The PRINTS Database of Protein Fingerprints: A Novel Information Resource for Computational Molecular Biology

T. K. Attwood,*,[†] H. Avison,[‡] M. E. Beck,[‡] M. Bewley,[‡] A. J. Bleasby,[§] F. Brewster,[†] P. Cooper,[‡] K. Degtyarenko,[‡] A. J. Geddes,[‡] D. R. Flower,[⊥] M. P. Kelly,[‡] S. Lott,[‡] K. M. Measures,[‡] D. J. Parry-Smith,[||] D. N. Perkins,[‡] P. Scordis,[†] D. Scott,[‡] and C. Worledge[†]

Department of Biochemistry and Molecular Biology, University College London, London WCIE 6BT, UK, Department of Biochemistry and Molecular Biology, The University of Leeds, Leeds LS2 9JT, UK, CLRC Daresbury Laboratory, Warrington, Cheshire WA4 4AD, UK, Department of Physical and Metabolic Sciences, Astra Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, UK, and Department of Discovery Biology, Pfizer Central Research, Sandwich, Kent CT13 9NJ, UK

Received October 1, 1996[⊗]

PRINTS is a compendium of protein motif fingerprints derived from the OWL composite sequence database. Fingerprints are groups of motifs within sequence alignments whose conserved nature allows them to be used as signatures of family membership. Fingerprints inherently offer improved diagnostic reliability over single motif methods by virtue of the mutual context provided by motif neighbors. To date, 650 fingerprints have been constructed and stored in PRINTS, the size of which has doubled in the last 2 years. The current version, 14.0, encodes 3500 motifs, covering a range of globular and membrane proteins, modular polypeptides, and so on. The database is now accessible via the UCL Bioinformatics Server on http://www.biochem.ucl.ac.uk/bsm/dbbrowser/. We describe here progress with the database, its compilation and interrogation software, and its Web interface.

INTRODUCTION

Bioinformatics, the computational analysis of protein and nucleic acid sequences, is one of the fastest growing areas of chemical information science. The development of rapid DNA sequencing techniques in the late 1970s has given rise to an ever-increasing flood of information. In the last few years this has been exacerbated by advances in automation technology and spurred on by various initiatives to sequence whole genomes. The human genome project, for example, expects to complete the sequencing of all 100 000 human genes by the first years of the new millennium.

The rapid accumulation of macromolecular sequence data, which is both abundant and, by its nature, complex, presents a significant challenge to chemical information science. The challenge resides not only in the management of this huge quantity of information but also in its analysis. One of the main goals of Bioinformatics, as an information science, is to uncover the knowledge implicit within these data. In practice, the principal means to achieve this is to identify relationships between the sequences that comprise the data set.

Often, the key step is discovering to which family a newly-identified gene belongs; from this devolves a wealth of insights into biological function. It is generally believed that the protein, rather than nucleic acid, sequence is the most sensitive level at which to seek these relationships; with its links to 3D structure, post-translational modifications, etc., the amino acid sequence is closest to biological function.

In the analysis of a novel protein sequence, the customary first step is to scan the full sequence against one of the many primary data sources (e.g., SWISS-PROT,[1] PIR,[2] translated GenBank[3,4]) or against a composite resource (e.g., NRDb,[5] MIPSX,[6] OWL[7]), using a general similarity search algorithm such as BLAST.[8] This will frequently allow outright identification of the query or at least allows its classification into a broad protein family. The practicalities of such searches, and the issues they raise, are reviewed well elsewhere.[9]

Sometimes, however, such diagnoses are not possible, either because there are no other related sequences in the primary sources, or because the target sequences are only partially similar and the relationship is lost in the so-called “Twilight Zone”,[10] the level of similarity below which global sequence alignments lose statistical significance. In such situations, it is important to bring a range of techniques to bear on the analysis in order to improve the chances of making a meaningful identification.

To this end, it is also becoming standard practice to search novel sequences against a variety of other secondary “value-added” databases, which distill sequence information from primary sources into a variety of potent family descriptors, including patterns and profiles (e.g., PROSITE[11]), motifs (e.g., BLOCKS[12]), and domains (e.g., ProDom[13]).

Of these techniques, regular expression patterns are probably the easiest to derive, involving the reduction of conserved motifs within alignments into single consensus expressions. PROSITE is the most comprehensive and widely-used database of this type, and version 13.0 contains 889 documentation entries describing 1167 patterns, rules, and profiles.

In terms of their performance in pattern recognition, the regular expressions that form the PROSITE database can be likened to Markush definitions used in chemical substructure

* To whom correspondence should be addressed.

[†] University College London.

[‡] The University of Leeds.

[§] CLRC Daresbury Laboratory.

[⊥] Astra Charnwood.

[||] Pfizer Central Research.

[⊗] Abstract published in *Advance ACS Abstracts,* May 1, 1997.

S0095-2338(96)00468-4 CCC: $14.00

searching. Although they allow flexibility, or fuzziness, in patterns to be matched, they are deterministic: theirs is an essentially binary, or "on/off", nature—i.e., a query sequence will either match the pattern or not, regardless of how similar it may be. Thus sequences that differ only slightly from the definition will be missed. This is a recognized drawback of regular expressions, and consequently more powerful discriminators (i.e., profiles) are being incorporated into PROSITE to try 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 used a different approach to pattern recognition that is simple to apply. Groups of conserved motifs are excised from sequence alignments and used as "fingerprints" of family membership. The motifs are used to make independent scans of the database, and the results are correlated to determine if any additional sequences have matched all the motifs. Where this is true, information from the new sequences is added to the initial motifs, and the database is searched again. Sequence information is thus augmented through iterative database scanning, so diagnostic performance increases with each database pass.[14,15] The constituent motifs are, in concept, similar to profiles and, like them, allow probabilistic matches; the match of a sequence segment to a motif is assigned a score reflecting quantitatively their agreement. The advantage of this approach is, first, that residue mismatches are tolerated within motifs, and second, where a motif is not matched, the diagnostic framework provided by neighboring motifs still allows reliable identification.

To facilitate sequence analysis and complement the PROSITE pattern/profile resource, we have recently made a range of unique protein fingerprints available in the PRINTS database.[16] In this paper, we describe the development of the PRINTS database system and its evolving role as an information resource in computational molecular biology. We include discussion of technical issues, such as the development of its indexing software, query language, and World Wide Web interface.

METHODS

Database Generation. Fingerprint construction commences with sequence alignment and excision of conserved motifs using an interactive multiple sequence alignment program, such as SOMAP.[17] The individual motifs are used to search OWL iteratively using the ADSP sequence analysis package.[14] OWL is a nonredundant composite of the major publicly-available primary sources:[7] SWISS-PROT,[1] PIR,[2] GenBank (translation),[3,4] and NRL-3D[18] (sequence data). Although strict redundancy criteria are applied to the amalgamation of the primary databases, error-checking of the sources themselves is not undertaken. OWL may thus include errors deriving directly from these sources: results of database searches must therefore be viewed in this context. ADSP is a suite of programs for database scanning and hit-list correlation.[14] The database-scanning algorithm interprets the aligned motifs essentially as a series of frequency matrices—i.e., identity searches are made, with no mutation or other similarity data to weight the results. Thus the weighting scheme is based on the calculation of residue frequencies for each position in the motifs, summing the scores of identical residues for each position of the retrieved match.

Database Format. The PRINTS database is currently generated in the form of a single ASCII (text) file. The contents are divided into a number of specific fields, relating to general information, bibliographic references, text, lists of matches, and the aligned motifs—each line of a field is assigned a distinct two-letter code, allowing the database to be indexed for fast querying of its contents. 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, which may be simple (if the fingerprint has only one element) or compound (if it contains several)—in this latter case, the number of constituent motifs is also indicated. To date, we have included only two single-component entries: these have been derived using a modification of the fingerprint technique and are thus best regarded as special cases. Finally, the general field provides cross-references to related entries in a variety of databases (e.g., PROSITE, BLOCKS, ProDom, NRL-3D,[18] scop,[19] and so on) together with entry creation and latest update information.

The example shown in Figure 1 depicts the PRINTS entry for prion proteins. This is an eight-element fingerprint, with cross-references to related entries in the PROSITE, BLOCKS, and SBASE[20] databases. References follow, together with text detailing the nature of the family under investigation and the manner in which the fingerprint was derived. A summary of the result is provided, indicating a total of 41 sequences to have matched the fingerprint in the specified version of OWL: 30 of these match all the fingerprint elements, and 11 match only a subset. The table following the summary breaks down this result to indicate how well individual motifs have performed: for example, we can see that of those sequences matching only six motifs, all fail to match motifs 1 and 8 (from which we might correctly deduce that these are fragments).

After the summary are listed the protein identification codes and titles of all true- and false-positive and partial matches. Of the 41 matches listed, 16 are not found in SWISS-PROT because OWL is a more comprehensive database. The scan history that follows indicates in which versions of OWL the fingerprint has been updated, how many iterations were required, what hit-list length was used, and the scanning method employed: in this case, the entry was derived on OWL18.0 and has been updated on three subsequent versions; the NSINGLE scanning method[14] was used, and the results reflect a hit-list length of 150. 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, i.e., the general code with the number of that particular motif appended. For convenience, only motif 1 (PRIONl) is shown in Figure 1.

Database Indexing and Query Language. The query language for the PRINTS database is called SMITE. Database fields are indexed for speed of information recovery and to enable flexibility within the query language. The primary index field is the PRINTS code. The code names are hashed into buckets and stored, in the index file, with their sequential position within the flat file. This primary index also maintains the offset within the flat file of the start of each entry. Secondary indices are constructed from other types of data in PRINTS. Text and sequence information is indexed in free format. Each three-letter combination is allocated a number or key, and a list of entry numbers containing the key is stored, along with the offset of that key within the indexed field. Numeric indices store a list of the valid entry numbers associated with a key. A key corresponding to the element number index, for example, is an integer giving the number of elements comprising a motif. All indices, other than the primary index, refer to entries by their sequential entry number and can thus locate the entry offset from the primary index file. SMITE exploits the database indices to achieve fast information retrieval. The flat file, which contains the collection of database fields, is only accessed when displaying results; SMITE determines the entry and field offsets of the required information from the field and primary index files.

Within SMITE, queries of varying complexity can be built using the syntax “command/qualifier function ‘string’”. The default command is “display”, whose attributes may be modified by the use of a variety of qualifiers. Thus, for example, the user may retrieve full information for a given query, brief information, just reference data, just authors, and so on. The principal functions refer to the entry code, text and sequence fields, other functions giving access, for example, to sequence title and code information. A typical query might thus be, “display/info seq ‘dryfs’”—display brief information on all entries containing the pentapeptide aspartic acid-arginine-tyrosine-phenylalanine-serine. Queries for sequence, text, and numerical data can be combined using Boolean logical operators (AND, OR, XOR, etc.). The program maintains a hit-list of entry codes matching a query. A hit-list can also be stored and logically combined later with results of other queries. SMITE is written in C and runs on all proprietary UNIX platforms.

DATABASE DEVELOPMENT

The fingerprint database is released in major and minor versions: major versions are database expansions, i.e., they denote the addition of new material to the resource; minor versions reflect updates of existing versions to bring the contents in line with the current version of OWL. To date, there have been 18 releases of the database: 14 major and four minor. We endeavor to make a major or minor version available quarterly.

The principal obstacle to the frequency of expansions, and particularly of updates, is the time-consuming nature of the approach. Deriving a fingerprint for a given protein family involves two principal threads: (i) a computational aspect, which involves initial alignment and maximization 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 an exhaustive technique but is consequently rigorous, and the precision of the resulting fingerprints, coupled with the quality of annotations, tends to justify the sacrifice of speed.

The current version of PRINTS, Release 14.0 (December 1996), contains 650 entries, encoding 3500 individual motifs. The complete contents list is available from the distribution sites and on the PRINTS WWW page (see later).

DATABASE DISTRIBUTION

PRINTS is available directly via the anonymous-ftp servers at Daresbury (on s-ind2.dl.ac.uk in pub/database/prints—this directory also supplies documentation and other information files, which contain details of the database contents, update statistics, references, and so on), and at EBI (ftp.ebi.ac.uk), NCBI (ncbi.nlm.nih.gov), EMBL (ftp.embl-heidelberg.de), and UCL (ftp.biochem.ucl.ac.uk). In addition, it is available on the EMBL suite of CD-ROMs. The database requires ~60 Mb of disc storage. The source code for the indexing software and the SMITE query language are also made available with the database.

There are several ways to access PRINTS interactively over the Internet. It is accessible via the SEQNET facility at Daresbury, where, together with OWL, it is part of an integrated database and software resource that also includes query languages for each of the databases and several other programs for sequence alignment,[17] pattern recognition,[14] and global similarity searching.[21]

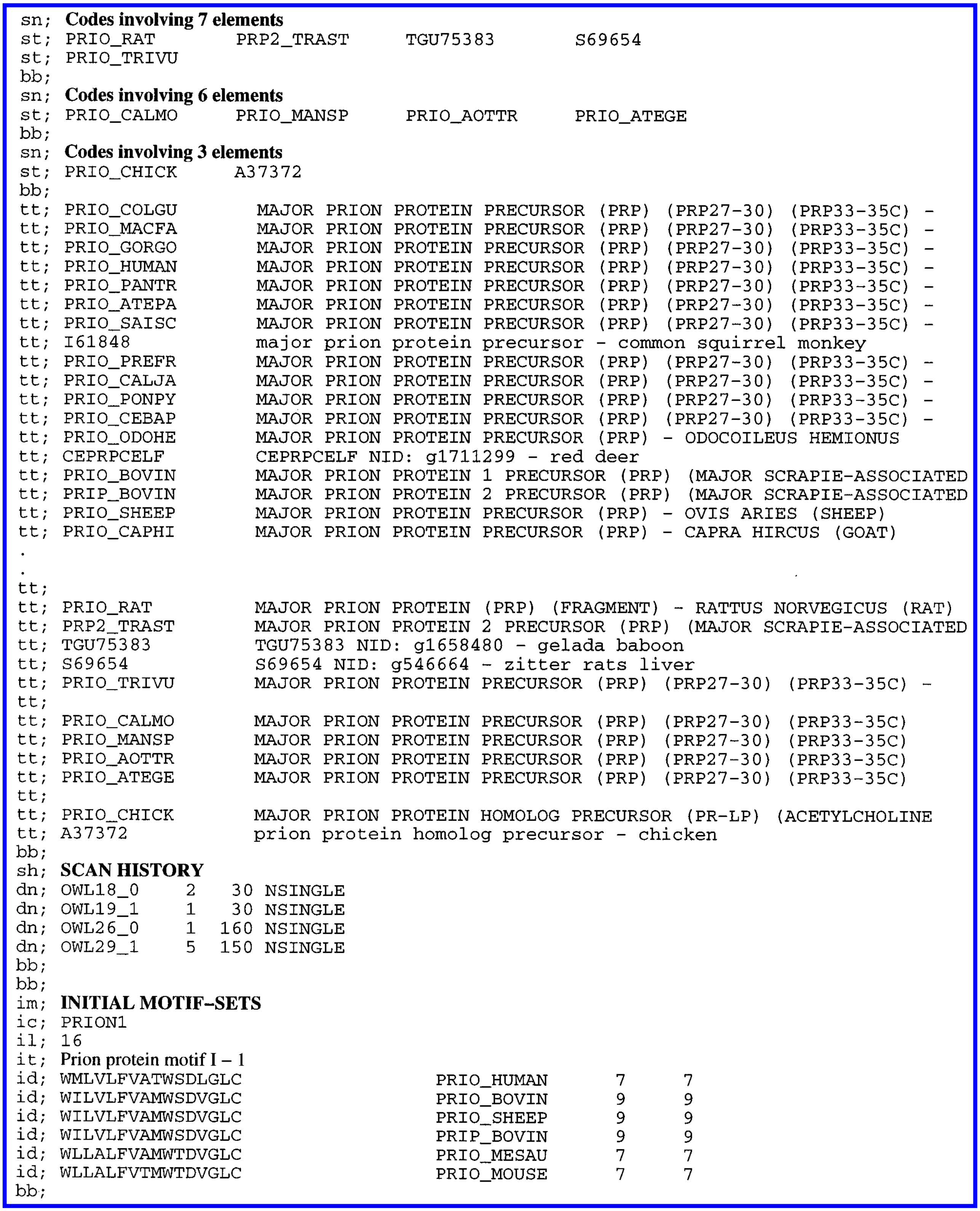
More recently, a database browser has been made accessible on the WWW, as part of UCL’s Bioinformatics Server, at http://www.biochem.ucl.ac.uk/bsm/dbbrowser.[22] The server primarily provides access to OWL, PRINTS, and ALIGN (the compendium of alignments used to create PRINTS entries), and one navigates through the facility by clicking on the appropriate hypertext link. Figure 2 shows part of the PRINTS home page, which offers the means to interrogate the database by 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. More complex queries are possible, however, by means of more explicit links to the query language logical operator functions.

The PRINTS home page also provides a facility to search PRINTS and PROSITE simultaneously, offering an instant diagnosis of any query sequence: the user supplies either the known database code or cuts and pastes a sequence from a file, and a fingerprint profile is returned in which the top-scoring matches and/or any completely matching fingerprints are plotted, as shown in Figure 3. Where results are of particular interest, the full entry may be retrieved from PRINTS to discover more about the matched fingerprint (Figure 1). Of particular importance here are the links to related databases, which allow further information to be accessed at the click of a mouse button. Such links are vital for communication between databases and effectively broaden the scope of the resource. The corollary is that the implementation of accession numbers in PRINTS, which remain static between releases, facilitates cross-referencing by other databases—PRINTS is now cross-referenced by SBASE and GCRDb[23] and is linked to by PROSITE and BLOCKS.

APPLICATIONS

The fingerprint technique has been used to study a wide range of globular and membrane proteins, modular polypeptides, and so on. Specific uses have included the development of a fingerprint for the lipocalins and fatty-acid binding

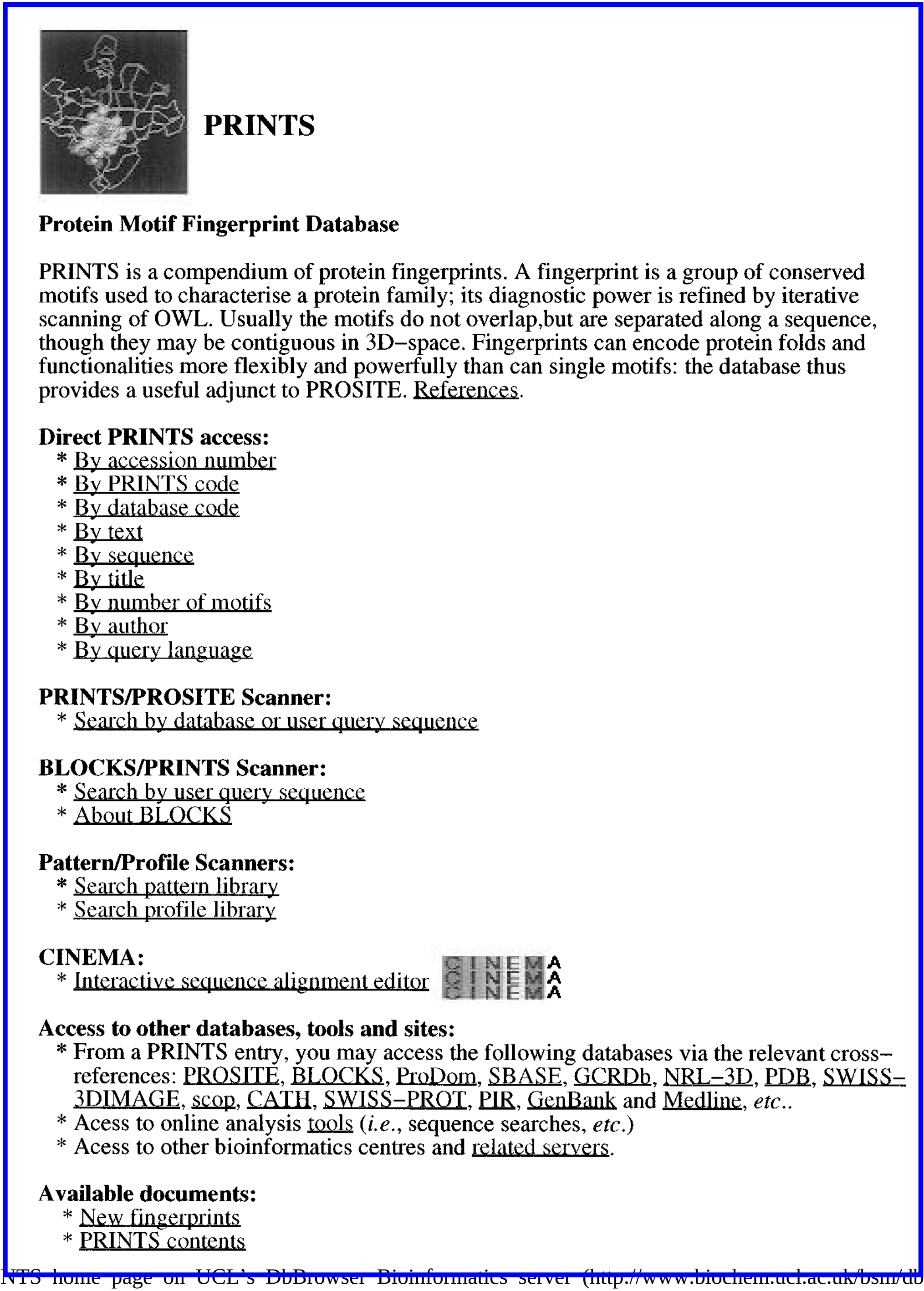
```
gc; PRION
gx; PR00341
gn; COMPOUND(8)
ga; 19-OCT-1992; UPDATE 22-JUN-1995
gt; PRION PROTEIN SIGNATURE
gp; PROSITE; PS00291 PRION_1; PS00706 PRION_2
gp; BLOCKS; BL00291
gp; SBASE; PRIO_BOVIN
bb;
gr; 1. STAHL, N. and PRUSINER, S.B.
gr; Prions and prion proteins.
gr; FASEB J. 5 2799-2807 (1991).
gr;
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 certain degenerative
gd; neurological diseases, e.g. sheep scrapie and bovine spongiform encephalopathy (BSE), and the human dementias Creutzfeldt-
gd; Jacob disease (CJD) and Gerstmann-Straussler syndrome (GSS). PrP is encoded in the host genome and is expressed both in
gd; normal and infected cells. During infection, however, the PrP molecules become altered and polymerise, yielding fibrils of
gd; modified PrP protein. 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 in other tissues,
gd; indicating that it may have different functions depending on its location.
gd;
gd; The primary 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 N-linked glycosylation; an essential disulphide bond; and 3 hydrophobic
gd; segments. These sequences show some similarity to a chicken glycoprotein, thought to be an acetylcholine receptor-inducing
gd; activity (ARIA) molecule. It has been suggested that changes in the octapeptide repeat region may indicate a predisposition to
gd; disease, but it is not known for certain whether the repeat can meaningfully 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 from an initial
gd; alignment of 5 sequences: the motifs were drawn from conserved regions spanning virtually the full alignment length, including
gd; the 3 hydrophobic domains and the octapeptide repeats (WGQPHGGG). Two iterations on OWL18.0 were required to reach
gd; convergence,  at which point a true set comprising 9 sequences was identified. Several partial matches were also found: these
gd; include a fragment (PRIO_RAT) lacking part of the sequence bearing the first motif, and the PrP homologue found in chicken -
gd; this matches  well with only 2 of the 3 hydrophobic motifs (1 and 5) and one of the other conserved regions (6), but has an
gd; N-terminal  signature based on a sextapeptide repeat (YPHNPG) rather than the characteristic PrP octapeptide. An update on
gd; OWL29.1  identified a true set of 30 sequences, and again several fragments and partial matches.
bb;
bb;
si; SUMMARY INFORMATION
sd;    30 codes involving  8 elements
sd;     5 codes involving  7 elements
sd;     4 codes involving  6 elements
sd;     0 codes involving  5 elements
sd;     0 codes involving  4 elements
sd;     2 codes involving  3 elements
sd;     0 codes involving  2 elements
bb;
bb;
ci; COMPOSITE FINGERPRINT INDEX
cd;  8|  30   30   30   30   30   30   30   30
cd;  7|   3    5    5    3    4    5    5    5
cd;  6|   0    4    4    4    4    4    4    0
cd;  5|   0    0    0    0    0    0    0    0
cd;  4|   0    0    0    0    0    0    0    0
cd;  3|   2    0    0    0    2    2    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_GORGO     PRIO_HUMAN
tp; PRIO_PANTR     PRIO_ATEPA     PRIO_SAISC     I61848
tp; PRIO_PREFR     PRIO_CALJA     PRIO_PONPY     PRIO_CEBAP
tp; PRIO_ODOHE     CEPRPCELF      PRIO_BOVIN     PRIP_BOVIN
tp; PRIO_SHEEP     PRIO_CAPHI     A34759         B34759
tp; PRIO_TRAST     PRIO_PIG       CDPRPCELF      PRIO_CERAE
tp; PRIO_MESAU     PRIO_MUSPF     PRIO_MUSVI     OCU28334
tp; PRIO_MOUSE     MUSPRPB
bb;
```



```
sn; Codes involving 7 elements
st; PRIO_RAT        PRP2_TRAST      TGU75383        S69654
st; PRIO_TRIVU
bb;
sn; Codes involving 6 elements
st; PRIO_CALMO      PRIO_MANSP      PRIO_AOTTR      PRIO_ATEGE
bb;
sn; Codes involving 3 elements
st; PRIO_CHICK      A37372
bb;
tt; PRIO_COLGU        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_MACFA        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_GORGO        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_HUMAN        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_PANTR        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_ATEPA        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_SAISC        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; I61848            major prion protein precursor - common squirrel monkey
tt; PRIO_PREFR        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_CALJA        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_PONPY        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_CEBAP        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt; PRIO_ODOHE        MAJOR PRION PROTEIN PRECURSOR (PRP) - ODOCOILEUS HEMIONUS
tt; CEPRPCELF         CEPRPCELF NID: g1711299 - red deer
tt; PRIO_BOVIN        MAJOR PRION PROTEIN 1 PRECURSOR (PRP) (MAJOR SCRAPIE-ASSOCIATED
tt; PRIP_BOVIN        MAJOR PRION PROTEIN 2 PRECURSOR (PRP) (MAJOR SCRAPIE-ASSOCIATED
tt; PRIO_SHEEP        MAJOR PRION PROTEIN PRECURSOR (PRP) - OVIS ARIES (SHEEP)
tt; PRIO_CAPHI        MAJOR PRION PROTEIN PRECURSOR (PRP) - CAPRA HIRCUS (GOAT)
.
.
tt;
tt; PRIO_RAT          MAJOR PRION PROTEIN (PRP) (FRAGMENT) - RATTUS NORVEGICUS (RAT)
tt; PRP2_TRAST        MAJOR PRION PROTEIN 2 PRECURSOR (PRP) (MAJOR SCRAPIE-ASSOCIATED
tt; TGU75383          TGU75383 NID: g1658480 - gelada baboon
tt; S69654            S69654 NID: g546664 - zitter rats liver
tt; PRIO_TRIVU        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) -
tt;
tt; PRIO_CALMO        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C)
tt; PRIO_MANSP        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C)
tt; PRIO_AOTTR        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C)
tt; PRIO_ATEGE        MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C)
tt;
tt; PRIO_CHICK        MAJOR PRION PROTEIN HOMOLOG PRECURSOR (PR-LP) (ACETYLCHOLINE
tt; A37372            prion protein homolog precursor - chicken
bb;
sh; SCAN HISTORY
dn; OWL18_0    2   30 NSINGLE
dn; OWL19_1    1   30 NSINGLE
dn; OWL26_0    1  160 NSINGLE
dn; OWL29_1    5  150 NSINGLE
bb;
bb;
im; INITIAL MOTIF-SETS
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 1. Sample data from PRINTS, showing the fingerprint for the prion protein precursor 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.

proteins,[24,25] for the diacylglycerol/phorbol-ester binding domain,[26] and for the five known families of G-protein-coupled receptors (GPCRs) and some of their many subfamilies.[15,27] This latter is particularly important as the growth of the GPCR "clan", in general and of the rhodopsin-like family, in particular, has been enormous—there are now >1000 rhodopsin-like GPCRs known, encompassing an enormously diverse range of sequences, to the extent that diagnosis of some family members is now difficult. The fingerprint facility on the Web provides an instant diagnostic tool for putative GPCRs and has allowed us to diagnose numerous sequences that are not identified by PROSITE.[15,27] This stems directly from the limitations inherent in the single-motif, regular expression approach: e.g., the third trans-membrane domain of the GPCRs alone provides the basis for the PROSITE pattern; any sequence containing only a



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 or user query sequence

BLOCKS/PRINTS Scanner:

* Search by user query sequence
* About BLOCKS

Pattern/Profile Scanners:

* Search pattern library
* Search profile library

CINEMA:

* Interactive sequence alignment editor

Access to other databases, tools and sites:

* From a PRINTS entry, you may access the following databases via the relevant cross−references: PROSITE, BLOCKS, ProDom, SBASE, GCRDb, NRL−3D, PDB, SWISS−3DIMAGE, scop, CATH, SWISS−PROT, PIR, GenBank and Medline, *etc*..
* Acess to online analysis tools (*i.e*., sequence searches, *etc*.)
* Acess to other bioinformatics centres and related servers.

Available documents:

* New fingerprints
* PRINTS contents

Figure 2. PRINTS home page on UCL's DbBrowser Bioinformatics server (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html). A range of direct access points is available, allowing simple queries by keyword searching or more complex queries using the query language logicals. A variety of complementary pattern database search tools is also provided, to afford users the opportunity to perform a comprehensive search, and novel visualization tools, such as the CINEMA alignment editor, allow interactive inspection of results.

single change in this domain will not be detected, and, accordingly, distant homologues are frequently missed. Using the fingerprint approach, however, it is possible to detect such Twilight relationships because of the diagnostic framework provided by neighboring motifs.

As a further example, consider the diagnostic performance of the prion protein fingerprint. As we have seen, this identifies 30 perfect and 11 partial matches. It is interesting to compare the profiles of some of these sequences, as illustrated in Figure 3. Within the profiles, the query sequence runs along the x-axis, the y-axis denoting the percent score for each of the eight constituent prion motifs. A perfect match is given by the human prion protein (PRIO_HUMAN), whose profile indicates six sharp individual peaks and two internally-repeating motifs (drawn from the characteristic octapeptide repeat region). A partial match is given by the prion protein precursor from the brush-tailed possum (PRIO_TRIVU): its profile indicates disparity with the "ideal" signature especially in the regions of motifs 2 and 8, and to an extent in the octapeptide repeat region. A more extreme example is revealed in the profile of the chicken prion protein homologue (PRIO_CHICK): the profile indicates only weak correspondence with motifs 1, 5, and 6 and suggests that the chicken homologue has a unique internal repeat signature. Nevertheless, in spite of the relative weakness of most of the peaks, the mutual context provided by the remaining motifs allows us to make a reliable assessment of family membership. This latter, rather poor

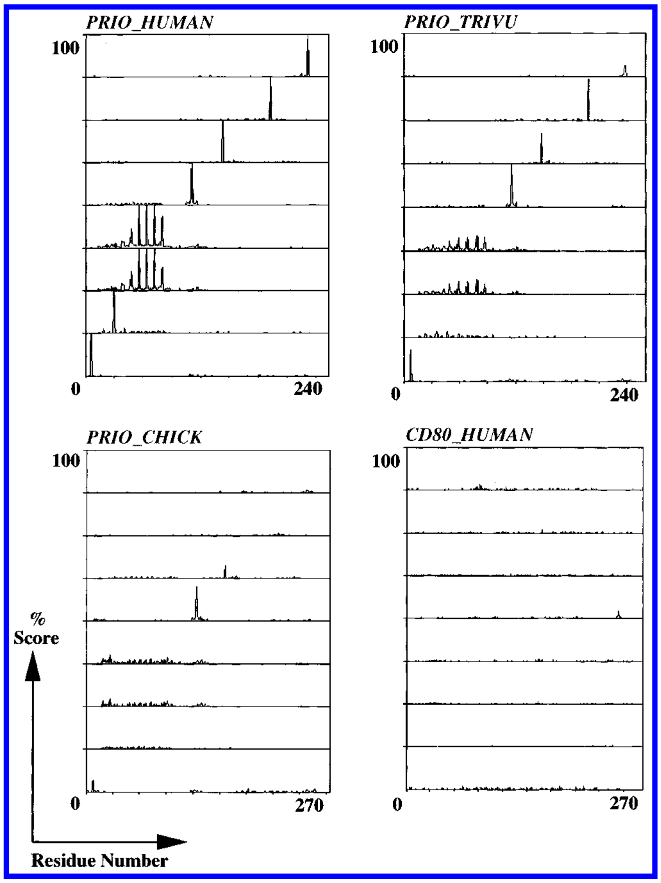


Figure 3. Fingerprint profiles returned by the PRINTS/PROSITE scanner. The horizontal axis represents the sequence, the vertical axis the percentage score of each fingerprint element (0−100 per element), and the peak a residue-by-residue match in the sequence, its leading edge marking the first position of the match. The profiles shown depict the prion fingerprint of human prion protein (PRIO_HUMAN); brush-tailed possum prion protein (PRIO_TRIVU); chicken prion protein homologue (PRIO_CHICK); and T-lymphocyte activation antigen CD80 (CD80_HUMAN). 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 of the fingerprint. These profiles illustrate how, by comparison with an "ideal fingerprint" (PRIO_HUMAN), it is possible to diagnose divergent family members (PRIO_TRIVU and PRIO_CHICK) and to distinguish them from nonfamily members, by virtue of the diagnostic framework provided by motif neighbors.

example may be contrasted with a prion fingerprint profile of another small, membrane-bound glycoprotein, the T-lymphocyte activation antigen CD80 (CD80_HUMAN), in which there are no significant peaks. The fingerprint is thus clearly a potent descriptor for the prion proteins, able to distinguish distant homologues effectively, and showing no cross-reaction with nonfamily members.

FUTURE DIRECTIONS

Just as circumstances arise where regular-expression patterns cannot unambiguously detect a particular protein family (usually because of their extreme sequence divergence), so fingerprints are not universally applicable. Sequences that have diverged to such an extent that no similarity remains will certainly escape detection by sequence-based methods of this type. We are therefore comparing the effects of applying substitution and mutation data matrices to investigate possible improvements in diagnostic performance. However, this is a complex process, as the additional information provided by such weighting schemes tends to compromise fingerprint potency by increasing the level of background noise.

CONCLUSION

Fingerprinting offers a powerful approach to the analysis of protein sequences: it inherently offers improved diagnostic reliability over single-motif methods by virtue of the mutual context provided by motif neighbors, and it allows rapid and striking visual diagnosis. Modern predictive methods increasingly exploit multiple alignments as input to prediction algorithms, since multiple sequence information can strongly enhance the signal (depending on the underlying structure of the data). In creating PRINTS, we recognized the importance of multiple sequence information from the outset, and accordingly results are stored in the form of multiply aligned motifs—these can then be the subject of detailed structure/function analyses, in a manner that is not possible with abstractions of sequence alignments such as regular expressions, profiles, and weight matrices.

Bioinformatics is one of the high frontiers of information science. It is a technically-demanding discipline in terms of both the nature and scale of the undertaking. It also promises enormous practical dividends as it begins to unlock the precious secrets of the human genome. Information resources, such as PRINTS, are the key to this endeavor; their breadth, depth, and subtlty make them powerful tools for establishing the relationships between sequences that underlie the identification of function.

ACKNOWLEDGMENT

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 AND NOTES

(1) Bairoch, A.; Apweiler, R. The Swiss-Prot Protein-Sequence Bata-Bank and Its New Supplement TREMBL. *Nucleic Acids Res.* **1996**, *24*, 21−25.
(2) George, D. G.; Barker, W. C.; Mewes, H.-W.; Pfeiffer, F.; Tsugita, A. The PIR-International Protein-Sequence Database. *Nucleic Acids Res.* **1996**, *24*, 17−20.
(3) Benson, D.; Boguski, M.; Lipman, D. J.; Ostell, J. GenBank. *Nucleic Acids Res.* **1996**, *24*, 1−5.
(4) Fickett, J. W. Correct Transmission of Protein Coding Regions in GenBank. *Trends Biochem. Sci.* **1986**, *11*, 190.
(5) Gish, W. National Center for Biotechnology Information server 1994.
(6) George, D. G.; Barker, W. C.; Mewes, H.-W.; Pfeiffer, F.; Tsugita, A., The PIR-International Databases. *Nucleic Acids Res.* **1993**, *21*, 3089−3092.
(7) Bleasby, A. J.; Akrigg, D.; Attwood, T. K. A Nonredundant Composite Protein-Sequence Database. *Nucleic Acids Res.* **1994**, *22*, 3574−3577.
(8) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(9) Altschul, S. F.; Boguski, M. S.; Gish, W.; Wootton, J. C. Issues In Searching Molecular Sequence Databases. *Nature Genetics* **1994**, *6*, 119−129.
(10) Doolittle, R. F. Proteins. *Sci. Am.* **1985**, *253*, 88−99.
(11) Bairoch, A.; Bucher, P.; Hofmann, K. The PROSITE Database, Its Status in 1995. *Nucleic Acids Res.* **1996**, *24*, 189−196.
(12) Pietrokovski, S.; Henikoff, S.; Henikoff, J. G. Automated Assembly of Protein Blocks for Database Searching. *Nucleic Acids Res.* **1996**, *24*, 197−200.
(13) Sonnhammer, E. L. L.; Kahn, D. Modular Arrangement of Proteins as Inferred from Analysis of Homology. *Protein Science* **1994**, *3*, 482−492.
(14) Parry-Smith, D. J.; Attwood, T. K. ADSP - A New Package for Computational Sequence-Analysis. *CABIOS* **1992**, *8*(5), 451−459.
(15) Attwood, T. K.; Findlay, J. B. C. Fingerprinting G-Protein-Coupled Receptors. *Protein Eng.* **1994**, *7*(2), 195−203.

(16) Attwood, T. K.; Beck, M. E.; Bleasby, A. J.; Degtyarenko, K.; Parry-Smith, D. J. Progress with the Prints Protein Fingerprint Database. *Nucleic Acids Res.* **1996**, *24*(1), 182−188.

(17) Parry-Smith, D. J.; Attwood, T. K. SOMAP - A Novel Interactive Approach to Multiple Protein Sequences Alignment. *CABIOS* **1991**, *7*(2), 233−235.

(18) Pattabiraman, N.; Namboodiri, K.; Lowrey, A.; Gaber, B. P. NRL-3D - A Sequence-Structure Database. *Protein Seq. Data Anal.* **1990**, *3*, 387−405.

(19) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP - A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* **1995**, *247*, 536−540.

(20) Murvai, J.; Gabrielian, A.; Fabian, P.; Hatsagi, Z.; Degtyarenko, K.; Hegyi, H.; Pongor, S. The SBASE Protein Domain Library, Release 3.0 - A Collection of Annotated Protein-Sequence Segments. *Nucleic Acids Res.* **1996**, *24*, 210−213.

(21) Akrigg, D.; Attwood, T. K.; Bleasby, A. J.; Findlay, J. B. C.; Maughan, N. A.; North, A. C. T.; Parry-Smith, D. J.; Perkins, D. N.; Wootton, J. C. SERPENT - An Information-Storage and Analysis Resource for Protein Sequences. *CABIOS* **1992**, *8*, 295−296.

(22) Michie, A. D.; Jones, M. L.; Attwood, T. K. DBBROWSER - Integrated Access to Databases Worldwide. *TiBS* **1996**, *21*, 191.

(23) Kolakowski, L. F. GCRDB - A G-Protein-Coupled Receptor Database. *Receptors and Channels* **1994**, *2*, 1−7.

(24) Flower, D. R.; North, A. C. T.; Attwood, T. K. Structure and Sequence Relationships in the Lipocalins and Related Proteins. *Protein Science* **1993**, *2*, 753−761.

(25) Flower, D. R.; North, A. C. T.; Attwood, T. K. Mouse Oncogene Protein-24P3 Is a Member of the Lipocalin Protein Family. *BBRC* **1991**, *180*, 69−74.

(26) Boguski, M.; Bairoch, A.; Attwood, T. K.; Michaels, G. S. PROTO-VAV and Gene-Expression. *Nature* **1992**, *358*, 113.

(27) Attwood, T. K.; Findlay, J. B. C. Design of a Discriminating Fingerprint for G-Protein-Coupled Receptors. *Protein Eng.* **1993**, *6*, 167−176.

CI960468E