Modeller PIR Format

Prof. Andrew C.R. Martin, UCL

2022

1 PIR File Format

The PIR sequence file format consists of two header lines followed by the amino acid sequence using 1-letter code ending with an asterisk (*).

The first header line is of the form:

```
>P1;xxxxxx
```

where `xxxxxx` is an identifier for the sequence. Any sequence of up to 6 characters may be supplied.

The second header line is a title describing the sequence. Optionally this may consist of two fields separated by a dash (-). If so, the second field describes the source of the sequence (e.g. 'human').

The sequence follows using the standard 1-letter code. Spaces and line breaks are ignored and an asterisk (*) marks the end of the sequence. Note that chain breaks are indicated with a slash (/) not an asterisk as used in the standard PIR format.

2 Alignment File Format

The alignment file format is an extension of the PIR format described in Appendix 1.

The alignment is simply created by using dash (-) characters to indicate deletions in the sequences.

The first header line is a standard PIR header line of the form:

```
>P1;xxxxxx
```

Note that the `xxxxxx` code (up to 6 characters) is used as an identifier by the **Target code:** and **Template PDB codes:** text boxes in MINT.

The second (comment) header line is modified and contains 10 fields separated by colons (:). These fields have the following meanings:

1. The type of structure associated with the sequence. This is specified as follows:

sequence	No structure available,
structureX	An X-ray crystal structure,
structureN	An NMR structure,
structureM	A model structure.

2. The filestem of the PDB file containing the associated structure. Any characters prepended onto the PDB code **are included**, but the extension and directory are not. Thus, if we are using crambin as a structure (PDB code 1crn) and we store the PDB files as `/pdb/pXXXX.pdb`, this would be specified as `p1crn`. This field is blank if there is no associated structure (i.e. this is the sequence to be modelled).

3. The residue number (in the PDB file) of the first residue of the sequence. Normally this will be the first residue number in the PDB file. This field is blank if there is no associated structure (i.e. this is the sequence to be modelled).

4. The chain name (in the PDB file) of the first residue in the sequence (or blank).

5. The residue number (in the PDB file) of the last residue of the sequence. Normally this will be the last residue number in the PDB file. This field is blank if there is no associated structure (i.e. this is the sequence to be modelled).

6. The chain name (in the PDB file) of the last residue in the sequence (or blank).

7. The name of the sequence. This is a text description of the protein and is normally the first COMPND record from a PDB file. This field may be left blank.

8. The source of the protein (e.g. 'HUMAN', 'MOUSE'). this field may be left blank.

9. The resolution of the crystal structure. This field is set to 0.00 if there is no associated structure (i.e. this is the sequence to be modelled) or this is an NMR structure.

10. The R-factor of the crystal structure. This field is set to 0.00 if there is no associated structure (i.e. this is the sequence to be modelled) or this is an NMR structure.

3 Frequently Asked Questions

1. Q: When doing full homology modelling using a PDB file in the current directory, why does MODELLER crash with a message of the form:

```
TOP______>    84  262 READ_ALIGNMENT FILE = SEGFILE, ALIGN_CODES = KNOWNS
rdseqpd_E> too many residues
recover__>  MODELLER_STATUS  >=  STOP_STATUS:   1  1
```

A: MODELLER is actually telling you that it cannot find the PDB file you specified for the template! The most probably reason for this is that the filenaming

convention you have used for the file in the current directory doesn't match that in the PDB.