Process of creating the data files for training LOGKPREDICT

Ben Hay, 10-20-2021

Provided are source codes for two fortran programs that can be used to make a training data files. These programs assume that mengine is installed on the system the same as it would be when running HostDesigner post-processing.

The first code is called **makedata**. The way this one works is that you enter working directory containing three input files and launch the code on the command line. The input files are named as follows:

input - an example is given below. The first line gives the name of the PCModel MM3-formatted input file for a ligand – metal complex containing a single metal ion and a single ligand. The second line gives the name of the ligand, the charge on the ligand, and the number of restricted bond rotations in the ligand. The fourth line gives the number of log K values. The remaining lines give the metal ion label, the charge of the metal ion, the log K value, the ionic strength, uncertainty of the value (one sigma based on average of available data), and number of available values.

```
me-en+M.pcm
me-en   0   3
7
Co   2   6.41   0.0  single value
Ni   2   7.37   0.0  +/- 0.10, 4 values
Cu   2  10.58   0.0  +/- 0.16, 5 values
Zn   2   5.80   0.0  +/- 0.15, 4 values
Ag   1   5.52   0.0  single value
Cd   2   5.42   0.0  +/- 0.10, 2 values
Pb   2   5.06   0.0  single value
```

free_ligand.pcm - This is a PCModel MM3-formatted file for the global minimum of the ligand. It is produced by removing the metal from the ligand-metal complex and conformer searching the ligand.

<name+M>.pcm - This is the PCModel MM3-formatted file for a generic metal-ligand complex. The <name> is the name given in the file named input. In the above example, the ligand name is ‘me-en’ and the name of the complex file is ‘me-en+M.pcm’. This structure is created by attaching a single ligand to a single metal ion atom, usually used Ni(II) but it does not matter, and conformationally searching the complex.

As output, a training data file will be created for each metal listed in the input file with format as previously defined in the YR8-Q10 progress report (also provided here). The directory will also contain the hydrated ligand-metal complex and aquo ion for each metal ion. Finally, the directory will contain an xyz formatted file for the free ligand global minimum.

The second code is called **makeall**. Whereas makedata creates data files for one ligand at a time, this code does the same thing for a series of ligands. The way this one works is that you create a working directory containing a subdirectories for the series of ligands, where each subdirectory contains the same three input files described above for a different ligand. The name of each subdirectory must be the same as the ligand name on the second line of the input file. The code is launched by entering the working directory and entering makeall on the command line. The output for each ligand, which will be the same as described above, is stored in the correspondingly named subdirectory.

Provided are two directories that each contain a series of ligand subdirectories. Each subdirectory contains the three input files as described above for 85 ligands.

The first directory is named dat_file_inputs. The input files contain the original lists of log K values with all data included.

The second directory is named dat_file_inputs_trimmed. Here complexes where LOGKPREDICT gives the top 5% worst predicted log K values have been removed by editing the list of metals in the appropriate input files.

Bottom line: If you want to generate training files for a single ligand, then enter one of the ligand subdirectories and launch makedata. If you want to generate a complete set of training data files, simple enter one of these directories and launch makeall.

Given that there have been changes to mengine (to add secondary electronegativity effects), to mm3.prm, as well as a few of the pcmodel input files for both ligands and complexes, my suggestion going forward would be to begin by remaking all of the training data files using the current version of mengine, mm3.prm, and the provided set of pcmodel input files before adding further ligand-metal complexes to the training data set.