Les fichiers PDB

Les fichiers PDB contiennent les informations qui vont permettre à des logiciels de visualisation moléculaire (ex : RasTop ou Jmol) d'afficher les molécules.

Un fichier au format PDB est un fichier texte composé de caractères ASCII. Il est donc possible d'accéder à l'information brute contenue dans ces fichiers en les ouvrant avec un éditeur de texte comme notepad++ (Logiciel libre téléchargeable gratuitement à cette adresse : http://notepad-plus.sourceforge.net/fr/site.htm). (Les éditeurs de texte comme Word sont ici à éviter.)

Les fichiers PDB contiennent les coordonnées cartésiennes des atomes qui constituent la molécule ainsi que des métadonnées. Ces métadonnées peuvent par exemple être la structure primaire de la molécule, ses éventuelles structures secondaires, la méthode expérimentale qui a permis d'obtenir les coordonnées des atomes, etc.

Sommaire :

1 – Le fichier PDB est composé de lignes de 80 colonnes où chaque colonne a une signification propre....2

2 – Différentes versions du format PDB....3

3 – Les champs dans les fichiers PDB....4

4 – Le champ « ATOM »....5

5 – Le fichier PDB permet le lien entre les données obtenues expérimentalement et les logiciels de visualisation moléculaire....6

6 – Limites du format PDB....8

7 – Le format mmCIF remplaçant du format PDB....8

ANNEXE....9
- Informations sur le format PDB....9
- Mémo, sélection de quelques champs (dans leur ordre d'apparition dans le fichier PDB)....10
- Le champ MASTER (mini résumé d'un fichier PDB)....10
- Précisions....10

Auteur : Stanislas Dorey – mail : stanislas.dorey@ens-cachan.fr

1 – Le fichier PDB est composé de lignes de 80 colonnes où chaque colonne a une signification propre

A l'origine, le format PDB a été dicté par la largeur de cartes perforées IBM pour ordinateur. En conséquence, chaque ligne contient exactement 80 colonnes, soit 80 caractères.

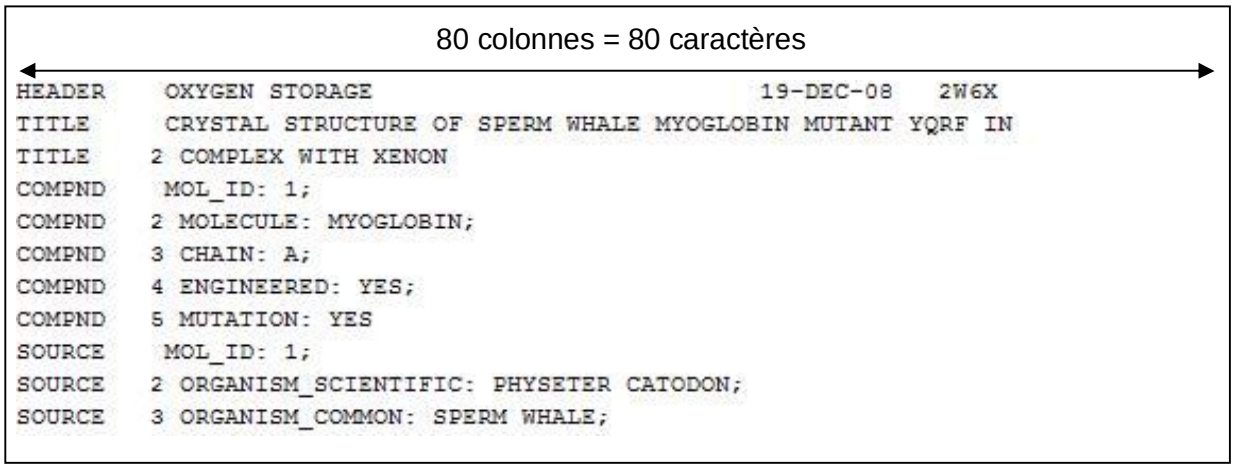

Figure 1 : Extrait des données brutes du début d'un fichier PDB.
Il est visible que chaque ligne contient exactement 80 colonnes, soit 80 caractères.

Chaque colonne possède sa signification, ainsi les 6 premières colonnes, c'est-à-dire les 6 premiers caractères pour une ligne donnée, déterminent le champ. Certains champs, qu'il est possible de trouver, sont présentés dans le tableau ci-dessous à titre d'exemple.

Nom du champ	Description du champ
TITLE	Titre de la macromolécule étudiée
KEYWDS	Les mots-clé de l'entrée
EXPDTA	Donne des informations sur la méthode expérimentale employée
SEQRES	La séquence de la protéine étudiée
ATOM	Coordonnées des atomes des résidus standard
HETATM	Coordonnées des atomes des résidus non standard (solvant, substrat, ion, détergent…)

Tableau 1 : Exemple de champs avec leur description.

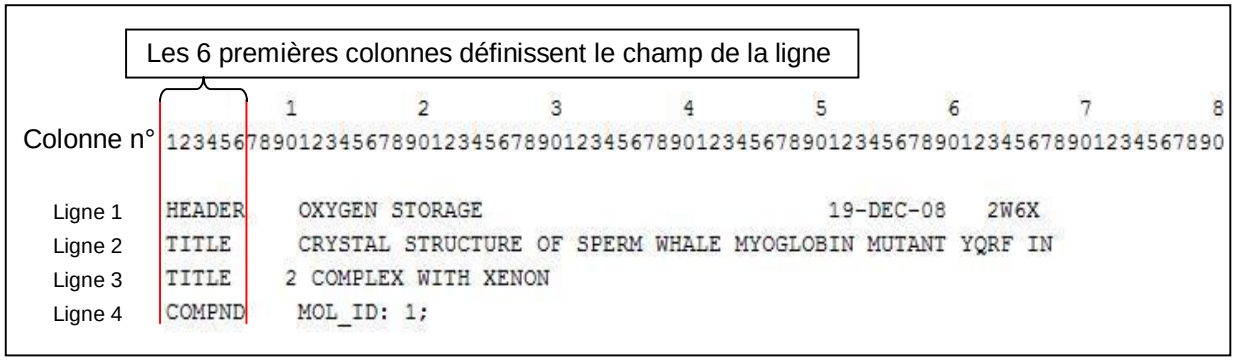

Figure 2 : Les 6 premières colonnes définissent le champ de la ligne.
Le champ de la ligne 1 est donc HEADER, celui des lignes 2 et 3 est TITLE et celui de la ligne 4 est COMPND.

2 – Différentes versions du format PDB

Il existe un guide du format PDB qui indique comment il doit être constitué. La version actuelle est la version 3.20, qui existe depuis le 15 septembre 2008 (la documentation sur les versions des fichiers PDB est disponible ici : www.wwpdb.org/docs.html).
Ici nous nous intéresserons à la version actuelle, cependant il est toujours possible de trouver des fichiers PDB de version antérieure. C'est pourquoi il est possible qu'en analysant le texte d'un fichier PDB vous trouviez des différences avec ce qui est décrit ici.
(Un moyen de connaître la version d'un fichier PDB est de rechercher le champ « REMARK 4 » qui l'indique. Néanmoins ce champ n'avait pas cette fonction dans des versions antérieures, et dans les versions actuelles il reste un champ optionnel.)

```
HEADER    OXYGEN TRANSPORT                         10-JUN-83   1HHO       1HHO   3
COMPND    HEMOGLOBIN A (OXY)                                              1HHO   4
SOURCE    HUMAN (HOMO SAPIENS)                                            1HHO   5
AUTHOR    B.SHAANAN                                                       1HHO   6
REVDAT   2   31-JAN-84 1HHOA   1       JRNL                               1HHOA  1
REVDAT   1   27-OCT-83 1HHO    0                                          1HHO   7
JRNL        AUTH   B.SHAANAN                                            1HHO   8
JRNL        TITL   STRUCTURE OF HUMAN OXYHAEMOGLOBIN AT 2.1 ANGSTROMS   1HHOA  2
JRNL        TITL 2 RESOLUTION                                           1HHOA  3
JRNL        REF    J.MOL.BIOL.                   V. 171    31 1983      1HHOA  4
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836                  070  1HHOA  5
REMARK   1                                                             1HHO  13
REMARK   1 REFERENCE 1                                                 1HHO  14
REMARK   1  AUTH   B.SHAANAN                                          1HHO  15
REMARK   1  TITL   THE IRON-OXYGEN BOND IN HUMAN OXYHAEMOGLOBIN       1HHO  16
REMARK   1  REF    NATURE                        V. 296   683 1982     1HHO  17
REMARK   1  REFN   ASTM NATUAS  UK ISSN 0028-0836                  006  1HHO  18
```

Ci-dessus, début d'un fichier PDB avec une ancienne version du format PDB pour la molécule 1HHO.

```
HEADER    OXYGEN TRANSPORT                         10-JUN-83   1HHO
TITLE     STRUCTURE OF HUMAN OXYHAEMOGLOBIN AT 2.1 ANGSTROMS
TITLE    2 RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: HEMOGLOBIN A (OXY) (ALPHA CHAIN);
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES;
COMPND   5 MOL_ID: 2;
COMPND   6 MOLECULE: HEMOGLOBIN A (OXY) (BETA CHAIN);
COMPND   7 CHAIN: B;
COMPND   8 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 MOL_ID: 2;
SOURCE   6 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
```

Ci-dessus, début d'un fichier PDB avec la version actuelle du format PDB (version 3.20) toujours pour la molécule 1HHO.

Figure 3 : Comparaison des données brutes de 2 fichiers PDB d'une même molécule (1HHO) mais de version différente.

3 – Les champs dans les fichiers PDB

Comme dit précédemment, les 6 premiers caractères pour une ligne donnée déterminent le champ. L'ordre des champs est toujours le même, cependant ils ne sont pas tous obligatoires. Le premier champ de tout fichier PDB est le champ « HEADER » et il est obligatoire. Cette première ligne va permettre d'identifier le fichier PDB – elle indique la molécule concernée par le fichier, la date à laquelle les coordonnées de la molécule ont été reçues par la PDB et l'identifiant de la molécule (unique au sein de la PDB).

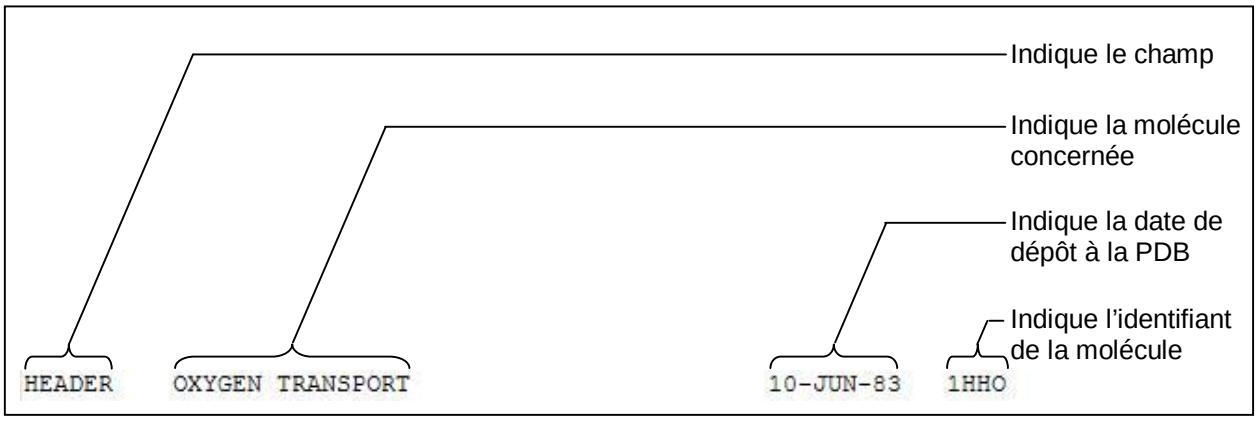

Figure 4 : Détail du champ « HEADER ». Ici ce champ indique que la molécule est Oxygen transport, que ses coordonnées ont été déposées sur la PDB le 10 juin 1983 et que son identifiant est 1HHO.

Les différents champs peuvent être regroupés en sections, comme montré dans le tableau ci-dessous.

Section	Description	Types de champs
Titre	Remarques descriptives résumées	HEADER, OBSLTE, TITLE, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, AUTHOR, REVDAT, SPRSDE, JRNL
Remarque	Bibliographie, divers détails	REMARKs 1, 2, 3 & annotations
Structure primaire	Séquence peptidique et/ou nucléotidique et la relation entre la séquence PDB et la séquence trouvée dans la/les base(s) de données	DBREF, SEQADV, SEQRES MODRES
Hétérogène	Description de groupes non standard	HET, HETNAM, HETSYN, FORMUL
Structure secondaire	Description des structures secondaires	HELIX, SHEET, TURN
Annotation sur les liaisons	Liaison chimique	SSBOND, LINK, HYDBND, SLTBRG, CISPEP
Divers caractéristiques	Caractéristique dans la macromolécule	SITE
Crystallographie	Description de la cellule cristallographique	CRYST1
Transformation de coordonnées	Les opérateurs de transformation de coordonnées	ORIGXn, SCALEn, MTRIXn, TVECT
Coordonnées	Données sur les coordonnées des atomes	MODEL, ATOM, SIGATM, ANISOU, SIGUIJ, TER, HETATM, ENDMDL
Liaison	Liaison chimique	CONECT
Comptabilité	Résumé des informations, marqueur de fin	MASTER, END

Tableau 2 : Les différentes sections qui regroupent les différents champs
(traduit et remanié d'après www.wwpdb.org/documentation/format2.3-0108-a4.pdf page 15).

La signification des colonnes de chaque champ est donnée dans ce document www.wwpdb.org/documentation/format2.3-0108-a4.pdf à partir de la page 18.

4 – Le champ « ATOM »

Le champ « ATOM » (www.wwpdb.org/documentation/changesv3.20.pdf décrit pages 150 à 152) donne les coordonnées des atomes pour les résidus standard (acides aminés, acides nucléiques et polysaccharide). Pour les atomes des résidus non-standard c'est le champ « HETATM » qui donne leurs coordonnées atomiques (décrit pages 161 et 162). La signification des colonnes pour le champ « ATOM » est indiquée dans le tableau ci-dessous.

Colonnes	Signification des colonnes
1 à 6	Indique que c'est le champ ATOM
7 à 11	Numéro de série de l'atome
13 à 16	Nom de l'atome
17	Indicateur d'une localisation alternative
18 à 20	Nom du résidu
22	Identifiant de la chaîne
23 à 26	Numéro du résidu
27	Code pour l'insertion des résidus
31 à 38	Coordonnées orthogonales pour X en Angstrom
39 à 46	Coordonnées orthogonales pour Y en Angstrom
47 à 54	Coordonnées orthogonales pour Z en Angstrom
55 à 60	Occupation
61 à 66	Facteur de température
77 à 78	Symbole de l'élément
79 à 80	Charge de l'atome

Tableau 3 : Description du champ « ATOM »
(d'après www.wwpdb.org/documentation/format2.3-0108-a4.pdf page 150 à 152).

```
                1         2         3         4         5         6         7         8
Colonne n° 12345678901234567890123456789012345678901234567890123456789012345678901234567890
         ATOM     80  CE  LYS A  11      16.865  23.437   0.587  1.00100.00           C
         ATOM     81  NZ  LYS A  11      16.300  24.743   1.031  1.00 50.42           N
         ATOM     82  N   ALA A  12      21.743  18.025   1.437  1.00 18.17           N
         ATOM     83  CA  ALA A  12      23.107  17.671   1.637  1.00 10.44           C
         ATOM     84  C   ALA A  12      23.555  16.291   1.157  1.00 15.53           C
         ATOM     85  O   ALA A  12      24.697  16.040   0.726  1.00 16.26           O
```

Figure 5 : Exemple de champs « ATOM » qu'il est possible de trouver dans un fichier PDB
(le numéro des colonnes a été ajouté à titre indicatif).

A noter que les atomes d'hydrogène sont rarement présents dans un fichier PDB.
Pour les protéines, les résidus sont listés de l'extrémité amine à l'extrémité carboxyle. Pour les acides nucléiques, les résidus sont listés de l'extrémité 5' à l'extrémité 3'. Il n'y a pas d'ordre spécifique pour les polysaccharides.

La liste des champs « ATOM » d'une chaîne se termine par un champ « TER ».

```
ATOM   1067  NH1 ARG A 141      -2.911  10.577  -1.995  1.00 52.81           N
ATOM   1068  NH2 ARG A 141      -0.604  10.395  -2.146  1.00 65.36           N
ATOM   1069  OXT ARG A 141      -6.729  15.170  -5.560  1.00 82.14           O
TER    1070      ARG A 141
ATOM   1071  N   VAL B   1       9.445 -18.730  -3.132  1.00 58.52           N
```

Figure 6 : Ici le champ « TER » marque la fin de la chaîne A.
On peut voir les coordonnées du premier atome de la chaîne B.

5 – Le fichier PDB permet le lien entre les données obtenues expérimentalement et les logiciels de visualisation moléculaire

C'est à partir de ces coordonnées d'atomes que le logiciel de visualisation moléculaire va pouvoir afficher un modèle de la molécule. Le fichier PDB ne donne pas d'information sur la liaison entre les atomes (sauf liaisons spécifiques comme les ponts disulfures). En effet, le logiciel va, à partir de l'ordre des atomes et de leur position spatiale (obtenu grâce aux coordonnées tridimensionnelles), calculer les liaisons entres les atomes.

Les coordonnées 3D rentrées dans le fichier PDB sont obtenues de manière expérimentale (ex : diffraction aux rayons X). Et ce sont donc ces coordonnées qui vont conditionner la représentation de la molécule affichée par le logiciel.
Le champ « EXPDTA », qui est obligatoire, indique la méthode expérimentale utilisée pour obtenir les coordonnées des atomes.

L'intérêt du logiciel de visualisation moléculaire est donc de rendre des informations plus parlantes.

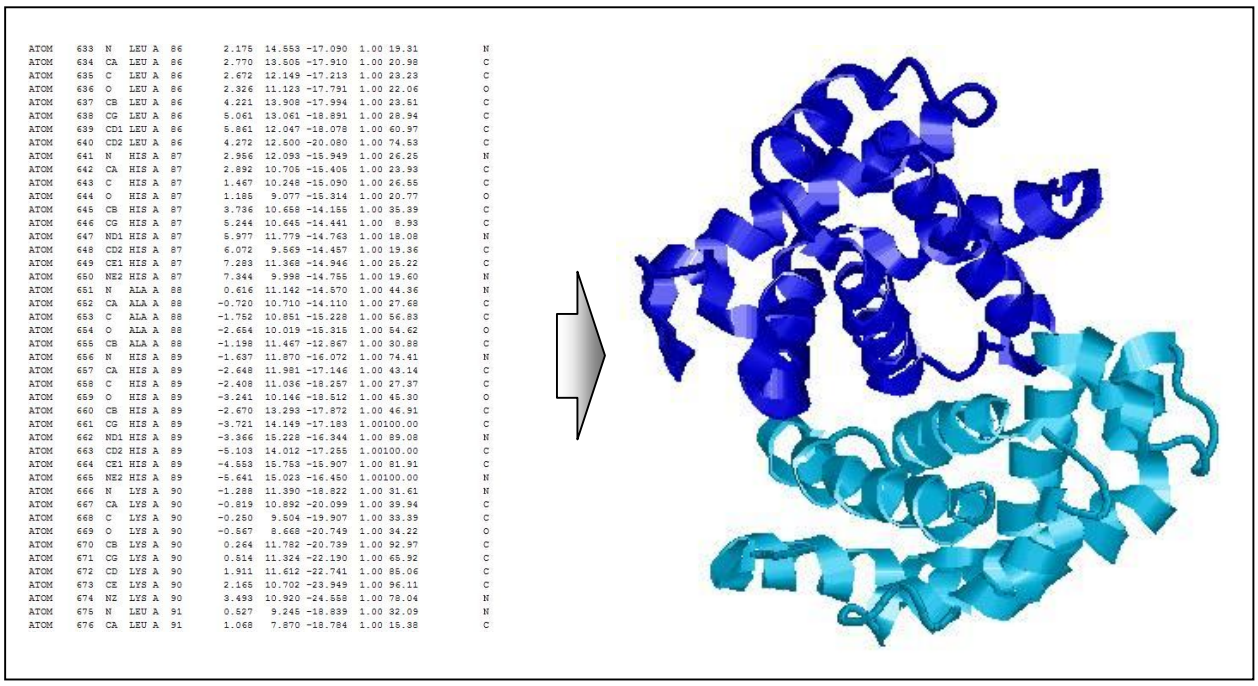

Figure 7 : Illustration de ce que permet le logiciel de visualisation moléculaire.

La molécule pourra avoir différents aspects, mais ceci sera uniquement dû au fait que l'utilisateur choisit différents types de modèle. Sachant que ces modèles ont été conçus par l'homme et permettent à l'utilisateur de mettre l'accent sur tel ou tel aspect de la molécule, aucun d'eux ne peut être une représentation fidèle de la molécule. Et ils sont tous basés sur les coordonnées 3D obtenues expérimentalement.

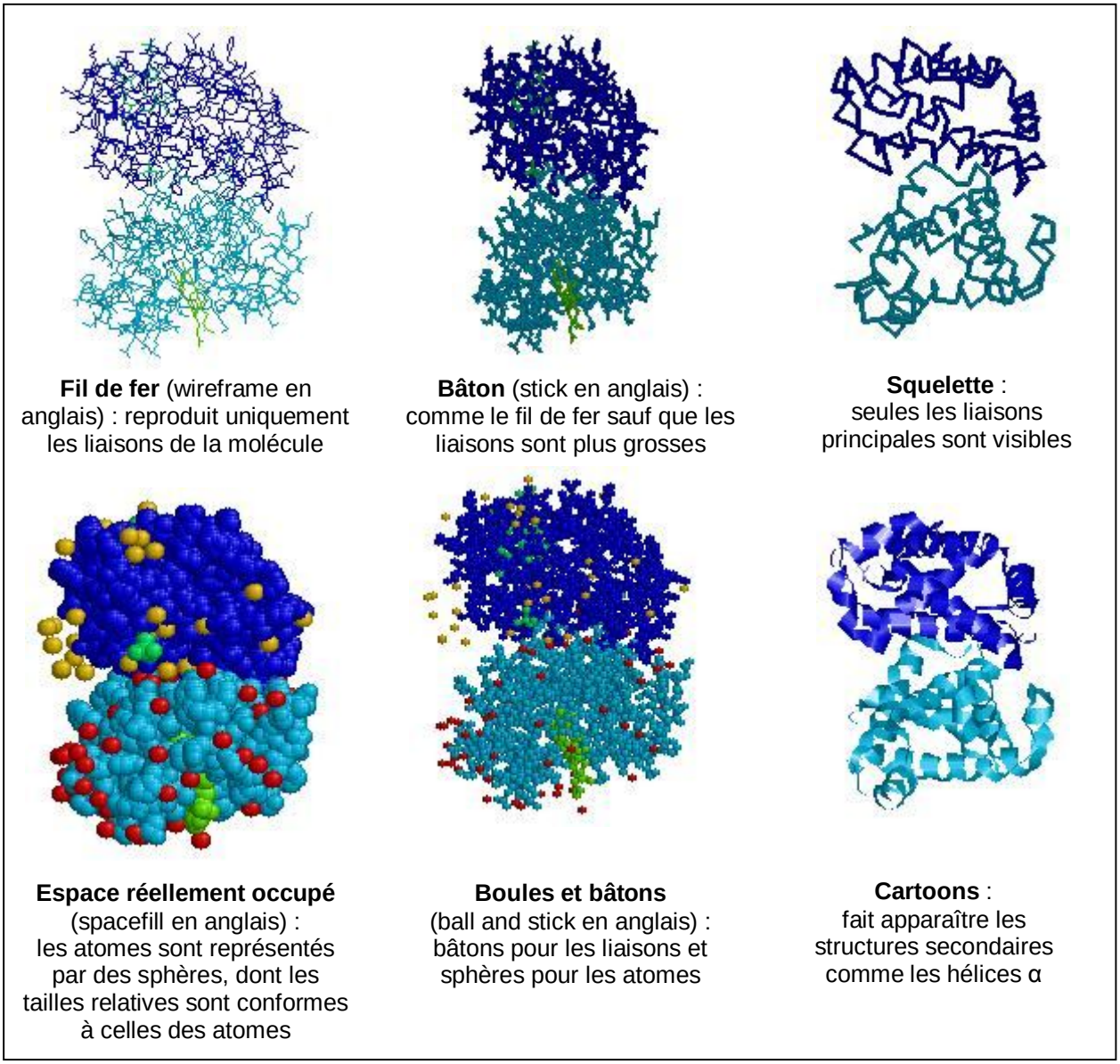

Figure 8 : Exemple de représentations moléculaire possibles (c'est la molécule 1HHO qui est utilisée).

Finalement le logiciel, à part le calcul des liaisons entre atomes, ne va rien « créer » / calculer de plus. Il va simplement aller chercher les informations dans le fichier PDB. Par exemple, si on demande l'affichage des hélices alpha, les informations sont stockées au niveau du champ « HELIX » dans le fichier PDB. Ou encore, si on veut afficher les liaisons disulfures ou les liaisons hydrogènes, les informations sont respectivement dans les champs « SSBOND » et « HYDBND ».

Les informations sont donc dans le fichier PDB, le logiciel ne fait que les rendre plus visuelles et ainsi beaucoup plus compréhensibles pour l'utilisateur. Néanmoins, l'analyse des informations brutes du fichier PDB peut être intéressante.

6 – Limites du format PDB

Bien que très utilisé, le format PDB présente certains défauts. En effet, le format en 80 colonnes est relativement restrictif.
Cela limite le nombre maximum d'atomes d'un fichier pdb qui est de 99999, vu qu'il n'y a que 5 colonnes allouées pour les numéros des atomes (colonnes n°7 à n°11). De même, le nombre de résidus par chaîne est au maximum de 9999 : il n'y a que 4 colonnes autorisées pour ce nombre (colonnes n°23 à 26). Le nombre de chaînes, lui, est limité à 62 (une seule colonne est disponible (colonne 22), et les valeurs possibles sont parmi les 26 lettres de l'alphabet, en minuscule ou en majuscule, ou l'un des chiffres de 0 à 9).

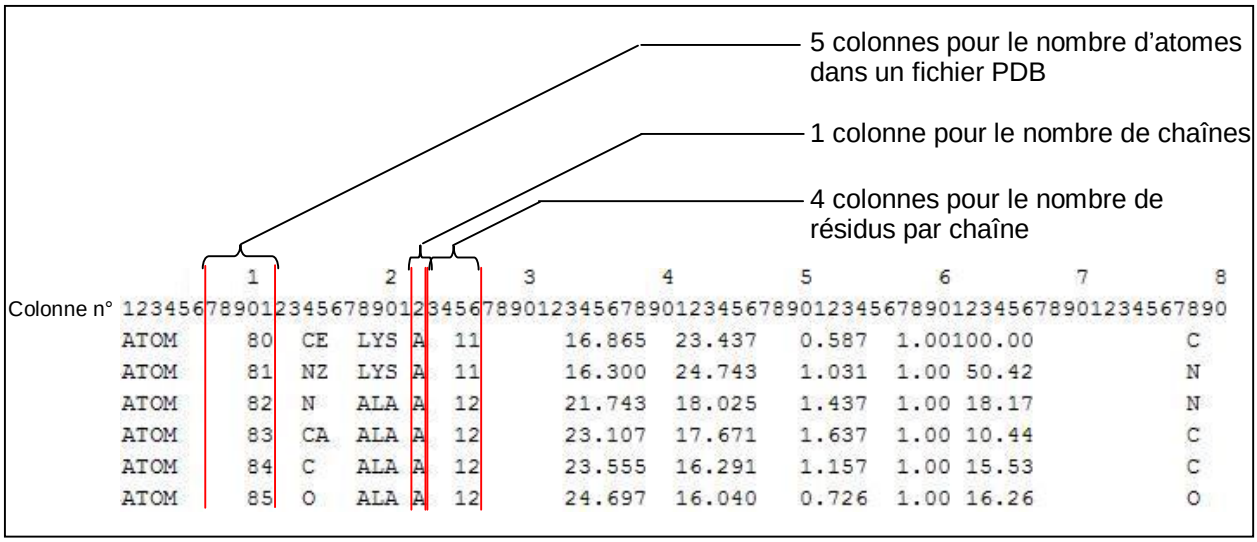

Figure 9 : Limitations du format PDB.

7 – Le format mmCIF remplaçant du format PDB

Quant ce format a été défini, ces limitations ne semblaient pas restrictives, mais elles ont plusieurs fois été franchies lors du dépôt de structures extrêmement grandes, comme des virus, des ribosomes ou des complexes multienzymatiques. C'est pour cette raison que lorsque une structure possède un trop grand nombre d'atomes pour être contenue dans un fichier PDB, elle va être coupée en deux entrées plus petites (utilisation du champ « SPLIT » www.wwpdb.org/documentation/changesv3.20.pdf page 2).

Etant donné les limitations du format PDB, l'Union internationale de la cristallographie (International Union of Crystallography, IUCr) a, en 1990, étendu aux macromolécules la représentation des données utilisées pour décrire les structures cristallographiques des molécules de faible poids moléculaire (appelée CIF, pour Crystallographic Information File).

A partir de celle-ci, le format mmCIF (macromolecular Crystallographic Information File) a donc été développé. La première version du format mmCIF a été publiée en 1996.
Ce format, mmCIF, permet une représentation plus structurée, uniformisée et non limitée. En mmCIF, chaque champ de chaque section d'un fichier PDB est représenté par une description d'une caractéristique d'un objet, qui comprend, d'une part, le nom de la caractéristique (par exemple : _struct.entry_id), et, d'autre part, le contenu de la description (pour _struct.entry_id ça sera le code pdb : 1cbn). On parle de paire « nom-valeur ».
Il est aisé de convertir, sans perte d'informations, un fichier mmCIF au format PDB, puisque toute l'information est directement analysable. Il n'est pas possible, en revanche, de complètement automatiser la conversion d'un fichier PDB au format mmCIF, puisque plusieurs descripteurs mmCIF sont, soit absents du fichier PDB, soit présents dans un champ « REMARK » qui ne peut pas toujours être analysé.
La documentation sur le format mmCIF peut être accessible à cette adresse : http://mmcif.pdb.org.

```
data_2W6X
#
_entry.id   2W6X
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    1.0675
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
```

Figure 10 : Exemple du début d'un fichier au format mmCIF.

ANNEXE

Informations sur le format PDB

La description du format PDB est disponible en ligne. Le guide de la version 2.3 décrit en intégralité ce format, tandis que les versions supérieures ne sont que des corrections de la version 2.3.

Guide de la version 2.3 :
www.wwpdb.org/documentation/format2.3-0108-a4.pdf
Correctif de la version 3.0 :
www.wwpdb.org/documentation/format3.0.1-dif.pdf
Correctif de la version 3.1 :
www.wwpdb.org/documentation/format3.1-20080211.pdf
Correctif de la version 3.15 :
www.wwpdb.org/documentation/changesv3.15.pdf
Correctif de la version 3.20 :
ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v32_A4.pdf
Lien vers l'ensemble des versions : www.wwpdb.org/docs.html

Mémo, sélection de quelques champs (dans leur ordre d'apparition dans le fichier PDB)

Champ	Description / Ce que le champ indique	Obligatoire
HEADER	Identifie le fichier PDB	Oui
SOURCE	Source chimique et/ou biologique de chaque molécule	Oui
KEYWDS	Termes en lien avec la molécule (mots-clé)	Oui
EXPDTA	Technique expérimentale utilisée pour obtenir les coordonnées des atomes	Oui
AUTHOR	Personnes responsables du contenu du fichier PDB	Oui
JRNL	La première citation dans la littérature, qui décrit l'expérience qui a mené à l'obtention des coordonnées	Non
REMARK 2	Donne la plus haute résolution (en Angstrom) pour construire le modèle	Oui
REMARK 4	Version du fichier PDB	Non
SEQRES	La séquence des acides aminés ou acides nucléiques dans chaque chaîne de la macromolécule	Non
HELIX	Identifie la position des hélices dans la molécule (résidus à laquelle l'hélice débute et finit ainsi que sa longueur)	Non
SHEET	Identifie la position des feuillets dans la molécule (résidus à laquelle le feuillet débute et finit ainsi que sa longueur)	Non
SSBOND	Identifie chaque pont disulfure	Non
SITE	Identifie les sites importants pour la macromolécule	Non
ATOM	Coordonnées des atomes des résidus standard	Non
HETATM	Coordonnées des atomes des résidus non-standard	Non
MASTER	Résumé d'informations	Oui
END	Marque la fin du fichier PDB	Oui

Le champ MASTER (mini résumé d'un fichier PDB)

Colonnes	Signification des colonnes
1 à 6	Indique que c'est le champ MASTER
11 à 15	Nombre de champs REMARK
16 à 20	
21 à 25	Nombre de champs HET
26 à 30	Nombre de champs HELIX
31 à 35	Nombre de champs SHEET
36 à 40	Nombre de champs TURN
41 à 45	Nombre de champs SITE
46 à 50	Nombre de champs de transformation de coordonnées (ORIGX + SCALE + MTRIX)
51 à 55	Nombre de champs de coordonnées atomiques (ATOM + HETATM)
56 à 60	Nombre de champs TER
61 à 65	Nombre de champs CONECT
66 à 70	Nombre de champs SEQRES

Précisions :

Des informations montrées dans ce document ont été obtenues, entre autre, à partir de ces sources :

www.umass.edu/microbio/rasmol/pdb.htm

http://fr.wikipedia.org/wiki/Protein_Data_Bank#Le_format_PDB

www.ensta.fr/~muguet/PPL97/31.html

Tout au long de ce document le terme « champ » est la traduction du terme anglais « record ».