CFM-Pipeline Manual

Introduction
CFM_id is a tool used to create insilico mass spectra based on SMILES or INCHI strings. With the use of this pipeline a file containing ID's and SMILES strings, and a file containing the options will be processed and turned in to a .mgf file that is ready for further use.

The difference between just throwing the file with ID's and SMILES directly in to cfm-id and using this pipeline is that multiple issues will be resolved such as charged smiles, aromatic smiles not having an inchikey, smiles containing multiple molecules. As well as adding extra data to the final .mgf file with things as InchIKeys the original smile, a separate ID. One more important thing this pipeline does is combining the spectra created on different energy levels.

CFM-Pipeline Options file.
This file contains the options for the cfm Pipeline. To edit just change the paths or variables behind the =. This file is included on github repository

config: the config file for cfm-id
parameters: the pretrained model for cfm-id which contains the probabilities of likely fragmentation points.
prob_thresh: the threshhold at which the
cutoff_intensity: every peak with an intensity under this value is not kept. (scale from 0-900)
molconvert_path: path to the molconvert install location.

Installation
To even get started with this pipeline the best thing would be to get a cfm-id installation on your workplace. This can be easily done by using the bash installer which can be found here: https://github.com/NP-Plug-and-Play-Scripts/Bash-scripts.git .
The script you need is called cfm-install.sh.

To instal cfm, run this script at a location where you wish to have your CFM_Workplace and run the script with ***./cfm-install.sh .*** This will install rdkit,boost and lpsolve as well as cfm id in the directory.
Next download the scripts from the following link:
https://github.com/NP-Plug-and-Play-Scripts/CFM-Pipeline
or
https://github.com/NP-Plug-and-Play-Scripts/Python_scripts.git
Save them on a location that you prefer.

In case molconvert is not instaled on the computer go to https://chemaxon.com/products/jchem-engines/download log in or create an account if you dont have one yet and download the appropriate version of Jchem base. In the case of linux download the .sh file and install it using "sh jchem_unix_18.28.sh" (depends on the version of jchem) and this should start the installer. The installer should be able to guide you through the installation process, just make sure to install it in accessible location.

This should be enough to run the pipeline. In case rdkit is not installed, either try to install it on the server or make a miniconda environment and install rdkit on it. Start the environment and then run the pipeline. https://conda.io/docs/user-guide/tasks/manage-environments.html or http://wiki.ab.wurnet.nl/index.php/How_can_I_install_software_for_my_own_use

Running the pipeline

1. **Get Data.** Before picking how to run the pipeline you first need data. This should be a comma separated file containing ID SMILE pairs. This file should then be placed in CFM_workplace/cfmData/smileFile/ , that way the files stay organised.
2. **Pick run method.** There are two ways to run the pipeline the first is by running it command line and the second way is to run the GUI and run it with the interface.
3. **Move to the python script directory.** In order to run the pipeline in the command line, move to the directory of the python scripts.
4. **Run program**. Next type "python cfmPipelineRunner.py" next add the path to the CFM_workplace, the path to the molconvert bin folder and after that the name of the file you wish to run (as long as it's located in CFM_workplace/cfmData/smileFile/) and press enter. The command should look something like this "python cfmPipelineRunner.py /mnt/scratch/ozing003/CFM_workplace/ mnt/scratch/ozing003/jchem/jchemsuite/bin/ Flavonoids.csv options.txt"
5. **Check output.** Once the run is done, the output of the run should be located at /CFM_workplace/cfmData/results.

The run will take a while depending on the size of the file, the length of the SMILES and the speed of the server. Best way to run it would be to run it in a screen so you can tab it away and continue on something else. Once the run is done (on average it should take about 24 hours with a file of 10000 SMILES) the results should be located in /CFM_workplace/cfmData/results/