3590–3596 *Nucleic Acids Research, 1994, Vol. 22, No. 17*

PRINTS — a database of protein motif fingerprints

T.K.Attwood[1]*, M.E.Beck, A.J.Bleasby[2] and D.J.Parry-Smith+
Departments of Biochemistry and Molecular Biology, The University of Leeds, Leeds LS2 9JT,
[1]University College London, London WC1E 6BT and [2]DRAL, Warrington, Cheshire WA4 4AD, UK

ABSTRACT

ᴾRINTS is a compendium of protein motif 'fingerprints'. A fingerprint is defined as a group of motifs excised from conserved regions of a sequence alignment, whose diagnostic power or potency is refined by iterative databasescanning (in this case the OWL composite sequence database). Generally, the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. The use of groups of independent, linearly- or spatially-distinct motifs allows protein folds and functionalities to be characterised more flexibly and powerfully than conventional single-component patterns or regular expressions. The current version of the database contains 200 entries (encoding 950 motifs), covering a wide range of globular and membrane proteins, modular polypeptides, and so on. The growth of the databaseis influenced by a number of factors: e.g. the use of multiple motifs; the maximisation of sequence information through iterative database scanning; and the fact that the database searched is a large composite. The information contained within PRINTS is distinct from, but complementary to the consensus expressions stored in the widely-used PROSITE dictionary of patterns.

INTRODUCTION

With the vast number of protein sequences now at our disposal, it has become increasingly desirable to rationalise this information in order to expedite sequence, and ultimately structure, analysis. This process has involved the compilation of secondary 'value-added' databases, which tend to house patterns,motifs, profiles, domains, etc., all of which have been derived from the primary sequence sources. Such databases currently offer the most practical means of predicting the biological functions and structures of newly-determinedproteins. The most comprehensive and widely-used database of this type is the PROSITE dictionary of patterns, which currently contains 715 documentation entries describing 926 different patterns [1]. Other resources include, for example, the SBASE protein domain library [2], Gribskov's profiles [3], theBLOCKS database of aligned sequence segments [4], the MBCRR protein pattern library [5], and the dictionary of sequence motifs or peptide fragments [6,7].

Facilitating the identification of motifs is the common principle behind the development of all such resources. Each exploits a slightly different approach to pattern recognition, and each tends to use its own nomenclature, although PROSITE and its primary source, SWISS-PROT [8], upon which several of these databases are based, are emerging as standards. For example: a PROSITE pattern is a consensus expression derived from a single conserved region of a sequence alignment — it is the minimum expression that defines a given structural or functional site; an SBASE domain is an annotated domain segment with a known structure or function, and includes a cross-reference to the appropriate PROSITE pattern; a BLOCK is a short aligned segment corresponding to a conserved region automatically excised from an alignment generated from aprotein family included in PROSITE; and an MBCRR pattern is a consensus-like sequence derived from sets of related sequences in SWISS-PROT.

Our fingerprints similarly adhere to these standards: where relevant, they are cross-referenced to corresponding PROSITE patterns, and they are derived from sequences in a composite database [9] in which SWISS-PROT now has the highest priority. By contrast with most of the above approaches, however, fingerprints embrace groups of motifs from different conserved regions in alignments. Exploiting sets of motifs allows individual components of a protein's architecture and/or of its functional activities to be encoded within unique patterns. The more components encapsulated within the fingerprint, the more powerful it becomes at recognising related patterns within the sea of biological variation and diversity, experimental error and other sources of noise that sequence databases represent.

As part of the concerted effort to rationalise the growing volume of available primary sequence data and to facilitate protein sequence and structure analysis, we have assembled and made available a compendium of unique protein fingerprints: this is the PRINTS database, which is described in the following pages.

SOURCE DATABASE AND METHODS

The database used to derive individual fingerprints is OWL [9], a non-redundantcomposite of the major publicly-available primary sources: SWISS-PROT [8], PIR [10], GenBank (translation) [11,12] and NRL-3D (sequence data) [13]. Although strict redundancy criteria are applied to the amalgamation of the primarydatabases, error-checking of the sources themselves is

*To whom correspondence should be addressed

+Present address: Department of Molecular Sciences, Pfizer Central Research, Sandwich, Kent CT13 9NJ, UK

GPCRMGL Metabotropic Glutamate Receptor Signature
Type of fingerprint: COMPOSITE with 7 elements
Created by T.K.Attwood, 9-OCT-1993

1. ATTWOOD, T.K. and FINDLAY, J.B.C.
Fingerprinting G-protein-coupled receptors.
PROTEIN ENGINEERING 7 (2) 195-203 (1994).
2. MASU, M., TANABE, Y., TSUCHIDA, K., SHIGEMOTO, R. and NAKANISHI, S.
Sequence and expression of a metabotropic glutamate receptor.
NATURE 349 760-765 (1991).
3. HOUAMED, K.M., KUIJPER, J.L., GILBERT, T.L., HALDEMAN, B.A., CHARA, P., MULVIHILL, E.R., ALMERS, W., HAGEN, F.
Cloning, expression and gene structure of a G-protein-coupled glutamate receptor from rat brain.
SCIENCE 252 (5010) 1318-1321 (1991).
4. ABE, T., SUGIHARA, H., NAWA, H., SHIGEMOTO, R., MIZUNO, N. and NAKANISHI, S.
Molecular characterisation of a novel metabotropic glutamate receptor MGLUR5 coupled to inositol phosphate/Ca2+ signal transduction.
J.BIOL.CHEM. 267 (19) 13361-13368 (1992).
5. TANABE, Y., MASU, M., ISHII, T., SHIGEMOTO, R. and NAKANISHI, S.
A family of metabotropic glutamate receptors.
NEURON 8 (1) 169-179 (1992).

G-protein-coupled receptors (GPCRs) constitute a vast protein family that encompasses a wide range of functions (including various autocrine, paracrine and endocrine processes). They show considerable sequence diversity, on the basis of which they fall into distinct groups. We use the term clan to describe the GPCRs, as they embrace families for which there are indications of evolutionary relationship but between which there is no significant sequence similarity [1]. Currently known clan members include the rhodopsin-like and secretin-like GPCRs, the cAMP and fungal mating pheromone receptors, and the metabotropic glutamate receptors.

The metabotropic glutamate receptors are functionally and pharmacologically distinct from the ionotropic glutamate receptors. They are coupled to G-proteins and stimulate the inositol phosphate/Ca2+ intracellular signalling pathway [2-5]. Like the rhodopsins and other GPCRs, their sequences contain high proportions of hydrophobic residues grouped into 7 domains. However, while a similar 3D framework has been proposed to account for this, they do not show significant sequence similarity to the rhodopsin-type superfamily: they thus bear their own unique '7TM' signature (cf. signatures GPCRRHOD, GPCRSEC, GPCRCAMP and GPCRSTE2).

GPCRMGL is a 7-element fingerprint that provides a signature for the metabotropic glutamate-type GPCRs. The fingerprint was derived from an alignment of 5 sequences: the motifs encode the 7 hydrophobic membrane-spanning regions. A single scan of OWL21.1 was required to reach convergence, no further sequences being identified beyond the starting set. An update on OWL23.0 identified 1 further sequence.

SUMMARY INFORMATION
6 codes involving 7 elements
0 codes involving 6 elements
0 codes involving 5 elements
0 codes involving 4 elements
0 codes involving 3 elements
2 codes involving 2 elements

COMPOSITE FINGERPRINT INDEX

```
7 | 6    6    6    6    6    6    6
6 | 0    0    0    0    0    0    0
5 | 0    0    0    0    0    0    0
4 | 0    0    0    0    0    0    0
3 | 0    0    0    0    0    0    0
2 | 0    0    1    0    1    1    1
----------------------------------------
  | 1    2    3    4    5    6    7
```

True positives...

MGR2_RAT	MGR5_RAT	MGR3_RAT	MGR1_RAT
MGR4_RAT	A46742		

PROTEIN TITLES

MGR2_RAT	METABOTROPIC GLUTAMATE RECEPTOR 2 PRECURSOR - RATTUS NORVEGICUS
MGR5_RAT	METABOTROPIC GLUTAMATE RECEPTOR 5 PRECURSOR - RATTUS NORVEGICUS
MGR3_RAT	METABOTROPIC GLUTAMATE RECEPTOR 3 PRECURSOR - RATTUS NORVEGICUS
MGR1_RAT	METABOTROPIC GLUTAMATE RECEPTOR 1 PRECURSOR - RATTUS NORVEGICUS
MGR4_RAT	METABOTROPIC GLUTAMATE RECEPTOR 4 PRECURSOR - RATTUS NORVEGICUS
A46742	metabotropic glutamate receptor, mGluR6 - rats

SCAN HISTORY

OWL21_1	1	50	NSINGLE
OWL23_0	1	100	NSINGLE

```
INITIAL MOTIFS
GPCRMGL1        Length of motif = 22  Motif number = 1
Metabotropic glutamate receptor motif I - 1
                                      PCODE        ST     INT
VGPVTIACLGALATLFVLGVFV                MGR2_RAT     568    568
IAAVVFACLGLLATLFVTVIFI                MGR5_RAT     579    579
IIAIAFSCLGILVTLFVTLIFV                MGR1_RAT     593    593
IGPVTIACLGFLCTCIVITVFI                MGR3_RAT     577    577
VLPLFLAVVGIAATLFVVVTFV                MGR4_RAT     588    588
.
.

FINAL MOTIFS
GPCRMGL1        Length of motif = 22  Motif number = 1
Metabotropic glutamate receptor motif I - 2
                                      PCODE        ST     INT
VGPVTIACLGALATLFVLGVFV                MGR2_RAT     568    568
IAAVVFACLGLLATLFVTVIFI                MGR5_RAT     579    579
IIAIAFSCLGILVTLFVTLIFV                MGR1_RAT     593    593
IGPVTIACLGFLCTCIVITVFI                MGR3_RAT     577    577
VLPLFLAVVGIAATLFVVVTFV                MGR4_RAT     588    588
ALPLLLAVLGIMATTTIMATFM                A46742       581    581
```

Figure 1. Sample data from PRINTS. The example shown is the metabotropic glutamate receptor fingerprint. For convenience, only the first motif is depicted here.

not undertaken. In its current form, OWL thus includes errors that derive directly from these sources:results of database searches must therefore be viewed in this context.

Fingerprint construction commences with sequence alignment and excision of conserved motifs using SOMAP [14]. The individual motifs are used to dredge OWLiteratively using the ADSP sequence analysis package, which is a suite of procedures for database scanning, hitlist correlation, output of new fingerprint elements, and for re-scanning the database [15,16]. Four database-scanning routines are available in ADSP, of which NSINGLE is the method chosen:the algorithm interprets the aligned motifs essentially as a series of frequency matrices — i.e. identity searches are made, with no mutation or other similarity data to weight the results. Thus the weighting scheme is based on the calculation of residue frequencies for each position in the motifs, summing the scores of identical residues for each position of the retrieved match [15,16].

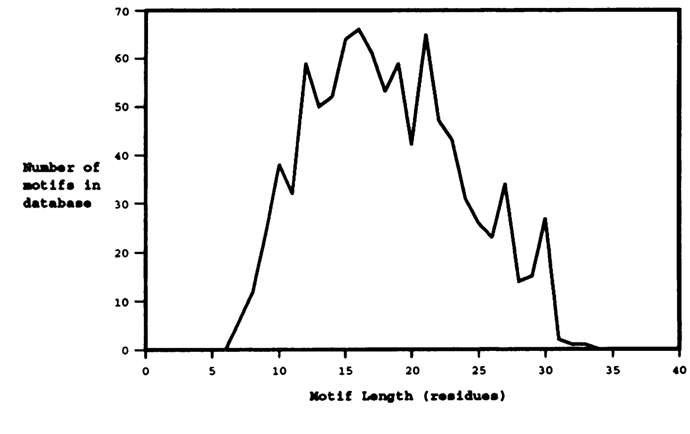

Figure 2. Graph showing the distribution of motif lengths in PRINTS 5.0. The most common motif lengths are between 12 and 22 residues.

DATABASE FORMAT

The PRINTS database is currently generated in the form of a single ASCII (text)file. This contains textual information documenting the particular protein family, details of how the fingerprint for that family was constructed and updated, and finally the aligned motifs themselves.

A sample entry is shown in detail in Figure 1. The contents are divided into a number of specific fields, relating to general information, bibliographic references, text, lists of matches, and the aligned motifs. In the general field at the top of the file, each entry is assigned a code by which it can be identified. This is followed by a description of the type of entry, which may be single (if the fingerprint has only one element) or multi-component (if it contains several) — in this latter case, the number of motifs contained is alsoindicated. To date we have included only two single-component entries: these have been derived using a modification of the fingerprint technique and are thus best regarded as special cases. Finally, the general field provides cross-references to corresponding PROSITE patterns, where relevant, together with entry creation and latest update information.

The example shown in Figure 1 depicts the fingerprint for the metabotropic glutamate receptors, a sub-class of the G-protein-coupled receptor (GPCR) superfamily. This is a seven-element fingerprint for which there is no corresponding PROSITE pattern. References are given, together with text detailing the nature of the family under investigation and the manner in which the fingerprint was derived, including cross-references to 4 other related PRINTS entries. Following the text is a summary, which indicates that 6 sequences match all the fingerprint elements and 2 make partial matches in thespecified version of OWL. This is followed by an analysis that indicates how well individual motifs have performed — in this case all have performed equally well, but four show a single additional match, presumed here

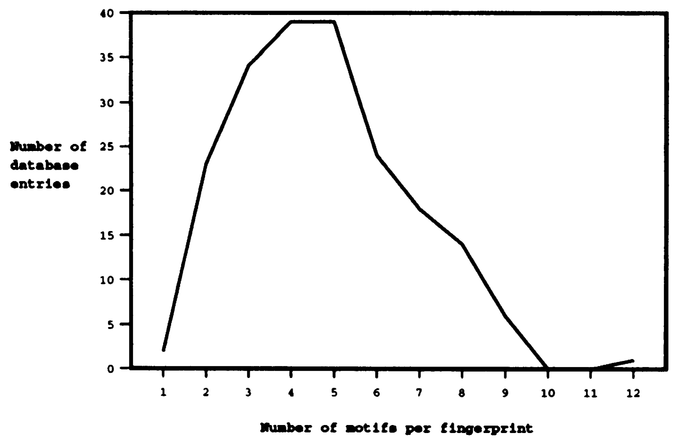

Figure 3. Graph showing the distribution of the number of motifs per fingerprint in PRINTS 5.0. The majority of fingerprints contain 4 or 5 motifs.

to be noise. All true- and false-positive and partial matches are then listed by means of the protein identification codes given in their respective source databases, together with a list of their titles.

The penultimate field provides a scan history, to indicate which versions of OWL have been scanned, how many iterations have been required, what length of hitlist has been used, and the scanning method employed: here, the entry was derived on OWL21.1 and has been updated on OWL23.0; the ADSP NSINGLE scanningmethod was used, and results reflect a hitlist length of 100.

The final field relates to the motifs themselves, listing both initial and final motifs, the motif lengths and their starting locations (ST). The intervals between adjacent motifs (INT) are also provided. Each motif isassigned a discrete code, which is the general code plus the number of that particular motif. For convenience, only the first motif (designated GPCRMGL1) is shown in Figure 1.

CONTENT OF THE CURRENT RELEASE

Release 5.0 of PRINTS (April 1994) contains 200 entries, encoding 950 individual motifs – a full list of entries is provided in Appendix 1. Figures 2 and 3 depict the nature of the database contents in terms of the distribution of motif lengths and the distribution of the number of motifs per fingerprint: from the graphs we find that motif lengths vary from 6 to 33 residues, the mostcommon lengths being between 12 and 22; and fingerprints contain from 1 to 12 motifs, the majority containing 4 or 5. The frequency of relatively long motifs, compared say with PROSITE in which pattern lengths peak at 10–12 residues [17], is largely a reflection of the number of membrane proteins with multiple transmembrane motifs included in the database.

DATABASE UPDATE AND GROWTH

The fingerprint database is released in major and minor versions: major versions are database expansions, i.e. they denote the addition of new materialto the resource; minor versions reflect updates of existing versions to bring the contents in line with the current version of OWL. To date, there have been 7 releases of the database: five major and two minor [18]. We endeavour to makea major or minor version available quarterly.

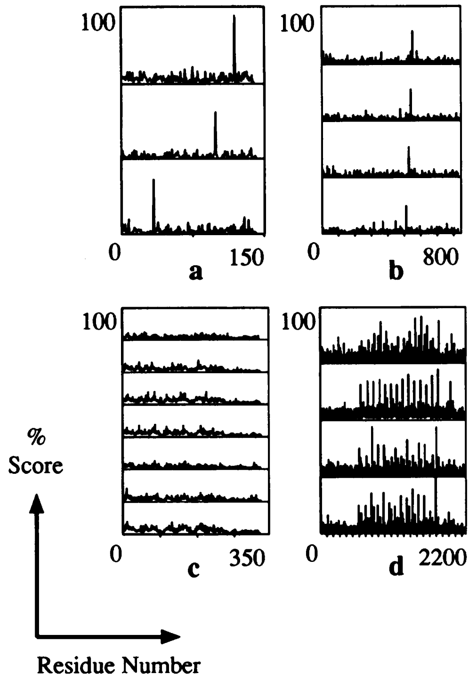

Figure 4. Fingerprint profiles. The horizontal axis represents the sequence, the vertical axis the percentage score of each fingerprint element (0–100 per element), and the peak a residue-by-residue match in the sequence, its leading edge marking the first position of the match. The profiles shown depict (**a**) the lipocalin fingerprint of mouse oncogene protein 24p3; (**b**) the diacylglycerol/phorbol ester-binding fingerprint of human vav oncogene; (**c**) the rhodopsin-like GPCR fingerprint of the *Dictyostelium dicoideum* cyclic-AMP receptor; and (**d**) the fibronectin type III repeat fingerprint of bovine fibronectin. Sharp peaks appearing in a systematic order along the length of the sequence and above the level of noise indicate matches with a given fingerprint, as evident in (a), (b) and (d), but not in (c).

The principal obstacle to the frequency of expansions, and particularly of updates, is the time-consuming nature of the approach. Deriving a fingerprint for a given protein family involves initial alignment and maximisation of sequence information through iterative scanning, with multiple motifs, of a large composite database. This is an exhaustive technique, but is consequently rigorous, and the precision of the resulting fingerprints tends to justify the sacrifice of speed.

DATABASE DISTRIBUTION

Interactive access to the PRINTS database can be achieved over the network via the SEQNET facility at Daresbury, where, together with OWL, it is part of an integrated database and software resource that also includes query languages for each of the databases, and several other programs for sequence alignment [14], pattern recognition [15] and global similarity searching [19].

PRINTS is also available directly via the anonymous-ftp servers at Daresbury (on s-ind2.dl.ac.uk in pub/database/prints — this directory also supplies documentation and other information files, which contain details of the database contents, update statistics, references, and so on), and at NCBI (ncbi.nlm.nih.gov) and EMBL (ftp.embl-heidelberg.de). In addition, it is available on the EMBL suite of CD-ROMs. The database requires ~4.6 Mb of disc storage.

APPLICATIONS

The fingerprint technique has been used to study a wide range of globular and membrane proteins, modular polypeptides, and so on. Specific uses have includedthe development of a fingerprint for the lipocalins, in which a previously unrecognised motif, bearing a striking spatial relationship to the well-known 'GxW' and 'TDY' motifs, was identified and used to characterise the family — the results revealed the hitherto unidentified mouse oncogene protein 24p3 to be a family member [20,21], as shown in Figure 4a. In a similar fashion, Figure 4b illustrates a fingerprint for the diacylglycerol/phorbol-ester binding domain, which was used to confirm that the human vav oncogene was a new member of the family [22]. By contrast, the design of a fingerprint for the rhodopsin-like G-protein-coupled receptors provided strong evidence that the Dictyostelium cyclic-AMP and yeast pheromone receptors, which had previously been ascribed to the group [23], were not sufficiently similar at the sequence level to be regarded as superfamily members [24] — Figure 4c; it also revealed significant differences between the sequences of the main superfamily and thoseof the olfactory receptors [24]. As a final example, to illustrate the flexibility of the approach, Figure 4d depicts the fingerprint derived for the fibronectin type III repeat [18], the sequences of which are too poorly conserved for regular expression-type patterns to be able to define them reliably [1].

FUTURE DIRECTIONS

Circumstances frequently arise in which regular-expression type patterns cannot unambiguously detect a particular protein family, usually because of their extreme sequence divergence (e.g. the fibronectin type III repeats). Similarly,fingerprints are not universally applicable. Sequences that have diverged to such an extent that no similarity remains will certainly escape detection by sequence-based methods of this type. In these cases, particularly in the analysis of membrane proteins, we aim to investigate the effects of substitution matrices, such as the BLOSUM series [25] and of newly-derived mutation data matrices [26,27], or we may consider the ANREP system, which exploits the concept of positional weights to express the relative importance of different parts of a motif [28], albeit with the caveat that such additionalinformation may actually compromise fingerprint potency by unacceptably increasing the level of background noise.

The use of fingerprints is also inappropriate when a region of conservation is confined to a single small patch of residues: because fingerprint potency improves with the number of motifs used, when applied to individual motifs, the method is likely to perform no better than any other single-motif method. But these are precisely the situations in which pattern methods tend to perform at their best — hence, for example, PROSITE is able to provide family-independent patterns for glycosylation and phosphorylation sites that contain only 3 or 4 residues, which PRINTS does not attempt to do. Pattern and fingerprint approaches are thus complementary and we aim to collaborate with Dr.Bairoch to integrate PRINTS with PROSITE. Software to search PRINTS is being made available via SEQNET, and we are also supplying a compendium of full sequence alignments (one for each PRINTS entry) to companion the resource.

CONCLUSION

Fingerprinting offers a powerful approach to the analysis of protein sequences: it inherently offers improved diagnostic reliability over single-motif methods by virtue of the mutual context provided by motif neighbours, and it allows rapid and striking visual diagnosis. Modern predictive methods are increasinglyexploiting multiple alignments as input to prediction algorithms, since multiple sequence information strongly enhances the signal [29]. The PRINTS database has not only been derived using this philosophy, but ultimately also stores its information in the form of alignments: these can themselves be the subject of detailed structure/function analyses, in a manner that is not possible with abstractions of sequence alignments such as regular expressions.

ACKNOWLEDGEMENTS

In particular we thank Professor J.B.C.Findlay for financial support (TKA and MEB), and we are grateful to everyone who has contributed entries to the database. The University of Leeds provided funds to initiate this project, which now continues with support from the Royal Society (TKA is a Royal Society University Research Fellow). The project benefits from use of the DRAL SEQNET facility.

REFERENCES

1. Bairoch, A. (1993) Nucleic Acids Res., 21 (13), 3097–3103.
2. Pongor, S., Skerl, V., Cserzo, M., Hatsagi, Z., Simon, G. and Bevilacqua, V. (1993) Nucleic Acids Res., 21 (13), 3111–3115.
3. Gribskov, M., Homyak, M., Edenfield, J. and Eisenberg, D. (1988) CABIOS, 4 (1), 61–66.
4. Henikoff, S. and Henikoff, J.G. (1991) Nucleic Acids Res., 19, 6565–6572.
5. Smith, R.F. and Smith, T.F. (1990) Proc.Natl.Acad.Sci.USA, 87, 118–122.
6. Ogiwara, A., Uchiyama, I., Seto, Y. and Kanehisa, M. (1992) Protein Engineering, 5 (6), 479–488.
7. Seto, Y., Ikeuchi, Y. and Kanehisa, M. (1990) Proteins, 8, 341–351.
8. Bairoch, A. and Boeckmann, B. (1993) Nucleic Acids Res., 21 (13), 3093–3096.
9. Bleasby, A.J. and Wootton, J.C. (1990) Protein Engineering, 3 (3), 153–159.
10. Barker, W.C., George, D.G., Mewes, H-W, Pfeiffer, F. and Tsugita, A. (1993) Nucleic Acids Res., 21 (13), 3089–3092.
11. Benson, D., Lipman, D.J. and Ostell, J. (1993) Nucleic Acids Res., 21 (13), 2963–2965.
12. Fickett, J.W. (1986) Trends Biochem.Sci., 11, 190.
13. Pattabiraman, N., Namboodiri, K., Lowrey, A. and Gaber, B.P. (1990) Protein Seq. Data Anal., 3, 387–405.
14. Parry-Smith, D.J. and Attwood, T.K. (1991) CABIOS, 7 (2), 233–235.
15. Parry-Smith, D.J. and Attwood, T.K. (1992) CABIOS, 8 (5), 451–459.
16. Attwood, T.K. and Findlay, J.B.C. (1993) Protein Engineering, 6 (2), 167–176.
17. Saqi, M.A.S. and Sternberg, M.J.E. (1994) Protein Engineering, 7 (2), 165–171.
18. Attwood, T.K. and Beck, M.E. (1994) Protein Engineering, 7 (7), 841–848.
19. Akrigg, D., Attwood, T.K, Bleasby, A.J., Findlay, J.B.C., Maughan, N.A., North, A.C.T., Parry-Smith, D.J., Perkins, D.N. and Wootton, J.C. (1992) CABIOS, 8 (3), 295–296.
20. Flower, D.R., North, A.C.T. and Attwood, T.K. (1993) Protein Science, 2, 753–761.
21. Flower, D.R., North, A.C.T. and Attwood, T.K. (1991) BBRC, 180 (1), 69–74.
22. Boguski, M., Bairoch, A., Attwood, T.K. and Michaels, G.S. (1992) Nature, 358, 113.

23. Chee, M.S., Satchwell, S.C., Preddie, E., Weston, K.M. and Barrell, B.G. (1990) Nature, 344, 774–777.
24. Attwood, T.K. and Findlay, J.B.C. (1994) Protein Engineering, 7 (2), 195–203.
25. Henikoff, S. and Henikoff, J.G. (1992) Proc.Natl.Acad. Sci.USA, 89, 10915–10919.
26. Jones, D., Taylor, W.R. and Thornton, J. (1992) CABIOS, 8 (3), 275–282.
27. Jones, D., Taylor, W.R. and Thornton, J. (1994) FEBS Letters, 339, 269–275.
28. Mehldau, G. and Myers, G. (1993) CABIOS, 9 (3), 299–314.
29. Persson, B. and Argos, P. (1994) J.Mol.Biol., 237, 182–192.

APPENDIX 1. Full list of entries in PRINTS 5.0. Indentations denote subfamilies.

Post-translational Modifications

P[a]	3[b]	Coagulation factor Gla domain signature
P	2	Bone matrix Gla domain signature

Domains

P	5	Cofilin/destrin family signature
P	4	Diacylglycerol/phorbol-ester signature
P	4	Type I EGF signature
	4	Type II EGF-like signature
	4	Type III EGF-like signature
	2	Fibronectin type I repeat signature
P	3	Fibronectin type II repeat signature
	4	Fibronectin type III repeat signature
P	2	Gram-positive coccus anchor signature
	1	Type I alpha-helix N-terminal signature
	1	Type I alpha-helix C-terminal signature
P	4	Kringle domain signature
	2	Leucine-rich repeat signature
	5	Small proline-rich repeat signature

DNA- or RNA-associated proteins

P	3	Homeobox signature
P	2	Homeotic antennapedia protein signature
P	2	Engrailed homeodomain signature
P	4	Paired box signature
P	5	POU domain signature
	4	Octamer-binding transcription factor signature
	2	Cro family helix-tum-helix signature
	2	Repressor protein helix-tum-helix signature
P	2	Bacterial regulatory protein araC signature
P	3	Bacterial regulatory protein asnC HTH signature
P	2	Bacterial regulatory protein crp HTH signature
P	2	Bacterial regulatory protein gntR HTH signature
P	2	Bacterial regulatory protein lacI HTH signature
P	2	Bacterial regulatory protein lacR HTH signature
P	3	Bacterial regulatory protein luxR HTH signature
P	3	Bacterial regulatory protein lysR HTH signature
P	3	Bacterial regulatory protein merR HTH signature
P	6	cAMP response element binding protein signature
P	5	Fos transforming protein signature
	5	Jun transcription factor signature
P	6	Myc proto-oncogene protein signature
P	5	Major sigma-70 factor signature
P	4	Sigma-54 factor signature
P	4	C4-type steroid receptor zinc finger signature
P	2	C2H2-type zinc finger signature
	4	Wilm's tumour protein signature
P	3	Cold shock protein signature
P	4	Fungal Zn-Cys binuclear cluster signature

Enzymes

Oxidoreductases

P	7	Catalase signature
P	4	Cu-Zn superoxide dismutase signature
P	4	Dihydrofolate reductase signature
P	5	Glyceraldehyde-3-phosphate dehydrogenases
P	3	Ribitol/alcohol dehydrogenase family signature
	6	Glucose/ribitol dehydrogenase family signature
P	4	Glu/Leu/Phe dehydrogenase signature
P	6	Hydroxymethylglutaryl-coA reductase signature
P	3	Lipoxygenase signature
P	3	Glutamine amidotransferase family signature
	7	Carbamoyl-phosphate synthase CPSase signature
	5	Carbamoyl-phosphate synthase GATase signature
P	4	Aspartate/omithine carbamoyltransferases
	6	Aspartate carbamoyltransferase signature
	5	Omithine carbamoyltransferase signature
	5	cAMP-dependent protein kinase signature
	6	cGMP-dependent protein kinase signature
P	3	Cytosine-specific methyltransferase signature
P	3	DNA-directed DNA-polymerase B signature
P	5	Thymidylate synthase signature
P	5	Tyrosine kinase catalytic domain signature

Hydrolases

	5	Alpha-amylase signature
	4	Alpha/beta hydrolase fold signature
P	3	Acylphosphatase signature
P	5	Alkaline phosphatase signature
P	5	Arginase signature
P	7	Beta-lactamase class A signature
	8	Putative herpesvirus protease signature
	6	Cation-transporting ATPase family signature
	5	H+-transporting ATPase signature
P	9	E1E2 sodium/potassium ATPase signature
	4	Vacuolar ATP synthase signature
P	5	Clp protease catalytic subunit P signature
P	4	Colipase signature
P	5	Cutinase signature
P	8	DNAse I signature
P	6	Fructose-1,6-bisphosphatase signature
P	6	Lysozyme/lactalbumin superfamily signature
	5	Lactalbumin signature
	6	Lysozyme signature
P	5	Matrixin signature
P	3	Asparaginase/glutaminase family signature
P	4	Proteasome component signature
P	8	RecA protein signature
P	7	Serine/threonine phosphatase family signature
P	6	Citrate synthase signature
P	4	DNA photolyase signature

Lyases

P	6	Enolase signature
	7	Phosphoenolpyruvate carboxylase signature
	6	RuBisCO small subunit signature

Others

P	2	AMP-binding signature

Electron transport proteins

P	3	Plant ferredoxin signature
	3	Rieske 2Fe-2S subunit signature
P	2	Rubredoxin signature

Other transport proteins

	7	Calcium channel signature
	6	Slow voltage-gated K+ channel signature
	8	Potassium channel signature
	7	Sodium channel signature
P	6	E.coli/Salmonella-type porin signature
	5	E.coli/Neisseria porin superfamily signature
	7	Neisseria sp. porin signature
P	4	Eukaryotic porin signature

P	5	L-lactate dehydrogenase signature
P	5	Bacterial luciferase signature
P	4	Nitrogenase component II signature
P	4	Tyrosinase copper-binding domain signature
P	5	Uricase signature
		Transferases
P	5	Adenylate kinase signature
	4	Anthranilate synthase component I signature
P	6	Anthranilate synthase component II signature
P	3	Lipocalin signature
	3	NMDA receptor/FABP signature
P	4	Fatty acid-binding protein signature
	5	NMDA receptor signature
P	9	Glutamate-aspartate transporter signature
P	5	Sugar transporter signature
	6	Glucose transporter signature
P	6	Sodium/alanine symporter signature
P	8	Sodium/neurotransmitter transporter signature

Other transport proteins contd.

P	4	Hemerythrin signature
P	5	Arthropod haemocyanin superfamily signature
P	5	Plant globin signature
P	4	Transthyretin signature
		Structural proteins
P	7	Actin signature
P	7	Annexin family signature
	8	Annexin type I signature
	8	Annexin type II signature
	8	Annexin type III signature
	8	Annexin type IV signature
	8	Annexin type V signature
	8	Annexin type VI signature
P	4	Amyloid A4 protein precursor signature
	3	Beta-amyloid peptide (beta-APP) signature
P	3	Cadherin signature
P	6	Connexin signature
	3	Flagellin signature
	3	Gliadin and LMW glutenin superfamily signature
	9	Alpha/beta gliadin family signature
	9	HMW glutenin signature
	7	Glutelin signature
	5	Myosin heavy chain signature
	5	Tropomyosin signature
	9	Herpesvirus major capsid protein signature
		Receptors
P	7	Bacterial opsin signature
	7	cAMP-type GPCR signature
	7	Metabotropic glutamate GPCR signature
P	7	Rhodopsin-like GPCR superfamily signature
	5	Olfactory receptor signature
P	3	Opsin signature
	6	Squid rhodopsin tail signature
P	7	Secretin-like GPCR family signature
	7	Yeast pheromone mating factor GPCR signature
P	5	Bacterial photosynthetic reaction centre signature
P	6	Low density lipoprotein (LDL) receptor signature
		Cytokines and growth factors
	3	Interleukin/heparin-binding growth factors
P	4	Interleukin 1 signature
P	4	Heparin-binding growth factor signature
P	5	Interleukin 2 family signature
P	3	Interferon alpha and beta subunit signature
P	4	Nerve growth factor signature
P	3	Pleiotrophin/midkine family signature
		Hormones and active peptides
P	3	Calcitonin signature
P	6	Erythropoietin signature
P	2	Galanin signature
P	4	Glycoprotein polypeptide hormone signature
P	2	Glucagon polypeptide hormone family signature
P	2	Insulin a chain signature
	2	Insulin b chain signature
P	2	Pancreatic hormone signature
		Toxins
P	2	Plant thionin signature
		Inhibitors
P	2	Disintegrin signature
P	4	Soybean trypsin inhibitor (Kunitz-type) signature
	3	Potato inhibitor I signature
		Protein secretion and chaperones
P	4	10 kD chaperonin signature
P	5	60 kD chaperonin signature
P	9	70 kD heat shock protein signature
		Others
P	5	Arrestin signature
P	4	NodO calcium binding signature
P	5	GTP-binding elongation factor signature
	3	Alpha G-protein (transducin) signature
	4	Beta G-protein (transducin) signature
P	3	G-protein beta WD-40 repeat signature
	3	Gamma G-protein (transducin) signature
	6	Haemagglutinin HA1/HA2 chain signature
	8	Haemagglutinin HA1 chain signature
	5	Haemagglutinin HA2 chain signature
	3	HMW kininogen signature
	5	Herpesvirus integral membrane protein signature
P	8	Prion protein signature
	12	Rhesus blood group protein signature
	4	Selectin complement binding repeat signature
	4	Bacterial sensor protein C-terminal signature
P	8	Tissue factor signature
P	6	Pathogenesis-related protein signature
P	3	Ubiquitin signature
	7	Virion infectivity factor signature

a denotes the existence of a corresponding PROSITE entry; **b** shows the number of motifs in the fingerprint.