**Internal Chemical Structure of InChI [1]**

This documentation is regarding the Internal chemical structure that is created after a Mol/Sd file is passed as an input to the InChI 1.061 (development version) [2]. Before getting directly into the processing of inputs (Mol/SDF) let’s have a brief understanding of what type of data they contain?

A Mol file, also called molecule file [3], contains a single molecular structure described by number of atoms, bonds, bonds connectivity and other chemical properties. On the other hand a SD file is the concatenation of mol files in a single file, where each molfile is separated by “$$$$” sign. InChI uses MOL/SD file as an input to create a unique InChI string for the identification of concerned molecular structure.

This process of getting from a MOL/SDF inputs to unique InChI string involves various steps. One of the most important step is the creation of an internal structure, from the input, that can be used as a stepping stone for other major steps like Normalization and Canonicalization afterwards.

In the InChI 1.061 development version source code there is a file named “mol2atom.c”. This C file has a method “**MakeInpAtomsFromMolfileData()**” which is responsible for creating the internal chemical structure from the input. The data from the input is saved to a predefined structure named “**MOL\_FMT\_DATA**” and to read this molfile data a pointer “**mfdata**” is used. Furthermore, a predefined structure “**inp\_ATOM**” is used for storing the internal chemical structure depending on the variables and properties defined in correspondence to the molfile data. To fill in the data a pointer “**at**” is used. The different variables in the inp\_ATOM structure definition holds the data for atoms, atoms number, bonds, bond type, bond stereo, valence and other chemical properties associated to a chemical compound which might be present in the initial molfile input data.

To get a better grasp of the above concept I will be sharing few example’s of different molfile’s having different properties. This will help us in analyzing the basic step by step creation of the internal chemical structure. These example’s have been created with a software tool BIOVIA Draw [4] and are just trivial hence only for the purpose of testing and documentation.

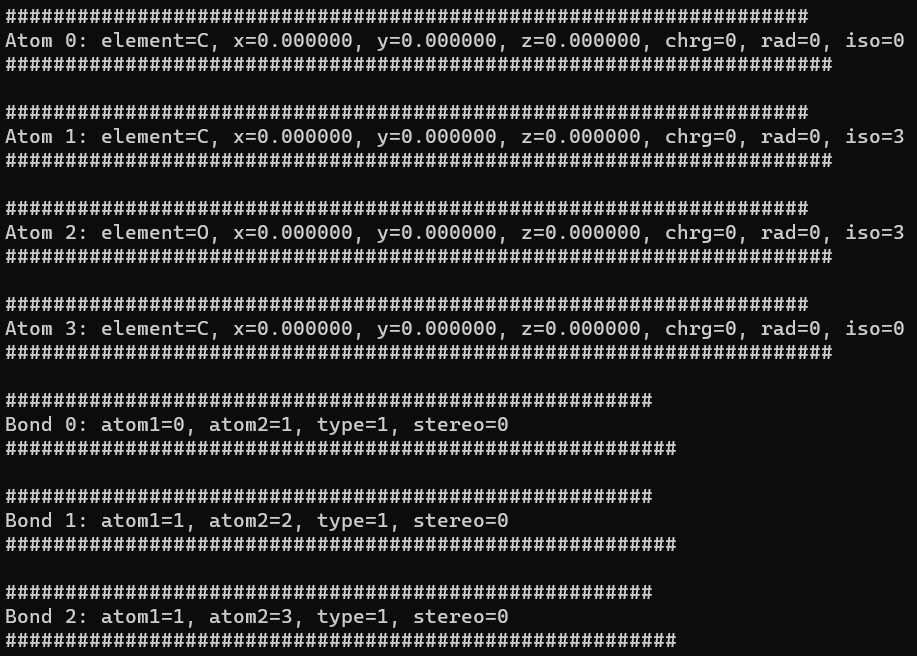
Figure 1.1 shows the molfile and structure of the chemical compound. Here, we can see that the total number of atoms are 4 and the number of bonds between them are 3. We can also see that in the properties block isotopic information is available. The isotopic information indicate that we have 2 isotopes in the molfile present at node 2 and 3.

A picture containing text, screenshot, font, line

Description automatically generated

Figure 1.1 : A compound with its molfile and isotopic information.

When we use the above molfile as an input for the InChI 1.06 development version we get the following output. Figure 1.2 shows the InChI software output, after a little improvisation in the code you can visually see it in the command line interface.

Figure 1.2 : InChI software output for the molfile with isotopic properties.

Another example we will be looking at is of a molfile containing charge information. Figure 1.3 shows a chemical compound and a molfile with charge information present in the property block of the mol file. The Charge information shows that we have 2 charges present at 2nd and 3rd nodes with charges +1 and -1 respectively. A picture containing text, font, screenshot, line

Description automatically generated

Figure 1.3 : A compound with its molfile and charge information.

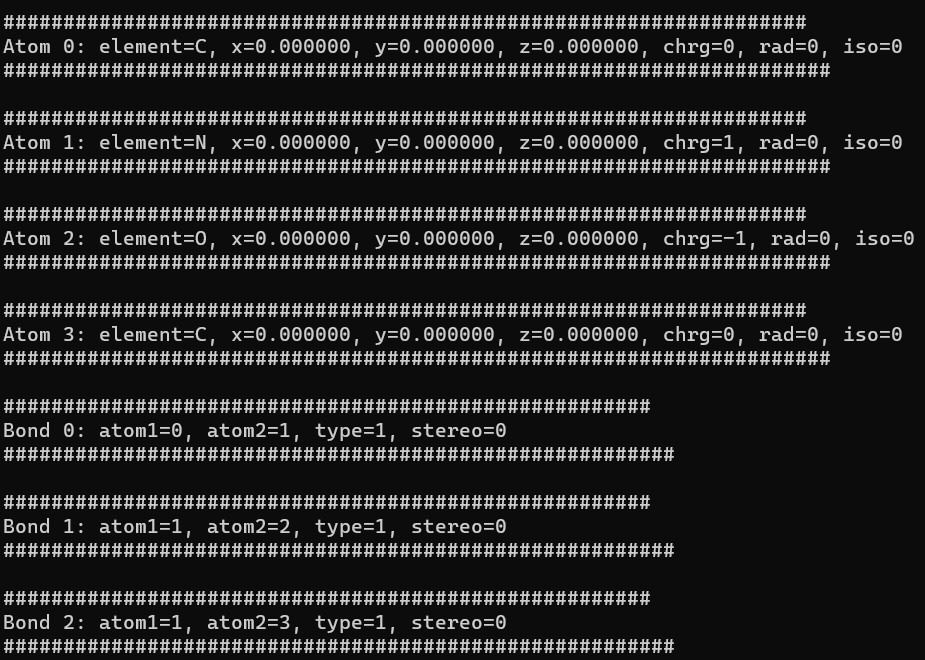
Figure 1.4 shows the InChI output of the input molfile having charge information. Please refer to the rows of atom block in the InChI output as shown in the figure below. 

Figure 1.4 : InChI software output for the mol file with charge properties.

The last example is about bond stereo property. This example consists of two parts, the first part is about the UP bond stereo and the second part is about the DOWN bond stereo. we will be using same chemical compound with UP and DOWN bonds and check for differences in the InChI output. We already know that for unspecified stereo the numerical value of 3 is used in InChI algorithm.

Figure 1.5a shows a chemical compound and its molfile with UP bond stereo. We can see that the numerical value given to UP bond is 1.

A picture containing text, screenshot, font, diagram

Description automatically generated

Figure 1.5a : A compound with its molfile and UP bond stereo.

Figure 1.5b shows a chemical compound with its molfile and DOWN bond stereo. Here, we can also see that the numerical value of 2 is used for DOWN bond.

A picture containing text, screenshot, font, diagram

Description automatically generated

Figure 1.5b : A compound with its molfile and DOWN bond stereo.

Figures 1.6a and 1.6b are the output results for the input molfile’s with UP and DOWN bond stereo configuration respectively. Please refer to the bond section of the InChI software’s output in the images shown below.



Figure 1.6a : InChI software output for the molfile with UP bond stereo configuration.



Figure 1.6b : InChI software output for the molfile having DOWN bond stereo configuration. The numerical value for DOWN bond shown here is 6.

Here, it is to be noted that due to conventional C programming methods the numbering of atoms and bonds are 0 indexed hence, I have also used 0 indexing for displaying the internal chemical structure i-e. Atoms and Bonds go from 0 to “n” numbers. Where “n” is the total number of atoms or bonds in the input mol/sd file.

Table 1 further shows the comparison of numerical representation used for stereo chemistry of the molecular structure. The comparison is between the input molfile generated by BIOVIA draw software and that of the internal chemical structure created by InChI software.

|  |  |  |
| --- | --- | --- |
| Stereo Possibilities | INCHI | BIOVIA |
| Stereo single UP | **1** | **1** |
| Stereo single EITHER | **4** | **4** |
| Stereo single DOWN | **6** | **2** |
| Stereo double EITHER | **3** | **-** |

Table 1: Comparison between the numerical representation of stereo chemistry.

**References**

1. <https://www.inchi-trust.org/> .
2. <https://github.com/IUPAC-InChI/InChI_Dev/> .
3. CTFile Formats by MDL Information Systems, Inc (“Elsevier MDL”).
4. <https://www.3ds.com/products-services/biovia/products/scientific-informatics/biovia-draw/> .