C1 File formats

Key Notes

Three common sequence formats	There are several conventions for representing nucleic acid and protein sequences, of which the NBRF/PIR, FASTA and GDE formats are widely used. These formats have limited facilities for comments, which must include a unique identifier code and the sequence accession number.
Files for aligned sequences	Aligned sequences can be represented in NBRF/PIR, FASTA or GDE formats but there are other formats devised especially for multiple sequence alignment, including MSF, PHYLIP and ALN.
Files of structural data	Structural data are maintained as flat files using the PDB format. Such files contain orthogonal atomic co-ordinates together with annotations, comments and experimental details.
Related topics	Annotated sequence databases (C2) Multiple sequence alignment and family relationships (F1) Obtaining, viewing and analyzing structural data (I4)

Three common sequence formats

If biological data is to be used by computer programs, it must be presented in a standard format that can be read by computer. It is very common to put data in text files. As the name suggests, these files contain text that can be read by a human being as well as a computer. They are rather like the files used by word-processing packages to hold documents, but there is one important difference: text files hold (almost) only the text and little auxiliary information about formatting (more details below). Here we discuss some standard formats and some more database specific formats are discussed in Topic C2.

Many bioinformatic databases and software applications are designed to work with sequence data, and this requires a standard format for inputting nucleic acid and protein sequence information. Three of the most common **sequence formats** are NBRF/PIR (National Biomedical Research Foundation/ Protein Information Resource), FASTA and GDE. Each of these formats has facilities not only for representing the sequence itself, but also for inserting a unique **code** to identify the sequence and for making **comments** which may include for example the name of the sequence, the species from which it was derived, and an accession number for GenBank or another appropriate database (Topic C2). *Figure 1* shows the same protein sequence, that of a guinea-pig serotonin receptor, represented in the three sequence formats listed above.

Figure 1a shows the **NBRF/PIR format**. Note that the first line begins with '**>P1;**' which specifies a protein sequence. If this was a nucleic acid sequence, it would begin with '**>N1;**'. The semicolon is followed by a code, in this case '5H1B_CAVPO', which is a unique sequence identifier. Serotonin is also known as 5-hydroxytryptamine, thus 5H1B identifies the protein as serotonin receptor 1B, while CAVPO identifies its source as the guinea-pig (*Cavia porcellus*). There

(a)

```
>P1;5H1B_CAVPO
Guinea pig serotonin receptor accession: O08892
MGNPEASCTP PAVLGSQTGL PHANVSAPPN NCSAPSHIYQ DSIALPWKVL LVVLLALITL
ATTLSNAFVI ATVYRTRKLH TPANYLIASL AFTDLLVSIL VMPISTMYTV TGRWTLGQAL
CDFWLSSDIT CCTASIMHLC VIALDRYWAI TDAVGYSAKR TPRRAAGMIA LVWVFSICIS
LPPFFWRQAK AEEEVLDCLV NTDHVLYTVY STGGAFYLPT LLLIALYGRI YVEARSRILK
QTPNKTGKRL TRAQLITDSP GSTSSVTSIN SRAPEVPCDS GSPVYVNQVK VRVSDALLEK
KKLMAARERK ATKTLGVILG AFIVCWLPFF IISLVMPICK DACWFHMAIF DFFTWLGYLN
SLINPIIYTM SNEDFKQAFH KLIRFKCTT
*
```

(b)

```
> 5H1B_CAVPO  O08892|guinea pig serotonin receptor
MGNPEASCTP PAVLGSQTGL PHANVSAPPN NCSAPSHIYQ DSIALPWKVL LVVLLALITL
ATTLSNAFVI ATVYRTRKLH TPANYLIASL AFTDLLVSIL VMPISTMYTV TGRWTLGQAL
CDFWLSSDIT CCTASIMHLC VIALDRYWAI TDAVGYSAKR TPRRAAGMIA LVWVFSICIS
LPPFFWRQAK AEEEVLDCLV NTDHVLYTVY STGGAFYLPT LLLIALYGRI YVEARSRILK
QTPNKTGKRL TRAQLITDSP GSTSSVTSIN SRAPEVPCDS GSPVYVNQVK VRVSDALLEK
KKLMAARERK ATKTLGVILG AFIVCWLPFF IISLVMPICK DACWFHMAIF DFFTWLGYLN
SLINPIIYTM SNEDFKQAFH KLIRFKCTT
```

(c)

```
%5H1B_CAVPO  O08892|guinea pig serotonin receptor
MGNPEASCTP PAVLGSQTGL PHANVSAPPN NCSAPSHIYQ DSIALPWKVL LVVLLALITL
ATTLSNAFVI ATVYRTRKLH TPANYLIASL AFTDLLVSIL VMPISTMYTV TGRWTLGQAL
CDFWLSSDIT CCTASIMHLC VIALDRYWAI TDAVGYSAKR TPRRAAGMIA LVWVFSICIS
LPPFFWRQAK AEEEVLDCLV NTDHVLYTVY STGGAFYLPT LLLIALYGRI YVEARSRILK
QTPNKTGKRL TRAQLITDSP GSTSSVTSIN SRAPEVPCDS GSPVYVNQVK VRVSDALLEK
KKLMAARERK ATKTLGVILG AFIVCWLPFF IISLVMPICK DACWFHMAIF DFFTWLGYLN
SLINPIIYTM SNEDFKQAFH KLIRFKCTT
```

Fig. 1. The sequence of a guinea-pig serotonin receptor in (a) NBRF/PIR format; (b) FASTA format; and (c) GDE format.

follows a **comment line**, and the rules allow this line to be of more or less any length so it can either be empty or far too wide to fit on a printed page. Then the sequence itself follows and is terminated by an asterisk (*). It is conventional to give files in this format the extension '.pir' or '.seq'.

Figure 1b shows the **FASTA format**. The first line begins with '>' but there is no designation of protein or nucleic acid sequence. The code is entered next and this is followed (on the same line) by comments, although it is conventional to delimit the comments with a '|' symbol. As with the NBRF/PIR format there is no limit to the length of the first line. One point to note about FASTA files is that they allow lower-case letters for the amino acids. Files in this format commonly have the extension '.fasta'.

Figure 1c shows the **GDE format**. This is essentially the same as the FASTA format, but the '>' symbol in the first line is replaced by '%'. Files in this format have the extension '.gde'.

All three file formats *ignore spaces and carriage returns*. This allows sequences to be typed out in a manner that is convenient for the user. In *Fig. 1*, for example, a space has been inserted every 10 amino acid residues and a carriage return after every 60, making it much easier to manually count the residues and identify amino acids at specific positions in the sequence. Note, however, that most standard word-processing software packages do not ignore blank spaces. For some purposes, it may be desirable or necessary to construct files from unpublished and preliminary sequence data, and if programs such as Microsoft Word or Corel WordPerfect are used, the results can be unpredictable. If using Word, use text only mode with a non-proportional font or, preferably, use a simple text editor such as Notepad.

To illustrate this point, consider the creation of the following very simple NBRF/PIR file:

```
>P1;MY_CODE
my cat
MYCATSATINMYLAP*
```

Despite the fact that this protein is clearly fictitious, the format is perfectly correct and it should be possible to search for the peptide sequence in other proteins. However, by typing this sequence into Microsoft Word and saving it as a Word document (cat.doc), the file proves to be over 19 thousand bytes in length, and therefore obviously contains much more than the simple text. By saving the file as *text with line breaks* (cat.txt), the file size is reduced to 39 bytes, which seems more reasonable. However, inspection of the contents of cat.txt reveals two extra characters at the end of each line and another at the very end of the file. It is therefore best to avoid word processors and use text editors for the preparation of sequences. If a word processor is used, the file should be saved as text and sent by ftp as ASCII (Topic O4), and a text editor should then be used on the computer where the sequence analysis is carried out, to check the integrity of the file. Another point to bear in mind is that the first line of a FASTA file and the second line of an NBRF/PIR file might be extremely long and it is essential not to cut it up by inserting carriage returns, otherwise the comments might be read as part of the sequence.

Files for aligned sequences

The output from **sequence-alignment programs** can be in any one of a number of formats. All three formats discussed above are suitable for dealing with aligned sequences but there are several formats designed specifically for alignment output. *Figure 2* shows partial results from the alignment of five serotonin receptor sequences, including the guinea-pig 5H1B receptor. In order to achieve the alignments, **gaps** must be introduced (Topic E3) and these are represented either by hyphens or dots. **Multiple sequence format (MSF)** is used by several software tools. **PHYLIP (phylogenetic inference package)** is the output format of the software of that name and CLUSTALW/X (F1) has its own **ALN format**. Multiple sequence alignment is discussed in more detail in Topic F1.

Files of structural data

The raw materials for bioinformatic studies on macromolecular structures are **PDB files.** These are text files using a format devised by the Protein Data Bank (Topic C4). Such files contain orthogonal atomic co-ordinates together with annotations, comments and experimental details. Examples of parts of such files are shown in *Fig. 3*. The most important aspect of PDB files is that the 'ATOM' lines are laid out in columns of characters *not* columns of words. Compare the first ATOM lines from *Fig. 3a, b* and *c* (only the left-hand parts of each line are displayed here):

```
ATOM    1   N   VAL     16   29.582    19.112   38.968
ATOM    1   N   ILE E   16   -9.947    23.613   20.817
ATOM    1   N   ALA      1   14.702   -10.824    3.425
```

The last three columns show the orthogonal co-ordinates (*x*, *y*, *z*) of the atom and they are deduced by counting the positions including the spaces, *not* by counting the words. This is because the *x* co-ordinate is the sixth word in the first and third cases but the seventh word in the second case, because a new

(a)

```
 MSF:  435  Type: P    Check:  2299   ..

 Name: 5H1A_MOUSE oo  Len:  435  Check:  7521  Weight:  0.166
 Name: 5H1A_RAT oo  Len:  435  Check:  8470  Weight:  0.250
 Name: 5H1A_HUMAN oo  Len:  435  Check:  8517  Weight:  0.166
 Name: 5H1B_CAVPO oo  Len:  435  Check:   829  Weight:  0.222
 Name: 5H1B_CRIGR oo  Len:  435  Check:  6962  Weight:  0.100

5H1A_MOUSE       MD......MF SLGQGNNTTT SLEPFG.... ..TGGNDTGL SNVTFSYQVI
5H1A_RAT         MD......VF SFGQGNNTTA SQEPFG.... ..TGGNVTSI SDVTFSYQVI
5H1A_HUMAN       MD......VL SPGQGNNTTS PPAPFE.... ..TGGNTTGI SDVTVSYQVI
5H1B_CAVPO       MGNPEASCTP PAVLGSQTGL PHANVSAPPN NCSAPSHIYQ DSIALPWKVL
5H1B_CRIGR       MEEQGIQCAP PPPAASQTGV PLVNLS...H NCSAESHIYQ DSIALPWKVL

5H1A_MOUSE       TSLLLGTLIF CAVLGNACVV AAIALERSLQ NVANYLIGSL AVTDLMVSVL
5H1A_RAT         TSLLLGTLIF CAVLGNACVV AAIALERSLQ NVANYLIGSL AVTDLMVSVL
5H1A_HUMAN       TSLLLGTLIF CAVLGNACVV AAIALERSLQ NVANYLIGSL AVTDLMVSVL
5H1B_CAVPO       LVVLLALITL ATTLSNAFVI ATVYRTRKLH TPANYLIASL AFTDLLVSIL
5H1B_CRIGR       LVALLALITL ATTLSNAFVI ATVYRTRKLH TPANYLIASL AVTDLLVSIL
```

{ rest of file omitted }

(b)

```
     5    435
5H1A_MOUSE MD------MF SLGQGNNTTT SLEPFG---- --TGGNDTGL SNVTFSYQVI
5H1A_RAT   MD------VF SFGQGNNTTA SQEPFG---- --TGGNVTSI SDVTFSYQVI
5H1A_HUMAN MD------VL SPGQGNNTTS PPAPFE---- --TGGNTTGI SDVTVSYQVI
5H1B_CAVPO MGNPEASCTP PAVLGSQTGL PHANVSAPPN NCSAPSHIYQ DSIALPWKVL
5H1B_CRIGR MEEQGIQCAP PPPAASQTGV PLVNLS---H NCSAESHIYQ DSIALPWKVL

           TSLLLGTLIF CAVLGNACVV AAIALERSLQ NVANYLIGSL AVTDLMVSVL
           TSLLLGTLIF CAVLGNACVV AAIALERSLQ NVANYLIGSL AVTDLMVSVL
           TSLLLGTLIF CAVLGNACVV AAIALERSLQ NVANYLIGSL AVTDLMVSVL
           LVVLLALITL ATTLSNAFVI ATVYRTRKLH TPANYLIASL AFTDLLVSIL
           LVALLALITL ATTLSNAFVI ATVYRTRKLH TPANYLIASL AVTDLLVSIL
```

{ rest of file omitted }

(c)

```
>P1;5H1A_MOUSE

MD------MFSLGQGNNTTTSLEPFG------TGGNDTGLSNVTFSYQVITSLLLGTLIF
CAVLGNACVVAAIALERSLQNVANYLIGSLAVTDLMVSVLVLPMAALYQVLNKWTLGQVT
CDLFIALDVLCCTSSILHLCAIALDRYWAITDPIDYVNKRTPRRAAALISLTWLIGFLIS
IPPMLGWRAPEDRSNPNECTISKDHG-YTIYSTFGAFYIPLLLMLVLYGRIFRAARFRIR
KTVKKVEKKGAGTSFGTSSAPPPKKSLNGQPGSGDCRRSAENRAVGTPCANGAVRQGEDD
ATLEVIEVHRVGNSKGDLPLPSESGATSYVPACLERKNERTAEAKRKMALARERKTVKTL
GIIMGTFILCWLPFFIVALVLPFCESSCHMPELLGAIINWLGYSNSLLNPVIYAYFNKDF
QHAFKKIIKCKFCR-
*
>P1;5H1A_RAT

MD------VFSFGQGNNTTASQEPFG------TGGNVTSISDVTFSYQVITSLLLGTLIF
CAVLGNACVVAAIALERSLQNVANYLIGSLAVTDLMVSVLVLPMAALYQVLNKWTLGQVT
CDLFIALDVLCCTSSILHLCAIALDRYWAITDPIDYVNKRTPRRAAALISLTWLIGFLIS
IPPMLGWRTPEDRSDPDACTISKDHG-YTIYSTFGAFYIPLLLMLVLYGRIFRAARFRIR
KTVRKVEKKGAGTSLGTSSAPPPKKSLNGQPGSGDWRRCAENRAVGTPCTNGAVRQGDDE
ATLEVIEVHRVGNSKEHLPLPSESGSNSYAPACLERKNERNAEAKRKMALARERKTVKTL
GIIMGTFILCWLPFFIVALVLPFCESSCHMPALLGAIINWLGYSNSLLNPVIYAYFNKDF
QHAFKKIIKCKFCRR
*
```

{ rest of file omitted }

Fig. 2. Partial results from the alignment of five proteins with CLUSTALW (Topic F1). The formats shown are (a) MSF output; (b) PHYLIP output; and (c) NBRF/PIR output.

(a) Trypsin

```
HEADER    HYDROLASE (SERINE PROTEINASE)            13-APR-88   1SGT       1SGT   3
COMPND    TRYPSIN (/SGT$) (E.C.3.4.21.4)                                  1SGT   4
SOURCE    (STREPTOMYCES $GRISEUS, STRAIN K1)                              1SGT   5
AUTHOR    R.J.READ,M.N.G.JAMES                                            1SGT   6
REVDAT   1   16-JUL-88 1SGT    0                                          1SGT   7
JRNL        There follow the literature references
REMARK   1  There follow several remarks of which only 1 is shown here    1SGT  21
REMARK   2 RESOLUTION. 1.7 ANGSTROMS.                                     1SGT  72
            The sequence (only 2 lines shown) follows
SEQRES   1    223  VAL VAL GLY GLY THR ARG ALA ALA GLN GLY GLU PHE PRO    1SGT  92
SEQRES   2    223  PHE MET VAL ARG LEU SER MET GLY CYS GLY GLY ALA LEU    1SGT  93
FTNOTE   1    There follow several footnotes                              1SGT 110
HET     CA    246          1      CALCIUM ++ ION                          1SGT 151
FORMUL   2   CA     CA1 ++                                                1SGT 152
FORMUL   3  HOH    *192(H2 O1)  The last 3 lines describe hetero atoms    1SGT 153
              there follow several lines of secondary structure assignment
              of which only the first is shown here
HELIX    1   A ALA      56  CYS      58  5                                1SGT 154
              There follow 7 lines describing the orthogonal coordinate system
ATOM      1  N   VAL    16        29.582  19.112  38.968  1.00 12.94      1SGT 199
ATOM      2  CA  VAL    16        30.031  20.461  38.668  1.00 15.43      1SGT
               ....the bulk of the file
ATOM   1618  CD1 LEU   245         2.571  16.977  47.866  1.00 40.15      1SGT1816
ATOM   1619  CD2 LEU   245         4.758  18.112  48.337  1.00 44.30      1SGT1817
ATOM   1620  OXT LEU   245         1.660  16.559  52.387  1.00 59.60      1SGT1818
TER    1621      LEU   245                                                1SGT1819
HETATM 1622 CA    CA   246        14.219  32.828  30.463  1.00 13.21      1SGT1820
HETATM 1623  O   HOH     1        22.919  19.524  42.538  1.00  8.79      1SGT1821
               ....the remaining water molecules up to ...
HETATM 1814  O   HOH   192        -3.192  30.325  46.346  0.68 57.70      1SGT2012
CONECT   72   70                             941                          1SGT2013
               ....the connectivity data
MASTER        79   41    1    5   14   15    1    6 1813    1   30   18  1SGT2043
END                                                                       1SGT2044
```

(b) Complex of a proteinase ("E") with a polypeptide inhibitor ("I")

```
HEADER    COMPLEX(SERINE PROTEINASE-INHIBITOR)     21-JAN-83   3SGB       3SGBE  1
COMPND    PROTEINASE B FROM STREPTOMYCES GRISEUS (/SGPB$)                 3SGB   4
COMPND   2 (E.C. NUMBER NOT ASSIGNED) COMPLEX WITH THIRD DOMAIN OF THE    3SGB   5
COMPND   3 TURKEY OVOMUCOID INHIBITOR (/OMTKY3$)                          3SGB   6
SOURCE    (STREPTOMYCES $GRISEUS, STRAIN K1) AND TURKEY (MELEAGRIS        3SGB   7
SOURCE   2 GALLOPAVO)                                                     3SGB   8
AUTHOR    R.J.READ,M.FUJINAGA,A.R.SIELECKI,M.N.G.JAMES                    3SGB   9
There follow many lines of remarks and details of lit. references
One such remark is important
REMARK   2 RESOLUTION. 1.8 ANGSTROMS.                                     3SGB  73
Start of sequence ...
SEQRES   1 E  185  ILE SER GLY GLY ASP ALA ILE TYR SER SER THR GLY ARG    3SGB  92
Start of ATOM entries for "chain E" ...
ATOM      1  N   ILE E  16      -9.947  23.613  20.817  1.00 16.42       3SGB 156
...end of "chain E" and start of "chain I"
ATOM   1310  OXT TYR E 242     -10.317  35.858  21.204  1.00 29.02       3SGB1465
TER    1311      TYR E 242                                                3SGB1466
ATOM   1350  N   ASP I   7      25.100  14.110  33.198  1.00 41.61       3SGB1467
ATOM   1351  CA  ASP I   7      25.863  15.369  33.122  1.00 40.76       3SGB1468
... to the end of the file .
```

(c) Intestinal fatty acid binding protein

```
HEADER    FATTY ACID-BINDING                              20-FEB-98   1A57
TITLE     THE THREE-DIMENSIONAL STRUCTURE OF A HELIX-LESS VARIANT OF
TITLE    2 INTESTINAL FATTY ACID BINDING PROTEIN, NMR, 20 STRUCTURES
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: INTESTINAL FATTY ACID-BINDING PROTEIN;
         Many lines of remarks including the authors but 3 are included:
EXPDTA    NMR, 20 STRUCTURES
AUTHOR    R.A.STEELE,D.A.EMMERT,J.KAO,M.E.HODSDON,C.FRIEDEN,
AUTHOR   2 D.P.CISTOLA
         Compare the following line with its counterparts in (a) and (b)
REMARK   2 RESOLUTION. NOT APPLICABLE.
         Start of the sequence

SEQRES   1    116  ALA PHE ASP GLY THR TRP LYS VAL ASP ARG ASN GLU ASN
          Start of the secondary structure assignment
SHEET    1   A 5 GLY     4  LYS     7  0
MODEL        1    start of the atomic coordinates for "model 1"
ATOM      1  N   ALA     1      14.702 -10.824   3.425  1.00  0.00           N
ATOM      2  CA  ALA     1      13.562 -10.618   2.552  1.00  0.00           C
ATOM      3  C   ALA     1      12.273 -10.914   3.355  1.00  0.00           C
         ... up to the end of that and start of "model 2"
ATOM   2132  OG  GLU   116      15.846   0.773   9.258  1.00  0.00           O
TER    2133      GLU   116
ENDMDL
MODEL        2
ATOM   2134  N   ALA     1      14.997  -8.697   2.368  1.00  0.00           N
ATOM   2135  CA  ALA     1      14.960  -9.149   3.746  1.00  0.00           C
        ... up to the end (there is a total of 20 such models).
```

Fig. 3. Parts of three PDB files showing (a) and (b) X-ray crystallographic data; and (c) NMR data. Comments not in the original files but added here for clarity are shown in italic.

word, the chain identifier (E), has been inserted. Other points that emerge from *Fig. 3* are: the atoms are numbered consecutively; amino acids are represented by three letters; the ends of the lines may or may not contain line numbers; NMR files do not have a *resolution* REMARK; and NMR files typically contain several models corresponding to different conformations.