304–308 *Nucleic Acids Research, 1998, Vol. 26, No. 1*

The PRINTS protein fingerprint database in its fifth year

T. K. Attwood*, M. E. Beck[1], D. R. Flower[2], P. Scordis and J. N. Selley

Department of Biochemistry and Molecular Biology, University College London, London WCIE 6BT, UK, [1]Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK and [2]Department of Physical and Metabolic Sciences, Astra Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RF, UK

Received September 22, 1997; Accepted September 24, 1997

ABSTRACT

PRINTS is a database of protein family 'fingerprints' offering a diagnostic resource for newly-determined sequences. By contrast with PROSITE, which uses single consensus expressions to characterise particular families, PRINTS exploits groups of motifs to build characteristic signatures. These signatures offer improved diagnostic reliability by virtue of the mutual context provided by motif neighbours. To date, 800 fingerprints have been constructed and stored in PRINTS. The current version, 17.0, encodes ~4500 motifs, covering a range of globular and membrane proteins, modular polypeptides, and so on. The database is accessible via the UCL Bioinformatics World Wide Web (WWW) Server at http://www. biochem. ucl.ac.uk/bsm/dbbrowser/ . We have recently enhanced the usefulness of PRINTS by making available new, intuitive search software. This allows both individual query sequence and bulk data submission, permitting easy analysis of single sequences or complete genomes. Preliminary results indicate that use of the PRINTS system is able to assign additional functions not found by other methods, and hence offers a useful adjunct to current genome analysis protocols.

INTRODUCTION

The last two decades have seen remarkable advances in molecular biology: 20 years ago sequencing a single gene was considered a monumental technical achievement; today, the sequencing of whole genomes has become almost routine. Advances in the fundamental techniques of sequencing, in concert with advances in laboratory automation and robotics, have led to the rapid and unprecedented accumulation of macromolecular sequence data. The challenge resides not just in the management of this huge quantity of information, but also in its analysis. One of the main goals of bioinformatics is to uncover the knowledge implicit within the data.

The decisive step in this knowledge-discovery process is often the identification of the family to which a newly-identified gene belongs; from this devolves a wealth of insights into function. With its links to 3D structure and post-translational modifications, and thus biological function, it is generally thought that the amino acid sequence, rather than the nucleic acid sequence, is the most appropriate level at which to seek such relationships.

Secondary, so-called value-added, databases are now standard tools in sequence analysis strategies. Such resources distil sequence information from the primary databanks into a variety of potent descriptors that aid family diagnosis: PROSITE, for example, houses regular expression patterns and a small number of profiles (1); the BLOCKS database stores aligned, weighted motifs, or blocks (2); Pfam offers a range of hidden Markov models (HMMs) (3); and PRINTS provides groups of aligned, unweighted sequence motifs, or fingerprints (4). Diagnostically, each of these types of descriptor has different strengths and weaknesses and hence different areas for optimum application. In terms of family coverage, the databases tend to differ in content, and the most effective search strategies should ideally combine them all.

The technique of protein fingerprinting (5,6) arose largely because of the limitations of single-motif regular expression pattern-matching methods: these give binary 'hit or miss', 'match or no match' diagnoses that provide no biological context with which to assess the significance of a result. However, 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 such 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. For example, a sequence that matches only three of six 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 those 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 fingerprints as a whole, renders fingerprinting a powerful diagnostic tool.

To facilitate sequence analysis and complement other secondary resources, we have made a range of protein fingerprints available in the PRINTS database (4). In this paper, we describe recent progress with the database, its new search software, and some applications.

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

SOURCE DATABASE AND METHODS

At present, the source database for PRINTS is OWL (7) (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/OWL/), a non-redundant composite of the major publicly-available primary sources: SWISS-PROT (8), PIR (9), GenBank (translation) (10) and NRL-3D (11).

Fingerprinting is an iterative procedure that commences with manual sequence alignment and excision of conserved motifs using SOMAP (12). The motifs are used to trawl OWL independently using the ADSP sequence analysis package (5,6). 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 parts of the signature are absent.

Database format

PRINTS is currently 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 the database to be indexed for fast querying of its contents (13). Entries are assigned both an identification code and an accession number to facilitate cross-referencing by other databases. Conversely, where relevant, cross-references are provided to other databanks (e.g., PROSITE (1), SBASE (14), scop (15), CATH (16), etc.) in order to promote efficient communication between related bioinformatics resources and effectively broaden the scope of sequence analysis strategies. The full format has been described previously (13,17,18), so will not be discussed further here.

Content of the current release

Release 17.0 of PRINTS (September 1997) contains 800 entries, encoding 4460 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 21 releases of the database: 17 major and four minor. We endeavour to make a major or minor version available quarterly; in the last year, we have achieved four major 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, where possible, linking sequence conservation information to known structural or functional data. This is a rigorous, exhaustive and thus time-consuming technique. But the precision of the results, coupled with the quality of annotations, has justified 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 UCL's DbBrowser Bioinformatics Server, at http://www.biochem.ucl.ac.uk/bsm/dbbrowser/ (19). The PRINTS home page (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/) allows keyword searching of database code, accession number, text, sequence, etc.. Such queries are made possible by links to the query language (13), but are presented in a manner that shields the user from its syntax, which is desirable for routine queries. Where results are of particular interest, the full entry may be retrieved to discover more about the fingerprint. As shown in Figure 1, hyperlinks allow the user to retrieve related information from a variety of bioinformatics resources. In addition, the parent alignment from which the fingerprint was derived may be downloaded via a link to the CINEMA colour alignment editor (20), allowing visualisation and interactive manipulation of the alignment of interest.

For local installation, the database may be retrieved directly from the anonymous-ftp servers at UCL (ftp.biochem.ucl.ac.uk in pub/prints), Daresbury (s-ind2.dl.ac.uk in pub/database/ prints), EBI (ftp.ebi.ac.uk in pub/databases), EMBL (ftp.embl-heidelberg.de) and NCBI (ncbi.nlm.nih.gov). In addition, it is distributed on the EMBL suite of CD-ROMs.

Derivative databases

A particular strength of the PRINTS database is that the underlying data are stored in the form of raw sequence alignments. This allows different implementations to be set up using a variety of alternative scoring methods and/or abstractions. For example, a BLOCKS-format version of the resource is available at the Fred Hutchinson Cancer Research Center (http://www.blocks.fhcrc.org/blocks_search.html); this exploits the powerful scoring method originally developed for the BLOCKS database (2). Alternatively, the protein function identification resource (IDENTIFY) at Stanford (http://dna.stanford.edu/identify/) overlays a fuzzy regular expression approach over the PRINTS multiply-aligned motifs and offers different levels of stringency from which to infer the significance of matches. Such derivative databases are useful for providing different perspectives on the same data set: they afford the opportunity to validate results, where there are corresponding matches in more than one resource; and they offer the chance to diagnose matches that may have been missed by the original implementation.

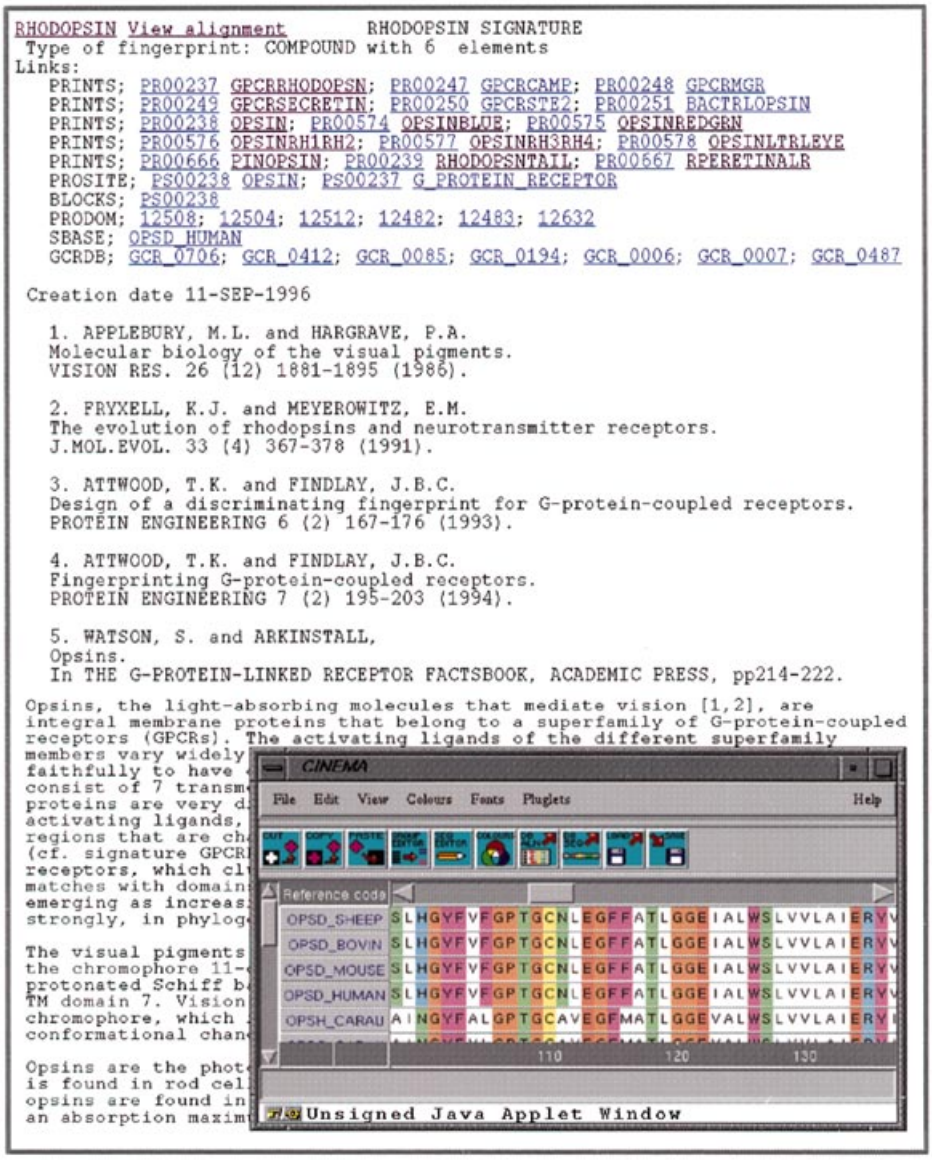


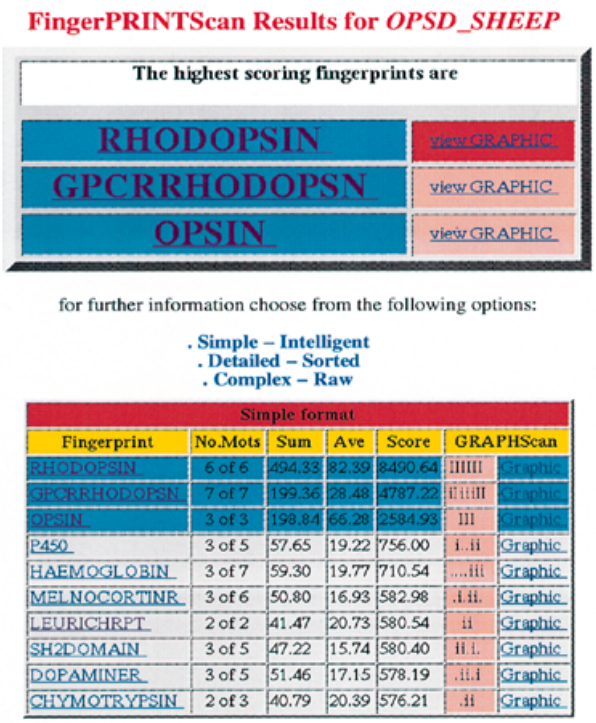
Figure 1. Sample data from PRINTS, showing part of the entry for the rhodopsin GPCR family. The information is separated into specific fields, relating to text, references, etc. The cross-references at the top of the file allow efficient coupling to related databases. The hyperlink for viewing the parent alignment invokes the CINEMA interactive alignment editor, as shown, allowing the user either to view or to augment the alignment as desired.

New search software

An important new facility has been added to the Web interface and deserves special mention. Secondary databases are of limited value without appropriate search tools. Our previous software (21) was limited to single sequence queries and could not differentiate between partial, but nevertheless true, fingerprint matches and random, high-scoring individual motif hits. We have addressed these problems with a new suite of programs, which provides facilities for: (i) interactive, individual query sequence submission against the full database; (ii) non-interactive, bulk query submission against the full database (with full genome analysis in mind); and (iii) interactive, individual sequence searching against a named fingerprint. Results from these programs are returned in distinct ways, with an attempt made to cater for both casual and expert users: the first offers an 'intelligent' best guess, based on the occurrence of the highest-scoring full or partial fingerprint match, but more detailed results are provided in different layers via an extended HTML table, as illustrated in Figure 2; the second facility provides only brief information, which is returned via email; and the third option provides a graphical cartoon view of a single fingerprint profile, offering an instant diagnosis of any query sequence, as shown in Figure 3.

Applications

The fingerprint technique has been used to study a wide range of globular, membrane, and modular proteins (6,22,23). In recent database releases, particular emphasis has been placed on the



FingerPRINTScan Results for *OPSD_SHEEP*

The highest scoring fingerprints are

RHODOPSIN	view GRAPHIC
GPCRRHODOPSN	view GRAPHIC
OPSIN	view GRAPHIC

for further information choose from the following options:

. Simple – Intelligent
. Detailed – Sorted
. Complex – Raw

Simple format

Fingerprint	No.Mots	Sum	Ave	Score	GRAPHScan	
RHODOPSIN	6 of 6	494.33	82.39	8490.64	IIIIII	Graphic
GPCRRHODOPSN	7 of 7	199.36	28.48	4787.22	iiiiiii	Graphic
OPSIN	3 of 3	198.84	66.28	2584.93	III	Graphic
P450	3 of 5	57.65	19.22	756.00	i..ii	Graphic
HAEMOGLOBIN	3 of 7	59.30	19.77	710.54	iii	Graphic
MELNOCORTINR	3 of 6	50.80	16.93	582.98	.i.ii.	Graphic
LEURICHRPT	2 of 2	41.47	20.73	580.54	ii	Graphic
SH2DOMAIN	3 of 5	47.22	15.74	580.40	ii.i.	Graphic
DOPAMINER	3 of 5	51.46	17.15	578.19	.ii.i	Graphic
CHYMOTRYPSIN	2 of 3	40.79	20.39	576.21	.ii	Graphic

Figure 2. Search output returned by FingerPRINTScan. For a given query sequence, the program makes an 'intelligent' best guess, based on the occurrence of the highest-scoring full or partial fingerprint match. The user may then choose to view different levels of matches, pushing further into the Twilight Zone, where results are no longer statistically significant. In this example, the query sequence, ovine rhodopsin, has been diagnosed as a member of the rhodopsin-like GPCR superfamily belonging to the opsin family, and is more specifically identified as a rhodopsin. In the next level of output, the top ten best-scoring matches are given. This table shows the number of motifs matched, the scores for individual motifs and for the fingerprint as a whole, and a thumb-nail sketch, which gives an instant visual diagnosis of the match; hyperlinks to the graphical output option allow such sketches to be visualised in more detail.

elucidation of discriminatory fingerprints for a range of G-protein-coupled receptor (GPCR) families and subfamilies (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/GPCR). This has become important as the growth of the rhodopsin-like family has soared; there are now >1000 rhodopsin-like GPCRs known and diagnosis of family outliers has become increasingly difficult. By expanding the range of GPCR families covered in PRINTS, the fingerprint facility on the Web effectively provides an instant diagnostic tool for putative GPCRs. This is illustrated in Figure 3, in which a *Caenorhabditis elegans* integral membrane protein from SWISS-PROT (SG12_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; BLAST (24) also fails to return any significant scores, and no matches are reported from searches of resources such as BLOCKS and Pfam. 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 or absence of several peaks in the fingerprint profile, the mutual context provided by the remaining fingerprint elements allows us to infer a distant family relationship.

The ability to detect distant familial relationships is particularly important in the context of complete genome analysis. Protocols based, for example, on the combination of BLAST and PROSITE alone, are likely to miss significant matches. Preliminary results from the examination of the *Saccharomyces cerevisiae* genome (25) suggest that application of the PRINTS system has been able to make family assignments for ~300 sequences designated as hypothetical proteins, i.e., the method has assigned potential functions to ~10% of uncharacterised sequences. This figure has to be set in the context of the size of PRINTS, which is small in comparison with the primary databases; as PRINTS grows, inevitably its impact in such applications will increase. But still, this is an encouraging early result and is the focus of an ongoing investigation.

Future directions

In order to cope more effectively with the information arising from the various genome projects, it is essential to reduce the manual burden inherent in our current database curation strategies and, where possible, increase levels of automation. To this end, developments are planned in a number of areas: e.g., we aim to (i) implement automated strategies for fingerprint derivation; (ii) design methods for automatic extraction of low-level annotations from the primary database; and ultimately, (iii) pool high-level documentations with those from PROSITE and Pfam, creating a central compendium of domain and family descriptions. This last will help to reduce duplication of effort in the rate-determining step of annotation, and aims to provide a one-stop shop for analysis of newly-determined sequences.

In the meantime, while largely-manual approaches are still in place, emphasis will continue to be placed on adding new families to PRINTS, rather than on routinely updating existing ones. The underlying philosophy here is to try to provide a more comprehensive diagnostic resource, with high-quality annotations, rather than simply to focus on providing an up-to-date look-up table of family membership (an impossible individual human task against the swelling tide of primary data).

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.

CONCLUSION

Secondary databases are an important part of the endeavour to harvest the abundant fruits of the various genome projects. The scope and subtlety of such resources make them powerful tools for diagnosing the relationships between sequences that underpin the inference of function. But none of these databases is an end in itself: none of the underlying analysis methods is yet infallible, and none of the resources is complete. But coupled with PROSITE, BLOCKS, Pfam, etc., PRINTS adds an important piece to the jigsaw in the challenging puzzle of sequence analysis.

ACKNOWLEDGEMENTS

We thank the authors of the database software and everyone who has contributed entries to the resource. PRINTS is built and

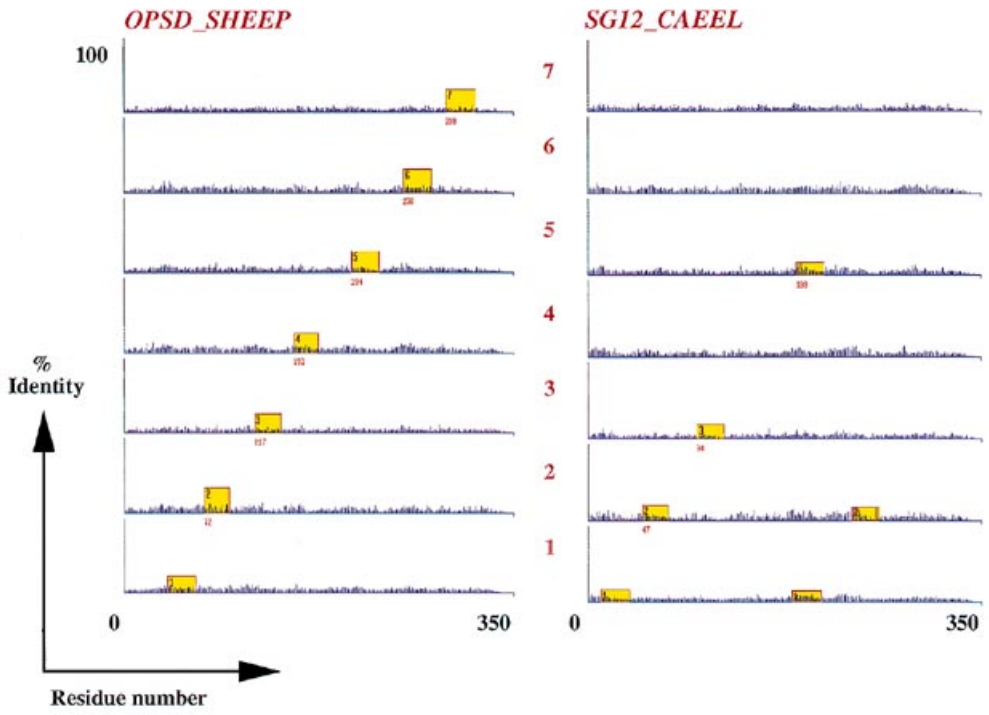


Figure 3. Graphical output returned by FingerPRINTScan. Within the profile, the horizontal axis represents the sequence, and the vertical axis the percentage score (identity) of each fingerprint element (0–100 per motif). Yellow blocks mark the positions of motif matches above a 15% threshold. The profiles depict rhodopsin-like GPCR fingerprints of ovine rhodopsin and of a *C.elegans* integral membrane protein. Blocks appearing in a systematic order along the length of the sequence and above the level of noise indicate matches with the constituent motifs. Ovine rhodopsin is a known true-positive family member, matching all seven transmembrane domains; the *C.elegans* sequence fails to make a complete match, but a relationship is apparent with the GPCR superfamily, as suggested by the correct sequence of matches with motifs 1–3 and 5. In the second profile, the two additional blocks highlight a degree of similarity between transmembrane domains 1 and 5, and between domains 2 and 6.

maintained at UCL with support from the Royal Society (TKA is a Royal Society University Research Fellow). PS is grateful to Astra Charnwood for a studentship. JS is grateful to Zeneca for a bioinformatics fellowship.The project benefits from use of the BBSRC SEQNET facility.

REFERENCES

1 Bairoch, A., Bucher, P. and Hofmann, K. (1997) *Nucleic Acids Res.*, **25**, 217–221
2 Henikoff, J.G., Pietrokovski, S. and Henikoff, S. (1997) *Nucleic Acids Res.*, **25**, 222–225 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 309–312].
3 Sonnhammer, E.L.L., Eddy, S.R. and Durbin, R. (1997) *Proteins Struct. Funct. Genet.*, **28**, 405–420.
4 Attwood, T.K., Beck, M.E., Bleasby, A.J., Degtyarenko, K., Michie, A.D. and Parry-Smith, D.J. (1997) *Nucleic Acids Res.*, **25**, 212–216.
5 Parry-Smith, D.J. and Attwood, T.K. (1992) *Comput. Applic. Biosci.*, **8**, 451–459.
6 Attwood, T.K. and Findlay, J.B.C. (1994) *Protein Engng*, **7**, 195–203.
7 Bleasby, A.J., Akrigg, D. and Attwood, T.K. (1994) *Nucleic Acids Res.*, **22**, 3574–3577.
8 Bairoch, A. and Apweiler, R. (1997) *Nucleic Acids Res.*, **25**, 31–36 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 38–42].
9 George, D.G., Dodson, R.J., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Sidman, K.E., Srinivasarao, G.Y., Yeh, L.S.L. *et al.* (1997) *Nucleic Acids Res.*, **25**, 24–27 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 27–32].
10 Benson, D.A, Boguski, M., Lipman, D.J. and Ostell, J. (1997) *Nucleic Acids Res.*, **25**, 1–6 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 1–7].
11 Pattabiraman, N., Namboodiri, K., Lowrey, A. and Gaber, B.P. (1990) *Protein Seq. Data Anal.*, **3**, 387–405.
12 Parry-Smith, D.J. and Attwood, T.K. (1991) *Comput. Applic. Biosci.*, **7**, 233–235.
13 Attwood, T.K., Avison, H., Beck, M.E., Bewley, M., Bleasby, A.J., Brewster, F., Cooper, P., Degtyarenko, K., Flower, D.R., Geddes, A.J. *et al.* (1997) *J. Chem. Info. Comput. Sci.*, **37**, 417–424.
14 Fabian, P., Murvai, J., Hatsagi, Z., Vlahovicek, K., Hegyi, H. and Pongor, S. (1997) *Nucleic Acids Res.*, **25**, 240–243.
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 Attwood, T.K. and Beck, M.E. (1994) *Protein Engng*, **7**, 841–848.
18 Attwood, T.K., Beck, M.E., Bleasby, A.J. and Parry-Smith, D.J. (1994) *Nucleic Acids Res.*, **22**, 3590–3596.
19 Michie, A.D., Jones, M.L. and Attwood, T.K. (1996) *Trends Biochem. Sci.*, **21**, 191.
20 Parry-Smith, D.J., Payne, A.W.R, Michie, A.D. and Attwood, T.K. (1997) *Gene Combis*, in Press.
21 Perkins, D.N. and Attwood, T.K. (1996) *Comput. Applic. Biosci.*, **12**, 89–94.
22 Flower, D.R., North, A.C.T. and Attwood, T.K. (1993) *Protein Sci.*, **2**, 753–761.
23 Attwood, T.K. and Findlay, J.B.C. (1993) *Protein Engng*, **6**, 167–176.
24 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D. (1990) *J. Mol. Biol.*, **215**, 403–410.
25 The Yeast Genome Directory (1997). *Nature* (supplement), 387.