BMI-8400 ASSIGNMENT II

Due Date: March 22nd, 2023 11:59PM
Points: 100[1]
Important: Please read this entire document very carefully, and check in with the instructor as soon as possible if something is not clear.

Background

The assignment focuses on implementing a simple protein-ligand docking algorithm and testing it on a real case. As discussed in class, protein-ligand docking algorithms explore the rotational, translational, and conformational space of a small molecule (keeping the protein fixed in space) and return the *pose* with the most favorable (i.e., **most negative**) interaction energy.

Docking algorithms rely on two main components:

(1) An energy function that takes as input the protein and the small molecule and returns a pseudoenergy.
(2) A search strategy.

In this assignment the search strategy will be limited to exploring rotations around the y-axis by systematically varying the β Euler angle and computing the energy for each rotation angle.

The energy function. You will be using a simple energy function that contains a Lennard-Jones term (capturing the van der Waals energy term) and an electrostatics term. The energy function computes the potential energy between the ligand and the protein using the following equation:

$$V = \sum_{ij} \left(V_{LJ}(r_{ij}) + V_E(r_{ij}) \right) \tag{1}$$

where the potential energy V is equal to the sum of a Lennard-Jones term V_{LJ} and an electrostatics term V_E over all pairs of atoms (*i.e.*, the atom i in the pair belongs to the ligand and the atom j to the protein). r_{ij} represents the Euclidean distance between the atom i in the ligand and the atom j in the protein.

The Lennard-Jones and the electrostatics term are expressed by the following two equations:

$$V_{LJ}(r_{ij}) = \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \tag{2}$$

$$V_E(r_{ij}) = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \tag{3}$$

The $C^{(6)}$ and $C^{(12)}$ parameters in the Lennard-Jones term depend on the chosen probe and the particular atom type and are taken from a matrix of LJ-parameters distributed with the GROMACS package[3]. For this assignment, they are contained in the `ffG43b1nb.params` file posted on Canvas. Each line in the file contains the $C^{(6)}$ and $C^{(12)}$ terms for each pair of atom types. **Please note that the PDB format uses angstroms as units. However, for the LJ calculations the units for the distance are nanometers, so you should divide the distance you obtain by 10 (one nanometer = 10 angstroms). The electrostatics term is in angstroms, so no conversion is needed for that**.

The dielectric constant $\frac{1}{4\pi\epsilon_0}$ should be set to 138.935485. q_i and q_j are the partial charges for atoms i and j, respectively, and they are contained in the last column of the modified PDB files (extension `.pdbm`) files posted on Canvas.

[1] See holistic rubric on grading programming assignments

The distance-dependent dielectric sigmoidal function has been taken from Solmajer and Mehler[2] and has the following form:

$$\epsilon(r_{ij}) = A + \frac{B}{1 + \kappa e^{-\lambda B r_{ij}}} \tag{4}$$

where $A = 6.02944$; $B = 72.37056$; $\lambda = 0.018733345$; $k = 213.5782$. When the distance r_{ij} between the atoms is $< 1.32\mathring{A}$, a dielectric constant of 8 is used. The parameters reported above for the distance-dependent dielectric have been taken from Cui et al.[1]

Search strategy. You will only explore the rotational space by sampling the β Euler angle. No translation of the small molecule or rotations around the other two axes will be required.

Files

On Canvas (under `Files/docking`) you will find the following files:

- `ffG43b1nb.params`: this file contains the Lennard-Jones parameters for each atom type
- `protein.pdb`: the PDB file for the protein
- `ligand_actual.pdb`: the PDB file for the "correct" pose of the small molecule (derived from the crystal structure of the complex)
- `ligand_starting.pdb`: the PDB file for the starting configuration of the small molecule
- `protein.pdbm`: a modified protein PDB file that contains atom types matching those in the `ffG43b1nb.params` file and partial charges in the last column
- `ligand_starting.pdbm`: same as above, but for the starting configuration of the ligand

1. Code

Requirements:

- Your code should read the `protein.pdbm` and `ligand_starting.pdbm` files
- It should produce an output that contains the β angle sampled and the corresponding docking energy (no particular format is required)
- It should be able to return the rotation matrix for the "best" docking pose

Suggestions:

- Making your code modular will be very helpful. For example, there should be a function to calculate the docking energy given the ligand and protein coordinates, another function that takes in the original ligand's coordinates and rotates them by β, etc.

2. Experiments

- Calculate and report the RMSD between the structures found in *ligand_actual.pdb* and *ligand_starting.pdb*, using only the non-hydrogen atoms in *ligand_starting.pdb*.
- Using your docking program, sample the β Euler angle from 0 to 2π using at least 100 increments, reporting the energy values for each β angle.
- Return the β angle and the coordinates for the ligand corresponding to the most favorable (most negative) interaction energy.
- Calculate the RMSD between this "most favorable" structure and the "true" pose found in `ligand_actual.pdb`.
- How long did it take to run your program? What does that tell you about the need for a more sophisticated search strategy? (*Keep in mind that for this example I only perturbed the "correct" structure by rotating it. In real-life situations we usually have to explore all angles and also the translational space of a molecule).*

3. What to hand in

The assignment should be submitted on Canvas as a single compressed file (either `zip` or `tar`), containing the following items:

(1) A **source code** folder with all necessary files to run the program and a README file with clear instructions on how to run (and if necessary, compile) the program.
(2) The rotation matrix and the ligand coordinates (x, y, z) corresponding to the most favorable docking energy.
(3) A document reporting your answers for the experiments listed above.

For more information on the PDB file format, see
`https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html`

While not required in this assignment, visualizing the initial configuration for the small molecule and comparing it with the actual one and your "best" docking pose will be very helpful, especially when you load both protein and small molecules together. You can use a software like `Pymol` to do that, available at https://pymol.org/2/.

References

[1] M. Cui, M. Mezei, and R. Osman. Prediction of protein loop structures using a local move Monte Carlo approach and a grid-based force field. *Protein Eng Des Sel*, 21(12):729–35, 2008.
[2] T. Solmajer and E.L. Mehler. Electrostatic screening in molecular dynamics simulations. *Protein Eng*, 4(8):911–7, 1991.
[3] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen. Gromacs: fast, flexible, and free. *J Comput Chem*, 26(16):1701–18, 2005.