Terri Attwood
is a Royal Society University Research Fellow and Professor of Bioinformatics at Manchester.

The PRINTS database: A resource for identification of protein families

Terri K. Attwood

Date received (in revised form): 10th June 2002

Abstract

The PRINTS database houses a collection of protein fingerprints, which may be used to assign family and functional attributes to uncharacterised sequences, such as those currently emanating from the various genome-sequencing projects. The April 2002 release includes 1,700 family fingerprints, encoding ~10,500 motifs, covering a range of globular and membrane proteins, modular polypeptides and so on. Fingerprints are groups of conserved motifs that, taken together, provide diagnostic protein family signatures. They derive much of their potency from the biological context afforded by matching motif neighbours; this makes them at once more flexible and powerful than single-motif approaches. The technique further departs from other pattern-matching methods by readily allowing the creation of fingerprints at superfamily-, family- and subfamily-specific levels, thereby allowing more fine-grained diagnoses. Here, we provide an overview of the method of protein fingerprinting and how the results of fingerprint analyses are used to build PRINTS and its relational cousin, PRINTS-S.

Keywords: *protein family, sequence alignment, similarity search, pattern recognition, function annotation*

INTRODUCTION

The first step in analysing a newly determined sequence usually involves trawling a sequence database with pairwise search tools such as BLAST[1] or FastA.[2] Such searches quickly reveal similarities between the query and a range of database sequences. The trick then lies in the reliable inference of homology (the presumption of divergent evolutionary descent) and hence of family ties and functional relationships. Ideally, a search output will show unequivocal similarity to a well-characterised protein over the full length of the query; at worst, it will reveal no significant hits; but the usual scenario is a list of weak matches to diverse proteins, many of them uncharacterised, some with dubious or contradictory annotations.[3]

Deciding how much functional annotation can legitimately be inherited by a query sequence and achieving consistent, reliable assignments can be a complicated process. As a result, in addition to routine searches of the sequence databases, it is now customary to extend search strategies to include a range of family or 'pattern' resources. These distil information within groups of related sequences into potent descriptors that aid diagnosis. In principle, searching family repositories is more powerful than sequence database searching because derived discriminators can detect weaker regions of similarity. Different analytical approaches have been used to create a bewildering array of discriminators, which are variously termed regular expressions, profiles, fingerprints, blocks, etc.[4,5] These different descriptors have been used to generate different family databases, which differ significantly in content. Here, we will describe the method that gives rise to the PRINTS database, whose current status we will review.

The database is accessible for BLAST, fingerprint and text searches.[6]

IDENTIFICATION OF PROTEIN FAMILIES

At the heart of the analysis methods that underpin family databases is the multiple sequence alignment. When building an

T. K. Attwood,
School of Biological Sciences &
Department of Computer Science,
The University of Manchester,
Oxford Road,
Manchester M13 9PT, UK

Tel: +44 (0) 161 275 5766
Fax: +44 (0) 161 275 5082
e-mail: attwood@bioinf.man.ac.uk

alignment, as more distantly related sequences are included, insertions are often required to bring equivalent parts of adjacent sequences into the correct register (see Figure 1). As a result of this gap-insertion process, islands of conservation emerge from a backdrop of mutational change. These conserved regions (typically around 10–20 residues in length) tend to correspond to the core structural or functional elements of the protein, and are commonly termed motifs.

Diagnostic opportunity

Motif

Several techniques have evolved to exploit the conservation encoded in alignments, all of which involve the derivation of some kind of discriminatory representation of the conserved elements. Broadly, these can be categorised into three main approaches: those that use single motifs to encapsulate the most conserved feature (or features) of an alignment; those that exploit multiple motifs to build a diagnostic signature of family membership; and those that encode complete domains, including both conserved regions and the gapped areas between them (an overview of the methods and the databases they underpin is shown in Figure 1). Each of these methods has different diagnostic strengths and weaknesses, and consequently optimum areas of application – none should be regarded as the best, as each offers a different perspective and a different (often complementary) diagnostic opportunity. We will now take a closer look at one of these approaches – namely protein fingerprinting.

Fingerprint

Domain

PROTEIN FINGERPRINTING

Within a multiple 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 or fingerprint. 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: eg 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 those expected of true neighbouring motifs, as illustrated in

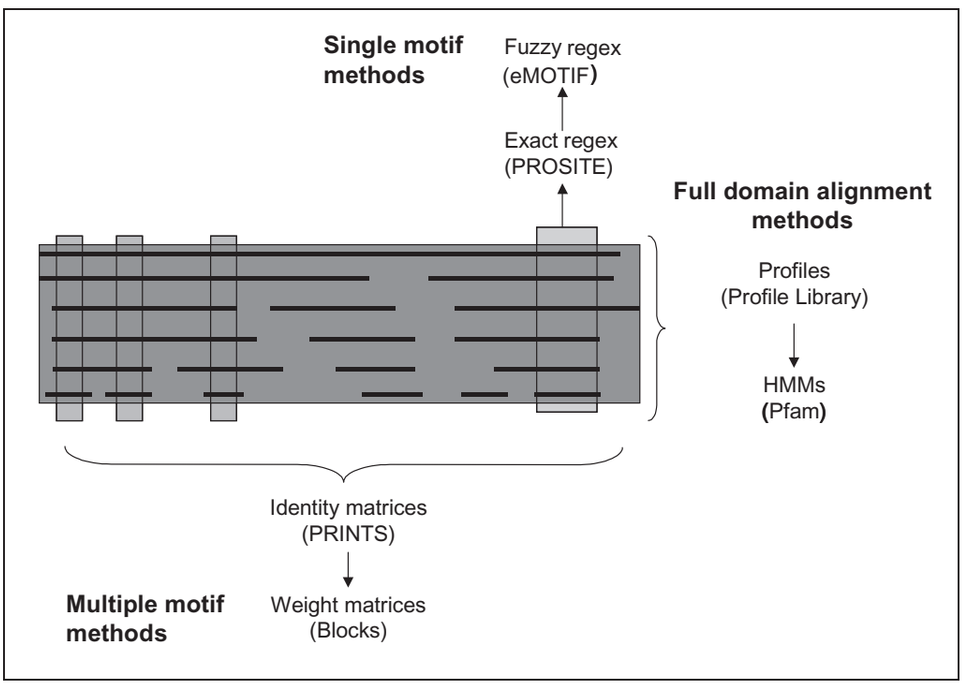


Figure 1: Overview of the three main sequence analysis approaches and the databases to which they give rise: single motif methods that exploit regular expressions (regexs) underpin PROSITE and eMOTIF; multiple motif approaches that use either identity or weight matrices are the basis of PRINTS and Blocks; and full-domain methods that exploit either absolute or probabilistic scores underpin Profiles and Pfam

Biological context

Familial hierarchies

Signature GPCR

Figure 2. The potency of fingerprints thus derives from the mutual context provided by motif neighbours – the more motifs it contains, the better able it is to identify distant relatives, even when parts of the signature are absent; conversely, the fewer the motifs, the poorer its diagnostic performance. Fingerprints with only two motifs are diagnostically little better than single-motifs, and are therefore more likely to make false-positive matches. Overall, fingerprinting is thus more flexible and powerful than single-motif approaches – the ability to tolerate mismatches, both at the level of individual residues within motifs, and at the level of motifs within the complete signature, renders it a powerful diagnostic approach.

The technique further departs from other pattern-matching methods by readily allowing the creation of fingerprints at superfamily-, family- and subfamily-specific levels. This is possible because the approach is manual and allows one to focus not only on regions of shared similarity (such as those that characterise superfamilies), but also on the regions of *difference* (such as those that resolve subfamilies from closely related siblings within a family, and/or that distinguish families from their parent superfamilies). This is crucial because it is the subtle differences between close relatives that largely determine their functional specificities. This hierarchical approach has been used to analyse a range of proteins, especially those of pharmaceutical interest, eg to resolve G protein-coupled receptor (GPCR) superfamilies into their constituent families and receptor subtypes,[7–9] and to finely classify a variety of channel proteins, transporters and enzymes. Fingerprinting thus provides a useful complement to profile-based and other 'catch-all' methods, which tend to specialise in the diagnosis of superfamilies.

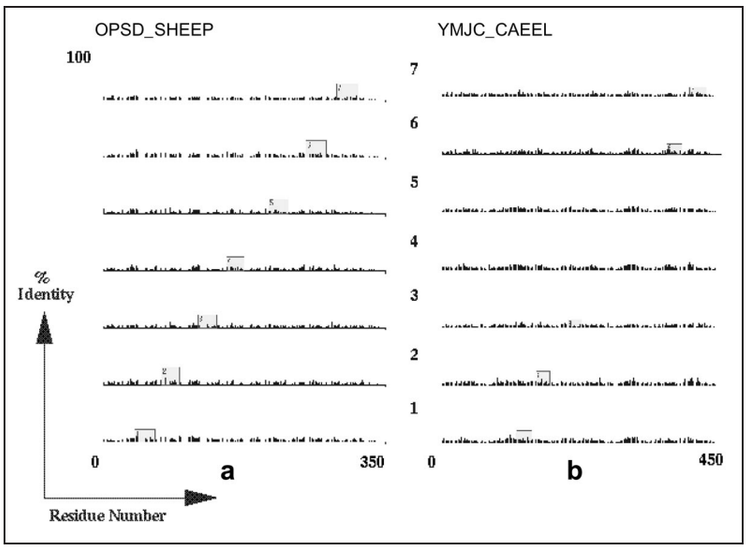


Figure 2: Graphical output from fingerprint searches illustrating both full and partial matches. Within the graphs, the *x*-axis represents the sequence and the *y*-axis the percentage score (identity) of each fingerprint element (0–100 per motif). Filled blocks mark the positions of motif matches above a 20 per cent threshold. Blocks appearing in a systematic order along the length of the sequence and above the level of noise indicate matches with the constituent motifs. Unequivocal family membership is denoted in (a) by strong matches to each of the seven motifs of the GPCR superfamily fingerprint. By contrast, (b) shows a partial match that exhibits characteristics, such as motifs being in the correct order and having acceptable inter-motif distances, that allow us to infer with a degree of confidence that it is a related family member, even though it fails to make significant matches with three of the seven GPCR superfamily motifs

THE FINGERPRINT METHOD

In detail, the method involves manual creation of a seed alignment, and location and excision of conserved motifs for searching the source database (a SWISS-PROT/TrEMBL[10] composite from which fragments have been extracted) – for historical reasons, the motifs may number up to a maximum of 15, with maximum length of 30 residues. The database-scanning algorithm converts the excised motifs into a series of frequency (identity) matrices – in other words, no mutation or other similarity data are used to weight the motifs. This is because the generation of fingerprints must be a selective process, to avoid being corrupted by spurious matches, and identity matrices are more stringent and produce cleaner discrimination than do similarity matrices, which are inherently noisy.[11] The scoring process uses a sliding-window approach, whereby each

Frequency matrix

motif is scanned across each database sequence in turn. For each position of the window (which, by definition, is the width of the motif), the algorithm simply sums the residue scores with reference to the motif frequency matrix. The best match is achieved when a position is found in the sequence where most of the residues within the sliding window match high-scoring terms in the frequency matrix.

Distance constraint

For each motif, results are stored in a hit-list that is rank-ordered by score. Diagnostic performance is enhanced by iterative database scanning: at each step, hit-lists are compared to determine which sequences have matched all the motifs in the fingerprint; if there are more matches than were in the initial alignment, the additional information from these new sequences is added to the motifs, and the database is searched again. The motifs therefore grow and mature with each database pass, as more sequences are matched and assimilated into the process. The procedure terminates when no more new sequences that match all the motifs can be identified between successive database scans, ie when the scans have converged.

Iterative search

Manual annotation

Convergence

An important point to note about the motif-matching process is that, unlike other methods, fingerprinting does not use an absolute scoring threshold to determine whether a match has been made or whether it is significant. During the iterative scanning procedure, the default hit-list length is 2,000 hits, but this can be varied by the user, depending on family size – if a family is thought to contain 1,000–2,000 members, hit-lists of 2,000 will clearly not be adequate. When the lists are compared to ascertain which sequences have matched all the motifs, the default comparison length is 300 (in other words, the top 300 hits are sliced off each hit-list and compared, irrespective of individual match scores). Thus the process only requires that a sequence appears within the given sample length, and makes no assumptions about score significance. However, the user may also vary this parameter – if too much noise appears in the result, the sample length can be reduced (eg by top-slicing only the first 100 hits); or, if true matches appear to have been missed, the sample length can be increased (eg to include the top 500 hits, or whatever). The approach is thus flexible with regard to score, the only rule being that the motifs must match in the correct order. Results can also be fine-tuned by imposing a distance constraint (ie that motif intervals should be consistent with those normally expected of true neighbouring motifs), but this option is usually used only as a cosmetic step to remove noise once the scans have converged – this avoids true matches being thrown away early in the process, which may later turn out to be outliers.

Once the scanning process has converged, and the results fine-tuned in the manner described above, they are then annotated manually (with biological information and literature, database cross-references, etc.) prior to inclusion in the database.[12] The complete fingerprint process is summarised in Figure 3.

DATABASE FORMAT

PRINTS is built as single ASCII (text) file – see Figure 4. The contents are separated into specific fields, relating to

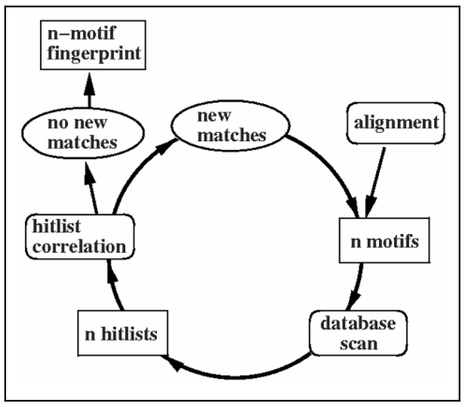


Figure 3: Overview of the iterative process by which fingerprints are generated from seed sequence alignments prior to annotation and deposition in PRINTS

```
gc; PRION
gx; PR00341
gn; COMPOUND(8)
ga; 19-OCT-1992; UPDATE 07-JUN-1999
gt; Prion protein signature
gp; INTERPRO; IPR000817
gp; PROSITE; PS00291 PRION_1; PS00706 PRION_2
gp; PFAM; PF00377 prion
bb;
gr; 1. STAHL, N. AND PRUSINER, S.B.
gr; Prions and prion proteins.
gr; FASEB J. 5 2799-2807 (1991).
gr;
gr; 2. BRUNORI, M., CHIARA SILVESTRINI, M. AND POCCHIARI, M.
gr; The scrapie agent and the prion hypothesis.
gr; TRENDS BIOCHEM.SCI. 13 309-313 (1988).
gr;
gr; 3. PRUSINER, S.B.
gr; Scrapie prions.
gr; ANNU.REV.MICROBIOL. 43 345-374 (1989).
bb;
bb;
gd; Prion protein (PrP) is a small glycoprotein found in high quantity in the brain of animals infected with
gd; certain degenerative neurological diseases, such as sheep scrapie and bovine spongiform encephalopathy (BSE),
gd; and the human dementias Creutzfeldt-Jacob disease (CJD) and Gerstmann-Straussler syndrome (GSS). PrP is
gd; encoded in the host genome and is expressed both in normal and infected cells. During infection, however, the
gd; PrP molecules become altered and polymerise, yielding fibrils of modified PrP protein.
gd;
gd; PrP molecules have been found on the outer surface of plasma membranes of nerve cells, to which they are
gd; anchored through a covalent-linked glycolipid, suggesting a role as a membrane receptor. PrP is also expressed
gd; in other tissues, indicating that it may have different functions depending on its location. The primary
gd; sequences of PrP's from different sources are highly similar: all bear an N-terminal domain containing multiple
gd; tandem repeats of a Pro/Gly rich octapeptide; sites of Asn-linked glycosylation; an essential disulphide bond;
gd; and 3 hydrophobic segments. These sequences show some similarity to a chicken glycoprotein, thought to be an
gd; acetylcholine receptor-inducing activity (ARIA) molecule. It has been suggested gd; that changes in the octa-
gd; peptide repeat region may indicate a predisposition to disease, but it is not known for gd; certain whether the
gd; repeat can be used as a fingerprint to indicate susceptibility.
gd;
gd; PRION is an 8-element fingerprint that provides a signature for the prion proteins. The fingerprint was derived
gd; from an initial alignment of 5 sequences: the motifs were drawn from conserved regions spanning virtually the
gd; full alignment length, including the 3 hydrophobic domains and the octapeptide repeats (WGQPHGGG). Two
gd; iterations on OWL18.0 were required to reach convergence, at which point a true set comprising 9 sequences was
gd; identified. Several partial matches were also found: these include a fragment (PRIO_RAT) lacking part of the
gd; sequence bearing the first motif, and the PrP homologue found in chicken - this matches well with only 2 of the
gd; 3 hydrophobic motifs (1 and 5) and one of the other conserved regions (6), but has an N-terminal signature
gd; based on a sextapeptide repeat (YPHNPG) rather than the characteristic PrP octapeptide.
gd;
gd; An update on SPTR37_9f identified a true set of 37 sequences, and 1 partial match.
bb;
bb;
si; SUMMARY INFORMATION
si; -------------------
sd;   37 codes involving  8 elements
sd;    0 codes involving  7 elements
sd;    0 codes involving  6 elements
sd;    0 codes involving  5 elements
sd;    0 codes involving  4 elements
sd;    1 codes involving  3 elements
sd;    0 codes involving  2 elements
bb;
bb;
ci; COMPOSITE FINGERPRINT INDEX
ci; ---------------------------
cr;
cd;  8|  37   37   37   37   37   37   37   37
cd;  7|   0    0    0    0    0    0    0    0
cd;  6|   0    0    0    0    0    0    0    0
cd;  5|   0    0    0    0    0    0    0    0
cd;  4|   0    0    0    0    0    0    0    0
cd;  3|   1    0    0    0    1    1    0    0
cd;  2|   0    0    0    0    0    0    0    0
cd; --+----------------------------------------
cd;   |   1    2    3    4    5    6    7    8
bb;
bb;
tp; PRIO_COLGU     PRIO_MACFA     PRIO_CEREL     PRIO_ODOHE
tp; PRIO_GORGO     PRIO_PANTR     PRIO_HUMAN     O46648
tp; PRIO_SHEEP     PRIO_CALJA     PRIO_BOVIN     PRP2_BOVIN
tp; PRIO_ATEPA     PRIO_SAISC     PRIO_PREFR     PRIO_PONPY
tp; O75942         PRIO_CAPHI     PRIO_CEBAP     PRIO_CAMDR
tp; PRIO_FELCA     PRP1_TRAST     PRIO_RABIT     PRP2_TRAST
tp; PRIO_PIG       PRIO_CANFA     PRIO_CRIGR     PRIO_CRIMI
tp; Q15216         PRIO_RAT       PRIO_CERAE     PRIO_MUSPF
tp; PRIO_MUSVI     PRIO_MESAU     PRIO_MOUSE     O46593
tp; PRIO_TRIVU
bb;
sn; Codes involving 3 elements
st; PRIO_CHICK
bb;
bb;
```

Figure 4: Sample data from PRINTS, showing the fingerprint for the prion protein family. For convenience, only the first motif is depicted. The two-letter code in the left-hand margin separates the information into specific fields (relating to text, references, motifs, etc.), which allows indexing of the data for rapid querying

```
tt; PRIO_COLGU       MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) - COLOBUS GUEREZA
tt; PRIO_MACFA       MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) - MACACA FASCICULA
tt; PRIO_CEREL       MAJOR PRION PROTEIN PRECURSOR (PRP) - CERVUS ELAPHUS (RED DEER)
tt; PRIO_ODOHE       MAJOR PRION PROTEIN PRECURSOR (PRP) - ODOCOILEUS HEMIONUS (MULE DEER) (BLACK-
tt; PRIO_GORGO       MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) - GORILLA GORILLA
tt; PRIO_PANTR       MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) - PAN TROGLODYTES
tt; PRIO_HUMAN       MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) (ASCR) - HOMO SAPI
.
.
.
tt;
tt; PRIO_CHICK       MAJOR PRION PROTEIN HOMOLOG PRECURSOR (PR-LP) (ACETYLCHOLINE RECEPTOR-INDUCIN
bb;
bb;
sh; SCAN HISTORY
sh; ------------
dn; OWL18_0     2    30 NSINGLE
dn; OWL19_1     1    30 NSINGLE
dn; OWL26_0     1   160 NSINGLE
dn; OWL29_1     1   150 NSINGLE
dn; SPTR37_9f   2   134 NSINGLE
bb;
bb;
im; INITIAL MOTIF-SETS
im; ------------------
ic; PRION1
il; 16
it; Prion protein motif I - 1
id; WMLVLFVATWSDLGLC                PRIO_HUMAN      7     7
id; WILVLFVAMWSDVGLC                PRIO_BOVIN      9     9
id; WILVLFVAMWSDVGLC                PRIO_SHEEP      9     9
id; WILVLFVAMWSDVGLC                PRIP_BOVIN      9     9
id; WLLALFVAMWTDVGLC                PRIO_MESAU      7     7
id; WLLALFVTMWTDVGLC                PRIO_MOUSE      7     7
bb;
```

Figure 4: (*continued*)

Database indices

Accession number

Partial match

Time-positive match

general information, bibliographical 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] In the general field at the top of the file, each entry is assigned a code by which it can be identified, and an accession number (which takes the form PR00000). This is followed by a description of the type of entry – the term 'compound' indicates that the fingerprint contains several elements, the number of constituent motifs being indicated in parentheses. Details of the creation and latest update information are then given, followed by a descriptive title, and cross-references to entries in a variety of other databases (InterPro,[14] PDB,[15] etc.). A list of bibliographical references is then provided – this relates to a detailed abstract of the family that describes its function and structure (where known), its disease associations, evolutionary relationships and so on. Every abstract also contains a technical description of how the fingerprint was derived.

Fingerprint diagnostic performance is indicated via a summary that lists how many sequences matched all the motifs and how many made only partial matches (ie failed to match one or more motifs) – the fewer the partial matches, the better the fingerprint. The table that follows the summary breaks down this result to indicate how well individual motifs have performed, from which it is possible to deduce which motifs are missing from any partial matches.

After the summary are listed the protein identification codes of all full and partial true- and false-positive matches, followed by their database titles. The scan history then indicates which version of the source database was used to derive the fingerprint, and on which versions it has been updated, how many iterations were required, what hit-list length was used, and the scanning method employed: the default scanning method is termed NSINGLE.[11]

The final field relates to the motifs themselves, listing both the initial and final motifs, the motif lengths and their starting locations. The intervals between adjacent motifs are also provided. Each motif is assigned a discrete code, ie the general identification code with the number of that particular motif appended.

For convenience, only initial motif 1 (PRIONl) is shown in Figure 4.

CURRENT RELEASE

PRINTS is released in major and minor versions: minor releases reflect updates, bringing the contents in line with the current version of the source database; major releases denote the addition of new material to the resource. Major releases are made quarterly, each release including 50 new families.

Gribskov profiles

BLOSUM matrices

To date, 1,700 fingerprints, encoding 10,342 motifs (version 34.0, April 2002), have been developed and deposited in PRINTS, making it the most comprehensive fully manually annotated protein family database available. Nevertheless, overall the database is still small relative to the number of protein families that exist, largely because the detailed documentation of entries is extremely time-consuming. However, the extent of manually crafted annotations sets it apart from the growing number of automatically derived resources, for which there is little or no biological documentation and/or result validation, and in which family groupings may change between database releases.

P-/E-values

SEARCH TOOLS

There are two main tools available for searching PRINTS: a BLAST server, which allows similarity searches against *sequences* matched in the current version of the database,[16] and the FingerPRINTScan suite,[17] which allows sequence searches against *fingerprints* contained in the current release. This is an important distinction, as the different search tools offer fundamentally different perspectives on sequence similarity: BLAST identifies generic similarities between sequences within a family and cannot recognise individual family traits, while fingerprints pinpoint the subtle (often structural or functional) differences that differentiate closely related family members. FingerPRINTScan thus affords greater specificity than the BLAST implementation and highlights the danger of relying on top BLAST hits to provide reliable functional annotation.[16]

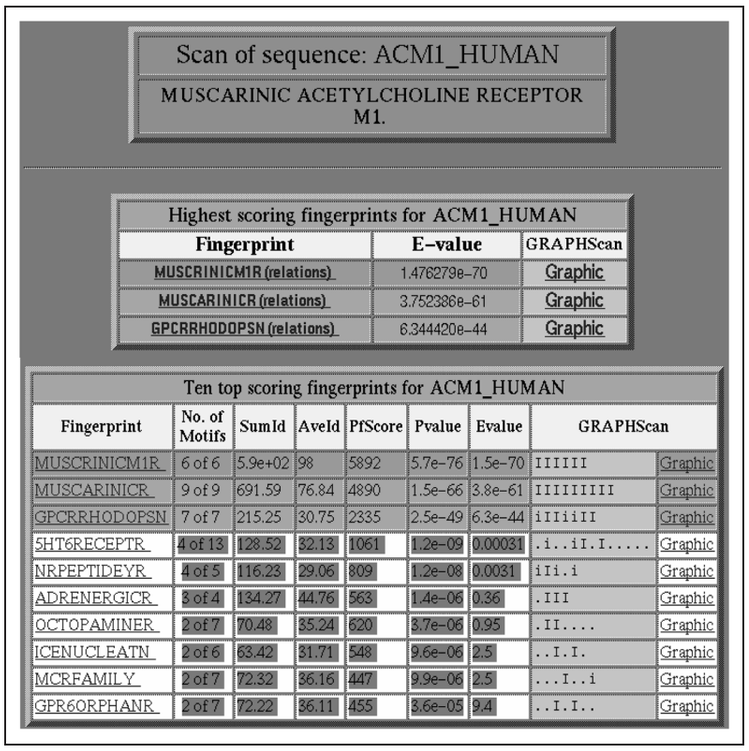
BLAST server

Fingerprint search

Modular/mosaic proteins

By contrast with the scanning method used to create fingerprints, which is highly selective, the algorithm designed to scan query sequences against PRINTS is more permissive, effectively allowing the user to cast a wider net and thereby maximise the number of potential matches. A sliding-window approach is once again used, but individual motifs are converted to Gribskov-type profiles,[18] without the inclusion of gaps, and residue scores are calculated with reference to the BLOSUM series of matrices.[19] As each motif is scanned across the query sequence, probability (*P*)-values are derived for each match; the algorithm then seeks the best combined set of matches that occur in the correct order with appropriate distances between them (true motif intervals are stored in PRINTS, from which the algorithm calculates maximum and minimum values). The overall significance of the result is expressed as the product of the *P*-values of each of the individual motifs, which is also expressed as an expect (*E*)-value.

A sample output is shown in Figure 5, which illustrates the result of searching PRINTS with the query sequence ACM1_HUMAN using default parameters. The output is returned on three levels: first, the program's 'best guess' at the correct fingerprint; next, a table of the 10 top-scoring fingerprints; and finally, the top 10 hits listed in greater detail, including the constituent motifs. Where multiple fingerprints are matched above the default *E*-value cut-off (0.0001), each of the results is reported in the 'best guess' table. This allows diagnosis both of family hierarchies, from superfamily down to subfamily, and also of modular and mosaic proteins, where multiple domains occur in the same sequence. In this example, the 'best guess' reveals a three-tiered diagnosis, indicating the sequence to be (i) a member of the rhodopsin-like GPCR superfamily; (ii) a member of the muscarinic receptor



Scan of sequence: ACM1_HUMAN

MUSCARINIC ACETYLCHOLINE RECEPTOR M1.

Highest scoring fingerprints for ACM1_HUMAN

Fingerprint	E–value	GRAPHScan
MUSCRINICM1R (relations)	1.476279e–70	Graphic
MUSCARINICR (relations)	3.752386e–61	Graphic
GPCRRHODOPSN (relations)	6.344420e–44	Graphic

Ten top scoring fingerprints for ACM1_HUMAN

Fingerprint	No. of Motifs	SumId	AveId	PfScore	Pvalue	Evalue	GRAPHScan	
MUSCRINICM1R	6 of 6	5.9e+02	98	5892	5.7e–76	1.5e–70	IIIIII	Graphic
MUSCARINICR	9 of 9	691.59	76.84	4890	1.5e–66	3.8e–61	IIIIIIIII	Graphic
GPCRRHODOPSN	7 of 7	215.25	30.75	2335	2.5e–49	6.3e–44	iIIiiII	Graphic
5HT6RECEPTR	4 of 13	128.52	32.13	1061	1.2e–09	0.00031	.i..iI.I.....	Graphic
NRPEPTIDEYR	4 of 5	116.23	29.06	809	1.2e–08	0.0031	iIi.i	Graphic
ADRENERGICR	3 of 4	134.27	44.76	563	1.4e–06	0.36	.III	Graphic
OCTOPAMINER	2 of 7	70.48	35.24	620	3.7e–06	0.95	.II....	Graphic
ICENUCLEATN	2 of 6	63.42	31.71	548	9.6e–06	2.5	..I.I.	Graphic
MCRFAMILY	2 of 7	72.32	36.16	447	9.9e–06	2.5	...I..i	Graphic
GPR60RPHANR	2 of 7	72.22	36.11	455	3.6e–05	9.4	..I.I..	Graphic

Figure 5: Hierarchical diagnosis returned from searching PRINTS with the human muscarinic acetylcholine M_1 receptor, ACM1_HUMAN. Three fingerprints have been matched, indicating the sequence to be a member of the rhodopsin-like GPCR superfamily (GPCRRHODOPSN), belonging to the muscarinic receptor family (MUSCARINICR), being specifically an M_1 receptor subtype (MUSCRINICM1R). The *E*-values in the centre of the table provide the measure of confidence in the result – here, all matches are statistically significant (ie above the threshold value of 10^{-4})

Protein clan

family; and specifically (iii) an M_1 receptor subtype.

PRINTS' RELATIONAL COUSIN, PRINTS-S

With the continued growth of the database, maintenance of the PRINTS flat-file was becoming increasingly inefficient and error-prone. An important development was therefore to migrate the resource to the PostgreSQL relational database management system (DBMS). The 'streamlined' version, termed PRINTS-S,[20] reduces redundancy, maintains consistency and facilitates routine maintenance. It also permits more complex queries of the underlying data, and allows the support of new display and flat-file formats. PRINTS-S is accessible for interactive use via the Web. The interface allows both strict keyword searching and more powerful queries using a combination of regular expressions and logical operators.

Relational database

Midnight zone

A valuable attribute of PRINTS-S is the ability to model relationships explicitly by defining parent–child and sibling relations within, and *implied* by, the PRINTS family hierarchy (see Figure 6).[21] This means, for example, that members of a clan (a group of families for which there are indications of an evolutionary relationship, but between which there is no statistically significant sequence similarity) can nevertheless be linked. Thus it is possible to transcend relationships evident at the sequence level and gain structural insights from a realm beyond the theoretical limits of conventional sequence analysis tools (this is the so-called 'midnight zone', the region of identity where sequence comparisons fail completely to detect structural similarities[22]).

Parent-child relationships

As an illustration, consider the relationships encoded in the database for the rhodopsin-like GPCRs shown in Figure 6(a). The FingerPRINTScan suite has been modified to exploit these relationships in such a way that when we search the database with a query sequence, all child/sibling/parent/grandparent relations between matched fingerprints are revealed. Take, for

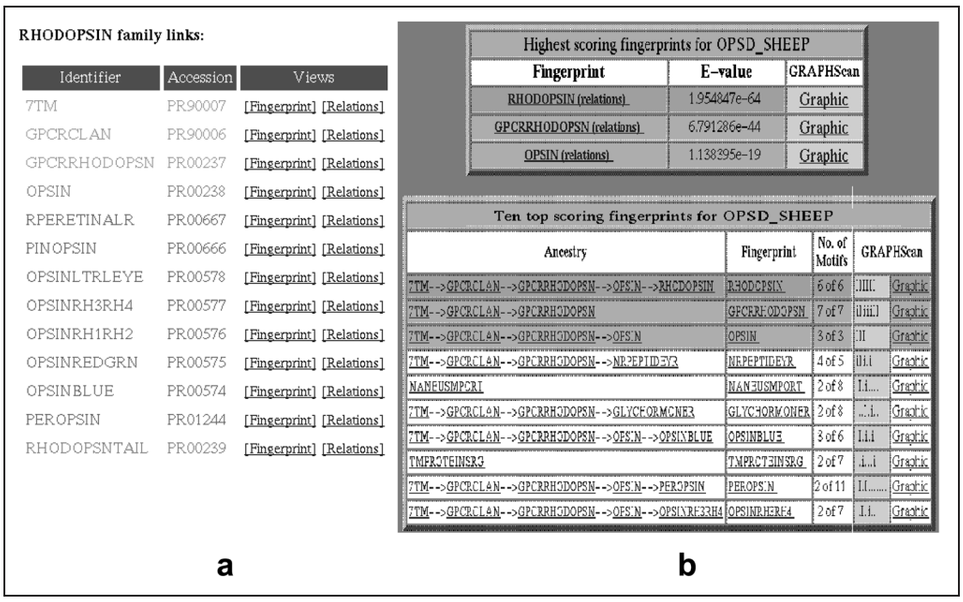


Figure 6: (a) The rhodopsin family hierarchy depicted by PRINTS-S. The hierarchy is colour-coded via the Web interface. Although not obvious here, RHODOPSNTAIL is a child; RPERETINALR, PINOPSIN, OPSINLTRLEYE, OPSINRH3RH4, OPSINRH1RH2, OPSINREDGRN, OPSINBLUE and PEROPSIN are siblings; OPSIN is the parent; and GPCRRHODOPSIN, GPCRCLAN and 7TM are grandparents and great-grandparents. (b) Result of searching PRINTS-S with the sequence of ovine rhodopsin using FingerPRINTScan. The table shows the top 10 matches (significant matches are highlighted), and traces the relationships between each matched fingerprint from its position in the familial hierarchy back to its most distant ancestor. Here, each rhodopsin-like GPCR match can be traced back through its parent superfamily, through the ancestral GPCR clan, ultimately to a presumed '7TM' architectural predecessor

example, the result of searching with the sequence of ovine rhodopsin, shown in Figure 6(b). RHODOPSIN, GPCRRHODOPSN and OPSIN are the only fingerprint matches with significant *E*-values highlighted in the table. For each of these matches, the relationships between them are traced back through the family hierarchy to the most remote putative ancestor. Thus, we see that RHODOPSIN is a child of the OPSIN family, which is a child of GPCRRHODOPSN (the rhodopsin-like GPCR superfamily), whose parent is the GPCR clan (which includes the secretin-like receptors, metabotropic receptors, etc.), which is derived from a putative '7TM' architectural ancestor. Such an 'ancestral perspective' is only possible because PRINTS-S models the biological associations between families within an internal relational structure, allowing a hierarchical representation of connections between database entries, including those outside the realm of sequence similarity searches.[21]

Local alignment

Architectural ancestor

Blocks

RELATED DATABASE DEVELOPMENTS

A particular strength of PRINTS is that its motifs are stored in the form of un-gapped, local alignments. An important consequence of storing the motifs in this 'raw' form is that, unlike with regular expressions or other abstractions, no sequence information is lost. Different scoring methods may thus be superposed onto the motifs, conferring different scoring potentials, and hence different perspectives, on the same data. Thus, a Blocks-format version of the resource that exploits Blocks scoring methods is

available at the Fred Hutchinson Cancer Research Center.[23] In addition, the eMOTIF database at Stanford overlays a permissive regular expression approach over PRINTS' multiply-aligned motifs, offering different levels of stringency from which to infer the significance of matches.[24] Because the Blocks and eMOTIF databases are derived automatically, their entries are not annotated, but links are made to the corresponding PRINTS files.

eMOTIF

Another landmark in the evolution of PRINTS builds on a decision made in 1991 to integrate it with PROSITE,[25] in order to create a unified protein family resource. This project has now been realised on a much larger scale, initially in the form of an international consortium including Profiles,[25] Pfam[26] and ProDom;[27] more recently, a number of other partners have entered the collaboration. This initiative, known as InterPro,[14] which primarily exploits the detailed family annotations provided by PROSITE and PRINTS, aims both to reduce duplication of effort in the laborious, bottle-necking process of annotation, and to facilitate communication between disparate resources. A particular strength of InterPro is the ability to compare results of simultaneous searches across all database partners, as shown in Figure 7. The graphical result returned by the search nicely illustrates the difference between the various motif- and domain-based approaches: in the example shown, it is evident how small is the region encoded by the regular expression; by contrast, both the profile and HMM span almost the complete sequence; similarly, the fingerprints are drawn from conserved regions spanning virtually the full sequence, but this method alone exploits groups of motifs that differentiate between regions of sequence that characterise the superfamily and those that characterise the family and subfamily, thereby offering important structural and functional insights.

Functional insight

A more recent development is a pilot project to provide an automatic supplement to PRINTS, termed prePRINTS. This exploits an automatic pipeline for sequence alignment, motif detection, iterative database searching and annotation. Interactive versions of parts of the pipeline are also being developed: (i) to allow users to create their own fingerprints for use in conjunction with FingerPRINTScan; and (ii) to generate annotation for groups of user-specified sequences – this is PRECIS (Protein Reports Engineered from Concise Information in SWISS-PROT).[28] prePRINTS and its associated tools will ultimately help to increase the family coverage of PRINTS and so improve its effectiveness as a sequence analysis tool.

prePRINTS

InterPro

Web search

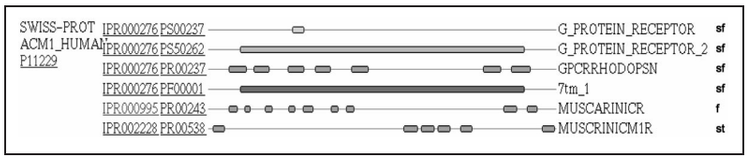


Figure 7: Graphical result from an InterPro search illustrating the difference between the various motif- and domain-based approaches: the regex encodes a single short motif (line 1), whereas the profile (line 2) and hidden Markov model (line 4) span almost the complete sequence. By contrast, fingerprints (lines 3, 5, 6) encode groups of motifs that differentiate regions of sequence that characterise the superfamily (sf) and those that typify the family (f) and receptor subtype (st). It is evident from this result that while PROSITE and Pfam furnish only superfamily diagnoses, PRINTS provides a more fine-grained result, thereby offering important structural and functional insights not apparent from the other methods.

AVAILABILITY

For local installation, PRINTS flat-files may be retrieved directly from the anonymous-ftp servers at the University of Manchester,[29] HGMP-RC,[30] EBI,[31] EMBL[32] and NCBI.[33] The database may be searched or queried via the Web – Table 1 summarises the locations of some of the PRINTS search tools and related resources.

CONCLUSION

Pattern databases provide powerful tools for analysing uncharacterised sequence data, in particular by placing individual sequences in a family context for a more

informed assessment of function than is possible with conventional pairwise searches. While there is some overlap between them, the contents of the family databases differ. It is therefore good practice to search all available repositories, to ensure that one's analysis is as comprehensive as possible and that it takes advantage of a variety of search methods. Where there is consensus, diagnoses can be made with greater confidence.

Unfortunately, creating and annotating family discriminators is time-consuming, so the databases have not kept pace with the deluge of sequence data, and PRINTS is no exception. Nevertheless, it is an evolving resource and the new developments help to increase its utility as a tool for sequence analysis. In addition, PRINTS-S sheds light on evolutionary relationships between families that were formerly hidden in PRINTS. Together, PRINTS and PRINTS-S are thus complementary tools that facilitate genome annotation, and add greater depth to sequence analyses by offering both unique hierarchical diagnoses and new ancestral perspectives on protein family relationships.

Genome annotation

Acknowledgments
I am grateful to the Royal Society for a University Research Fellowship and to the seq group.

Table 1: Accessing PRINTS and its related search tools and resources

PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
Current contents	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/printscontents.html
Keyword search	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/QuizPRINTS.html
FingerPRINTScan	http://www.bioinf.man.ac.uk/dbbrowser/fingerPRINTScan/
FingerPRINTScanfam	http://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/fingerPRINTScan/muppet/FPScan_fam.cgi/
PRINTS BLAST	http://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/PRINTS/printsBLAST.cgi
PRINTS-S	http://www.bioinf.man.ac.uk/dbbrowser/sprint/
Blocks-format-PRINTS	http://www.blocks.fhcrc.org/blocks/blocks_search.html
eMOTIF	http://motif.stanford.edu/emotif/
InterPro	http://www.ebi.ac.uk/interpro/scan.html/
prePRINTS	http://www.bioinf.man.ac.uk/dbbrowser/prePRINTS/
PRECIS	http://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/precis/precis.cgi

References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res.,* Vol. 25(17), pp. 3389–3402.
2. Pearson, W. R. (1998), 'Empirical statistical estimates for sequence similarity searches', *J. Mol. Biol.*, Vol. 276(1), pp. 71–84.
3. Hofmann, K. (1998), 'Protein classification and functional assignment', in 'Trends Guide to Bioinformatics', Elsevier Science, pp. 18–21.
4. Attwood, T. K. (1997), 'Exploring the language of bioinformatics', in Stanbury, H. Ed., 'Oxford Dictionary of Biochemistry and Molecular Biology', Oxford University Press, Oxford, UK, pp. 715–723.
5 . Attwood, T. K. (2000), 'The role of pattern databases in sequence analysis', *Briefings in Bioinformatics*, Vol. 1, pp. 45–59.
6. URL: http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
7. Attwood, T. K. and Findlay, J. B. C. (1994), 'Fingerprinting G protein-coupled receptors', *Protein Eng.*, Vol. 7(3), pp. 195–203.
8. Attwood, T. K. (2001), 'A compendium of specific motifs for diagnosing GPCR subtypes', *Trends Pharmacol.Sci.*, Vol. 22(4), pp. 162–165.
9. Attwood, T. K., Croning, M. D. R. and Gaulton, A. (2001), 'Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors', *Protein Eng.*, Vol. 15(1), pp. 7–12.
10. Bairoch, A. and Apweiler, R. (1999), 'The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000', *Nucleic Acids Res.,* Vol. 28(1), pp. 45–48.
11. Parry-Smith, D. J. and Attwood, T. K. (1992), 'ADSP – A new package for computational sequence analysis', *CABIOS,* Vol. 8(5), pp. 451–459.
12. Attwood, T. K. and Beck, M. E. (1994), 'PRINTS – a protein motif fingerprint database', *Protein Eng.*, Vol. 7(7), pp. 841–848.
13. Attwood, T. K., Avison, H., Beck, M. E. *et al.* (1997), 'The PRINTS database of protein fingerprints: A novel information resource for computational molecular biology', *J. Chem. Inf. Comp. Sci.*, Vol. 37, pp. 417–424.
14. Apweiler, R. Attwood, T. K., Bairoch, A. *et al.* (2001), 'The InterPro database, an integrated documentation resource for protein families, domains and functional sites', *Nucleic Acids Res.*, Vol. 29(1), pp. 37–40.
15. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B. *et al.* (1997), 'The protein data bank: A computer based archival file for

macromolecular structures', *J. Mol. Biol.,* Vol. 112, pp. 535–542.

16. Wright, W. and Attwood, T. K. (1999), 'BLAST PRINTS – alternative perspectives on sequence similarity', *Bioinformatics*, Vol. 15(6), pp. 523–524.
17. Scordis, P., Flower, D. R. and Attwood, T. K. (1999), 'FingerPRINTScan: intelligent searching of the PRINTS motif database', *Bioinformatics*, Vol. 15(10), pp. 799–806.
18. Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987), 'Profile analysis: detection of distantly related proteins', *Proc. Natl Acad. Sci. USA*, Vol. 84(13), pp. 4355–4358.
19. Henikoff, J. G. and Henikoff, S. (1992), 'Amino acid substitution matrices from protein blocks', *Proc. Natl Acad. Sci. USA*, Vol. 89, pp. 10915–10919.
20. Attwood, T. K., Croning, M. D. R., Flower, D. R. *et al.* (2000), 'PRINTS-S: the database formerly known as PRINTS', *Nucleic Acids Res.*, Vol. 28(1), pp. 225–227.
21. Attwood, T. K., Blythe, M. J., Flower, D. R. *et al.* (2002), 'PRINTS and PRINTS-S shed light on protein ancestry', *Nucleic Acids Res.*, Vol. 30(1), pp. 239–241.
22. Rost, B. (1999), 'Marrying structure and genomics', *Structure*, Vol. 6, pp. 259–263.
23. Henikoff, J., Greene, E. A., Pietrokovski, S. and Henikoff, S. (2000), 'Increased coverage of protein families with the Blocks Database servers', Nucleic Acids Res., Vol. 28(1), pp. 228–230.
24. Huang, J. Y. and Brutlag, D. (2001), 'The EMOTIF database', *Nucleic Acids Res.*, Vol. 29(1), pp. 202–204.
25. Falquet, L., Pagni, M., Bucher, P. *et al.* (2002), 'The PROSITE database, its status in 2002', *Nucleic Acids Res.,* Vol. 30(1), pp. 235–238.
26. Bateman, A., Birney, E., Durbin, R. *et al.* (1999), 'Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins', *Nucleic Acids Res., Vol.* 27(1), pp. 260–262.
27. Gouzy, J., Corpet, F. and Kahn, D. (1999) 'Recent improvements of the ProDom database of protein domain families', *Nucleic Acids Res.,* Vol. 27(1), pp. 263–267.
28. Reich, J. R., Mitchell, A., Goble, C. A. and Attwood, T. K. (2001), 'PRECIS: Protein Reports Engineered from Concise Information in SWISS-PROT', *IEEE Intelligent Systems*, Vol. 16(6), pp. 42–51.
29. URL: ftp://ftp.bioinf.man.ac.uk/pub/prints
30. URL: ftp://ftp.hgmp.mrc.ac.uk/pub/database/prints
31. URL: ftp://ftp.ebi.ac.uk/pub/databases
32. URL: ftp://ftp.embl-heidelberg.de
33. URL: ftp://ncbi.nlm.nih.gov