Sequence File Formats

MacPattern accepts a wide range of different sequence formats. In most cases, one file may contain more than one sequence. MacPattern does not allow you to select specific sequences from a multi-sequence file; instead, MacPattern will search through all the sequences in a file sequentially.

Therefore, MacPattern may be used to scan a protein database for patterns, but be warned: the execution time is considerable! MacPattern was designed for the analysis of single sequences (or small batches of sequences), not for database searching. On the other hand, searching the whole of SWISS-PROT directly on the EMBL CD-ROM for one pattern takes of the order of 20 min on a Mac IIcx, which I think is reasonably fast. A search can also be performed as a background task.

In principle, apart of DNASTAR and DNAid files, all sequence files must be of type 'TEXT', so if you use a word processor to create these files make sure to use the ‘text only’ option when you save them.

MacPattern understands the following formats:

- EMBL/SWISS-PROT format

```
ID   entryname
...  (optional lines with a non-blank character in the first column)
     sequence
//
```

e.g.

```
ID   MYSAMPLE
DE   This is an example of EMBL format
OS   Homo sapiens
     ggav
//
```

Sequences in EMBL/Swiss-Prot format must have an ID line with an entryname (max. 10 chars), plus the // delimiter at the end of the sequence. Other line types are ignored. If the entryname does not contain a dollar or an underscore (_) character, the sequence is assumed to be nucleotide.

- NBRF/PIR format

```
>P1;entryname - species name
comment line
sequence*
Optional comment lines
```

e.g.

```
>P1;MYTEST - Homo sapiens
This is an example of PIR format
ggav*
C; This is a comment
```

Sequences in NBRF/PIR format must have a > line with an entryname (max. 10 chars), a comment line, and the * delimiter at the end of the sequence. Comment lines after the sequence are ignored. If an entry starts with >D or >R, the sequence is assumed to be nucleotide.

- Pearson’s Fasta format

```
>entryname - comment
sequence
```

e.g.

```
>MYTEST - This is an example of FASTA format
ggav
```

Sequences in Fasta format must have a > line with an entryname (max. 10 chars).

- DNA-Strider format

```
; comment line 1
; comment line 2
; comment line 3
sequence
//
```

e.g.

```
; This is an example
; of
; DNA-Strider format
ggav
//
```

MacPattern can read DNA-Strider sequences, if they were saved using the ‘Write ASCII’ option. DNA-Strider sequences must have three comment lines and the // and the end of the sequence.

- DNASTAR format

MacPattern understands protein sequence and nucleotide sequence files created by DNASTAR software.

- DNAid format

MacPattern will read DNAid protein sequences (file type 'PROS') and DNA sequences (file type

'DNAS'). If a protein sequence is used for input, only residues up to the first stop codon indicator ('_'; underscore character) will be treated as part of the sequence.

- IG format

```
; comment line 1
; comment line 2
...
; comment line n (n>3)
entryname
sequence
```

e.g.

```
; This is
; an example
; of
; IG format
;
MYTEST
ggav
```

Sequences in IG format must have more than three comment lines and an entryname (max. 10 chars). The sequence may end with a “1” or “2”.

- GenBank format

```
LOCUS       entryname
...  (optional lines)
ORIGIN
    sequence
//
```

e.g.

```
LOCUS       MYTEST
ORIGIN
    1 ggav
//
```

Sequences in GenBank format must have a LOCUS line, an ORIGIN line and a // line. Other line types and optional sequence numbering are ignored. Note that GenBank format is only used for nucleotide sequences!

- ASCII text

After opening a sequence file, MacPattern tries to determine the format automatically. If it is not one of those listed above, it will read in the file as a simple series of ASCII characters, treating every valid character as part of the sequence. No comments are allowed simply because they are treated as part of the sequence.

Files in EMBL/SWISS-PROT, NBRF/PIR, Fasta, DNA-Strider, or IG format may contain more than one sequence.

MacPattern will also correctly read sequence files from the EMBL CD-ROM, in both EMBL/SWISS-PROT and PIR format. PIR format is recommended for this purpose because it is more compact than EMBL/SWISS-PROT format.

Nucleotide Sequences

MacPattern accepts nucleotide sequences as input. First, it tries to determine the molecule type from the file format. For example, PIR-formatted protein sequences begin with >P or >F, while PIR-formatted nucleotide sequences start with >D or >R. If it cannot succeed this way, it analyses the sequence itself and assumes that it is a nucleotide sequence if the content of A, C, G, T, and U is higher than 85%.

The General Options menu allows you to adjust the way MacPattern processes nucleotide sequences. It can translate the given strand in all three reading frames, translate both strands in all six frames, or use the sequence as is.The standard genetic code is use for all translations.

To indicate the fact that results stem from the analysis of conceptual translations of nucleotide sequences, sequence names are modified by adding the suffix RF1, RF2 etc.

Note:The numbering of residues refers to the translation product, not the original nucleotide sequence.