

Free energy calculations of protein-water complexes with Gromacs

M. A. Shahzad ^{1,2}

¹ *Department of Physics, Hazara University, Pakistan.*

² *Dipartimento di Fisica, Università di Camerino, Camerino, Italy*

(Dated: April 20, 2018)

We used GFP (Green Fluorescent Protein) to understand its basic structure, adding solvent water around the GFP, minimize and equilibrating it using molecular dynamics simulation with Gromacs. Gromacs is an open source software and widely used in molecular dynamics simulation of biological molecules such as proteins, and nucleic acids (DNA AND RNA-molecules). We employ the CHARMM (Chemistry at HARvard Molecular Mechanics) program for the force fields which enable the potential energy of a molecular system to be calculated rapidly. In this particular simple molecular mechanics interaction fields (CHARMM), the force fields consists of stretching energy, bending energy and torsion energy. The non-bond interaction energy is modeled by Lennard-Jones potential. The `pdb2gmx` gromacs command is implemented to obtain the basic coordinate file and topology for the particular system from the GFP PDB file (1gfl.pdb). We enclosed water molecule in a rhombic dodecahedron box having size $0.5nm$, and protein are embedded in solvent water. The steepest descent method (first-order minimization) are implemented to calculate the local energy minimum with 10^4 -steps. The stability of protein with solvent water molecule are analyzed by measuring the root-mean square displacement (RMSD) of all atoms. It was shown with help of figure that RMSD initially increases rapidly in the first part of simulation, but become stable around $0.14nm$, roughly the resolution of the X-ray structure. The difference is partly due to the moving and vibration of atoms around an equilibrium structure. Secondary structure of GFP protein is presented with the help of DSSP-program.

I. INTRODUCTION

Gromacs is an engine to perform molecular dynamics simulations and energy minimization of biomolecular systems which usually consists of several tens to thousands of amino acid residues. Molecular dynamics simulation presents the time evolution of a molecular systems such as protein, by numerically solving Newton's equation of motion for all atoms in the system. Molecular dynamics (MD) is a powerful tool which can be used to simulate the time evolution of molecular systems with great accuracy. Gromacs is one of application which able to do molecular dynamics simulation based on equation of Newton's law [1–3]. The algorithm used for molecular dynamics assigns positions and velocities to every atom in the simulation system. The forces are computed, which can be used to evolve the positions and velocities through Newtons law, using a given finite time step. The non-bonded interactions between particles model behavior like van der Waals forces, or Coulombs law.

Here we used molecular dynamics simulation on protein with water solvent. We minimize and equilibrate the system utilizing energy minimization technique with Gromacs. The stability of the system is characterized by studying the RMSD and secondary structure of protein.

II. RESULT

We used Green fluorescent protein (GFP)-water complexes in our simulation. The Green fluorescent protein (GFP) [4–7] is an important proteins currently used as a marker of gene expression and protein localization, and

as a bio-sensor [8]. The green fluorescent protein (GFP) is composed of 238 amino acid residues and a molecular weight of 26.9 kdalton that exhibits bright green fluorescence when exposed to light [9].

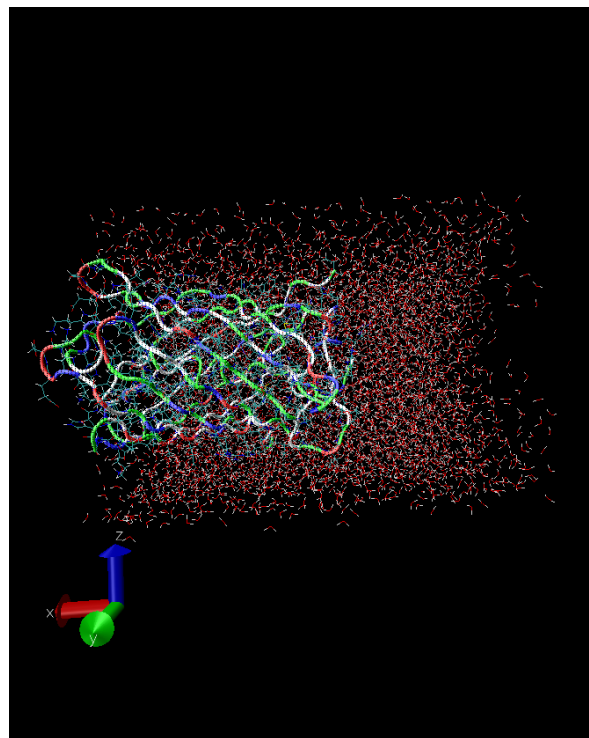


FIG. 1. GFP, with colors that vary according to the residues type, embedded in water molecule. Figure produced by VMD-software .

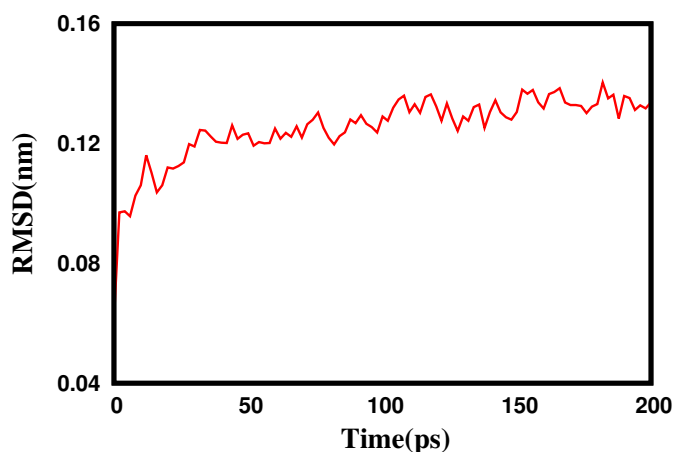


FIG. 2. Root mean square displacement of all atoms fitted on the Protein_H (protein except hydrogens) group.

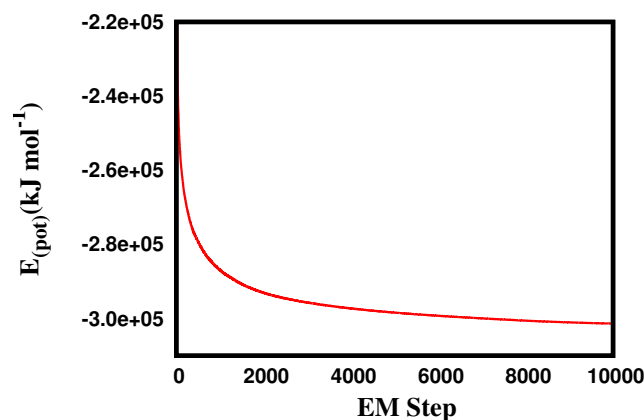


FIG. 3. Average potential energy E_{pot} as a function of energy minimization steps.

The native structure of a GFP is taken from the protein data bank (PDB ID:1gfl) file and the interaction between the amino acids are grouped into native and non-native state. The protein data bank (PDB) format file contains information about atomic coordinates for various types of proteins, small molecules, ions and water. It provides a standard representation for macromolecular structure data which is derived from X-ray diffraction and nuclear magnetic resonance (NMR) studies.

We used `pdf2gmx` command in Gromacs to convert the pdb (GFP pdb:1gfl.pdb) file format into Gromacs coordinate format, and adds water around GFP. We choose a CHARMM-force fields, given by the following equation [10, 11]:

$$\begin{aligned}
 V = & \sum_{bonds} k_b(r - r_0)^2 + \sum_{angels} k_\theta(\theta - \theta_0)^2 \\
 & + \sum_{dihedrals} k_\phi[1 + \cos(n\phi - \delta)] \\
 & + \sum_{impropers} k_\omega(\omega - \omega_0)^2 + \sum_{Urey-Bradley} k_\mu(\mu - \mu_0)^2 \\
 & + \sum_{nonbonded} \epsilon \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^6 - \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^5 \right] + \frac{q_i q_j}{\epsilon r_{ij}} \quad (1)
 \end{aligned}$$

The first term accounts for the bond stretches, with force constant k_b and $r - r_0$ is the distance from equilibrium. The second term in the equation accounts for the bond angles where k_θ is the angle force constant. The third term is for the torsion angles where k_ϕ is the dihedral force constant. The fourth term accounts for the improper, that is out of plane bending, where k_ω is the force constant and $\omega - \omega_0$ is the out of plane angle.

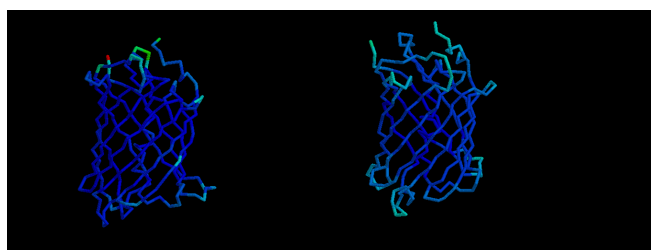
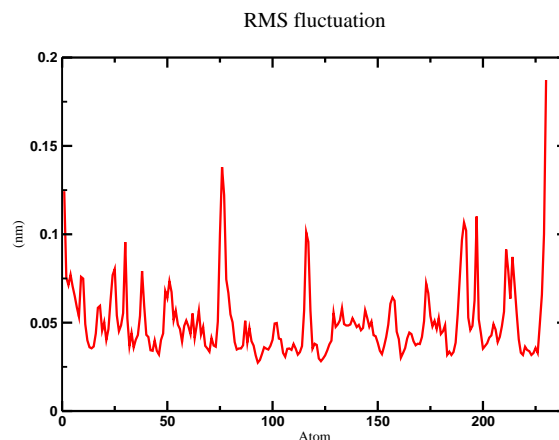


FIG. 4. Upper Panel: Root mean square value of fluctuation as a function of protein residue. Lower Panel: Backbone of Green fluorescent protein with color that vary according to temperature. The right panel with in lower panel show the comparison with the input GFP (PDB:1gfl.pdb). High fluctuation are represented in red color while the blue color represent the low fluctuation .

The Urey-Bradley component (cross-term accounting for angle bending using 1,3 nonbonded interactions) comprises the fifth term, where k_U is the respective force constant and U is the distance between the 1,3 atoms in the harmonic potential. Nonbonded interactions between pairs of atoms (i, j) are represented by standard

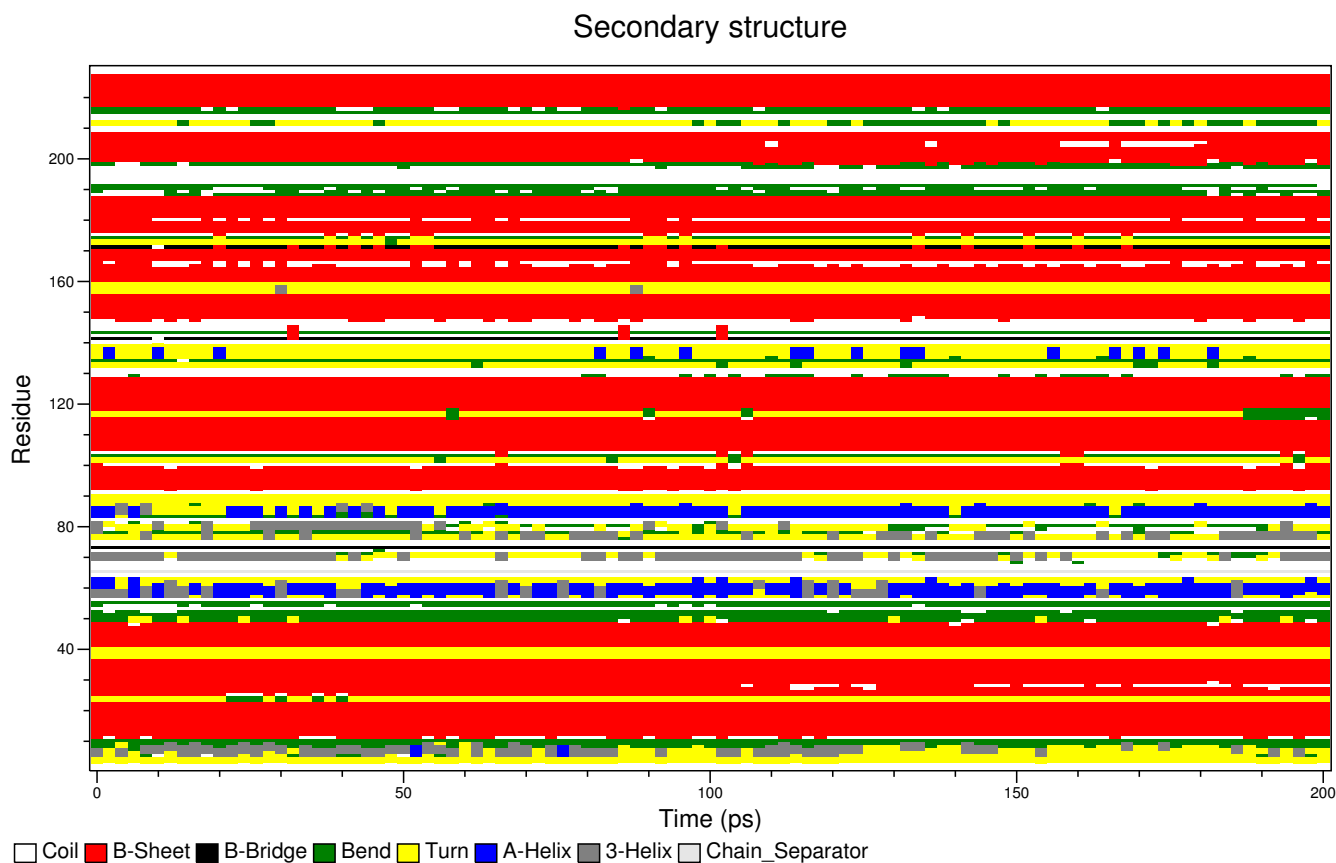


FIG. 5. Secondary structure of GFP per residue versus time using dssp-software

12-6 Lennard-Jones potential and the electrostatic energy with a Coulombic potential. We used periodic boundary conditions for the protein to be simulated to keep track of the motion of all particles. We define a rhombic dodecahedron box of size $0.5nm$. To mimic the physiological environment for a protein, we add water molecule around the protein. Fig. (1) illustrate the water molecules around the protein in a box having size $0.5nm$.

We used md-integrator, a leap-frog algorithm for integrating Newton's equations of motion. Gromacs also provide a leap-frog stochastic dynamics integrator (sd), and Euler integrator for Langevin dynamics (bd). The leap-frog algorithm implemented in Gromacs uses position \mathbf{r} and velocity vector \mathbf{v} at time t and $t - \Delta t/2$, respectively [12–14]. Mathematically:

$$\mathbf{v}(t + \Delta t/2) = \mathbf{v}(t - \Delta t/2) + \frac{\mathbf{F}(t)}{m} \Delta t \quad (2)$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t - \Delta t/2) \Delta t \quad (3)$$

where $\mathbf{F}(t)$ represent the force as a function of time t . In our simulation, we used $nsteps=10^4$, maximum numbers of steps to integrate. The time steps for integration is $dt=0.002$. The cut-off distance and neighbor list are $rlist=1.0$ and $nstlist=10$, respectively.

To understand the stability of the constructed system used as starting structure for the molecular dynamic

simulations, the Protein-H root mean square deviation (RMSD) with respect to the Protein-H group were calculated. The RMSD in Gromcas can be calculated by using the formula,

$$RMSD(t_1, t_2) = \left[\frac{1}{M} \sum_{i=1}^N m_i \| \mathbf{r}_i(t_1) - \mathbf{r}_i(t_2) \|^2 \right]^{\frac{1}{2}} \quad (4)$$

where $\mathbf{r}_i(t)$ is the position of i th atom at time t , and $M = \sum m_i$ represent the total mass. Fig. (2), represents the root mean square deviation (RMSD) of GFP-water complex over the equilibration time of $200ps$. Initially the RMSD increase rapidly, but stabilizes around $0.14nm$. To see that the system (GFP-water) is at an energy minimum, we plot the average potential energy E_{pot} as a function of energy minimization steps. Fig. (3) illustrate stability and convergence of the potential energy.

We also calculate the root mean square fluctuation (RMSF) which is a measure of the deviation between the position of i th-particle and some reference position with over the time T average. The root mean square fluctuation (RMSF) as a function of per residue over the `c.alpha` group is shown in fig. (4:Upper Panel). The figure show the vibration around the equilibrium which depend on the local structure flexibility. The protein backbone is shown

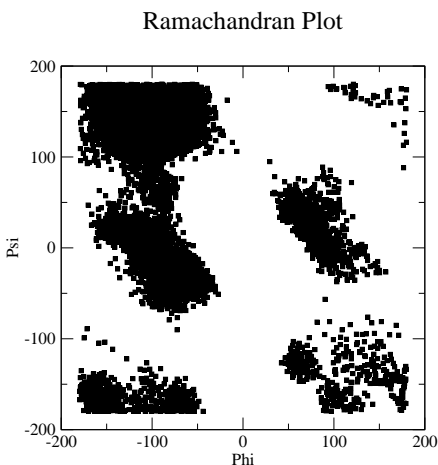


FIG. 6. Ramachandran plot of a GFP .

in fig. (4:Lower Panel) with the flexibility encoded in the

colour that vary with temperature. The red regions are at high fluctuation and the blue regions are relatively at low fluctuation.

To further check the stability of the system, we plot the secondary structure of GFP per residue versus time using dssp (Dictionary of Secondary Structure for Proteins)-software, as shown in Fig. (5). The β -sheet and α -helix are highlighted with red and blue color respectively. Fig. (6) shows the Ramachandran plot of GFP. The figure describe the projection of the structure on the two dihedral angles ϕ ($-180 : 180$) and Ψ ($-180 : 180$) of the protein backbone. In Gromacs, `g_rama` command is implemented to obtained the Ramachandran plot.

III. CONCLUSION

In summary, we discuss molecular dynamics simulation of GFP(Green Fluorescent Protein) with solvent water. We used steepest descent method to calculate the local energy minimum with 10^4 -steps. To determine whether the protein is close to the experimental structure during and after simulations, we calculate the root-mean-square deviation (RMSD) of the Protein-H atoms with respect to the X-ray structure. Secondary structure and Ramachandran plot of GFP protein is also presented.

-
- [1] B. Hess B, C. Kutzner C, D. van der Spoel, E. Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory. Comput.*, 4(3):435-47, 2008.
- [2] S. Pronk, S. Pll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. Shirts, J. Smith, P. Kasson, D. van der Spoel, B. Hess, E. Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845-54, 2013.
- [3] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E. and Berendsen, H. J. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26(16): 1701-1718, 2005.
- [4] Fan Yang, Larry G. Moss and George N. Phillips, Jr. The molecular structure of green fluorescent protein. *Nature Biotechnology*, 14, 1246:1251, 1996.
- [5] R. Heim, Prasher, D. C. Tsien, R. Y. Wavelength mutations and post-translational autoxidation of green fluorescent protein. *Proc. Natl Acad. Sci. USA*, 91, 12501:12504, 1994.
- [6] A. B. Cubitt, Heim, R., Adams, S. R., Boyd, A. E., Gross, L. A. Tsien, R. Y. Understanding, improving and using green fluorescent proteins. *Trends Biochem. Sci.*, 20, 448:455, 1995.
- [7] M. Ormo, Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y. Remington, S. J. Crystal structure of the Aequorea victoria green fluorescent protein. *Science*, 273, 1392:1395, 2006.
- [8] Renato J. Aguilera, Jessica Montoya, Todd P. Primm, Armando Varela-Ramirez. Green Fluorescent Protein as a Biosensor for Toxic Compounds. *Reviews in Fluorescence*, 463:476, 2006.
- [9] Shimomura, Martin Chalfie and Roger Y. Tsien. for the discovery and development of the green fluorescent protein, GFP. *The Nobel Prize in Chemistry*, 2008. https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2008/advanced-chemistryprize2008.pdf
- [10] MacKerell et. al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586-3616, 1998.
- [11] Jr. A.D. MacKerell and N. Banavali. All-atom empirical force field for nucleic acids: 1) parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comp. Chem.*, 21:86-104, 2000.
- [12] R. W. Hockney, S. P. Goel, and J. W. Eastwood. Quite high-resolution computer models of plasma. *J. Comp. Phys.*, 14:148, 1974.
- [13] W. F. van Gunsteren and H. J. Berendsen, Algorithms for macromolecular dynamics and constraint dynamics, *Mol. Phys.* 34: 1311-1327 (1977).
- [14] L. Verlet, Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules, *Phys. Rev.* 159: 98-103 (1967).