University of Crete

Department of Mathematics and Applied Mathematics

Master Thesis

Molecular Dynamics simulations of proteins

Author:
Maria Arnittali

Commitee:
Associate Prof. Evangelos Harmandaris
Prof. Michael Kokkinidis
Dr. Anastasia Rissanou

October 24, 2018

Contents

Figures

Tables

Acknowledgements

First and foremost I offer my sincerest gratitude to my supervisors Prof. Evangelos Harmandaris as well as Dr. Anastassia Rissanou, for their support, patience as well as their guidance in my research. The door to their offices were always open whenever I had a problem with my research or writing. They consistently allowed this thesis to be my own work, but steered me in the right direction whenever they thought I needed it.

I would also like to acknowledge Prof. Michael Kokkinidis of the Department of Biology at University of Crete as the second reader of this thesis, and I am gratefully indebted to his for his very valuable comments and suggestions on this thesis.

Special thanks also to Dr. Maria Ambrazi of the Department of Biology at the University of Crete, for her valuable help during my thesis.

Finally, I must express my very profound gratitude to my parents Chrysoula and Thanasis and dedicate it to them, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Author

Maria Arnittali

Abstract

Proteins are large biomolecules, which perform a great variety of functions in the living systems. The structure of biological macromolecules, like proteins can be provided by experiments. However it is very difficult to get insight into the atomic structures and the way that their various properties change in time experimentally.

Molecular Dynamics simulations have become a powerful technique in studying biological macromolecules, like proteins. In the current work, we investigate two proteins in the native state, the repressor of primer (Rop), and its loopless mutation (RM6), in aqueous solution. We report data concerning hydrogen bonding, structural properties and stereochemical quality of the systems. Our results provide information both in molecular and in atomic level.

The native state is stable in Rop protein as well as in RM6 protein. In terms of Rop protein, Ala is the only amino acid among the residues of the loop that hydrogen bonding with both helices simultaneously. There are slightly different between Rop protein and RM6 protein. Comparisons between Rop protein and its mutation RM6 have been also performed.

The analysis of our simulations show that native state of proteins is stable under specific conditions. Additionally, hydorgens bonds are responsible for the stability of the secondary structure in both proteins. Electrostatic interactions are very important for the stability of the native state. Moreover, we could demonstrate the structure of our model with the Ramachandran plot.

Chapter 1

Introduction

1.1 Proteins

Proteins are complex biological macromolecules, that not only attain a wide variety of functional roles but also they are the most abundant molecules in the living systems [11], [12], [20]. The building blocks of proteins are some simple monomeric subunits, called amino acids, which provide the key to the structure of the thousands of different proteins. Additionally, all proteins are constructed from the same set of 20 different amino acids, which are linked covalently to one another in characteristic linear sequences. The amino acids are linked to one another by peptide bonds forming a long chain of proteins.

To understand how the protein gets its final shape or conformation, we need to understand the four levels of protein structure:

- **Primary Structure** is the linear sequence of amino acids in a peptide or protein. By convention, the primary structure of a protein is reported starting from the amino-terminal (N) end to the carboxyl-terminal (C) end.

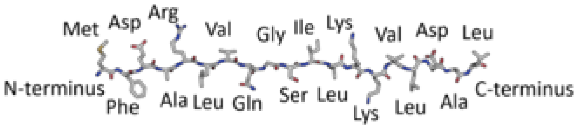


Figure 1.1: Primary Structure [60].

- **Secondary Structure** refers to three dimensional form of local conformation of proteins. Secondary structure focuses on particularly arrangements of amino acid residues giving rise to recurring structural patterns of the polypeptide backbone. A few types of secondary structure are particularly stable and occur widely in proteins. The most prominent are the α-helix and β-sheet conformations, though β turns occur as well.

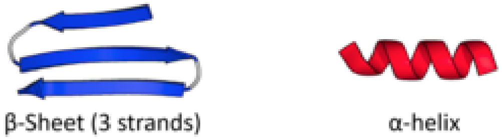


Figure 1.2: Secondary Structure [60].

- **Tertiary Structure** describes the three dimensional (3D) shape of protein. The tertiary structure will have a single polypeptide chain "backbone" with one or more protein secondary structures, the protein domains. Amino acid side chains may interact and bond in a number of ways. The interactions and bonds of side chains within a particular protein determine its tertiary structure. The protein tertiary structure is defined by its atomic coordinates. These coordinates may refer either to a protein domain or to the entire tertiary structure. A number of tertiary structures may fold into a quaternary structure.

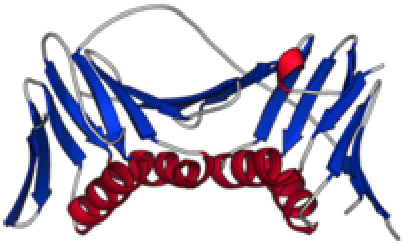


Figure 1.3: Tertiary Structure [60].

- **Quaternary Structure** is the number and arrangement of multiple folded protein subunits in a multi-subunit complex. It includes organisations from simple dimers to large homooligomers and complexes with defined or variable numbers of subunits. It can also refer to biomolecular complexes of proteins with nucleic acids and other cofactors.

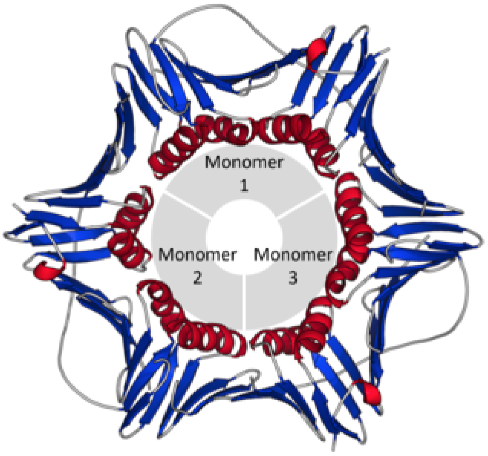


Figure 1.4: Quaternary Structure [60].

The spatial arrangement of atoms in a protein is called its **conformation**. The possible conformations of a protein include any structural state that can be acheived without breaking covalent bonds. Native proteins are the proteins in any of their functional folded conformations, which is called native state.

Peptide Bond

A peptide bond is a covalent bond formed between two amino acids. When amino acids combine to form proteins, an amine group and a carboxyl group form a covalent bond between them with subsequent elimination of a molecule of water [11], [12], [20]. The chemical reaction of amino acids to form a peptide bond is presented in Figure (1.5).

As we can see, two amino acids are shown in Figure (1.5), these are able to form a peptide bond, when two Hydrogens and an Oxygen (the elements of water) are removed from them. From the first amino acid a carboxyl group loses a hydroxyl group, whereas the amine group of the other amino acid loses a hydrogen. Then, the Nitrogen (N) substitutes amide linkages. So, after the formation of the peptide bond, we see a dipeptide and a water molecule.

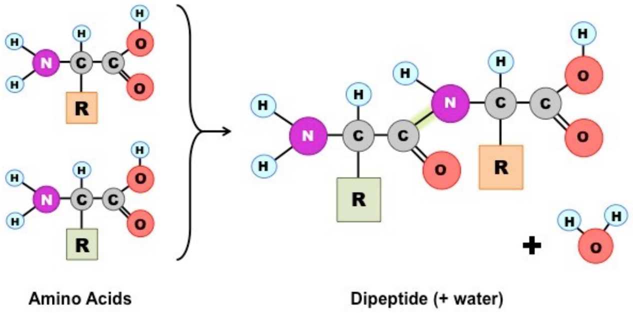


Figure 1.5: Formation of a peptide bond between two amino acids [51].

In this way amino acids can be linked to form polypeptides, of almost any length. Take into account that there are 20 amino acids to work with, an unlimited number of chains can be imagined in which either the sequence or number of amino acids is different; hence we can explain the chemical basis for the wide diversity of proteins.

By having different compositions or sequences, proteins can be formed to have totally different structures and functions.

Amino acid structure and its classification

There are 20 amino acids [11], [12], [20]. All 20 of amino acids are α-amino acids. They have a **the carboxyl group** (COOH) and **an amino group** (NH_2) bonded to the same carbon atom (α carbon). They differ from each other in their side chains (**R groups**), which vary in structure, size, and electric charge, and which influence the solubility of the amino acids in water.

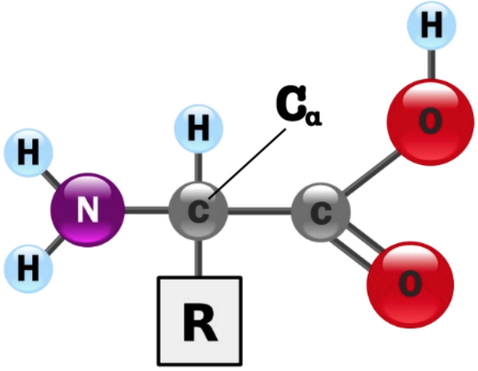


Figure 1.6: General structure of an amino acid.

In Figure (1.6), the atom of Carbon (C), which is the center atom, and is bonded with N, R, H and C, we are going to call it as C_α from now on.

In Table (1.1), we present the 20 amino acids, with their three and one letter code name. In Figure (1.7), we show the chemical type of the twenty amino acids.

Full name of amino acid name	**3-letter name**	**1-letter name**
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamate	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Iso	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Table 1.1: The twenty amino acids.

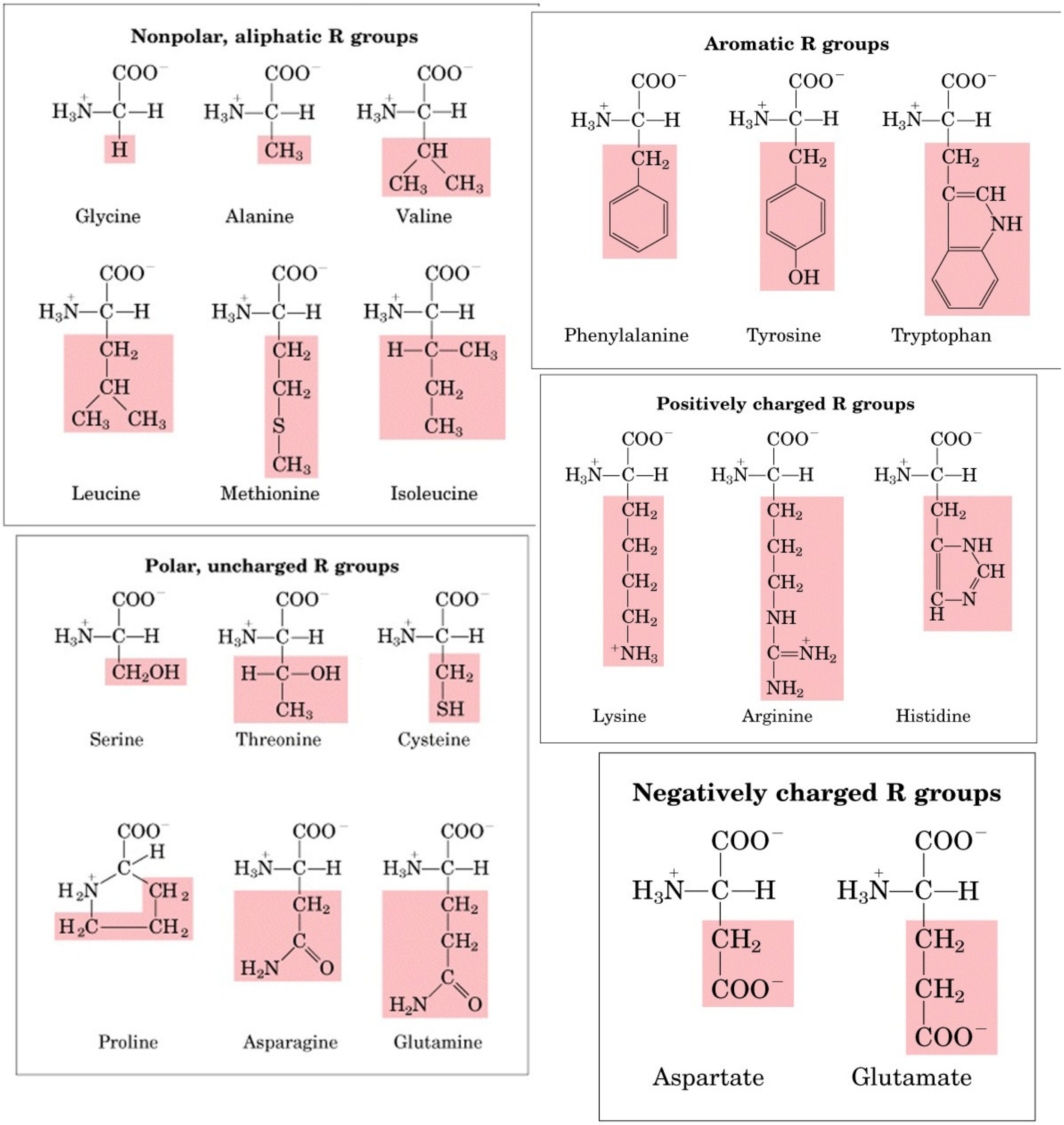


Figure 1.7: The chemical type of the 20 amino acids [52].

1.2 The Protein Folding Problem

Proteins must fold up into the native state before they can actively function in the living cell [7]. According to the famous Levinthal paradox, if a peptide bond can adopt only two conformations, and let us consider that we have a relatively short protein contained a hundred residues. Then, all the possible forms of it are around $2^{100} \approx 10^{30}$. If the protein should explore conformational space randomly with nearly one conformer per ps, which is possibly the speed limit, then the time to find the native state should take more than the age of the universe. So now, it is clear that there must be some other principles at work, to make proteins fold in fractions of time. The path a protein follows to its native state is known as the protein folding problem. We can summarize it in two basic subjects.

1. Can we predict the native structure of a protein from its amino acid sequence?
2. How does the protein find its path toward the native state?

1.3 Rop Protein

The Repressor of Primer (Rop) protein is a small homodimeric (Figure 1.8a) RNA binding protein, which is involved in the regulation of copy number of ColE1 plasmids [1], [2], [3], [4], [6]. Both methods X-Ray crystallography and NMR are used to study the structure of Rop Protein. Each monomer of Rop protein consists of 63 amino acids (Figure 1.9), that form two a-helices (in Figure 1.8b the yellow) connected by a loop of four amino acids (L29, D30, A31, D32), see in Figure (1.8b) the green region. Moreover, the molecular weight of its monomer (a chain) is about 7.500 Da. The two monomers (the two chains) related to a two-fold symmetry, which means that the protein looks the same if we turn it 180 around an axis.

Rop serves as a paradigm of a canonical 4-a-helical bundle. The apparent structural simplicity of its folding motif led various groups to believe that Rop is a model system that we can use for the investigation of the sequence-structure relationships in the folding and dynamics of four a-helix bundles. The four helices packed in an antiparallel style of hydrophobic and hydrophilic amino acids, as well as follow a specific pattern, known as heptad pattern, which repeated every seven residues and it has the type (a,b,c,d,e,f,g). The positions a and d are generally hydrophobic and their side chains packed in the central part of the structure according to the ”knobs in holes” model forming the hydrophobic core. There is a disruption of the heptad periodicity only once and leads to the formation of the loop. The role of Ala31 is crucial in the establishment of the loop region because among the amino acids it is the only one that forms Hydrogen Bond to both helices simultaneously.

In Figure (1.8), we present in more details the Rop protein by a cartoon representation. In Figure (1.8a), we show the two chains of Rop with different color. Then, we take one of the chains (Figure 1.8b) and we colored it with different colors, in order to present the different parts of the chain. With yellow is the two α-helices of the chain, with red is the tail and finally, with green is the loop region.

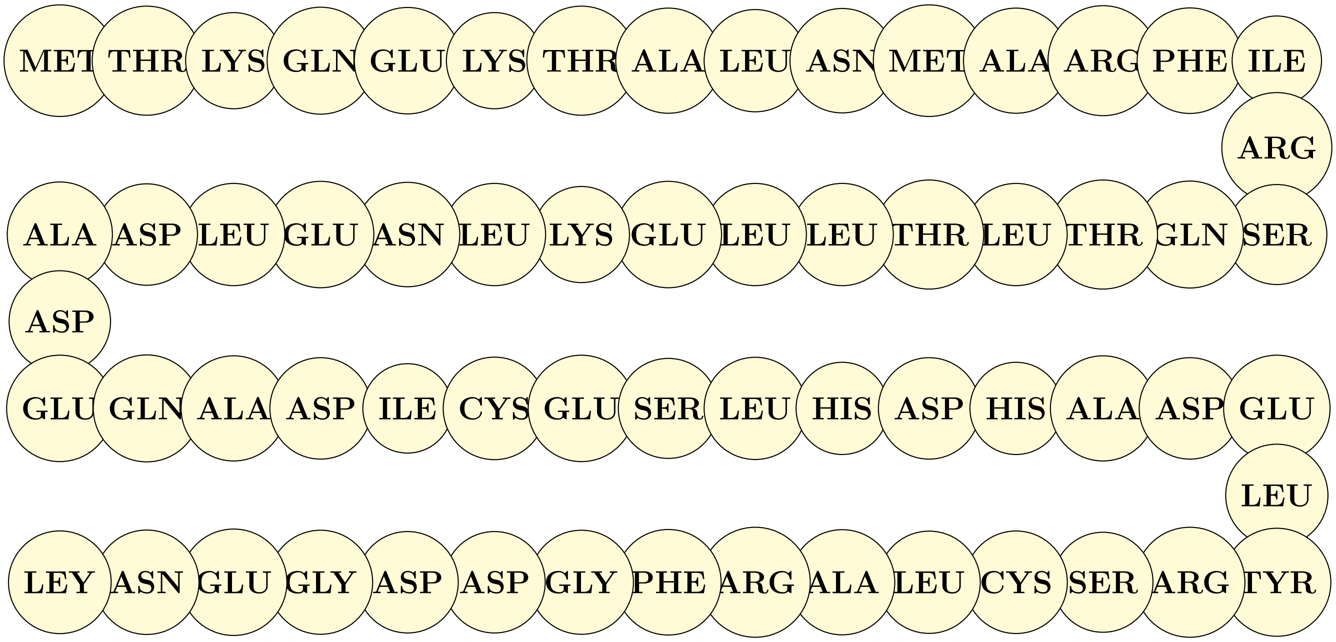
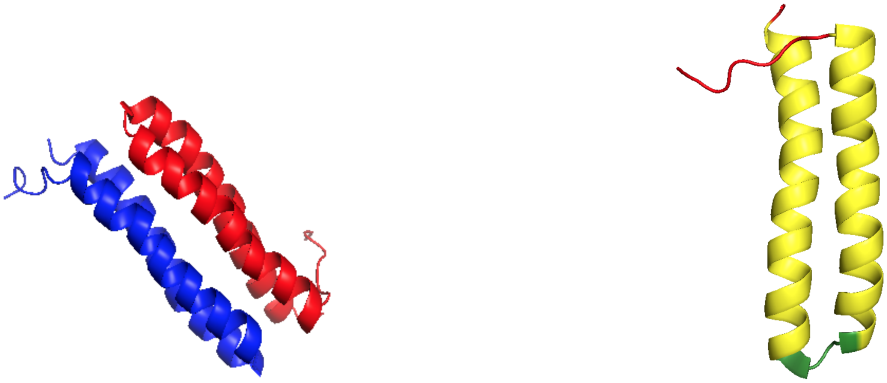


Figure 1.9: The sequence of amino acids of the Rop monomer.



(a) A cartoon representation of Rop protein.

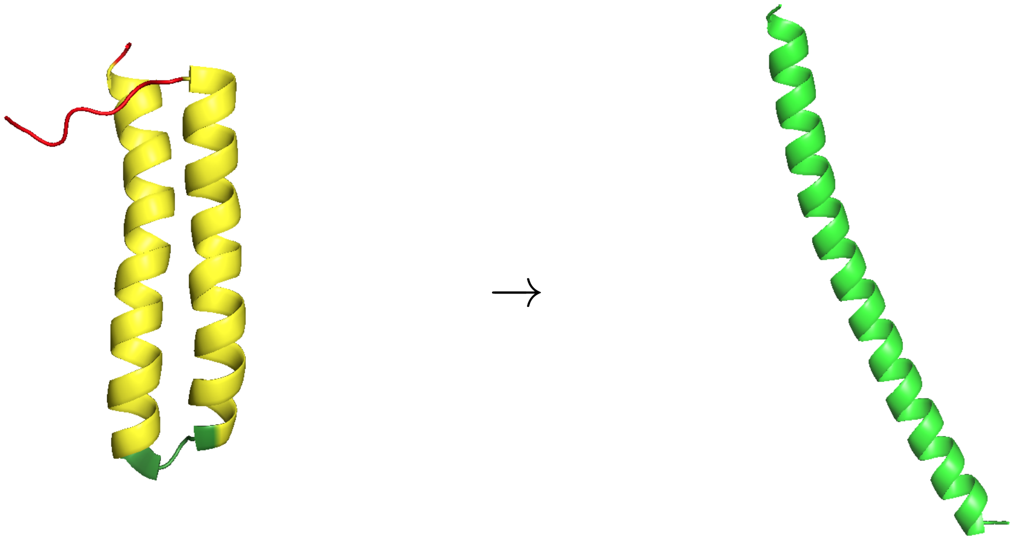
(b) A cartoon representation of a single chain of Rop protein, which we distinguish with different color the different parts of it.

Figure 1.8: Rop protein.

1.4 Loopless Rop (RM6) Protein

The loopless mutation of Rop, known as RM6 protein, created in order to study the role of heptad motif and the hydrophobic core of Rop protein [2], [5], [4]. As we have already mentioned, the heptad motif is interrupted in the loop, where there are only five residues instead of seven. So, RM6 is a loop mutation, in which an uninterrupted pattern of heptad repeats established through a five-residue deletion in the loop (from 30ASP up to and including 34GLN). The monomer of this loopless mutant converted into a single helix. The whole RM6 molecule is a homotetrameric, all anti-parallel, left-handed four-α-helix bundle, totally reorganized relative to the dimeric Rop. Thereby, it is becoming a hyper-thermostable protein. Each monomer of the tetramer corresponds to one Rop polypeptide chain and forms a continues α-helix in contrast with Rop where we have the form α-helix-turn-α-helix hairpin.

In the Figure (1.10), we present a chain of Rop protein, from which we remove 5 amino acids, from the loop region and then, we present a chain of RM6. We observe that the chain of RM6 looks like a linear line.



(a) A chain of Rop protein.

(b) A chain of RM6 protein.

Figure 1.10: The transformation of the chain of Rop with the remove of the 5 residues from the loop region to the linear chain of RM6.

In Figure (1.11), we represent with cartoon scheme the RM6 tetramer. We show each chain of RM6 with different colour. Moreover, in Figure (1.12), we present the amino acid sequence of RM6 monomer.

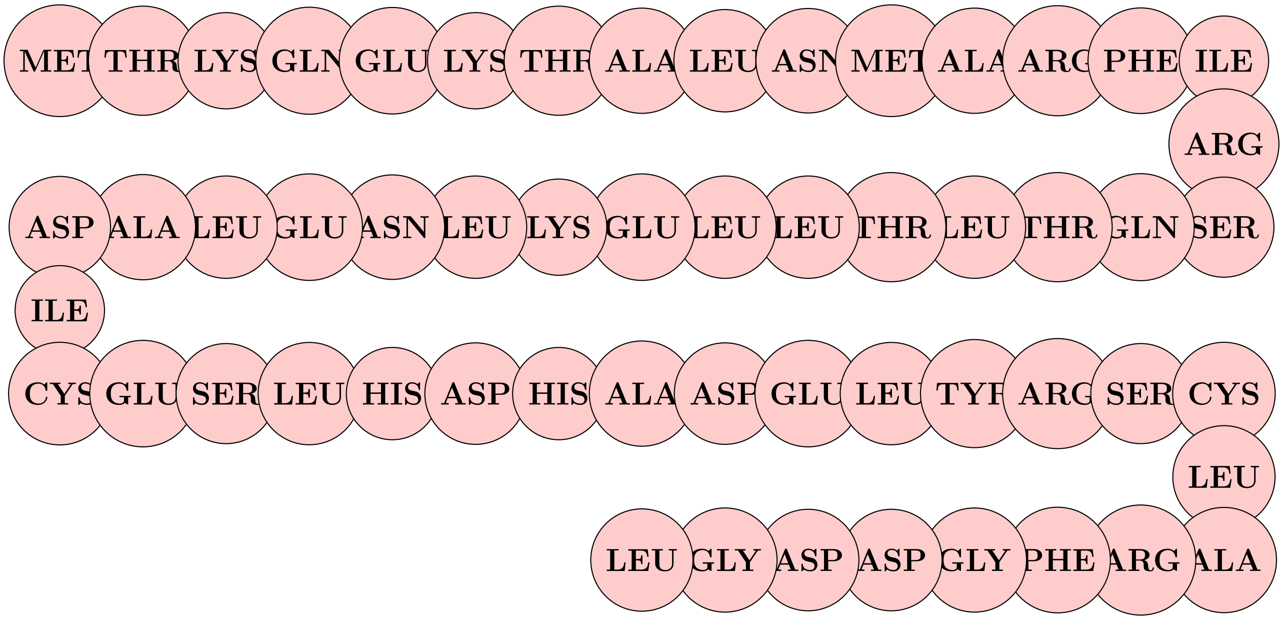


Figure 1.12: The sequence of amino acids of an RM6 monomer.

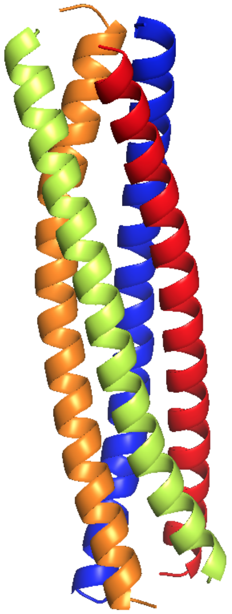


Figure 1.11: A cartoon representation of RM6 protein.

1.5 Goals of this thesis

The structure of biological macromolecules like proteins can be provided by experiments. However it is very difficult to get insight into the atomic structures and the way that their various properties change in time experimentally. Computer simulations constitute a powerful technique, which is based on fundamental physics providing information on how proteins and different biomolecules move, vibrate, interact and generally operate [8]. We can analyze the simulated structure in atomic detail, to investigate what is critical in determining biological activity. Simulations have been assumed as the key in developing the theoretical context which is now the core of biomolecular science. This is the understanding of the way that biological molecules move, their structure and conformations. Biological macromolecules like proteins are very complex and challenging in modeling.

In biology, one of the most important but still unanswered problems, is known as the protein folding problem (Section:1.2), we can summarize it in two basic questions: 1) can we predict the structure and function of proteins only from its amino acid sequence alone? 2) How does the protein find its way toward the native state? As we have already mentioned in Section (1.1), the function of protein is determined by its three-dimensional (3D) native state. Biological function of proteins is based on molecular interactions of a consequence of their structure. So, the study of the proteins structure is a key point in the understanding of biology. The knowledge that governs the protein folding will allow us to design proteins with desirable structural properties. The second question related with the first one, because the protein's stability and folding kinetics are crucial components of the structure and function prediction.

However althought nowadays, atomistic simulations can reach the time scale of microseconds (μs) for small proteins in solution, in the case of larger systems, simulations are limited to tens up to hundreds of nanoseconds. This limitation in time can be overcome through proper coarse grained models coming from hierarchical multi-scale modelling techniques.

In the current work, we explore, through atomistic Molecular Dynamics (MD) simulations (Chapter:2), the native state of two proteins. The first one is the repressor of primer (Rop) protein (Section: 1.3), and the second one is its loopless mutation (RM6) protein (Section: 1.4), in aqueous solution. MD generates a trajectory by numerical integration of classical equations of motion (Section: 2.3.1). The produced trajectory, contains all dynamical information, that we need for our analysis, in addition, we can characterize the time evolution of the molecular structure, their fluctuations and interactions. Dynamics plays also a key role in their functionality. Proteins undergo significant conformational changes while performing their function and this could lead to their rearrangement in the 3D space. The main purpose of our study is to investigate the stability of the native state of these two proteins, and to understand which are the driving forces for this stability. A clear advantage of performing MD simulations over real experiments is the fact that we can obtain information in microscopic level and express them in macroscopic properties. Concerning the stability of the native state, we measure the root mean square deviation (RMSD), with respect to a reference (native state) structure . Then, we measure the radius of gyration (Rg), which gives us information about the compactness of the 3D conformation of our protein molecules. We perform a more detailed analysis in atomic level where the creation of hydrogen bonds among proteins is detected. Additionally, the stereochemical quality of our two models, is very significant, therefore we calculate the well known Ramachandran plot.

In the current work we study the Rop protein and its RM6 mutation, where five amino acids are removed from the loop region. A totally different conformation compared to Rop is observed in the native state of RM6. According to literature ([2]), even a single mutation of an amino acid can drive to a complete different native state.

Molecular simulations provide the oportunity to explore the driving forces for the native state conformation of various mutations as well as their structural and dynamical properties. Therefore they provide information for the creation of mutation with desired functions.

Chapter 2

Molecular Dynamics

2.1 Computer Simulations

Computer simulations are executed on a computer and are the imitation of a real-world process or system over time [14], [16], [18]. If we want to perform a computer simulation, we should develop a good model for the desired system we want to investigate. Computer simulations have become a useful part of mathematical modeling of many natural systems in physics, chemistry, as well as biology, to gain insight into the operation of those systems. We perform Computer Simulations in the belief of understanding the properties of assemblies of molecules, concerning their structure, as well as, the microscopic interactions between them. We can learn through simulations something that cannot found out in other ways, like real experiments. So, we can say that they serve as an addition to the conventional operations. The two main families of simulation technique are molecular dynamics (MD) and Monte Carlo (MC).

In Figure 2.1, we can see that computer simulation act as a link between microscopic length and time scale and the macroscopic world of the laboratory. Therefore, if someone wants to perform a real experiment in a laboratory is difficult or even impossible, for the reason that the materials we want to use are too expensive or too dangerous for use. So, computer simulations solve these kinds of problems, and we can study properties of systems, which we could not through real experiments. If we want to compare directly experimental measurements made on specific molecules (like proteins) to the results of simulation work, it is essential to have a good model of molecular interactions.

In this work we focus on Molecular Dynamic Simulations, so we are going to discuss about it in the next sections.

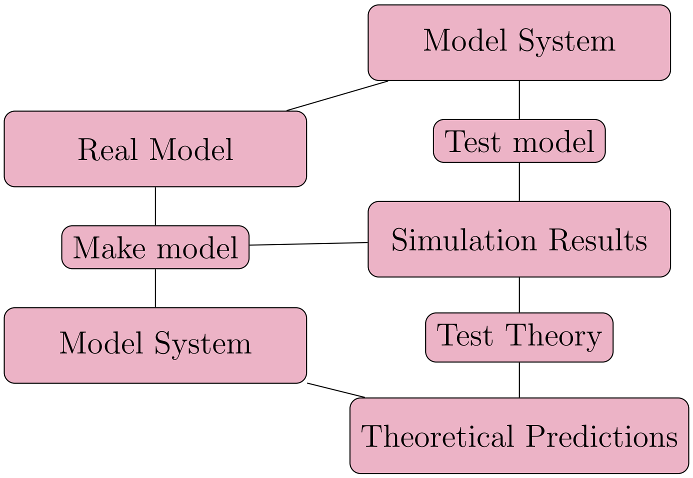


Figure 2.1: Computer simulations act as a bridge between theory and experiment

2.2 Molecular Dynamics Simulations

Let us consider that we have a many-body system consisting of N particles that we want to study [14], [16], [18]. Molecular Dynamics (MD) Simulation is a powerful technique of computing the equlibrium and transport properties of such systems. The solution of the classical equations of motion, which are integrated numerically gives us information about the positions and velocities of atoms in the system, is at the heart of this technique. Before we perform a real experiment, we follow specific steps, the same we do in a MD simulation. First, we should prepare our sample. We select our model we want to study, consisting of N particles and we solve Newton's equations numerically for this system until we equilibrate it, which means that the properties of our system no longer change with time. When we reach equilibration, we perform the actual measurement. If we want to measure a quantity, firstly, we should express it as a function of the positions and momenta of the particles.

After equilibration we perform the actual measurement. A simple flow diagram of a standard MD algorithm is shown in the Figure 2.2 and includes the following steps:

STEP 1 First, a model configuration representing a molecular-level snapshot of the corresponding physical system is chosen or constructed and is initialized (initial positions, velocitites of each particle within the system).

STEP 2 Then, we compute the total force acting on each particle within our system.

STEP 3 We integrate the equations of motion by choosing an appropriate method.

STEP 4 After the system has reached equilibration, we performed the actual measurements (positions, velocities, energie, etc, are stored) after the system has reached equilibration, periodically every N_k steps.

STEP 5 After we complete the N steps of our loop, the averages of the measured quantities that we want to study are calculated and printed.

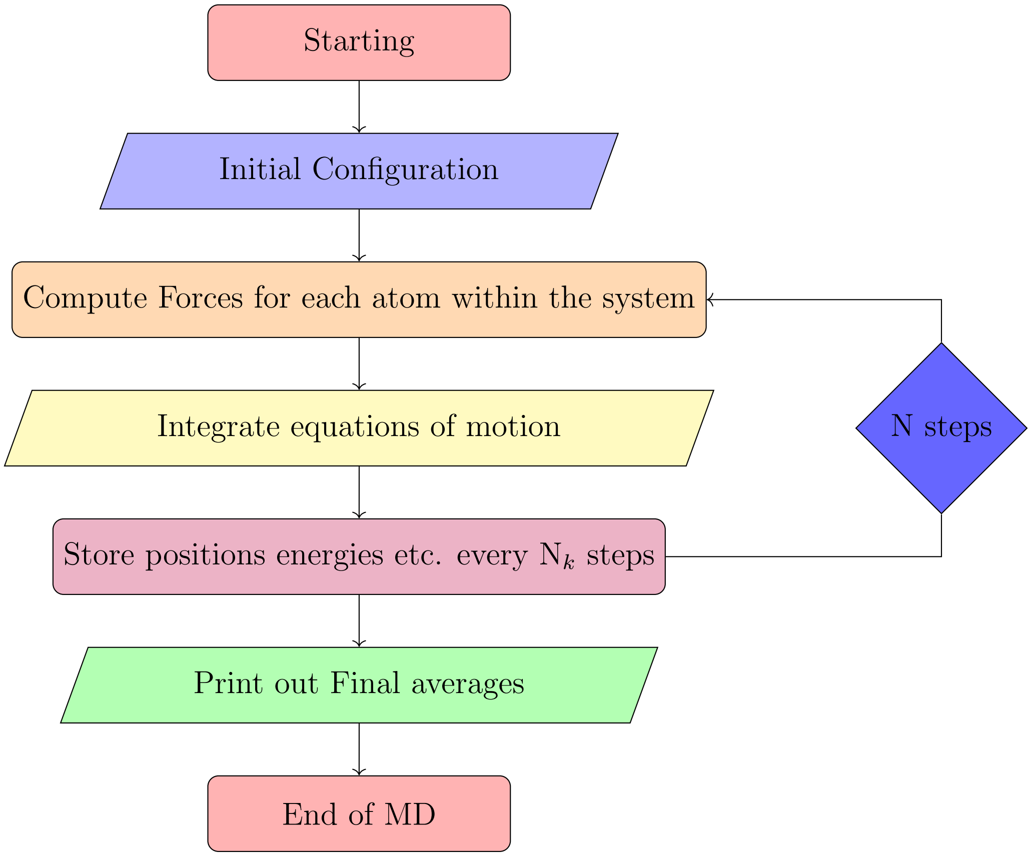

Starting
Initial Configuration
Compute Forces for each atom within the system
Integrate equations of motion
N steps
Store positions energies etc. every N_k steps
Print out Final averages
End of MD

Figure 2.2: A simple flow diagram of standard MD algorithm

2.3 Force Calculation

In Molecular Dynamics simulations, the calculation of the force acting on every particle is the most time-consuming part [14].

2.3.1 Classical equations of motion

In Molecular Dynamics simulation, the most time-consuming part is the calculation of the forces [14], [16], [18], [15], [13]. The solution of the classical equations of motion, which are integrated numerically and gives information for the positions and velocities of atoms in the system, is the heart of MD simulation.

Let us assume a system of N interacting molecules interacting via a potential energy function

V. If we consider a system of atoms with Cartesian coordinates then, we can write the equations of motion in the following form:

$$m_i \ddot{\mathbf{r}}_i = F_i \tag{2.1}$$

$$F_i = -\frac{\partial V(\mathbf{r})}{\partial \mathbf{r}_i}$$

where:

- m_i is the mass of atom i
- F_i is the force acting on atom i
- $\mathbf{r}_i$ is the coordinates of atom i
- $\mathbf{r}_i = r_i^x, r_i^y r_i^z$
- $V(\mathbf{r})$ is the potential energy, where $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ represents the complete set of 3N atomic coordinates.
- $i = 1, \ldots, N$

In the equations dots denote time derivatives.

We can write the above equations (2.1) in various ways. While the equations are valid for a broader formulation of classical mechanics enables us to develop equations of motion for any coordinate system. For that reason, we should denote the generalized coordinates as $\mathbf{q}$ and $\dot{\mathbf{q}}$. The latter coordinates describe the molecular configuration, as well as their time derivatives, respectively. The most popular formalisms are the Lagrangian and the Hamiltonian.

In the Lagrangian formalism, the trajectory $\mathbf{q}(t)(= q_1(t), q_2(t), \ldots, q_k(t), \ldots)$ satisfies the following set of differential equations:

$$\frac{d}{dt}(\frac{\partial L}{\partial \dot{q}_k}) - \frac{\partial L}{\partial q_k} = 0$$

where $L = L(\mathbf{q}, \dot{\mathbf{q}}, t)$ is the Lagrangian funtion of the system. The Lagrangian is defined in terms of the kinetic K and the potential energy V as:

$$L = L(\mathbf{q}, \dot{\mathbf{q}}, t) \equiv K(\dot{\mathbf{q}}) - V(\mathbf{q}).$$

As we can see, the kinetic energy (K) is expressed as a function of the derivative of $\mathbf{q}$ and the potential energy (V) as a function of the $\mathbf{q}$ coordinates themselves. The generalized momenta p_k conjugate to the generalized coordinates q_k are defined as

$$p_k = \frac{\partial L}{\partial \dot{q}_k}.$$

Alternatively, we can use the Hamiltonian formalism, which is defined in terms of the generalized coordinates and momenta. These obey Hamilton's equations of motion:

$$\dot{q}_k = \frac{\partial H}{\partial p_k}$$

$$\dot{p}_k = -\frac{\partial H}{\partial q_k}$$

where H is the Hamiltonian of the system, which is strictly defined by the following equation:

$$H(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{3N} \dot{q}_k p_k - L(\mathbf{q}, \dot{\mathbf{q}}) \tag{2.2}$$

Now, if the potential energy V is independent of velocities and time, then H becomes equal to the total energy of the system: $H(\mathbf{p}, \mathbf{q}) = K(\mathbf{p}) + V(\mathbf{q})$. For the Cartesian coordinates, Hamilton's equations of motion become:

$$\dot{\mathbf{r}}_i \equiv v_i = \frac{\mathbf{p}_i}{m_i}$$

$$\dot{\mathbf{p}}_i = -\nabla_{r_i} V \equiv -\frac{\partial V}{\partial r_i} = F_i$$

hence

$$m_i \ddot{r}_i \equiv m_i \dot{q}_i = F_i \tag{2.3}$$

where F_i is the force acting on atom i.

The classical equations of motion contain some impressive properties. The conservation law is the most important one. Now, if we assume that $K(\dot{\mathbf{q}})$ and $V(\mathbf{q})$ do not depend explicitly on time, then it is easy to see that $\dot{H} = \frac{dH}{dt}$ is zero, i.e the Hamiltonian is a constant of the motion. This conservation law is satisfied in actual calculations when the forces acting on the system are no explicitly time- or velocity-dependent.

A second important property is that Hamilton's equations of motion are time reversible. The time reversible means that, if we change the signs of all velocities, we will cause the molecules to retrace their trajectories backward. The trajectories generated by the computer should also contain this property.

2.3.2 Integration of equations of motion

Numerical Algorthms

The molecular dynamics simulations based on the numerical solution of the set of Newton's equations of motion [14], [16], [18], [15], [13]. In principle, the use of a proper algorithm for the numerical integration of these equations seems to be very important. There are a lot of different methods for solving ordinary differential equations. For that reason, a proper algorithm should satisfy the following criteria:

- algorithm should be executed fast, require low memory and be easy to implement.
- It should permit to use a large timestep.
- Additionally, the trajectory should be reproducible and time reversible.
- Finally, it has to converse the total energy.

Verlet Algorithm

The Verlet integrator was first introduced by Verlet in 1967. Verlet-like integrators are symplectic. A symplectic integrator preserve phase space, as well as time-reversibility. For distinctness, we will derive the Verlet equations according to a single coordinate, but we should keep in mind that the following relations could be generalized to vector Cartesian coordinates. Let us consider a Taylor expansion of the position coordinate in two distinct directions in time:

$$r(t+dt) = r(t) + dtv(t) + \frac{dt^2}{2}\ddot{r}(t) + \frac{dt^3}{6}\dddot{r}(t) + O(dt^4) \tag{2.4}$$

$$r(t-dt) = r(t) - dtv(t) + \frac{dt^2}{2}\ddot{r}(t) - \frac{dt^3}{6}\dddot{r}(t) + O(dt^4) \tag{2.5}$$

Summing these two equations (2.4) and (2.5), we obtain:

$$r(t+dt) + r(t-dt) = 2r(t) + dt^2\ddot{r}(t) + O(dt^4)$$

Now, we move the term $r(t+dt)$ of the above equation to the right hand side:

$$r(t+dt) = 2r(t) - r(t-dt) + dt^2\ddot{r}(t) + O(dt^4) \tag{2.6}$$

Here, we calculate the term $\ddot{r}(t) = \frac{f(t)}{m}$ from the forces at the current positions, by the use of the force field.

The accuracy of the equation (2.6) is of order dt^4. The equation (2.6) is the basis of the Verlet algorithm for MD.

It is notable that no velocities are necessary to compute the new positions. The knowledge of trajectory help us to derive the velocity by using:

$$r(t+dt) - r(t-dt) = 2v(t)dt + O(dt^3)$$

or

$$v(t) = \frac{r(t+dt) - r(t-dt)}{2dt} + O(dt^2). \tag{2.7}$$

The expression of velocity has an accuracy of order dt^2.

A disadvantage of the Verlet algorithm is that we have to store in memory two sets of positions, $r(t)$ and $r(t-dt)$.

In the next subsections, we discuss two modifications of Verlet algorithm, the first one is the velocity Verlet algorithm and the second is the leapfrog algorithm, which is the one we use in our simulations.

Velocity Verlet Algorithm

The Velocity Verlet algorithm introduced in 1982, which is more commonly in use. It is a reformulation of Verlet scheme. More specifically the derivation way is pretty similar. As we discussed in above subsection, by the Taylor expansion we have:

$$r(t+dt) = r(t) + v(t)dt + \frac{f(t)}{2m}dt^2 \tag{2.8}$$

$$v(t+dt) = v(t) + \frac{f(t+dt)+f(t)}{2m}dt \tag{2.9}$$

Let us show how the Velocity Verlet algorithm and the Verlet scheme are equivalent, we note that the equation

$$r(t+2dt) = r(t+dt) + v(t+dt)dt + \frac{t+dt}{2m}dt^2$$

and the equation (2.8) can be written as

$$r(t) = r(t+dt) - v(t)dt - \frac{f(t)}{2m}dt^2$$

By summing we obtain

$$r(t+2dt) + r(t) = 2r(t+dt) + [v(t+dt) - v(t)]dt + \frac{f(t+dt)-f(t)}{2m}dt^2$$

Then, by substitution of equation (2.9) assigns

$$r(t+2dt) + r(t) = 2r(t+dt) + \frac{t+dt}{m}dt^2,$$

which, actually, is the coordinate version of Verlet algorithm.

A general strategy of this algorithm is the following:

Step 1 Given an initial setup of the velocities ($v(t)$) and positions ($r(t)$) of our system, we first compute the forces on each atom using the force field.

Step 2 Compute the new position:

$$r(t+dt) = r(t) + v(t)dt + \frac{f(t)}{2m}dt^2$$

Step 3 Do a computation of the intermediate velocity:

$$v(t+dt/2) = v(t) + \frac{f(t)}{2m}dt$$

Step 4 Since the new positions $r(t+dt)$ are known, calculate the new forces $f(t+dt)$

Step 5 Finish with the calculation of the new velocity:

$$v(t+dt) = v(t+dt/2) + \frac{f(t+dt)}{2m}dt$$

Step 6 Return to step 1.

Table 2.1: A strategy of Velocity Verlet algorithm.

Leap Frog Algorithm

There are several algorithms equivalent to the Verlet algotihm. The simplest among them is the leapfrog algorithm, which evaluates the velocities at half-integer time steps ($dt/2$) and uses these velocities to compute the new positions of the particles. Before, we derive the Leap Frog scheme from the Verlet algorithm, we should define the velocities at half-integer time steps:

$$v(t-dt/2) \equiv \frac{r(t) - r(t-dt)}{dt}$$

as well as

$$v(t+dt/2) \equiv \frac{r(t+dt) - r(t)}{dt}$$

From the above equation we obtain an expression for the new positions, which bases on the old positions and velovities:

$$r(t+dt) = r(t) + dtv(t+dt/2) \tag{2.10}$$

According to the Verlet scheme, we get the following expression for the update of the velocities:

$$v(t+dt/2) = v(t-dt/2) + dt\ddot{r}(t) \tag{2.11}$$

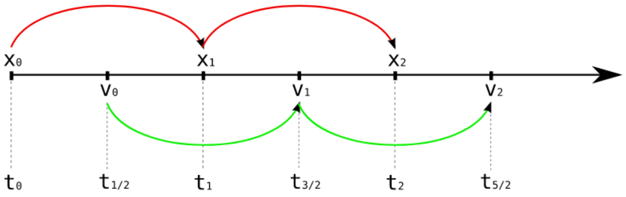


Figure 2.3: An illustration of Leapfrog method.

There is a disadvantage of Leap Frog algorithm. As we notice the velocities and positions are not defined at the same time. As a result, kinetic and potential energy are not defined at the same time, and hence we are not able to compute the total energy at any one point of time.

2.3.3 Statistical Ensembles

An ensemble is a collection of all possible systems which have different microscopic states but have an identical macroscopic or thermodynamic state [14], [16], [18], [15], [13]. A thermodynamic state is usually defined by a small set of parameters, like the number of particles N, the temperature T, and the pressure P. In the previous section, we described methods, which address the solution of Newton's equation of motion in the microcanonical ensemble (NVE). There is usually the need to perform MD simulations under specified conditions such as temperature and/or pressure. Let us consider that we want to investigate other ensembles, in practice there are three ensembles in common use:

- The microcanonical ensemble (NVE); in this ensemble the number of molecules N, the volume V, and the energy E are kept constant,
- The canonical ensemble (NVT); here the number of molecules N, the volume V and the temperature T are maintained constant,
- The isothermal-isobaric ensemble (NPT); in this ensemble the number of molecules N, the pressure P, and the temperature are kept constant.

In the literature, there are a great variety of methodologies for performing MD simulations under NVT or NPT ensemble. Most of these constitute a reformulation of the classical equation of motion, in order to include the constarints of constant T and/or P. Before proceeding to overview the details of the extended system method, it is better to separate two issues: pressure and temperature control.

2.3.4 Thermostats and Barostat

The temperature (T) is an important parameter because many experimental observables that can be compared with the information obtained from MD simulations depends on it . In order to control properly the temperature of the system we use the thermostats. The most widely used class of thermostat algorithms is based on rescaling of atomic velocities. In this section we are going to discuss about Berendsen thermostat and the V-rescale thermostat.

Berendsen Thermostat

Berendsen proposed a weak coupling method to an external heat bath at constant temperature, which is called Berendsen thermostat [9], [**?**], [10]. Berendsen thermostat is trying to correct the deviations of the actual temperature T form the prescribed one T_o. To achieve that the system is forced to obey the following equation:

$$\frac{dT}{dt} = \frac{1}{\tau_T}(T_o - T) \tag{2.12}$$

where T is the desired temperature, τ_T is time constant characterizing the frequency of the system coupling to temperature bath, and T_o is the instantaneous value of temperature, calculated from the momenta of the system. The velocities are scaled each time step by the following scaling factor:

$$\lambda = \left(1 + \frac{dt}{\tau_T}\left(\frac{T_o}{T} - 1\right)\right)^{\frac{1}{2}}$$

V-rescale Thermostat

V-rescale thermostat is essentially a Berendsen thermostat with an additional stochastic term that ensures a correct kinetic energy distribution by modifying it according to

$$dK = (K_o - K)\frac{dt}{\tau_T} + 2\sqrt{\frac{KK_o}{N_f}}\frac{dW}{\sqrt{\tau_T}} \tag{2.13}$$

where K is the kinetic energy, N_f the number of degrees of freedom, dW a Wiener process. Without the stochastic term $\boxed{2\sqrt{\frac{KK_o}{N_f}}\frac{dW}{\sqrt{\tau_T}}}$ the equation (2.13) reduces to that of the standard Berendsen thermostat [10].

Berendsen Barostat

Berendsen also proposed an algorithm for performing isobaric MD simulations, by coupling the system into a pressure bath [16]. If we want to achiece this, the system is forced to obey the following equations:

$$\frac{dP}{dt} = \frac{1}{\tau_P}(P_o - P) \tag{2.14}$$

whrere P is the desired pressure value, τ_P is the time constant characterizing the frequency of the system coupling pressure bath. P_o is the instantaneous value of pressure calculating from the configuration of the system. The positions are scaled at each time step by the following factor:

$$\mu = 1 - \beta_T \frac{dt}{\tau_T}(P_o - P) \tag{2.15}$$

where β_T is the isothermal compressibility of the system.

2.3.5 Periodic Boundary Conditions

Usually, the number of particles of the systems we want to study is at least N$\sim 10^{23}$ [14], [13]. In such systems, the effect of the walls on particles near them is negligible. When we perform a MD simulation, the number of particles involved are $10^3 - 10^5$, due to limited CPU power. For these relatively small systems, there are surface effects to the particles near the walls. To overcome the surface effect of our simulated systems, we apply periodic boundary conditions (PBCs), pretending that boundaries do not exist.

The simplest way to understand how (PBCs) work, we should think the game Pacman, see Figure (2.4). When Pacman reaches one side of the box, it re-enters on the opposite side. The same thing happens in a molecular dynamics simulation, with the exception that instead of Pacman, we have particles.

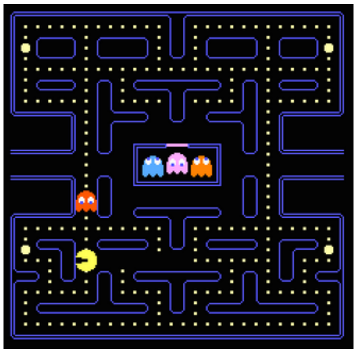


Figure 2.4

The simulation takes place in a computational box, which is virtually surrounded by an infinite number of identical copy boxes (Figure (2.5)). In our case, we care about behavior only of one box, the 'central' one, all the other copies, behave the same as the main box. When PBCs are applied particles may freely pass the boundaries of the box. For each particle leaving, simultaneously an identical one from a copy box enters at the opposite side. The particles of an MD system, are influenced by them who are in same simulation box as well as in replicas boxes.

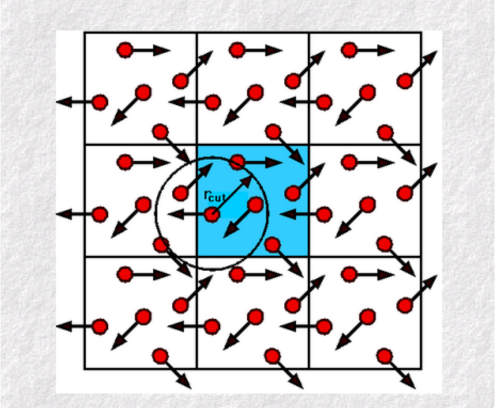


Figure 2.5: 2D Periodic Boundary Condition. The main box (blue) surrounded by copies.

In 3D space, there are five types of a box that we are able to use: the triclinic box, the hexagonal prism,two types of dodecahedrons, and the truncated octahedron. By the conclusion of what we mentioned above, a studied system with periodic boundary conditions is like an infinite system.

2.4 Initial Conditions

2.4.1 Initial Coordinates and Velocities

Before the production of an MD simulation, we need the initial positions and the initial velocities of all atoms in the studied system [13], [14], [15]. The initial positions correspond to the Cartesian coordinates of the initial geometry of the system. In the current work we study proteins, the initial 3D protein structure determined by NMR method (Rop) and X-ray crystallography (RM6).

The initial velocities generated by selecting velocities from the Maxwell Boltzmann distribution, for a specified temperature. The Maxwell Boltzmann distribution, p(u), provides the range of values the program selects the initial velocities of all atoms. This distribution function has the following form:

$$p(u_i) = \sqrt{\frac{m_i}{2\pi k_b T}} exp(-\frac{m_i u_i^2}{2k_b T}) \tag{2.16}$$

In equation (2.16), we denote as m_i the mass of each atom i, k_b the Boltzmann constant, T the temperature and finally u_i the velocity of each atom i.

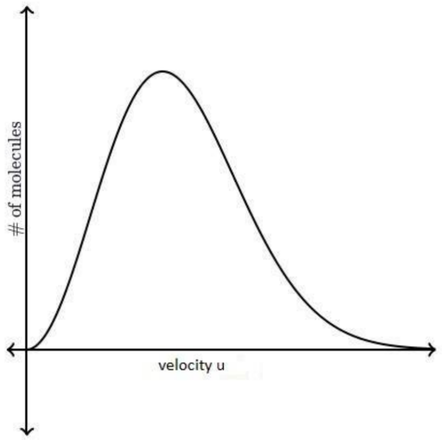


Figure 2.6: Maxwell Boltzmann Distribution.

2.4.2 Force Field

The accuracy of the MD simulations is straightly related to the potential energy function used to describe the interactions between molecules [22],[13], [15], [18]. Most commonly these interactions are split into two terms, bonded interactions and non-bonded interactions. The interatomic potential energy is:

$$U(\mathbf{r}) = U_{bonded}(\mathbf{r}) + U_{nonbonded}(\mathbf{r})$$

A Force Field constituted by a set of these interactions, this set included into a topology file.

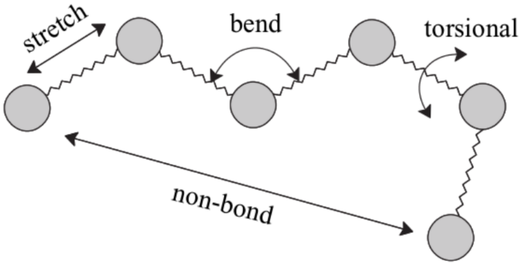


Figure 2.7: Force Field.

Bonded Interactions

The bonded interactions may be described by the following terms:

$$U_{bonded}(\mathbf{r}) = U_{bonds}(\mathbf{r}) + U_{angles}(\mathbf{r}) + U_{dihedrals}(\mathbf{r}) + U_{improper}(\mathbf{r}) \tag{2.17}$$

If we want to model a covalent bond in a molecular structure, we use the harmonic potential. The harmonic potential is the most common, and it can be used in any molecular dynamics program.

$$U_{bonds} = \sum_{bonds} k_b (r - r_0)^2 \tag{2.18}$$

In the above equation k_b is the harmonic force constant, $r = |r_i - r_j|$ is the current bond length, where i and j are the two consequent along the bond and r_0 is the equilibrium bond length.

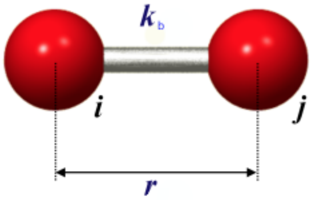


Figure 2.8: Bond [61].

An angle likely defined between two bonds sharing a common atom.

$$U_{angles} = \sum_{angles} k_\theta (\theta - \theta_0)^2 \tag{2.19}$$

where θ_0 is the reference angle and k_θ is the force constant for the harmonics version of the angle potential.

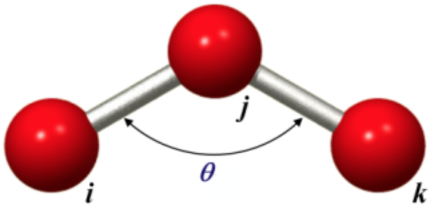


Figure 2.9: Angle [61].

There are two types of torsion potentials, and we can distinguish them into two terms: dihedral angle potential and improper torsions. Both of them depends on a quartet of atoms. The first potential relies on four consecutive bonded atoms, whereas the second depends on three atoms centered around a fourth atom. First, we discuss the proper dihedral type and subsequently the improper type.

The proper torsion potential is expressed as a cosine series expansion:

$$U_{proper} = \sum_{proper} k_{\phi}(1 + cos(n\phi + \phi_0)) \tag{2.20}$$

where k_{ϕ} is the force constant belonging to the cosine type of potential, ϕ is the torsional angle, ϕ_0 determines where the torsion angle passes through its minimum value and n is called multiplicity and gives the number of minimum points in the function as the bond is rotated through 360°.

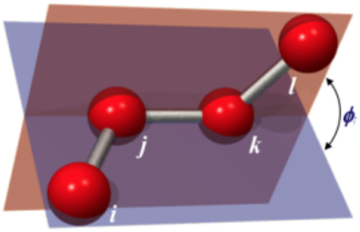


Figure 2.10: Proper dihedral angle [61].

The improper torsion potential is mainly used to maintain the planarity in a molecular structure. It only has one minimum and a harmonic potential may be used. Therefore the improper torsion is given by the equation:

$$U_{improper} = \sum_{improper} k_{\xi}(\xi - \xi_0)^2 \tag{2.21}$$

where k_{ξ} is the force constant for the harmonics version of the improper potential, ξ is the torsional angle between the two planes and ξ_0 the angle where the potential passes through its minimum value.

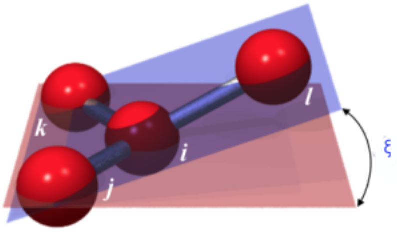


Figure 2.11: Improper dihedral angle [61].

Non Bonded Interactions

Force Field also contains non-bonded interactions. The non-bonded inter- actions depend on the distance between interacted molecules and can express as the sum of Lennard-Jones potential and Coulomb potential:

$$U_{nonbonded}(\mathbf{r}) = U_{LJ}(\mathbf{r}) + U_{coulomb}(\mathbf{r})$$

Lennard-Jones Interactions

The van der Waals interactions are often modeled via the Lennard–Jones potential and is given by the following equation:

$$U_{LJ} = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.22}$$

where

- U is the intermolecular potential between two atoms or two molecules.
- ϵ_{ij} is the well depth and a measure of how strongly the two particles attract each other.
- σ_{ij} is the distance at which the intermolecular potential between the two particles is zero and it is called van der Waals radii.
- $r_{ij} = |r_i - r_j|$ is the distance of separation between both particles.

The Lennard-Jones model consists of two parts: a steep repulsive term r_{ij}^{-6} and smoother attractive term r_{ij}^{-12}.

If we want to calculate the σ_{ij} and ϵ_{ij} parameters we could use the combination rules of Lorent-Berthelot :

$$\sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj})$$
$$\epsilon_{ij} = (\epsilon_{ii}\epsilon_{jj})^{\frac{1}{2}}$$

It is obviously, that the parameters σ_{ij} and ϵ_{ij} depend on the type of molecules.

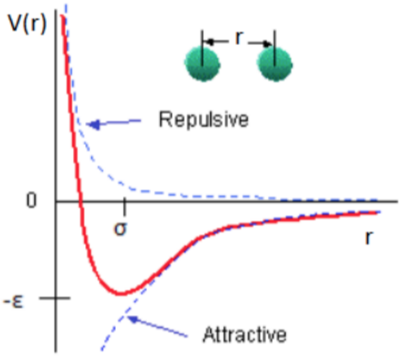


Figure 2.12: The Lennard Jones potential.

Coulomb Interactions

Coulomb interaction describes the electrostatic interaction between two stationary ions (i.e. charged particles). In the Coulomb interaction, the force is repulsive for the same-charged ions and attractive for the opposite-charged ions. The magnitude of the repulsive/attractive forces increases as the charge increases or the separation distance decreases.

The electrostatic potential between two charges Q_1 and Q_2 is described by the following equation:

$$U_{coulomb}(r_{ij}) = \frac{Q_i Q_j}{2\pi\epsilon_0 r_{ij}} \tag{2.23}$$

where Q_i Q_j are the charges of atoms i and j respectively, and ϵ_0 is the permittivity of free space and $r_{ij} = |r_i - r_j|$.

The Coulomb interaction is considered to be a long-range interaction.

Cut-off

The computational cost associated with molecular dynamics simulations is very large and it is a significant obstacle to the study of complex biological processes using MD simulation technique [14].

The calculation of nonbonded interactions, both electrostatic and Van der Waals acting between all pairs of atoms, is the most computationally expensive part in MD simulations. MD simulations trade face a tradeoff between computational efficiency and accuracy. We want to perform MD simulations less expensively or on longer timescales, in order to achieve it a number of approximations of the potential energy function are often employed. There is a common approach to decrease the computational cost, which is to ignore any interaction between atoms separeted by more than some cutoff distance. This approach is accepted in general as being sufficiently accurate for Van der Waals forces, which decay rapidly to zero as the distance increases (Figure 2.12).

Verlet List

Let us consider that we want to simulate a large system, and we use a cutoff that is smaller than the simulation box [14]. then many particles do not contribute to the energy of a particle i.

In this method, we use a second cutoff radius $r_v > r_c$, and before the calculation of the interactions, a list if made, known as the Verlet list, of all particles within a radius r_v of particle i. In the following calculations of the interactions, we have to consider only those particles in this list. We have not save CPU time until now, which is our purpose. We gain such time when we next calculate the interactions. We compute the maximum displacement of the particles, and if this is less than d, then we have to count only the particles in the Verlet list. But if one of the particles is displaced more than d, we should update the Verlet list.

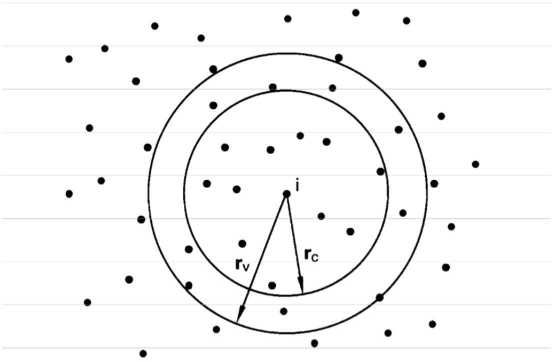


Figure 2.13: The Verlet list: a particle i interacts with those particles within the cutoff radius r_c. The Verlet list contains all the particles within a sphere with radius $r_v > r_c$.

2.4.3 Particle Mesh Ewald Method (PME)

We can simulate biological systems which may contain 10^5 particles [14], [13], [46], [47], [48]. If we have such systems, it is crucial to avoid computing all pair interactions as otherwise the computational cost would be $O(N^2)$. This issue is particularly relevant for long-range interactions,such as Coulombic potential, although is simple. So, it is necessary to find an efficient technique for computing the long-range part of the intermolecular interactions. Moreover, the computation of electrostatic interactions is very complicated, for the reason that we apply periodic boundary conditions, and it requires the calculations for the Coulombic interactions to happen in numerous replicated simulation boxes.

The Smooth Particle Mesh Ewald method (SPME) is an efficient scheme for computing the long-range electrostatic forces which exhibits improved O(NlogN) scaling in comparison to the O(N^2) scaling of standard Ewald summation, differing from the latter in the use of B-spline interpolation and the 3D FFT in computing the reciprocal space summation. The use of B-spline interpolation leads to an energy conservation. In the following, we introduce the SPME method and B-spline interpolation.

Smooth Particle Mesh Ewald (SPME)

If we want to calculate the electrostatic interaction with periodic boundary conditions , we should consider that our system is contained in a unit cubic cell U [46], [47]. Let U defined by the edge vectors $\mathbf{a}_1$,$\mathbf{a}_2$,$\mathbf{a}_3$. We can also denote for later the conjugate reciprocal vectors $\mathbf{a}_1^*$,

$\mathbf{a}_2^*$,$\mathbf{a}_3^*$ which are defined by the following relations $\mathrm{a}_i^*\mathrm{a}_j$=$\delta_{ij}$ (Kronecker delta) for i,j=1,2,3. The unit cell U contained all the N points charges $q_1, \ldots, q_N$ with positions $\mathbf{r}_1, \ldots, \mathbf{r}_N$, as well as, within the unit cell U charge neutrality $\sum_{i=1}^N q_i = 0$ is satisfied. Now, we can define the fractional coordinates of charge q_j by s_{lj}=a, where l=1,2,3.

The charges interact according to Coulomb's law $F = k\frac{q_i q_j}{r_{ij}^2}$, where k is the Coulomb's constant, $r_{ij} = |r_i - r_j|$ and q_i,q_j are the charges of points, with periodic boundary conditions. Thus a point charge q_i at position r_i interacts with other charges q_j, j $\neq$ i at positions r_j as well as with all of their periodic imags at positions $\mathbf{r}_j$+$n_1\mathbf{a}_1$+$n_2\mathbf{a}_2$+$n_3\mathbf{a}_3$

We can write the electrostatic energy of a neutral unit cell U is given by:

$$E(\mathbf{r}_1, \ldots, \mathbf{r}_N) = \frac{1}{2} \sum_n{}' \sum_{i=1}^N \sum_{j=1}^N \frac{q_1 q_j}{|\mathbf{r}_i - \mathbf{r}_j + \mathbf{n}|} \tag{2.24}$$

For simlpicity of notation, we are omitting all factos of $4\pi\epsilon_o$. The prime on the outer sum indicates that summation term with i=j and **n** are omitted. The reason is that the particle does not interact with itself. The $\frac{1}{2}$ in front of the sum is to avoid double counting. The infinite sum in equation (2.24) not only converges very slowly but also is conditionally convergent, meaning that the result depends on the order of the summation. The Ewald method evaluates E by transforming it into summations that converges absolutely: a direct sum in Cartesian space, and a reciprocal sum in Fourier space .

In most MD simulations certain non-bonded interactions within the same molecule are omitted, being handled by different terms in the potential calculation. The corresponding particle pairs (i,j) for which nonimaged nonbond interactions are not calculated are said to be stored in a mask pairlist called M. The masked list correction should be subtraced from the direct and reciprocal energy sums. Usually the masked pairs are explicitly skipped over during the direct sum calculation. The other contribution to the correction term is the charges acting on themselves.

If we want to deal with the Fourier space, we should define the reciprocal lattice vectors **m** by **m**=$m_1\mathbf{a}_1^*$+$m_2\mathbf{a}_2^*$+$m_3\mathbf{a}_3^*$, with m_1, m_2, m_3 integers not all zero. The so called structure factor is defined as:

$$S(\mathbf{m}) = \sum_{j=1}^N q_j exp(2\pi i \mathbf{m} \cdot \mathbf{r}_i) = \sum_{j=1}^N q_j exp(2\pi i (m_1 s_{1j} + m_2 s_{2j} + m_3 s_{3j})) \tag{2.25}$$

where s_{lj} are the fractional coordinates atom j and they defined above.

The electrostatic energy in equation (2.24) can be rewritten as follows:

$$E = E_{dir} + E_{rec} + E_{corr}$$

where

$$E_{dir} = \frac{1}{2} \sum_n{}^* \sum_{i,j=1}^N \frac{q_i q_j erfc(\beta|\mathbf{r}_i - \mathbf{r}_j + \mathbf{n}|)}{|\mathbf{r}_i - \mathbf{r}_j + \mathbf{n}|} \tag{2.26}$$

$$E_{rec} = \frac{1}{2\pi V}\sum_{\mathbf{m}} \frac{exp(-\frac{\pi^2\mathbf{m}^2}{\beta^2})}{\mathbf{m}^2} S(\mathbf{m})S(-\mathbf{m}) \tag{2.27}$$

$$E_{corr} = -\frac{1}{2}\sum_{(i,j)M} \frac{q_i q_j erf(\beta|\mathbf{r}_i - \mathbf{r}_j|)}{|\mathbf{r}_i - \mathbf{r}_j|} - \frac{\beta}{\sqrt{\pi}}\sum_{i=1}^{N} q_i^2 \tag{2.28}$$

In the equation (2.26), the asterisk over the n summation indicates that the terms with n=0 and i=j are omitted. The reciprocal energy E_{rec} requires a generalization of the strucure factor $S(\mathbf{m})$ to include the multipolar interactions. Moreover, V is a unit cell volume ($V = a_1 \cdot a_2 \times a_3$)

The Coulombic force acting on atom i can be obtained by differentiating the sum

$$E(\mathbf{r}_1, \dots, \mathbf{r}_N) = E_{dir} + E_{rec} + E_{corr}$$

with respect to $\mathbf{r}_j$. We are going to refer to the individual terms :

$$-\frac{\partial E_{dir}}{\partial \mathbf{r}_i}, -\frac{\partial E_{rec}}{\partial \mathbf{r}_i}, -\frac{\partial E_{corr}}{\partial \mathbf{r}_i},$$

as the "direct force", "reciprocal force", and "correction force", accordingly.

B-splines interpolation on a grid

In order to approximate the complex exponentials appearing in the structure factors (2.25), we use interpolation [46], [47]. Let us assume that we have a discrete 3D grid of sizes K_1, K_2, K_3 of the unit cell U. We can write the scaled fractional coordinates of the particle with position $\mathbf{r}$ as: $u_j = K_j \mathrm{a}_j^* \cdot \mathbf{r}$, for j=1,2,3. Therefore,

$$exp\Big(2\pi i\mathbf{m} \cdot r\Big) = exp\Big(2\pi i\frac{m_1 u_1}{K_1}\Big) \cdot exp\Big(2\pi i\frac{m_2 u_2}{K_2}\Big) \cdot exp\Big(2\pi i\frac{m_3 u_3}{K_3}\Big) \tag{2.29}$$

In the smooth PME method, we use cardinal B-splines M_l for the interpolation which produces sufficiently smooth energy functions, allowing for analytical differentiation to arrive at the forces. If we use the Euler exponential spline, the exponential can be represented for interpolation of order l as:

$$exp\Big(2\pi i\frac{m_i u_i}{K_i}\Big) \approx b_i(m_i)\sum_{k\in\mathbb{Z}} M_l(u_i - k) exp\Big(2\pi i\frac{m_i u_i}{K_i}\Big) \tag{2.30}$$

where $b_i(m_i)$ is given by,

$$b_i(m_i) = exp\Big(2\pi i\frac{m_i}{K_i}(l-1)\Big) \times \Big[\sum_{k=0}^{l-2} M_l(k+1) exp\Big(2\pi i\frac{m_i k}{K_i}\Big)\Big]^{-1} \tag{2.31}$$

Using the equation (2.30), the structure factor can be approximated as

where F is the Fourier transform,

$$Q(k_1, k_2, k_3) = \sum_{i=1}^{N} \sum_{n\in\mathbb{N}^3} q_i M_l(u_{1i} - k_1 - n_1 K_1) \times M_k(u_{2l} - k_2 - n_2 K_2) \cdot M_l(u_{3i} - k_3 - n_3 K_3)$$

We can approximate the reciprocal energy by:

$$\tilde{E}_{rec} = \frac{1}{2\pi V} \sum_{m\neq\mathbb{O}} B(\mathbf{m}) F(Q)(\mathbf{m}) F(Q)(-\mathbf{m}) = \frac{1}{2} \sum_{m_1=0}^{K_1-1} \sum_{m_2=0}^{K_2-1} \sum_{m_3=0}^{K_3-3} Q(m_1 m_2 m_3)(\theta_{rec}{*}Q)(m_1 m_2 m_3) \tag{2.32}$$

where θ_{rec} is the pair potential and is given by

$$\theta_{rec} = F(B \cdot C)$$

and

$$B(\mathbf{m}) = |b_1(m_1)|^2 \cdot |b_2(m_2)|^2 \cdot |b_3(m_3)|^2 \tag{2.33}$$

The reciprocal force does not depend on particle positions, so we can calculate the resulting reciprocal force using the spline parameters:

$$\mathbf{F}_{ij} = -\frac{\partial \tilde{E}_{rec}}{\partial \mathbf{r}_{ij}} = \sum_{m_1=0}^{K_1-1} \sum_{m_2=0}^{K_2-1} \sum_{m_3=0}^{K_3-3} \frac{\partial Q}{\partial \mathbf{r}_{ij}}(m_1 m_2 m_3)(\theta_{rec} * Q)(m_1 m_2 m_3) \tag{2.34}$$

Chapter 3

Simulated Systems and Definition of Measured Quantities

In the current chapter, we describe the studied systems and the parameters of the simulations. Then, we define the quantities that we have investigated.

3.1 Systems

In Table 3.1 we present our simulated systems and conditions. Listed are the simulation identifiers, the total number of atoms in the simulation box, the number of water molecules, the number of chains which each proteins has, the number of ions (Na^+), which we added to neutralize our systems, the total time of simulation for each system and the temperature.

System	Protein Atoms	Protein Chains	Total Atoms	Water Mol.	Ions (Na^+)	Time (ns)	T (K)
Rop (Native State)	1280	2	73861	24189	14	100	300
RRC (Remote chains of Rop)	1280	2	72160	23622	14	300	300
RM6 (Mutation of Rop)	2376	4	243199	80269	16	200	300

Table 3.1: Details of the simulated systems.

Atomistic molecular dynamics (MD) simulations were performed in (NPT) statistical ensemble, using Gromacs software package (version 5.0.7) [54],[55], with gromos53a6 force field [45]. In this force field the non polar hydrogen atoms are treated as united atoms together with the carbon to which they are attached. The proteins were placed in a cubic box. The proteins were solvated using the Extended Simple Point Charge (SPC/E) water model. The solvated proteins were energy minimized using a steepest descent algorithm [13]. Simulations were carried out using 1fs time step and periodic boundary conditions have been applied. Initial velocities were assigned randomly applying Maxwell distribution at 300K. Long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) summation, with a 10$\mathring{A}$ cut-off for the direct space of sums, a 1.6$\mathring{A}$ FFT grid spacing and a 4-order interpolation polynomial for the reciprocal space sums. The temperature was kept at 300K by separately coupling the protein and solvent to an external temperature bath (t=0.1ps). The thermostat we used was the V-rescale. The pressure was kept constant at 1 bar by weak coupling (t=1.0ps) to a pressure bath according to Berendsen barostat using isotropic mode. The pressure was maintained with a compressibility of 4.6 10^{-5}/bar.

MD Simulations		EM Simulations	
dt (ps)	0.001	emtol	1.0
Tcoupling	Yes/V-rescale	emsteps	5000
tau_t (K)	0.1		
Pcoupling	isotropic/Berendsen		
tau_p (ps)	1.0		
compressibility (bar^{-1})	4.5 10^{-5}		
ref_p (bar)	1.0		

Table 3.2: A summary of the parameters used in the energy minimization (EM) amd MD simulations.

3.2 Rop and RM6 in Native State

Generation and Equilibration of the model system

Before the produce MD runs, models system are generated and equilibrated through the following steps.

Rop

1. We take the structure file of the protein from the protein data bank. This file contains the coordinates of the protein. Then, we check pdb file to ensure that there are no missing atoms or amino acids. If there are missing atoms or amino acids it is written inside the pdb file. In the case of Rop the pdb file is fine.

2. Once, we are sure that there are no missing atoms or residues we create the topology file. This file contains all the necessary information to define the molecule within a simulation. This information includes nonbonded parameters (atom types and charges) as well as bonded parameters (bonds, angles, and dihedrals). The topology file is created with the gromacs tool (gmx pdb2gmx). Except from the topology this tool creates a post-processed structure file in gro format (.gro file).

3. We want to simulate our systems in aqueous solution. This step is splitted into two actions.

(a) We define the dimensions of the simulated box and centere our protein molecule.

(b) Then, we fill the box with water molecules.

4. The Rop protein has a net charge of -14e (based on the amino acid chain composition). We neutralize the system adding ions. Again, with a tool of gromacs (called genion), we replaced water molecules with 14 positive charged ions Na^{+}.

5. Before we perform the production of MD run, we had to ensure that it had no steric clashes or inappropriate geometry. For that, we performed energy minimization.

6. After we completed all the above steps, we performed the production of MD simulation run for time reported in Table (3.1).

7. When our simulation finished, we analyzed the simulated system.

8. In terms of Rop protein we keep 1600 configurations (we keep a configuration every 62.5 ps). In terms of RM6 protein we keep 1342 configurations (we keep a configuration every 156.3ps).

RM6 protein

For the RM6 protein we also followe the same procedure describe above. However, in steps 1, 4 and 8 we do something different. In terms of step 1, there are missing residues in RM6 protein. So the process we follow to prepare the RM6 is the following:

1. The pdb file contain only the two chains of the tetramer. So, we have to build the tetramer, this is done with the help of the tool called pisa [56].

2. When we create the tetramer, we check it once more. Then, we seethat the last seven residues are missing from all the four chains, as well as the first four amino acids but only from two of them. We surpass this issue with the help of a useful tool named PyMol [57]. With this tool we add the missing amino acids.

3. Also, there are no Hydrogen atoms. This problem is easy to be solved because the tool of gromacs (pdb2gmx) adds the missing Hydrogens, where they should be.

Now about step 4, accordingg to the amino acid chain composition RM6 has a net charge of -16e. So, we add 16 ions Na^{+}, in order to neutralize the system.

Finally, for the step 8, we keep 1342 configurations (we keep a configuration every 156.3ps).

3.3 Measured Quantities

3.3.1 Bioinformatics Analysis: Root Mean Square Deviation (RMSD)

Quantitative comparisons of protein has a central importance in structural biology [41],[42],[43],[44]. For example, biomolecular activity is often coupled to changes of biomolecular structure and dynamics. Therefore, it is useful to quantify the evolution of three-dimensional (3D) structures of a protein molecule in time or to compare structure of similar proteins (i.e mutations), coming from Molecular Dynamics simulations trajectories of proteins, are commonly analyzed by comparison of each snapshot with respect to a reference structure. We often take as reference structure the starting structure (coming from NMR or X-ray diffraction measurements). The conformational changes of the protein during MD simulations were checked by the root mean square derivations (RMSD). RMSD is the most widely used measure for comparison protein structures; it is a measure of molecules mobility, from its reference state. It is calculated by translating and rotating the coordinates of the instantaneous structures in order to superimpose with the reference with a maximum overlap. In order to use RMSD, the two structures (the reference and the one we want to compare) should have identical number and types of atoms.

RMSD is generally computed after a least-squares fit of the structure to the reference structure to by removing the effects of global translation and rotation, which minimizes the differences between the two structures.

The RMSD is defined by the following formula:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} m_i(\mathbf{r}_i - \mathbf{r}_{ref})^2}{\sum_{i=1}^{N} m_i}} \tag{3.1}$$

where:

- m_i is the mass of atom i,
- $\mathbf{r}_i = (r_{i,x}, r_{i,y}, r_{i,z})$ is the coordinates of atom i at a certain instance,
- $\mathbf{r}_{ref} = (r_{ref,x}, r_{ref,y}, r_{ref,z})$ is the coordinates of atom i at its reference state.

In the following subsections, we explain what we mean by fitting and least-squares fit.

Superposition (fitting) of protein structures

We should reflect the internal motion of proteins, in order to achieve it we should have more or less the same orientation in space [41],[42],[43]. This is done by fitting the models on the reference structure.

Least square fit

In a mathematical way, the superposition is actually a problem of minimize RMSD between two models (two sets of identical atomic positions). In order to minimize the RMSD, we need a rotation matrix $\mathbf{R}$ and we have:

$$RMSD_{min} = \sqrt{\frac{\sum_{i=1}^{N} m_i(\mathbf{r}_i - \mathbf{R}\mathbf{r}_{ref})^2}{\sum_{i=1}^{N} m_i}}$$

In general, as least square fit is called any operation that has as result an appropriate rotation matrix $\mathbf{R}$.

An algorithm for achieving a minimum RMSD is introduced by W. Kabsch in 1976. Let us assume, that we have two sets of centered atomic positions which are given by $\mathbf{r}_i$ and $\mathbf{r}_{ref,i}$. If the points are not centered to the origin, it is very important to translate them in order to their average coincides with it.

The 3×3 rotation matrix $\mathbf{R}$ is used to rotate the points of $\mathbf{r}_{ref,i}$ into $\mathbf{r}_i$. First two $N \times 3$ matrices $\mathbf{P}$ and $\mathbf{Q}$ are produced, which contain the coordinates of $\mathbf{r}_i$ and $\mathbf{r}_{ref,i}$ as row vectors accordingly.

$$\mathbf{P} = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_N & y_N & z_N \end{pmatrix}$$

$$\mathbf{Q} = \begin{pmatrix} x_{ref,1} & y_{ref,1} & z_{ref,1} \\ x_{ref,2} & y_{ref,2} & z_{ref,2} \\ \vdots & \vdots & \vdots \\ x_{ref,N} & y_{ref,N} & z_{ref,N} \end{pmatrix}$$

After that, we calculate the cross-covariance matrix as follows:

$$\mathbf{A} = \mathbf{P}^T\mathbf{Q}$$

Then, the rotation matrix is given by:

$$\mathbf{R} = (\mathbf{A}^T\mathbf{A})^{1/2}\mathbf{A}^{-1}$$

But, we can not guarantee that the matrix $\mathbf{A}$ has an inverse. In terms of these special cases, that $\mathbf{A}$ has not an inverse, is carried out

$$\mathbf{A} = \mathbf{V}\mathbf{S}\mathbf{W}^T$$

So, the rotation matrix **R** is given by:

$$\mathbf{R} = \mathbf{W}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix}\mathbf{V}^T$$

where $d = sign(det(\mathbf{W}\mathbf{V}^T))$

If we have two identical protein configurations, then we expect that the function (3.1) equals to zero [7]. In general, this is impossible due to the offset of translation and rotation. So, for example, a configuration with an RMSD value less than $2.5\mathring{A}$ is considered a very close approximation to the native state. The RMSD is getting even worst for configurations away from the native state.

3.3.2 Radius of Gyration (Rg)

Radius of gyration (Rg) is a measure which we generally use in order to predict the dimensions of a macromolecule [29], [30]. It is an indicator of the protein size and of the protein compactness. The radius of gyration is calculated using the formula:

$$Rg = \sqrt{\frac{\sum_{i=1}^{N} m_i r_i^2}{\sum_{i=1}^{N} m_i}} \tag{3.2}$$

where:

- m_i is the mass of the atom i
- r_i is the distance of atom i from the center of mass of the protein.

3.3.3 Hydrogen Bonds

Atoms in proteins are held together by various types of bonds with different strength including very strong covalent bond, moderate hydrogen bonds and weak Van der Waals bonds [31], [32], [33], [34]. A Hydrogen Bond tends to be stronger than Van der Waals bonds, but weaker than the covalent bonds or ionic bonds.

Hydrogen bonding constitutes one of the most significant inter-atomic interaction in biology. Moreover, it plays an important role in the stability of conformations in secondary structures of proteins.

The typical definition of a hydrogen bond (HB) demand the presence of 3 atoms. Generally, a HB within a protein molecule is an attractive interaction between an electronegative donor (D) and a hydrogen (H) that is covalently bonded to an acceptor (A) atom of a non-adacent residue. Following this convention, we represent the HB within dotted lines, as follows:

$$\textbf{D-H} \cdots \textbf{A}$$

where the dash line between the D atom and the H represents a chemical bond.

There are two kinds of criteria, that are often used for determining the HBs i.e, by energy and by geometry.

In the current work, we are going to use the geometric criteria to detect the existance of a HB. With geometric criteria, the HB is determined by the relative configuration of the two molecules. The geometric parameters involve interatomic distances and angles. A set of geometric parameters are illustrated in Figure 3.1. The geometric criterion we used to determine a HB is the following:

- $|D - A| \leq 3.50$ Å, where $|D - A|$ describes the distance between Donor and Acceptor.
- $\angle$HDA $\leq 30^o$, where the angle $\angle$HDA describes the Hydrogen-Donor-Acceptor angle.

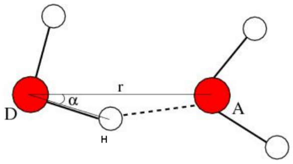


Figure 3.1: Geometric Parameters of Hydrogen Bonds.

Intermolecular and Intramolecular Hydrogen Bonds

Typical HB donors would make only one HB, whereas typical oxygen acceptors can make two HBs [19]. We separate two types of hydrogen bonds, namely:

1. **intermolecular** hydrogen bonds between two molecules
2. **intramolecular** hydrogen bonds within the molecule.

The intermolecular and intramolecular hydrogen bonds are represented in Figure (3.2).

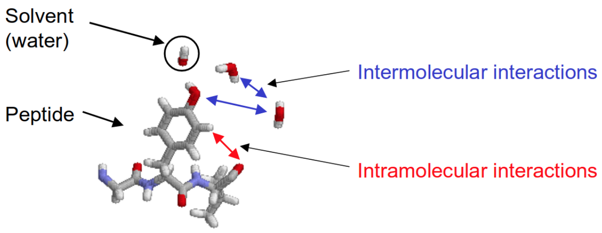


Figure 3.2: Intermolecular and Intramolecular Hydrogen Bonds.

3.3.4 Ramachandran Plot

In 1968, Ramachandran and Sasisekharan proposed that the conformation of amino acids in proteins and peptides could be described by two torsion angles ϕ and ψ [35], [36], [37], [38], [39], [40]. The phase diagram of the two torsion angles is known as the Ramachandran plot. This plot was first used to predict the possible conformations of the main chain (i.e. backbone of protein). We should note that the Ramachandran plot constructed before the first protein structure had been determined to atomic resolution by X-ray diffraction. In the following we define the dihedral angles in order to make clear the description of Ramachandran plot. The torsion (or dihedral) angle is applicable to any molecule has a system of four atoms. Let us denote them as A, B, C and D connected by covalent bonds. In Figure (3.3) we show a schematic representation of them.

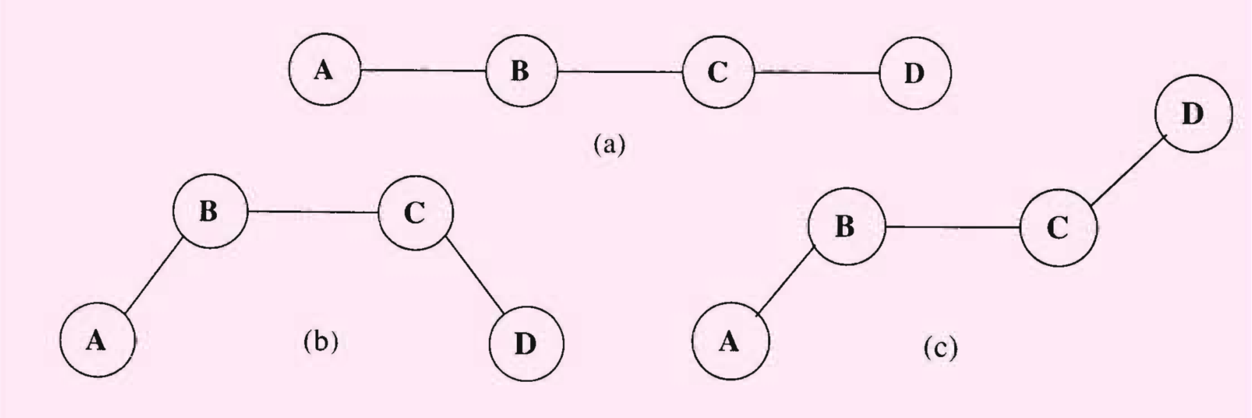


Figure 3.3: (a)A representation of a system of 4 atoms bonded linearly.Orientation in the system ABCD (b) & (c) .In (b) we have 'cis' form. In (c) we have the 'trans' form.

In principle, a dihedral angle represents the angle between two intersecting planes as shown in Figure (3.4) by subsequent triples of points (atoms), namely between the planes spanned by A, B, C and by B, C, D [19]. We should keep in mind that a dihedral angle is an angle of rotation about the line BC at the intersection of two planes.

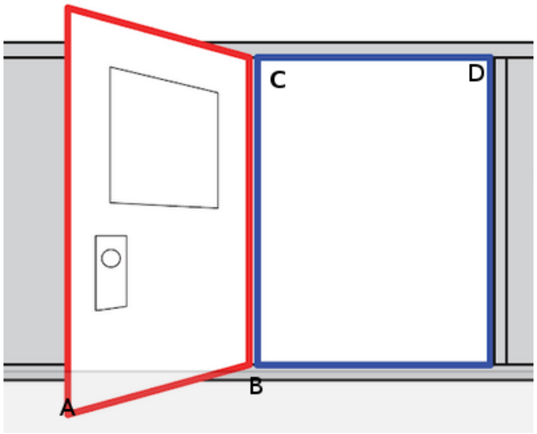


Figure 3.4: A schematic representation of torsion angle.

Dihedral is defined in terms of the angle between the normal vectors for the two planes. The normal vector to the plane spanned by A, B, C, as well as by B, C, D is proportional to $n_1 = \vec{AB} \times \vec{CB}$ and $n_2 = \vec{BC} \times \vec{DC}$ respectively, where × denotes the vector ”cross” product in $\mathbb{R}^3$. The angle θ between these two normal vectors is given by the following formula:

$$cos\theta = \frac{n_1 n_2}{|n_1||n_2|}$$

where $|n_i|$ denotes the Euclidean length of n_i. Let θ =[A, B, C, D] denotes a dihedral (or torsion) angle.

Ramachandran angles

Once, we briefly introduced what is a torsion angle, we are going to explain the two angles of the Ramachandran plot. There are two bond whose rotations are permissible. These two bonds are $N - C_\alpha$, as well as, $C_\alpha - C$, hence torsion angles have been defined with respect to these bonds. These angles are going to be symbolized as ϕ and ψ and we define them as follows:

- ψ(i)=[N(i),C$_\alpha$(i),C(i),N(i+1)]
- ϕ(i)=[C(i),N(i+1),C$_\alpha$(i+1),C(i+1)]

where i denotes the amino acid number.

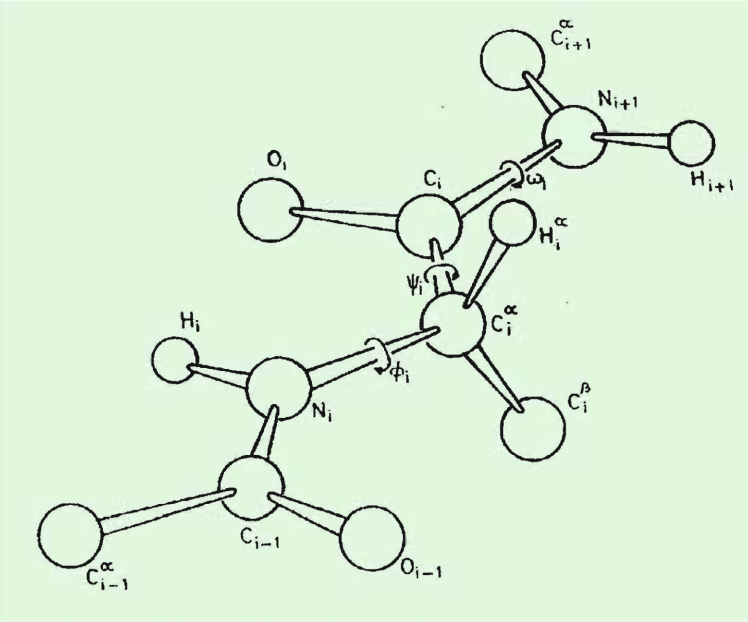
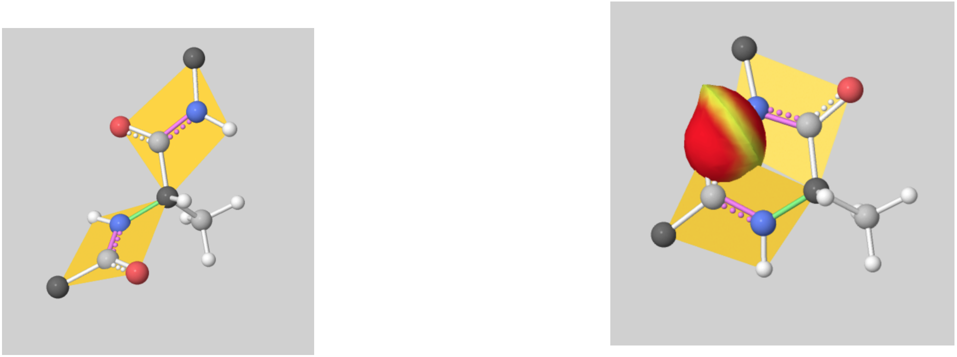


Figure 3.5: A schematic representaion of the two torsion angle in a dipeptide.

It is significant to note that the internal rotation around these bonds is restricted by possible steric collisions in the produced conformations.

Moreover, the bond between C(i) and N(i+1), known as the peptide bond has partial double-bond character. The peptide bond cannot rotate freely, contrary it is held in a planar orientation, so this restriction drives it to have two possible configurations, the 'cis' and the 'trans' presented in Figure (3.3 (b) and (c), respectively) and in Figure (3.7). Due to steric and electronic interactions the trans form is the most common, whereas the cis form is very rare. In the cis form, it is very common to have a steric clash between the side chains. As we have already mentioned, the basic unit of the peptide group comes in two forms (Figure 3.7) that are related by a rotation around the peptide bond. For this reason this angle, which defined as ω(i)=[C_{α}(i),C(i),N(i+1),C_{α}(i+1)], could take two possible values. These are the following:

- ω=0^o in the 'cis' form and
- ω=180^o in the 'trans' form.



(a) There is no clash.

(b) There is a clash during the rotation of the two bonds.

Figure 3.6: A schematic representation of a polypeptide chain. In the center we have a complete alanine residue. The alanine is covalently bonded to other amino acids through peptide bonds (pink). In Figure 3.6a there is no overlapping among the atoms, whereas in Figure 3.6b there is an ovelapping, which is physically impossible. [53]

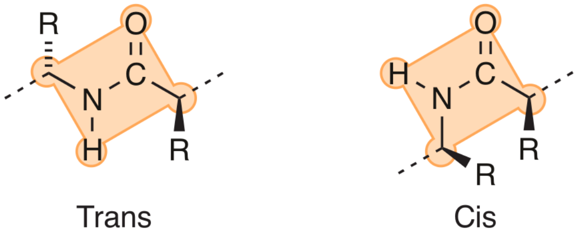


Figure 3.7: The trans and cis conformations.

Since the angle ω could obtain only two possible configurations, so primarily, the conformation of a protein described by the combination of angles ϕ and ψ.

Regions of Ramachandran plot

It is known that the atoms in a structure could not approach each other more closer than a distance of above the sum of their Van der Waal's radii. For some combinations of ϕ and ψ, with the trans conformation of the peptide bond, atoms would collide, but as we said this is a physical impossibility, for that reason ramachandran plot was used to plot the sterically allowed regions. For example for alanine-like residues there are three major "allowed" regions that they can occupy. In this family do not belong Glycine and Proline. But the other 18 residues have a common Ramachandran plot. We should note that Glycine and Proline are outliers since they are the extreme points in terms of conformational freedom, so these two have their own Ramachandran plots.

There are two major allowed regions of residue conformation, with the largest extend. These are the $\alpha - region$ and the $\beta - region$ and match to the two major classes of secondary structure: the α-helix and the β-sheet respectively. If there is a succession of residues in the $\alpha - region$ the generated structure is an α-helix. The α-helix is right-handed. Additionally, a sequence of residues in the $\beta - region$ produces a strand that is nearly fully extended. This is the conformation of β-sheet. Therefore, there is a much smaller region known as $\alpha_L - region$ representing backbone conformations that are mirror images of those in the $\alpha - region$. These regions are "allowed", in the sense that, when the peptide given standard radii they don not collide. There is also an additional region called the bridge region because it acts like a bridge between α- and β- regions. It becomes approved if the atoms are given smaller radii that represents the least values that could be considered plausible.

Let us explain the Ramachandran plot, which is shown in Figure 3.8. There are regions shaded with dark blue. These areas reflect conformations that involve no steric overlap and thus are fully allowed (and we will call them the favoured regions). Medium blue indicates conformations allowed at the extreme limits for unfavourable atomic contacts. Moreover, there is area shaded with the lightest blue which reflects the conformations that are permissible if a little flexibility is allowed in the bond angles. Finally, the unshaded region indicates sterically disallowed conformations. In total we can sum up that the gradient of blue color from dark to light regions indicates the passage from the most to the less favorable conformations.

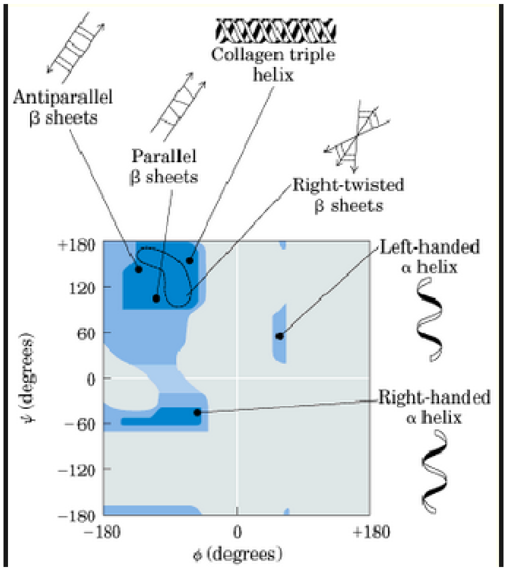


Figure 3.8: A Ramachandran Plot for alanine-like residues. The values of ϕ and ψ for various allowed regions [59].

Finally, we should note that in a Ramachandran plot, we do not plot the first and last amino acids, because for the first amino acid we need a residue before it and for the last residue a residue after it, in order to be able to find the (ϕ-ψ) angles.

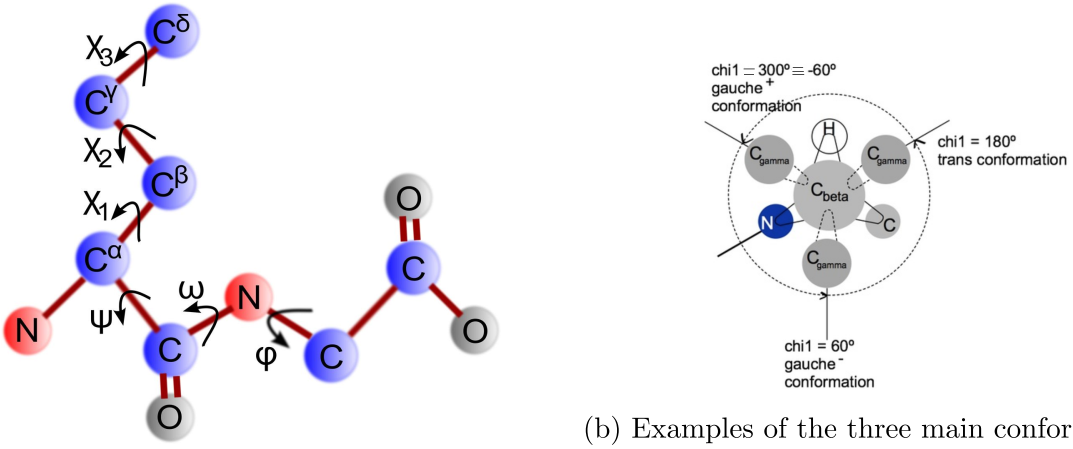
Based on all the above, nowadays, Ramachandran plot is converted into a useful tool for assessing the correctness of a protein structure determination.

3.4 Sidechain Rotamers

The sidechains of an amino acid are flexible. So, the geometric description of sidechain requires additional information [19]. As we remeber from the above section, the ϕ and ψ angles define the positions of the backbone relative to the position of the N-terminal end, so we can also use dihedral angles to describe the positions of the sidechains. For instance, we can define them as follows:
$\chi_1 = [N, C_\alpha, C_\beta, C_\gamma]$, $\chi_2 = [C_\alpha, C_\beta, C_\gamma, C_\delta]$, and so forth, for all sidechains that have a C_β atom attached to an additional "heavy atom" (not Hydrogen). From the sidechain rotamers glycine and alanine are excluded.

The knowledge of the conformation of side chains is an important feature of protein architecture [27]. There are factors that affect the side-chain conformations and the knowledge of them is significant, both for the understanding of protein folding and for the successful design of mutated proteins.

Side chains in proteins prefer certain conformations, as shown by Janin et al. [23]. The preferred conformations correspond to energy minima and generally represented by three regions, which are known as **gauche+**, **gauche-**, and **trans**, corresponding to mean χ_1 values of -60 degrees, +60 degrees, and 180 degrees, respectively, as shown in Figure (3.9b). There are studies [24],[26],[25] , that revealed that the preferences of the rotamer of side chains are strongly affected by the secondary structure.



(a) The dihedral angles of sidechains.

(b) Examples of the three main conformations.

Figure 3.9: In Figure: (3.9a) we present the sidechain dihedral angles, and in Figure: (3.9b), we present the three possible conformations (gauche -, trans, gauche +).

In Figure (3.10), we present the nine regions, which the compinations of angles (χ_1-χ_2) may have. The nine regions are: (g-,g+), (g-,t), (g-,g+), (t,g+), (t,t), (t,g-), (g+,g+), (g+,t) and (g+,g-).

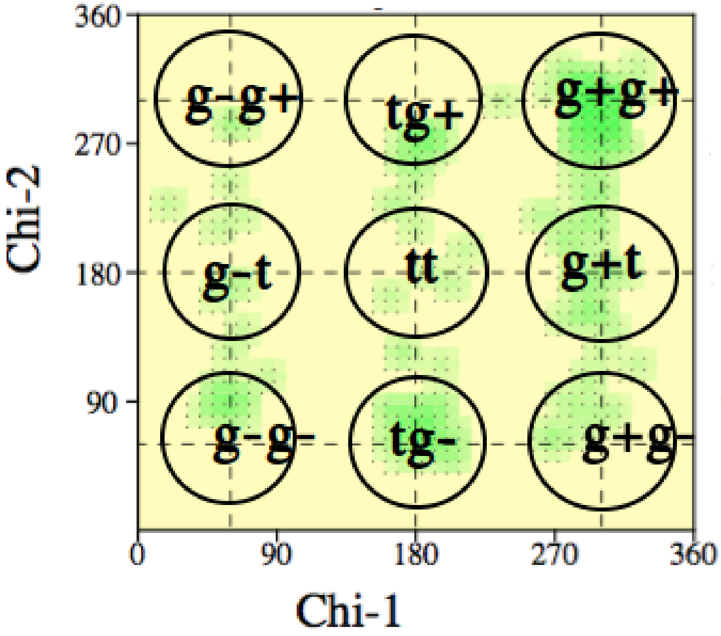


Figure 3.10: The nine regions of Janin plot of (χ_1-χ_2) tosion angles.

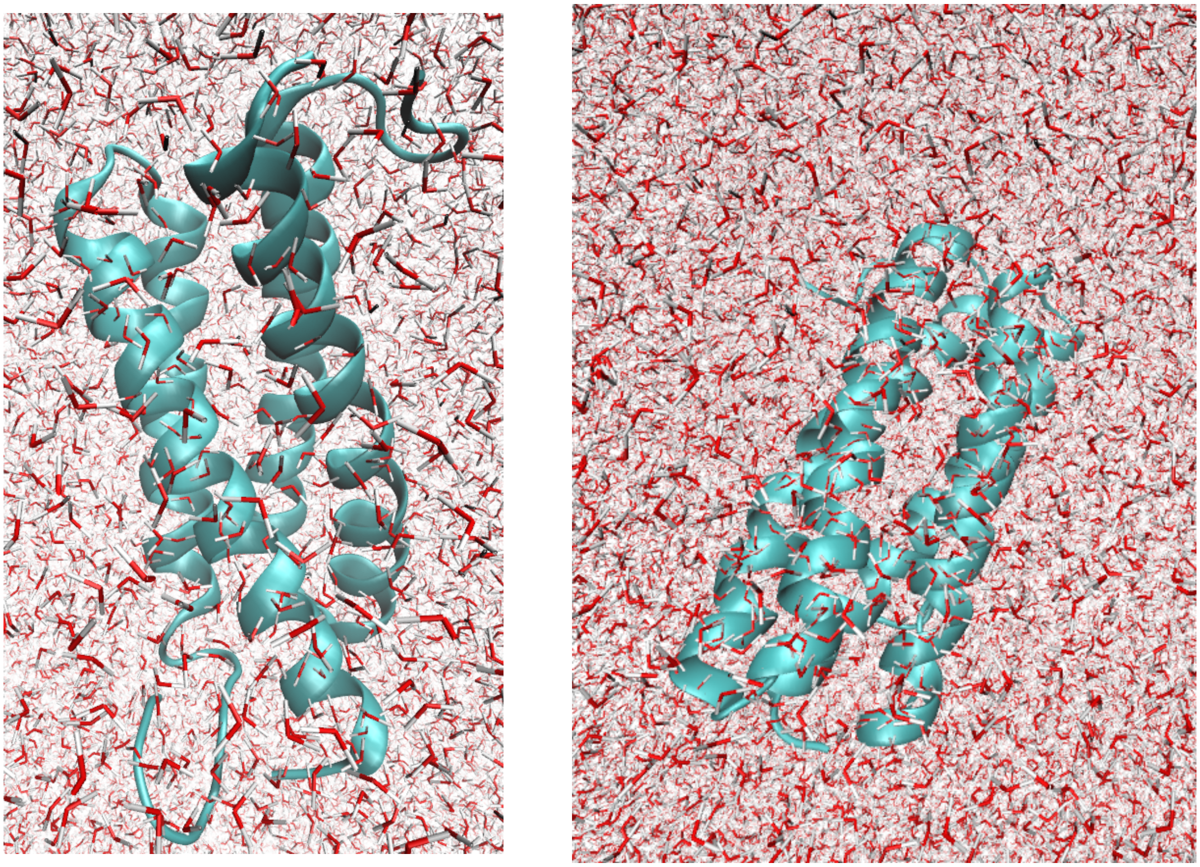
Chapter 4

Results

In this chapter, we present the results from our simulation work. In Chapter 3, we are defined analytically the measured quantities . As we have already mentioned, we have simulated two different systems. The results are going to be presented as follows (1). We present each system separately, and then we perform a comparison between them (2). For each system we provide the initial configuration snapshot (3). We calculate the root mean square deviation (RMSD) as an indicator of the stability of the native state of our proteins (4). We quantify the mean size of the proteins using the radius of gyration (5). In the following we carry out a more detailed analysis, which is the calculation of the number of hydrogen bonds, which are formed among the components of each system during the simulation (6). Finally, we check the stereochemical quality of our model by plotting the Ramachandran plot.

4.1 Rop in Native State

The first system that we are going to present is the Rop protein in aqueous solution. In Figure 4.1, we present two different snapshots one of the initial configuration of our system and one of the final after the completion of the simulation run (100 ns). Not any observed difference in the structure can be reported on the snapshot since the protein was already in its native state (i.e stable).



(a)

(b)

Figure 4.1: Snapshots of Rop system in water solution, (4.1a) initial configuration and (4.1b) final configuration.

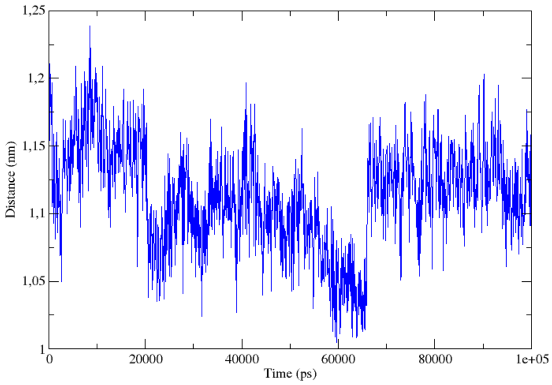


Figure 4.2: Plot of the distance between the centers of mass of two chains of Rop.

In the native state,the distance between the centers of mass of the two chains is $\Delta R_{cm} = 1.2$

nm, where $\Delta\mathbf{R}_{cm} = \|\mathbf{R}_{cm,chain1} - \mathbf{R}_{cm,chain2}\|$. As shown in Figure 4.2 it is almost constant during the simulation.

4.1.1 RMSD

We calculate the root mean square deviation (RMSD) as an indicator of the accuracy of our model. This could also ensure that our system is in equilibrium. The RMSD is a measure of the difference between two structures. Each configuration from a trajectory file is compared to a reference structure. In our study, the reference conformation is the initial configuration of the simulation. All the measurements of the RMSD are done based only on the C_α atoms, since these atoms are often used in typical computations of RMSD.

The RMSD of all C_α backbone atoms with respect to the reference structure as a function of time is shown in Figure 4.3. The RMSD of the protein at the end of the simulation is 0.258 nm= 2.58$\mathring{A}$. According to the literature (ref), this is a very good result since a configuration with a value of RMSD below 2.5$\mathring{A}$ thought as a close approximation of the native state. RMSD after the first steps (very short times) attains values between $[0.2 - 0.3]$

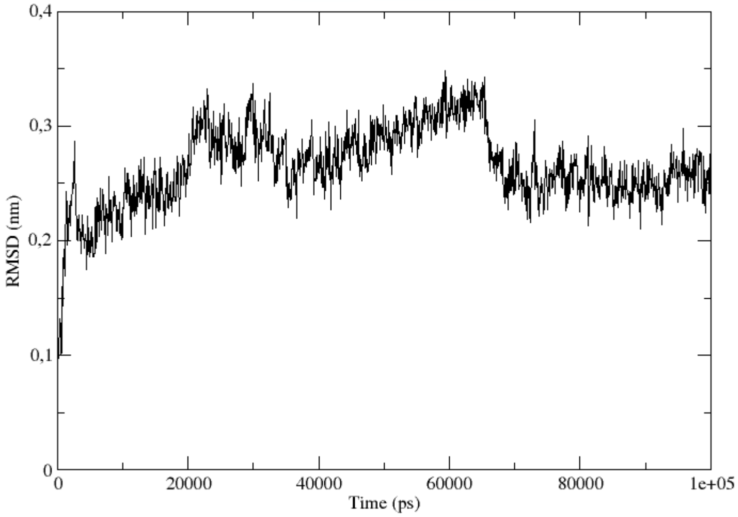
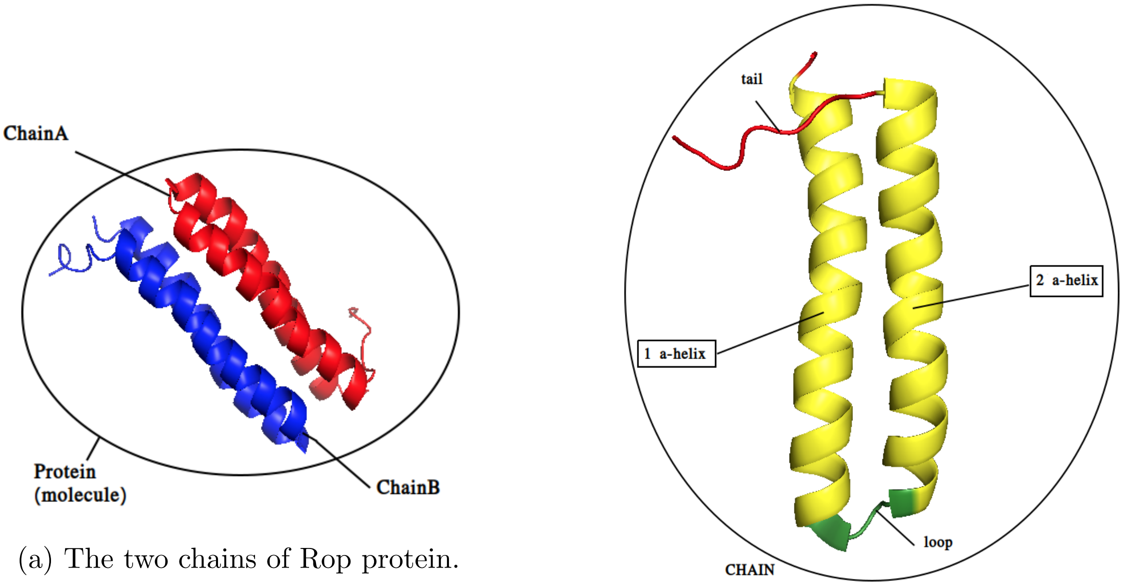


Figure 4.3: Root Mean Square Deviation (RMSD) plot showing the RMSD based on C_α atoms as a function of time for the simulation of 100 ns for Rop in water.



(a) The two chains of Rop protein.

(b) The three different parts of a chain.

Figure 4.4: Different parts of Rop.

Measured Quantities	RMSD (nm)
loopA	0.023 ± 0.0017
loopB	0.013 ± 0.0008
tailA	0.156 ± 0.0234
tailB	0.350 ± 0.0445
ChainA	0.173 ± 0.0060
ChainB	0.317 ± 0.0256
Two Helices of chainA	0.119 ± 0.0059
Two Helices of chainB	0.131 ± 0.0049
Protein	0.275 ± 0.0201

Table 4.1: Results are averaged over all the trajectory. The unit of RMSD is in nm.

The RMSD analysis shows that the system is in an equilibrated structure. Moreove, in Table: 4.1 we present a detailed analysis for the RMSD of the various parts of the protein separeted into helix, loop, tail, chain. As we have already mentioned in Section 1.3, the protein has two chains, and we are going to call them as ChainA and ChainB (Figure:4.4a). Each chain consisting of the the following parts (Figure:4.4b), two α-helices, a loop, and a tail. So, we marked with A the parts of ChainA and with B the parts of ChainB. Sometime, we are going to refer as a molecules the protein consisting of two chains. We observe that the RMSD of the last seven residues (the tail) of each chain is bigger than the rest chain, which comes in agreement with th results from experiments, which say that the tail is very flexible. We also calculate RMSD for the two tails and the four α-helices, and as we can see the conformations of them remain almost the same. Therefore, it is the flexibility of the two tails, which renders the value of RMSD of the whole protein quite high.

4.1.2 Radius of Gyration (Rg)

The Radius of gyration (Rg) is a measure of the compactness of the protein conformation in biomolecular simulations and quantifies their size. As we can see in Figure (4.5) the value of Rg is relatively stable over time, which means that the Rop is stably folded at the equilibrium conformation (native state).

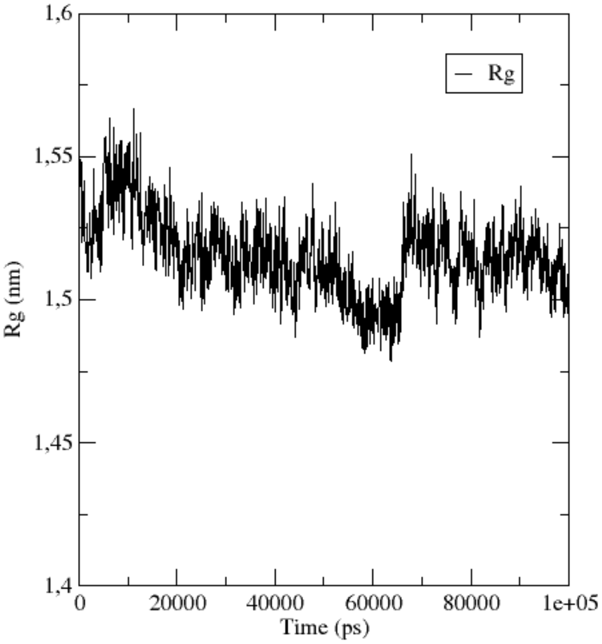


Figure 4.5: Radius of gyrarion as a function of time for Rop in water.

The Rg in (nm) was computed by taking the average over all the trajectory and is presented in Table (4.8).

Measured Quantities	Rg (nm)
ChainA	1.404 ± 0.004
ChainB	1.410 ± 0.003
Protein	1.511 ± 0.004

Table 4.2: The average value of Rg over all the trajectory of Rop in water.

In the first column of Table (4.8), the ChainA and ChainB correspond to the first and second chain of Rop protein accordingly, and finally, the Protein corresponds to the whole Rop protein.

4.1.3 Hydrogen Bonds

Next, we analyze hydrogen bonds using gromacs tool (gmx hbond). For each pair of atoms, which form hydrogen bond, the bond is counted once. We split the analysis of hydrogen bonds into intermolecular and intramolecular hydrogen bonds, as we discussed in Section: 3.2. More specifically in our case about Rop, we further categorize HB as follows: we split the intramolecular hydorgen bonds into intechain hydorgen bonds between the two chains (ChainA-ChainB), as well as intrachain hydrogen bonds within the same cahin (ChainA-ChainA or ChainB-ChainB). The intramolecular hydrogen bonds are within the same molecule, in our case, the molecule is the Rop protein, which has two chains, let us denote them A and B. So, we are going to call intramolecular hydrogen bonds, the bonds within chain A, as well as within chain B. Additionally, we have the intermolecular hydrogen bonds involve two or more molecules, such as Protein-Water, ChainA-Water, ChainB-Water, Water-Water. The notation we are going to use for the protein is the letter P and the W for water. The geometric criterion we used to calculate the number of hydrogen bonds per each frame is: $\angle$DHA $\leq 30^o$ & $|D - A| \leq 3.50$ Å. According to the previous geometric criterion a hydrogen bond exist if these two conditions are satisfied simultaneously. Then, we computed the average number over all the trajectory. The average number of Hydrogen Bonds over all the trajectory is in the Tables (4.3).

Measured Quantities	**<HB>**
<ChainA-ChainA>	61.777 ± 0.2552
<ChainB-ChainB>	57.93 ± 1.7470
<ChainA-ChainB>	7 ± 0.2602
<P-W>/W	0.0126 ± 0.0000
<W-W>/W	1.79 ± 0.0002
<ChainA-W>/W	0.0061 ± 0.0000
<ChainB-W>/W	0.0064 ± 0.0001

Table 4.3: The average number of hydrogen bonds over all the trajectory of the simulation of Rop protein in water.

Errors are calculated based on standard block averaging . In more detail, the run has been divided in 3 blocks of equal distant time. Then, we have calculated the average value in each block, and after that we have calculated the error of these three values. The average number is calculated from the average values of the three blocks.

As we can notice from the values in Table (4.3), we can see that the number of intrachain hydrogen bonds within the same chain is much higher than the number of interchain HB between the ChainA and ChainB. So, HB plays a key role in maintained the secondary structure of the proteins, whereas it is not the driving force that retains the two chains in their specific cm-cm distance (native state). In the later case electrostatic interactions constitute dominant factor as we eill describe through an MD test presenting in the following subsection.

According to the reference [28] [21], the average number of hydrogen bonds of water per water molecule was computed following, that each water molecule participates in, which technically, is twice the average number of hydrogen bonds per molecule. In order to be able to compare our average value of hydrogen bonds (in Table 4.3) to the average value in the bulk system, we doubled it. So, the doubled value is 3.58, which is too close to the bulk [28].

According to the reference [1], 31ALA is the only amino acid that forms HBs to both helices simultaneously. We performed an analysis of the average number of hydrogen bonds, which produced during the simulation, for each of the four residues of the loop (29LEU, 30ASP, 31ALA, 32ASP). In Tables (4.4) and (4.5), we summarized the average number of HBs over all the trajectory configurations. The Table (4.4) is for the Chain A, whereas the Table (4.5) is for Chain B of the Rop protein. As we already mentioned in the introduction, each chain has two α-helices, so we are going to call 1-αhelix the first one and 2-αhelix the second one.

Measured Quantities	**<HB>**
<29LEU-1-αhelix>	0.96
<29LEU-2-αhelix>	0
<30ASP-1-αhelix>	0.54
<30ASP-2-αhelix>	0
<31ALA-1-αhelix>	0.69
<31ALA-2-αhelix>	0.95
<32ASP-1-αhelix>	0
<32ASP-2-αhelix>	1.42

Table 4.4: The average number of hydrogen bonds of the four amino acids of the loop region with each of the α-helices of Chain A.

Measured Quantities	**<HB>**
<29LEU-1-αhelix>	0.95
<29LEU-2-αhelix>	0
<30ASP-1-αhelix>	0.16
<30ASP-2-αhelix>	0
<31ALA-1-αhelix>	0.16
<31ALA-2-αhelix>	0.78
<32ASP-1-αhelix>	0
<32ASP-2-αhelix>	1.55

Table 4.5: The average number of hydrogen bonds of the four amino acids of the loop region with each of the α-helices of Chain B.

So, we can conclude from the above results (Tables:4.4 and 4.5), that 31ALA is the only amino acid of the loop creating HB with both of the helices of each chain.

Rop withouts charges

As we have already mentioned about hydrogen bonds, they play a key role in the maintance of the secondary structure of protein, but they are not responsible for retaining the native state stable. For that reason, we perform a test simulation, where we remove all the charges from our topology file, in order to check the role of the electrostatic interactions. In Figure (4.6), we present the distance between cm-cm of the two chains and as we ca see the distance decreasing immediatly.

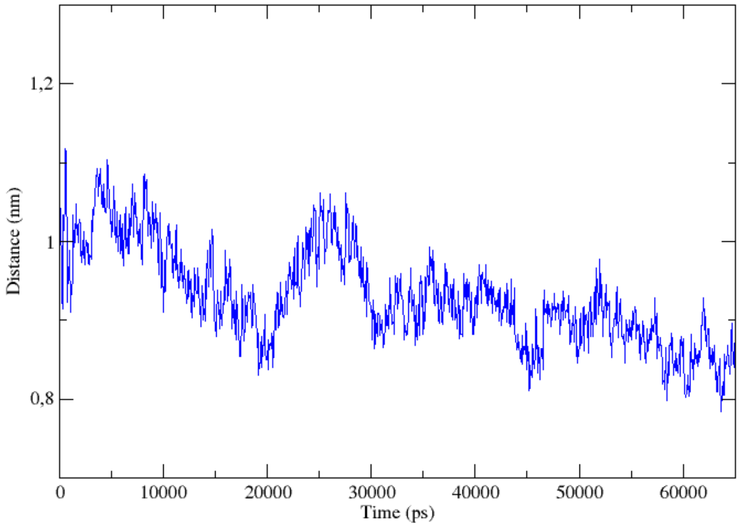
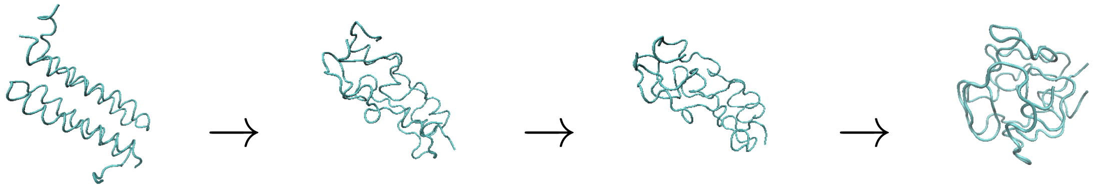


Figure 4.6

Moreover, in Figure (4.7) we present the time evolution of conformation of the Rop protein through some snapshots. As we can see the conformation of Rop change in 62.5 ps too fast until it reaches in the last configuration where the change is obvious. In the last snapshot (4.7d), we can not distinguish the two chains, the protein looks like ”a sphere”.



(a) Snapshot at t=0 ps.

(b) Snapshot at t=62.5 ps.

(c) Snapshot at t=687.5 ps.

(d) Snapshot at t=65.5 ns.

Figure 4.7: The change evolution of the conformation of Rop protein without charges.

4.1.4 Ramachandran Plot

Based on the Ala theroy, the Ramachandran plot produced by PROCHECK [58] to validate the backbone structure of the Rop protein. PROCHECK provides a detailed analysis of the stereochemical quality of the 3D protein structure. In Figure (4.8), we present the Ramachandran plot of Rop protein. As we can notice, there are triangles and squares in the diagram. All the residues of Rop identified by squares except Glycine residues, which we shown separately by triangles, because as it is explained in Section : (3.3.4) Gly constitutes an exception. Each black square represents the conformation of the main chain of one residue of the protein, as well as the same for each black triangle represents each Gly residue of the protein. Moreover, in Figure (4.8), we show with the lines the triangles which belong to Gly residue. We represent the different regions of the plot by shadings. By shading, we actually mean the different areas with different colour. The darker they are, the more favourable the $\phi - \psi$ combination.

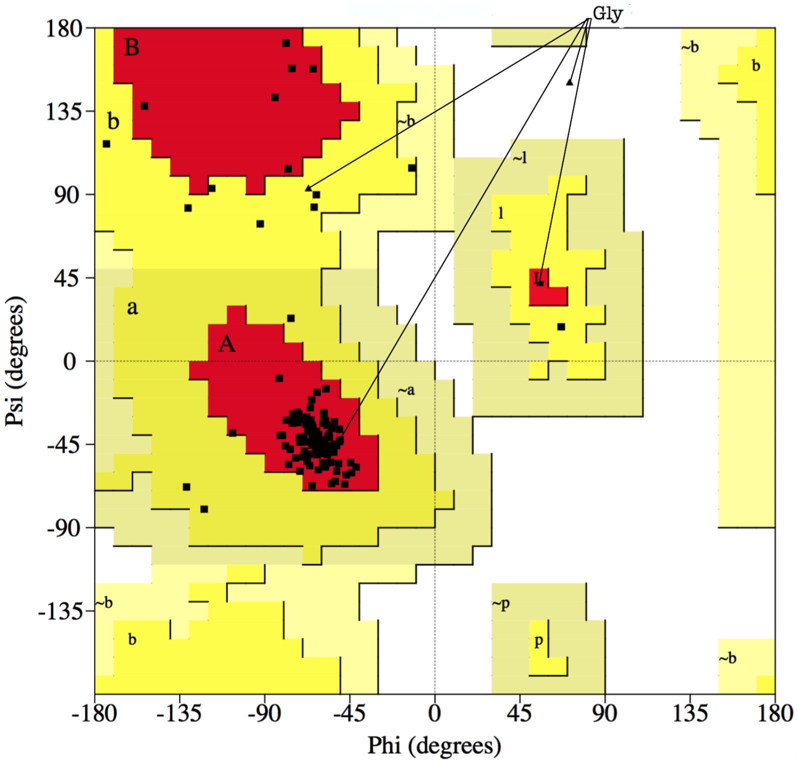


Figure 4.8: Ramachandran plot of Rop. This plot produced by PROCHECK.

The ’core’ regions, which are the darkest and are represented with the red color, are marked them A (corresponds to the α-helix) and B (referred to residues in a β strand conformation of the main chain) and L (it is the smallest red area and corresponds to the left-handed α-helix conformation). The 90% of the residues of the protein structure should be ideally there. In our case, we have in the most favored region 90.7% of them, that is 107 amino acids. Moreover, we marked with a,b,l, and p the additional allowed regions (yellow). More precisely, we marked with a the ”allowed alpha regions”, with b ”the allowed beta”, with l the ”allowed left-handed alpha” and with p the ”allowed epsilon”. In these regions, there are 9.3% of the amino acids, which are 11 residues. Then, with the lightest yellow are the in generously allowed regions distinguished with ∼a (generous alpha), ∼b (generous beta), ∼l (generous left-handed aplha), and ∼p (generous epsilon). The disallowed areas are with white color. In Rop protein, we have four α-helices so, we expect to see the majority of the amino acids to be in the area A. As we can see from the diagram in Figure 4.8, the majority of them are clustered in this area. It has also 4 Gly residues, and as we can notice from the diagram combinations of the $(\phi - \psi)$ angles exist in the red, yellow and in the white region. This happens because Gly has no side chain and is much less sterically prevented than other amino acids. So, Gly is able to adopt pairs of ϕ, ψ angles that are not allowed for any other residue. We should note here that this Ramachandran plot is for Ala-like residues and Gly has its one Ramachandran plot with less regions disallowed. Finally, we do not plot the first and last amino acids in the Ramachandran plot, because for the first amino acid we need a residue before it and for the last a residue after it, in order to be able to find the $(\phi - \psi)$

angles.

In Figure (4.9), we present separate Ramachandran plots for each one of the 20 amino acids, which has Rop protein. Above each plot, there is the name of the corresponding amino acid. There is a number in brackets, following the name of each residue, shows the total amount of data points on this graph. The darker the shaded area on each plot represents the more favourable region. We may see a letter (A or B) following a number above some points. The number and the letter refer to residue-number and the chain of the protein accordingly. This notation stands for the amino acids which are in unfavourable regions of the plot.

We can also observe that there are squares that have a different color. Those with red means it is located in unfavorable regions, whereas the squares with yellow are them in favorable regions. Moreover, there are square points with different accents in orange color. The darker it becomes the orange so farther are the squares from the favorable areas.

Let us now observe each plot, we notice, that the majority of points are clustered in the area of a-helix, but there are also diagrams of particular interest.

As we have already mentioned, Rop protein has two chains, which have two loops, and in the loop region, there is a residue of Ala. So, these two points may be these two residues, which are in the unfavoured area. One more diagram of particular interest is the one of Asp residue since it has points in different areas. Finally, Gly residue is a special residue and as we said it has its own Ramachandran plot, because of its sidechain which is only a hydrogen atom it can adopt more combinations of $(\phi - \psi)$ angles. So, if we notice the plot of Gly in contrast with the others, we will see that it has more shaded regions than the others.

In Figure (4.10), we present the plots of the torsion angles χ_1 vs χ_2, which belong to the sidechain angles, for all residue types whose sidechains are long enough to have both these angles. The plots of $(\chi_1 - \chi_2)$ angles are produced by PROCHECK.

If we notice, there is no plot for Cys residues, whereas we have 4 of them in our protein. This happens because the sidechain of Cys is not long enough in order to calculate the χ_2 angle. Additionally, we do not have this plot for Ala and Gly residues. As we have already discussed in Section (3.4), these two amino acids are excluded from the calculations of χ_1 and χ_2 angles, for the reason that Gly has no C_β atoms in its sidechain and in the case of Ala its C_β atom does not bond with another heavy atom except the Hydrogens.

In Figure (4.10), we present the plots of the torsion angles χ_1 versus χ_2, which belong to the sidechain angles, for all residue types whose sidechains are long enough to have both these angles.The ideal regions are represented by the points where dashed lines are crossed: the gauche minus, the trans, as well as the gauche plus for the $\chi_1 - \chi_2$ dihedral angles. The shading is based on data has come from a data set of 163 non-homologous, high-resolution protein chains, that are chosen from structures solved by X-ray crystallography. The average value of these data we are going to call them “ideal values”. The center of the crosses is at the mean and the length represents the ±1 standard deviation of these ideal values. The points marked with a letter (A, B) and a number. The number implies the residue-number,

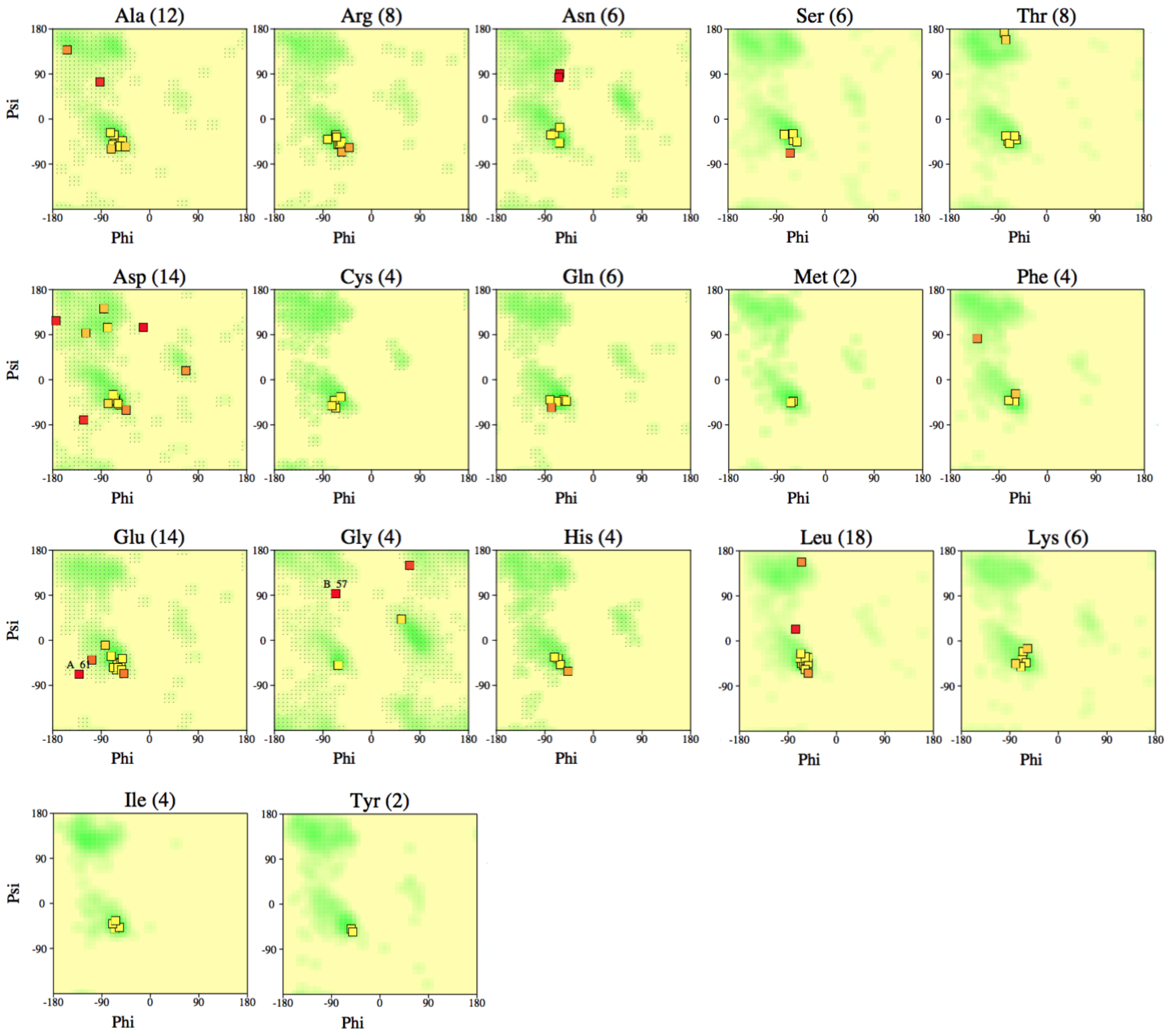


Figure 4.9: The above plots show a Ramachandran plot (ϕ vs ψ) for each residue.

and the letter refers to the protein chain. Again, the number in brackets, following the name of amino acids, shows the total number of points on this graph.

From the plots of ($\chi_1 - \chi_2$) angles in Figure (4.10), we notice that the most favored is the (t,g-) region. After this region, the next one where the points clustered is the (g+, t) region. We also notice that at the regions where $\chi_1 = +60$ (gauche -) there are no points.

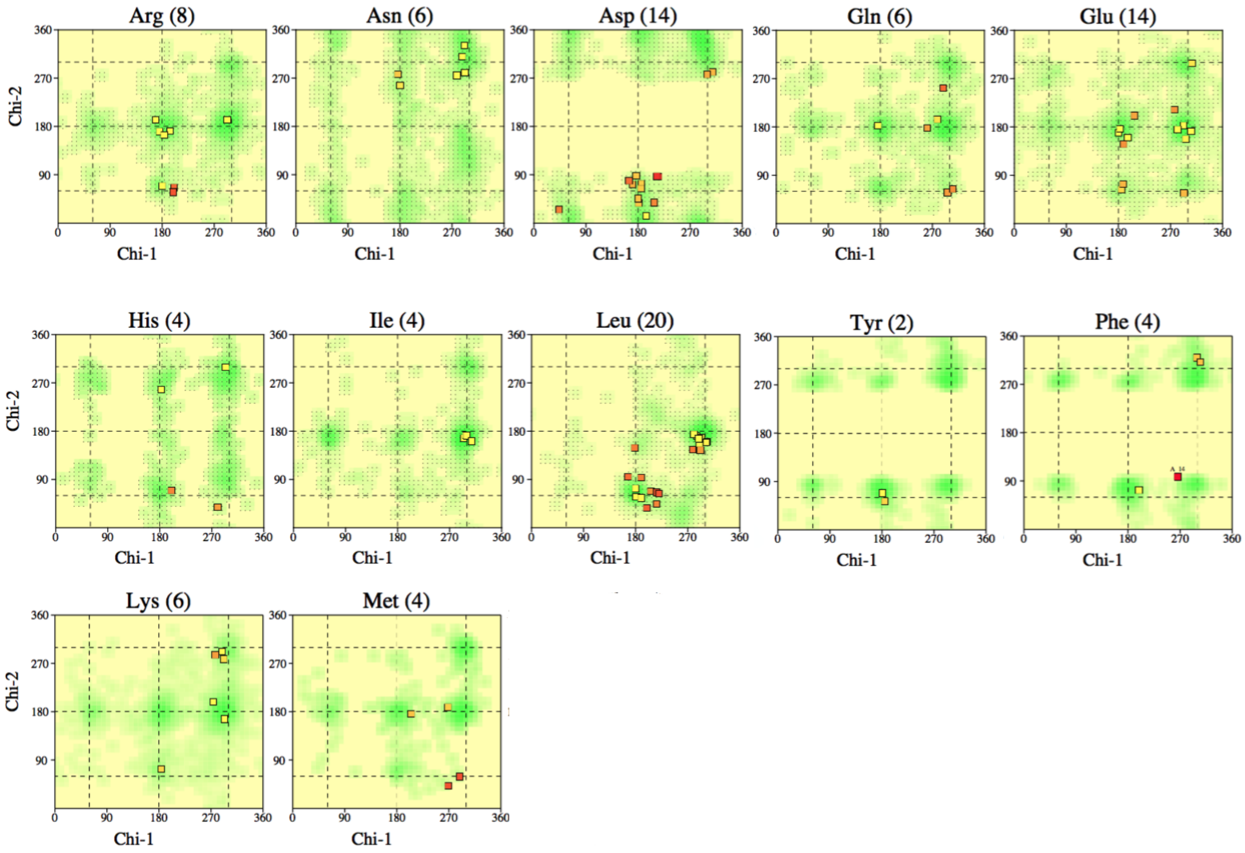
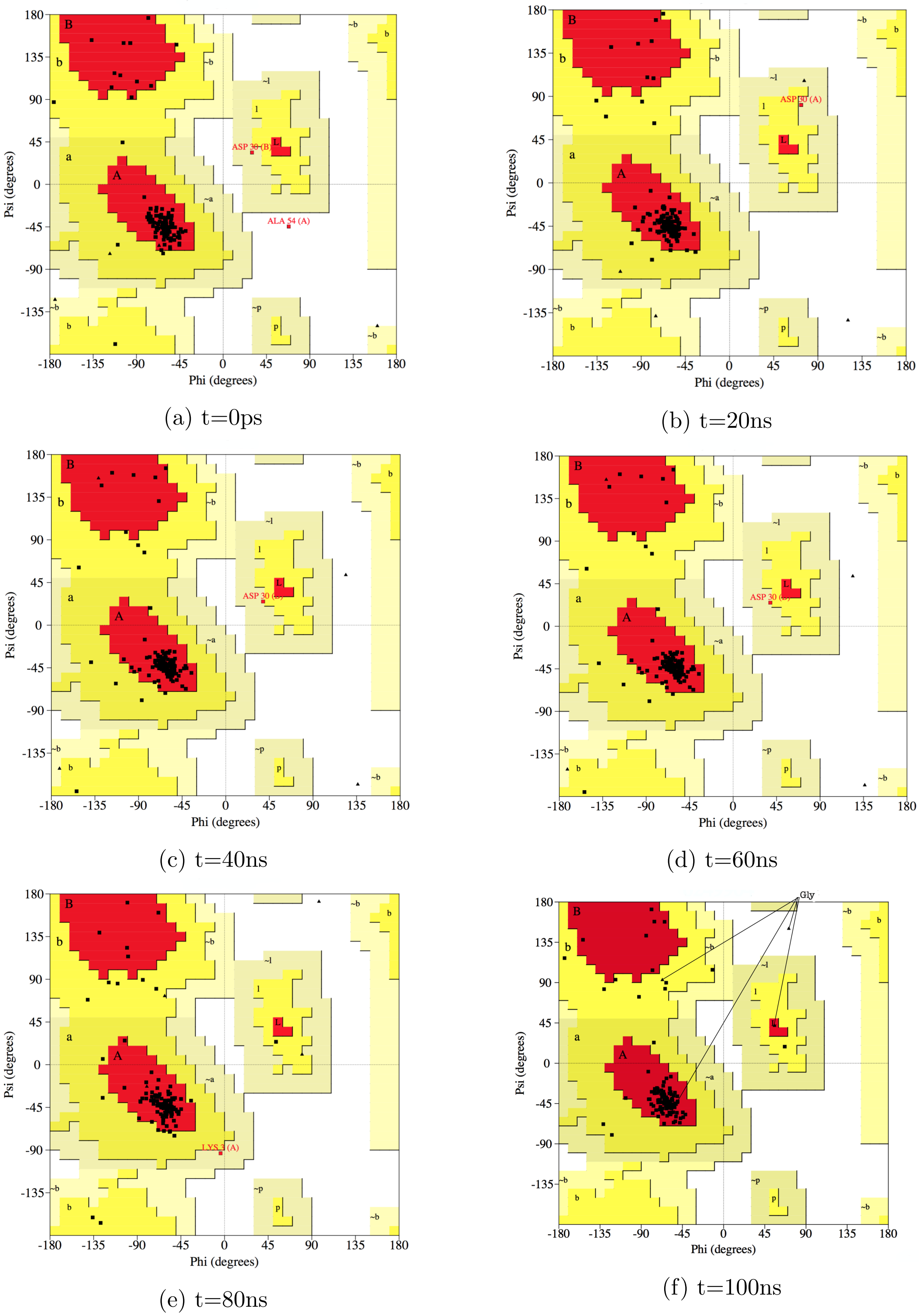

Arg (8)
Asn (6)
Asp (14)
Gln (6)
Glu (14)
His (4)
Ile (4)
Leu (20)
Tyr (2)
Phe (4)
Lys (6)
Met (4)

Figure 4.10: Plot of (χ_1 vs χ_2)for each residue of Rop protein.

In order to study, the time evolution of the Ramachandran plot, we chose from our trajectory six equal distant configurations. The total simulation time was 100 ns. Then, we performed the analysis of Ramachandran plot in each of them. In Figure (4.11), we present the results of this analysis. As we can see, in all plots there are one or two amino acids marked in red lettering. These are in the generously allowed and disallowed regions. Only, the last frame has not labeled residues. Therefore, we should note that in all time frames, the majority of amino acids clustered in A region.



(a) t=0ps

(b) t=20ns

(c) t=40ns

(d) t=60ns

(e) t=80ns

(f) t=100ns

Figure 4.11: The time evolution of Ramachandran Plot of Rop protein.

4.2 RM6 in Native State

The second system that we are going to present is RM6 protein in aqueous solution. RM6 is a mutant of Rop protein in which we removed five amino acids from the loop region. We present two different snapshots, one of the initial configuration of our system and one of the last after we performed a simulation run of 209 ns. If we notice the two different snapshots, we see no difference in the structure of our protein, due to the stability of the native state 4.12.

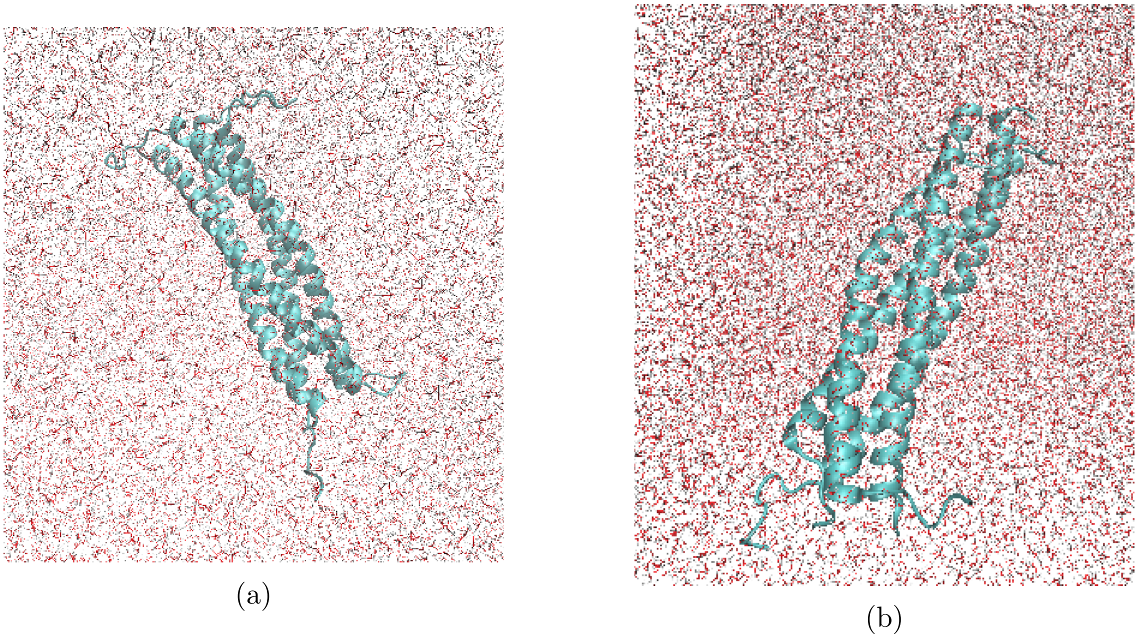


Figure 4.12: Snapshots of RM6 system in water solution, (4.12a) initial configuration and (4.12b) final configuration.

4.2.1 RMSD

The root mean square deviation (RMSD) was calculated as an indicator of accuracy of our model or as an indicator that our model reach in equilibrium. As we said, the RMSD is a measure comparing two structures. Each conformation from a trajectory file is compared to a reference structure. In our case, the reference structure is the initial configuration of the simulation. All the measurements of the RMSD are done based on the C_α atoms.

The RMSD of all C_α backbone atoms with respect to the reference structure as a function of time is shown in Figure 4.13. As we notice from the diagram (4.13), we see a great deviation from the starting structure. Such difference was expected, since the seven last residues in each chain, as well as the first four in two of them were added using the tool Pymol.

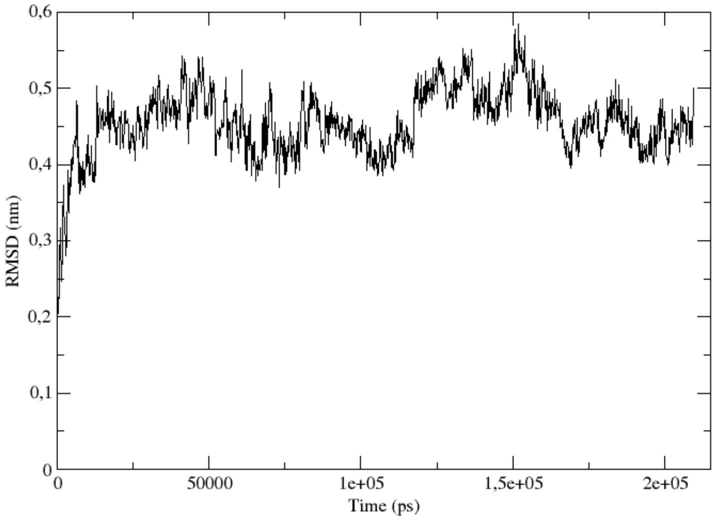


Figure 4.13: Root Mean Square Deviation (RMSD) plot showing the RMSD based on C_α atoms as a function of time for the simulation of 209ns for RM6 protein in water.

4.2.2 Radius of Gyration (Rg)

The radius of gyration (Rg) is a measure of the compactness of the protein conformation in biomolecular simulations and quantifies their size. As we can notice in Figure (4.14), the value of Rg starts being relatively stable after 50ns, which means that the RM6 is stably folded at the equilibrium conformation. In the beginning, the value of Rg was higher this may happen, for the reason that, we added seven last residues to all chains, and it needed some time until finding an equilibrium conformation.

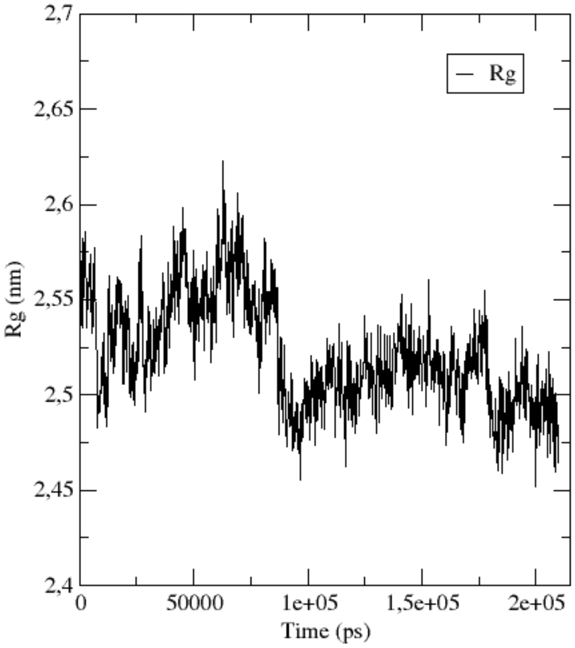


Figure 4.14: Radius of gyrarion as a function of time for RM6 in water.

As we have already mentioned in the Introduction of RM6 protein, it is a tetramer protein, which consisting of two molecules each of them has two chains. These two molecules we are going to call them MoleculeA and MoleculeB. The RM6 has 4 chains, which we are going to name them as ChainA, ChainB, ChainC, ChainD. The chains (ChainA and ChainB) belongs to the MoleculeA, and the other two chains (ChainC and ChainD) corresponds to the MoleculeB. By Protein we mean the whole molecule of RM6. After 50ns, the value of Rg is relatively stable over time, which means that the RM6 is stably folded at th equilibrium conformation.

The Rg in nm was computed by taking the average over all the trajectory and is presented in Table (4.6).

Measured Quantities	**Rg (nm)**
MoleculeA	2.510 ± 0.018
MoleculeB	2.506 ± 0.034
ChainA	2.445 ± 0.032
ChainB	2.528 ± 0.017
ChainC	2.500 ± 0.030
ChainD	2.464 ± 0.040
Protein	2.520 ± 0.026

Table 4.6: The average value of Rg over all trajectory frames of the simulation of RM6 protein.

4.2.3 Hydrogen Bonds

Next, we analyze hydrogen bonds using gromacs tool (gmx hbond). For each pair of atoms, which form hydrogen bond, the bond is counted once. We split the analysis of hydrogen bonds into intermolecular and intramolecular hydrogen bonds, as we discussed in Section: 3.2. More specifically in our case about RM6, we are going to further categorize HB as follows. We split the intramolecular hydorgen bonds into four different categories:

1. intechain hydorgen bonds between the two chains (ChainA-ChainB, ChainA-ChainC, ChainA-ChainD, etc.)
2. intrachain hydrogen bonds within the same cahin (ChainA-ChainA, ChainB-ChainB, ChainC-ChainC, ChainD-ChainD)
3. intramolecularmol within the same molecule (MoleculaA-MoleculeA, MoleculeB-MolceculeB)
4. intremolecularmol between the two molecules (MoleculaA-MoleculeB).

The intramolecular hydrogen bonds are within the same molecule, in our case, the Protein is the RM6 protein, which has four chains, let us denote them ChainA, ChainB, ChainC, ChainD. Additionaly, it has two molecules and we are going to name them MoleculeA, which has the chains: ChainA and ChainB, and MoleculeB, with the chains ChainC and Chain D. Additionally, we have the intermolecular hydrogen bonds involve two or more molecules, such as Protein-Water, ChainA-Water, ChainB-Water, ChainC-Water, ChainD-Water, Water-Water. The notation we are going to use is the letter W for water. The geometric criterion we used to calculate the number of hydrogen bonds per each frame is: $\angle$DHA $\leq 30^o$ & $|D-A| \leq 3.50$ Å. According to the previous geometric criterion a hydrogen bond exist if these two conditions are satisfied simultaneously. Then, we computed the average number over all the trajectory. The average number of Hydrogen Bonds over all the trajectory is in the Tables (4.7).

Measured Quantities	**<HB>**
<ChainA-ChainA>	46.523 ± 0.6596
<ChainB-ChainB>	50.620 ± 0.3144
<ChainC-ChainC>	43.520 ± 1.364
<ChainD-ChainD>	51.578 ± 2.1132
<ChainA-ChainB>	8.779 ± 1.317
<ChainC-ChainD>	10.479 ± 0.963
<ChainA-ChainC>	5.694 ± 0.183
<ChainA-ChainD>	1.250 ± 0.001
<ChainB-ChainC>	4.943 ± 0.100
<MoleculeA-MoleculeB>	12.495 ± 0.549
<Protein-W>/W	0.0065 ± 0.00001
<W-W>/W	2.1862 ± 0.00005
<MoleculeA-W>/W	0.0032 ± 0.0000
<MoleculeB-W>/W	0.0033 ± 0.0000
<ChainA-W>/W	0.0016 ± 0.0000
<ChainB-W>/W	0.0016 ± 0.0000
<ChainC-W>/W	0.0017 ± 0.0000
<ChainD-W>/W	0.0015 ± 0.0000

Table 4.7: The average number of HB over all the trajectory frames of the simulation of RM6 protein in aqueous solution.

We notice that the average number of intrachain HB of all the chains are high, which means that the HB maintain stable the conformation of each chain. But again, as in Rop protein, we observe that, between the two molecules MoleculeA and MoleculeB there are no many HB. So, the hydrogens bonds are not responsible for the stable conformation of the native state.

Additionaly, we observe that the chains do not hydrogen bonging with water molecules.

According to the reference [28], [21], the average number of hydrogen bonds of water per water molecule was computed following, that each water molecule participates in, which technically, is twice the average number of hydrogen bonds per molecule. In order to be able to compare our average value of hydrogen bonds (in Table 4.7) to the average value in the bulk system, we doubled it. So, the doubled value is 4.37, which is too close to the bulk [28].

4.2.4 Ramachandran Plot

The Ramachandran plot produced by PROCHECK [58] to validate the backbone structure of the RM6 protein. PROCHECK provides a detailed analysis of the stereochemistry quality of the 3D protein structure. In Figure 4.15, we present the Ramachandran plot of mutation

RM6. As we can notice, there are both triangles and squares in this diagram. All the residues of RM6 identified by squares except Glycine residues, which we shown by triangles. Each black square represents the conformation of the main chain at one residue of the protein, as well as the same for each black triangle represents each Gly residue of the protein. Moreover, in Figure (4.15), we show with the lines the triangles which belong to Gly residue. We represent the different regions of the plot by shading. By shading, we actually mean the different areas with different colour. The darker the are, the more favourable the $\phi - \psi$ combination.

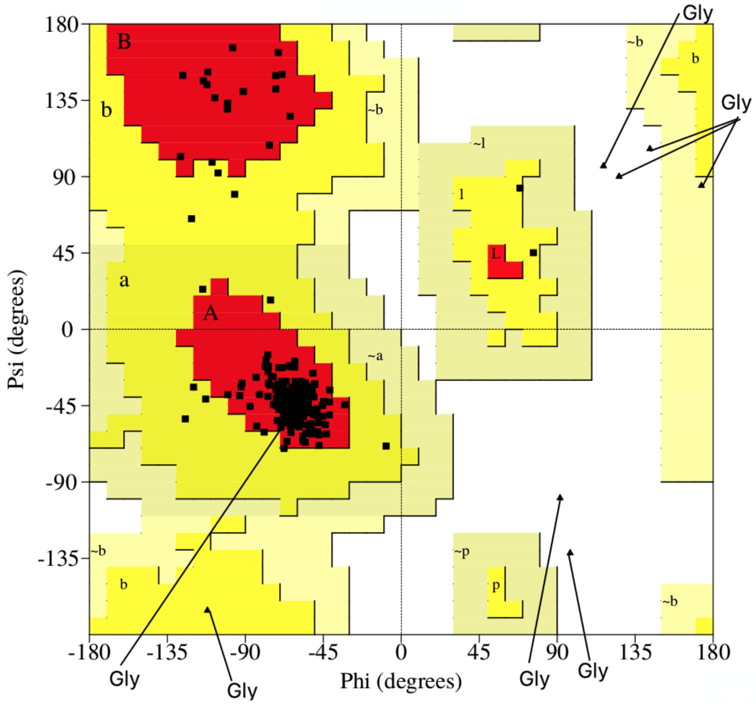


Figure 4.15: Ramachandran plot of RM6.

As we can see from the diagram in Figure (4.15), the majority of them clustered in this area.

The 'core' regions, which are the darkest and are represented with the red color, are marked them A (corresponds to the α-helix) and B (referred to residues in a β strand conformation of the main chain) and L (it is the smallest red area and corresponds to the left-handed α-helix conformation). The 90% of the residues of the protein structure should be ideally there. In our case, we have in the most favored region 93.5% of them, that is 202 amino acids. Moreover, we marked with a,b,l, and p the additional allowed regions (yellow). More precisely, we marked with a the "allowed alpha regions", with b "the allowed beta", with l the "allowed left-handed alpha" and with p the "allowed epsilon". In these regions, there are 6.5% of the amino acids, which are 14 residues. Then, with the lightest yellow are the in generously allowed regions distinguished with ∼a (generous alpha), ∼b (generous beta), ∼l (generous left-handed aplha), and ∼p (generous epsilon). The disallowed areas are with white color.

It has also 8 Gly residues, and as we can notice from the diagram combinations of the $(\phi-\psi)$ angles exist in the red, yellow and in the white region. This happens because Gly has no side chain and is much less sterically prevented than other amino acids. So, Gly is able to adopt pairs of ϕ, ψ angles that are not allowed for any other residue. We should note here that this Ramachandran plot is for Ala-like residues and Gly has its one Ramachandran plot with less regions disallowed.

The plots in Figure (4.16) shows separate Ramachandran plots for each of 20 different amino acids. There is a number in brackets, following the name of each residue, shows the total amount of data points on this graph. The darker the shaded area on each plot represents the more favourable region. We may see a letter (A or B or C or D) following a number above some points. The number and the letter refer to residue-number and chain of the protein accordingly. These marked amino acids are in unfavourable regions of the plot.

We can also observe that there are squares that have a different color. Those with red means it is located in unfavorable regions, whereas the squares with yellow are them in favorable regions. Moreover, there are square points with different accents in orange color. The darker it becomes the orange so farther are the squares from the favorable areas.

If we observe each, we see, that the majority of points are clustered in the area of a-helix, but there are also diagrams of particular interest.

Gly residue, as we have already said above, is a special residue and as we said it has its own Ramachandran plot, because of its sidechain which is only a hydrogen atom it can adopt more combinations of $(\phi-\psi)$ angles. So, if we notice the plot of Gly in contrast with the others, we will see that it has more shaded regions than the others. Additionally, the majority of the points of Gly seems to be in unfavorable regions.

In Figure (4.17), we present the plots of the torsion angles χ_1 versus χ_2, which belong to the sidechain angles, for all residue types whose sidechains are long enough to have both these angles. The three ideal regions are represented by the points where dashed lines are crossed: the gauche minus, the trans and the gauche plus for the $\chi_1-\chi_2$ dihedral angles. The shading is based on data has come from a data set of 163 non-homologous, high-resolution protein chains, that are chosen from structures solved by X-ray crystallography. The average value of these data we are going to call them "ideal values". The center of the crosses is at the mean and the length represents the ±1 standard deviation of these ideal values. We observe that there are points labeled, these are points placed in more than 2.5 deviation away from the ideal values. The points marked with a letter (A or B or C or D) and a number. The number implies the residue-number, and the letter refers to the protein chain. Again, the number in brackets, following the name of amino acids, shows the total number of points on this graph. The darker the shaded area on each plot represent the more favorable regions.

In terms of Asp residue, we notice that the majority of points are clustered at the (t,g-) region. Additionally, Asp residue has a point at (g-,g-) region. As we can see, the combinations of angles $(\chi_1-\chi_2)$ for Glu residue prefer the regions (t,t) and (g+,t). In addition, there are one or two points at the following regions: (g+,g+), (g+,g-) and (t,g+) All the points of Ile

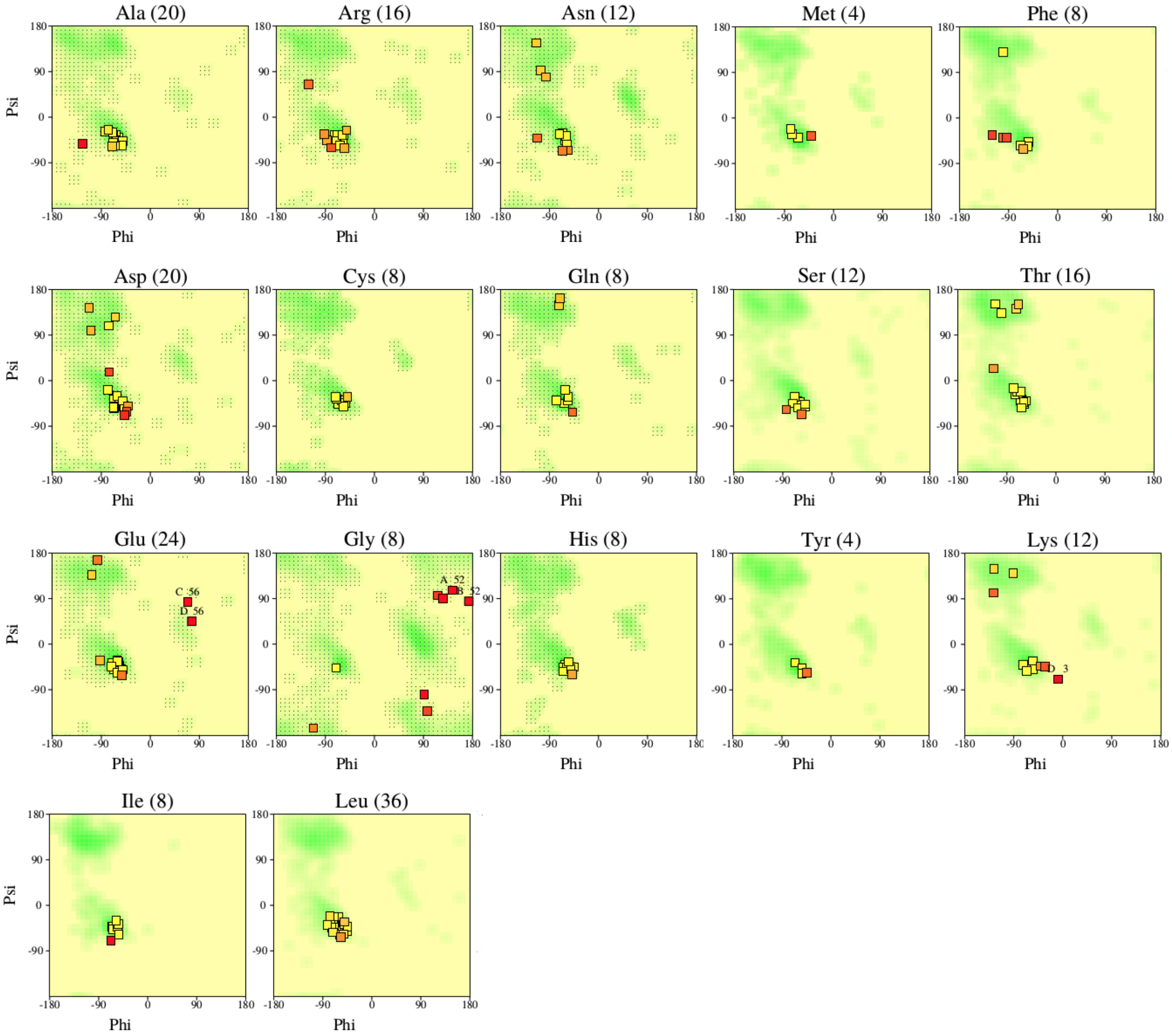


Figure 4.16: The above plots show a Ramachandran plot (ϕ versus ψ) for each residue.

residue are in the region (g+,t). Leu residues have two favoured regions, where the points are clustered, these are (t,g+) and (g+,t).

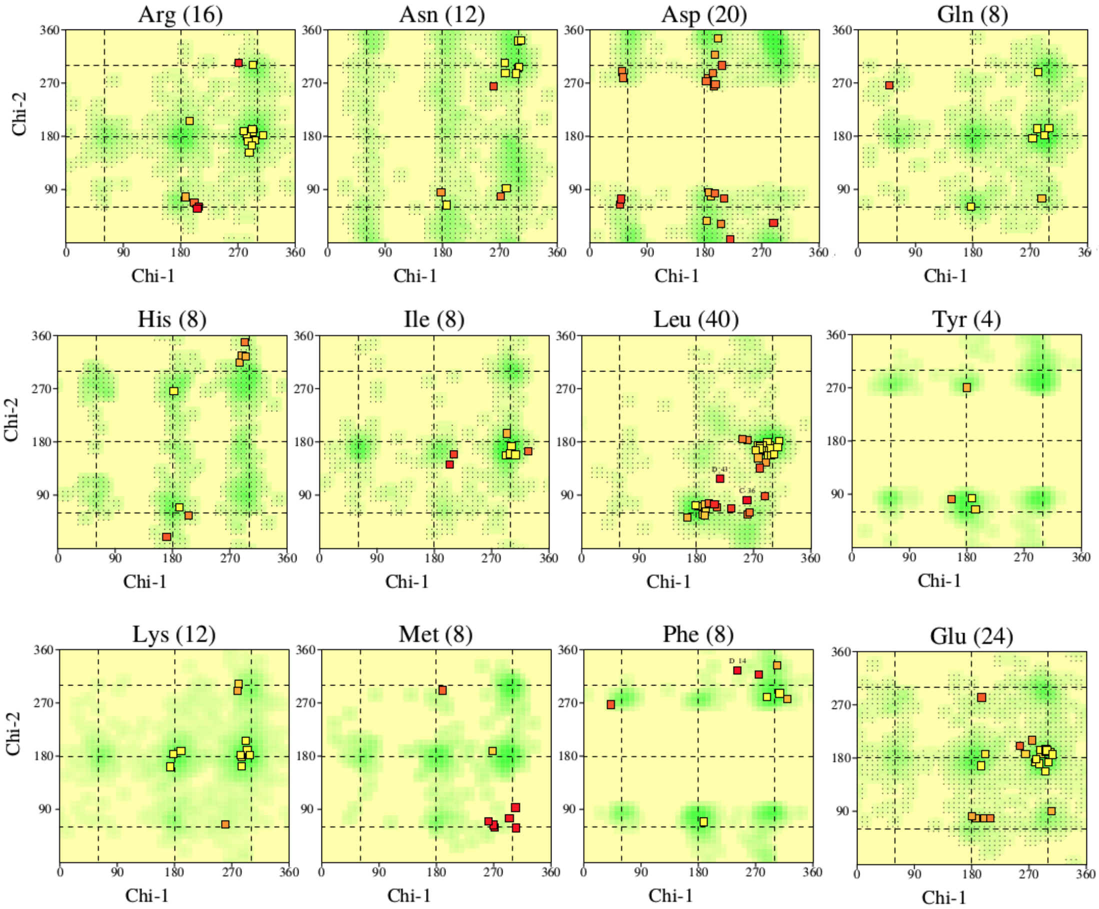
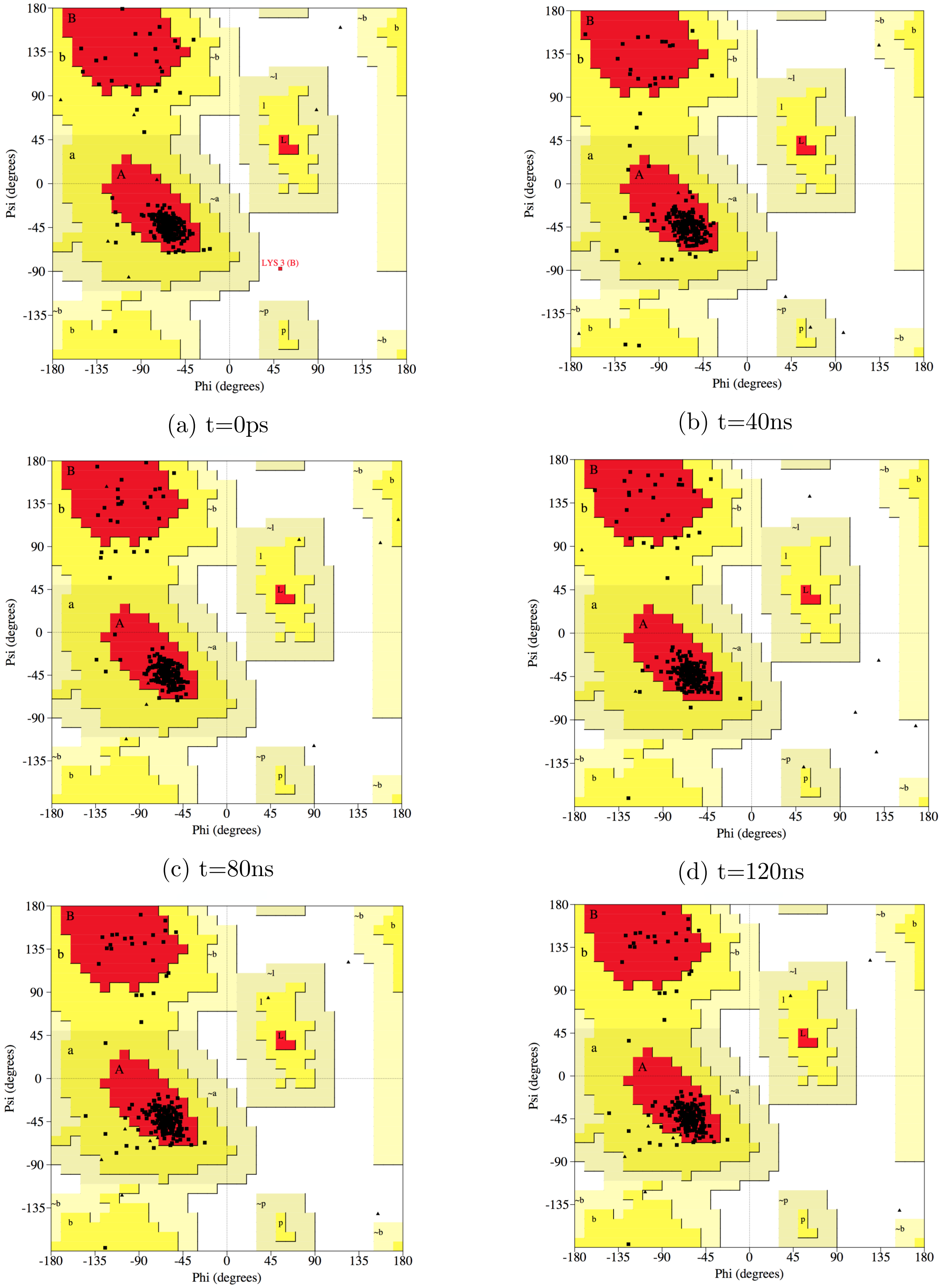


Figure 4.17: The above plots show a plot χ_1 vs χ_2 for each residue of RM6 protein, where it is applicable.

In order to study, the time evolution of the Ramachandran plot, we chose from our trajectory six equal distant configurations. The total simulation time was 209 ns. Then, we performed the analysis of Ramachandran plot in each of them. In Figure (4.18), we present the results of this analysis. As we can notice, in the first plot, which is our initial configuration, there is a combination of $(\phi - \psi)$ angles that is in the disallowed region and marked in red lettering. As simulation time passes, and as we observe the following Ramachandran plots, we see a good enough stereochemical quality of our model. It is noticeable, that in all time panels the majority of residues clustered in A region (α-helix region).



(a) t=0ps

(b) t=40ns

(c) t=80ns

(d) t=120ns

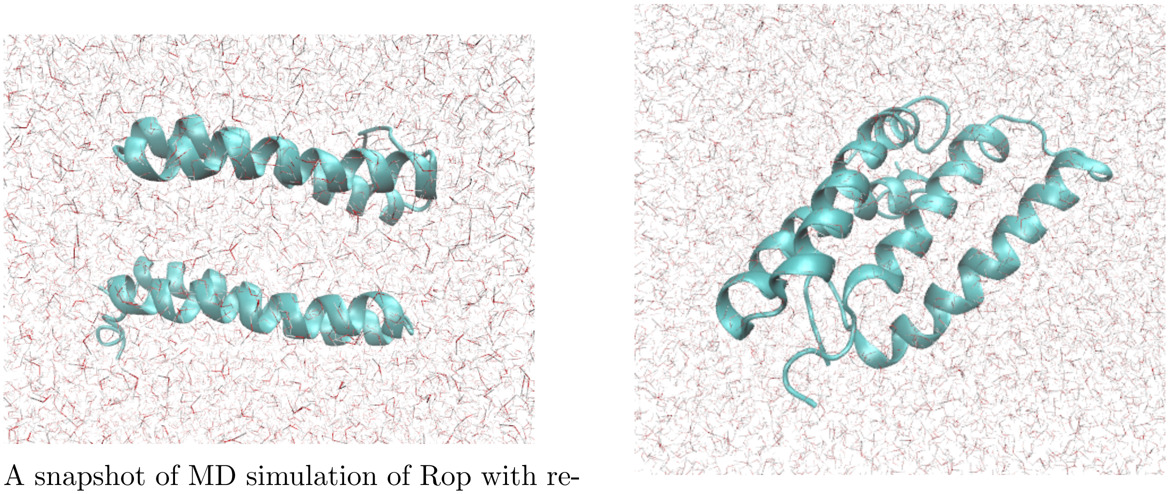
(e) t=160ns

(f) t=200ns

Figure 4.18: The time evolution of Ramachandran plot of RM6 protein.

4.3 Remote Chains of Rop

This system has been used as a test-case for validation of the force field and the model that we used. With the intention of this check, we did the following process. First of all, we found the initial distance between the centers of mass of the two chains in the native state of the Rop protein was 1.2 nm. Then, we kept the coordinates of the first one constant, and we moved away the second one in a cm-cm distance equal to 2.027 nm. We started a simulation of this system in aqueous solution. A snapshot of the initial conf is depicted in Figure (4.19).



(a) A snapshot of MD simulation of Rop with remote chains in water at t=0ps.

(b) A snapshot of MD simulation of Rop with remote chains in water at t=271ns.

Figure 4.19: Two snapshots of the RRC system with: (4.19a) is the two chains in water solution at t=0ps and (4.19b) is the two chains in water at t=271ns.

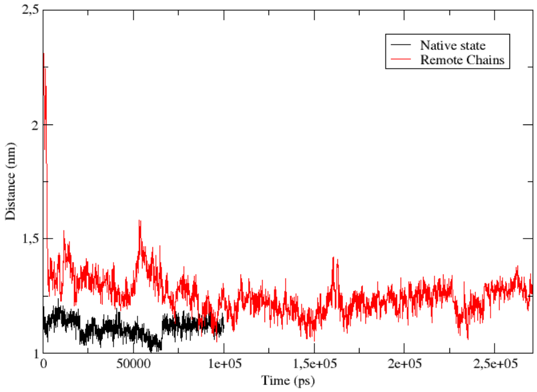
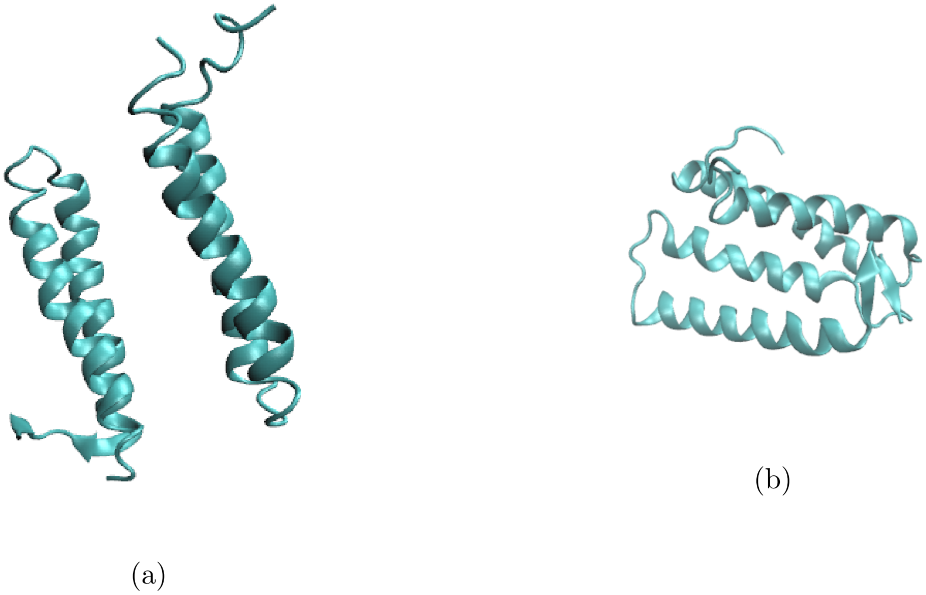


Figure 4.20: Distance between centers of mass as a function of time.

In Figure (4.20), the black line is the distance between the centers of mass in the native state and the red line is the distance between the com in remote chains. The x-axis is the time in ps and y-axis is the distance in nm. In Figure, we present the plot of the distance between cm-cm of the two chains. As we can see, there are two lines, a blue one, and a red. The blue belongs to the Rop protein in the native state, whereas the red one is of the system with the moved chains. Moreover, the x-axis refers to the time of our simulation, whereas the y-axis is the distance between the cm-cm of the two chains. We should also note, that the simulation time of the remoted chains is longer than the time in the native state. The time of the performed simulation of the RRC system (red line) was 271ns in contrast to the Rop in native state (black line) which was 100ns. As we can notice from Figure (4.20), the two remote chains tend to come close enough in a very short period of time. Specifically, at t=2.500 ps the distance between the two centers of mass of the chains is 1.505nm, a slower decrease of the cm-cm distance follows and beyond 100ns values approach the distance of the native state ∼1.1nm still being somehow higher after 271ns. There are strong fluactuations in distance values and the procedure evolves slowly.

In Figure (4.21), we present two snapshots of the RRC system, but only the protein. The Figure (4.21a) is the initial structure of our protein at t=0ns and the Figure (4.21b) is the last configuration at t=271ns. As we can see, there is a clear difference between the cm-cm distance of chains.



(a)

(b)

Figure 4.21: Two snapshots of the RRC system with: (4.21a) is the two chains at t=0ps and (4.21b) is the two chains at t=271ns.

In Figure (4.22), we present the root mean square deviation (RMSD) of our system. If we observe the plot (Figure 4.22) thoughtfully, we are going to see, that the reference structure starts differing with the structure of each consequent frame during the simulation time. This is because the two chains tend to come closer to each other trying to find the native state (equilibrium conformation), which is very different from the initial one of the current simulation.

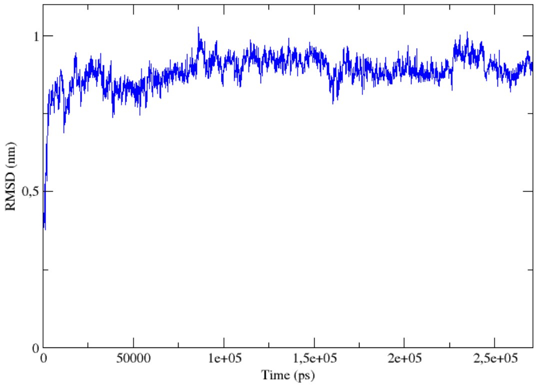


Figure 4.22: Rmsd of system RRC.

4.4 Comparison Between the Rop and RM6

In the following we are going to do a comparison between Rop protein and its mutation RM6. The differences between the two proteins are highloghted.

Radius of gyration

As we can observe from the diagram in Figure 4.23 the value of Rg of RM6 protein is bigger than the Rg of Rop protein. This is expected, due to in Rop protein we have only two chains and the two α-helices connected with a loop, whereas in RM6 which is a tetramer we have 4 chains in which we remove 5 residues from the loop and the two helices are in a linear line.

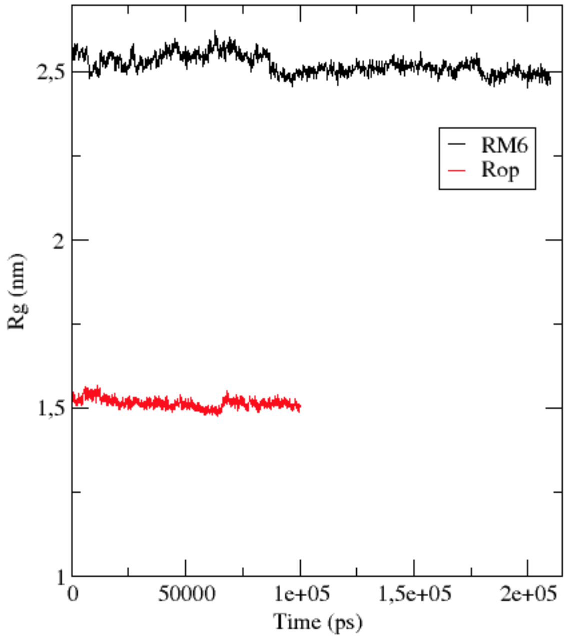


Figure 4.23: Radius of gyrarion (Rg) plot.

Measured Quantities	Rop	RM6
ChainA	1.404 ± 0.004 (nm)	2.445 ± 0.032 (nm)
ChainB	1.410 ± 0.003 (nm)	2.528 ± 0.017 (nm)
ChainC	-	2.500 ± 0.030 (nm)
ChainD	-	2.464 ± 0.040 (nm)
MoleculeA	1.511 ± 0.004 (nm)	2.510 ± 0.018 (nm)
MoleculeB	-	2.506 ± 0.034 (nm)

Table 4.8: Comparison of Rg values between RM6 and Rop.

In Table (4.8), in the system of Rop we denote as MoleculeA the whole protein, because it has two chains. Moreover, we have ”-” in the positions of ChainC and ChainD because Rop has only two chains.

We see that the values of Rg in RM6 is much greater than in Rop. In the case of the chains this is norma because in Rop there is a loop so the chains are more ”compact”, than in RM6 in which they look like a linear line. Additionaly, the Molecules in RM6 are less compact than in Rop.

Hydrogen Bonds

The average number of hydrogen bonds among water molecules is greater in RM6 system than in Rop protein. Therefore Rop structures causes bigger disturbance in the HB network of water compared to RM6 structure.

Additionally, the average number of HB between ChainA-ChainB in RM6 is greater than in Rop protein, as well as the number of ChainC-ChainD. The chains in Rop tend to create more hydrogen bonds with themselves than the chains in RM6.

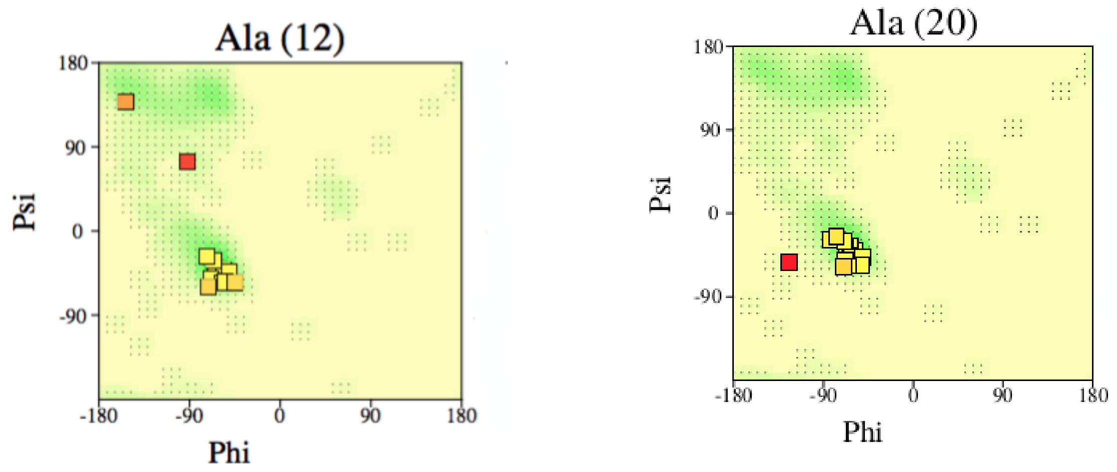
Measured Quantities	**Rop**	**RM6**
<ChainA-ChainA>	61.777 ± 0.2552	46.523 ± 0.6596
<ChainB-ChainB>	57.93 ± 1.7470	50.620 ± 0.3144
<ChainC-ChainC>	-	43.520 ± 1.364
<ChainD-ChainD>	-	51.578 ± 2.1132
<Water-Water>/Water	1.79 ± 0.0002	2.1862 ± 0.00005

Table 4.9: Hydrogen Bonds

In Table (4.9), the ”-” in Rop protein in terms of ChainC and ChainD is because Rop protein has only two chain.

Ramachandran Plot

In Figure (4.24), we compare the Ramachandran plot of Ala residue. In RM6 we have 20 amino acids of Ala because we have four chains, but we have removed the Ala residue from the loop region. The digram (4.24a) belongs to Rop and as we can see there are two points away from the A region, this may caused because we have two residues of Ala in the loop which has no α-helix conformation, compared to RM6 (4.24b) where all the points are clustered in the A region.



(a) Ramachandran plot of residue Ala in Rop.

(b) Ramachandran plot of residue Ala in RM6.

Figure 4.24: Comparison of Ramachandran plot of amino acid Ala between Rop and RM6.

Chapter 5

Conclusions and Future Work

Conclustions

We have studied the Rop protein and its loopless mutatio RM6 in aqueous solution. Between them there is a big difference in the topology. RM6 is a tetramer, whereas Rop is a homodimer. We see that the removal of five amino acids from the loop region drives to a complete different topology.

In terms of the stability of the native state of the two proteins. According to the analysis of rmsd and the Rg, we see that the two structures are stable. We should note that RM6 has greater Rg than Rop.

The results above about hydrogen bonding of the residues in the loop regions shows that indeed Ala is the only residue among them, which hydrogen bonding with both α-helices of the chain simultaneously. According to hydorgen bonding analysis the secondary conformation is stable because of the hydrogen bonds, but in the stability of the native state of the protein the elecrostatic interactions play a fundamental role. The average number of hydorgen bonds among water molecules is greater in RM6 system than in Rop protein.

Electrostatic interactions are very important for the stability of the native state of proteins.

Finally, from the analysis of Ramachandran plot we see that the stereochemical quality of our models is good.

Future Work

Very recent results have been extracted related to the thermostability of RM6 which has been observed experimentally. Our plans for the future is to continue the analysis of Rop protein properties and we focus on the following topics:

1. Srudy of trapped water molecules between the two chains has been done, and shows no trapped molecules but a more systematic amalysis is going to be done in future.
2. A simulation of a mutation of a single amino acid in the loop region (mutate 31Ala with Pro) is in progress, because according to experiments this mutation shows a change in the topology of the protein.
3. Another interesting topic for the future is to study the thermal stability of Rop protein.

In our future plans, we are going to calculate the free energy of Rop protein. Additionally, another interesting for future work is to examine if RM6 protein creates aggergations.

Bibliography

[1] D.W.Banner, M.Kokkinidis, D.Tsernoglou *Structure of the ColE1 Rop Protein at 1.7 ÅResolution*; J.Mol.Biol. 196, 657-675, (1987).

[2] M.Ambrazi *Production of biomaterials based on specific interactions between protein elements of secondary structures*; Phd Thesis, University of Crete, ITE, (2012) (in Greek).

[3] M. Amprazi, D. Kotsifaki, M. Providaki, E.G. Kapetaniou, G. Fellas, I. Kyriazidis, J. Pérez and M. Kokkinidis. *Structural plasticity of 4-α-helical bundles exemplified by the puzzle-like molecular assembly of the Rop protein*; PNAS, 111, 11049-11054, (2014).

[4] H.P. Kresse, M. Czubayko, G. Nyakatura, G. Vriend, C. Sander and H. Bloecker. *Four-helix bundle topology re-engineered: monomeric Rop protein variants with different loop arrangements*; Protein Engineering, 14, 897-901, (2001).

[5] N.M. Glykos, Y.Papanikolau, M. Vlassi, D. Kotsifaki, G.Cesareni and M. Kokkinidis *Loopless Rop: Structure and Dynamics of an Engineered Homotetrameric Variant of the Repressor of Primer Protein* Biochemistry, 45, 10905-10919, (2006).

[6] W. Eberle, A. Pastore, C. Sander and P. Rösch. *The structure of ColE1 rop in solution*; J. Bio NMR 007, 1, 71-82, (1991).

[7] P.G. Bolhuis *Sampling Kinetic Protein Folding Pathways using All-Atom Models*; Springer-Verlag Berlin Heidelberg, Lect. Notes Phys. 703, 393-433, (2006).

[8] A.J.Mulholland *Introduction. Biomolecular simulation*; J.R Soc Interface 5, 3, 169-172, (2008).

[9] P.H. Hünenberger *Thermostat algorithms for molecular dynamics simulations* Adv. Polymer. Sci. 173, 105-149, (2005).

[10] G. Bussi, D. Donadio and M.Parrinello. *Canonical sampling through velocity-rescaling* J. Chem. Phys., 126, 014101, (2007).

[11] D.L.Nelson, M.M.Cox *Lehninger Principles of Biochemistry*; W. H. Freeman, 4th edition, (2004).

[12] H.P.Gajera, S.V.Pater and B.A.Golakiya *Fundamentals of Biochemistry: A Textbook*; Interncational Book Distributing Co., 1st edition, (2008).

[13] A.R.Leach. *Molecular Modelling: PRINCIPLES AND APPLICATIONS*; Prentice Hall, 2nd edition, (2001).

[14] D.Frenkel and B.Smith. *Understanding Molecular Simulations: From Algorithms to Applications*; Academic Press: New York, (1996).

[15] M.P.Allen and D.J.Tildesley. *Computer Simulation of Liquids*; Clarendon Press: Oxford, (1991).

[16] V.A. Harmandaris, *Molecular Dynamics Simulations of Polymers*, V.G. Mavrantzas, *Chapter in Book Simulation Methods for Polymers*, Edited by M.J. Kotelyanskii and D.N. Theodorou, Marcel Dekker, New York, (2004).

[17] M.E.Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford University Press, New York, (2010).

[18] M.P.Allen. *Introduction to Molecular Dynamics Simulation*; NIC Series, 23, 1-28, (2004).

[19] L.R. Scott and A. Fernández *A Mathematical Approach to Protein Biophysics*; Springer International Publishing AG, Cham, (2017).

[20] R. Gabler *Electrical Interactions In Molecular Biophysics*; AP, New York, (1978).

[21] A.N.Rissanou, E.Georgilis, E.Kasotakis, A.Mitraki and V.Harmandaris *Effect of Solvent on the Self-Assembly of Dialanine and Diphenylalanine Peptides*; J.Phys.Chem. B, 117, 3962-3975, (2013).

[22] M.A. González. *Force fields and molecular dynamics simulations*; Collection SFN, 12, 169–200, (2011).

[23] J. Janin, S. Wodak, M. Levitt and B. Maigret. *Conformation of amino acid side-chains in proteins*; J.Mol.Biol. 125, 357-386, (1978).

[24] M.N.G. Jame, A.R. Sielecki. *Structure and refinement of penicillopepsin at 1.8 Å resolution*; J.Mol.Biol. 183, 299-361, (1983).

[25] J.W. Ponder, and M. Richards *Tertiary Templates for Proteins Use of Packing Criteria in the Enumeration of Allowed Different Structural Classes* ; J.Mol.Biol. 193, 775-791, (1987).

[26] M. J. McGregor, S.A. Islam and M.J.E. Sternberg. *Analysis of the relationship between side-chain conformation and secondary structure in globular proteins*; J.Mol.Biol. 198, 295-310, (1987).

[27] V. E.Fadouloglou, N.M.Glykos, and M. Kokkinidis. *Side-chain conformations in 4-α-helical bundles.*; 4, 321-328, (2001).

[28] R. Kumar, J.R. Schmidt, and J.L. Skinner. *Hydrogen bonding definitions and dynamics in liquid water.*; J. Chem. Phys. 126, 204107, (2007).

[29] M.Monajjemil and A.R. Oliaey. *Gyration Radius and Energy Study at Different Temperatures for Acetylcholine Receptor Protein in Gas Phase by Monte Carlo, Molecular and Langevin Dynamics Simulation*; J.Phys. Theor. Chem. 1AU Iran, 5, 195-201, (2009).

[30] M.Yu. Lobanov, N.S. Bogatyreva and O.V. Galzitskaya. *Radius of Gyration as an Indicator of Protein Structure Compactness*; J. Mol. Biol., 42, (2008).

[31] Z.Shahbazi *Mechanical Model of Hydrogen Bonds in Protein Molecules*; American J. Mech. Eng., SciEP, 3, 47-54, (2015).

[32] A.L. Cuff, R.W. Janes and A.C.R. Martin. *Analysing the Ability to Retain Sidechain Hydrogen-bonds in Mutant Proteins*; Bioinformatics, 22, 1464–1470, (2006).

[33] S. Canossa. *Teaching Tutorial: Hydrogen bond Definition, examples, special cases*; Course of General and Inorganic Chemistry, 00, 00, 1-7, (2005).

[34] C. Chen, W.Z. Li, Y.C. Song, L.D.Weng and N. Zhang. *The effect of Geometrical Criteria on Hydrogen Bonds Analysis in Aqueous Glycerol Solutions*; J. Mol Imag Dynamic, 1:101, (2011).

[35] A.L Morris, M.W. MacArthur, E.G. Hutchinson and J.M. Thornton. *Stereochemical Quality of Protein Structure Coordinates*; Wiley-Liss, INC. 12, 345-364, (1992).

[36] S. Hovmöller, T. Zhou and T. Ohlson. *Conformations of amino acids in proteins*; Acta Cryst. D58, 768-776, (2002).

[37] C. Ramakrishnan *Ramachandran and his Map*; Indian Institute of Science, (2001).

[38] S.A. Hollinsworth and P.A. Karplus *A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins*; BioMol, 1, 271-283, (2010).

[39] G.N. Ramachandran, C. Ramakrishnan and V. Sasisekharan. *A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins*; J. Mol. Biol., 7, 95–9, (1963).

[40] R.J. Anderson, Z. Weng, R.K. Campbell, X. Jiang. *Main-chain conformational tendencies of amino acids*; Proteins, 60, 679–89, (2005).

[41] W.Kabsch. *A solution for the best rotation to relate two sets of vectors*; Acta Crystallographica Section A, 32, 922-923, (1976).

[42] W.Kabsch. *A discussion of the solution for the best rotation to relate two sets of vectors*; Acta Crystallographica Section A, 32, 922-923, (1978).

[43] B.Kovács *Computing RMSD and fitting protein structures: how i do it and how others do it*; Pázmány Péter Catholic University, (2016).

[44] I. Kufareva and R. Abagyan. *Methods of protein structure comparison*; Methods Mol Biol. 857, 231–257, (2012).

[45] L.D. Schuler, X. Daura, W.F.V. Gunsteren. *An Improved GROMOS96 Force Field for Aliphatic Hydrocarbons in the Condensed Phase*; J. Comp. Chem., 22, 1205–1218, (2001).

[46] T. Darden, D. York and L. Pedersen *Particle mesh Ewald: An $N \cdot log(N)$ method for Ewald sums in large systems*; J. Chem. Phys. 98, 10089, (1993).

[47] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen *A smooth particle mesh Ewald method*; J. Chem. Phys. 103, 19, (1995).

[48] G.A. Cinseros, M. Karttunen, P. Ren and C. Sagui. *Classical Electrostatics for Biomolecular Simulations*; Chem. Rev. 114, 779-814, (2014).

[49] Lectures: Molecular Simulations in Chemical Engineewring (Web), `http://nptel.ac.in/courses/103103036/24`

[50] `http://biochemistrycourse.blogspot.com/2011/05/primary-structure-of-proteins.html`

[51] `http://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/pepti`

[52] `https://amit1b.wordpress.com/the-molecules-of-life/about/amino-acids/`

[53] `http://bioinformatics.org/molvis/phipsi/`

[54] *Gromacs Manual: 5.0.7*

[55] `http://www.gromacs.org/`

[56] `http://www.ebi.ac.uk/pdbe/pisa/`

[57] `https://pymol.org/2/`

[58] `http://servicesn.mbi.ucla.edu/PROCHECK/`

[59] `http://easylifescienceworld.com/ramachandran-plot/`

[60] `https://en.wikipedia.org/wiki/Protein_structure`

[61] `http://cbio.bmt.tue.nl/pumma/index.php/Theory/Potentials`