PRINTS prepares for the new millennium

T. K. Attwood¹,²,*, D. R. Flower³, A. P. Lewis⁴, J. E. Mabey¹,², S. R. Morgan², P. Scordis¹,², J. N. Selley¹,² and W. Wright¹,²

¹School of Biological Sciences, University of Manchester, Manchester M13 9PT, UK and ²Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK, ³Department of Physical and Metabolic Sciences, Astra Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RF, UK and ⁴Bioinformatics Unit, Glaxo Wellcome Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, UK

Received September 28, 1998; Revised and Accepted October 16, 1998

ABSTRACT

PRINTS is a diagnostic collection of protein fingerprints. Fingerprints exploit groups of motifs to build characteristic family signatures, offering improved diagnostic reliability over single-motif approaches by virtue of the mutual context provided by motif neighbours. Around 1000 fingerprints have now been created and stored in PRINTS. The September 1998 release (version 20.0), encodes ∼5700 motifs, covering a range of globular and membrane proteins, modular polypeptides and so on. The database is accessible via the DbBrowser Web Server at http://www.biochem.ucl.ac.uk/bsm/dbrowser/. In addition to supporting its continued growth, recent enhancements to the resource include a BLAST server, and more efficient fingerprint search software, with improved statistics for estimating the reliability of retrieved matches. Current efforts are focused on the design of more automated methods for database maintenance; implementation of an object-relational schema for efficient data management; and integration with PROSITE, profiles, Pfam and ProDom, as part of the international InterPro project, which aims to unify protein pattern databases and offer improved tools for genome analysis.

INTRODUCTION

The last decades have witnessed a rapid, unprecedented accumulation of sequence data and, hand in hand with this, the birth of the field of bioinformatics. To take bioinformatics beyond the realm of mere data collection, concerted efforts are needed to analyse the amassed sequence collections, to uncover the biochemical and biophysical information implicit in the data. Today’s challenge, and the imperative for the new millennium, is thus to be able to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences.

If we do not yet understand this language sufficiently to be able to read the words in a sequence sentence that form a particular protein structure, we can at least recognise patterns of words that characterise given families. Such patterns are now used routinely to detect similarity between newly-determined and known sequences, and hence to infer related structures and functions.

Several distinct pattern-recognition techniques have emerged, spawning a range of secondary databases. Such resources distil sequence information from the primary databanks into a variety of potent descriptors that aid family diagnosis: for example, PROSITE houses regular expressions 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 descriptors 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 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. The ability to tolerate mismatches, both at the level of residues within individual motifs, and at the level of motifs within 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

*To whom correspondence should be addressed at present address: School of Biological Sciences, University of Manchester, Manchester M13 9PT, UK. Email: attwood@biochemistry.ucl.ac.uk
available in the PRINTS database (4). Here, we describe recent progress with the resource, and outline the major changes we are making in preparation for the new millennium.

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, for example, SOMAP (12) or CINEMA (13). 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 (14). Entries are assigned both an identification code and an accession number to facilitate cross-referencing by other databases. Conversely, where possible, cross-references are provided to other databanks [e.g., PROSITE (1), SBASE (15), SCOP (16), PDBsum (17), CATH (18), etc.] in order to promote efficient communication between related bioinformatics resources. The full format has been described previously (14,19), so will not be discussed further here.

Content of the current release

Release 20.0 of PRINTS (September 1998) contains 990 entries, encoding ~5700 individual motifs. A complete list of contents is available from the distribution sites and from the PRINTS home page (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/printscontents.html ).

Database update and growth

PRINTS is released in major and minor versions, the former denoting database expansions (i.e., the addition of new material to the resource), and the latter reflecting updates of existing entries to bring results in line with the current version of OWL. To date, there have been 24 releases of the database: 20 major and four minor. We endeavour to make a major or minor version available quarterly: in the last year, we have achieved three 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 is a rigorous, exhaustive technique, involving both a computational aspect (sequence alignment, motif detection, iterative database searching, etc.), and an annotation component (researching the family, discussion of motifs and their structural or functional significance, and so on). Until now, the precision of results and extent of family annotations has tended to justify the sacrifice of speed, setting the database apart from the growing number of automatically-derived pattern resources, for which there are no annotations and no appropriate mechanisms for result validation. Today, however, the overwhelming numbers of sequences flooding from the genome projects demand more efficient approaches to data analysis and storage, and projects are now in hand to address these issues.

Database distribution

PRINTS is available for interactive use via the DbBrowser Bioinformatics Server, at http://www.biochem.ucl.ac.uk/bsm/dbbrowser/ . 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 a query language (14), the syntax of which is conveniently hidden from the user beneath the Web interface. To discover more about a fingerprint, the full entry may be retrieved: as shown in Figure 1, hyperlinks allow retrieval of related information from a variety of bioinformatics resources, and the parent alignment may be downloaded via a link to the CINEMA colour alignment editor (13), allowing interactive manipulation of the alignment and, where co-ordinates are available, visualisation of motifs in a 3D context.


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 alternative scoring methods or data abstractions. Thus, a BLOCKS-format version of the resource is available at the Fred Hutchinson Cancer Research Center (http://www.blocks.fhcrc. org/blocks_search.html ), which exploits the scoring method 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 PRINTS’ multiply-aligned motifs, offering different levels of stringency from which to infer the significance of matches (20). 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 lysozyme family. The information is separated into specific fields, relating to text, references, etc. The cross-references at the top of the file are efficient coupling to related databases. The hyperlink for viewing the parent alignment invokes the CINEMA colour interactive editor, as shown, allowing the user to manipulate the alignment and, where co-ordinates are available, to rationalise conserved motifs in three dimensions. In the example shown, the acidic calcium-binding motif of equine lysozyme is highlighted within the alignment and visualised in a structural context with the 3D viewer.

Search software

The most recent facility to have been added to the Web interface is a BLAST server (http://www.biochem.ucl.ac.uk/cgi-bin/wright/printsBLAST.cgi). A FASTA-format database was created from all sequences in PRINTS, and SRS indexing (21) used to extract the relevant fingerprint information, together with sequence details from OWL. An implementation of BLAST (22) allows searches of the sequence file with either protein or DNA queries. Part of a typical BLAST output is shown in Figure 2. The result is divided into three fields: the first of these is a table summarising the top 10 most-frequently occurring fingerprint matches; the second is the standard BLAST summary, with a modified description line consisting of the PRINTS ID name and accession number, the number of motifs matched by the query and the total number of motifs in the fingerprint, the primary database ID and accession number, and finally, the matching sequence description; the third field contains the pairwise sequence alignments. Constructing the description line in the manner shown in Figure 2 ensures that matches reported by BLAST can be related directly back to a given fingerprint, and further information easily retrieved from both PRINTS and OWL via their hyper-linked ID names.

Further search options are available via the fingerPRINTScan suite (http://www.biochem.ucl.ac.uk/cgi-bin/fingerPRINTScan/fingerPRINTScan.cgi), which provides facilities for individual and bulk data searches against the full database, and for single sequence searches versus named fingerprints. In an attempt to cater for both novice and expert users, results of individual searches against the database are returned on different levels: first an ‘intelligent’ best guess is provided, based on the occurrence of the highest-scoring full or partial fingerprint match; more detailed results are then provided in different layers via extended HTML tables, as illustrated in Figure 3. By contrast, the bulk data submission facility provides only brief information, which is
Figure 2. Excerpt from a PRINTS BLAST output. Results are shown for a search with the hypothetical protein UL78_HCMVA: scores and P-values are low, but all are to the rhodopsin-like family of GPCRs. Those familiar with BLAST output will recognise many of the output features; new are the fingerprint summary file, showing the 10 most frequently matched fingerprints; and, in the description line, links to the ID and accession number of fingerprint matches, and a note of the number of motifs matched from the total number of motifs in the fingerprint.

returned via Email; and the result of searching a single sequence with a named fingerprint takes the form of a graphical cartoon of the fingerprint profile, offering an instant diagnosis of the query, as shown in Figure 4.

FingerPRINTScan has recently been extensively optimised to improve its efficiency, making it more readily applicable to searches of large datasets. The latest version is currently being used within the EDITtoTrEMBL suite as part of the EBI's automatic protocol to annotate TrEMBL entries (http://www.uni-koeln.de/math-nat-fak/biochemie/gcb98 ).

Applications

The fingerprint technique has been used to study a variety of G-protein-coupled receptor (GPCR) families and subfamilies (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/GPCR/ ), and a wide range of other membrane, globular, modular proteins and so on (6,23,24). Particular emphasis has been placed on the elucidation of discriminatory fingerprints for GPCRs, because they constitute a pharmacologically important family and their numbers in sequence databases have consequently soared—there are now >1500 rhodopsin-like GPCRs available 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 provides an instant diagnostic tool for putative GPCRs, as illustrated by the examples shown in Figure 4. The

Figure 3. 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 matches; deeper levels of matches, pushing further into the Twilight Zone, are presented in additional HTML tables. In this example, the program’s best guess confirms the query sequence (GPCR_LYMST) as being a member of the rhodopsin-like GPCRs and, in addition, diagnoses it as containing both leucine-rich and LDL-receptor repeats. In the next level of output, the top 10 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.

Figure shows a number of complete and partial matches to the rhodopsin-like fingerprint, which encodes the seven transmembrane (TM) domains: the first (OPSD_SHEEP) is a known family member, matching all seven TM regions; the second (NY5R_HUMAN) is again a clear family member, but is not diagnosed by PROSITE because it contains changes in the third TM domain, which alone provides the basis for the PROSITE pattern; the third (YMJC_CAEEL) is a partial match, lacking significant matches with TM domains 4 and 5 (this sequence not only fails to match the PROSITE pattern, but also fails to be annotated as a false negative); and the final example (UL78_HCMVA) is a poor Twilight match, for which PROSITE again returns no diagnosis at all (positive or negative), and for which, initially, BLAST similarly fails to return any significant scores—although no matches are reported from searches of resources such as Pfam and ProDom (25), the BLOCKS databases do suggest a relationship with the rhodopsin-like GPCRs, as do subsequent iterations using PSI-BLAST (22). 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 a fingerprint profile, the mutual context provided by associated fingerprint elements allows us to infer distant family relationships. This is particularly
Figure 4. Graphical output returned by FingerPRINTScan. Within each 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, human type 5 neuropeptide Y receptor, and Caenorhabditis elegans and human cytomegalovirus hypothetical proteins. 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 and human neuropeptide Y receptor are known true-positive family members, matching all seven TM 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 6–7; and the human cytomegaloviral sequence is an outstanding partial match that demonstrates a poor, but nevertheless probable relationship with the rhodopsin-like GPCRs.

important in the context of genome analysis, providing a useful adjunct to protocols based, for example, on the combination of PROSITE and HMMs, which alone are likely to miss significant matches.

Future directions

To cope more effectively with information pouring from the genome projects, we are attempting to reduce the manual burdens inherent in our current database maintenance strategies by increasing levels of automation. To this end, projects are now in hand to automate, where possible, processes for: (i) fingerprint derivation; (ii) extraction of low-level annotations from the primary databases; and (iii) database update. As part of this effort, we are designing an object-relational schema to allow more efficient data management, but with sufficient flexibility to provide continued support for the current flat-file format and to generate alternative (e.g., PROSITE-compatible) output formats in the future. We are also considering methods to link PRINTS with other bioinformatics applications; with this in mind, we are exploring options to incorporate CORBA into the schema.

Further, with a view to improving the scoring potential of the fingerprint technique, we have been developing PRISMs (PRINTS Substitution Matrices), based on the methodology developed by Henikoff and Henikoff in the derivation of the BLOSUM matrices (26). This will allow us to provide position-specific scoring matrices for individual motifs in the near future.

A further important project is the international initiative to unite the efforts of the secondary database providers. The project, known as InterPro, aims to pool the high-level documentations within PRINTS and PROSITE into a central compendium of domain and family descriptions based in Switzerland; from this will stem different satellite pattern resources, maintained at their current sites (e.g., including PROSITE, profiles, PRINTS, Pfam and ProDom). This project will help to reduce duplication of effort in the rate-determining step of annotation, it will facilitate communication between the disparate databases, and ultimately aims to provide a one-stop shop for the analysis of newly-determined sequences.

The radical changes to our current methods of database administration, coupled with the implementation of increased levels of automation, constitute a significant undertaking. While the methods are under development and 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. Our philosophy is to provide a more comprehensive, annotated resource, rather than attempting to generate an up-to-date look-up table of family membership, a Sisyphean task in the face of the ever-growing mass of primary data.

CONCLUSION

Although effective tools for diagnosing the relationships between sequences that underpin the inference of function, none of the secondary databases is complete and none of the underlying analysis methods is infallible. Therefore, in concert with PROSITE, profiles, BLOCKS, Pfam, etc., we offer PRINTS as a powerful component of a complementary suite of resources for sequence analysis. The extensive scheme of modernisation outlined here should place the database in a stronger position to embrace the genome era heralded by the new millennium.

ACKNOWLEDGEMENTS

PRINTS is now built and maintained at the University of Manchester with support from the Royal Society (T.K.A. is a Royal Society University Research Fellow), having been hosted for the last 5 years by University College London. We are grateful for individual support from Astra Charnwood (P.S.), Zeneca
Pharmaceuticals (J.N.S.), Glaxo Wellcome (W.W.), the BBSRC (J.E.M. and W.W.) and Pfizer (S.R.M.). We further acknowledge the support of the SEQNET and EMBnet user communities.

REFERENCES