Novel developments with the PRINTS protein fingerprint database

T. K. Attwood*, M. E. Beck[1], A. J. Bleasby[2], K. Degtyarenko[1], A. D. Michie and D. J. Parry-Smith[3]

Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WCIE 6BT, UK, [1]The University of Leeds, Leeds LS2 9JT, UK,, [2]CLRC Daresbury Laboratory, Warrington, Cheshire WA4 4AD, UK and [3]Department of Molecular Sciences, Pfizer Central Research, Sandwich, Kent CT13 9NJ, UK

Received September 27, 1996; Accepted October 8, 1996

ABSTRACT

The PRINTS database of protein family 'fingerprints' is a diagnostic resource that complements the PROSITE dictionary of sites and patterns. Unlike regular expressions, fingerprints exploit groups of conserved motifs within sequence alignments to build characteristic signatures of family membership. Thus fingerprints inherently offer improved diagnostic reliability by virtue of the mutual context provided by motif neighbours. To date, 600 fingerprints have been constructed and stored in PRINTS, representing a 50% increase in the size of the database in the last year. The current version, 13.0, encodes ~3000 motifs, covering a range of globular and membrane proteins, modular polypeptides, and so on. The database is accessible via UCL's Bioinformatics World Wide Web (WWW) server at http://www.biochem.ucl.ac.uk/bsm/dbbrowser/ . We describe here progress with the database, its Web interface, and a recent exciting development: the integration of a novel colour alignment editor (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA), which allows visualisation and interactive manipulation of PRINTS alignments over the Internet.

INTRODUCTION

In the analysis of novel protein sequences, in addition to routine searches of the primary data sources, it is now customary to extend search strategies to include a range of 'secondary' databases. These distil sequence information in the primary databanks into a variety of potent descriptors that aid family diagnosis: for example, PROSITE houses regular expression patterns and a number of profiles (1), and the BLOCKS database stores aligned, weighted motifs (2).

Of the available analysis methods, regular expressions are probably the easiest to understand. Their derivation involves the reduction of conserved motifs within sequence alignments into simple, single consensus expressions, in which all but the most signficant residue information is discarded. PROSITE is now the most comprehensive and widely-used secondary database, version 13.0 containing descriptors for 889 families and functional sites.

In terms of their performance in pattern recognition, regular expressions have certain limitations. Patterns may themselves encode flexibility, or fuzziness, but require query sequences to match them exactly. Thus sequences that differ only slightly from the definition will be missed. In view of this draw-back, more powerful discriminators (i.e., profiles) have been incorporated into PROSITE to provide an alternative means of diagnosis where patterns are likely to fail. Profiles are highly complex descriptors, generally encoding the full sequence length and allowing gap insertion in generating pairwise alignments between profile and target sequence; their numbers in PROSITE are therefore still relatively small.

We have developed a different approach to pattern recognition, which we term 'fingerprinting' (3,4). Within a sequence alignment, it is usual to find not one, but several motifs that characterise the aligned family. Diagnostically, it makes sense to use many, or all, of the conserved regions to build a family signature. In a database search, there is then a greater chance of identifying a distant relative, whether or not all parts of the signature are matched. Thus, for example, a sequence that matches only four of seven motifs may still be diagnosed as a true match if the motifs are matched in the correct order in the sequence, and the distances between them are consistent with that expected of true neighbouring motifs. The ability to tolerate mismatches, both at the level of residues within individual motifs, and at the level of motifs within the fingerprint as a whole, renders fingerprinting a very powerful diagnostic technique.

To facilitate sequence analysis and complement other secondary resources, we have made a range of unique protein fingerprints available in the PRINTS database (5). This paper describes recent progress with the PRINTS system and its evolving role as an information resource in computational molecular biology.

SOURCE DATABASE AND METHODS

PRINTS' source database is OWL (6) (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/OWL/OWL.html), a non-redundant

* To whom correspondence should be addressed. Tel: +44 171 419 3879; Fax: +44 171 380 7193; Email: attwood@bsm.bioc.ucl.ac.uk

composite of the major publicly-available primary sources: SWISS-PROT (7), PIR (8), GenBank (translation) (9) and NRL-3D (10).

Fingerprinting commences with sequence alignment and excision of conserved motifs using SOMAP (11). The motifs are used to dredge OWL independently using the ADSP sequence analysis package, a suite of procedures for iterative database scanning and hit-list correlation (3,4). The scanning algorithm interprets the 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. The weighting scheme is thus 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. Diagnostic performance is enhanced by iterative database scanning. The motifs therefore grow, and become more mature, with each database pass, as more sequences are matched and assimilated into the process. Full potency is gained from the mutual context provided by motif neighbours, which allows sequence identification even when some parts of the signature are absent.

Database format

PRINTS is built as a single ASCII (text) file. The contents are separated into specific fields, relating to general information, bibliographic references, text, lists of matches, and the motifs themselves—each line of a field is assigned a distinct two-letter code, allowing us to index the database for fast querying of its contents. In the general field, each entry is assigned an identification code and an accession number (of the form PR00000), followed by an indication of the number of constituent motifs in the fingerprint. Finally, where relevant, the general field provides cross-references to corresponding entries in a variety of other bio-databanks, including PROSITE, ProDom (12), SBASE (13), NRL-3D, SWISS-3DIMAGE (14), scop (15), cath (16), etc. Such links are vital for efficient communication between related databases and effectively broaden the scope of the resource. Similarly, the use of static accession numbers itself facilitates cross-referencing by other databases—PRINTS is now cross-referenced by BLOCKS, SBASE and GCRDb (17), and is linked to by PROSITE.

The full format has been described previously (18,19), so will not be discussed further here.

Content of the current release

Release 13.0 of PRINTS (September 1996) contains 600 entries, encoding ~3000 individual motifs. The complete contents list is available from the distribution sites and on the PRINTS WWW page (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/printscontents.html).

Database update and growth

PRINTS is released in major and minor versions: major releases are database expansions, i.e., they denote the addition of new material to the resource; minor releases reflect updates of existing entries to bring the contents in line with the current version of OWL. To date, there have been 17 releases of the database: 13 major and four minor. We endeavour to make a major or minor version available quarterly: in the last year, we have achieved four major and two minor releases.

The principal obstacle to the frequency of expansions, and particularly of updates, is the time-consuming nature of the approach. Deriving a fingerprint involves two major threads: (i) a computational aspect, which involves initial alignment and maximisation of sequence information through iterative scanning, with multiple motifs, of a large composite database; and (ii) an annotation component, which involves researching each family, and linking sequence conservation information to known structural or functional data. This is a rigorous, exhaustive technique. The precision of the results, coupled with the quality of annotations, tends to justify the sacrifice of speed, and sets the database apart from the growing number of automatically-derived pattern resources, for which there are no annotations, and hence no appropriate mechanisms for result validation.

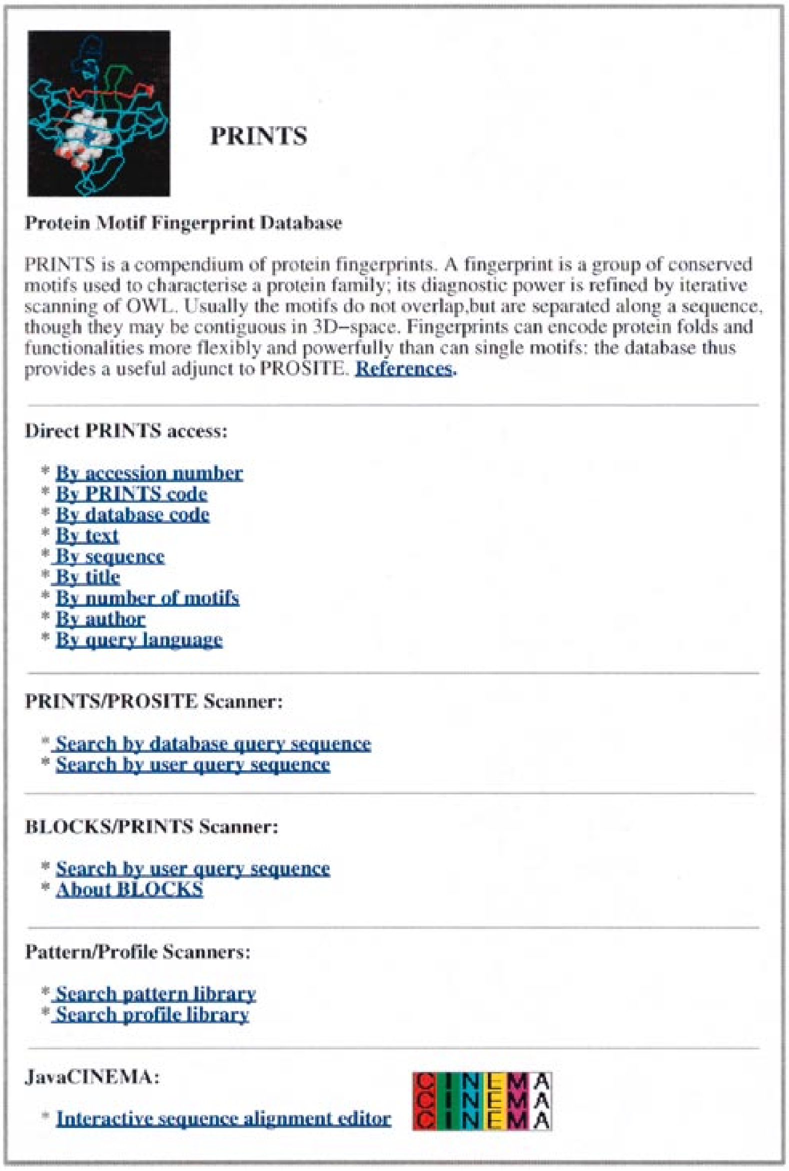
Database distribution

PRINTS is available for interactive use via the SEQNET service. It may be retrieved directly from the anonymous-ftp servers at Daresbury (s-ind2.dl.ac.uk in pub/database/prints), NCBI (ncbi.nlm.nih.gov), EBI (ftp.ebi.ac.uk in pub/databases), EMBL (ftp.embl-heidelberg.de) and UCL (ftp.biochem.ucl.ac.uk in pub/prints). In addition, it is distributed on the EMBL suite of CD-ROMs. The database requires ~60 Mb of disk storage.

In addition, the database is accessible from UCL's DbBrowser Bioinformatics Server, at http://www.biochem.ucl.ac.uk/bsm/dbbrowser/ (20). The server primarily provides access to OWL, PRINTS and ALIGN, the compendium of alignments used to create PRINTS entries (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/ALIGN/ALIGN.html). Figure 1 shows part of the PRINTS home page (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS. html), which allows keyword searching of database code, accession number, text, sequence, etc. Such queries are made possible by links to the query language, but are presented in a manner that shields the user from its syntax, which is desirable for routine, trivial queries. Where query results are of particular interest, the full entry may be retrieved to discover more about the fingerprint, as shown in Figure 2.

The PRINTS home page also provides a facility to search PRINTS and PROSITE simultaneously, offering an instant diagnosis of any query sequence (21). Results may be viewed in the form of fingerprint profiles (22), as shown in Figure 3. A variety of complementary pattern database search tools is also provided, to afford users the opportunity to perform comprehensive searches, e.g. by additionally scanning BLOCKS and BLOCKS- format PRINTS (http://www.blocks.fhcrc.org/blocks_search.html), the profile library (http:// ulrec3.unil.ch/software/profilescan.html), and so on.

An important new facility has been added to the Web interface and deserves special mention. As described above, associated with each fingerprint is a parent alignment, from which conserved motifs are selected for database scanning. These alignments are stored in PRINTS' companion compendium, ALIGN. For each PRINTS entry, an explicit link has now been made to ALIGN via the JavaCINEMA colour alignment editor (23). This is a significant advantage over previous releases: where formerly we offered only static PostScript or Gif images for viewing alignments, now, using Java, for the first time we are able to provide facilities for visualisation and interactive manipulation of alignments over the Internet.



PRINTS

Protein Motif Fingerprint Database

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of OWL. Usually the motifs do not overlap,but are separated along a sequence, though they may be contiguous in 3D–space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs: the database thus provides a useful adjunct to PROSITE. References.

Direct PRINTS access:

* By accession number
* By PRINTS code
* By database code
* By text
* By sequence
* By title
* By number of motifs
* By author
* By query language

PRINTS/PROSITE Scanner:

* Search by database query sequence
* Search by user query sequence

BLOCKS/PRINTS Scanner:

* Search by user query sequence
* About BLOCKS

Pattern/Profile Scanners:

* Search pattern library
* Search profile library

JavaCINEMA:

* Interactive sequence alignment editor

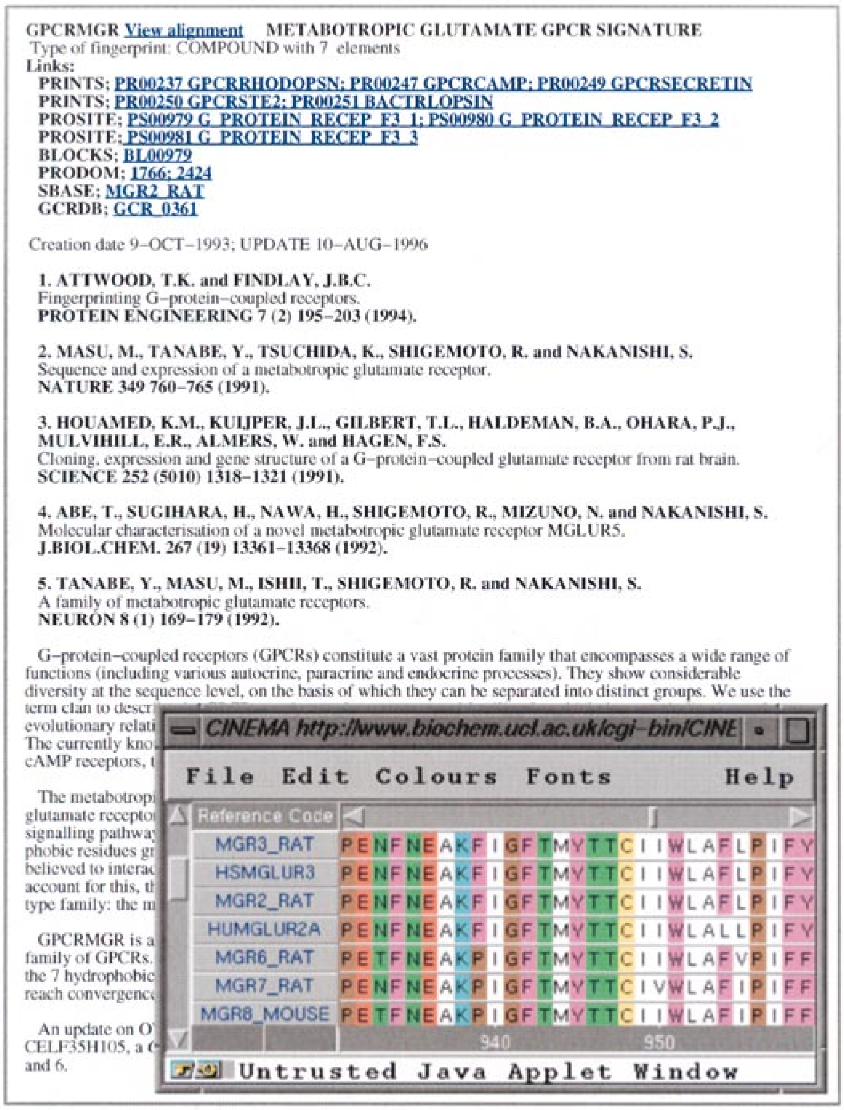
Figure 1. Part of the PRINTS home page. A range of direct access points is available, allowing simple queries by keyword searching, or more complex queries using the query language. A variety of complementary pattern database search tools is also provided, to afford users the opportunity to perform comprehensive searches. The most recently-added tool is the JavaCINEMA interactive sequence alignment editor.

Applications

The fingerprint technique has been used to study a wide range of globular and membrane proteins, modular polypeptides, and so on (4,24–27). In recent database releases, particular emphasis has been placed on the elucidation of discriminatory fingerprints for a range of G-protein-coupled receptor (GPCR) families and subfamilies (4,27). This has become increasingly important as the growth of the rhodopsin-like family has soared: there are now >1000 rhodopsin-like GPCRs known and diagnosis of certain family outliers has become more and more difficult. By expanding the range of GPCR families covered in PRINTS, the fingerprint facility on the Web now effectively provides an instant diagnostic tool for putative GPCRs—this is illustrated in Figure 3, in which a hypothetical *C.elegans* protein from SWISS-PROT (YMJC_CAEEL) is shown to make a partial match with the rhodopsin-like fingerprint, which encodes the seven transmembrane domains. The sequence is not diagnosed by PROSITE because it contains changes in the third transmembrane domain, which alone provides the basis for the PROSITE pattern. Using the fingerprint approach, it is possible to detect such Twilight relationships because of the diagnostic framework provided by neighbouring motifs. Thus, in spite of the relative weakness of several peaks in the fingerprint profile, the mutual context provided by the remaining fingerprint elements allows us to make a reliable assessment of family membership.

Future directions

In order to address more effectively the flood of information arising from the various genome projects, it is essential to increase levels of automation, and relieve many of the current manual



GPCRMGR View alignment METABOTROPIC GLUTAMATE GPCR SIGNATURE
Type of fingerprint: COMPOUND with 7 elements
Links:
PRINTS; PR00237 GPCRRHODOPSN; PR00247 GPCRCAMP; PR00249 GPCRSECRETIN
PRINTS; PR00250 GPCRSTE2; PR00251 BACTRLOPSIN
PROSITE; PS00979 G_PROTEIN_RECEP_F3_1; PS00980 G_PROTEIN_RECEP_F3_2
PROSITE; PS00981 G_PROTEIN_RECEP_F3_3
BLOCKS; BL00979
PRODOM; 1766; 2424
SBASE; MGR2_RAT
GCRDB; GCR_0361

Creation date 9–OCT–1993; UPDATE 10–AUG–1996

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., KULJPER, J.L., GILBERT, T.L., HALDEMAN, B.A., OHARA, P.J., MULVIHILL, E.R., ALMERS, W. and HAGEN, F.S.
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.
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 diversity at the sequence level, on the basis of which they can be separated into distinct groups. We use the
term clan to descr
evolutionary relati
The currently kno
cAMP receptors, t

The metabotropi
glutamate recepto
signalling pathway
phobic residues gr
believed to interac
account for this, th
type family: the m

GPCRMGR is a
family of GPCRs.
the 7 hydrophobic
reach convergence

An update on O
CELF35H105, a G
and 6.

CINEMA http://www.biochem.ucl.ac.uk/cgi-bin/CINE

Reference Code	
MGR3_RAT	PENFNEAKFIGFTMYTTCIIWLAFLPIFY
HSMGLUR3	PENFNEAKFIGFTMYTTCIIWLAFLPIFY
MGR2_RAT	PENFNEAKFIGFTMYTTCIIWLAFLPIFY
HUMGLUR2A	PENFNEAKFIGFTMYTTCIIWLALLPIFY
MGR6_RAT	PETFNEAKPIGFTMYTTCIIWLAFVPIFF
MGR7_RAT	PENFNEAKPIGFTMYTTCIVWLAFIPIFF
MGR8_MOUSE	PETFNEAKPIGFTMYTTCIIWLAFIPIFF
	940 950

Figure 2. Sample data from PRINTS, showing part of the entry for the metabotropic glutamate receptor family. The information is separated into specific fields, relating to text, references, etc. The cross-references at the top of the file allow efficient coupling to related databases. The hyperlink for viewing the parent alignment invokes the JavaCINEMA interactive alignment editor, as shown, allowing the user either simply to view or to augment the alignment as desired.

burdens inherent in database maintenance. This is already imperative, given the difficulties in attracting funding for database curation. In the short term, emphasis will be placed on adding new families to PRINTS, rather than on routinely updating existing ones. Attention will then be given to developing more automated curation strategies. We will, however, maintain a balance between manual input (especially at the stage of annotation) and automatic processing. In practical applications, the power of secondary databases derives not only from the reliability of their diagnostic performance, but also from the extent and quality of their family documentations. Annotated databases tend to be more reliable than their fully-automated counterparts, which are more error prone and provide little or no validation either for the patterns they house or for matches to those patterns.

In addition to addressing the practicalities of database maintenance, we also aim to enhance the range of analysis tools available, to make the information within PRINTS more readily accessible to users. For example, we are extending the alignment applet to include a structural viewer so that, for families for which coordinates are available, their fingerprints may be visualised in a 3D context.

CONCLUSION

Bioinformatics is a technically-demanding discipline, in terms of both the nature and the scale of the undertaking, and promises enormous practical dividends as it begins to reveal the hidden jewels of the human genome. Secondary databases, such as PRINTS, are an important part of this endeavour: their scope and subtlety make them powerful tools for diagnosing the relationships between sequences that underlie the identification of function. But none of these resources is sufficient in itself. No single analysis method is yet infallible, and no pattern database

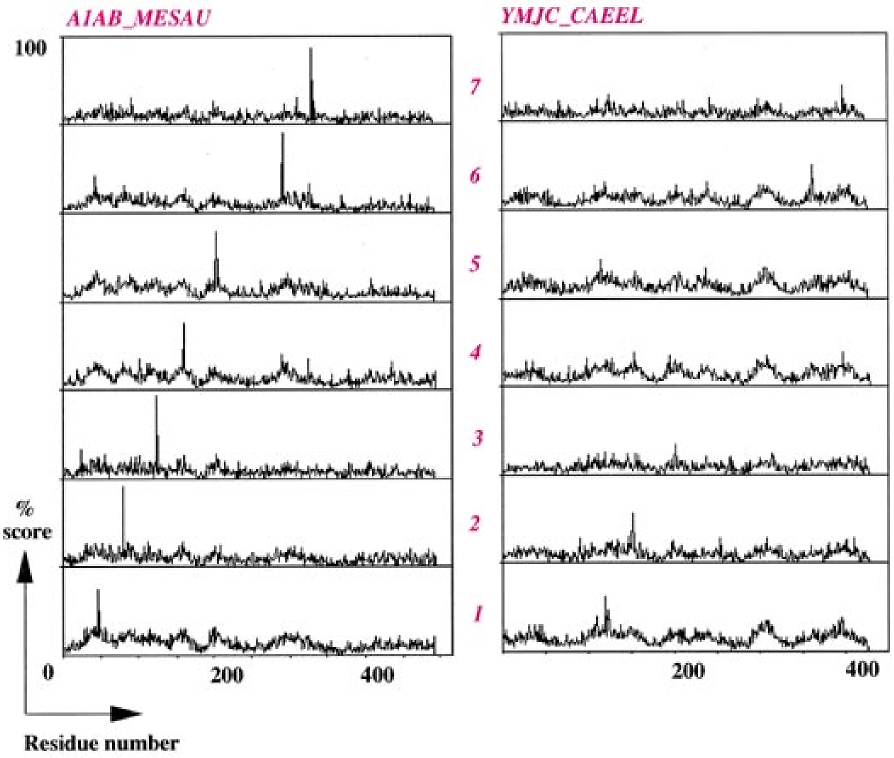


Figure 3. Fingerprint profiles returned by the PRINTS/PROSITE scanner. The horizontal axis represents the sequence, the vertical axis the percentage score of each fingerprint element (0–100 per element), and the peak a residue-by-residue match in the sequence, its leading edge marking the first position of the match. The profiles depict rhodopsin-like GPCR fingerprints of hamster α-1B adrenergic receptor and of a *C.elegans* hypothetical protein. Sharp peaks appearing in a systematic order along the length of the sequence and above the level of noise indicate matches with the constituent motifs. The adrenergic receptor is a known true-positive family member, matching all seven transmembrane domains; the *C.elegans* sequence fails to make a complete match, but can still be reliably identified with the GPCR superfamily because of the diagnostic framework provided by motif neighbours.

complete. Together with PROSITE, BLOCKS, the profile library, etc., PRINTS thus provides one of several potent weapons in the sequence analyst's armoury.

ACKNOWLEDGEMENTS

We thank everyone who has contributed entries to the database. PRINTS is built and maintained at UCL with support from the Royal Society (T.K.A. is a Royal Society University Research Fellow). The project benefits from use of the BBSRC SEQNET facility.

REFERENCES

1 Bairoch,A., Bucher,P. and Hofmann,K. (1996) *Nucleic Acids Res.*, **24**, 189–196.
2 Pietrokovski,S., Henikoff,S. and Henikoff,J.G. (1996) *Nucleic Acids Res.*, **24**, 197–200.
3 Parry-Smith,D.J. and Attwood,T.K. (1992) *CABIOS*, **8**, 451–459.
4 Attwood,T.K. and Findlay,J.B.C. (1994) *Protein Engng*, **7**, 195–203.
5 Attwood,T.K., Beck,M.E., Bleasby,A.J., Degtyarenko,K. and Parry-Smith,D.J. (1996) *Nucleic Acids Res.*, **24**, 182–188.
6 Bleasby,A.J., Akrigg,D. and Attwood,T.K. (1994) *Nucleic Acids Res.*, **22**, 3574–3577.
7 Bairoch,A. and Apweiler,R. (1996) *Nucleic Acids Res.*, **24**, 21–25.
8 George,D.G., Barker,W.C., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1996) *Nucleic Acids Res.*, **24**, 17–20.
9 Benson,D.A, Boguski,M., Lipman,D.J. and Ostell,J. (1996) *Nucleic Acids Res.*, **24**, 1–5.
10 Pattabiraman,N., Namboodiri,K., Lowrey,A. and Gaber,B.P. (1990) *Protein Seq. Data Anal.*, **3**, 387–405.
11 Parry-Smith,D.J. and Attwood,T.K. (1991) *CABIOS*, **7**, 233–235.
12 Sonnhammer,E.L.L. and Kahn,D. (1994) *Protein Sci.*, **3**, 482–492.
13 Murvai,J., Gabrielian,A., Fabian,P., Hatsagi,Z., Degtyarenko,K., Hegyi,H. and Pongor,S. (1996) *Nucleic Acids Res.*, **24**, 210–213.
14 Peistch,M.C., Wells,T.N.C., Stampf,D.R. and Sussman,J.L. (1995) *Trends Biochem. Sci.*, **20**, 82–84.
15 Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
16 Michie,A.D., Hutchinson,E.G., Laskowski,R.A., Orengo,C.A. and Thornton,J.M. (1995) Proceedings of the CCP4 Meeting, Chester.
17 Kolakowski,L.F. (1994) *Receptors and Channels*, **2**, 1–7.
18 Attwood,T.K. and Beck,M.E. (1994) *Protein Engng*, **7**, 841–848.
19 Attwood,T.K., Beck,M.E., Bleasby,A.J. and Parry-Smith,D.J. (1994) *Nucleic Acids Res.*, **22**, 3590–3596.
20 Michie,A.D., Jones,M.L. and Attwood,T.K. (1996) *Trends Biochem. Sci.*, **21**, 191.
21 Perkins,D.N. and Attwood,T.K. (1996) *CABIOS*, **12** (2), 89–94.
22 Attwood,T.K and Parry-Smith,D.J. (1996) In Pickover,C. (ed.), *Visualizing Biological Information.* World Scientific Publishing Co., Singapore, pp. 145–157.
23 Parry-Smith,D.J., Payne,A.W.R, Michie,A.D. and Attwood,T.K. (1996) *Gene Combis*, submitted.
24 Flower,D.R., North,A.C.T. and Attwood,T.K. (1993) *Protein Sci.*, **2**, 753–761.
25 Flower,D.R., North,A.C.T. and Attwood,T.K. (1991) *Biochem. Biophys. Res. Commun.*, **180**, 69–74.
26 Boguski,M., Bairoch,A., Attwood,T.K. and Michaels,G.S. (1992) *Nature*, **358**, 113.
27 Attwood,T.K. and Findlay,J.B.C. (1993) *Protein Engng*, **6**, 167–176.