Psu-Alpha Documentation

Protein Structure Utility Version Alpha

Rachel Alcraft, 6 Feb 2020

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.

CALPHA = TRUE

Contact map report based on pdb 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.

RMSDFIX=TRUE (Not yet implemented)

This takes 2 pdb files and calculated an RMSD based on the CAlphas. Where an alignment file is given in FASTA format it uses this to match off the CAlphas. This does not move the structures.

RMSDOPT=TRUE (Not yet implemented)

This takes 2 pdb files and calculated an RMSD based on the CAlphas. Where an alignment file is given in FASTA format it uses this to match off the CAlphas. This **does** move the structures with optimisation leaving the structures rigid as given. Algorithm details link here (not implemented).

Algorithms

RMSD

I can't see that anything would be more optimal than simply drawing a cuboid around the structure, finding the geometric centre and orthogonal axes, and then lining up on those. I'll test around that point to see.

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.

Development Plans

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.

RMSD taking a FASTA format alignment

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/?)

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/