UNIVERSITY OF L'AQUILA
UNIVERSITY OF NICE-SOPHIA ANTIPOLIS
ERASMUS MUNDUS MATHMODS PROGRAM

MASTER THESIS

Molecular Modelling and Simulations of Photosystems I Complex

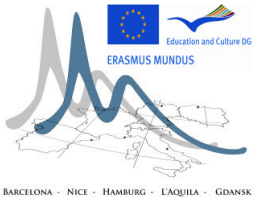
Student:
Kwame Atta GYAMFI

Advisors:
Prof. Leonardo GUIDONI
Dr. Daniele NARZI

A thesis submitted in partial fulfilment of the requirements for the degree of
Master of Science Mathematical Modelling in Engineering: *Theory, Numerics and Application*
in the group of

Computational Biochemistry Biophysics Chemistry Group

With the support from of the Erasmus Mundus Program of the European Union
and partial funding from ERC.



Data discussione tesi: 11.09.2014

Declaration of Authorship

I, Kwame Atta Gyamfi, declare that this thesis titled, 'Molecular Modelling and Simulations of Photosystems I Complex' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a master degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 11.09.2014

UNIVERSITY OF L'AQUILA

Abstract

Faculty of Science

Department of Information Engineering, Computer Science and Mathematics (DISIM)

Master of Science Mathematical Modelling in Engineering: *Theory, Numerics and Application*

Molecular Modelling and Simulations of Photosystems I Complex

by Kwame Atta Gyamfi

Two protein/cofactors complexes embedded in the thylakoid membrane, Photosystem I (PSI) and Photosystem II (PSII), play a central role in the conversion of solar energy into chemical energy. The process is initiated when a network of cofactors in PSII are photo-induced at specific wavelengths enabling the catalytic core of PSII to oxidize two water molecule into molecular oxygen and four equivalent of H^+ and $\mathbf{e}^-$. The four electrons extracted from the water are transferred through the electron transfer chain (ETC) of PSI to ultimately reduce $2NADP^+$ to 2NADPH. The light induced oxidation of the water in the Photosystem II (PSII) protein complex is catalyzed by the Mn_4Ca cluster which has been the focus of our previous studies on the photosynthetic apparatus [11]. In this work, first we modelled the structure of PSI starting from an uncompleted x-ray structure. We provide an energy minimized geometry of the PSI structure using standard energy minimization schemes. We follow up with vacuum MD simulations restraining the positions of the atoms present in the crystal structure allowing the added atoms to rearrange their starting positions in conformations which is energetically more favourable. This work provide us with a complete starting structure to study for the first time by classical MD simulations and in subsequent studies, by mixed QM/MM calculations the dynamics and the function of the PSI complex at physiological conditions.

Acknowledgements

The ultimate thanks goes to the Most High God Jehovah for carrying me through all these times. I'm also grateful to my supervisors for their advice during this training. Such a wonderful experience with the group

Contents

List of Figures

List of Tables

I dedicate this to my wonderful family

Chapter 1

Introduction

1.1 Introduction

Photosynthesis is a term used to describe the process by which photosynthetic organisms efficiently convert solar energy from sunlight into chemical energy. This extraordinary chain of events is started and regulated by light harvesting complexes and chromophores. Energy from sunlight is captured by these complexes that subsequently funnel it to reaction centres on timescales of 10-100 picoseconds. The excitation energy is used to oxidise water molecules in the reaction center of Photosystems II. Thus generating a proton gradient across the thylakoidal membrane and at the same time producing electrons for subsequent reactions. In eukaryotes (e.g plants and algae), the primary reactions of photosynthesis takes place in a special cell organelle called **chloroplast** in which chlorophyll play a very key role. Within the thylakoid membranes of the chloroplast there are two gaint protein complexes namely Photosystems I (PSI)and photosystem II(PSII). The nomenclature indicates the order in which they were historically discovered and not their physiological order of their existence in the photosynthetic apparatus.[31] These two biomolecular systems work in series and are functionally coupled by cytochrome b_6f complex which basically mediates electron transfer between them. Under normal circumstances electrons flow from PSII to PSI via light absorption at wavelengths 680nm and 700nm respectively. PSII uses the light-induced energy to oxidize two molecules of H_2O into molecular oxygen (O_2) and 4 equivalents of protons and electrons in a process called *water splitting*. Thus up to $4\mathbf{e}^-$ are removed from the water molecules and are then transferred through the electron transfer chain(ETC) of PSI to ultimately reduce 2 oxydoreductase ($NADP^+$) to NADPH. During the electron transfer process the proton gradient generated across the membrane which is the driving force for the synthesis of Adenosine tryphosphate (ATP). ATP is a nucleoside triphosphate used in cells often

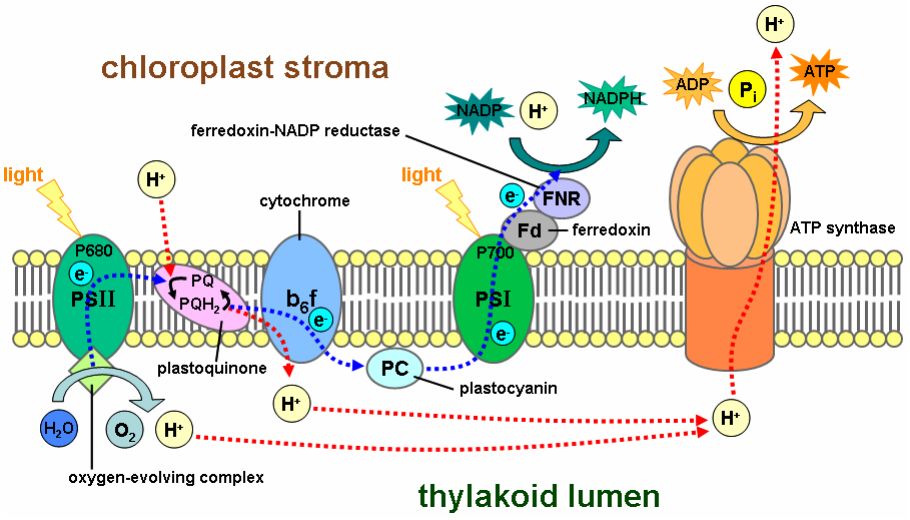


FIGURE 1.1: Photosynthetic electron transport chain of the thylakoid membrane. [6]

called the "molecular unit of currency" of intracellular activities. It is believed that the principles learned from studies of the various natural antenna organelle involved in this process suggest how to elucidate strategies for designing efficient light-harvesting systems artificially [34]. By conducting theoretical and experimental studies on the photosynthetic apparatus at the atomic scale, scientist hopes to apply results to solve the earth's energy problem while reducing CO_2 emission concurrently.

Computer simulation methods play an important role in the investigation of biological processes and functions at such a scale. Moore's law suggests that computer power and speed doubles every eighteen months or less, implying that molecular dynamics simulations can be extended to larger systems regardless of it's fast dynamics [25]. This makes it possible to reproduce the motions of biomolecules and to obtain information that would otherwise be inaccessible from experiment in physiological conditions. This work aims to model the whole molecular structure of the PSI starting from it's partially solved X-ray structure in order to subsequently perform a classical and QM/MM equilibrations.

1.2 Overview of Photosynthesis

Even though there is a third biological process that mediates photosynthesis in which light energy is utilized, oxygenic and anoxygenic kinds are the most widely known. Oxygenic photosynthesis – the photosynthetic process which involves the production of O_2 resulting from the oxidation of water as the main electron donor, is often carried out by plants, algae and cyanobacteria. Anoxygenic photosynthesis, gains all of its electrons from sources other than H_2O (for example from hydrogen sulphide, H_2S).

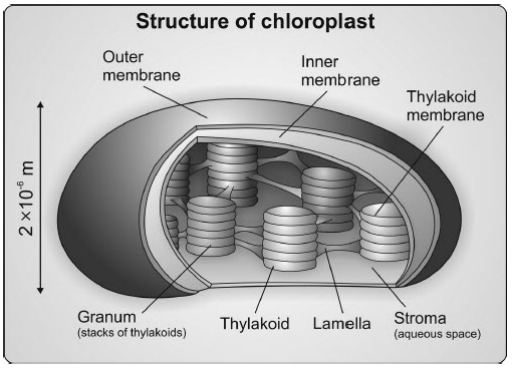


FIGURE 1.2: 3D diagramatic view of a chloroplast. Picture adapted from: [9]

However the name anoxygenic suggests the absence of oxygen and therefore cannot exist in cyanobacteria (See [21] and [9]). Whilst these kinds of photosynthesis can occur in bacteria and involves the production of ATP for cell activities, the latter is only present in *halophilic Archea* and involves the extruding of Cl^-. Principally, this process cannot be categorized as photosynthesis since the main reaction center doesn't contain chlorophyll, the pigment typically common in all photosynthetic membrane. In this article, we would generally use oxygenic photosynthesis to represent photosynthesis regardless of the existence of the others.

The primary reactions of photosynthesis in eukaryotes (e.g. plants and algae) takes place in a special cell organelle called *chloroplast*. The chloroplast is enclosed by a double membrane which separates the innermost **stroma** from the outermost cell cytoplasm. The stroma is an aqueous space within which the enclosed membrane vesicle called **thylakoid** is found. Thylakoids form a physically continuous three-dimensional network (see figure on page 3) enclosing an aqueous space called the **lumen** and can be differentiated into two distinct physical domains: cylindrical stacked structures called **granna** and interconnecting single membrane regions **stroma lamellae**. For a full review of the thylakoidal architecture see the [28] and [7]. The protein complexes that catalyze electron transfer and energy transduction are unevenly distributed in thylakoids: PSI is located in the stroma lamellae, PSII is found almost exclusively in the grana, the F-ATPase is located mainly in the stroma lamellae, and the cytochrome b_6f complex is found in grana and grana margins. [28].

The initial charge separation occuring in PSI is commenced by a special chlorophyll electron donor pair, **P700**, consisting mainly of a chloropyll *a* molecule coupled to another chlorophyll *a'*. From **P700**, the electron is transferred stepwise to **A** (a chlorophyll *a* molecule), A0 (another chlorophyll *a* molecule), **A1** (a phylloquinone molecule) and then subsequently to the three iron-sulphur clusters, named **FX**, **FA** and **FB**. After the docking of terminal cluster **FB**, the electron is transferred to Fe_2S_2 cluster of flavodoxin which transfers the electron to the $NADP^+$-reductase to be reduced to NADPH. In order

to complete the cycle, $P700^+$ is re-reduced. In the next section, we give an overview of the architectural arrangements of protein subunits and co-factors of PSI.

1.3 Photosynthesis I Architecture

Our knowledge regarding the structure [17] [18] of PSI is based on the X-ray crystallised structure of Thermophilic cyanobacteria *Synechococcus elongatus*. Cyanobacterial PSI exists in photosynthetic membrane *in vivo* as both trimeric and monomeric form depending on the organism and some light conditions. Our computational study employed in this article take the monomeric form of [17]. See figure on page 5 below. Each monomeric unit consists of 12 protein subunits (PsaA, PsaB, PsaC, PsaD, PsaE, PsaF, PsaI, PsaJ, PsaK, PsaL, PsaM and PsaX) to which 127 non-covalently bonded cofactors. PsaA and PsaB are the two most largest subunits located at he center of PSI. These co-factors accounts for about 30% of the total mass of PSI which 356KDa. The process catalyzed by PSI can be divided into light capturing, excitation energy transfer and electron transfer processes. The initial light capturing process is performed by a large antenna system that consists of 90 antenna chlorophylls and 22 β-carotenoids. Excitation energy is transferred to the center of the complex, where the electron transport chain (ECT) is located. The ECT is functionally the most important part of PSI. Located at the heart of PSI, it consists of 6 chlorophylls, 2 phylloquinones and all 3 4Fe4S clusters. Iron-sulfur clusters FA and FB are carried by PsaC which forms the docking site for ferredoxin/flavodoxin together with PsaD and PsaE. Majority of these cofactors are harboured by the PsaA and PsaB. The overall PSI complex extends into the stroma by 90Åwith protein subunits PsaC, PsaD and PsaE which provide the docking site for the ferredoxin/flavodoxin. It must be noted that not all the co-factors of the ETC as well as several amino acid residues has been determined spectroscopically in experiment. See [31], [18], and [13]. In this work, we attempt to provide a model structure for the undetermined and partially solved molecular structures.

1.4 Theory and Modelling methods in Photosynthesis

Molecular dynamics (MD) simulations provide information concerning the thermal atomic motions of proteins biomolecules. They calculates the time-dependent molecular trajectories of motion with considerable accuracy and that have the potential to surpass the amount of information contained in static x-ray structures. Specific computational methods with this focus have already contributed significantly to photosynthesis research. There is the need for a multi-scale approach in the effort to describe processes in both

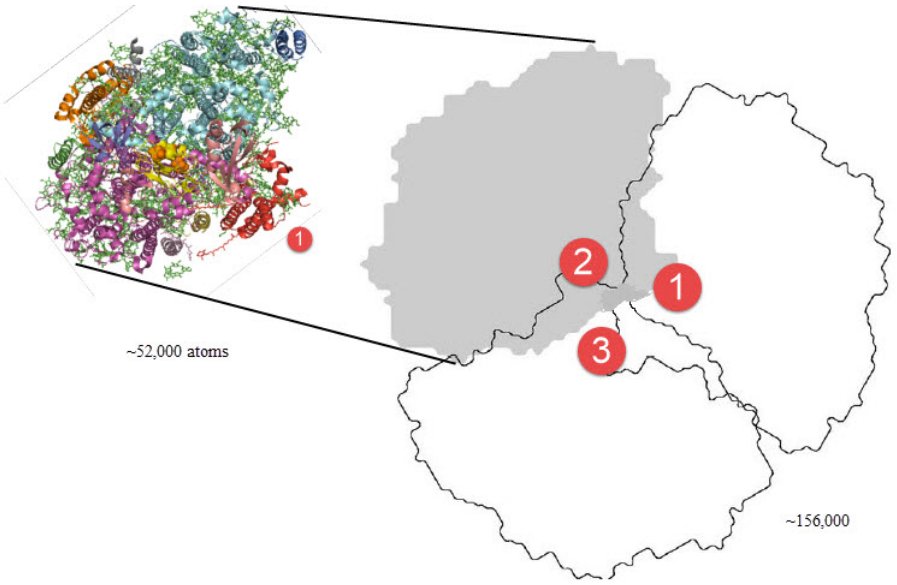


FIGURE 1.3: View of C_3 symmetry axis of the maximum lateral extent of trimeric PSI(outlined) and it's corresponding monomer (colored-left). Outlined image adapted from [13] and monomer is taken with PyMOL visualization software.

all the photosynthetic apparatus in order to access several order of magnitudes in time scale and system size [13]. It is undeniable that using quantum-mechanical dynamical evolution of the system would adequately describe the phenomena in photosynthesis. However, in spite of the considerable increase in computation power, that approach still remains a dream and therefore some approximations should be made. In this section, we outline the computational methods commonly employed in photosynthesis research.

1.4.1 First-principles Approaches

Calculations by first principles quantum-mechanical approach is by far one of the most difficult tasks to perform as one needs to take into account the complex pigment-pigment and pigment-protein interactions so abundant in all the photosystems. In this context, accurate highly corelated wavefunction-based methods are computationally expensive and can hardly deal with the large molecular model of interest[12]. Therefore the most successful quantum chemical method mostly used for applications in large molecular complex is the density functional theory (DFT). The main component in DFT is the electron density which is a non-negative function dependent on three spatial coordinates but can be measured experimentally [3] (for e.g. by X-ray diffraction methods). It is capable of describing the electronic ground states and excited states. In photosynthetic research, several references could be given with regards to the application of DFT methods. Canfield and Dahlbom *et al* (2006) has successfully developed a system-wide optimization scheme for the 150,000-atom of the PSI trimer. Another study is based a

time-dependent DFT(TD-DFT) by [29] which addressed the issue of the environmental effects on the excitation energies and photophysical properties of LH2 complexes.

1.4.2 QM/MM methods

Almost more than half, if not all of the processes of interest in natural photosynthesis, are characterized by huge pigment-protein complexes with atomic counts amounting to hundreds to thousands spanning over several orders of magnitude from picoseconds to milliseconds. Regardless of the considerable progress made in DFT method based approaches, there's still the need to develop novel multiscale approaches with minimal complexity and computational costs. Quantum Mechanical/Molecular Mechanical (QM/MM) interface between classical and the quantum mechanical approaches with the view to overcoming diring challenges encountered in MD approaches. The first step in QM/MM simulation is to divide the system in two subsystems: One "inner" (usually a small) region which is treated with principles of quantum mechanics (QM) and an "outer" region treated with classical molecular mechanics (MM). The basis for this system level separation is that the region where QM approach is used is usually limited to a relatively small region where the electronic structure changes significantly (for e.g. bond-making, bond-breaking processes) [8]. In the context of photosynthesis a typical QM/MM application has recently been done describing the catalytic cycle of the oxygen evolving complex in photosystem II. [11]

1.4.3 Ab initio MD

Ab initio molecular dynamics seeks to provide approximate solution to the electronic Schrodinger equation via the Born-Oppenheimer approximation for each nuclear configuration. This scheme can be defined by the coupled equations:

$$M_1 \frac{\mathrm{d}^2 R_1}{\mathrm{d}t^2} = -\nabla_1 \langle \Psi | H_e | \Psi_0 \rangle \tag{1.1}$$

$$H_e \Psi_0 = E_0 \Psi_0 \tag{1.2}$$

The above equation is the Newton's second law of motion for nucleus I of mass M_1 and position R_1. The force on the right hand side is obtained by calculating the gradient ∇_1 of the total energy with respect to the nuclear coordinates to obtain H_e, the expected value of the electronic Hamiltonian H_e. An effecient scheme to solve (1.1) has been in use since 1985 and is referred to as the Car-Parrinello molecular dynamics method (CPMD). See [14]. In the CPMD method, DFT is generally used for computing the electronic ground-state energy.

1.4.4 Classical Molecular Dynamics

Classical molecular dynamics (MD) presents a simple yet very powerful approach to describe trajectories of atomic motions in macromolecules. In this approach the Newtonian equations of motions are numerically solved by evolving in time the positions and velocities of each particle. See chapter 2 for detailed implementation of this method. This numerical technique has been applied to study the reorganization energy of the initial electron-transfer step in photosynthetic bacterial reaction centers (BCR). Moreover, atomistic simulations can be used to estimate parameters needed in the so-called coarse-grained models. [29]. See references for details. In this article, we seek to model an initial structure of PSI complex to perform classical MD simulations generated molecular trajectories with a duration of a few picoseconds.

1.5 Summary

This article is organized in three chapters outlined as follows: Chapter one has provided an introductory background to the article. In chapter two, the theoretical background of the numerical calculations performed in this thesis are presented and closes the work with a presentation and analysis of results in chapter three.

Chapter 2

Chapter Two

Introduction

Molecular modelling is the science and art of studying molecular structure and it's function through model building and computation [8]. With such a simple definition, only few could imagine the complexity of the methods and techniques developed and implemented in application. These methods are diverse and complicated as the three main disciplines that bridges this field.

The computations aspect encompass *ab initio* or semi-empirical quantum mechanics, empirical (molecular) mechanics, molecular dynamics, Monte Carlo, free energy and solvation methods, structure/activity relationships (SAR), chemical/biochemical information and databases, and many other established procedures. The so-called model building component comprises mainly of experimental techniques such as nuclear magnetic resonance (NMR) or X-ray crystallography, spectroscropy among others. Without much regards to the topic of the study, computational component of molecular modelling is the main approach employed in the study. The study seeks to apply standard methods of molecular dynamics simulations in search of a stable, minimized structure of a monomeric PSI complex.

In this chapter, we provide a brief historic background and review of computational methods (classical molecular dynamics) applied to the study of our biomolecular system. The section which follows discusses theoretical foundations to methods used in application. Thus providing as close as possible, all the derivation of standard techniques from the discussion of the fundamental principles of physics that makes classical approximation to the biomolecular structure possible.

2.1 Background History

As late as in the 1930s, protein crystals had already been obtained. After Robert Brown (1827) successfully described thermal motion of particles in solution ('Brownian' motion), René J. H. Dutrochet(1828) firmly believed that biological processes could be explained in terms of physics and chemistry. His conviction was heavily based on his extensive study he had conducted on osmotic pressure in living systems. However, the first high resolution structure of proteins - myoglobin and hemoglobins – were solved by and [22, 30] making it possible for the analysis and study of the 'anatomy' of protein structure and functions at atomic details by computational methods.

Decades later molecular dynamics (MD) simulations has emerged as one of the most effective approaches or methods required to obtain the dynamic properties of many-body system. The first MD was first accomplished for a molecular system of hard spheres by Alder[1] [10] in their studies on the dynamics of liquids. They considered the 'liquid particles' to be moving at constant velocities between elastic collisions while solving the equations of motion without making any approximation within the limits imposed by the 'machine'. It was not until several years later before Rahman (1964) made a successful attempt to solve the equations of motion for a set of Leannard-Jones particles. Due to revolutionary advances in computer technology and algorithmic improvements, MD has subsequently become a valuable tool in many areas of physics and chemistry. Since the 1970s MD has been used widely to study the structure and dynamics of macromolecules, such a proteins or nucleic acids. It's success can be attributed to it's solid foundations in Physics, Chemistry and Mathematics. For the remaining of the chapter, we elaborate on the theoretical foundations that serve as the building block of this computational tool so prevalent in applied research community.

2.2 Theoretical Foundations of Molecular Dynamics Simulations

In this section we provide theoretical derivations of computational techniques employed from physics, chemistry and mathematics that intersect to be used in modelling and simulation of protein complexes. Most derivations provided here are geared towards the standard MD as it is the main computational tool employed in the study. We begin first with its roots in quantum mechanics and justify the need for classical approximations to

[1]Berni Julian Alder is a Swiss born in Germany and spent almost all his entire academic pursuit as at University of California. He is physicist and has specialized in statistical mechanics but currently considered a pioneer of numerical simulation in physics.

incorporate all atoms of huge biological complex as PSI. The development of the potential surface function for force field parametrisations is presented in later sections. These derivations are not expected to be complete, for a complete and detailed derivations, readers is advised to see the references provided herein.

2.3 Quantum Mechanical Foundations

Equations from first principle provides an accurate description of the properties of molecules such as protein. To obtain the stationary properties of a molecule (or many body) consisting of nuclei and electrons, it is necessary to solve the time-independent (stationary) Schrödinger equation:

$$H\Phi = E\Phi \tag{2.1}$$

where

$$H = \frac{\hbar}{2m}\nabla^2 + U$$

is the Hamiltonian. The first term in the Hamiltonian is the contribution for the kinetic energy term $\mathcal{T}$ and the second term is the potential energy term, U. Φ and E represents the wave function and energy of the system respectively which can be treated as the eigenfunction and eigenvalue of the Hamiltonian operator H. For full derivation of (2.1) see [1]

Due to the complexity of equation (2.1), an analytical solution is only possible for a system with a few number of atoms (e.g. the hydrogen atom). Therefore for macromolecules, direct numerical solution of equation (2.1) still remains a dream despite continuous increase in computational power and speed. In the light of these, there is the need to provide a framework to simplify the quantum description of a molecular system via appropriate physics and mathematical approximations. The most prominent and common among such approximations is the Born-Oppenheimer (BO) approximation.

This approximation decouples the motion of the heavy nuclei from that of light electron movements. This implies that nuclei is considered fixed whereas only the electronic motions are considered. As a result the E and Φ just reduces to electronic properties. BO approximation also serves as the foundation for *ab-initio* and semi-empirical quantum calculations in molecular dynamics studies. In the next section we attempt to give the mathematical formulation that motivates this approach.

2.3.1 BORN-OPPENHEIMER APPROXIMATION

Let's consider a system consisting of N_{nuc} atoms (each atom has one nucleus) with nuclear charge $Z_1 \dots Z_{N_{nuc}}$, location at $\mathbf{R}_{nuc}$ in the Cartesian positioning and momenta $\mathbf{P}_{nuc}$. If there are N_{el} electrons with positions and momenta r_{el} and p_{el} respectively, the Hamiltonian has the form:

$$H = \mathcal{T}_{nuc} + U_{el-el} + \mathcal{T}_{el} + U_{nuc-nuc} + U_{el-nuc} \tag{2.2}$$

where the terms has the following meanings: $T_{el} = \sum_{j=1}^{N_{el}} \frac{p_j^2}{2m_e}$ and $T_{nuc} = \sum_{j=1}^{N_{nuc}} \frac{P_j^2}{2M_j}$, m_e represents the electron mass and M_j is the mass of the jth nucleus. The interactions between the nuclei and electrons are expressed via the Coulombic force to obtain the potential terms for electron-electron interaction $\mathbf{U}_{el-el}$, the nucleus-nucleus interaction $\mathbf{U}_{nuc-nuc}$ as well as for the nucleus-electron interaction U_{nuc-el} as:
$U_{el-el} = \frac{1}{2}\sum_{j\neq i} \frac{e^2}{|r_i - r_j|}$; $U_{nuc-nuc} = \frac{1}{2}\sum_{j\neq i} \frac{Z_i Z_j e^2}{|R_i - R_j|}$ and $U_{nuc-el} = -\frac{1}{2}\sum_{j\neq i} \frac{Z_j e^2}{|r_i - R_j|}$.
The Hamiltonian expression of (2.2) gives the full version of the Schrödinger equation for any many body system. The BO approximation is one of the several empirical approximations capable of providing a framework to describe both nucleic and electronic properties of such systems. The method assumes that the electrons move in the field generated by the statics of the nuclei. This assumption is valid due to the mass disparity between electrons and the nuclei. Thus the electrons are assumed to be relatively 'lighter' than the nuclei, and will therefore be able to respond instantaneously to any perturbation in the nuclear configuration allowing the electronic Hamiltonian to be represented in the form which parametrically depends on the nuclear coordinates, $\mathbf{R}_{nuc}$ giving:

$$H_e = \mathcal{T}_{el} + U_{el-el} + U_{el-nuc} \tag{2.3}$$

Trivially, $\mathcal{T}_{nuc-nuc} = 0$. Almost all quantum chemistry algorithms seek to solve this electronic Hamiltonian resulting in the electronic energies $E_n^{el}(\mathbf{R})$ and the wave function $|\Phi_n(r, \mathbf{R})\rangle$ which will depend parametrically on the nuclear geometry and is written in the form:

$$|H_n(r, \mathbf{R})\rangle = E_n^{el}(\mathbf{R})|\Phi_n(r, \mathbf{R})\rangle \tag{2.4}$$

When inter-nuclear repulsions are added, the solution of the electronic Schrödinger equation in (2.4) for different nuclear coordinates $\mathbf{R}$ results in the potential hyper-surface, $U(\mathbf{R})$, which is constant with respect to the electronic coordinates and therefore any non-adiabatic coupling effects could be neglected. The elimination of the non-adiabatic electronic coupling is the core of the BOA leading to the nuclear Schrödinger's equation:

$$H_n|\Phi(\mathbf{R})\rangle = (\mathcal{T}_{el} + U_n(\mathbf{R})|\Phi_n(\mathbf{R})\rangle = E_n^{nuc}(R)|\Phi_n(\mathbf{R})\rangle \tag{2.5}$$

The equation given by (2.5) describes the geometry of the nuclei in the field generated by the fast moving electrons even though each electronic state generates a different nuclear potential. The dynamics of the nuclei is well described reasonably well with the nuclear potential $U(\mathbf{R})$ within the BOA as long as the potential surfaces belonging to different states are well separated and serves as the origin of the potential functions or *force-fields* used in MD. (See section 2.5 in this article). This force-fields can be parameterised to account for both bonded and non-bonded interactions. The classical approximation to the quantum mechanical description of a molecule and their interactions are not directly derived from 'first principles' approaches, but rather, is the results of adapting both structure and potential function to a variety of different kinds of information. This includes the results of quantum mechanical energy calculations, experimental data obtained by thermodynamics and various kinds of spectroscopic means, the structure of the crystalline state, measurements of transport properties, collision studies using molecular beams, and so on [5]. MD generally adopts a classical point of view, typically representing atoms or molecules as point masses interacting through the forces that depend on the positions of these objects. In section 2.5, we provide the theoretical background to what terms are included in the potential function in MD. Regardless of the approach (classical or quantum) used it is still necessary to apply to some extent, quantum chemical calculations to derive parameters to cater for charges of the metal centres so prevalent in the structure of PSI (for e.g iron-sulfur clusters). This is done in the next section.

2.4 Restrained Electrostatic Potential (RESP)

Electrostatic energy represents an important term in the potential energy function of almost all *force fields* and therefore an accurate representation is important for obtaining good results. Within the partial charge model, the atomic charges are normally assigned by fitting the molecular electrostatic potential calculated by an electronic structural method. The electronic potential U at a point $\mathbf{R}$ is given by the nuclear charges and an electronic wave function as in the following:

$$U(\mathbf{R}) = \sum_{\alpha}^{N_{nuc}} \frac{Z_\alpha}{|\mathbf{R}_\alpha - \mathbf{R}|} - \int \frac{\phi^2(\mathbf{R}')}{|\mathbf{R}' - \mathbf{R}|} d\mathbf{R}' \tag{2.6}$$

Where $\phi(\mathbf{R}')$ is the electrostatic field potential. The fitting is done by minimizing an error function of the following form, with the constraint that the sum of the partial

charges $Q_i = Z_\alpha$.

$$ErrF(Q) = \sum_{r}^{N_{points}} \left(\phi(r) - \sum_{\alpha}^{N_{atoms}} \frac{Q_\alpha}{|\mathbf{R}_\alpha - \mathbf{R}|} \right)^2 \tag{2.7}$$

The electrostatic potential is sampled at a few thousand points in the vicinity of the molecule. Usually the set of equations arising from minimizing the error function are often poorly conditioned so the calculated partial charges are sensitive to small details in fitting the data. The physical reason for this is that the electrostatic potential is primarily determined by the atoms near the surface of the molecule, while the atoms buried within the molecule have very little influence on the external electrostatic potential. A straight forward fitting procedure often results in unrealistic small charges for the non-surface atoms. To some extent, this problem could be avoided by adding a hyperbolic penalty term for having non-zero partial charges, since this ensures that only those charges that are important for the electrostatic potential have values significantly different from zero. This scheme is known as Restrained Electrostatic Potential(RESP) fitting and is currently considered the most efficient by majority of force field developers. We have successfully applied this methods to derive partial charges for cofactor molecules that are present in the PSI complex with the GAUSSIAN [16] and ANTECHAMBER [36] software programs. The detailed procedures are outlined in chapter 3.

2.5 Interatomic Interactions Potential Function

The potential energy function, U_n in section 2.3.1 can be expressed as the sum of local (or bonded) and non-local (or non-bonded) terms. Local interactions are induced by bond structure of the molecule whereas the *long-range* local terms considers non-bonded interactions such as electrostatic and van der Waals interactions. A simple but widely-used form of the potential function in MD comes from Ryckaert-Bellemans potential for modelling alkane chain [32]. For convenience, the bold face symbol $\mathbf{q}_i \in \mathbb{R}^3, i = 1, ..., N$ shall be used to represent the position for the $i-$th atom throughout the text. The local interaction terms are [19]:

Bond stretching Covalent bonds between atoms i and $i+1$ is modelled by harmonic stretching and is described by $\mathbf{U}_{bd}(\mathbf{q}_i, \mathbf{q}_{i+1}) \propto (r_i - r_{eq})$ where $\|\mathbf{q}_{i+1} - \mathbf{q}_i\|$

Bond rotation Angles formed by the covalent bonds between three successive atoms are modelled by $\mathbf{U}_{ba} = (\mathbf{q}_{i-1}, \mathbf{q}_i, \mathbf{q}_{i+1}) \propto (\psi_i - \psi_{eq})^2$ where the bond angle comes from $\psi_i = \angle(\mathbf{q}_{i-1}, \mathbf{q}_i, \mathbf{q}_{i+1})$

Torsion angles The third term in the potential function considers the motion of the torsion angles ω_i between two planes spanned by three atoms and therefore depends on the positions of four successive atoms given as $\mathbf{U}_{ts} = \mathbf{U}_{ts}\left(\mathbf{q}_{i-1}, \mathbf{q}_i, \mathbf{q}_{i+1}, \mathbf{q}_{i+2}\right)$

The corresponding *non-bonded* interactions are similarly outlined:

Electrostatic interaction This term takes into account the electrostatic interaction from the charges of the atoms j and k which is usually given by the Coulomb potential $\mathbf{U}_C = (\mathbf{q}_j, \mathbf{q}_k) \propto \frac{1}{d_{jk}}$ where $d_{jk} = \|\mathbf{q}_j - \mathbf{q}_k\|$.

van der Waals interaction : In the van der Waals term, interactions between polarizable atoms are modelled by a Lennard-Jones potential $\mathbf{U}_{LJ}\left(\mathbf{q}_j, \mathbf{q}_k\right) \propto \frac{1}{d_{jk}^{12}} - \frac{1}{d_{jk}^{6}}$ and contains both long range and short range interactions.

Thanks to the additivity principle (see the book [8]), all of these terms could be summed up to give:

$$\begin{aligned}\mathbf{U}(\mathbf{q}) = &\sum_i \mathbf{U}_{\text{bd}}\left(\mathbf{q}_i, \mathbf{q}_{i+1}\right) + \sum_j \mathbf{U}_{\text{bd}}\left(\mathbf{q}_{j-1}, \mathbf{q}_{j+1}, \mathbf{q}_{l+1}\right) \\ &+ \sum_k \mathbf{U}_{\text{TA}}\left(\mathbf{q}_{k-1}, \mathbf{q}_k, \mathbf{q}_{k+1}, \mathbf{q}_{k+2}\right) + \\ &\sum_{k,l} \mathbf{U}_C\left(\mathbf{q}_k, \mathbf{q}_l\right) + \sum_{k,l} \mathbf{U}_{\text{LJ}}\left(\mathbf{q}_k, \mathbf{q}_l\right)\end{aligned} \tag{2.8}$$

Even though this potential doesn't capture interactions which involve more than four atoms modern force fields are parametrized to overcome this challenge. It has several counts of success in applications allowing computational studies on systems of 100,000 atoms or more [24]. Other possibly relevant interactions that are not directly contained in the potential can be accounted for by adjusting the potential function parameters in an appropriate way. This parametrization yield what is popularly known as the *force-field* parametrization. Empirical force field parametrizations can be and are always developed with the view to incorporating solvent effects in order to give a realistic description of the biomolecular environment. See [24], [27] and [38] for close comparisons of different force fields. A typical modern empirical force field derived from (2.8) takes the form

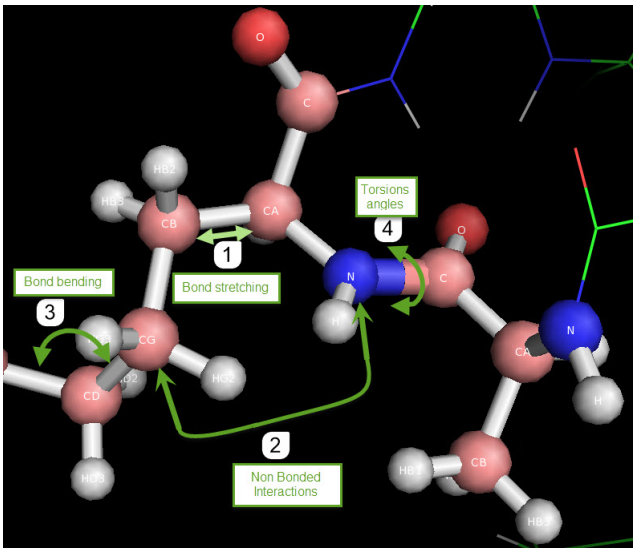


FIGURE 2.1: Illustration of MM force field parameterization

given below:

$$\begin{aligned}\mathbf{U(q)} = &\sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \\ &\sum_{dihedrals} \times K_\chi(1+\cos(n\chi-\delta)) + \sum_{improper} K_{imp}(\varphi-\varphi_0)^2 + \\ &\sum_{non-bonded} \varepsilon_{ij}\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{Q_iQ_j}{\varepsilon r_{ij}}\end{aligned} \tag{2.9}$$

Where b is the bond length; θ is the valence bond angle; χ represents the dihedral or torsion angle; φ, the improper angle; and r_{ij} is the actual distance between atoms i and j. The terms that represent the actual force field, include the bond force constant and equilibrium distance, K_b and b_0 respectively; the valence angle force constant, multiplicity and phase angle, K_χ, n, and δ, respectively; and the improper force constant and equilibrium improper angel, K_φ and ψ_0 respectively. These values collectively represent the internal or intra-molecular parameters.

Non-bonded parameters between atoms i and j include the partial atomic charges, Q_i and the LJ well-depth, ε_{ij}, and minimum interaction radius, $R_{min,i}$ are obtained for individual **atom types** and then combined to yield ε_{ij} and $R_{min,ij}$ for the interacting atoms via some combining rules. The dielectric constant, ε is typically set to 1 to correspond to permittivity of vacuum, for the incorporation of explicit solvent representations. Alternative methods to treat the solvent environment is an active area of research currently being undertaken by scientist in this speciality [38]. As evident, equation (2.9) are just simple functions which are used to describe the minimal set of forces that can be used to describe molecular structures. Several MD packages (e.g. The AMBER [23] in the

AMBER software[36]) employ this standard force field to describe bonds, angles and out-of-plane distortions in molecules. During our the modelling of initial structures, we used the AMBER force fields [23] for the standard proteins and the GAFF[37] for cofactor molecules in the PSI complex.

2.6 Foundations of Classical Mechanics (MM)

In the theory of molecular mechanics (MM) the biomolecular system is characterised as a microscopic mechanical system in which atoms are linked by mechanical springs which controls their covalent bonds, angle between successive bonds, rotations around the bonds, etc. The atoms are assumed to interact with each other (attraction or repulsion) according to non-bonded potentials that determines the non-bonded inter-atomic forces. A potential function is required to mathematically deduce their inter-atomic and macroscopic thermodynamic properties by exclusively incorporating classical terms. The concept is based on the basic formulations of Lagrange and Hamilton in classical mechanics. This potential (or energy) function is then used to compute all the relevant forces for equations of motions which ultimately describe the microscopic motions of the atoms in the molecular system of interest.

2.6.1 Equations of Motion

Lagrangian Mechanics

Consider a molecule with configuration space $Q \subseteq \mathbb{R}^n$ and molecular configurations $\mathbf{q} = \left(q^1, \ldots, q^n\right)^T$. Since a typical biomolecular system usually involve a large number N of atoms, the spartial dimensions $n = 3N$ is large. Let $U : Q \longrightarrow \mathbb{R}$ be a smooth molecular interaction potential and assume the system is *bounded* or *infinity at infinity*: (thus $U \longrightarrow \infty$ as $\|q\| \longrightarrow \infty$). This assumption ensures that individual particles doesn't escape to infinity or the system is *periodic* in the sense that $Q \cong \mathbf{T}^n$ for any $\mathbf{T} \subset \mathbb{R}^n$. In these cases, the Lagrangian takes the form

$$\mathcal{L}\left(\mathbf{q}, \dot{\mathbf{q}}\right) = \frac{1}{2}\langle M\dot{\mathbf{q}}, \dot{\mathbf{q}}\rangle - U\left(\mathbf{q}\right), \tag{2.10}$$

where $M \in \mathbb{R}^{n\times n}$ is a the diagonal, and positive-definite mass matrix and $\langle ., . \rangle$ is the standard inner product between vectors in $\mathbb{R}^n$. In addition, let a curve $q(t) \in Q$ with $t \in [a, b]$ and fixed endpoints $q(a) = q_a, q(b) = q_b$. The Hamilton's principle of least

action states that if a curve $q(t)$ minimizes

$$\int_a^b \mathcal{L}\left(\mathbf{q}(t), \dot{\mathbf{q}}(t)\right) dt, \tag{2.11}$$

then it is a solution of the corresponding Euler-Lagrange equations:

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{q}^i} - \frac{\partial \mathcal{L}}{\partial q^i} = 0, i = 1, \ldots, n. \tag{2.12}$$

A special case of (2.11) and (2.12) in Cartesian space is the familiar Newtons equations of motion. In this case $n = 3$, $Q = \mathbb{R}^{3N}$ and we let m_i be the corresponding atomic mass, such that $M = diag\left(m_1, m_1, m_1, \ldots, m_N, m_N, m_N\right)$, then (2.12) becomes:

$$m_i \ddot{q}_i = -\frac{\partial U}{\partial q_i},\ i = 1, \ldots, N. \tag{2.13}$$

One main advantage of the Lagrangian formulation is that it is invariant under any coordinate system transforms - a property which the Newtons formulation lacks. The proof of these equations can be found in any classical mechanic textbook. A good source of information can be found in [2].

Hamiltonian Mechanics

From the molecular Lagrangian (2.11), the Hamiltonian formalism can be introduced via the conjugate momentum variable:

$$p_i = \frac{\partial \mathcal{L}}{\partial \dot{q}_i},\ \ i = 1, \ldots, n. \tag{2.14}$$

assuming the transformation $(\mathbf{q}, \dot{\mathbf{q}}) \mapsto (\mathbf{q}, \mathbf{p})$ is invertible and requires the Hessian matrix $\frac{\partial^2 \mathcal{L}}{\partial \dot{q}_i \partial q_j} = M$ to be invertible. The conjugate pair $(\mathbf{p}, \mathbf{q})$ is called *phase space* variables. In this way, the Hamiltonian is just the Legendre transform of the Lagrangian $\mathcal{L}$ as:

$$H(\mathbf{q}, \mathbf{p}) = \langle \dot{\mathbf{q}}, \mathbf{p} \rangle - \mathcal{L}(\mathbf{q}, \dot{\mathbf{q}}),$$

As usual the Hamiltonian describes the energy of the system. Once it is defined, the Hamilton's equations of motion is straight forward:

$$\begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i} \\ \dot{p}_i &= -\frac{\partial H}{\partial \dot{q}_i}\ , \text{for}\ \ i = 1, \ldots, n. \end{aligned} \tag{2.15}$$

The straight consequence of the above is that:

$$\frac{\partial H}{\partial q_i} = -\frac{\partial \mathcal{L}}{\partial q_i} \tag{2.16}$$

which shows the equivalence of Hamilton's equations and the Euler-Lagrange equations shown above.

2.7 Foundations of Statistical Mechanics

Statistical mechanics is the branch of physics by which macroscopic properties of systems are studied from microscopic or molecular standpoint. It aids in the understanding and prediction of macroscopic properties from the properties of the individual molecules that make up the whole system [26]. Thus with statistical mechanics, basic concepts and assumptions are introduced and when applied to an N-body system, reproduces their thermodynamic states and function. In this section, we discuss some of these assumptions which are widely used in MD simulations and for that matter is employed in all simulations of this work. We provide the link between these discussions to the partition function and how it can be used to reproduce all the MD ensemble at different degrees of freedom.

2.7.1 Thermodynamic States

As discussed above in section 2.3, the Schrödinger equation is solvable only for elementary systems (e.g H_2). For many body systems, the solution would take the form:

$$\psi_{\text{total}} = \mathbf{C}\left(\psi_a(1), \psi_b(2), \psi_c(3), \ldots\right) \tag{2.17}$$

where ψ_a means a particle in state a with an energy E_a defined by equation (2.1). In statistical mechanics, there are two state representation of an N-body system, namely **microstate** and **macrostate**. A system is said to be in microstate when all parameters (for e.g. position and momentum) of the constituent particles are specified whereas for a macrostate, only the distribution of particles over the energy levels needs to be specified. Many microstates can exist for each state of the system specified through macroscopic variables (for e.g. energy, volume, number of atoms, etcetera) and there are many parameters for each state. From classical mechanics perspective, a microstate is usually represented by the phase space representation, position, $\mathbf{q} = (q_x, q_y, q_z)$ and momentum, $\mathbf{p} = (p_x, p_y, p_z)$ giving $6N$ degrees of freedom for a system constituted by N particles. For a system in equilibrium, only three macroscopic variables (P, V, T)

or (P, V, N) or (E, V, N), where P=pressure, V=volume, T=temperature, N=Number of particles and E=energy, are needed to describe the state of the system of interest. There must exists an equation of state for the system which relates the 3 variables to a fourth variable. For an ideal gas, this equation is given by $PV = NkT$ where k is the universal gas constant. Statistical mechanics formalism suggests that the equilibrium tends towards a macrostate which is the most stable and depends on the perspective of microstates. By an **a priori** assumption, *the macrostate which is the most stable contains the overwhelming majority of microstates.* This introduces probabilistic approaches in statistical mechanics which defines how energy is distributed among all microstates.

2.7.2 Distribution of Energy

The probability of finding the system in it's j^{th} quantum state at a specific temperature, T is given by the Boltzmann population formula:

$$\mathbf{P}_j = \frac{\Omega_j \exp(\frac{-\mathrm{E}_j}{kT})}{Z} \tag{2.18}$$

where

$$Z = \sum_j \Omega_j \exp(\frac{-\mathrm{E}_j}{kT}) \tag{2.19}$$

is called the partition function and Ω_j is the degeneracy of the jth state. The classical mechanical equivalence of (2.18) for a system of M coordinates and M momenta is given by:

$$\mathbf{P}(\mathbf{q},\mathbf{p}) = \frac{\exp\left(\frac{-\mathrm{H}(\mathbf{q},\mathbf{p})}{kT}\right)}{Z} h^{-M} \tag{2.20}$$

where H is the classical Hamiltonian, h is the Planck's constant, and the classical partition function is also given by:

$$Z = \hbar^M \int \exp\left(\frac{-\mathrm{H}(\mathbf{q},\mathbf{p})}{kT}\right) d\mathbf{q}d\mathbf{p} \tag{2.21}$$

Notice that (2.18) gives a non-zero probabilities for populating all states from the lowest to the highest. States of higher energy E_j are disfavored by the Boltzman factor, $\exp\left(\frac{-\mathrm{E}_j}{k\mathrm{T}}\right)$. This implies that if states of higher energy have higher degeneracies Ω_j(which they usually do), the overall population of such states may not be low. However, the **a priori assumption** states *all microstates of a given E are equally probable and thus the equilibrium macrostate must have overwhelming* Ω. Though partition function is known, many of the macroscopic properties of the system can be calculated using standard equations of statistical mechanics of different degrees of freedom. For

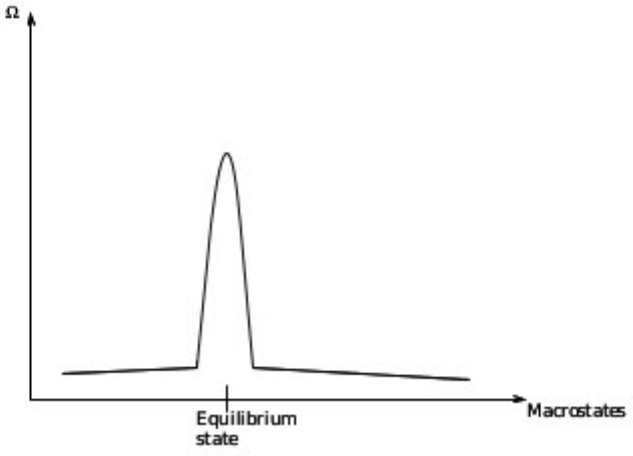


FIGURE 2.2: A plot of the population of microstate (in this case represented by the disorder number, Ω) macrostate at equilibrium. The disorder number is defined as the number of microstate available to a macrostate.

example it can be used to show that the kinetic and potential energies associated with translational, rotational and vibrational degrees of freedom are all quadratic.

2.7.3 Ensemble Averages

A molecular dynamics simulation generates a sequence of points in phase space as a function of time. Usually these points belongs to the same ensemble and they correspond to the different conformations of the system and their respective momenta. An ensemble is a collection of all possible systems which have different microscopic states but have an identical macroscopic or thermodynamic state. [26]. There exists different ensembles with different characteristics as described below: **Microcanonical ensemble**(NVE) is a thermodynamic state characterized by a fixed number of atoms, N, fixed volume, V and a fixed energy, E. This corresponds to an isolated system. **Canonical ensemble**(NVT) represents a collection of all systems whose thermodynamic state is characterized by a fixed number of atoms, N, a fixed volume, V, and fixed temperature, T. Grand canonical ensemble (mVT) is the thermodynamic state characterized by a fixed chemical potential, m, a fixed volume and a fixed temperature. The ensemble averages are given by

$$\langle A \rangle_{ensemble} = \int \int d\mathrm{p}^N d\mathrm{q}^N A\left(\mathrm{q}^N, \mathrm{p}^N\right) \rho\left(\mathrm{q}^N, \mathrm{p}^N\right) \tag{2.22}$$

where $A\left(\mathrm{q}^N, \mathrm{p}^N\right)$ is the experimental observable quantity of interest and $\rho\left(\mathrm{q}^N, \mathrm{p}^N\right)$ is the probability density function defined by (2.20). In practice all these integral equations are extremely difficult to solve because all possible states of the system must be

calculated. So far all discussion has not taken time into account, a concept which is inevitable in every MD simulations and is the topic of the next section.

2.8 Time Averages and Ergodicity

In molecular dynamics simulations, the *thermodynamic observable* is usually associated with a state function defined $f : \Gamma \longrightarrow \Re$ on Γ which characterize the thermodynamic states of the system. Notable exceptions are temperature and entropy which need the invariant probability distribution over the phase space (*microstates*) in order to be properly introduced. This has been discussed in preceding sections of this chapter.

Lemma 2.1. *Given an initial condition $\zeta = (\mathbf{q}, \mathbf{p}) \in \Gamma$,for Hamiltonian equations of motion, (2.15) there exist a unique solution*

$$\zeta(t) = T^t\zeta = (\mathbf{q}(t), \mathbf{p}(t))$$

for all $t \in \mathbb{R}$ and the compact set $\mathcal{O}_\zeta = \{\zeta(t) | t \in \Re, \zeta(0) = \zeta\}$ is the orbit of ζ in phase space.

Compactness of $\mathcal{O}_\zeta$ is critical in order to reproduce all constants of motion (e.g. energy, $H(\zeta(t))$ as functions of time. See texts in dynamical systems (e.g. [4]) for proof of the lemma. Given a phase space function, f corresponding to a macroscopic physical quantity, precise measurement values $f(\zeta(t))$ are almost infeasible since detailed knowledge of positions and momenta of the particles of the system would be necessary. It is therefore supposed that the results of a measurement is the time average of f. Each measurement of a macroscopic observable at time, t_0, usually takes time to be realized, the microscopic state variable $\zeta(t)$ changes and so different values of $f(\zeta(t))$ are generated and the time average:

$$\frac{1}{t}\int_{t_0}^{t_0+t} f(T^s\zeta)\, ds, \tag{2.23}$$

emerge as "constant"(*i.e* independent of t_0 and t). This macroscopic interval of time for the measurement is extremely large from the microscopic point of view, and so it is practical to take limit $t \longrightarrow \infty$ in (2.23):

$$f^*(\zeta) = \lim_{t\longrightarrow 0} \frac{1}{t}\int_{t_0}^{t_0+t} f(T^s\zeta)\, ds. \tag{2.24}$$

It must be noted that this limit does not depend on the initial time t_0 and the system "visits" all open sets of the phase space Γ during the measurement process for longer microscopic time intervals. This is the benchmark of Ergodicity applied in molecular

dynamics and its clearly expressed in the lemma given below. For measurable state function f, the limit in (2.24) coincides with the average value of f over Γ, defined by:

$$\langle f \rangle = \int_{\Gamma} f(\Gamma) d\zeta. \tag{2.25}$$

where f is assumed to be integrable and measurable which leads to the famous Ergodic theorem.

Lemma 2.2. *If particle motion is restricted to a bounded domain, then for many initial conditions there exist an ensemble (or probability measure) such that the time-average value of the observable equals an ensemble average:* $f^*(\zeta) = \langle f \rangle$,

The lemma implies that when the system is allowed to evolve in time indefinitely, it will eventually pass through all possible states. One goal, therefore, of a molecular dynamics simulation is to generate enough representative conformations such that this theorem holds. Equation (2.24) is comparable to eq. (2.22) and is written as:

$$\langle A \rangle_{time} = lim_{\tau \longrightarrow \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A\left(\mathbf{q}^N(t), \mathbf{p}^N(t)\right) dt \sim \frac{1}{M} \sum_{t=1}^{M} A\left(\mathbf{p}^N, \mathbf{q}^N\right) \tag{2.26}$$

Here t is the simulation time, M is the number of steps in the simulation and $A\left(\mathbf{p}^N, \mathbf{q}^N\right)$ is the instantaneous observable value of A.

2.9 The Algorithm for Molecular Dynamics

Molecular Dynamics (MD) allows for the computing of equilibrium and non-equilibrium properties of the systems, which obey the laws of classical physics. It is especially a good idea to apply MD simulations when characterizing a system with timescale $\tau > \tau_q \simeq$ 0.2ps. The essence of the simulation is the use of computer to model a physical system. Calculations implied by a mathematical model are carried out by the 'machine' and the results are interpreted in terms of physical properties. Since computer simulation deals with models it may be classified as a theoretical method. On the other hand, physical quantities can in a sense be measured on a computer, justifying the term *'computer experiment'*. [20] The implementation of an MD as compared to a real experiment is illustrated in figure 2.3: In this section, we attempt to present the core algorithm of MD.

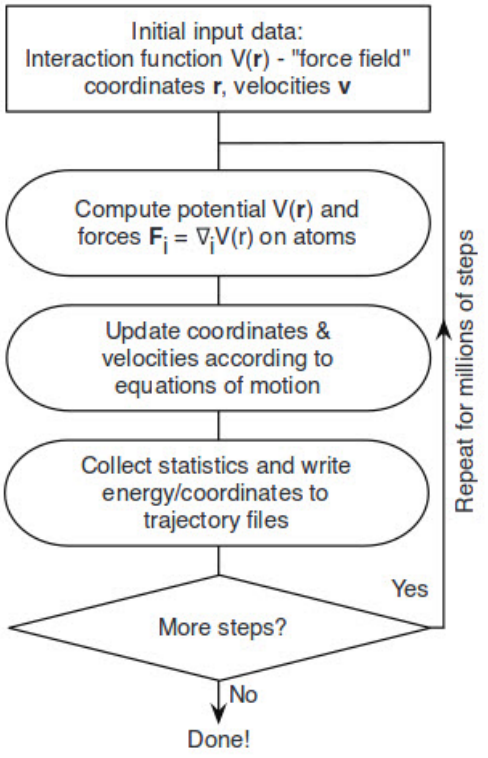


FIGURE 2.3: Main components of the execution of MD simulations

2.9.1 Initial Velocity Distribution

All MD integrators, require some initial generation of velocity assigned to the starting atomic coordinate. It is important that this conformation does not include atom overlaps, unusual local conformations which may result in large forces usually causing instability in MD integration. In almost all cases, the starting conformation is derived by experimental procedures such as X-ray diffraction and NMR spectroscopy. To initial MD integrators, initial velocity distribution is first drawn from the Maxwell-Boltzman distribution:

$$P\left(\nu_{i,\alpha}\right) = \left(\frac{m}{2\pi k_B T}\right)^{\frac{1}{2}} \exp^{\frac{m\nu_{i,\alpha}^2}{2k_B T}} \tag{2.27}$$

where $\nu_{i,\alpha}$ is the $\alpha(= x, y, z)$ component of the velocity of atom i . The distribution is used to define the instantaneous temperature $T(t)$ via the equipartition theorem which states that *the energy of a molecular system is shared equally among all energetically accessible degrees of freedom (or accessible modes of motion) of the system.* Specifically each quadratic degree of freedom will on average possess an energy of $\frac{1}{2}kT$, where T is the temperature. See the book of [26] for derivation and proof of this theorem and section 2.7 for brief discussion. Since the kinetic and potential energy has a quadratic dependency on the velocity and position respectively, it relates the energy of the system

via the equation given below:

$$\langle \frac{m\nu_\alpha^2}{2} \rangle = \frac{1}{2} k_B T, \tag{2.28}$$

All terms have the usual meaning ($\langle \ldots \rangle$ means ensemble averages). Because the ensemble average corresponds to the average over all velocity of atoms, the instantaneous temperature, *T(t)* can be defined:

$$k_B T(t) = \frac{1}{N_f} \sum_{i,\alpha} m\nu_{i,\alpha}, \tag{2.29}$$

N_f is the number of degrees of freedom. Therefore the velocities are generated stochastically and the instantaneous temperature doesn't coincide with the initial temperature T but must be controlled through other mechanisms.
A general practice is velocity re-scaling via:

$$\nu'_{i,\alpha} = \sqrt{\frac{T}{T(t)}} \nu_{i,\alpha} \tag{2.30}$$

It can be shown that the temperature scales with the temperature rescaling with $T'(t) \cong T$ otherwise the relative fluctuations of temperature in the system of N atoms is given by:

$$\frac{\Delta T(t)}{\langle T(t) \rangle} = \frac{\sqrt{(\langle T^2(t) \rangle - \langle T(t) \rangle^2)}}{\langle T(t) \rangle} \sim \sqrt{N} \tag{2.31}$$

In a similar fashion, pressure is also controlled during MD simulation runs.

2.9.2 Integration Algorithms

As outlined earlier, the main goal of a classical molecular dynamics simulation programs is to solve the Newton's equation of motion to write atomic coordinates and velocities of the particles as function of time:

$$m_i \dot{\nu_{i,\alpha}} = -\frac{\partial U(\mathbf{R})}{\partial r_{i,\alpha}} \text{ for } i = 1, \ldots, N \text{ and } \alpha = (x, y, z) \tag{2.32}$$

To discretize and numerically solve this IVP, a time step $\delta t > 0$ is chosen and the sampling point sequence $t_n = n\delta t$ considered. The main objective here would be to construct a sequence of points $\mathbf{R}_n$ that closely follow the points $\mathbf{R}(t_n)$ on the trajectory of the exact solution. There are several numerical and discretization schemes that have been developed for this problem at different order of numerical accuracy namely: Euler, Verlet, Leap frog algorithms. In this section we will discuss the derivation of the Verlet algorithms as it is the most widely used scheme in many MD programs.

Verlet algorithm

Given a time-dependent atomic coordinate $r(t) \equiv r_{i,\alpha}$, the forward and backwards Taylor expansion in time gives:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{F(t)}{2m}\delta t_2 + \dddot{r}(t)\delta\frac{\delta t_3}{3!} + \mathcal{O}(\delta t_4) \tag{2.33}$$

$$r(t-\delta t) = r(t) - v(t)\delta t + \frac{F(t)}{2m}\delta t_2 - \dddot{r}(t)\delta\frac{\delta t^3}{3!} + \mathcal{O}(\delta t^4) \tag{2.34}$$

summing up equations (2.33) and (2.34), the positions for the next time step δt is obtained:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + \frac{f(t)}{2m}\delta t^2 + \mathcal{O}(\delta t^4) \tag{2.35}$$

Thus the new position at $t + \delta t^4$ is approximated at order of $\mathcal{O}(\delta t^4)$. However it must be noted that (2.35) does not use the velocity to compute new position coordinates but can be derived from (2.35) itself and is given by:

$$v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{2\delta t} + \mathcal{O}\delta t^2 \tag{2.36}$$

It must be noted that from (2.35) and (2.36) the kinetic energy and thus instantaneous temperatures (see **??** for more) at time t cannot be calculated until the positions are known at next time step $t+\delta t$. In addition, since the interactions and forces between the atoms in the system must be taken into account, the total nergy and momentum must be conserved. During a typical MD simulations there must be a mechanism to explicitly check these conditions. Other methods such as Leap frog and Velocity Verlet algorithms are developed to cater for this computational delay deficiency. A typical implementation of (2.35) and (2.36) is given below.

```
subroutine integrate(f,en)      !subroutine to integrate equation of motion
sumv=0
sumv2=0
do i=1,npart            !loop over all particles
        do a=1,3                !loop over the 3D cartesian coordinates
                rr=2*r(i,a)-rm(i,a)+dt**2*f(i,a)/m(i)   !Verlet algorithm
                v=(rr-rm(i,a))/(2*dt)   !velocity equation
                sumv=sumv+v     !velocity center of mass
                sumv2=sumv2+m(i)*v**2   !total K. E
                rm(i,a)=r(i,a)  !update postions for previous time
                r(i,a)=rr       !update current time
        enddo
enddo
temp=sumv2/(3*npart*kb)         !instantaneous temperature
```

```
etot=(en+0.5*sumv2)/npart        !average total energy per particle
return
end subroutine
```

2.10 Minimization of the Potential Function

Before starting a MD simulation, it is necessary to allow the initial structure to reach a locally stable energy configuration. At that energy, the first derivative of the potential function 0 and the second derivative is positive (i.e. $\frac{\partial U_p}{\partial \mathbf{r}} = 0$ and $\frac{\partial^2 U_p}{\partial \mathbf{r}^2} > 0$ were $\mathbf{r}$ is the atomic coordinates.) The set of atomic coordinates that satisfy these equations are saved and adapted as the energetically favourable starting configuration for equilibration and production runs. This is a very important step in MD simulation, but very complex problem to solve for macromolecules, because of the large number of atoms and interactions involved. In particular it is very difficult to explore large areas of conformational space by crossing saddle point barriers(or energy barriers). However, various mathematical optimization algorithms have been developed to address this problem at different levels of approximation. The conjugate gradient and steepest descent algorithms are among the most commonly used in structural analysis of macromolecules. They involve the minimization of a quadratic cost function (energy potential), subject to some constraint pathways or not for a specified *force-fields*. Thus given an N-dimensional system with the potential energy functional $U(r) : \mathbb{R}^N \longrightarrow \mathbb{R}$, defined over all atomic coordinates, $r \in \mathbb{R}^N$, the optimization algorithms seeks to find

$$\min U(r),$$

provided $U(r)$ is a continuous and differentiable function with $g(r) = \nabla U(x)$ as it's gradient. Steepest descent and conjugate gradient algorithms seeks to generates a sequence $r^0, r^1, r^2, r^3, \ldots$ such that $U(r^0,) > U(r^1,) > U(r^2,) > U(r^3,) > \ldots$ In general these algorithms takes the form:

$$r^{k+1} = r^k + \alpha^k d^k \tag{2.37}$$

here $\alpha^k > 0$ is a step length and d^k is a search direction for a kth guess. In order to converge these algorithms, α^k must be chosen to satisfy certain conditions, like the Wolfe line search conditions for conjugate gradient algorithms:

$$U(r^k + \alpha^k d^k) - U(r^k) \leq \sigma_1 \alpha^k (g_k^T) d_k) \tag{2.38}$$

and

$$\nabla U(r^k + \alpha^k d^k)^T d^k \geq \sigma_2 (g_k^T) d^k \tag{2.39}$$

where $0 < \sigma_1 \leq \sigma_2 < 1$. Different versions of conjugate gradient and steepest descent algorithms are implemented in all MD simulation packages. These algorithms are sufficient enough to find a local minima for molecular topologies supplied to it. The energy minimization algorithms implemented in AMBER software package is run by specifying the IMIN flag to 1 or 2 without performing any molecular dynamics.

A rather more rigorous scheme called simulated annealing is also commonly used to find an initial stable structure. It is based on the sequential 'heating' and 'cooling' of the system. This practice is widely used in experimental procedures in the prediction of proteins. In a typical simulated annealing implementation, the temperature of the system is raised to an enormous temperature and cooled down abruptly during the dynamics. The working principle is that by raising the temperature, there's a high chance to overcome the energy barrier to sample out the stable configuration in the final step. Simulated annealing has a strong theoretical formulation in statistical mechanics and combinatorial mathematics[35]. Several MD packages have diverse molecular dynamics protocols to implement simulated annealing with acceptable accuracy. In this work

2.11 Analysis of Molecular Dynamics Result

In order to demonstrate the accuracy of the integration methods, it is usually useful to carry out a series of performance indication runs with several time step values. Some of these indicators are, the temperature, radius of gyration, root mean square deviation (RMSD), Mean square fluctuation and Debye-Waller factors, Diffusion coefficients, among others. In this section we discuss only a selected few which is used in this work. These frameworks are not entirely new but have their theoretical backings in mathematical statistics.

2.11.1 Root Mean Square Deviations (RMSD)

The RMSD follows the evolution of the structure during the simulation. It is defined by:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_i^{\text{ref}})^2} \tag{2.40}$$

where $\mathbf{r}_i^{ref}$ are the coordinate vectors of the reference structure (usually, the starting structure, used for the simulation) and N is the total number of atoms. The RMSD plotted during the trajectory of a structure as it is allowed to equilibrate at a constant temperature should a plateau value in times of the order of 100ps.

2.11.2 Radius of Gyration

The radius of gyration,R_G is a useful parameter to define the overall conformation of the system. It can be calculated in two steps. First the center-of-mass coordinates, $\mathbf{R}_C$, are obtained from

$$\sum m_i(\mathbf{r}_i - \mathbf{R}_c) \tag{2.41}$$

and then the radius of gyration is related to the moment of inertia and defined by:

$$R_G^2 = \frac{\sum m_i(\mathbf{r}_i - \mathbf{R}_C)}{M} \tag{2.42}$$

where M is the total mass of the system It is also possible to obtain R_G from the system coordinates, without going through the center-of-mass calculations, from:

$$R_G^2 = \frac{\sum\sum m_i m_j(\mathbf{r}_i - \mathbf{r}_j)^2}{M^2} \tag{2.43}$$

Chapter 3

METHODOLOGY: THE MODELLING AND SIMULATIONS

Introduction

In this chapter, the main results obtained are presented. The chapter begins by outlining the techniques and procedures used which has a strong theoretical backing from the formulations discussed in the previous chapter. In the final sections we discuss the relevance of the work and lay foundation for future works to be done on the PSI complex.

3.1 Preparation of Initial Structure

The initial structure used for all simulations was taken from the crystal structure of PSI complex of the thermophylic cyanobacterium *Synechococcus elongatus* crystalized at 2.5Å[17] by X-ray diffraction experimental method. The structure described therein provides atomic details of 12 protein subunits and 127 cofactors. Specifically they comprise of 96 chlorophylls, 2 phylloquinones, 3 Fe_4S_4 clusters, 22 β-carotenoids, three 1,2 dipalmitoylphosphatidylglycerole (LHG), one 1,2 distearoylmonogalactosyldiglyceride (LMG) (these two co-factors would be referred to as lipids for the remaining part of the text), a putative Ca^{2+} ion and 201 naturally occurring crystal water molecules. As discussed in chapter 1, PSI can exists either as a monomer or trimer or both. For

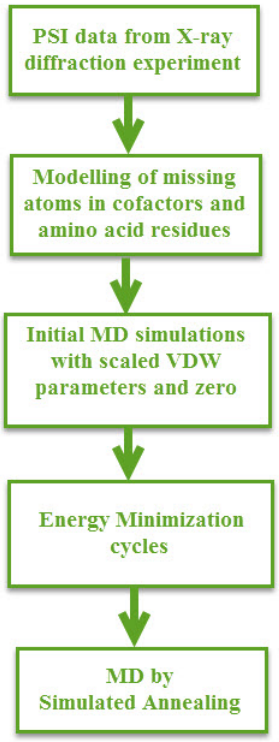


FIGURE 3.1: flow chart of the implementations used in this work

the purpose of our study we consider just the monomer reported above. The entry corresponding to this description from the Protein Data Bank is **1JB0**.

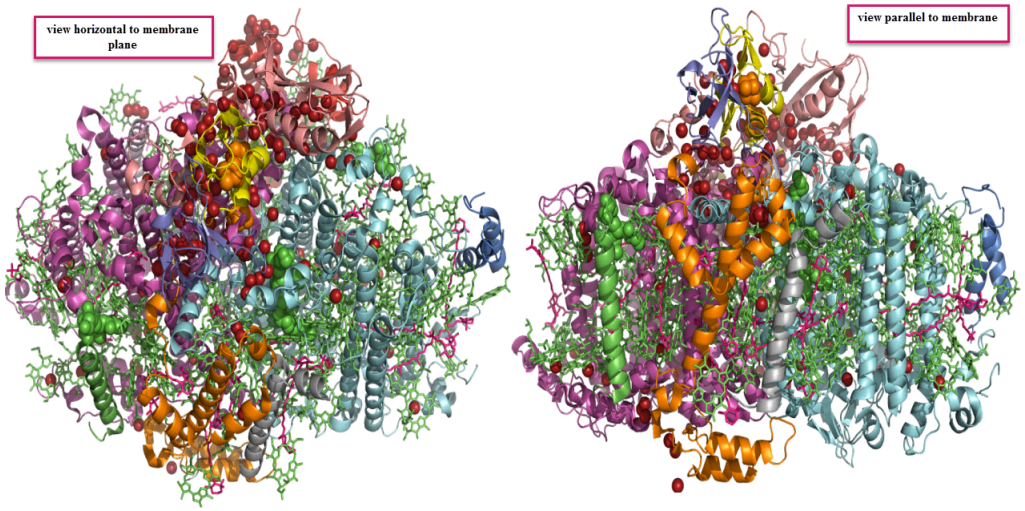


FIGURE 3.2: visualization of PSI as obtained from the protein data bank. View on the left is horizontal to the thylakoid membrane while view on the right is parallel to the thylakoid membrane.

Photosystems I protein complex is particularly interesting yet challenging protein complex system to model. So far the structure we have adopted, cyanobacterial PSI, from

Jordan *et al* (2001) represents the highest resolution X-ray crystallization at 2.5Å available in 3D so far. Even though this resolution isn't high enough, it serves as a good starting point for computational studies on the protein complex. That also implies that in order to obtain the Hamiltonian of the system which is to be integrated during simulations, all molecules partially solved during the X-ray experiment has to completely resolved by computational means.

In the monomer of PSI, there was a large space of positions with unassigned atomic coordinates. There are 35 amino acids with missing atoms at the N-terminals and C-terminal of the residues. In addition, ∼91 amino acids residues were reported as missing and were not located during experiment. Most of these residues are expected to complete the protein sequencing, therefore their absence creates a gap in the intramolecular interaction and could lead to huge forces during MD integration. For example the the protein subunit PsaK at residues numbered 44-54 contains up to 10 amino acid residues not completely resolved during X-ray crystallisation (see figure below). This was

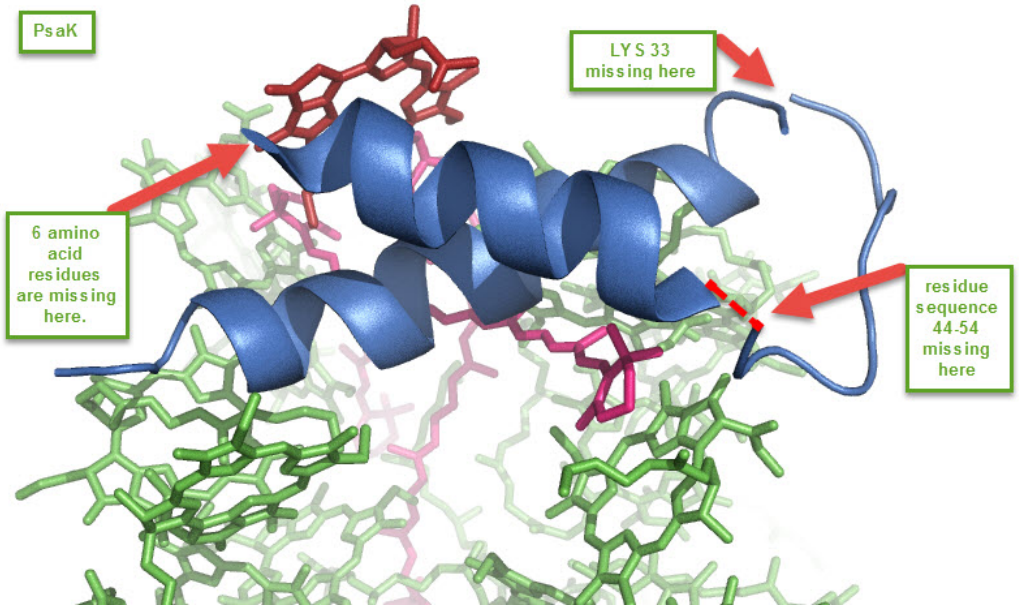


FIGURE 3.3: visualization of PsaK subunit from X-ray data

successfully complete modelled in this work by the MODELLER[15] and GROMACS molecular modelling software package and results shown on page 32 .

Moreover, some of non-standard molecules were also reported to have missing atoms. In total, 52 cofactors were reported to have missing atoms. The breakdown is as follows: 1 $\beta-$carotenoid BCR4009(B), two lipids LHG5003(A) and LHG5004(B) and 49 chlorophylls. We have successfully dealt with this with the aid of the XLEAP utility of AMBER12. For each of these molecules, we first prepare the topology of a full molecule and then supplied this input to XLEAP which automatically completes the sections of the

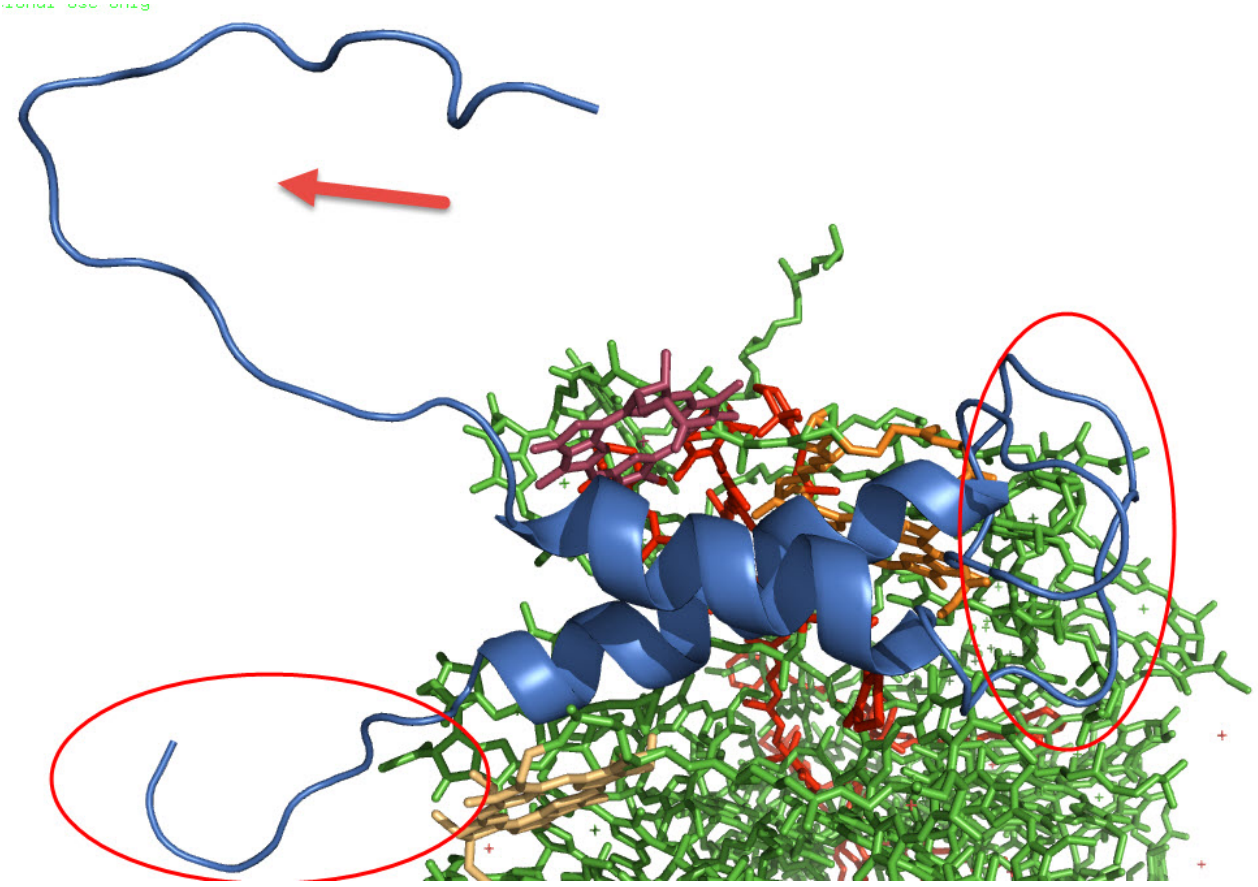


FIGURE 3.4: visualization of PsaK subunit after adding missing amino acid residues

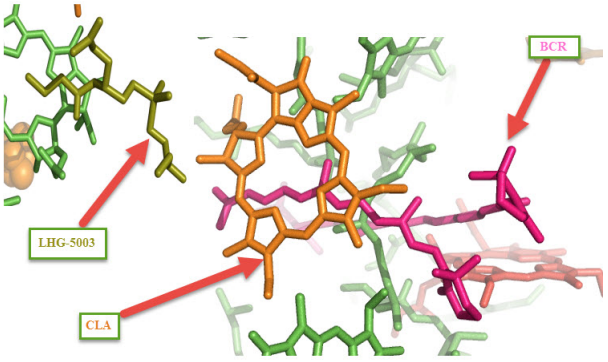


FIGURE 3.5: visualization of some missing co-factors in PSI with labelling

molecules not completely modelled in X-ray crystallization without considering neighbour atoms. After all structural modelling procedures, atomic count amounts to a total of ∼52700 (including all hydrogen atoms and the naturally occurring water molecules). Since during tries to generates a reasonable coordinates for the missing atoms to complete the structure and without taking into account the electrostatic contribution of it's environment, there is large collection of atoms superimposed either with already existing atoms in the X-ray structure or with the positions of the already completed atoms. This problem was successfully dealt with during our energy minimization.

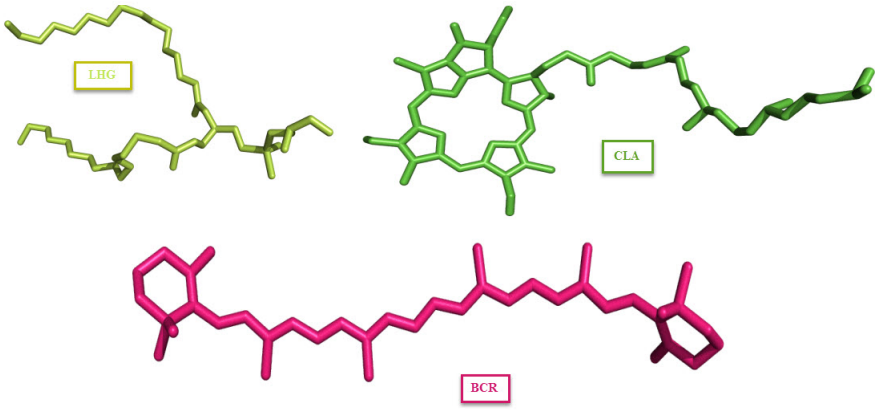


FIGURE 3.6: complete structures of Chlorophyll, $\beta-$carotene, and lipid molecule

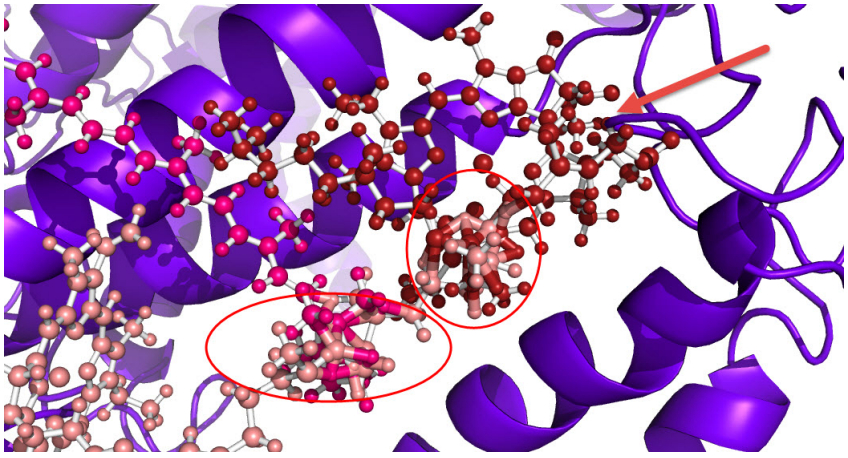


FIGURE 3.7: illustration of superimposition of atomic coordinates after modelling of missing cofactors. BCR is shown in pink-red and other two chlorophylls are shown in red.

3.1.1 Assignment of protonation states

Histidines of PsaA numbered His-215, 300, 396, 461 and of PsaB identified His-192, 275, 377, 442 were considered to be protonated at the delta positions. All other histidines are protonated at the epsilon positions. It must be noted that not all missing amino acids was taken care of in our modelling procedure. For these residues, the missing side chain were immediately terminated with hydrogen atoms. Thus some of the complete missing amino acid residues reported at the terminal positions were ignored.

3.1.2 Charge determination of cofactor molecules

As noted in the previous section, the PSI complex contains 127 non-standard protein amino acids. With regards to the topology preparation for cofactors CLA, BCR, LHG and LGM, we used the *ab initio* calculations previously developed by [11]. However, the

TABLE 3.1: Other protonation states of specific residues considered. [13]

Residue name	Protonation state	Reason
HIS33-A	protonated	On outside
HIS135-A	protonated	On outside, increased H bonding
HIS633-A	protonated	On outside
HIS33-B	protonated	not clear, in hydrophobic region
HIS121-B	protonated	Near ASP367-B
HIS205-B	protonated	Near ASP209-B and ASP133-B
HIS241-B	protonated	close to ASP367-B
HIS368-B	protonated	In hydrophobic region
HIS95-D	protonated	Hydrophobic region
HIS63-E	protonated	On outside
HIS50-F	protonated	Near ASP46-F and GLU459-B
HIS3-J	protonated	Near GLU26-A and LYS30-A

force field topology preparation for PQN and SF4 was separately prepared by utilizing the ANTECHAMBER[36] tool of ABMER12 and GAUSSIAN[16] as outlined below.

First we modelled the electrostatic potential field of these cofactors with GAUSSIAN and used the RESPGEN tool to generate R.E.S.P charges of all heavy atoms of PQN. This was accomplished by assigning partial charges to each atom. In the initial step, the LEAP tool of AMBER12 program was used to assign GAFF [37] and ff99SB force field [23] parameters which account for bond lengths, proper and improper dihedral angles. Using Antechamber, the resulting molecular structure served as the preparation input files for GAUSSIAN calculations. In GAUSSIAN, the molecular structures were first energy-minimized at a density functional level of theory with the Hartree-Fork Self Consistent Field Method (HF-SCF). This approximation is known to provide an accurate description of minimized geometries. To maintain consistency with the AMBER-ff99SB force field, we used the 6-31G* basis set. Partial atomic charges were fitted to reproduce electrostatic potential on this set of grid points by a restrained electrostatic potential fit, as implemented in the RESP program from AMBER12 suite. Atomic charges of the iron sulphide cluster (SF4) were assigned according to the redox state of its atoms in the dark-adapted (S_1) state as follows: S1–S4 (sulphur), -2; Fe1-Fe4 (non-heme irons), +2 non-heme iron was set to +2. We then ensured that equilibrium geometries of the cofactors were accurately reproduced by the force field. We used the PARMCHK tool

of ANTECHAMBER to automate the calculation of any missing force field parameters (bond lengths, angles, and dihedrals). These two structures were successfully added to the PSI complex for subsequent calculations. The net charge of the PSI monomer evaluates after all these was -8.

3.2 Initial MD Simulations

Before the minimization step, a short MD simulation was performed at low temperature coupling of 250K for 20ps and 0.2fs timestep. Due to the high number of missing residues and incomplete chains, it was deemed appropriate to turn off partial atomic charges for all atoms in the system. Particularly, we wanted to avoid system crash due to division by zero during simulations as all atomic van der Waal radii overlapped after completion of the missing atoms and residues with already existing residues. In addition to this, the van der Waal's radii were scaled by about 90 percent for all atoms. The figure on page 35 shows changes in minimum distances between CLAs and protein atoms during the entire minimization cycles. On the figure, Cycle 0 corresponds to the minimum distance after adding missing atoms from the X-ray data. At that point, most of the phytyl tails of the chlorophylls overlapped with some atoms of the proteins. The distance approaches an average value of close to 1.7Åwhich is the average distance for H-bonds.

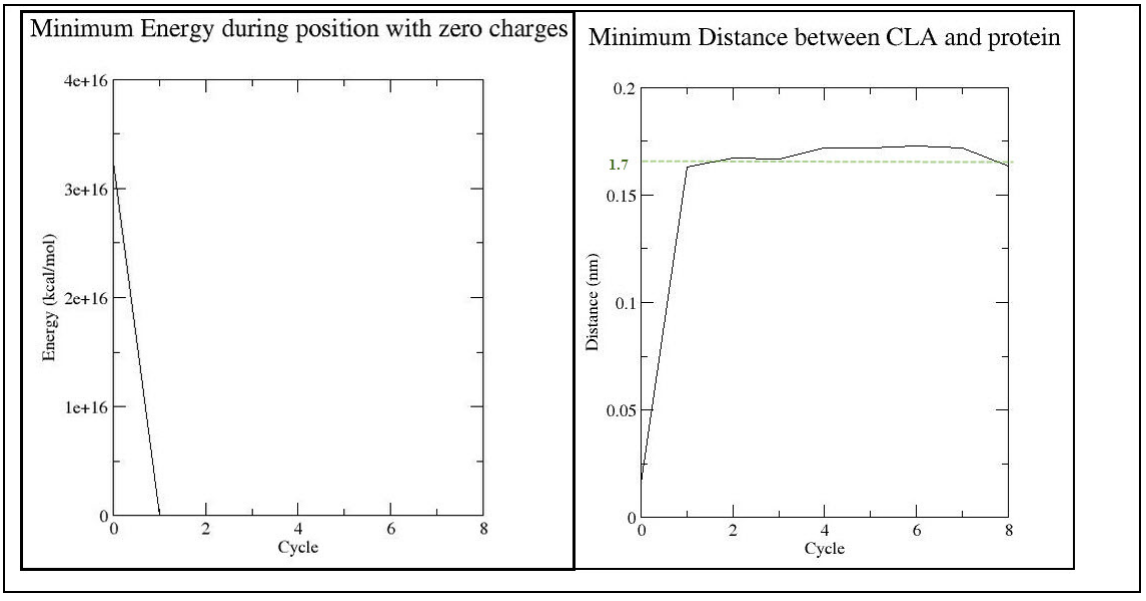


FIGURE 3.8: A plot showing minimum distance between all chlorophyll and protein atoms during minimization cycles.

Energy Minimization(EM) Cycles

Eight cycles of steepest descent and conjugate gradient energy minimization of the modelled or prepared structure was performed in stages with the SANDER program of AMBER12 in vacuum. As was done in the initial MD simulation run, charges were zeroed and VDW parameters were scaled to allow all atoms in an unstable conformation (particularly completed missing atoms) to re-orient to stable conformation during the minimization. With these parameters, the first to third minimization cycles was successfully performed. These cycles were begun with first 500 steps of EM by the steepest descent algorithm without position restraint on all atoms. To maintain consistency with the initial positions for all the tails of the cofactors and proteins, harmonic position restraint was enforced on both the protein and all cofactors. In the second cycle a restraint of 1000Kcal/mol was applied to all the proteins and 150Kcal/mol for all tails of cofactor protein and the Ca^{2+}. This restraint was weakened gradually up to the 6th cycle. However, in the last cycle, no position restraint was applied but the cycle was had to include the chlorin heads to the list of restrained molecules with a force of 150Kcal/mol while the *phytyl* tails allowed to move during the next round of 1000 cycles of energy minimization. After these, the atomic charges of the simulations were then turned on for the simulated annealing simulations. (see figure for full parameters used for simulated annealing). A simulated annealing MD simulations protocol was used to

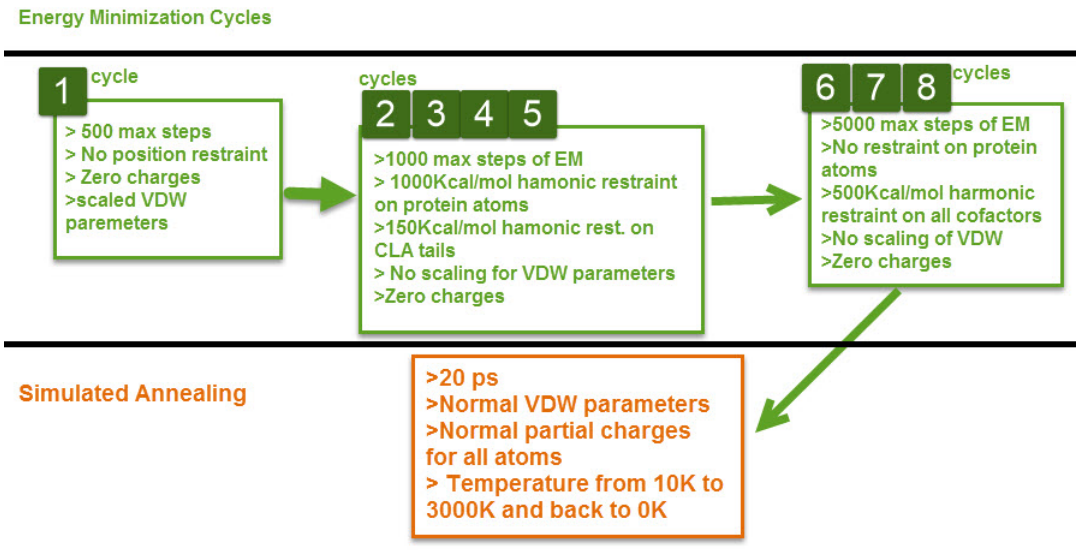


FIGURE 3.9: schematic view of minimization cycles and simulated annealing

find lowest-energy of the complete structure of PSI. The structure used for the rest of the simulations had a final energy minimum at **-98088.9112Kcal/mol*A^2**.

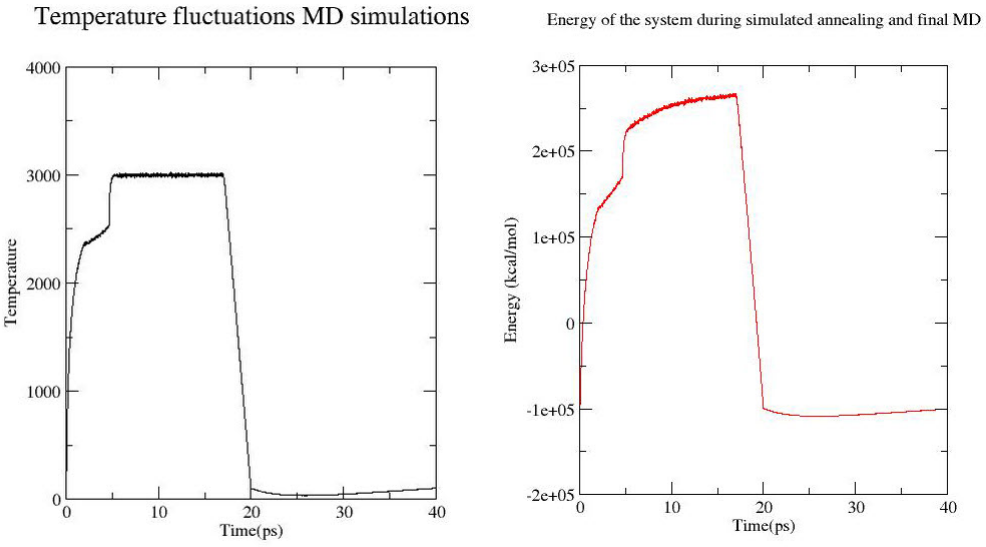


FIGURE 3.10: Energy and temperature fluctuations in simulated annealing MD followed by a final 20ps MD simulations. Energy was stabilized in the last 20ps of the simulations.

3.2.1 MD simulations protocol

The SANDER module of AMBER 12 package was used for all MD simulations at this stage. The simulations were performed using periodic boundary conditions and the SHAKE algorithm in vacuum. Constant pressure simulations (NPT) were run at a time step of 2fs for 20ps. These simulations were carried out at a temperature of 283 K with Berendsen weak temperature coupling. [33]. Long range electrostatics were treated with the Particle Mesh Ewald method and the Lenard-Jones interactions were evaluated with 10.0Å cut-off value. Positional restraints were of 50kcal/(mol*A^2) were applied to the protein backbone and 30kcal/(mol*A^2) for all cofactor molecules. The collected structural data was analysed with the PTRAJ module of AMBER 12 software while the visualization was done with PyMOL visualization software. This optimized structure of PSI complex showed a remarkable stability during the short 20ps MD simulation run as indicated in the RMSD plot in figure on 38. See also figure on 37 to compare with energy fluctuations during the final 20ps MD simulation run. The RMSD approaches an equilibrium of 1.7Å in the final time steps.

3.3 Conclusions

In this work we have successfully provided an optimized structure of the PSI protein complex and tested the structure for short simulation of 20ps in vacuum. The initial structure was adapted from the 3D crystal structure solved by Jordan *et al* (2001)

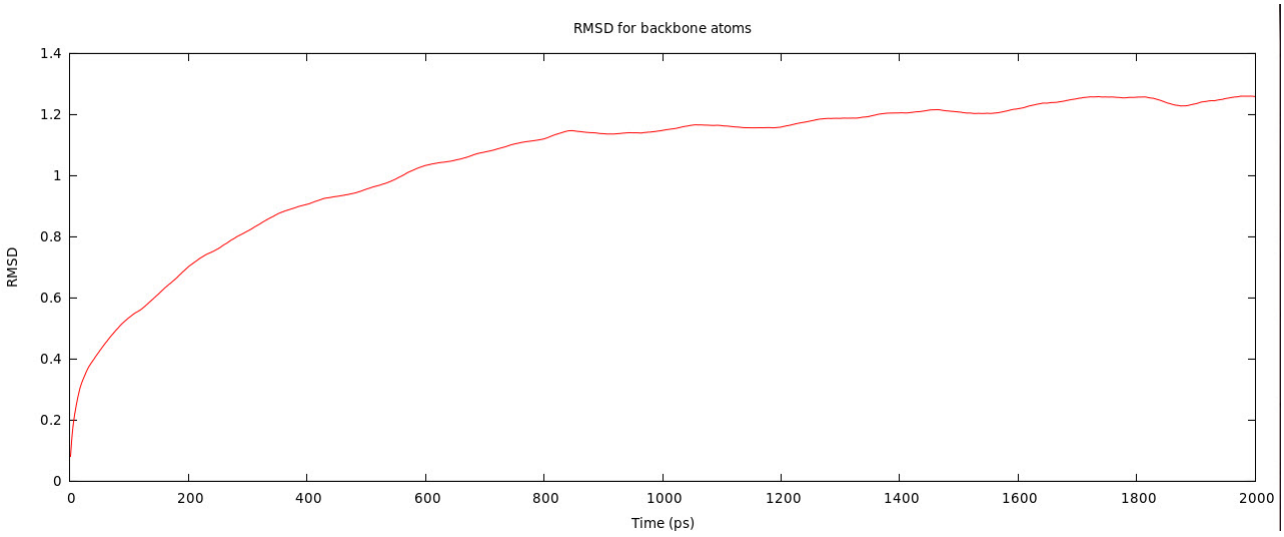


FIGURE 3.11: RMSD plot on protein backbone atoms during final MD simulation run

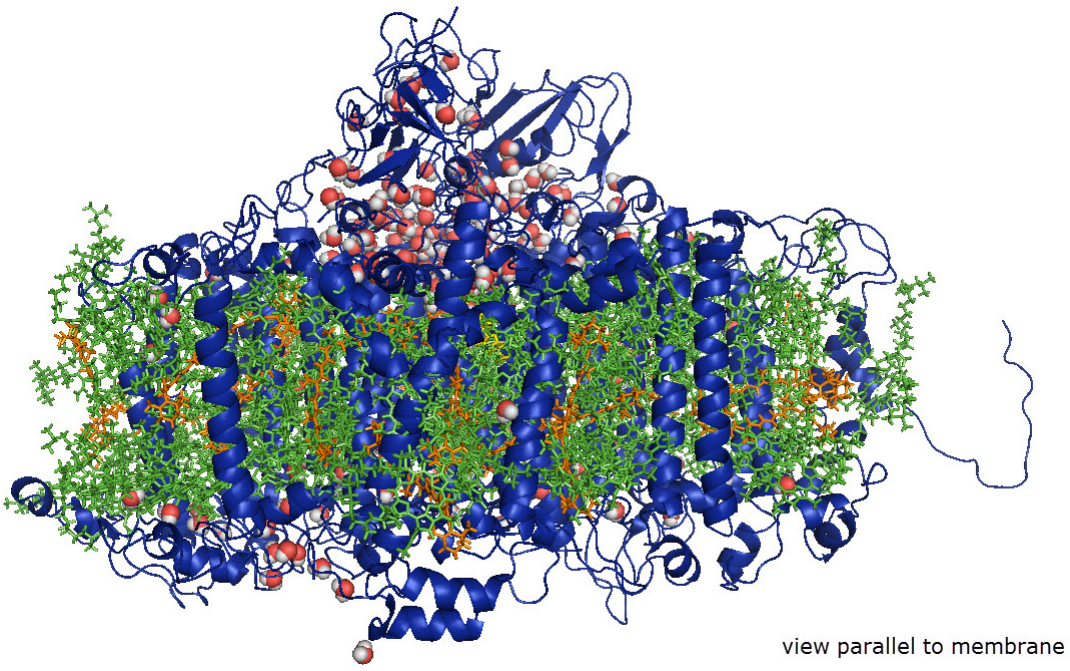


FIGURE 3.12: Our minimized structure of PSI after minimization, simulated annealing and MD simulation.

at 2.5Å using X-ray diffraction. The low resolution implied that several atoms and molecular fragments were not completely solved during the experiment. However, using sophisticated techniques, we have completed all these missing atoms and obtained an optimized structure of this huge protein complex. The final structure was showed remarkable stability during our MD simulation runs laying foundation for further production runs in physiological environment. It is our hope that this structure can serve as starting structure for QM/MM equilibration to incorporate most of the active sites accessible at short time scales.

Bibliography

[1] *Introduction to Quantum Mechanics.* McGraw-Hill Book Company, Inc, 1935.

[2] *Classical Mechanics Systems of Particles and Hamiltonian dynamics.* Verga Harri Deutsch, 1989.

[3] *Density Functional Theory of Atoms and Molecules.* Oxford University Press, 1989.

[4] *Differential Equations and Dynamical Systems.* Springer, 2000.

[5] *The Art of Molecular Dynamics Simulations.* Cambridge University Press, 2004.

[6] *Plant Physiology.* 2007.

[7] *Photosynthetic Protein Complex.* Wiley-Blackwell, 2008.

[8] *Molecular Modeling and Simulation.* Springer Science, 2010.

[9] *Natural and Artificial Photosynthesis.* John Wiley and Sons Inc, 2013, ch. 2.

[10] Alder, B. J., and Wainwright, T. E. Studies in molecular dynamics i general methods. *Journal of Chemistry and Physics* (1959).

[11] Bovi, D., Nazi, D., and Guidoni, L. Magnetic interactions in the catalyst used by nature to split water: a DFT+ U multiscale study on the Mn_4Cao_5 core in photosystem ii. *New Journal of Physics IOPscience* (2014).

[12] Buda, F. Introduction to theory/modeling methods in photosynthesis. *Photosynthesis Research* (2009).

[13] Canfield, P., and Reimers, R. Density-functional geometry optimization of the 150000-atom photosystem-i trimer. *The Journal of Chemical Physics* (2006).

[14] Car, R., and Parrinello, M. Unified approach for molecular dynamics and density function theory. *Physical Review Letters* (1985).

[15] Fiser, A., and Sali, A. Modeller: generation and refinement of homology-based protein structure models. *Meth. Enzymol.* (2003).

[16] Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Montgomery, Jr., J. A., Vreven, T., Kudin, K. N., Burant, J. C., Millam, J. M., Iyengar, S. S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G. A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J. E., Hratchian, H. P., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Ayala, P. Y., Morokuma, K., Voth, G. A., Salvador, P., Dannenberg, J. J., Zakrzewski, V. G., Dapprich, S., Daniels, A. D., Strain, M. C., Farkas, O., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Ortiz, J. V., Cui, Q., Baboul, A. G., Clifford, S., Cioslowski, J., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Gonzalez, C., and Pople, J. A. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[17] Fromme, P., and Jordan, P. Three-dimensional structure of cynobacterial photosystem i at 2.5åresolution. *Nature* (2001).

[18] Grotjohann, I., and Fromme, P. Structure of cyanobacterial photosystem i. *Photosynthesis Research* (2005).

[19] H, C. *Model Reduction in Classical Molecular Dynamics*. PhD thesis, Free University Berlin, 2007.

[20] Jarosaw, M. Molecular dynamics. *Encyclopedia of Life Science* (2001).

[21] Jolley, C. C. *Structure and Dynamics in Photosystem I.* PhD thesis, Arizona State University, 2007.

[22] Kendrew, J., and Philips, D. C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature Science Journal* (1958).

[23] Lindroff-Larsen, K., Piano, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Wiley InterScience* (2010).

[24] Mackerell, A. D. Empirical force fields for biological macromolecules: Overview and issues. *J Comput Chem* (2004).

[25] Martin, K., and McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature Publishing Group* (2002).

[26] McQuarie, D. A. *Statistical Mechanics, Second Edition.* Harper and Row, New York, Evansto, San Fransisco, London, 2000.

[27] Mu, Y., Daniil, S. K., and Stock, G. Conformational dynamics of trialanine in water. 2. comparison of amber, charmm, gromos, and opls force fields to nmr and infrared experiments. *J. Phys. Chem.* (2003).

[28] Nelson, N., and Yocum, C. F. Structure and function of photosystems i and ii. *Annual Review of Plant Biology* (2006).

[29] Neugebauer, J. Photophysical properties of natural light-harvesting complexes studied by subsystem density functional theory. *J. Phys Chem. B* (2008).

[30] Perutz, M. F. X-ray analysis of haemoglobin. Tech. rep., Noble Prize Lecture, 1962.

[31] Petra, F. Unraveling the photosystem i reaction center: a history, or the sum of many efforts. *Photosynthesis Research* (2004).

[32] Ryckaert, J. P., and Bellemans, A. Molecular dynamics of liquid n-butane near its boiling point. *Chemical Physics Letters* (1975).

[33] Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. Numerical integration of the cartesian equations of motions of a system with constraints: Molecular dynamics simulation of n-alkanes. *J. Comput. Phys.* (1977).

[34] Scholes, D. G., Fleming, R. G., Olaya-Castro, A., and Riankvan, G. Lessons from nature about solar light harvesting. *Nature Chemistry* (2011).

[35] Vecchi, M. P., Gelatt, C. D., and Kirkpatrick, S. Optimization by simulated annealing. *Science* (1983).

[36] Wang, J., Wang, W., Kollman, P. A., and Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* (2006).

[37] Wang, J., Wolf, R. M., Caldwell, J. W. K., Kollman, P. A., and Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* (2004).

[38] Wang, W., and Kollman, P. A. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-protein, and protein-nucleic acid non-covalent interactions. *Ann Rev Biophys Biomol Struct. 30* (2001), 211–43.