Annexe : Quelques formats de données courantes utilisées en SVT

Exemple des fichiers de séquences pour Anagène ou Geniegen :

Un fichier edi pour Anagène est un fichier texte contenant quelques informations structurées permettant leur utilisation dans le logiciel Anagène. Il peut être ouvert avec un logiciel éditeur de texte comme le bloc-notes de windows (notepad). Il comporte 2 éléments : une ligne d'entête puis une série de lignes correspondant à une séquence, soit une donnée.
Chaque ligne débute par un caractère particulier (;).
Une première ligne indique la nature de la donnée (*Anagène - Fenêtre Edition)*, c'est l'entête du fichier.
Puis une série de ligne indiquant le nom de la séquence (*alphacod.adn*), son type (adn, arn ou protéine, *Type 1, 2 ou 3*), le décalage éventuel de la séquence (*Dec*), un commentaire, la séquence en code à 1 lettre de la molécule, une ligne indiquant la fin de la donnée (;-).
Une nouvelle série de ligne correspond à la deuxième donnée et ainsi de suite.
Exemple de fichier .edi correspondant aux séquences codantes des hémoglobines alpha et bêta.

```
; Anagène - Fenêtre Edition
; alphacod.adn
; Type 1
; Dec  0
; Partie strictement codante du gène de la globine alpha humaine (brin
;  non transcrit).
;
ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGCTGGCG
AGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTT
CGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCC
GTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGG
TGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGA
GTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAA
TACCGTTAA
;-
; betacod.adn
; Type 1
; Dec  0
; Partie strictement codante du gène codant pour la globine bêta humai
; ne.
;
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAA
;-
```

Les mêmes données au format fasta. Ce format est un standard international de type texte. Il est composé de deux parties une première ligne, d'entête, démarrant par le signe **>** qui indique des informations diverses pour identifier la donnée. Puis la séquence en code à une lettre est fournie. C'est un format minimaliste mais d'une grande souplesse. Anagène ne peut pas lire ce format de fichier mais Geniegen en est capable.

```
> <Nom:alphacod.adn>_non_transcrit)._<Nom:alphacod.adn>
ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGCTGGCGAGTATGGTGC
GGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTG
CCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCG
CTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCT
GGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGA
GCACCGTGCTGACCTCCAAATACCGTTAA
> <Nom:betacod.adn>_ne._<Nom:betacod.adn>
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGA
GGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATG
CTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCT
GGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAG
TGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA
```

Le format pdb (protein data bank) :

Ce format permet de décrire la structure tridimensionnelle de molécules utilisable avec Rastop ou Libmol. Il est au format texte. Il peut être ouvert avec un logiciel éditeur de texte comme le bloc-notes de windows (notepad). Il comporte un entête souvent très long comportant de nombreuses informations liées notamment à la publication scientifique rattachée au fichier.
Le début de chaque ligne indique la nature de cette ligne (HEADER, TITLE, AUTHOR, …)
Les données, proprement dites, correspondent aux lignes ATOM et HETATM.
Exemple :

ATOM 1 N LEU A 5 62.864 31.008 17.506 1.00 31.88 N

Données :	Signification	Dans l'exemple
ATOM	Indique les caractéristiques d'un atome	Cette ligne donne les caractéristiques d'un atome
1	Indice de l'atome dans le fichier	Premier atome décrit dans le fichier
N	Nature de l'atome	Atome d'azote (N)
LEU	Résidu comportant l'atome	Atome appartenant à une leucine du fichier
A	Chaine comportant l'atome	Atome appartenant à la chaine A du fichier
5	Position du résidu dans la séquence	5ème acide aminé dans la séquence
62.864	Position selon axe X	Position tridimensionnelle de l'atome
31.008	Position selon axe Y	
17.506	Position selon axe Z	
1.00	Taux d'occupation	Facteur décrivant la fréquence de la présence de l'atome dans la molécule. 1 soit 100 % indique un atome toujours présent.
31.88	Facteur B ou facteur de température	Facteur décrivant la mobilité de l'atome dans la molécule en général l'agitation moléculaire liée à la température.
N	Rappel de la nature de l'atome (vérification)	Atome d'azote (N)

d'après https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/dealing-with-coordinates

Les lignes ATOM correspondent aux atomes des molécules pour lesquelles le fichier est initialement conçu à savoir les protéines ou les acides nucléiques (ADN ou ARN) constitués de deux types d'éléments (appelés résidus) les acides aminés ou les nucléotides.
Les lignes HETATM correspondent aux atomes des molécules non peptidiques ou non nucléiques décrites dans le fichier (eau, ions, glucides, ….). La structure des lignes HEATM est identique à celles des lignes ATOM.

Extrait d'un fichier IA3I.pdb (524 lignes au total) :

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...

ATOM      1  N   LEU A   5      62.864  31.008  17.506  1.00 31.88           N
ATOM      2  CA  LEU A   5      61.929  31.199  18.665  1.00 34.48           C
ATOM      3  C   LEU A   5      61.824  29.931  19.514  1.00 39.10           C
ATOM      4  O   LEU A   5      61.596  28.842  18.977  1.00 41.59           O
ATOM      5  CB  LEU A   5      60.539  31.576  18.166  1.00 28.24           C
ATOM      6  CG  LEU A   5      59.741  32.509  19.068  1.00 24.83           C
ATOM      7  CD1 LEU A   5      60.406  33.872  19.086  1.00 26.21           C
ATOM      8  CD2 LEU A   5      58.319  32.643  18.556  1.00 26.59           C
ATOM      9  N   GLU A   6      61.988  30.073  20.829  1.00 41.42           N
ATOM     10  CA  GLU A   6      62.143  28.912  21.699  1.00 41.10           C
ATOM     11  C   GLU A   6      60.817  28.503  22.321  1.00 39.93           C
ATOM     12  O   GLU A   6      60.473  28.949  23.420  1.00 44.06           O
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
```