his document is arranged in sections that explain each of the features of MacPattern in detail. The section names can be found under the Chapters menu item and by choosing the names in the menu with the mouse, you will be taken to that chapter. You can also click into the box in the lower left corner of the window to get a table of contents. This document is formatted for DIN A4 paper size.

Introduction

MacPattern is not a normal sequence analysis program. It does not provide the standard features for, e.g., composition analysis, secondary structure prediction, etc. Instead, it supplements standard programs by offering tools that can give clues to the functions of newly discovered proteins.

MacPattern allows you to use the PROSITE and BLOCKS databases (or any other databases with the same format) for searching protein and nucleotide sequences for the occurrence of known patterns or similarities to protein blocks. Nucleotide sequences can automatically be translated on the fly. You may search for all entries in the database, for selected entries or you can create “entry sets”. In addition to its searching capabilities, MacPattern also provides an integrated, powerful database browser tool. Individual database entries can be easily accessed, and the complete information stored in the database can be viewed on screen, printed, extracted to disk files, etc.

The provision of methods for identification of statistically significant sequence features is still experimental.

Pattern Analysis

MacPattern was originally developed to utilise the PROSITE pattern database for the Apple Macintosh.

The identification of conserved short stretches of residues (so-called motifs or patterns) can be the first step in finding a biological role for a newly determined protein sequence. These patterns frequently correspond to functional important regions of a protein, for instance, catalytic sites, DNA-binding sites, or target peptides. The detection of a known sequence pattern can therefore immediately suggest a protein function and direct further experimental work.

An excellent database of protein patterns is maintained by Amos Bairoch, University of Geneva: PROSITE (Bairoch, 1992). This database is available in computer-readable form from EMBL and several other sources. It does not only contain about 800 different entries but also includes detailed textual information about all these patterns.

Here is a typical example of a PROSITE pattern, ADH_ZINC:

```
PA   G-H-E-x(2)-G-x(5)-G-x(2)-[VC].
```

Spelled out this means: a G, followed by an H and an E, then any two residues, a G, any five residues, a G again, and after two more residues either a V or a C.

So if your sequence contains, e.g., GHEALGAGCTCGTSC, MacPattern will report a pattern match, and it is quite likely that your protein is a Zinc-containing alcohol dehydrogenase. If you use MacPattern to examine the ADH_ZINC entry and its documentation, you will find that this pattern identifies all known Zinc-containing alcohol dehydrogenases in SWISS-PROT and no unrelated proteins, which makes it even more likely that your sequence is indeed a member of this protein family.

Block Searches

Pattern searches are very attractive because they provide clear yes/no answers: either you find a pattern match or you don’t. The drawback is that it provides only clear yes/no answers. Even slight variations of a pattern will not be detected. An alternative approach is to use aligned blocks of protein subsequences which characterise protein families.

Based on information extracted from PROSITE and using Smith’s MOTIF algorithm (Smith et al., 1990), Steven Henikoff and colleagues have produced a database of protein blocks: BLOCKS. For (almost) every PROSITE pattern there is one or more corresponding BLOCKS entries. The PROSITE documentation is used to build groups of related proteins, and for each group motif blocks that identify these families are identified in a multi-step process (Henikoff & Henikoff, 1991). The sequences that make up these blocks are stored in the BLOCKS database entries, e.g.

```
ID   3HCDH; BLOCK
AC   BL00067A; distance from previous block=(7,315)
DE   3-hydroxyacyl-CoA dehydrogenase proteins.
BL   GAV motif; width=28; 99.5%=782; strength=1971
CRYL_RABIT  (      8)  VLIVGSGLVGRSWAMLFASGGFRVKLYD

   ECH_RAT  (    298)  VGVLGLGTMGRGIAISFARVGISVVAVE

FADB_ECOLI  (    316)  AAVLGAGIMGGGIAYQSAWKGVPVVMKD
```

```
  HCDH_PIG  (     18)  VTVIGGGLMGAGIAQVAAATGHTVVLVD

//
```

This is one of four blocks that, together, characterise the 3-hydroxyacyl-CoA dehydrogenase protein family. Compare this to the corresponding PROSITE entry:

```
ID   3HCDH; PATTERN.
AC   PS00067;
DT   APR-1990 (CREATED); MAY-1991 (DATA UPDATE); JUN-1992 (INFO UPDATE).
DE   3-hydroxyacyl-CoA dehydrogenase signature.
PA   G-F-[LIVMF]-x-N-R-[LIVM]-x(2)-[AP].
NR   /RELEASE=22,25044;
NR   /TOTAL=4(4); /POSITIVE=4(4); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0(0);
CC   /TAXO-RANGE=??EP?; /MAX-REPEAT=1;
DR   P14755, CRYL_RABIT, T; P07896, ECH_RAT   , T; P00348, HCDH_PIG  , T;
DR   P21177, FADB_ECOLI, T;
DO   PDOC00065;
//
```

The way in which blocks are used for searches differs very much from Prosite pattern searches. The sequences stored in a BLOCKS entry are converted to a scoring matrix on the fly (Henikoff et al., 1990). This matrix is then, after some adjustments, used in a sliding-window technique to generate a score for each fragment of the input sequence equal in length to the block width, and the highest scoring fragments are identified. See ‘Block Databases’ for more information.

Scores higher than the block’s “strength” value, which is stored in every database entry, are probably significant. Because for most protein families there is more than one block, the occurrence of multiple high-scoring blocks from one family in the correct order greatly increases the significance of a match.

Statistical Analysis

Even in the absence of strong similarities to patterns or blocks, it may still be possible to get some indication about functionally important regions of a protein using statistical methods that identify “unusual” regions in a sequence.

Karlin and colleagues have developed a variety of methods to estimate the statistical significance of sequence features (Karlin & Brendel, 1992). MacPattern includes support for one of these methods, the maximal segment score analysis (Karlin & Altschul, 1990).

This approach is useful for the identification of characteristic charge clusters, hydrophobic regions, transmembrane or DNA-binding domains, and so on. It is based on the theorem that for sequences of length $N>100$ the probability of finding one or more distinct segments with scores greater than or equal to S is closely approximated by $1 - \exp(-KN \exp[-\lambda S])$ where K and λ are dependent on the scoring scheme and the letter frequencies (Karlin & Altschul, 1990). Another application of this approach is the BLAST algorithm used for sequence database searches (Altschul et al., 1990).

A second method included in MacPattern is based on an algorithm proposed by Eguchi & Seto (1992) which identifies the most dissimilar oligomer in a sequence, assuming that it contains the highest information. A sequence is dissected into overlapping oligomers and for each the degree of dissimilarity to the rest of the protein is calculated. These dissimilarity scores are smoothed using a triangular window approach (Claverie and Daulmerie, 1991) and graphically displayed or listed in tabular form.

Speed

All tests were performed with MacPattern v3.0 using a Macintosh IIcx under System 7.1.

- Pattern analysis using PROSITE Rel. 10 (main set, 797 patterns)

  vs. SWISS-PROT TMAS_HUMAN (325 residues): 5 sec
  vs. SWISS-PROT 104K_THEPA (924 residues): 6 sec

- Blocks analysis using Blocks Rel. 6 (2,302 blocks)

  vs. SWISS-PROT TMAS_HUMAN (325 residues), exact search: 7:46 min

sensitivity 5: 5:51 min

sensitivity 1: 4:56 min

- Maximal segment score analysis (all scoring schemes):

  vs. SWISS-PROT TMAS_HUMAN (325 residues): 2:30 min
  vs. SWISS-PROT SODF_ECOLI (192 residues): 0:55 min

- Database search using the PROSITE AMINO_ACID_PERMEASE
  vs. SWISS-PROT Release 24 (28,154 entries),
  directly on EMBL CD-ROM (PIR format): 20 min

How to Cite MacPattern

If you publish results obtained with the help of this program, please cite the following publication:

> Fuchs, R. (1991) MacPattern: Protein pattern searching on the Apple Macintosh. Comput. Applic. Biosci. 7, 105-106.

Support

Although the author is affiliated with the EMBL Data Library, MacPattern started off as a private research project and is no official EMBL Data Library product. Therefore, the Data Library is in no way responsible for any problems that might occur when using this program and cannot provide any support.

However, I am willing to maintain and improve this program, so feel free to contact me directly with comments, bug reports, questions, etc... Please, use electronic mail if possible.

E-mail address: Fuchs@EMBL-Heidelberg.DE.

Postal address:

Rainer Fuchs
EMBL Data Library
Postfach 10.2209
69012 Heidelberg
Germany