Appendices

This section contains the details of some file formats that have been used in examples in these notes. They are included for the student who wants to work on projects that use such file formats.

Appendix I: PDB file format

The national Protein Data Bank (PDB) file format is extremely complex and contains much more information than we can ever hope to use for student projects. We will extract the information we need for simple molecular display from the reference document on this file format to present here. From the chemistry point of view, the student might be encouraged to look at the longer file description to see how much information is recorded in creating a full record of a molecule.

There are two kinds of records in a PDB file that are critical to us: atom location records and bond description records. These specify the atoms in the molecule and the bonds between these atoms. By reading these records we can fill in the information in the internal data structures that hold the information needed to generate the display. The information given here on the atom location (ATOM) and bond description (CONECT) records is from the reference. There is another kind of record that describes atoms, with the keyword HETATM, but we leave this description to the full PDB format manual in the references.

ATOM records: The ATOM records present the atomic coordinates for standard residues, in angstroms. They also present the occupancy and temperature factor for each atom. The element symbol is always present on each ATOM record.

Record Format:

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
73 - 76	LString(4)	segID	Segment identifier, left-justified.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

The "Atom name" field can be complex, because there are other ways to give names than the standard atomic names. In the PDB file examples provided with this set of projects, we have been careful to avoid names that differ from the standard names in the periodic table, but that means that we have not been able to use all the PDB files from, say, the chemical data bank. If your chemistry program wants you to use a particular molecule as an example, but that example's data file uses other formats for atom names in its file, you will need to modify the `readPDBfile()` function of these examples.

Example:

```
         1       2       3       4       5       6       7       8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM      1  C           1      -2.053   2.955   3.329  1.00  0.00
ATOM      2  C           1      -1.206   3.293   2.266  1.00  0.00
ATOM      3  C           1      -0.945   2.371   1.249  1.00  0.00
ATOM      4  C           1      -1.540   1.127   1.395  1.00  0.00
ATOM      5  C           1      -2.680   1.705   3.426  1.00  0.00
ATOM      6  C           1      -2.381   0.773   2.433  1.00  0.00
ATOM      7  O           1      -3.560   1.422   4.419  1.00  0.00
ATOM      8  O           1      -2.963  -0.435   2.208  1.00  0.00
ATOM      9  C           1      -1.455  -0.012   0.432  1.00  0.00
ATOM     10  C           1      -1.293   0.575  -0.967  1.00  0.00
ATOM     11  C           1      -0.022   1.456  -0.953  1.00  0.00
ATOM     12  C           1      -0.156   2.668   0.002  1.00  0.00
ATOM     13  C           1      -2.790  -0.688   0.814  1.00  0.00
ATOM     14  C           1      -4.014  -0.102   0.081  1.00  0.00
ATOM     15  C           1      -2.532   1.317  -1.376  1.00  0.00
ATOM     16  C           1      -3.744   1.008  -0.897  1.00  0.00
ATOM     17  O           1      -4.929   0.387   1.031  1.00  0.00
ATOM     18  C           1      -0.232  -0.877   0.763  1.00  0.00
ATOM     19  C           1       1.068  -0.077   0.599  1.00  0.00
ATOM     20  N           1       1.127   0.599  -0.684  1.00  0.00
ATOM     21  C           1       2.414   1.228  -0.914  1.00  0.00
ATOM     22  H           1       2.664   1.980  -0.132  1.00  0.00
ATOM     23  H           1       3.214   0.453  -0.915  1.00  0.00
ATOM     24  H           1       2.440   1.715  -1.915  1.00  0.00
ATOM     25  H           1      -0.719   3.474  -0.525  1.00  0.00
ATOM     26  H           1       0.827   3.106   0.281  1.00  0.00
ATOM     27  H           1      -2.264   3.702   4.086  1.00  0.00
ATOM     28  H           1      -0.781   4.288   2.207  1.00  0.00
ATOM     29  H           1      -0.301  -1.274   1.804  1.00  0.00
ATOM     30  H           1      -0.218  -1.756   0.076  1.00  0.00
ATOM     31  H           1      -4.617   1.581  -1.255  1.00  0.00
ATOM     32  H           1      -2.429   2.128  -2.117  1.00  0.00
ATOM     33  H           1      -4.464   1.058   1.509  1.00  0.00
ATOM     34  H           1      -2.749  -1.794   0.681  1.00  0.00
ATOM     35  H           1       1.170   0.665   1.425  1.00  0.00
ATOM     36  H           1       1.928  -0.783   0.687  1.00  0.00
ATOM     37  H           1      -3.640   2.223   4.961  1.00  0.00
ATOM     38  H           1       0.111   1.848  -1.991  1.00  0.00
ATOM     39  H           1      -1.166  -0.251  -1.707  1.00  0.00
ATOM     40  H           1      -4.560  -0.908  -0.462  1.00  0.00
```

CONECT records: The CONECT records specify connectivity between atoms for which coordinates are supplied. The connectivity is described using the atom serial number as found in the entry.

Record Format:

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"CONECT"	
7 - 11	Integer	serial	Atom serial number
12 - 16	Integer	serial	Serial number of bonded atom
17 - 21	Integer	serial	Serial number of bonded atom
22 - 26	Integer	serial	Serial number of bonded atom
27 - 31	Integer	serial	Serial number of bonded atom
32 - 36	Integer	serial	Serial number of hydrogen bonded atom
37 - 41	Integer	serial	Serial number of hydrogen bonded atom
42 - 46	Integer	serial	Serial number of salt bridged atom

```
47 - 51         Integer          serial          Serial number of hydrogen bonded
                                                 atom
52 - 56         Integer          serial          Serial number of hydrogen bonded
                                                 atom
57 - 61         Integer          serial          Serial number of salt bridged
                                                 atom
```

Example:

```
          1         2         3         4         5         6         7
12345678901234567890123456789012345678901234567890123456789012345678901234567890
CONECT 1179  746 1184 1195 1203
CONECT 1179 1211 1222
CONECT 1021  544 1017 1020 1022 1211 1222      1311
```

As we noted at the beginning of this Appendix, PDB files can be extremely complex, and most of the examples we have found have been fairly large. The file shown in Figure 17.2 below is among the simplest PDB files we've seen, and describes the adrenalin molecule. This is among the materials provided as `adrenaline.pdb`.

```
HEADER    NONAME 08-Apr-99                                                NONE   1
TITLE                                                                     NONE   2
AUTHOR    Frank Oellien                                                   NONE   3
REVDAT   1  08-Apr-99     0                                               NONE   4
ATOM      1  C           0      -0.017   1.378   0.010  0.00  0.00           C+0
ATOM      2  C           0       0.002  -0.004   0.002  0.00  0.00           C+0
ATOM      3  C           0       1.211  -0.680  -0.013  0.00  0.00           C+0
ATOM      4  C           0       2.405   0.035  -0.021  0.00  0.00           C+0
ATOM      5  C           0       2.379   1.420  -0.013  0.00  0.00           C+0
ATOM      6  C           0       1.169   2.089   0.002  0.00  0.00           C+0
ATOM      7  O           0       3.594  -0.625  -0.035  0.00  0.00           O+0
ATOM      8  O           0       1.232  -2.040  -0.020  0.00  0.00           O+0
ATOM      9  C           0      -1.333   2.112   0.020  0.00  0.00           C+0
ATOM     10  O           0      -1.177   3.360   0.700  0.00  0.00           O+0
ATOM     11  C           0      -1.785   2.368  -1.419  0.00  0.00           C+0
ATOM     12  N           0      -3.068   3.084  -1.409  0.00  0.00           N+0
ATOM     13  C           0      -3.443   3.297  -2.813  0.00  0.00           C+0
ATOM     14  H           0      -0.926  -0.557   0.008  0.00  0.00           H+0
ATOM     15  H           0       3.304   1.978  -0.019  0.00  0.00           H+0
ATOM     16  H           0       1.150   3.169   0.008  0.00  0.00           H+0
ATOM     17  H           0       3.830  -0.755  -0.964  0.00  0.00           H+0
ATOM     18  H           0       1.227  -2.315  -0.947  0.00  0.00           H+0
ATOM     19  H           0      -2.081   1.509   0.534  0.00  0.00           H+0
ATOM     20  H           0      -0.508   3.861   0.214  0.00  0.00           H+0
ATOM     21  H           0      -1.037   2.972  -1.933  0.00  0.00           H+0
ATOM     22  H           0      -1.904   1.417  -1.938  0.00  0.00           H+0
ATOM     23  H           0      -3.750   2.451  -1.020  0.00  0.00           H+0
ATOM     24  H           0      -3.541   2.334  -3.314  0.00  0.00           H+0
ATOM     25  H           0      -4.394   3.828  -2.859  0.00  0.00           H+0
ATOM     26  H           0      -2.674   3.888  -3.309  0.00  0.00           H+0
CONECT    1    2    6    9    0                                           NONE  31
CONECT    2    1    3   14    0                                           NONE  32
CONECT    3    2    4    8    0                                           NONE  33
CONECT    4    3    5    7    0                                           NONE  34
CONECT    5    4    6   15    0                                           NONE  35
CONECT    6    5    1   16    0                                           NONE  36
CONECT    7    4   17    0    0                                           NONE  37
CONECT    8    3   18    0    0                                           NONE  38
CONECT    9    1   10   11   19                                           NONE  39
CONECT   10    9   20    0    0                                           NONE  40
CONECT   11    9   12   21   22                                           NONE  41
CONECT   12   11   13   23    0                                           NONE  42
CONECT   13   12   24   25   26                                           NONE  43
END                                                                       NONE  44
```

Figure 17.1: Example of a simple molecule file in PDB format

Appendix II: CTL file format

The structure of the CT file is straightforward. The file is segmented into several parts, including a header block, the counts line, the atom block, the bond block, and other information. The header block is the first three lines of the file and include the name of the molecule (line 1); the user's name, program, date, and other information (line 2); and comments (line 3). The next line of the file is the counts line and contains the number of molecules and the number of bonds as the first two entries. The next set of lines is the atom block that describes the properties of individual atoms in the molecule; each contains the X-, Y-, and Z-coordinate and the chemical symbol for an individual atom. The next set of lines is the bonds block that describes the properties of individual bonds in the molecule; each line contains the number (starting with 1) of the two atoms making up the bond and an indication of whether the bond is single, double, triple, etc. After these lines are more lines with additional descriptions of the molecule that we will not use for this project. An example of a simple CTfile-format file for a molecule (from the reference) is given in Figure 17.2 below.

Obviously there are many pieces of information in the file that are of interest to the chemist, and in fact this is an extremely simple example of a file. But for our project we are only interested in the geometry of the molecule, so the additional information in the file must be skipped when the file is read.

```
L-Alanine (13C)
GSMACCS-II10169115362D 1 0.00366 0.00000 0

6 5 0 0 1 0 3 V2000
-0.6622  0.5342 0.0000 C    0 0 2 0 0 0
 0.6220 -0.3000 0.0000 C    0 0 0 0 0 0
-0.7207  2.0817 0.0000 C    1 0 0 0 0 0
-1.8622 -0.3695 0.0000 N    0 3 0 0 0 0
 0.6220 -1.8037 0.0000 O    0 0 0 0 0 0
 1.9464  0.4244 0.0000 O    0 5 0 0 0 0
1 2 1 0 0 0
1 3 1 1 0 0
1 4 1 0 0 0
2 5 2 0 0 0
2 6 1 0 0 0
M CHG 2 4 1 6 -1
M ISO 1 3 13
M END
```

Figure 17.2: Example of a simple molecule file in CTfile format

Appendix III: the STL file format

The STL (sometimes called StL) file format is used to describe a file that contains information for 3D hardcopy systems. The name "STL" comes from stereo lithography, one of the technologies for 3D hardcopy, but the format is used in several other hardcopy technologies as described in the hardcopy chapter.

The .stl or stereolithography format describes an ASCII or binary file used in manufacturing. It is a list of the triangular surfaces that describe a computer generated solid model. This is the standard input for most rapid prototyping machines as described in the chapter of these notes on hardcopy. The binary format for the file is the most compact, but here we describe only the ASCII format because it is easier to understand and easier to generate as the output of student projects.

The ASCII .stl file must start with the lower case keyword `solid` and end with `endsolid`. Within these keywords are listings of individual triangles that define the faces of the solid model. Each individual triangle description defines a single normal vector directed away from the solid's surface followed by the xyz components for all three of the vertices. These values are all in Cartesian coordinates and are floating point values. The triangle values should all be positive and contained within the building volume. For this project the values are 0 to 14 inches in x, 0 to 10 in the y and 0 to 12 in the z. This is the maximum volume that can be built but the models should be scaled or rotated to optimize construction time, strength and scrap removal. The normal vector is a unit vector of length one based at the origin. If the normals are not included then most software will generate them using the right hand rule. If the normal information is not included then the three values should be set to 0.0. Below is a sample ASCII description of a single triangle within an STL file.

```
solid
...
facet normal 0.00 0.00 1.00
   outer loop
     vertex  2.00  2.00  0.00
     vertex -1.00  1.00  0.00
     vertex  0.00 -1.00  0.00
   endloop
 endfacet
...
endsolid
```

When the triangle coordinates are generated by a computer program,it is not unknown for roundoff errors to accumulate to the point where points that should be the same have slightly different coordinates. For example, if you were to calculate the points on a circle by incrementing the angle as you move around the circle, you might well end up with a final point that is slightly different from the initial point. File-checking software will note any difference between points and may well tell you that your object is not closed, but that same softward will often "heal" small gaps in objects automatically.

Vertex to vertex rule

The most common error in an STL file is non-compliance with the vertex-to-vertex rule. The STL specifications require that all adjacent triangles share two common vertices. This is illustrated in Figure 17.3. The figure on the left shows a top triangle containing a total of four vertex points. The outer vertices of the top triangle are not shared with one and only one other single triangle. The lower two triangles each contain one of the points as well as the fourth invalid vertex point.

To make this valid under the vertex to vertex rule the top triangle must be subdivided as in the example on the right.

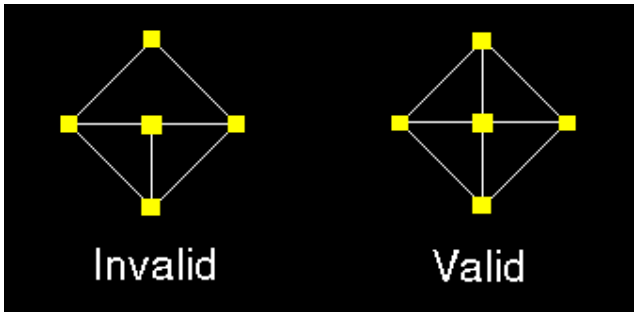

Figure 17.3

References:

CTFile Formats, MDL Information Systems, Inc., San Leandro, CA 94577, 1999. Available by download from `http://www.mdli.com/`

Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description, version 2.1, available online from `http://www.pdb.bnl.gov`