PSU-Alpha Documentation

Protein Structure Utility Version Alpha

Rachel Alcraft

Doc creation 6 Feb 2020
Doc update 16-Feb-20

Summary

PSU is a student first attempt at creating some protein structure analysis and creation tools based on the MSc Bioinformatics course at Birkbeck College 2019/2020.

The initial tools are to load pdb files and examine structure, the goal is to simulate movement of structures for RMSD and more interestingly optimising single structure with energy functions. I have no idea how the spatial algorithms will work until I get there. I anticipate a notifier pattern but how not to make it infinitely recursive I don't know yet.

Ultimately, I would like to be able to reverse engineer contact maps to structure and, separately, build hypothetical small molecules and examine their interaction with proteins using molecular modelling techniques.

The project is written in C++ using Visual Studio, but I have put most of the classes in a shared library that should be easily portable into a CMake project. Data is output from the project in either a pdb format or a data frame format for loading into R.

The project is run simply from a light commandline exe using a config file for its run information and relying on configuration data.

Input Files

The config file lives next to the exe.

Within the config file 2 paths are set, for the inputs and for the outputs.

```
CONFIGPATH=F:\\PSUA\\Config\\
```

Within these locations the format is fixed as:

```
CONFIGPATH\\data_aminoinfo.csv    //r data frame format amino acid info REQUIRED
CONFIGPATH\\PDB\*.pdb             //where it looks for pdb files
CONFIGPATH\\Fasta\*.fasta         //where it looks for fasta files
(more to come eg force field)
```

Output Files

Within the config file 2 paths are set, for the inputs and for the outputs.

```
OUTPUTPATH=F:\\PSU\\Data\\
```

Each time the utility is run a jobid is created based on the time.

```
OUTPUTPATH\\JobId\\Logger.txt     //the log file is here
OUTPUTPATH\\JobId\\Reports\\       //All the reports and data are output here
```

Functionality

RAMA=TRUE
Ramachandran Plots, an R compatible data file is saved, with corresponding R script available for creating a Ramachandran plot, or any Chi plots and variations of. 3D psi/chi/omega available but not very interesting.

CONTACT = TRUE
Contact map report based on pdb1 file. Currently not configurable because more data makes better plots in R, the distance is set quite high as the corresponding R report can filter to desired distance for display. High distance is useful for a heat map. The R report will show blobs coloured on distance or any other property of the amino acid of interest.

CONTACTCHAIN1=A
CONTACTCHAIN2=B
In this case the contact map will be selected on each given chain. It will not look anything like a regular contact map, it won't be symmetrical for example. It is intended to show closeness of chains for Protein-Protein-Interaction where a complex is given. I have no idea if it means anything yet.

RMSDFIX=TRUE
An RMSD report will be ruin without moving the structures at all, so will just calculate on C-alphas in order up to the minimum possible – unless an alignment is either given or requested (not implemented that yet).

RMSDOPT=TRUE
Both structures will be moved into a central location about the axes and the origin as per an algorithm that seeks to find the orthogonal extremities of each structure, iterating through them to account for outliers. Without an alignment it will optimise on the best matching c-alphas in order, then shift the entire structures. Implemented but doesn't make much sense without an alignment.

RMSDCONTACT=TRUE
A contact map is done for the 2 structures against each other. Although, a reverse contact map would make more sense – TODO.

ALIGNMENT=TRUE
In this case either an alignment will be done with a Needleman-wunsch algorithm or a fasta alignment can be passed in. This will then be used for the C-Alpha matches for RMSD reports. Not implemented.

Algorithms

RMSD
Distances between all c-alphas are found, and the greatest distance forms the first axis. Then the furthest c-alpha orthogonal to this axis is found to make a final atom in the transformation shape (GeoTripod). With these 3 points, they are mapped into the centre: the first onto the origin, the next onto the x-axis and the third rotated over the x-axis to meet it. This transformation is calculated from the 3 points in the GeoTripod and then applied to all atoms in the structure, for both pdb files separately, mapping them both into the same space. The RMSD is calculated, then the space is searched by iterating through different versions of the tripods – the second largest magnitude, the third greatest distance from the orthogonal, etc. It does not work will without a good alignment because it will never overlay secondary structures.

Development Plans

RMSD: needs to be able to detect structural similarity if proteins are mirror images. After adding a sequence alignment, consider also allowing flipping across all exes and creating an alignment based on structure looking at closest atoms in the structure.

Visual interface is needed, consider https://www.qt.io/developers

Add a contact map for 2 structures to show where they are close (binding sites?) Or for 2 selected chains in a single pdb.

Add a calculation for solvent accessible surface area (SASA). Code found here:
https://www.cgl.ucsf.edu/chimera/data/sasa-nov2013/sas_area.cpp
https://www.cgl.ucsf.edu/chimera/data/sasa-nov2013/sasa.html

Enable the rotation of 1 protein to another (binding sites?)

Add an algorithm to find an area of similar closeness (binding site possibility)

How do I distinguish between surface and core? Find hydrophobic surface molecules.

RMSD taking a FASTA format alignment or implement Needleman-Wunsch

Create hydrogen bonds – will it depend on the quality/resolution of the structure?

Geometric manipulation of structures

Conversion between angles and coordinates of structure descriptions

RMSD calculation optimisation

Secondary structure prediction from sequence

Geometric unravelling of sequence using SS prediction and most likely geometry

Conformation optimised by energy calculations

Add a visual interface to watch conformations change (http://www.gnuplot.info/?)

Energy Optimisation

First version plan - Given the forcefield I have calculated, see below in development plans, and given a structure either generated ☺ or retrieved, there will be min and max possibilities for every aa triple in every secondary structure for bond length, angles, phi, psi, omega and the chis. Start with values most out of range of the chosen boundaries, put them on a queue, move them to boundary minimum, and all their neighbours appropriately. Then for all the neighbours, looking at their other neighbours, any effected bonds/angles that are now out of range go on the queue, etc.

Generate my own forcefield taking the top 1000(?) best structures from pdb, calculating all angles and bond lengths, and assuming those are the best structures so calculating parameters to fit them. Or is that a statistical model?

Ball library has CHARMM energy model:
https://github.com/BALL-Project/ball/blob/master/data/CHARMM/param22.ini

References

Dihedral angles: http://www.ccp14.ac.uk/ccp/web-mirrors/garlic/garlic/commands/dihedrals.html

Amino Acid properties: http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/IMGTclasses.html

Amino Acid structures: http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/formuleAA/

SASA calculation (not implemented): https://www.cgl.ucsf.edu/chimera/data/sasa-nov2013/sasa.html