate sets of input data. Firstly, six fragments were entered into the program, including those that were conserved in *M. gallisepticum* and the majority of other EF molecules studied. The results (Fig. 1) showed *M. gallisepticum* as the top match from the search. Cross-species matches appear ranked at 4, 10, 12 and 13. When six non-conserved fragments were used, no cross-species matches appeared (*M. gallisepticum* remained ranked first). When all 12 fragments were searched, EF-Tu molecules were successfully ranked at positions 1, 2, 3, 4 and 24 (data not shown).

Peptide-mass fingerprinting of *S. melliferum* EF-Tu

As an example of data acquired experimentally, the mass profile obtained for EF-Tu from an organism poorly defined at the molecular level (*S. melliferum*) is shown in Fig. 2. The MOWSE database was searched using the following peaks (Da): 858, 1014, 1123, 1390, 1674, 1977, 2154 and 2728. Results from the MOWSE program are shown in Fig. 3. The nearest cross-species match for an EF-Tu molecule is ranked 10 (from *Thermatoga* species.). The match was successful due to the presence of a single 2728 Da fragment, consistent with residues 76–90 in most EF proteins (see Table 3). The other matching peptides were 1674 and 1977 Da, two peptides that do not appear to be conserved in a majority of EF proteins and may therefore fall within the category of 'fluke' matches. No other conserved peptides could be seen in the mass profile of *S. melliferum* EF-Tu. These results may also reflect experimental errors due to non-specific tryptic cleavage, incomplete digestion and the unpredictability of cleavages such as Lys/Arg-Pro. Matching based on the specificity of just one or two peptides is reflected in the low ranking attained by the nearest peptide-mass homologue for *S. melliferum* EF-Tu in Fig. 3.

DISCUSSION

Many authors (see above) have now attested to the utility of peptide-mass fingerprinting, yet little thought has been given to the significance of data referring to protein identities detected across species boundaries. It is noteworthy that similar approaches[10,49,50] have yet to be applied with confidence across species boundaries

due to inherent amino acid substitutions and/or post-translational modifications. In single species database searches, any and all fragments can be seen to produce homologous data sets. The same process of matching peptide masses applies when detecting identities across species boundaries (heterologous species matching). However, these identities may take on far greater significance if they refer to conserved and functionally important motifs, even if this function has yet to be determined. The ability to identify such motifs between phylogenetically distant organisms may shed light upon molecules for which function has yet to be attributed and may further influence experimental design to confirm these initial assumptions, for example, enzyme assays and/or mutagenesis within specific regions of open reading frames (ORFs) detected by sequencing initiatives. As a complete sequence becomes available for several species[51] this need will be even further accentuated. Such means for problem-solving in biology are destined to become increasingly important research tools. For example, of the approximate 6000 ORFs detected within the genome of *Saccharomyces cerevisiae*, some 2000 lack an attributable function.[52] Therefore, determining function for the remainder will constitute a daunting task.[53]

The results presented for peptide fragments were taken from proteins covering a significant range (29.4–80.9%) of total amino acid sequence identity and some 66 cross-species analyses. Data obtained by both experimental and theoretical endoproteinase digestion were found to be equally amenable to manipulation. Although only members of one well-studied family of molecules were examined here (elongation factors), the range of total protein identities investigated would suggest that results can be interpreted as being indicative of those expected generally within protein databases. Other molecules selected for detailed study (results not shown) were unable to equal such diversity in sequence identity. The results obtained for molecules showing $>60\%$ total amino acid sequence identity demonstrated that peptide-mass fingerprinting works most efficiently for proteins sharing significant similarity. This is because the peptide fragments upon which protein identification is based (a small portion of the total number of fragments produced following endoproteinase digestion) refer to the fragments containing, or being near to, conserved motifs present across species boundaries. The latter effectively provide a 'profile' (as defined by Bairoch[44]) between proteins sharing significant similarity. Reliable results were observed at levels as low as 29.6% of the total amino acid sequence identity. It is unlikely that peptide-mass fingerprinting could function much below this level.

This can be compared to the work of Johnson *et al.*[54] who demonstrated functionally homologous proteins, as confirmed by X-ray crystallographic studies, can share as little as 9% of the total amino acid sequence identity, while others have speculated on the importance of domain shuffling and the role of introns in the evolution of proteins and protein function.[55–60] Yet it will remain extremely difficult to detect low identity proteins without three-dimensional molecular comparisons. Nonetheless, a number of strategies have been developed to home in on conserved portions when searching both gene and protein databases.[44,61–64] The present approach can now be numbered amongst these. An example used by Johnson *et al.*[54] further reinforces this finding, namely, a comparison of bovine thrombin (385 amino acid residues) and *Streptomyces* protease (299 amino acid residues) revealed just 20.9% of the amino acid sequence identity. On only three occasions did as many as three amino acid residues align (data not shown), thus effectively precluding detection of conserved sequence or functional motifs in the absence of accurate predictions or information relating to tertiary structure.

Some information is available regarding the 808.9 Da peptide fragment (in Table 3) classified as unknown. As shown in the Results section, the residues in positions 340–351 in *T. aquaticus* EF-Tu (residues 328–339 in *E. coli* EF-Tu) are highly conserved in EF-Tu molecules from several species. They are found in domain III[48] of the protein and little is known of their significance with respect to structural integrity or function. These residues form a loop structure that passes into the interface between domains I and III. The conservation of residues 340–351 suggests an important role for the motif with respect to protein function or structural integrity. Hypothesised activities of EF-Tu for which no correlation to structure have been made include: binding to the inner membrane;[65] initiation of viral RNA synthesis;[66] and the formation of EF-Tu aggregates.[67] Substitutions at positions R333 and R334 (in *E. coli* EF-Tu) lead to resistance to the antibiotic pulvomycin, which in wild-type *E. coli* prevents the binding of aminoacyl-tRNA to the ribosome.[68] Arginine residues are positively charged and may allow the formation of salt bridges with the tRNA ribose-phosphate backbone. Therefore these residues may be part of the aa-tRNA binding site. A mutation at position 329 (*E. coli* EF-Tu) increases resistance to kirromycin, an antibiotic which prevents EF-Tu release from the ribosome after GTP hydrolysis. Kirromycin is known to bind in the domain I–III interface, and also reduces the affinity for aa-tRNA.[69] It is therefore likely that this conserved motif is involved in the binding of aa-tRNA; however, further work is needed to clarify this question. More importantly, however, the present study has been able to demonstrate that even for a molecule well studied at the level of tertiary structure, regions of potential significance can be highlighted by an approach based on peptide-mass fingerprinting.

Furthermore, this study has shown that peptide-mass fingerprinting of molecularly undefined material relies heavily upon the presence of conserved peptide fragments. Once the protein homolog has been obtained, further information with regards to function may be found from within and surrounding the region of sequence similarity. When this conservation was maintained in three or more species, there was good evidence to suggest that the conserved peptides contained or were near to, functionally significant motifs. Probability theory also suggests that evolutionary pressures would not conserve regions of protein molecules in a purely random manner.

Factors which limit the efficiency of PMF include: (1) the size of potentially conserved peptides across species boundaries (larger fragments will become less likely to

be conserved through evolutionary time, while smaller fragments will go undetected because of their low specificity); (2) the overall frequency of the amino acid(s) targeted by endoproteinases; (3) the incidence of a particular peptide mass within databases; (4) the molecular mass of the intact protein, i.e. large fragments are perhaps less likely to occur and be conserved in proteins with low molecular mass and (5) the cut site with respect to the amount of the conserved motif left intact. The latter phenomenon reinforces the utility of digesting proteins with one or more endoproteinases. To date, we have shown the usefulness of this approach for proteins from *M. pneumoniae* where both trypsin and Glu-C (*Staphylococcus aureus* V8 protease) were used routinely for PMF (manuscript in preparation).

If protein molecules are to maintain their functional integrity through evolutionary time, then genetic drift and mutational forces must be restricted to regions of the molecule other than functional motifs. Thus, other regions of molecules are more likely to vary considerably with respect to phylogenetic distance and thereby show specificity for a given organism. Due to this phenomenon, peptide-mass fingerprinting can successfully aid in the identification of homologous proteins across species boundaries. However, as shown previously,[27] amino acid residue substitutions across species boundaries, as a result of phylogenetic distance, reduces the efficiency of this approach. For this same reason, a lower mass stringency (± 6 Da) must be employed during database searches across species boundaries than for searches of homologous species databases. Depending upon the size of the peptide fragment, this error window in matching can only account for a small percentage of amino acid substitutions and some experimental error. The larger the fragment, the more substitutions that can be accommodated by the mass window. In the final analysis, however, it will be amino acid sequence similarity which determines the relevance of peptide mass identity.

In conclusion, we have demonstrated that peptide-mass fingerprinting can be used to identify conserved peptide masses across species boundaries. Sequences corresponding to these masses and adjacent regions (upstream/downstream) of protein molecules could then be compared to identify conserved amino acid sequence from within existing databases. Once this sequence had been conserved across three or more species, our working hypothesis was that a conserved motif, possibly related to function, had been detected. Database searches conducted on a sequence corresponding to one or more conserved motifs can then provide a far more effective tool for screening protein databases. PMF has the advantage of being able to detect several such motifs simultaneously ('profile') in closely related molecules. At less than 30% total sequence identity between proteins, it was shown to be unlikely that anything less than tertiary structure or accurate predictions thereof will be able to achieve similar results. Apart from attributing putative function to hypothetical proteins detected during genome sequencing,[70] this approach has the potential to identify regions within molecules for mutagenesis studies, as opposed to the more crude strategy of total gene knock-out to confirm function. The utility of this approach has been further demonstrated[71] to attribute putative gene-function otherwise overlooked by Fraser *et al.*[19] when undertaking total genomic analysis of *M. genitalium.* Peptide-mass fingerprinting has therefore been shown to be more than simply an analytical tool for use in protein characterization.

Since submission on 23 May 1996, a further four genomes have been fully sequenced and a large number of sequencing initiatives commenced world-wide.[72–76] In addition, the tertiary structure of the previously unknown 808.9 Da peptide motif discovered here has been shown to be most likely responsible for maintaining the structural integrity between the three domains of the Elongation Factor Tu molecule from *Thermus aquaticus.*[77]

Acknowledgements

The authors wish to thank Carolyn Bucholtz and Alex Reisner from the Australian National Genome Information Service and Peter Jungblut and Bridget Mabutt for their valuable input. Scott Patterson, Matthias Mann and Darryl Pappin are gratefully acknowledged for discussions and their critical appraisal of this manuscript. SJC is the recipient of an Australian Postgraduate Award. This work was supported in part by funding from the University of Sydney, GlaxoWellcome Australia, Australian Research Council and the Australian Government's Cooperative Research Centre Programme.

REFERENCES

1. D. F. Hunt, H. Michel, T. A. Dickinson, J. Shabanowitz and A. L. Cox, *Science* **256**, 1817 (1992).
2. K. K. Mock, C. W. Sutton and J. S. Cottrell, *Rapid Commun. Mass Spectrom.* **6**, 233 (1992).
3. B. Spengler, D. Kirsch, R. Kaufmann and E. Jaeger, *Rapid Commun. Mass Spectrom.* **6**, 105 (1992).
4. J. R. Yates III, S. Speicher, P. R. Griffin and T. Hunkapiller, *Anal. Biochem.* **214**, 397 (1993).
5. E. Mørtz, O. Vorm, M. Mann and P. Roepstorff, *Biol. Mass. Spectrom.* **23**, 249 (1994).
6. D. J. C. Pappin, P. Højrup and A. J. Bleasby, *Curr. Biol.* **3**, 327 (1993).
7. M. Mann, P. Højrup and P. Roepstorff, *Biol. Mass. Spectrom.* **22**, 338 (1993).
8. W. J. Henzel, T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley and C. Watanabe, *Proc. Natl Acad. Sci. USA* **90**, 5011 (1993).
9. M. Barlet-Jones, W. A. Jeffery, H. F. Hansen and D. J. C. Pappin, *Rapid. Commun. Mass Spectrom.* **8**, 737 (1994).
10. M. Mann and M. Wilm, *Anal. Chem.* **66**, 4390 (1994).
11. M. M. Vestling and C. Fenselau, *Anal. Chem.* **66**, 471 (1994).
12. M. M. Vestling and C. Fenselau, *Biochem. Appl. Mass Spectrom.* **22**, 547 (1994).
13. S. D. Patterson, *Anal. Biochem.* **221**, 1 (1994).
14. M. Mann and M. Wilm, *TIBS* **20**, 219 (1995).
15. S. D. Patterson and R. Aebersold, *Electrophoresis* **16**, 1791 (1995).
16. C. W. Sutton, K. S. Pemberton, J. S. Cottrell, J. M. Corbett, C. H. Wheeler, M. J. Dunn and D. J. C. Pappin, *Electrophoresis* **16**, 308 (1995).
17. J. Klose and U. Kobalz, *Electrophoresis* **16**, 1034 (1995).
18. R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W.

FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith and J. C. Venter, *Science* **269**, 496 (1995).
19. C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison III and J. C. Venter, *Science* **270**, 397 (1995).
20. P. Edman, *Acta. Chem. Scand.* **4**, 283 (1950).
21. R. M. Hewick, M. W. Hunkapiller, L. E. Hood and W. J. Dreyer, *J. Biol. Chem.* **256**, 7990 (1981).
22. J. Vandekerckhove, G. Bauw, M. Puype, J. Van Damme and M. Van Montagu, *Eur. J. Biochem.* **152**, 9 (1985).
23. P. Matsudaira, *J. Biol. Chem.* **262**, 10035 (1987).
24. B. Bauw, J. Van Damme, M. Puype, J. Vandkerckhove, B. Gesser, G. P. Ratz, J. B. Lauridsen and J. E. Celis, *Proc. Natl Acad. Sci. USA* **86**, 7701 (1989).
25. J. Rosenfeld, J. Capdevielle, J. C. Guillemot and P. Ferrara, *Anal. Biochem.* **203**, 173 (1992).
26. J. E. Celis, H. H. Rasmussen, E. Olsen, P. Madsen, H. Leffers, B. Honore, K. Dejgaard, P. Gromov, H. Vorum, A. Vassilev, Y. Baskin, X. Liu, A. Celis, B. Basse, J. B. Lauridsen, G. P. Ratz, A. A. Andersen, E. Walbum, I. Kjaergaard, I. Andersen, M. Puype, J. Van Damme and J. Vandekerckhove, *Electrophoresis* **15**, 1349 (1994).
27. S. J. Cordwell, M. R. Wilkins, A. Cerpa-Poljak, A. A. Gooley, M. Duncan, K. L. Williams and I. Humphrey-Smith, *Electrophoresis* **16**, 438 (1995).
28. V. C. Wasinger, S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams and I. Humphery-Smith, *Electrophoresis* **16**, 1090 (1995).
29. C. Eckerskorn, P. Jungblut, W. Mewes, J. Klose and F. Lottspeich, *Electrophoresis* **9**, 830 (1988).
30. P. Jungblut, M. Dzionara, J. Klose and B. Wittman-Leibold, *J. Prot. Chem.* **11**, 603 (1992).
31. G. Shaw, *Proc. Natl Acad. Sci. USA* **90**, 5138 (1993).
32. U. Hobohm, T. Houthaeve and C. Sander, *Anal. Biochem.* **222**, 202 (1994).
33. P. Jungblut, A. Otto, E. Zeindl-Eberhart, K.-P. Pleissner, M. Knecht, V. Regitz-Zagrosek, E. Fleck and B. Wittmann-Liebold, *Electrophoresis* **15**, 685 (1994).
34. J. R. Frey, L. Kuhn, J. R. Kettman and I. Lefkovits, *Mol. Immunol.* **31**, 1219 (1994).
35. J. I. Garrels, B. Futcher, R. Kobayashi, G. I. Latter, B. Schwender, T. Volpe, J. R. Warner and C. S. McLaughlin, *Electrophoresis* **15**, 1466 (1994).
36. A. Galat, F. Bouet and S. Rivière, *Electrophoresis* **16**, 1095 (1995).
37. M. R. Wilkins, J.-C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser and K. L. Williams, *Biotechnol. Genet. Engng Rev.* **13**, 19 (1996).
38. H. H. Rasmussen, E. Mørtz, M. Mann, P. Roepstorff and J. E. Celis, *Electrophoresis* **15**, 406 (1994).
39. B. F. C. Clark, M. Kjeldgaard, T. F. M. LaCour, S. Thirup and J. Nyborg, *Biochim. Biophys. Acta* **1050**, 203 (1990).
40. A. Weijland, K. Harmark, R. H. Cool, P. H. Anborgh and A. Parmeggiani, *Mol. Micro.* **6**, 683 (1992).
41. A. Weijland and A. Parmeggiani, *Science* **259**, 1311 (1993).
42. A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **20**, 2019 (1992).
43. W. R. Pearson and D. J. Lipman, *Proc. Natl Acad. Sci. USA* **85**, 2444 (1988).
44. A. Bairoch, *Nucleic Acids Res.* **21**, 3097 (1993).
45. B. Bjellqvist, J.-C. Sanchez, C. Pasquali, F. Ravier, N. Paquet, S. Frutiger, G. J. Hughes and D. Hochstrasser, *Electrophoresis* **14**, 1375 (1993).
46. B. Bjellqvist, C. Pasquali, F. Ravier, J.-C. Sanchez and D. F. Hochstrasser, *Electrophoresis* **14**, 1357 (1993).
47. J.-C. Sanchez, F. Ravier, C. Pasquali, S. Frutiger, N. Paquet, B. Bjellqvist, D. Hochstrasser and G. J. Hughes, *Electrophoresis* **13**, 715 (1992).
48. H. Berchtold, L. Reshetnikova, C. O. A. Reiser, N. K. Schirmer, M. Sprinzl and R. Hilgenfeld, *Nature* **365**, 126 (1993).
49. J. K. Eng, A. L. McCormack and J. R. Yates III, *J. Am. Soc. Mass Spectrom.* **5**, 976 (1994).
50. J. R. Yates III, J. K. Eng, A. L. McCormack and D. Schieltz, *Anal. Chem.* **67**, 1426 (1995).
51. R. Nowak, *Science* **268**, 1273 (1995).
52. D. Butler, *Nature* **380**, 660 (1996).
53. N. Williams, *Science* **268**, 1560 (1995).
54. M. S. Johnson, J. P. Overington and T. L. Blundell, *J. Mol. Biol.* **231**, 735 (1993).
55. P. Bork and R. F. Doolittle, *Proc. Natl Acad. Sci. USA* **89**, 8990 (1992).
56. M. W. Smith, D.-A. Feng and R. F. Doolittle, *TIBS* **17**, 489 (1992).
57. R. F. Doolittle and P. Bork, *Sci. Am.* Oct, 50 (1993).
58. W. Gilbert and M. Glynias, *Gene* **135**, 137 (1993).
59. J. S. Mattick, *Curr. Opin. Genet. Dev.* **4**, 823 (1994).
60. A. Stoltzfus, D. F. Spencer, M. Zuker, J. M. Lodgson Jr. and W. F. Doolittle, *Science* **265**, 202 (1994).
61. S. Henikoff and J. G. Henikoff, *Nucleic Acids Res.* **19**, 6565 (1991).
62. T. K. Attwood, M. E. Beck, A. J. Bleasby and D. J. Parry-Smith, *Nucleic Acids Res.* **22**, 3590 (1994).
63. E. L. Sonnhammer and D. Kahn, *Protein Sci.* **3**, 482 (1994).
64. R. L. Tatusov, S. F. Altschul and E. V. Koonin, *Proc. Natl Acad. Sci. USA* **91**, 12091 (1994).
65. A. F. M. Cremers, L. Bosch, J. E. Mellema and A. P. Sam, *J. Mol. Biol.* **153**, 477 (1981).
66. T. Blumenthal, T. A. Landers and K. Weber, *Proc. Natl Acad. Sci. USA* **77**, 866 (1972).
67. J. Weiser, K. Mikulík, Z. Zizka, J. Stastná, I. Janda and A. Jiránová, *Eur. J. Biochem.* **129**, 127 (1982).
68. L. A. H. Zeef, L. Bosch, P. H. Anborgh, R. Cetin, A. Parmeggiani and R. Hilgenfeld, *EMBO J.* **13**, 5113 (1994).
69. J. P. Abrahams, M. Van Raaij, G. Ott, B. Kraal and L. Bosch, *Biochemistry* **30**, 6705 (1991).
70. E. V. Koonin, R. L. Tatusov and K. E. Rudd, *Proc. Natl Acad. Sci. USA* **92**, 11921 (1995).
71. S. J. Cordwell, D. J. Basseal, J. D. Pollack and I. Humphery-Smith, *Gene* (1997) (in press).
72. C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J-F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, K. M. Roberts, M. A. Hurst, B. P. Kaine, M. Borodovsky, H-P. Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, J. C. Venter, *Science* **273**, 1058 (1996).
73. T. Kaneko, S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, S. Tabata, *DNA Res.* **3**, 109 (1996).
74. T. Kaneko, S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, S. Tabata, *DNA Res.* **3**, 185 (1996).
75. R. Himmelreich, H. Hilbert, H. Plagens, E. Pirkl, B.-C. Li, R. Herrmann, *Nucleic Acids Res.* **24**, 4420 (1996).
76. A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S. G. Oliver, *Science* **274**, 546 (1996).
77. The genome of *Escherichia coli* was completed in January 1997, Blattner and Mori, personal communication.
78. I. Humphery-Smith, and W. Blackstock, *J. Protein Chem.* (1997) (in press).

