Clustering MDMix hotspots

The script here 'surf_cluster.py' is a clustering algorithm that uses surface distances to identify clusters of hotspots (from MDMix) that can point towards the most druggable pockets in a protein.

1. Installing the environment & dependencies

The script was originally made by Daniel Alvarez, and uses some outdate libraries which are dificult to find nowadays. I have patched it, but it still needs a very specific environment to funcion.

The script is writen in python 2.7 and I suggest you to use anaconda2. The version I used was the following:

```
$ wget https://repo.anaconda.com/archive/Anaconda2-2019.03-Linux-x86_64.sh
```

In the ./installation folder, there is an yml file describing all the dependecies which can be used to install the environment. Notice that you should change the prefix on the last line, to specify where you want to store the environment.

You will require that file to install and then activate the environment using the followign commands:

```
$ conda env create -f environment_droplet.yml
$ conda activate surf_cluster
```

WARNING If your terminals are opened by default with anaconda3 (check the ~/.bashrc), you will need to deactivate that environment first and initialize the anaconda2 before activating the environment.

2. Usage

The script we will use is the 'surf_cluster.py' but the 'asa_calculations.py' needs to be in the same folder, as it is a dependecy in the previous one. Both are in the ./lib folder.

```
 $ python lib/surf_cluster.py

Usage: surf_cluster.py [options]

Options:
  -h, --help            show this help message and exit
  -n NAME, --name=NAME  Project name. Used as prefix for output file.
  -r REF, --ref=REF     Reference structure PDB file name.
  -t TOP, --top=TOP     Mol2 file of the reference PDB, from which the
                        atomtypes are extracted.
  -p POINTS, --points=POINTS
                        Points to cluster

  Optional args:
    Use them if you know what you are doing :)

    -f, --force         Force recalculation of surface distance matrix.
                        Default False.
    -i NEIGH, --neighc=NEIGH
                        Neighbour cutoff distance for node connection in
                        subgraphs. Surface distance. Default=6 A
    -j JOIN, --joinc=JOIN
                        Neighbour cutoff distance for subgraph joining.
                        Surface distance. Default=8 A
    -m MIN, --min=MIN   Minimum number of nodes to consider a subgraph.
                        Default = 3 nodes.
```

The compulsory files to have are the reference PDB (-r), the same file but in mol2 format (-t), the hotspots from mdmix in pdb format (-p). The other options can be changed at will, but the defaults are a good starting point.

WARNING The hotspots parsing is quite hardcoded and the file (-p) has to be in a very specific format. I will leave an example in the ./example folder so you can compare.

If your system is very big, it might be good to remove the hydrogens from the reference pdb and mol2 files, as they tend to multiply the memory usage (as a distance matrix is calculated).