Abstract

A tool for searching pattern and fingerprint databases is described. Fingerprints are groups of motifs excised from conserved regions of sequence alignments and used for iterative database scanning. The constituent motifs are thus encoded as small alignments in which sequence information is maximized with each database pass; they therefore differ from regular-expression patterns, in which alignments are reduced to single consensus sequences. Different database formats have evolved to store these disparate types of information, namely the PROSITE dictionary of patterns and the PRINTS fingerprint database, but programs have not been available with the flexibility to search them both. We have developed a facility to do this: the system allows query sequences to be scanned against either PROSITE, the full PRINTS database, or against individual fingerprints. The results of fingerprint searches are displayed simultaneously in both text and graphical windows to render them more tangible to the user.

Introduction

Searches of primary sequence databases and of their secondary ('value-added') counterparts currently offer the most practical means of predicting the biological functions of newly-determined proteins. Value-added databases contain highly-specialized information, from regular-expression patterns to aligned motifs or domains, and are quick to search because they are generally smaller than their primary sources. The types of resource now available include the PROSITE dictionary of patterns (Bairoch, 1994); the BLOCKS database of aligned sequence segments (Henikoff and Henikoff, 1993); the SBASE domain library (Pongor et al., 1993); the ProDom domain database (Sonhammer and Kahn, 1994); and the PRINTS database of protein motif fingerprints (Attwood and Beck, 1994; Attwood et al., 1994).

The most comprehensive and widely-used pattern database is PROSITE, release 12.2 of which contains 785 documentation entries describing 1029 patterns, rules and profiles. Version 8.0 of PRINTS is around half this size, encoding 1686 motifs in 350 fingerprints. A fingerprint is a group of motifs excised from conserved regions of a sequence alignment and used to dredge OWL, a non-redundant composite of the major publicly-available primary sources (Bleasby et al., 1994). The process is iterative, sequence information being maximized with each cycle, and the fingerprint consequently matures as additional matches are included (Attwood and Findlay, 1994, 1993; Parry-Smith and Attwood, 1992). Once the fingerprint is fully-refined, its constituent motifs are stored in the PRINTS database as local alignments—they therefore differ from conventional regular-expression type patterns, in which sequence information in conserved regions of alignments is distilled into minimal consensus strings.

Databases are of limited value if appropriate search tools are not available. Accordingly, several programs have been developed to search PROSITE, such as MacPattern (Fuchs, 1994) Scrutineer (Sibbald et al., 1991) and PATMAT (Wallace and Henikoff, 1992). Interrogation of PRINTS, on the other hand, has largely only been possible via the SEQNET service, which provides access to its query language, SMITE (Bleasby et al., 1994; Akrigg et al., 1992). SMITE allows direct interactive exploration of the database: e.g. it is possible to perform text, amino acid string and database code searches, and to combine these functions to develop more complex queries using logical operators. However, the program does not provide the means to search PRINTS with a full query sequence, and the lack of such search tools in general is a barrier to the usefulness of the database.

Searching PROSITE is relatively straightforward, involving translation of the stored patterns into Unix-style regular expressions, which can be searched with the...
AWK pattern-matching language. Sequence identification is then a binary process, in which target sequences are recognized if they match the pattern exactly, otherwise they are discarded. Results must be analysed manually to decide if they constitute true- or false-positive hits, and/or whether true hits have been falsely discarded. Searching PRINTS is more difficult. Ideally, a target sequence will match with all constituent motifs in any given fingerprint. But fragments, for example, often lack parts of a fingerprint, or very distant relatives may make only partial matches, because they fail to make significant matches with one or more of the motifs. It is important to be able to distinguish these cases, and it is advantageous to be able to resolve them easily from false-positive matches. Another complicating factor arises when motifs from different fingerprints match similar areas in the query sequence. How then can we judge which is correct? To address these problems, we have developed a new program (XFINGER), which is described in the following pages.

System and methods

XFINGER is written in ansi standard C. The graphics and IO facilities are implemented using the X and Motif programming libraries; Silicon Graphics versions utilise GL for 3D graphics, while other versions (e.g. Suns and Alphas) use the PEX extension to X. Sequence files are read in standard NBRF-PIR format, either directly from OWL (if the OWL indices are mounted locally) or from users’ local files. For the purposes of 3D display, coordinates in standard PDB format are accepted—if a PDB file is given, sequence information is extracted directly from that file.

XFINGER is currently accessible via the SEQNET facility, where it can be used in conjunction with the PRINTS query language, SMITE, and will be made available with the database itself from the international anonymous ftp servers. A version of the program is also now accessible via the UK PRINTS World Wide Web server, which provides an interactive fingerprinting facility over the network (see http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html).

Program operation

Initially, the program generates a dialogue window, in which various options are provided to determine the input search parameters and output requirements: these are selected by depressing appropriate buttons, moving sliders or typing into dialogue boxes (see Figure 1). When the search parameters have been chosen, the results are output to the screen, listing first the top-scoring motif hits, followed by a list of any complete fingerprint matches. If the plot option is chosen, a fingerprint profile is produced onscreen, in which the x-axis denotes the query sequence and the y-axis depicts the percentage score for each of the motifs matched; this allows instant visual diagnosis of all individual motifs or of complete fingerprints matching the sequence (Parry-Smith and Attwood, 1992). Where the 3D display is chosen, the structure is depicted as an α-carbon trace in a separate window, and the user selects from the hit-list which of the motifs to visualize.

Input options

Input filename or database code. The program accepts as input either a valid database identification code or a local file name, which is typed into the dialogue box, as illustrated in Figure 1.

PDB file. Where a structure is known, the program accepts a relevant PDB file, the name of which is typed into the dialogue box provided.

Target database. This option allows the search database (PROSITE or PRINTS) to be selected by depressing the appropriate button.

Minimum length of PROSITE pattern. When PROSITE is the selected database, this option allows a minimum length of pattern to be set—effectively, this provides a noise filter, excluding from the search some of the smaller non-specific patterns (e.g. rules for glycosylation and phosphorylation sites), which often litter hit-lists and are not always relevant.

Scanning method. Two scanning methods are supported. In both, the aligned motifs are converted to raw frequency data, but in one, diagnostic performance is enhanced by weighting motif scores in proportion to the number of identities shared between the probe and target sequences (Parry-Smith and Attwood, 1992). This has the effect of enhancing the signal to noise ratio, and is thus the default.

Motif overlaps. This option determines whether motifs from different fingerprints can overlap. In searching a query sequence against an individual fingerprint, the constituent motifs would be expected to provide a unique, non-overlapping signature. In the context of the entire database, however, a query sequence may make high scores with motifs from a number of fingerprints, which may overlap. In such situations, the highest-scoring motif is not necessarily the correct match (e.g. this is likely to be the case when the query sequence encounters very short motifs in the database). This option is therefore...
Search tool for protein fingerprints and patterns

Fig. 1. XFINGER dialogue window, showing input and output options, which are selected by manipulating the appropriate buttons and sliders. Filenames and database codes are supplied in the dialogue boxes provided.

important to avoid losing potential true matches to higher-scoring random matches.

Fingerprints to check? This determines whether a single named fingerprint is searched against the query sequence, or whether the full database is used. For individual fingerprints, an appropriate PRINTS code must be supplied. The default is a full database search.

Threshold. Using a slider, a cut-off can be set for individual motifs, below which scores are discarded. The default value is 15% identity, which is at the level of noise.

Threshold for all motifs to match. Again, using a slider, a cut-off can be set for complete matches, below which scores are discarded. The default value is 35% identity, which experience has shown discriminates well between true- and false-positive matches.

Output options

PLOT top-scoring motifs. When searching the database, a query sequence may match a number of motifs without matching any particular fingerprint completely, or it may make both complete and partial matches. This option plots profiles of the best-scoring individual motifs above the specified threshold.

PLOT all complete fingerprints? To facilitate visualization of true matches, without the intrusion of spurious, high-scoring hits, this option allows profiles to be plotted only
for fingerprints making complete matches with the query sequence.

Output file for results. The results of a search are directed to a file by providing a suitable file name, or by accepting the default (prints.out).

Structure display. If a PDB file has been provided, when the search is initiated, a separate window is generated in which the structure is depicted as an α-carbon backbone. Characteristics of the display are changed using pull-down menus, which provide scale, rotation and translation options, full main-chain display, etc. The motifs to be visualised are chosen by clicking on the hit-list provided, and the locations of the motifs are then highlighted with different colours.

Action

OK. Once the desired input and output parameters have been selected, the OK option initiates the search. It also allows new options to be selected and the search resubmitted, without leaving the program.

Quit. Quit terminates the session.

Results and discussion

Fingerprints are powerful diagnostic tools, providing the means rapidly to identify or characterize query sequences. Their particular strengths are (i) they encode multiple motifs independently, allowing the construction of signatures for given protein families; (ii) sequence information is maximised through iterative database scanning, rendering them more potent with each database pass; and (iii) mismatches are tolerated both at the level of individual motifs and of the full fingerprint. Thus, when a novel sequence is searched, it has a relatively high chance of being recognized by a fingerprint—even if it does not match the signature exactly, the context provided by the particular combination of motifs nevertheless provides a diagnostic framework.

Two important ramifications arise from these conditions: first, partial sequences have a high chance of being characterized by the technique, hence offering the means to identify and classify, for example, expressed sequence tags (ESTs) and thus to explore EST databases, which are rapidly growing sources of partial sequence data; and second, highly diverged sequences that show variations within motifs that are otherwise characteristic of particular protein families are not discarded—such sequences would not be recognized by regular expression patterns that had not ‘seen’ these variations before, but here they are viewed as part of a wider picture in which neighbouring motifs provide suitable reference points. To this extent then, fingerprinting offers a complementary approach to regular expression pattern matching.

To illustrate this last point, we turn to a group of proteins that has already been successfully fingerprinted, namely the GPCRs (Attwood and Findlay, 1994, 1993), an extensive family of cell-surface receptors involved in intercellular communication. GPCRs are characterized by seven hydrophobic regions, which are believed to form transmembrane (TM) helices that create a binding pocket for small ligands. For large ligands, the N-termini of the receptors play a part in receiving and presenting the molecule to the binding site, a role reflected in their possession of large N-terminal extensions. For example, the GPCR from the great pond snail, Limnaea stagnalis (accession number S40241) (Tensen et al., 1994) has an N-terminal domain of ~700 residues: this contains repeats similar to cysteine-rich motifs of LDL receptors and to leucine-rich motifs of various extracellular proteins, which are believed to be involved in protein-protein interactions.

When we use XFINGER to scan this sequence against the PRINTS database, three fingerprints are shown to make complete matches, namely those for LDL motifs, leucine-rich repeats and rhodopsin-like GPCRs (see Figure 2). Closer inspection reveals that this mosaic is characterized by 12 LDL motifs, 40 residues apart, in the N-terminal 500 residues; 8 Leu-rich repeats, 25 residues apart, in the central 150 residues; and a rhodopsin-like ‘7TM’ architecture in the C-terminal 400 residues. The match with TM domain 5 is poor, barely above the level of noise, but in the context of the full GPCR signature, it is clearly a true match (closer inspection shows that the fifth TM domain lacks well-conserved proline and tyrosine residues that characterize other GPCRs in this region). The third TM domain makes significant matches also with the Leu-rich repeats, highlighting a periodic pattern of leucines in this portion of the sequence.

This result neatly illustrates the occurrence and distribution of the motifs known to be present in this sequence: it belong to the rhodopsin-like GPCR superfamily (clearly matching no other GPCR fingerprints, such as those for the metabotropic glutamate or secretin-like receptors, etc.), and it contains N-terminal LDL and leucine-rich repeats. By contrast, the results of searching PROSITE revealed no matches. This was unexpected, since the sequence is clearly a GPCR. However, in the region of the acid-Arg-aromatic triplet that characterises TM domain 3, it is unusual in having an Asp-Arg-Leu motif—the current PROSITE pattern only tolerates Phe, Tyr, Trp, Cys, Ser or His in the third position, so the sequence fails to make a match because of the ‘all or nothing’ outcome of typical regular expression matching.
The visualization options of XFINGER are important because they provide informative images of matched fingerprints in 2- or 3-dimensions: as we have seen, they can show different fingerprints occurring in the same query sequence; they can reveal the occurrence of multiple tandem repeats; they show the spacing between such repeats and their precise location in the query sequence; and they illustrate the quality of match of individual motifs. The system therefore provides a flexible and powerful diagnostic tool.

The effectiveness and widespread application of regular expressions in encoding functional sites in proteins is well-established, and PROSITE is now an essential research tool. The result outlined here, however, highlights one of the problems of using single motif pattern methods in which only exact matches are permitted: any variation in a query sequence not catered for in such a pattern will prevent its diagnosis. The use of multiple motifs to build up characteristic fingerprints obviates this difficulty because, where an individual motif matches poorly, the context of its companion motifs nevertheless provides a framework for reliable diagnosis.

The program described here has been designed in an attempt to provide the best of both worlds, combining the ease and convenience of pattern searching with the diagnostic rigour of protein fingerprinting. Thus, a motif missed by a pattern search may be picked up by a fingerprint search (provided, of course, it exists in the database). XFINGER thus provides an important link between the information stored in the PROSITE and PRINTS databases. For PRINTS in particular, whose widespread use has been hampered by the lack of appropriate search tools, it allows the database to be exploited more effectively, and may encourage future use of the resource.

Acknowledgements
DNP wrote this software in the fourth year of his PhD, which was funded by the MRC. TKA is a Royal Society University Research Fellow. This work benefits from the use of the DRAL SEQNET facility.
D.N. Perkins and T.K. Attwood

References


Received on June 30, 1995; accepted on January 1996