pdb_analysis

Jackson Sheppard
CH E 210D, Exercise 1
10/05/22

Here we present an analysis of 309 protein sequences, computing structural quantities such as the radius of gyration and statistical interaction potentials. The input sequences are stored in The Protein Data Bank `*.pdb` format and reside in the `proteins/` directory at the root of this repository. Each file consists of a protein X-ray crystal structure and thus includes atomic `X`, `Y`, `Z` coordinates of each amino acid (residue) comprising the protein. In this analysis, we read the sequences of our 309 proteins and compute the radius of gyration considering all residues along with that for hydrophobic residues only. We then plot both of these quantities along with their ratio against the total number of residues in the structure. Finally, we compute and visualize the "statistical" interation potential for the 20 amino acid types by considering amino acid contacts present in this data set.

Installation and Usage

Clone this repository and navigate to its root. Install dependencies from the `enironment.yml` file using `conda`:

```
conda env create --name envname --file=environment.yml
```

If instead files are downloaded individually, ensure the `*.pdb` files are in the relative path of working directory and that necessary dependencies are installed.

Run the code to generate results as follows:

```
$ python exercise1.py
```

This yields both plots of the radius of gyration and statistical interaction potentials computed from sequences in this data set. It also gives console output in the following format:

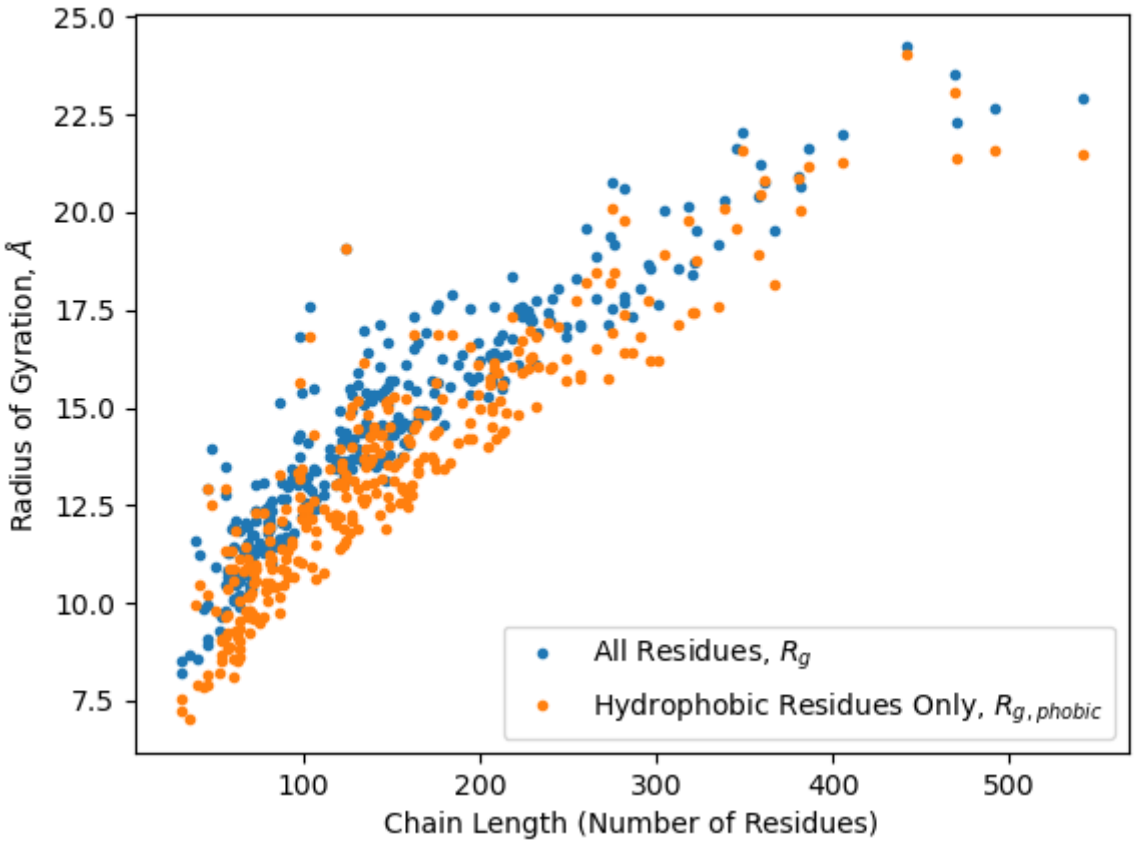
```
filename sequence_length Rg_phobic Rg_all Rg_phobic/Rg_all
...Repeats for all *.pdb...

Lowest Interaction Energies:
RES-RES : Energy (kcal/mol)
---------------------------
...5 Lowest Interactions and Corresponding Energies...

Highest Interaction Energies:
RES-RES : Energy (kcal/mol)
```

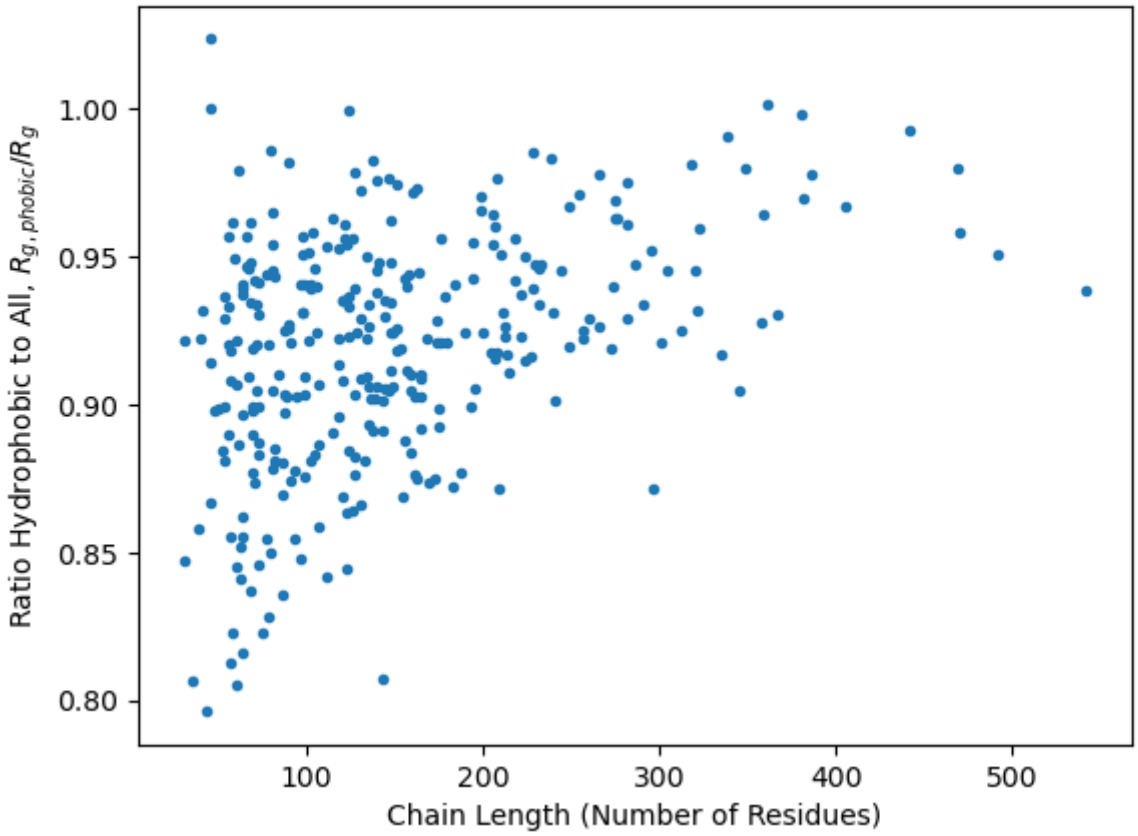
```
--------------------------
...5 Highest Interactions and Corresponding Energies...
```

Results

Radius of Gyration

We first coarse grain each protein structure in the data set by filtering to include only alpha-carbon residues, those fixed to the backbone chain of the protein. We the compute the radius of gyration for the coarse-grained sequence by first summing over all residues and then over hydrophobic residues only. We then create a scatter plot of each protein radius of gyration (for both all residues and hydrophobic residues only) versus the protein chain chain length (total number of alpha-carbon amino acids). We then compute and plot the ratio of these radii, `Rg_phobic/Rg_all`, and again plot the result against protein chain length. The results are shown below.

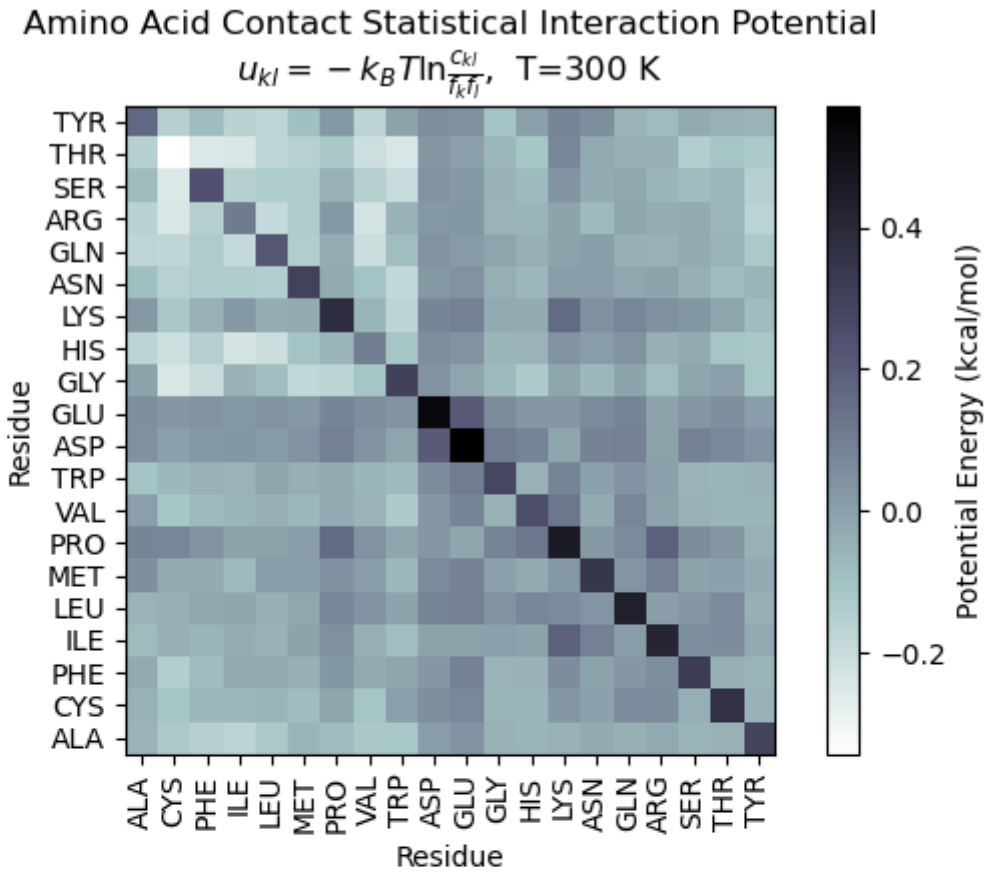




We thus see the radius of gyration for hydrophobic residues only is in general roughly 85%-95% of that for all residues. This could be due to the fact that hydrophobic residues are more likely to be closer to the protein backbone as opposed to its aqueous environment. However, as the chain length increases, both radii increase and this hydrophobic property has a less significant effect.

Statistical Interaction Potential

We compute the statistical interaction potential by computing the fraction of all amino acids appearing in our coarse-grained structures along with the fraction of all "contact pairs" present. Since there are 20 amino acids, there are 210 possible residue pairs. A pair of residues is then declared a "contact" if theire separation is less than some cutoff distance, here 9 Angstroms. We thus compute the statistical potential energy for each amino acid contact pair (assuming T = 300 K) from the statistics of our data set and normalize the result such that the mean potential energy is zero. This yields a symmetric matrix of interaction potentials, since the interaction potential between amino acids (i, j) is the same as that between amino acids (j, i). We visualize the result as a heat map below:



Lowest Energy Interactions (RES-RES)	Lowest Energies (kcal/mol)	Highest Energy Interactions (RES-RES)	Highest Energies (kcal/mol)
CYS-CYS	-0.343109	GLU-GLU	0.571927
CYS-PHE	-0.247388	ASP-ASP	0.530072
CYS-ILE	-0.242653	LYS-LYS	0.458285
CYS-TRP	-0.240138	GLN-GLN	0.438962
ILE-VAL	-0.225613	ARG-ARG	0.417976

Comparing this result with Macromolecules 18, 534 (1985), we find reasonable agreement. We also find our most favorable (most negative) interaction potential to be that for the Cys-Cys contact and the least favorable (most positive) to be that for the Arg-Arg, Lys-Lys, Asp-Asp, and Glu-Glu contacts.