Pattern Databases

MacPattern was originally developed to provide Macintosh software that utilises Amos Bairoch's PROSITE database (Bairoch, 1992). Nonetheless, MacPattern is not restricted to PROSITE but will accept any database file that uses PROSITE format. A description of the PROSITE pattern syntax can be found in Appendix B of this manual. For a more detailed description of the PROSITE database format see the PROSITE user manual.

PROSITE

PROSITE is contained on the EMBL Databases CD-ROM. Additionally, it can be obtained electronically from the EMBL File Server in the directory PROSITE. If you are not familiar with the EMBL File Server, first send a normal electronic mail message to NetServ@EMBL-Heidelberg.DE containing only the single line:
HELP
to get introductory information.

Prosite can also be obtained by anonymous ftp from ftp.embl-heidelberg.de in /pub/databases/prosite or from ncbi.nlm.nih.gov in /repository/prosite.

PROSITE consists of two main parts: prosite.dat and prosite.doc.

Prosite.dat contains basic information about protein patterns in a format similar to the EMBL and SWISS-PROT databases. In addition to patterns, PROSITE contains "rules" and (announced for future releases) "matrices" to describe protein consensus sequences. Since rules are free text and the matrix format has not been defined yet, MacPattern will ignore those entries.

Prosite.doc contains textual information that fully documents the consensus sequences contained in prosite.dat. The .dat and .doc files are cross-referenced.

To use PROSITE with MacPattern, follow the instructions given in the 'Installation' chapter of this help file. Always keep the PROSITE database files and the index file in one folder if you want to have access to the documentation.

Private Pattern Databases

You may construct your own database files as well. MacPattern's indexing routines require that these files carry the extensions .dat and .doc, like the original PROSITE database files.

In fact, the .doc file is optional. A .dat file is sufficient, but if you have a .doc file keep it in the same folder.

Make sure that you strictly follow the PROSITE syntax when you construct your files!
The only line types required by MacPattern are:

.dat file:
ID line
AC line
PA line(s)
DO line (only required if you also have a .Doc file)
// line

.doc file:
PDOCnnnnn line
END line

Results of Pattern Searches

In a pattern search each input sequence is compared against all patterns selected from a database browser window or against a pattern entered manually using the Enter Pattern command. For more details see the chapter ‘Algorithms’.

The output shows for each sequence independently all matches to these patterns. If there is no match at all against an input sequence, this sequence will not be mentioned in the output. The menu commands Sort Results by Pattern and Sort Results by Location allow you to modify the order in which matches are listed.

For each match, the output lists the pattern description (i.e. accession number and entry name), the sequence position at which the match begins, and the matching region of the input sequence.

At the end there is a short summary of the number of sequences and residues searched, the total number of hits found, and the execution time of the search.

Nucleotide Pattern Databases

Although MacPattern was designed to work with protein sequences and patterns, it also works with nucleotide sequences and patterns. No part of the pattern search routines is protein-sequence specific.

A relational database of transcription factors (TFD) is maintained by David Ghosh from the National Center for Biotechnology Information (NCBI), Bethesda. A program called tfd2prosite is available to reformat several tables of this database into PROSITE format which can then be used with MacPattern. Tfd2prosite can be downloaded as an ANSI C source file from the EMBL File Server (NetServ@EMBL-Heidelberg.DE). The TFD database can be obtained most

conveniently via anonymous ftp from ncbi.nlm.nih.gov.

Although it is certainly possible to use MacPattern with databases like TFD to identify patterns in nucleotide sequences, this approach is probably not adequate. Nucleotide consensus sequences are in general less conserved and more ambiguous than protein consensus sequences and the PROSITE/MacPattern approach for characterising/identifying such sequences is, in my opinion, not appropriate for identifying nucleotide patterns.

Important: If you want to analyse nucleotide sequences, make sure that the “Translation” option in the General Options dialog is set to “none”!