Heterocyle Isostere Explorer (HCIE)
Research Notes

Matthew Holland
Brennan Group
Target Discovery Institute, University of Oxford
matthew.holland@spc.ox.ac.uk

29 July 2021

1 ShaEP: Shape and Electrostatic Comparison

Previously electroshape parameters had been used to find the top 100 molecules from the VEHICLE database that matched well the input molecule. However, this is a crude measure, only taking three points on the molecule, and not accurately representing the 3D shape of the molecule, or its distribution of electronic potential.

ShaEP is a tool developed by Dr Mikko Vainio in the Department of Biochemistry and Pharmacy at the Åbo Akademi University in Finland. It aligns molecules by calculating a field graph representation of each molecule, and assiging two values to each node on the field graph; the value of the electrostatic potential (ESP) at the node, and a local shape descriptor. The local shape descriptor is a histogram vector of signed distances of atoms from a plane tangential to the molecular shape-density (surface) at the vertex coordinates $\mathbf{r} \in \mathbb{R}^3$.

1.1 Usage

ShaEP is distributed as a binary, unix executable. This makes it easy to use in the command line, but access to the source code is not possible, so it is currently a bit of a black box, and makes it difficult (but not impossible) to integrate into HCIE. An interface between the python script and a bash script for calculating ShaEP parameters will be needed, and the necessary information will need to be automatically extracted from the .txt output file.

In order to calculate the ESP at each of the points on the field graph, ShaEP needs the coordinates of each of the atoms in the molecule, and the partial charge on each atom. The coordinates are included in the .pdb files generated from VEHICLE, but the partial charges need to be calculated - initially I have used Gasteiger charges as implemented in RDKit, but it would be possible to do this using XTB/Orca DFT, or using the DNN as Ewa used previously for electroshape.

One issue was that the only accepted input file format that can contain both atomic coordinates and partial charges was TRIPOS Mol2 format. Initially I could not find an open-source program to convert .pdb files to .mol2 files, so I wrote one myself, which is freely available on GitHub. It turns out OpenBabel, a python library available on

GitHub, can convert pdb to mol2 files, but trying to convert the full Vehicle database into a single .mol2 file with OpenBabel proved to be errorprone, so I am using my own converter.

Initial tests at screening a single target molecule with ShaEP against the full library of 24 867 VEHICLe structures generated the following error each time:

```
>> shaep --maxhits 100 -q query.mol2 --output-file similarity.txt
    target_library.mol2
Assertion failed: (boost::math::isfinite( grad[0] )), function operator()
    , file /Users/runner/runners/2.165.0/work/1/s/include/overlapper.icc,
    line 265.
Abort trap: 6
```

Contacting Mikko Vainio, he confirmed that this was a bug in the code, and published a new build of ShaEP for Mac (1.3.1) which, when the above command was run again, generated the desired outcome.

To screen 1 small organic molecule (7 heavy atoms, 1 aromatic 6m ring) against a library of 24 867 (the entire VEHICLE database) took 00:01:38.67, with an average time of 3.69 ms per structure.

1.2 Comparing ShaEP and Electroshape

The current implementation of HCIE uses electroshape as a quick measure of screening input molecules against the database. To see if ShaEP might be able to offer an improvement, I compared the results from the current electroshape implementation of HCIE to those of the newest build of ShaEP.

I used 1,3,5-triazine (S80) as a trial input molecule, and ran this in HCIE, and in ShaEP as a .mol2 file (generated with malt) against the full VEHICLE database in a .mol2 file (again generated in malt).

1.2.1 Problems with Coordinates in mol2 and pdb files

I noticed that several of the entries in the .mol2 file had coordinates of 0.0 for every atoms

- S1603
- S1605
- S5534

Inspection of the .pdb files showed that this error had originated from there, with the pdb files containing no meaningful atomic coordinates. Inspection of the xtb geometry optimisation output files (.xyz) showed that the optimisation on these molecules appeared to be successful, with the molecular structures appearing chemically sensible.