Introduction to the LWO model with CG electrostatics

Fernando Miguel Yrazu

April 30th, 2016

1 Introduction

This document contains a description of the adapted Lammert-Wolynes-Onuchic (LWO)[1] model to include coarse-grained electrostatics.

2 Description of the model

- The idea of the model is to represent the excluded volume of all the heavy atoms, while coarse-graining the long range non bonded interactions. Long range interactions are those established between residues which are separated by more than 8 amino acids in the sequence.
- From now on by "atoms" we refer to heavy atoms. All atoms have the same weight and excluded volume radius.
- Short range non bonded interactions are all-atom, while long range ones are coarse grained and established between the C_β atom of each residue (with the exception of glycine, which has no C_β).
- The implementation of the model was done by using an adaptation of the `SBM_AA+gaussian` template from the SMOG2 software package[2, 3].
- The values for the bonded constants are kept from the default SMOG2 setup, with the exception of the angles constant which is raised from a

value of 80 to a value of 100 (a 25% increase)

$$V^{bonded} = \Big\{ \sum_{bonds} V^B + \sum_{angles} V^A + \sum_{impropers/planars} V^I + \sum_{backbone} V^{BB} + \sum_{side-chains} V^{SC} \Big\} \quad (1)$$

- More specifically, the terms of the bonded potential are mostly harmonic ones, with the exception of the backbone and side chain dihedrals

$$V^{bonded} = \sum_{bonds} \varepsilon_r (r - r_0)^2 + \sum_{angles} \varepsilon_\theta (\theta - \theta_0)^2 + \sum_{impropers/planars} \varepsilon_\chi (\chi - \chi_0)^2 + \sum_{backbone} \varepsilon_{BB} F_D(\phi) + \sum_{side-chains} \varepsilon_{SC} F_D(\phi) \quad (2)$$

- The specific function for the backbone and side-chain dihedrals is:

$$F_D(\phi) = [1 - cos\,(\phi - \phi_0)] + \frac{1}{2}\,[1 - cos\,(3\,(\phi - \phi_0))] \quad (3)$$

- The `SBM_AA+gaussian` template of SMOG2 gives a list of non bonded contacts between all heavy atoms. Our own `Smog_AA` package coarse-grains all contacts established between residues separated by more than 8 amino acids in the sequence. The coarse grained contacts are Gaussian as well, but their equilibrium distance is that between the corresponding C_{beta} atoms.

- The non bonded potential for non native contacts is purely repulsive and equal to:

$$V^R = \varepsilon_{NC}\,(d/r_{i,j})^{12} \quad (4)$$

- For native contacts we apply a Gaussian potential. The third term is meant to keep the repulsive wall fixed in place; it tends to zero when taking into account the relatively large distances between C_β atoms and thus is not needed for coarse grained contacts.

$$V^C = V^R\,(r_{i,j}) + \varepsilon_{i,j} V^G\,(r_{i,j}) + V^R\,(r_{i,j})\,V^G\,(r_{i,j})) \quad (5)$$

- And the Gaussian is defined as:

$$V^G = exp\left[-\left(r_{i,j} - r_{i,j}^N\right)^2 / \left(2\sigma_{i,j}^2\right)\right] \tag{6}$$

- The excluded volume for the atoms participating in coarse grained contacts remains the same (we do not want to modify the steric characteristics of the protein), but the width of the Gaussian σ is enlarged together the corresponding equilibrium radius r_0. As usual, $\sigma^2 = \left(r_0^{ij}\right)^2 / (50 \ln 2)$. This assures that the contact potential diminishes to one half within $\pm 20\%$ of r_0.

- The value of the nonbonded interaction parameters (ε_{ij}) are kept at 0.4 for the CG long range interactions and at 2.0 for the all-atom short range interactions.

- Electrostatic interactions are added to the C_β atom of each residue, with the exception of glycine. The net charge of each residue is calculated using MOPAC. Then these net charges are implemented by adding a tabulated Debye-Huckel (DH) potential to Gromacs.

- The DH potential is implemented following Wang et al.[4]:

$$E_{elec}^{ij} = \frac{q_i q_j}{4\pi r \varepsilon_0 \varepsilon_r} e^{-r/\sqrt{\varepsilon_0 \varepsilon_r k_B T / 2e^2 I}} \tag{7}$$

3 Simulation details

- In order to identify the folding temperature, a first array of runs is submitted separated by 1K intervals. These runs are 50,000 ps or 10^8 steps long. Once they are finished, the number of native coarse grained native contacts Q per frame is calculated (based on the crystal structure). After applying WHAM to the Q per frame trajectory, a tentative folding temperature is defined.

- Then, a set of approximately 11 temperatures is defined in a narrow 2-3 K range centered on the tentative folding temperature. Six simultaneous runs are performed for every temperature in the set, each starting from a different frame obtained from the preliminary run which is closest to the estimated folding temperature. This is done in order

to achieve a greater number of folding and unfolding transitions, while at the same time keeping the runs relatively short and the size of the output files tractable.

- Using a neighbor searching radius for Gromacs of $rlist = 2.0$ in NOTS, each run on a single node takes about 80 hours to finish, giving a minimal time for each iteration step of 160 hours or about 8 days. By using a value of $rlist = 1.2$, the run times are reduced to 50 and 100 hours respectively, taking a total of 4-5 days. Wang et al. [4] used a *rlist* value of 3.0; I have not yet compared the effect of either setting.

- Sometimes though, the initial estimation of the folding temperature is not equilibrated enough, giving an estimate that is too far from the actual value. In the latter case the 6-run sets have to be re-run, adding 2 or more days of simulations.

- As mentioned in the previous section, bonded constants are determined by the SMOG2 setup, in this case using the `SBM_AA+gaussian` template. In the particular case of the spectrins, raising the angles bonded constant from 80 to 100 seems to increase the size of the barrier in r16 but not necessarily in r15.

- In order to be able to edit the topology file more easily, I've splintered smog.top into the different sections (e.g. `smog_atoms.top`, etc.), including the division between long-range and short-range non-bonded contacts. All constituting parts are then gathered by `smog_main.top`.

- In summary, the values of the bonded and non bonded constants are the following:

  - Bonds $\varepsilon_r = 10000$
  - Angles $\varepsilon_\theta = 100$
  - Dihedrals-improper $\varepsilon_\chi = 10$
  - Dihedrals harmonic aromatic $\varepsilon_\chi = 40$
  - Dihedrals-proper backbone $\varepsilon_{BB} = 10$
  - Dihedrals-proper side chain $\varepsilon_{SC} = 10$
  - Repulsive wall constant $4 * \varepsilon_{NC} * d^{12} = 5.9605 * 10^{-10}$

- Short-range (all-atom) $\varepsilon_{ij} = 2.0$
- Long-range (coarse-grained) $\varepsilon_{ij} = 0.4$
- Permittivity of vacuum $\varepsilon_0 = 5.73 * 10^{-4}(e^2.mol)/(kJ.nm)$
- Relative permitivity of water $\varepsilon_r = 80$
- Boltzmann's constant $K_B = 1.380648 * 10^{-26}(kJ/K)$

(Note: These values yield a Debye length of approximately 30.4 nm for an ionic strength of 0.1M)

References

[1] Heiko Lammert, Peter G Wolynes, and José N Onuchic. The role of atomic level steric effects and attractive forces in protein folding. *Proteins*, 80(2):362–373, February 2012.

[2] Paul C Whitford, Jeffrey K Noel, Shachi Gosavi, Alexander Schug, Kevin Y Sanbonmatsu, and José N Onuchic. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins*, 75(2):430–441, May 2009.

[3] Heiko Lammert, Alexander Schug, and José N Onuchic. Robustness and generalization of structure-based models for protein folding and function. *Proteins*, 77(4):881–891, December 2009.

[4] Qian Wang, Pengzhi Zhang, Laurel Hoffman, Swarnendu Tripathi, Dirar Homouz, Yin Liu, M Neal Waxham, and Margaret S Cheung. Protein recognition and selection through conformational and mutually induced fit. *Proc. Nac. Acad. Sci.*, 110(51):20545–20550, 2013.