ECS 129

Assignment: Option 5 (Programming)

Validating protein structure models

Due: Friday, March 13th, 2020

This assignment is meant to combine what you have learned about protein structure and protein structure modeling in the lectures. The general outline is very simple. You are given two models for a protein structure and you would like to identify which of the two is closer to the actual structure of that protein.

The protein sequence is:

DDKTWTTILLFAHEKSQDDTKKQIKVLQDAGAGGDKVPPIGARPEHGKIVVQKCPSVLRAAALVPWKNEN
AGIDGTKAVKVFQTAGMNWAVVGLHYLESSESWDLSLLGKIFSVAQKKGEGALVCISGKRERVENGHIYE
EHARKVKLFIAFGCGFGAALLAMVATIALDQGGDAPDVFLRSCAKIIHRKVQNYLSESVAKTIAIPAGHEVE
NGTGQAVGRADEETGVVIWGEGFDESYVKIAKIGH

The two model structures, model1.pdb and model2.pdb are available from the web site but you will probably not use those files (see below). To evaluate these two models, we will write a program that computes the total energy of the structures they represent.

If a molecule has N atoms, its total energy is the sum of its vdW energy, its electrostatics energy, and its solvation energy (ignoring the bonded interactions):

$$U = \sum_{i=1}^{N} \sum_{j=i+1}^{N} flag(i,j) \left(\epsilon_{ij} \left(\left(\frac{s_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{s_{ij}}{r_{ij}} \right)^{6} \right) + \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_i q_j}{r_{ij}} \right) + \sum_{i=1}^{N} ASP(i) ASA(i)$$

where:

- f*lag(i,j) = 0* if *i* and *j* are atoms involved in a chemical bond, or connected through two chemical bonds, and f*lag(i,j)=1* otherwise
- $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$ and ε_i are given for each atom
- $s_{ij} = \frac{s_i + s_j}{2}$ and s_i are given for each atom
- r_{ij} is the distance between the two atoms i and j
- q_i and q_j are the charges of atoms *i* and *j*, respectively (given)
- $\frac{1}{4\pi\varepsilon_0} = 332.$
- $\varepsilon_r = 4$ (dielectric constant for water)
- *ASP(i)* : atomic solvation parameter of atom *i* (given)
- *ASA(i):* accessible surface area of *i* (see below)

To simplify your task, I have parsed the PDB files for model1 and model2, and generated "summary" files model1.crd and model2.crd, that should be much easier to use. These files are organized as follows;

Line 1: number of atoms N
Line2: comments (to be ignored)
Line 3 to Line N+2: infos for each atom of the molecule. Each of these lines contains:

Atom # X Y Z R (r_i) EPS (ε_i) SIGMA (s_i) Charge (q_i) ASP(i)

Line N+3 : comments (to be ignored)
Line N+4 : comments (to be ignored)
Line N+5 to Line 2N+5: FLAG info for each atom:

Atom # Nex j1 j2 ….. jNex

Where *Nex* is the number of atoms *j* with *flag(i,j) = 0*, and *j1* to *jNex* are these *Nex* atoms; those are the atoms that are in close chemical proximity to atom I and should not be considered for computing non bonded interactions. (look at one of the files, they should be easy to understand.)

The only thing missing in the crd files are the accessible surface areas, ASA, for all atoms. As a rough approximation, we will use:

$$ASA_i = 0.2 * 4 * \pi * \left(r_i + R_{H2o}\right)^2$$

Where r_i is the vdW radius of atom I (given as *R* in the crd files), and R_{H2o} is the radius of a water molecule (we will use R_{H2o} = 1.4 Angstroms).

NOTE: in the two files model1.crd and model2.crd, numbering starts at 1.

Write a program that computes the total energies for model1 and model2. Based on these energies, can you conclude on which model is best? Explain.

Please turn your program in.

Handing things in:

Remember that you are submitting the report for a final project. There is no need to send a lengthy write-up, but it should definitely include an introduction, results and analysis, a conclusion, and references to published work, if needed.

Good luck!