Block Databases

MacPattern v2.0 introduced support for S. Henikoff's BLOCKS database (Henikoff & Henikoff, 1991). For a detailed description of the BLOCKS database format see the BLOCKS user manual and Henikoff & Henikoff (1991).

BLOCKS

BLOCKS is contained on the EMBL Databases CD-ROM. Additionally, it can be obtained electronically from the EMBL File Server in the directory BLOCKS. 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.

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

The only file required by MacPattern is blocks.dat.

Blocks.dat contains information about protein blocks in a format similar to the EMBL and SWISS-PROT databases. Inside each database entry the alignment of all SWISS-PROT sequences which comprise this block is stored.

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

Private Block Databases

You can construct your own database files as well. MacPattern's indexing routines require that these files carry the extension .dat, like the original BLOCKS database file.

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

.dat file:
ID line
AC line
BL line(s)
the protein sequence fragments

// line

Results of Block Searches

In a block search each input sequence is compared against all blocks selected from a database browser window. For more details, see the chapter ‘Algorithms’.

Each block is converted to a scoring matrix on the fly. Sequences are weighted differently to adjust for the contributions of multiple closely related sequences. The matrix entries are then weighted again to compensate for bias by using an amino acid frequency table calculated from SWISS-PROT Release 22. The resulting matrix is used to assign a score to each fragment of the input sequence(s) of the width of the block. For each pattern and sequence the highest score is kept.

Raw scores are adjusted to allow the comparison of results obtained with different blocks (Wallace & Henikoff, 1992). Otherwise, wider blocks would yield higher scores simply because they are wider. MacPattern uses the “99.5%” value stored in each BLOCKS entry, which is the 99.5th percentile of scores of true negative sequences detected by this block during calibration. Raw scores are divided by the 99.5% value and multiplied by 1000. An adjusted score of 1000 therefore means that this match is probably not important.

The output displays for each sequence independently all fragments which yield high scores against the matrices constructed from the input blocks. For each sequence, MacPattern shows a fixed number of high-scoring fragments, sorted by score. The number of fragments displayed can be changed using the Search Options command from the Blocks menu.

For each fragment, the output lists the score, the block’s strength, the fragment description (i.e. accession number and entry name), the sequence position at which the match begins, and the matching region of the input sequence. An uppercase character means that this residue appears in at least one of the database sequences that comprise the block. A lowercase character indicates that this residue is not found in any of these sequences.

MacPattern assists in the evaluation of the results of block searches:

If a score exceeds the block's "strength" it is likely that the hit is significant, and MacPattern puts a plus sign (+) in front of the score. If a score is less than 1000, the significance of the hit is rather unlikely, so MacPattern puts a minus sign (-) in front of the score.

Note: Keep in mind that this only serves as a hint. Please refer to Henikoff & Henikoff (1991) and Wallace & Henikoff (1992) for a detailed discussion of the interpretation of block search results.

At the end of the output 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.