THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING
SCHOOL OF BIOLOGICAL SCIENCES

A Molecular Dynamics Study of the Allosteric Control Mechanisms of the Glycolytic Pathway

A Dissertation submitted for the degree of

Doctor of Philosophy

by Ankita Naithani

Structural Biochemistry Group

Institute of Structural and Molecular Biology

School of Biological Sciences

University of Edinburgh

Scotland

United Kingdom

October 2014

Declaration

The work presented in this dissertation is the original work of the author. This dissertation has been composed by the author and has not been submitted in whole or in part for any other degree.

Ankita Naithani

This thesis is dedicated to my guru Maharajji His Holiness "Shree Shree Swami Keshavanand Puri Ji Maharaj Ji"

Degree of Doctoral of Philosophy – The University of Edinburgh - 2014

UNIVERSITY OF EDINBURGH

ABSTRACT
COLLEGE OF SCIENCE AND ENGINEERING
SCHOOL OF BIOLOGICAL SCIENCES

Doctor of Philosophy

A Molecular Dynamics Study of the Allosteric Control Mechanisms of the Glycolytic Pathway

by Ankita Naithani

There is a growing body of interest to understand the regulation of allosteric proteins. Allostery is a phenomenon of protein regulation whereby binding of an effector molecule at a remote site affects binding and activity at the protein's active site. Over the years, these sites have become popular drug targets as they provide advantages in terms of selectivity and saturability. Both experimental and computational methods are being used to study and identify allosteric sites. Although experimental methods provide us with detailed structures and have been relatively successful in identifying these sites, they are subject to time and cost limitations.

In the present dissertation, Molecular Dynamics Simulations (MDS) and Principal Component Analysis (PCA) have been employed to enhance our understanding ofallostery and protein dynamics. MD simulations generated trajectories which were then qualitatively assessed using PCA. Both of these techniques were applied to two important trypanosomatid drug targets and controlling enzymes of the glycolytic pathway - pyruvate kinase (PYK) and phosphofructokinase (PFK).

Molecular Dynamics simulations were first carried out on both the effector bound and unbound forms of the proteins. This provided a framework for direct comparison and inspection of the conformational changes at the atomic level. Following MD simulations, PCA was run to further analyse the motions. The principal components thus captured are in quantitative agreement with the previously published experimental data which increased our confidence in the reliability of our simulations. Also, the binding of FBP affects the allosteric

mechanism of PYK in a very interesting way. The inspection of the vibrational modes reveals interesting patterns in the movement of the subunits which differ from the conventional symmetrical pattern. Also, lowering of B-factors on effector binding provides evidence that the effector is not only locking the R-state but is also acting as a general heat-sink to cool down the whole tetramer. This observation suggests that protein rigidity and intrinsic heat capacity are important factors in stabilizing allosteric proteins. Thus, this work also provides new and promising insights into the classical Monod-Wyman-Changeux model of allostery.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

3D : Three-dimensional

ADP: adenosine diphosphate

AMBER: Assisted Model Building with Energy Refinement

ANM: anisotropic network model

ATP: adenosine triphosphate

CN: contact number

CO: cumulative overlap

DOF: degree of freedom

ENM: elastic network model

F26BP: fructose-2,6-bisphosphate

F16BP:fructose-1,6-biphosphate

FBS: fragment-based screening

FF: force-field

GNM: Gaussian network model

GROMACS: GROningen MAchine for Chemical Simulations

HPC: High Performance Computing

KNF:Koshland–Nemethy–Filmer

lmPYK: Leishmania mexicana Pyruvate Kinase

MD: molecular dynamics

MDS: Molecular Dynamics Simulations

MWC: Monod–Wyman–Changeux

NMA: normal model analysis

NMR: nuclear magnetic resonance

OXL: Oxalate

PBC: Periodic Boundary Conditions

PBS : Parallel Batch System

PC: principal component

PCA: principal component analysis

PDB: protein data bank

PFK: Phosphofructokinase

PME: particle mesh ewald

PYK: Pyruvate Kinase

RMSD: Root Mean Square Deviation

RMSF: Root Mean Square Fluctuation

RMSIP: root mean-square inner product

SPC: Simple Point Charge

SVM: support vector machine

TTS: tertiary two-state

VMD: Visual Molecular Dynamics

SOFTWARES AND RESOURCES

ARCHER

Chemsketch (reaction figures)

Endnote

GROMACS v4.4.5

GROMACS v4.6

HecToR

Linux system

Marvinsketch (chemical structures)

MATLAB (to plot the eigenvector graphs and 2D structures)

Phostoshop

Powerpoint (pictures)

Pymol (structure figures)

VMD (movies and graphics)

CCP4

Windows8 Operating System

CHAPTER 1

INTRODUCTION

1. Introduction

"In the drama of life on a molecular scale, proteins are where the action is".

A. M. Lesk,
Introduction to Protein Architecture[4]

1.1 Proteins

All living cells are composed of proteins which are a complex multi-particle system constituted by the connection of several building blocks, called amino acids. Chemically, a protein is a homogenous class of organic molecule composed of secondary structures like the helices, beta strands, and loops that form the ensemble. Each protein is dedicated to performing a specific task which ranges from serving as a supporter for other molecules or essential breakdown of food components to the synthesis of new molecules required for functioning. The structural elements are arranged in a thermodynamically favourable conformation, variants of which provide flexibility to the protein structure. Proteins contain a large number of restraints such as covalent bonds, steric interactions, and hydrogen bonds which limit the available configurational space. Despite these restraints, proteins adopt different conformations and retain their flexibility while performing their functions. For example, in the case of ligand binding, the different conformations of binding pockets provide different levels of accessibility for ligand binding. Thus, every task has its own specific conformations which make the understanding of conformational changes in proteins a prerequisite to understanding the biochemical processes on an atomic level.

This nature of flexibility forms the basis of regulation as the binding sites of the protein could be altered with the variation in overall shape of the protein structure. In comparison to other biological molecules such as carbohydrates, lipids and nucleic acids, it is the sheer versatility and capability to assume different shapes of proteins which makes them nature's preferred tool to perform

the somewhat complicated duties in a single living cell. Some of these include playing the role of transporters, defending organisms as antibodies, transmitting information as hormones, controlling gene expression, transcribing genetic information, protecting fellow proteins to acquire their tertiary structure as chaperones, etc. As proteins perform a vast amount of functions in a living cell, it is vital to understand the normal functioning and probable malfunctioning of proteins for therapeutic purposes. In the context of rational drug design, they elucidate a number of tasks which range from blocking the synthesis of proteins of bacterial ribosomes[5, 6], malfunctioning of some proteins may result in many pathological diseases like cancer[7], neurodegenerative disorders caused by abnormal protein aggregation like Alzheimer[8], Huntington[9] or motor neuron diseases[10]. Also, proteins play an important role in attacking the vital proteins of pathogens of HIV[11, 12], SARS[13, 14], etc. Due to this participation in almost every task that is essential for life, protein science has increasing importance for the development of modern medicine.

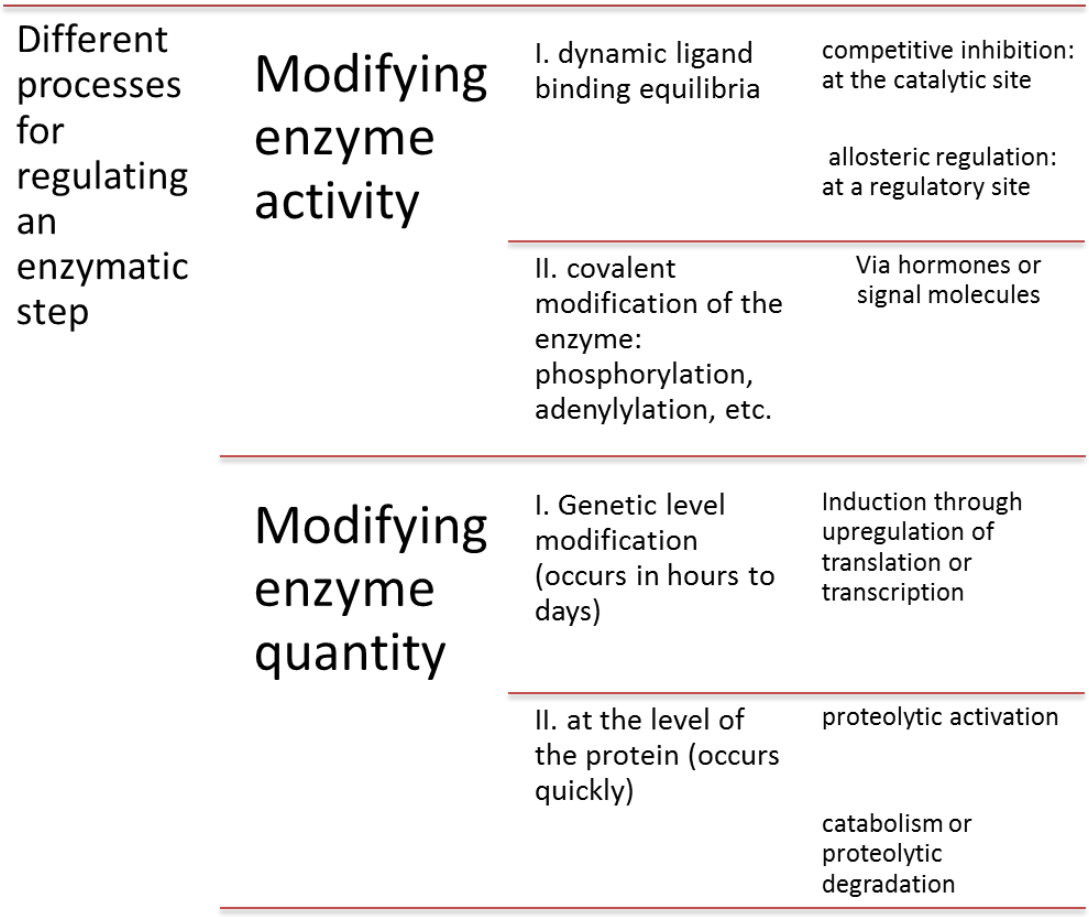


Figure 1.1: Regulation of enzymes by different biochemical processes.

To sustain the molecular machinery of our body, binding events need to take place. This is facilitated by proteins which can bind to small molecules (often called ligands) or other proteins and fulfil various tasks. However, a proper control needs to be exerted over the binding event and there are multiple ways which alter the behaviour. This is done by effector molecules which change the binding behaviour of the protein (Figure 1.1). Depending upon the type of effectors, these can increase the binding affinity or catalysis rate thereby acting as activators or do the opposite and act as inhibitors and suppress protein activity. If an effector binds in the same site as the ligand it is affecting, the regulation is termed as orthosteric regulation[15].

The effector can directly act on the ligand, or it can directly manipulate the binding site. However, if an effector binds at a site distant from the site whose binding affinity it is changing, the regulation is termed as allosteric and the effector is called an allosteric effector and the whole phenomenon is referred to as allostery. This is one of the most intriguing and well-studied mechanisms which control the functioning of proteins (Figure 1.2). This thesis focuses on understanding the conformational changes, behaviour and allosteric regulation of two of the proteins involved in glucose metabolism reaction.

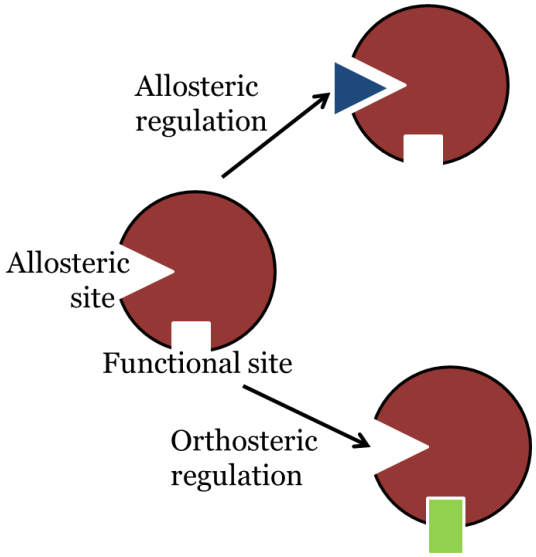
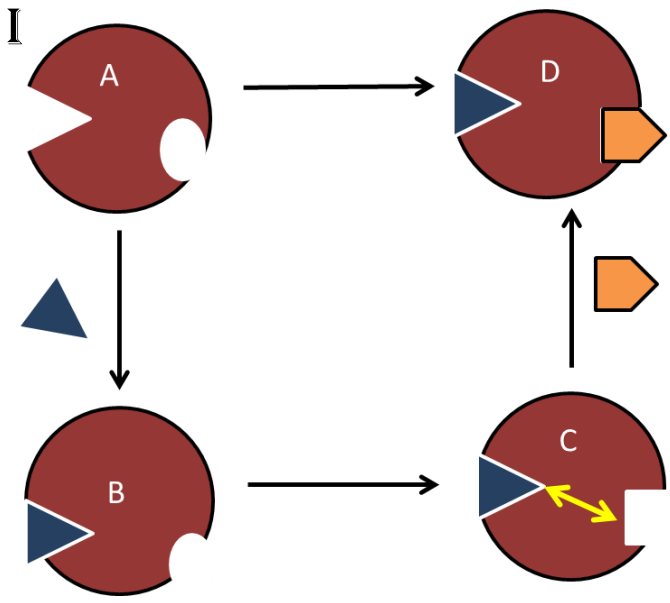


Figure 1.2: Illustration of orthosteric and allosteric mode. In case of orthosteric regulation, the perturbation originates at the functional site of a cellular target. However, if the perturbation is located at a site that is not in the vicinity of the functional site, it is an allosteric event. The Blue triangle represents the allosteric ligand and the green square represents the substrate molecule.

1.2 Allostery

Allostery in its most basic sense refers to a change in conformation brought about by the binding of an effector molecule to a protein at a site which differs from the active site, namely the allosteric site. It is derived from a combination of two Greek words: *allos* = other, and *stereos* = shape. When the effector binding site coincides with the substrate, it may increase the substrate binding affinity (allosteric activator) or decrease it (allosteric inhibitor). Between the binding sites, there needs to be communication in order to transmit the information (Figure 1.3). The importance of allostery emanates from the fact that it aids the proteins to transmit the regulatory effects induced by the binding of a ligand at one site to a different and distant site which in turn governs the function The past five decades have seen a rapid growth of interest in allostery, from merely a biochemical phenomenon to an area of potential drug development. It has been aptly titled as "second secret of life"[16-18]. There exists a plethora of ways in which the cells are controlled allosterically, which range from covalent modifications such as phosphorylation, acetylation[19], many binding interactions involving ions, lipids, light absorption mechanisms and environmental stimulus like pH, temperature, etc. [20-22] (Figure 1.4).



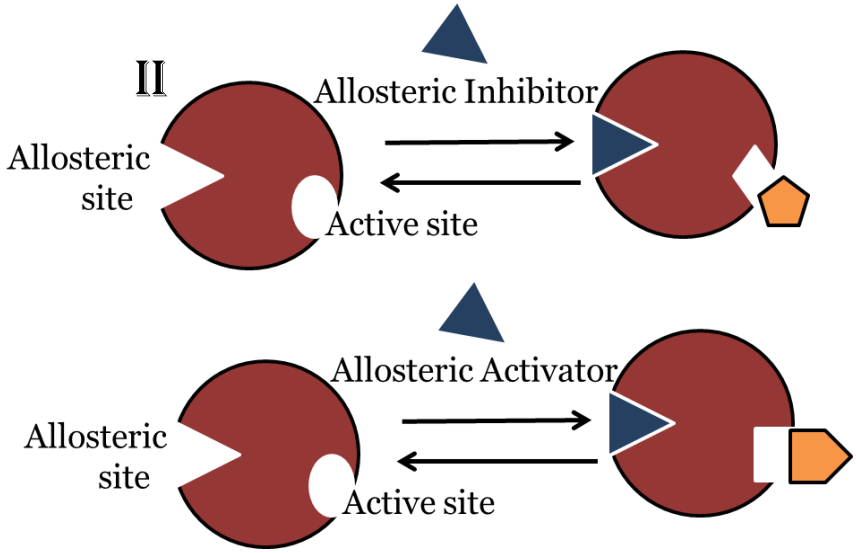


Figure 1.3: A cartoon representation of a typical allosteric event and regulation. I: Sketch of a typical allosteric interaction wherein the allosteric effector molecule binds at the allosteric site and induces some change in the substrate site to facilitate it's binding.
II: The allosteric regulator can act both as an inhibitor and activator. *Top*: An allosteric inhibitor alters the active site or binding site conformation in an unfavourable way, thereby decreasing substrate affinity or catalytic efficiency. *Bottom:* On the other hand an allosteric activator results in increased substrate affinity. Blue triangle represents the allosteric ligand and the orange structure represents the substrate molecule.

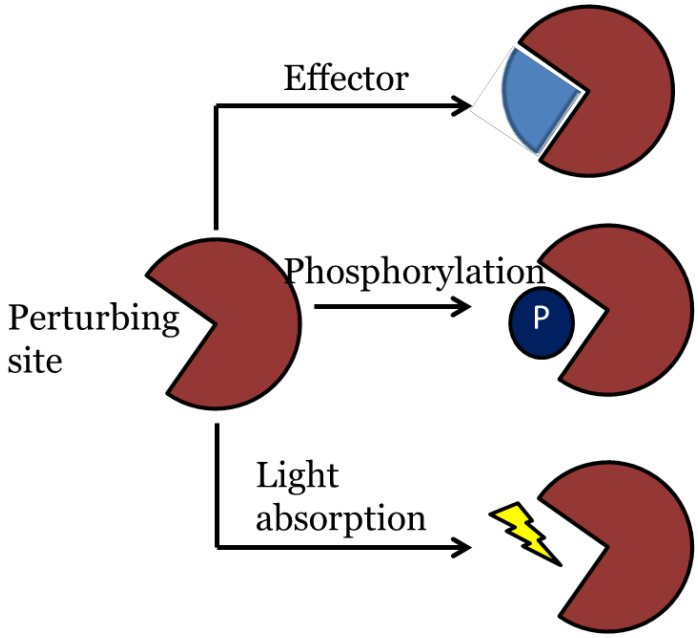


Figure 1.4: Initiation of structural changes in a macromolecule through local perturbations such as a binding event, posttranslational modification, mutation (not illustrated), or light absorption.

There is a growing body of evidence demonstrating the regulation of protein activity by communication between two sites – active site and the site of modification or allosteric site[23]. The original definition of allostery as the regulation of a protein by a small molecule that differs in shape from the substrate has been broadened to mean a change of state (flexibility or structure) caused by an interaction with another protein or small molecule[24]. This definition incorporates the original Cooper-Dryden model[25] in which changes in vibrational states of the protein without any conformational changes could cause allosteric effects.

A mechanistic view of allostery or allosterism has been defined in a number of ways. For instance, in some cases the lipid bilayer with its associated material properties (thickness, intrinsic lipid curvature, and the elastic compression and bending moduli) has been proven similar to an allosteric regulator of membrane protein function[26]. From a thermodynamics perspective, enthalpy and entropy define the behaviour of a system. Binding of a ligand/effector/any molecule reduces the thermal motions and mobility of the binding partner and thus becomes entropy-unfavourable but enthalpy-favourable owing to the increases in interactions. Since, the nature of allostery is fundamentally thermodynamic, the communications may be mediated both by enthalpy (conformational changes) and entropy (dynamic fluctuations about mean structure)[27-29]. This means that the perturbations at an allosteric site may or may not be followed by a conformational change. If there is no conformational change, then the entropy loss at the binding site can propagate dynamically[30].

This perturbation may further increase or decrease the affinity of the substrate, where an increased affinity is termed positive cooperativity and the decreased affinity is termed as negative cooperativity. Conformational change at the substrate binding site can be a result of positive cooperativity dominated by enthalpy whereas the unfavourable entropy drives the negative cooperativity.

It also involves three interacting role-players: an **effector molecule** (a ligand which binds to the protein receptor of interest), **allosteric site**, that

transduces the thermodynamic allosteric energy to the active site (the other site). The allosteric effects are also reciprocal in the sense that the guest receives the energy of an effector molecule from the transporter and in turn returns energy via the allosteric site back to the effector molecule[31] .The mechanism gives proteins a way to sense the environment and react to it.

In the traditional view, an energetic "hot wire" has also been proposed, which links the allosteric encryptor site with the guest binding site[32, 33] (Figure 1.5). Recent studies have also argued that there exists an ensemble of protein conformations rather than just two conformational states[34]. This manner of binding of different ligands in different ways to a common site can then lead to a plethora of allosteric consequences for the receptor. The environmental conditions may lead to the different structural and functional states in an allosteric protein.

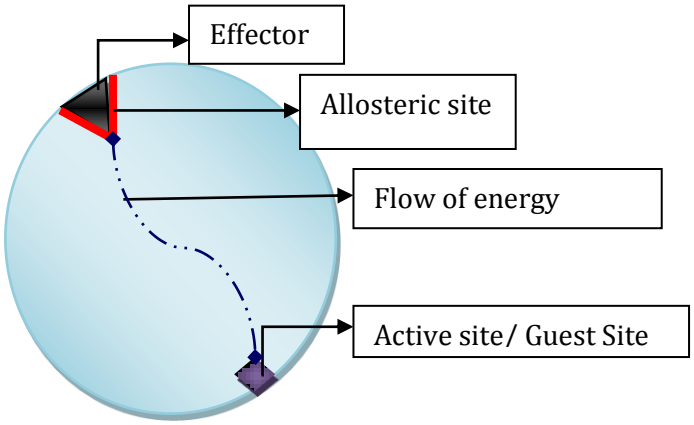


Figure 1.5: Schematic representation of the key interactions during an allosteric event. Also shown is one of the channels of allosteric communication by the flow of energy from allosteric site to the active site.

According to the two initial models, i.e. MWC and KNF model (explained in the following section), allosteric proteins are oligomeric and adopt two structural states called T (Tight and is biochemically inactive) and R (Relaxed and is biochemically active). Depending upon the type of allosteric transition, a protein may adapt to any of the above mentioned states. The intriguing and challenging area concerning allosteric interactions is that there is no general answer to how this information flow is manifested in different systems. Despite a heterogenous array of information available on allosteric mechanisms, the underlying basis of the transmission of the allosteric signal across the protein has not been clearly defined. A typical perturbation signal produced at the allosteric site is transcribed across the tertiary structure of the protein to its active site with the help of various interatomic fluctuations, domain motions or residue networks thereby altering the protein function[35, 36]. A deviation from this intricately designed system of allosteric communication may result in several pathogenic diseases like cancer, diabetes or Alzheimer's disease[37]. This forms the basis to understand the role of allosteric networking within a protein structure and perhaps to design effective drugs to either block or support the mechanisms in order to prevent the diseases. Additionally, it has been observed that the allosteric drugs provide greater selectivity, fewer side effects and low toxicity in comparison to the orthosteric drugs which bind to the same proteins[15, 38-41]. This pervasive biological occurrence and advantage of allosteric regulation makes it an area of interest to exploit for therapeutic purposes.

1.2.1 Models of Allostery

The initial models of allostery emerged 50 years back and were described by Monond, Wyman & Changueux (MWC model) in 1965 followed by K Nemethy & Filmer (KNF model) in 1966. This superseded the purely lock and key analogy of protein ligand binding originally postulated by Fischer over 100 years ago. Our views have broadened from the classical theories and evolved from the traditional rigid receptor-substrate binding as shapes and conformations

of bindings sites can be modified considerably by degree and specificity of substrates[42, 43]. There exist a number of theoretical models to explain the allosteric regulation, some of which are summarised below (Figure 1.6):

- **MWC Model**

A classical and foremost model of allostery was presented by Monod, Wyman and Changeux in 1965[24, 44, 45]. According to the MWC model, allosteric proteins are oligomers made of symmetrically arranged identical monomers. They exist in equilibrium between two conformational states, R state i.e. relaxed state and the T state i.e. tensed state (Fig. 1.6). The equilibrium can be shifted to either of the two states through binding of a ligand to a site different from the active site. MWC model is also known as the two-state concerted model in which the change in one subunit is conferred to all the subunits. This model was extensively used to describe the chaperons such as GroEL, CheA, etc[46, 47]. However, this model was extended from the initial two-state to tertiary two state model (TTS) in order to account for the allosteric cooperativity under various experimental conditions. According to the TTS model, there exists equilibrium between the two differing high and low affinity tertiary conformations of individual subunits that are present in both the T and R quaternary structures. This model could explain the varying affinity of oxygen in haemoglobin as regulated by the allosteric inhibitors[48].

- **KNF Model**

The KNF model is a sequential model in which the subunits do not necessary exist in the same conformation and so the conformational changes are not propagated to all the subunits. Also, the substrate binding at one subunit alters the structure of other subunits to facilitate the binding to adjacent subunits. It also negates the idea of any conformational change in the absence

of a ligand. One of the most common example of cooperative binding is Haemoglobin molecule [49, 50].

➢ Population Shift Model

With the rapid extension of MWC model there have been a few theories which state that the apo form of proteins exists in an ensemble of conformations[34]. This is characterized by a free-energy landscape whose dynamics and the relative populations can be altered allosterically. In the presence of an allosteric effector a complementary substrate binding site of a fluctuating protein will be selected. This complementary binding of allosteric effector to the allosteric site would result in a redistribution of the conformational states towards the conformation most favoured by the effector and hence undergo a population shift [51, 52].

➢ Morpheein Model

Apart from the three above mentioned models, this is a more recent model which is based on the homooligomeric forms with distinct functionality i.e. morpheein forms. The interconversion of these morpheein forms requires the dissociation of higher order multimers into a lower order multimer which is preceded by a conformational change at the lower order state. According to this model, an allosteric effector can bind to a multimer and shift the interconversion equilibrium to a more preferred lower order multimer[53, 54].

➢ Dynamically Driven Model

The Dynamically driven model is more specifically observed in the catabolic activator protein (CAP) where the mutant seems to activate the protein for DNA binding. This allosteric activation of mutant is carried out without the change in protein structure but rather the dynamic behaviour is altered.

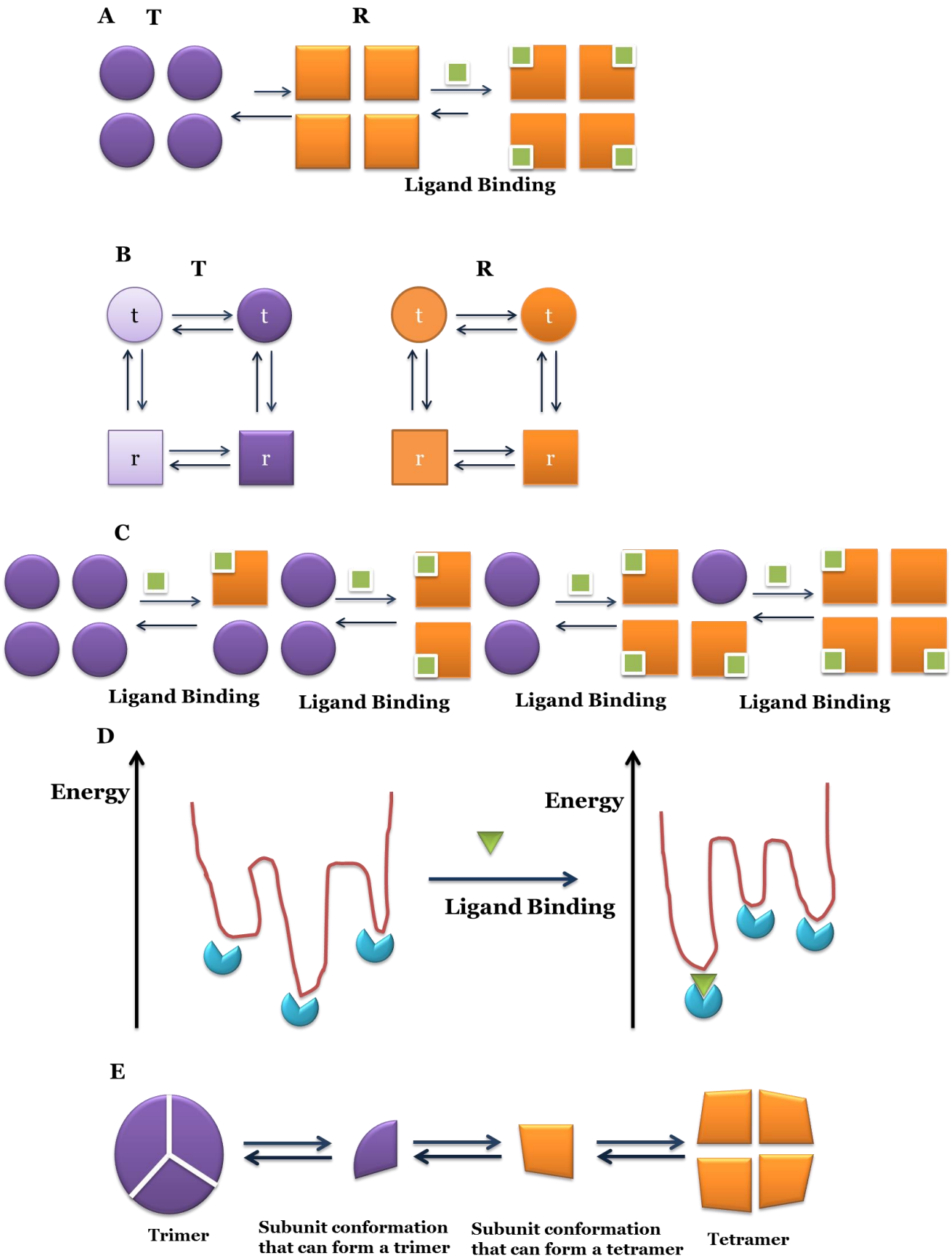


Figure 1.6: An illustration of the various proposed models of allostery . A: MWC model, B: Tertiary Two State Model (open symbols represent unliganded subunit; filled symbols represent liganded subunit; circles, *r* tertiary conformation; squares, *t* tertiary conformation, C: KNF Model, D: Population shift Model, E: Morpheein model[3]. Purple diagrams represent the unliganded structures and orange shows the active liganded structures.

All these models and studies have rather revolutionised the initial perception of allostery. From pure structure mediation, it has been broadened to be mediated predominantly by changes in protein's dynamics[55, 56].

1.2.2 General features of allosteric proteins

With the extensive research going on in the field of allostery, an increasing number of allosteric proteins have been discovered using such experimental procedures as X-ray crystallography and NMR spectroscopy. The increasing number of available high-quality protein structures has facilitated the determination of characteristics of these proteins both computationally and experimentally. Considering the array of allosteric proteins, we can briefly enlist some of the common features such as :

- Allosteric enzymes consist of multiple polypeptide chains, with multiple active and allosteric sites[57].
- Allosteric enzymes have the ability to respond to several different conditions in their environments[58].
- Allosteric sites tend to be more hydrophobic as they are enriched with hydrophobic residues in comparison to the catalytic and orthosteric sites [59, 60].
- Allosteric proteins undergo quaternary structural changes which includes the rearrangement of the subunits and domains[45].
- There are site-site interactions over large distances and complex kinetic patterns in classical allosteric proteins[61].
- The allosteric modulators have been observed to obey the Lipinski’s rule of 5 in a much more strict fashion[62].

- Allosteric proteins show a sigmoidal behaviour and thus deviate from a typical Michaelis-Menten model. This is due to the fact that allosteric enzymes have multiple coupled domains/subunits a d show cooperative binding. This results in the sigmoidal dependence on the concentration of their substrates

- Substrate concentrations can influence the equilibrium of allosteric enzymes.

- Furthermore, the presence of other molecules can regulate and influence the working of allosteric enzymes[63].

Even though we have enlisted some of the features, we cannot limit allostery to these which seems to be evolving at an enormous rate and new features and insights are being discovered rapidly.

1.2.3 Methods to identify and characterize allosteric sites

I. Experimental Methods

a) Tethering

Also known as disulfide trapping, tethering is a rather direct approach to characterise the allosteric sites on a given protein using small molecules. In this technique, a cysteine modified protein is screened against a library of disulphide containing compounds under partially reducing conditions. This cysteine residue is typically adjacent to the site of interest. Upon screening, a disulphide bond is formed between the small molecules and a cysteine residue in the proximity of the allosteric site which can be easily detected via mass spectrometry[64, 65].

b) Flourescent labels

This experimental technique is based on attaching a fluorophore covalently to the biological macromolecules. A reactive derivative of the fluorophore is used to selectively bind to a functional group of the target molecule. However, the important part of this technique involves the selection of an appropriate amino acid position that is solvent exposed and displays movement upon ligand binding[66, 67].

c) High Throughput Screening (HTS)

High throughput screening (HTS) is a sequential drug-discovery technique for target validation and is based on assaying a large number of potential biological modulators against a chosen set of defined targets. HTS follows a combinatorial approach of several techniques like liquid handling and robotic automation, multi-platform plate readers and more recently high content imaging. This provides an advantage over conventional methods in terms of time and money [68, 69].

d) Fragment based Screening (FBS)

This is a complementary approach to HTS and provides several advantages over HTS in terms of higher chemical diversity, higher ligand efficiency and hit rates. It is based on identifying the fragments or low molecular weight compounds that generally bind with weak affinity to the target of interest. The fragments that form high quality interactions are then optimized to lead compounds with high affinity and selectivity[70, 71].

II. Computational Methods

a) Simulations

In terms of computational resources, large scale unbiased MD simulations are being exploited rapidly to uncover the allosteric sites. It has been successfully employed in finding the allosteric sites of GPCRs amongst several other proteins[72].

b) Allosteric Toolkit (AST)

This webservice predicts the allosteric sites and has been optimised using a SVM model (Support Vector Machine). It is based on the set of allosteric proteins whose allosteric sites are unknown, and is capable of predicting potential allosteric sites which affect the orthosteric functions of the proteins[73].

c) Binding Leverage

This approach focuses on detecting allosteric sites based on intrinsic protein dynamics without performing full-scale simulations. It is based on the notion that a typical allosteric regulation involves conformational transitions or fluctuations between a few closely related states which can be triggered by the binding of effector molecules to stabilize a native conformation. [74, 75]

d) Theoretical models like NMA, Go models

There are several theoretical models which help in identifying the allosteric sites. For example, the go model is based on the coarse-grained two states of protein. It is based on population shift model where the energy landscape has been tuned to bias the active state[76].

The NMA model, i.e. the Normal Mode Analysis model monitors the conformational changes in protein flexibility upon ligand binding to predict the presence and location of allosteric sites[77].

e) SID analysis

Another computational method to identify allosteric binding has been developed by researchers at the University of Strathclyde. Simple Intrasequence Difference analysis (SID) is based on grading the individual

residue position in a protein 3D structure according to the topology in the folded chain. This in turn generates an expression of potential contribution of each residue position along with the neighbouring residues towards the molecular conformation. It is this internal arrangement of chain interfacing which helps in predicting the potential for site-specific inductions via ligand binding or mutations or allosteric binding. Thus this analysis helps in comprehending the properties of protein fold topologies[78]. Structures of free and inhibited human secretory phospholipase A2 have been studied to determine the folds and binding[79]. Also, the closeness of residues in a protein structure during the primary sequence for Lactate dehydrogenase is studied. It also has been studied to observe and determine the folds of bovine pancreatic trypsin inhibitor, phospholipase A2, chymotrypsin and carboxypeptidase A

Several of these models and techniques are described in detail in the materials and methods chapter.

1.3 Glycolytic pathway

Several types of enzymes are present within an organism to carry out a spectrum of reactions. These reactions are carried out in the form of biochemical pathways[80]. A series of chemical reactions occurring within a cell can be defined in terms of metabolic pathways. This involves the step by step modification of an initial molecule to form a final product. The substrate is converted to a product by a first enzyme in the pathway and this product then acts as a substrate for the next enzyme, thus a series of chemical reactions follow until the final product is obtained. A metabolic flux can be regulated by:

a) Availability of substrate
b) Concentration of enzymes responsible for rate-limiting steps
c) Allosteric regulation of enzymes
d) Covalent modification of enzymes (e.g. phosphorylation)

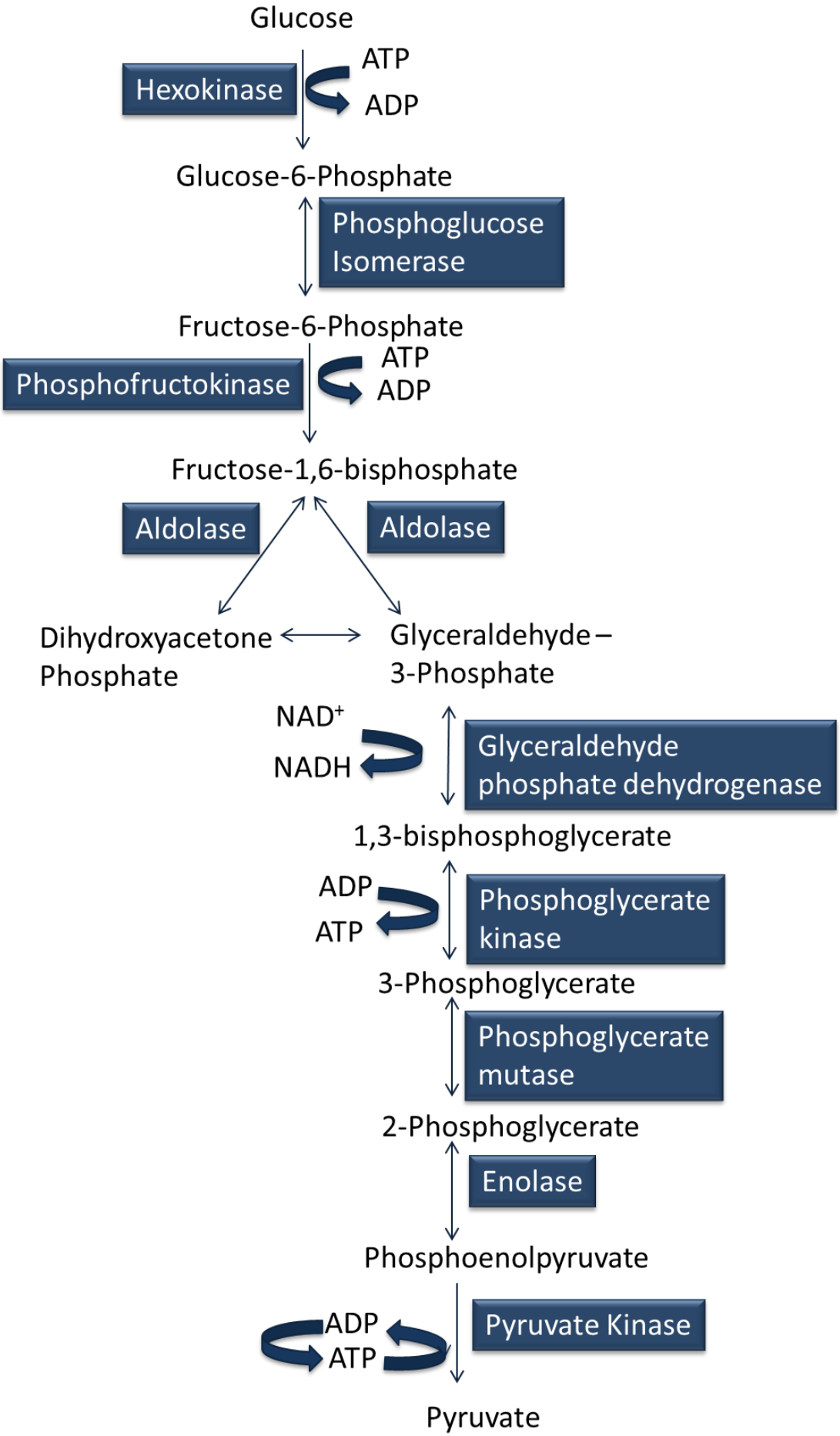
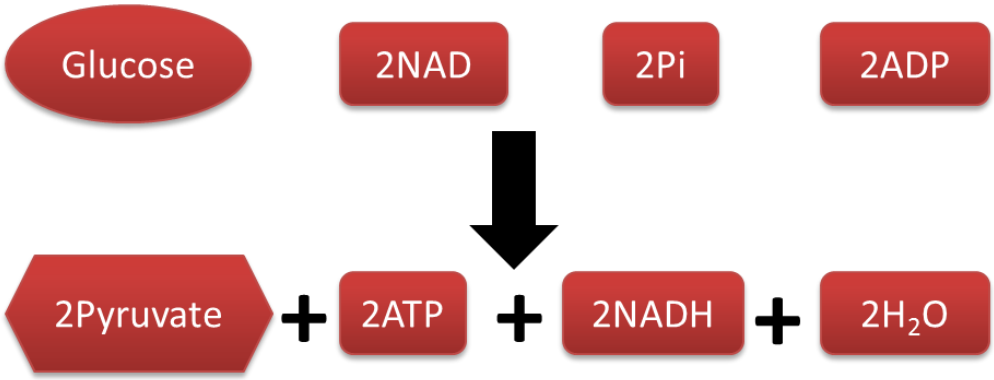


Figure 1.7: The ten enzymes of a glycolytic metabolic pathway of an organism

One of the first metabolic pathways discovered was glycolysis. The glycolytic pathway is common to all cells, both prokaryotic and eukaryotic. It is a core metabolic pathway which can be considered to comprise three main stages. In the first stage, glucose is converted to fructose 1,6-bisphosphate and consists of three steps. The second stage involves the cleavage of fructose 1, 6-bisphosphate and the final stage leads to the generation of ATP when pyruvate is formed[63]. The glycolytic pathway, also called the Embden-Meyerhof Pathway, is universally present in all living organisms. In eukaryotes, cytoplasm is the site of action where the glucose is oxidised to pyruvate thereby generating ATP and NADH by a series of 10 main steps (Figure 1.7).

The Net Reaction of glycolysis is as follows:



Regulation of glycolysis occurs at three points of the pathway. As is the case in metabolic pathways, enzymes catalysing irreversible reactions serve as control points. The three points of pathway correspond to the largest negative free energy changes (i.e. most exergonic - negative ΔG). It is the magnitude of the ΔG for these steps which makes them essentially irreversible and thus potential sites of control. The three rate-limiting enzymes serving as control points of glycolysis are hexokinase, phosphofructokinase and pyruvate kinase.

Enzyme	Activator	Inhibitor
Hexokinase	AMP/ADP	Glucose-6-phosphate
Phosphofructokinase	AMP/ADP, Fructose-2,6-bisphosphate	ATP, citrate
Pyruvate kinase	AMP/ADP, Fructose-1,6-bisphosphate	ATP, Acetyl CoA, Alanine

Table 1.1: Regulation of glycolysis by the three control enzymes

The activities of these control enzymes are monitored through covalent modifications, transcription and also a variety of allosteric effectors. Out of these, PFK activity is modulated by the allosteric inhibitors such as ATP and citrate which signal increased energy levels whereas the allosteric activators signal the low energy levels. Also, PFK has a major control on glycolytic flux since it serves as a key rate-limiting enzyme. Hexokinase is subject to product inhibition by glucose 6-phosphate. It catalyzes the phosphorylation of glucose. In conditions of low PFK activity, the increase in relative concentration of fructose 6-phosphate also increases the levels which slows down the rate of catalysis by hexokinase, thus it is regulated by excess G6P levels. In the case of liver, however, glucokinase allows brain and muscles to utilize glucose prior to its storage as glycogen. The third key enzyme, PYK, is controlled allosterically through different isozymes. Generally, fructose-1,6-bisphosphate which is also the product of PYK serves as allosteric activator whereas ATP and alanine are the allosteric inhibitors. PYK is inhibited under low glucose conditions and F-1,6-BP drives the PYK activity forward.

As extensive work on LmPYK and PFK has been carried out in the laboratory and also due to the intrinsic key positioning of these two enzymes in the glycolytic pathway, it made them our first point of interest. This thesis focuses on the case studies presented on the two regulatory valves of glycolytic pathway.

1.4 Outline of this thesis

The focus of this work is to explore and understand the conformational changes incurred upon ligand binding. The case in study is of two important enzymes of the glycolytic pathway whose dynamics have been explored using standard Molecular Dynamics techniques.

In this chapter I have outlined the basis of this dissertation by focussing on the fundamental concepts of allostery. Through this dissertation I have tried to understand the nature of allostery using molecular dynamics simulations and principal component analysis. Also, the thesis has employed the computational approach to probe the dynamics of protein conformations under allosteric control. The details of the methods and framework are introduced in chapter 2,

Over the years, there have been numerous case studies on protein conformations using molecular dynamics simulations to understand the effect of allostery. This technique has been utilised on one of the most important enzymes of the glycolytic pathway, pyruvate kinase, Chapter 3 presents the first case study and the results of our simulations which have been applied to both the tetrameric and monomeric systems.

Chapter 4, presents the second case study on yet another enzyme from the glycolytic pathway, phosphofructokinase. The simulations are run both on the monomer and the tetramer and follow a similar approach as that of pyruvate kinase to analyse the results. Phosphofructokinase is another key enzyme of the glycolytic pathway and we have tried to understand the effect of effector molecule on the conformational dynamics of this enzyme.

Chapter 5 follows a brief conclusion and overall summary of the results as obtained through MD simulations following which is the bibliographical list of the works cited in this thesis.

1.5 Therapeutic potential

There is a rapid progress in the field of cancer metabolism and drug development. With the clinical evidences linking progression of cancer with cell metabolism, metabolic enzymes are being sought as potential drug targets for cancer therapy. Extensive research is being carried out to target the metabolic pathway as means of anticancer strategy. Predominantly, it has been observed that the cancer cells exhibit increased glycolysis and utilise this pathway as a main source of energy. The role of glycolysis in regulating cell proliferation makes the pathway a good potential drug target for a diverse range of diseases including cancer and parasite infection. Also, increased aerobic glycolysis is commonly seen in a wide spectrum of human cancers and thus development of novel glycolytic inhibitors as a new class of anticancer agents is likely to have broad therapeutic applications[37].

The two key regulatory enzymes of glycolytic pathway are phosphofructokinase (PFK) and pyruvate kinase (PYK). Thus, these serve as hot spots for anti-cancer drug development strategies. Both the systems are quite large and understanding the molecular basis of regulation of these key enzymes can help in designing effective molecules to block the pathway. Despite extensive crystallographic and experimental studies on these enzymes, the mechanism of regulation is quite elusive particularly in terms of allosteric regulation. For instance, it is not known exactly how the binding of an allosteric activator triggers changes in the neighbouring subunits of PYK despite being 40 Å away from the active site of the enzyme. The main thrust of this work thus lies in combining and utilising the molecular dynamics techniques and statistical measures like Principal Component Analysis to reveal the details and properties of atomic motions. The linkage of molecular structure to function in complex biochemical systems has contributed majorly to the success of molecular biology and the introduction of MD has further expanded the horizons[81]. In order to validate our observations we have compared the computational results with the previously published experimental data. The key question of how the binding of an effector at a remote site affects the overall structure and function of a protein is addressed.

The prime focus has been to understand the phenomena of allostery and track the associated conformational changes following ligand perturbation. The dynamics of PYK and PFK at the atomic level have been examined and attempts to understand the regulating driving forces have been made.

Although the allosteric signals are very subtle to track, we have still been able to complement our simulation results with experimental evidence i.e. the results obtained after experiments in the form of crystal structure information or behaviour of bonds and angles. This has been done by combining the information of various PCs to yield a description of concerted physical modes of motion. The results presented in this work have provided new insights into the modes of protein regulation and certainly accentuated the fact that MD simulations and associated techniques can be well exploited to understand the molecular basis of communication. However, the next step from our study would be to apply the standard MD-PCA combinations to the two systems and also to other enzymes of the glycolytic pathway along a much longer time scale. The analysis of the PCs and eigenvalue spectrum would further provide substantial understanding of underlying cell metabolism. This would help in identifying novel drug targets and aid in the design of putative effector molecules against the receptors of interest.

CHAPTER 2:

THEORY AND METHODS

2 THEORY AND METHODS

2.1 CLASSICAL TECHNIQUES

In order to understand the mechanism of biological processes, an understanding of protein dynamics is essential[82]. The influence of conformational dynamics on catalysis[83], allostery[84] and molecular recognition[85] can be further probed with a description of atomic motions [Figure 2.1]. However, numerous intermolecular interactions and complexity in protein conformations makes this study quite complex. A number of techniques are employed, both experimental and computational which provide insights into protein dynamics. Some of them are listed below:

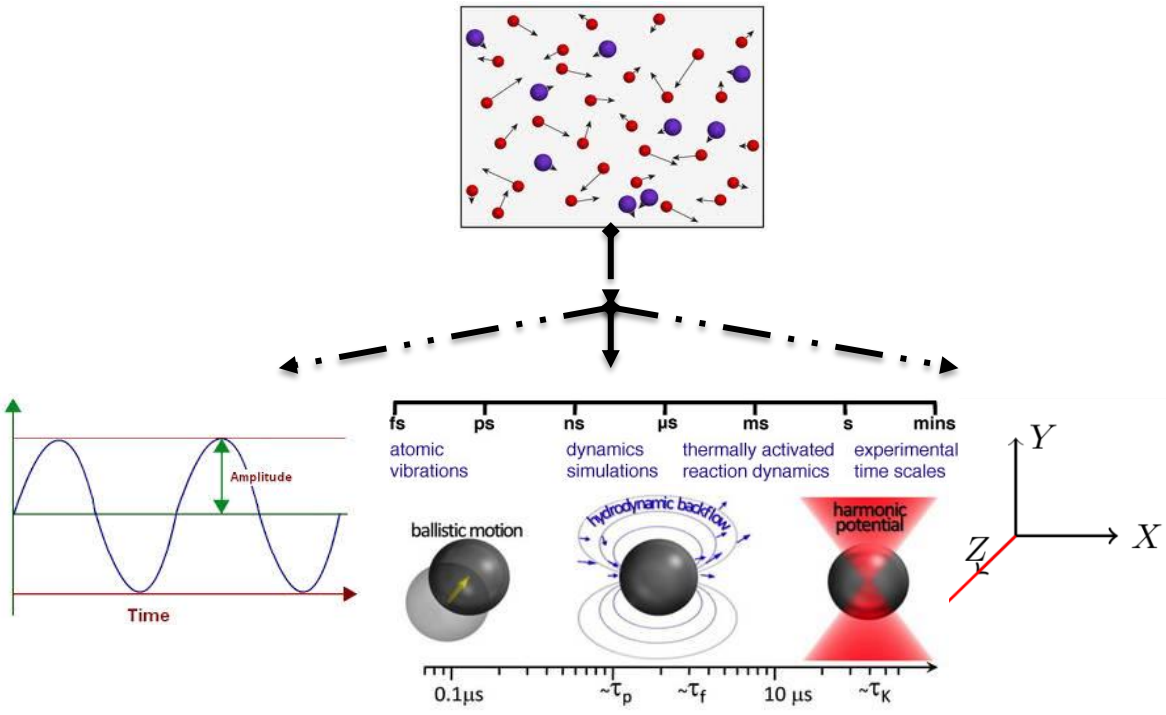


Figure 2.1: Components of atomic motions; Left to Right: Amplitude of the motion, the timescale and direction.

2.1.1 X-RAY CRYSTALLOGRAPHY

X-ray crystallography, also known as X-ray diffraction, is one of the most common methods to investigate molecular structures enabling us to visualize protein structures at the atomic level and augment the understanding of protein function. It also enables to study protein-protein interaction, conformational changes and catalysis. A phenomenon called X-ray diffraction takes place when the X-rays hit the crystal. Diffraction occurs when a wave (of any nature) encounters an obstacle, which can be any material object. This results in bending of the wave around that object, also called scattering of waves. Also, diffraction could be caused when a wave encounters a small opening, a small hole or a slit. This causes spreading of the wave in all directions. In both of these cases, the hole/slit start to act as a new wave source, sending around waves with slightly different direction of propagation, as compared to the original wave. The "new" scattered waves interact with each other, resulting in other physical phenomena called interference, which simply means addition of waves. It is a multi-step process involving three distinct stages. It begins with crystal production in a wet lab where the protein is extracted, purified and crystallized. The next step is in an X-ray facility where the crystal is mounted and the data is collected after shooting with an x-ray beam. The final step is computing the obtained density maps which are processed via multiple stages of refinement before the final model is built[86] (Figure 2.2). Nowadays, almost all centres are equipped with their own x-ray generators which facilitate the preliminary testing of crystals and selection of the best ones for synchrotron sources. In order to obtain X-ray data from a crystal, it needs to be placed in a monochromatic (single wavelength) X-ray beam with a wavelength roughly equal to a typical interatomic distance. This intense monochromatic X-ray beam is produced in the synchrotron sources, which have highly sophisticated optics for generating high resolution structures [87-89]. Many allosteric sites such as those of HIV transcriptase[90], haemoglobin[91], glucokinase, etc. have been identified by this technique.

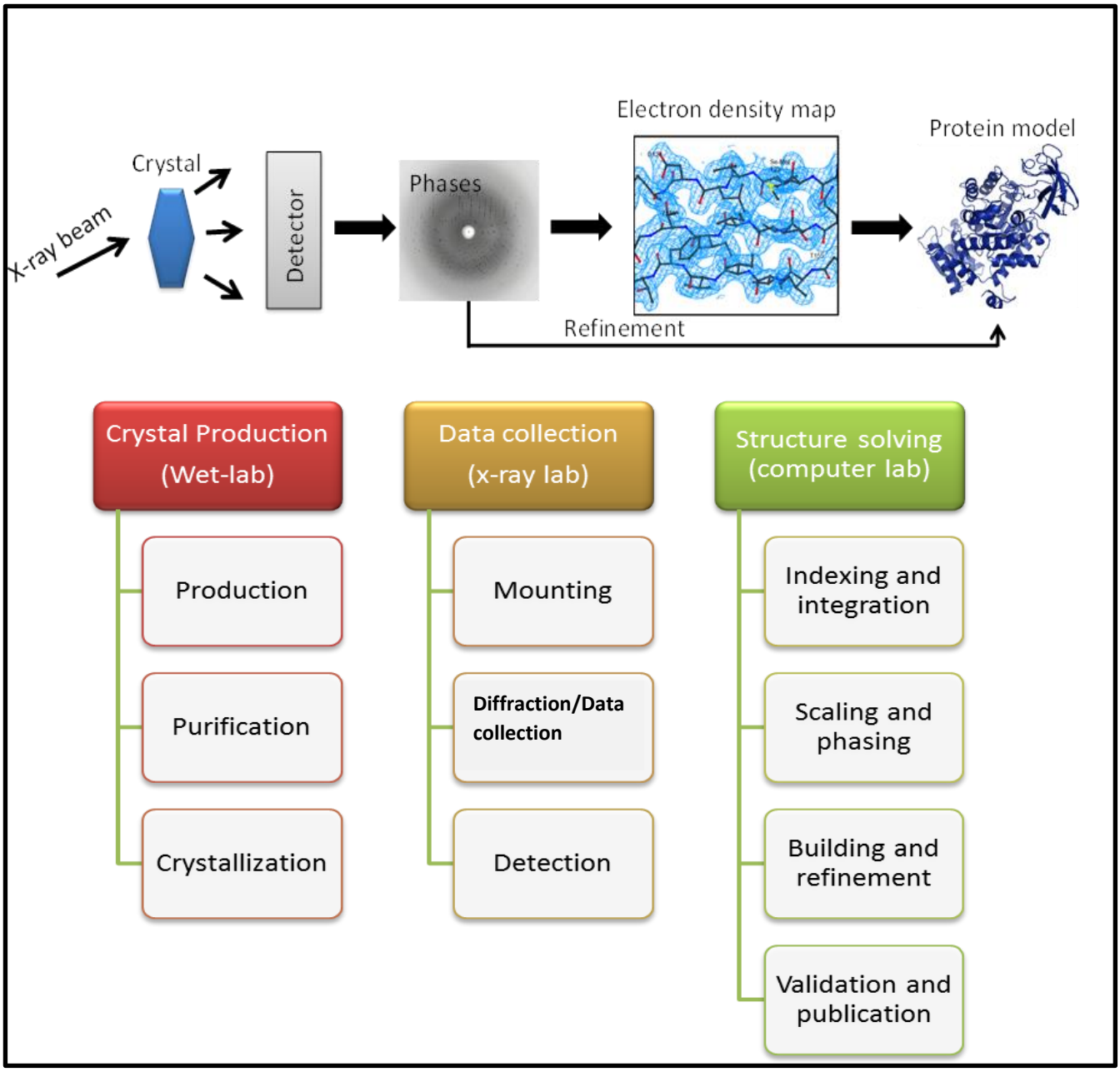


Figure 2.2: Schematic representation of the steps involved in crystal structure determination (top) and list of various processes performed in the three stages ; crystal production, data collection and structure sorting (bottom). The first step is to grow crystals of the molecule of interest which is then put in an X-ray beam. The crystal scatters the X-rays onto an electronic detector, which functions as a recorder. With specialized computer programs, electron density maps are constructed which tell about the 3-D structure of the molecule. This is then used to design protein models which are then validated.

2.1.2 NUCLEAR MAGNETIC RESONANCE

Nuclear Magnetic Resonance (NMR)[92] spectroscopy is a complementary technique to X-ray crystallography with both having their advantages and disadvantages. While X-ray involves study of solid crystals, NMR studies molecules in solution. This can be an advantage as conducting the analysis in solutions similar to physiological conditions allows the study of dynamic properties of the proteins. However, the limitation of protein size (<20kDa) when using NMR proves a major disadvantage in terms of studying large and complex systems[80]. But still, the method has been implemented successfully to identify residues involved in protein-ligand interactions[93]. Together, both these techniques give an insight into the structural, functional and dynamical aspect of the molecules (Figure 2.3).

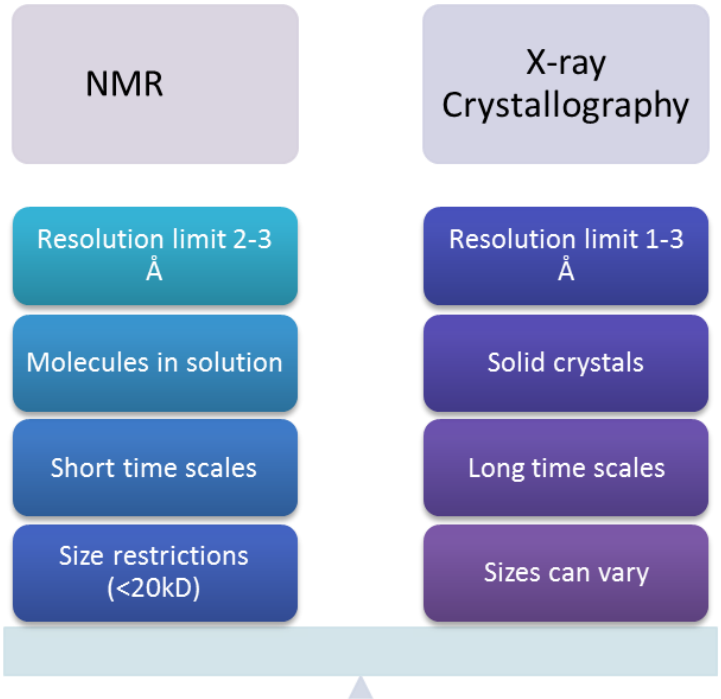


Figure 2.3: Some of the features of the two complementary techniques (X-ray and NMR).

NMR measures relaxation times and determines the order parameters, chemical exchange rates covering a broad timescale from ps to ms[94], which captures the fast and slow motions (figure 2.4). A number of 2D techniques are applied to extend the usability of NMR. Some of these include transverse relaxation optimised spectroscopy (TROSY) detection method [95, 96] which uses spectroscopic means to reduce transverse relaxation; NMR can be studied for molecules of up 100 kDa weight. Interpretation of NMR spectra is quite complicated with the most common problem being overlapping peaks. This can be overcome by using NMR techniques like nuclear over Hauser effect spectroscopy (NOESY)[97] which measures distance-dependent coupling, and total correlation spectroscopy (TOCSY) which measures coupling between covalently bonded atoms. Strategies have been optimized and employed to identify allosteric inhibitors[98] by NMR and further, extensive work is being carried out to increase the throughput of structure determination[99].

2.1.3 FLUORESCENCE RESONANCE ENERGY TRANSFER

Fluorescence resonance energy transfer (FRET) is a technique in which a laser beam is used to excite two fluorophores attached to the molecule of study[100]. It is a photophysical phenomenon and a process involving energy transfer between an excited donor fluorophore and an acceptor fluorophore. It is a sensitive technique which can identify small distance fluctuations between atoms in a macromolecule which can be measured with respect to time, thereby providing information about conformational dynamics. Recently, this technique has been employed to probe the conformational dynamics of caM-regulated proteins (calcium binding protein calmodulin)[101].

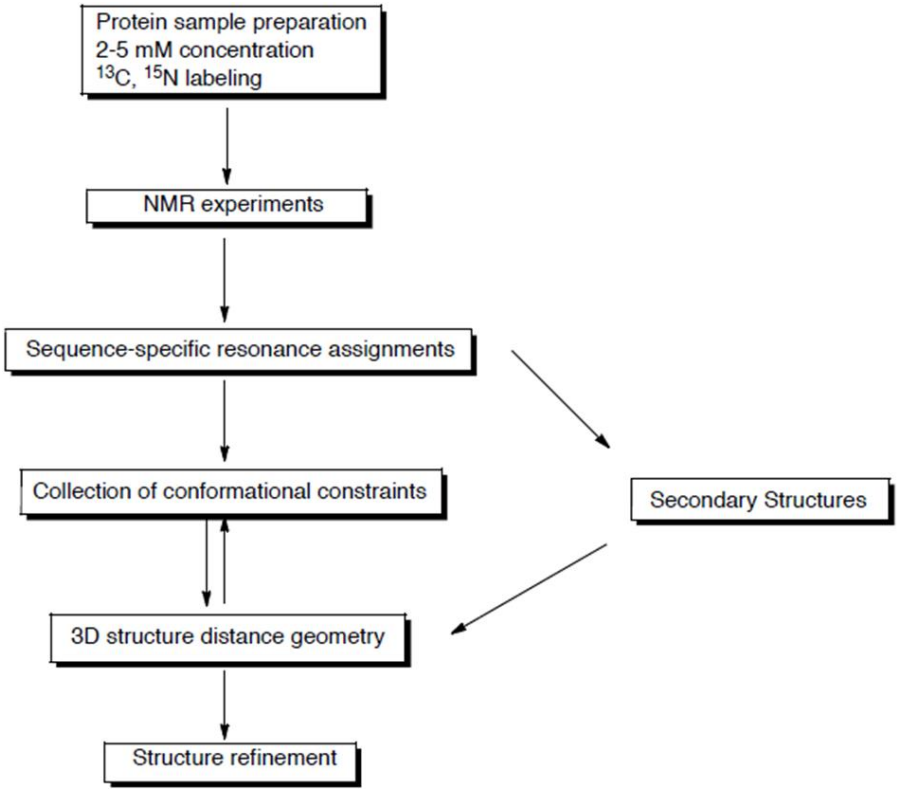


Figure 2.4: Schematic representation and outline of steps to obtain a crystal structure using Nuclear Magnetic Resonance Spectroscopy. Adapted from [90].

2.1.4 ATOMIC FORCE MICROSCOPY

Atomic Force Microscopy (AFM) involves studying particles in solution without the prerequisite of fixing them[102]. It has three different approaches namely, contact, oscillating and force to gather information about the molecule. In the contact approach[103], the mechanical probe is in direct contact with the molecule while in oscillating approach[104], low force is applied and the molecule is touched briefly. The force approach as the name suggests involves measuring the forces between the molecule and the AFM tip at picoNewton resolution and has also been applied to study the allosteric transition in 20S

proteasome gate[105]. Furthermore, the allosteric effects of calcium-dependent signal transducer calmodulin (CaM) were studied using force spectroscopy[106].

2.2 COMPUTATIONAL APPROACHES

Although rapid development has been made in optimizing and enhancing the modern experimental techniques to understand the dynamics of macromolecules, time and cost still pose a limitation. Therefore, many conventional computational methods like normal mode analysis (NMA), molecular dynamics (MD), etc. alongwith structural analysis tools aim to provide information about the protein function and dynamics which otherwise is elusive of the experimental approaches.

2.2.1 PREDICTIVE METHODS

Predictive methods like COREX [107, 108] are used to monitor protein stability. This algorithm characterises a thermodynamic ensemble into regions of folded and non-folded states. This method has been used to monitor the effect of ligand binding and it identifies the most probable state i.e. minima in an energy landscape. COREX also reveals functionally relevant thermodynamic couplings based on the relative distributions of residue folding states. It models the native-state as ensembles of microstates in which residues may exist in either a folded or unfolded state. These microstates can be considered analogous to thermal fluctuations or local pico- to nanosecond-scale motions that allow the protein to sample conformations separated by low-lying energy barriers. It uses a reduced model for the degrees of conformational freedom available to a residue, as each residue exists in either a folded or unfolded state. [109, 110].

Another method called the Statistical Coupling Analysis (SCA) developed by Lockless and Ranganathan[32], makes use of functional information buried within [111]the evolutionary record from the sequences of a family of

proteins [112].This method divides the family members based on allosteric regulation by examining the sequences within a family and thus has been implemented to validate allosteric signal networks. SCA however has the drawback of being specific to the evolutionary co-conservation of residues which might not be an essential property of allosterically coupled residues[32].Another closely related technique is the evolutionary trace analysis which is based on multiple sequence alignment and a sequence identity based evolutionary tree[113].

2.2.2 NETWORK COUPLING METHODS

Investigating the coupling among residues is being rapidly used to study the new structural communication pathways[114]. In one of the studies, networks of contacts in 15 pairs of allosteric proteins were calculated which led to a finding of changes being initiated from allosteric site to the active site for at least five of the pairs[57]. The deviation for the other 10 pairs suggested that the large scale conformational changes like rigid-body motions must be accompanied by local contact rearrangements.

Another protein network based method called the Markov random walk has been applied to predict the changes in which signal transmission is modelled through conformational space where the state of the system changes randomly between steps[115].

2.2.3 FEATURE PREDICTION MODELS

These methods involve construction of databases of known protein-protein interactions by integrating the fields of proteomics and bioinformatics[116, 117]. One of the first mechanistic study of protein-protein interactions was done by construction of a 2300 alanine residue mutant database where the binding hot spots were identified[118]. This work was further extrapolated into elucidating

the cooperativity among these residues while they participated in binding at protein-protein interfaces[119]. Furthermore, advances have been made in prediction of protein interaction hotspots from sequence data, accessible surface area, residue mapping, protein structure.[120, 121]

2.2.4 DYNAMIC METHODS

Methods based on dynamics generate an ensemble of conformations which are then analysed via a number of structural techniques such as cross-correlations[122, 123], contact correlations, principal components[123], etc. Also, there are several other dynamic methods that account for both quasi-harmonic and anharmonic correlations in Cartesian space[124, 125].

One of the most common dynamic methods is normal mode analysis (NMA). It is based on protein vibrations around a static structure and approximates the energy surface.[81, 126]

MutInf is one of the approaches to analyse protein conformational ensembles that are generated by molecular dynamics or Monte Carlo simulations[127]. This entropy based approach quantifies the amount of conformational dependence between protein residues and calculates the configurational entropies[128].

2.2.5 Methods identifying potential binding sites

There are three categories for identifying potential binding sites using computational methods, namely: geometry based, energy-based and docking/ probe mapping algorithms. The probe mapping algorithm can be used for de Novo design of suitable ligands[129-132].

Methods like SURFNET[133], LIGSITE[134], PASS[135], CAST[136], PocketPicker [137] and STP [138] are geometry based and employ different algorithms such as virtual sphere fitting or grid point fitting into a binding site.

PocketFinder[139] and Q-SiteFinder[140] are the two energy based methods which retain probes with favourable interactions. While PocketFinder sorts the clusters by their volume, QSiteFinder ranks the cluster according to their total interaction energies.

2.3 MOLECULAR DYNAMICS SIMULATIONS

Although experimental techniques such as X-ray, NMR, etc. provide information about the structure, mechanism and function of a protein, a static view, or a selection of static views, of the biomolecule are presented. However, in order to understand the activities of biological interest, it is imperative to use simulation techniques which render visualization to these static images. Many multiscale techniques such as quantum chemical, classical molecular dynamics, coarse grained molecular dynamics and meso scale techniques are used to study the micro scale to meso scale properties of the biological systems. Molecular Dynamics (MD) simulation[141] is a widely used computational biophysics method which efficiently provides insight into many biochemical processes at an atomistic level (structure-property relations, mechanism, dynamical, thermo dynamical properties). It generates a step-by-step trajectory of all particles in the system using a force field to approximate the potential energy of a given protein[1] and is used to understand the binding and folding events of the protein at an atomistic level. This provides insight into the evolution of the system with time, thereby providing detailed information on the atomic fluctuations and conformational changes of the protein. MD simulations are a bridge between the microscopic and macroscopic world of the laboratory (Fig. 2.5).

MD methods[142] are classified into two major groups according to the mathematical model chosen to represent a system. The classical mechanics approach or classical MD, resembles a 'ball and spring' model. Atoms correspond to soft balls and elastic sticks correspond to bonds. The laws of classical mechanics define the dynamics of the system. However, another group/family is the 'quantum' or 'first-principles' MD, started in 1980s by Car and Parinello[143], which involves the quantum nature of the chemical bonds.. The electron density function for the valence electrons that determine bonding in the system is computed using quantum equations, whereas the dynamics of ions (nuclei with their inner electrons) is followed classically. This approach provides improvements over the classical MD albeit at a more expensive computational cost. This makes classical MD practical for simulations of biomolecular systems comprising many thousands of atoms over time scales of nanoseconds.

All-atom classical MD is the central method which has been applied in this research to obtain the trajectories and conformations of our proteins. We have used this technique to study the allosteric mechanisms of the proteins and visualize the trajectories in order to gain insight into the communication when the effector molecule binds at the allosteric site.

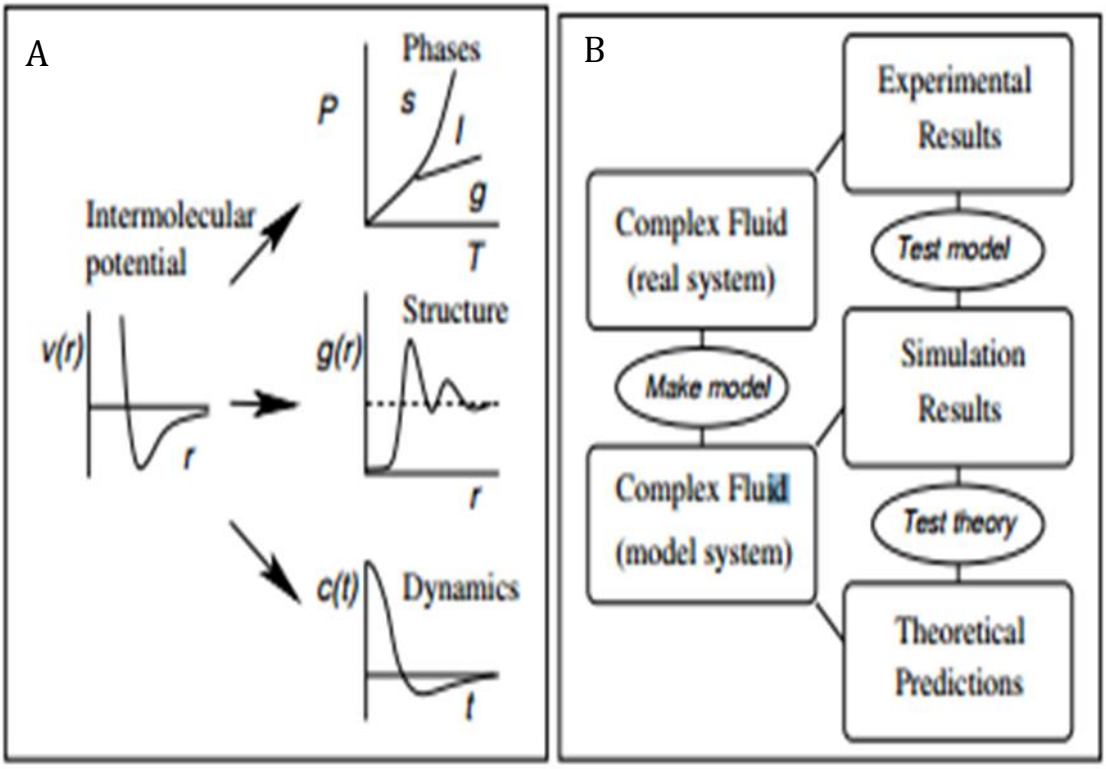


Figure 2.5: MD simulations bridge the gap between the two worlds:
(a) Microscopic and macroscopic observables; (b) theoretical and experimental techniques. Adapted from [1]

2.3.1 BACKGROUND

The history of MD simulations dates back to late 1950's when the technique was first introduced by Alder and Wainwright[144, 145] (Figure 2.6). They elucidated useful information about the behaviour of liquids. Following them, Rahman[146] carried out the first simulation using a realistic potential for liquid argon. After this advancement, Rahman and Stillinger[147] did the first molecular dynamics simulation of a realistic system of liquid water . However, it

was only in 1970's that the first protein simulation was carried out. In 1977, the first molecular dynamics (MD) simulation of a protein at atomic level was performed by J. Andrew McCammon, and Martin Karplus[148] of a small bovine pancreatic trypsin inhibitor (BPTI), ~500 atoms, in vacuum for 9.2 ps. However, despite the time limitation factor, the significant atomic fluctuations showed by the BPTI paved the way for involvement of classical molecular dynamic simulations to protein structures. A variety of water models arose in the early 1980's. The former simulations were mainly restricted to small proteins, ps timescales and vaccum conditions due to the high expenses in terms of computation. However, the transition from the ps to the ns and μs timescales (Figure 2.7) started around 1988 which has further been quantified and advanced on to nearly identical physiological conditions and visualization of binding and folding events. A number of groups have carried out MD simulations since the pioneering work of Karplus, for e.g. those of Brooks[149] ,van Gunsteren[150], Levitt[151], Jorgensen[152], Daggett[153], Kollman[154], Pande[155], Berendsen[156], Baker[157], McCammon[158] , among others.

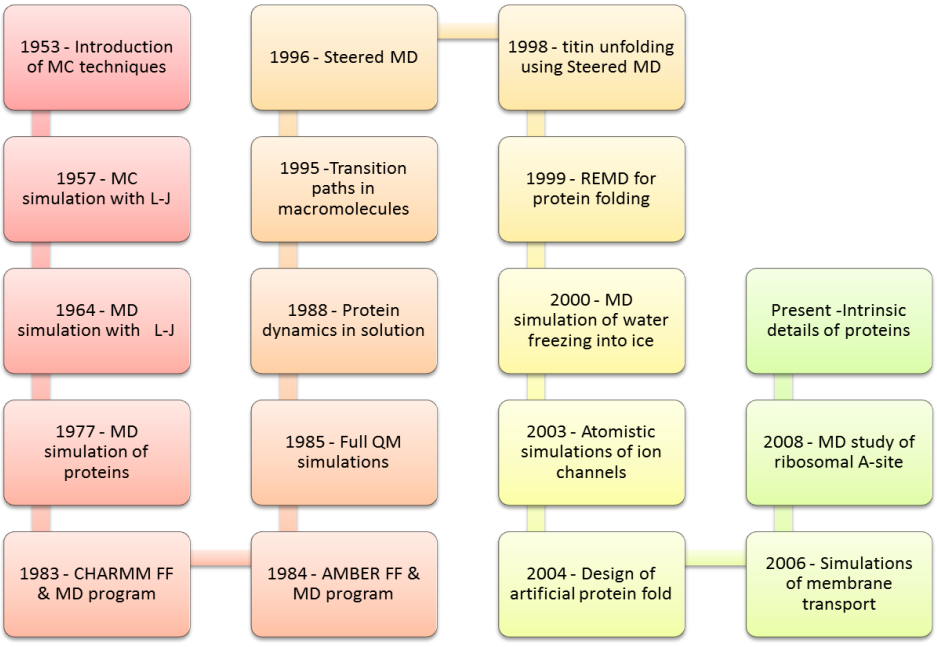


Figure 2.6: 50 years of simulations; A brief history of simulations over the years.

A

Motion	Timescale (s)
Bond stretching	10^{-14}
Global DNA twisting	10^{-12}
Surface-sidechain rotation	10^{-11} - 10^{-10}
Collective subgroup motion (hinge bending, allosteric transitions)	10^{-11} - 10^{-7}
Global DNA bending	10^{-10} - 10^{-7}
Interior sidechain rotation	10^{-4} – 10^{0}
Protein folding	10^{-5} - 10^{1}

B

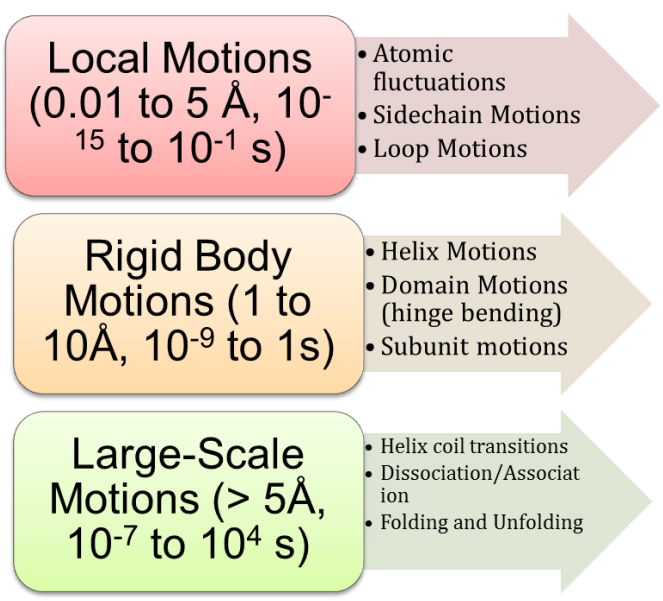


Figure 2.7: a: Overview of broad spectrum of characteristic timescales of biomolecules , b: Biological molecules exhibit a wide range of time scales over which specific processes occur

2.3.2 Basic workflow and approximations of MD

Molecular dynamics (MD) simulation is a technique by which one generates the atomic trajectories of a system of N particles by numerical integration of Newton's equation of motion, for a specific interatomic potential, with certain initial condition (IC) and boundary condition (BC) while satisfying thermodynamical (macroscopic) constraints i.e. time evaluation of a set of N interacting particles by solving Newton's equations of motion:

$$\boldsymbol{F = ma} \tag{2.1}$$

$$\boldsymbol{F_i = m_i \frac{d^2 r_i(t)}{dt^2}} \tag{2.2}$$

Where F_i is the force acting on atom i at time t and m_i is the mass of that atom and $r_i(t) = (x_i(t), y_i(t), z_i(t))$ is the position vector of atom i.

In terms of the potential energy function, V, of a system, the force on atom i can be determined as the negative of the gradient of potential energy.

$$\boldsymbol{F_i = -\nabla V} \tag{2.3}$$

In physical world of many body systems, these forces depend on the position of the particle which would change whenever the particle moves or interacts. This requires a continuous sequence of states updated with time (Figure 2.1) i.e. from the current state of atom (r(t),v(t)) to the next state $(r(t+\delta t), v(t+\delta t))$. These positions and velocities are integrated using finite difference method which assumes that position and dynamic properties of the atom can be approximated as Taylor's series expression: There exist several algorithms for

integrating the equations of motion all of which follow the Taylor series convention.

$$\boldsymbol{r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \frac{1}{6}b(t)\delta t^3 + \cdots} \quad (2.4)$$

$$\boldsymbol{v(t+\delta t) = v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \frac{1}{6}c(t)\delta t^3 + \cdots} \quad (2.5)$$

$$\boldsymbol{a(t+\delta t) = a(t) + b(t)\delta t + \frac{1}{2}c(t)\delta t^2 + \cdots} \quad (2.6)$$

$$\boldsymbol{b(t+\delta t) = b(t) + c(t)\delta t + \cdots} \quad (2.7)$$

Where r is the position, v is the velocity (first derivative of the position), a is the acceleration (second derivative of the position) and b is the first derivative of acceleration (third derivative of the position) with respect to time and so on.

The basic working of finite difference method is that the integration is broken down into many small stages, each separated in time by a fixed time step δt. The total force on each particle in the configuration at a time t is calculated as the vector sum of its interactions with other particles. From the force (which is assumed to be constant during the time step), we can determine the accelerations of the particles, which are then combined with the positions and velocities at time t to calculate the positions and velocities at time t + δt.

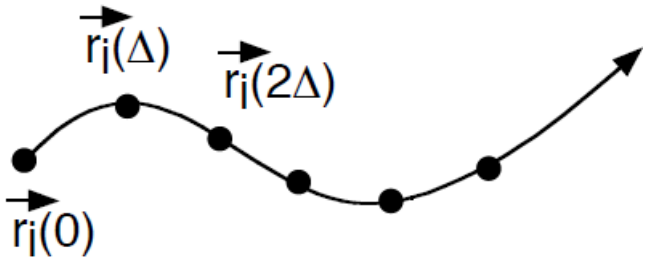


Figure 2.8: Progression of the algorithm

There are several criteria to consider when integrating the Newton's equations of motion and any good algorithm must abide by those. Some of these are:

1. Speed: This is not important since most time is spent on calculating nonbonded interactions and forces rather than integrating the equations of motion.
2. Accuracy: This is required for large timesteps. The longer the time step, the fewer evaluations of the energies and forces are needed.
3. Energy conservation: This is an important criterion to distinguish between short-time and long-time energy conservation.
4. Reversibility: Newton's equations of motion are time reversible, and so should be the integration algorithms. Non reversible algorithms will have serious long-term energy drift problems.

One of the simplest algorithms to solve the Newton's equations of motion is the Verlet algorithm (155). This algorithm uses the positions and accelerations at time t, and the positions from previous step, r(t – δt), to calculate new positions at time (t+ δt), r (t+ δt).

So,

$$\boldsymbol{r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots} \qquad (2.8)$$

$$\boldsymbol{r(t-\delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots} \qquad (2.8)$$

So, according to Verlet algorithm, $r(t+\delta t) = r(t+\delta t) - r(t-\delta t)$

i.e.

$$\boldsymbol{r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2} \qquad (2.9)$$

In order to obtain velocities, we need to divide the difference between the positions at $(t+\delta t)$ and $(t-\delta t)$ by$(2\delta t)$

$$\boldsymbol{v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{(2\delta t)}} \qquad (2.10)$$

So, in a Verlet algorithm, Given $r(t-\delta t)$ and $r(t)$,

1. Compute a(t) as a function of $r(t)$
2. $r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2$
3. $v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{(2\delta t)}$

As the velocity at time t cannot be calculated until the coordinate at time $t+\delta t$ is calculated. The velocities need to be estimated which indicates that the method is not self-starting. In order to overcome this limitation, there are several variations and algorithms which have improved upon the Verlet scheme. One of the modified Verlet scheme in which velocities appear explicitly is the Velocity Verlet algorithm in which the positions, velocities and acceleration are all stored at the same time without any loss of precision:

$$\boldsymbol{r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2} \qquad (2.11)$$

$$\boldsymbol{v(t+\delta t) = v(t) + \frac{1}{2}\delta t[a(t) + a(t+\delta t)]} \qquad (2.12)$$

So, unlike the Verlet algorithm, Velocity-Verlet algorithm would work as following:

Given (r(t), v(t)) :

1. Compute a(t) as a function of $r(t)$
2. $v\left(t+\frac{\delta}{2}\right) = v(t) + \frac{\delta}{2}a(t)$
3. $r(t+\delta t) = r(t) + v\delta\left(t+\frac{\delta}{2}\right)$
4. Compute a $(t+\delta t)$ as a function of r$(t+\delta t)$
5. $v(t+\delta t) = v\left(t+\frac{\delta}{2}\right) + \frac{\delta}{2}a(t+\delta)$

Other improvements to Verlet algorithm include the leap-frog, and Beeman algorithms. When choosing an integration algorithm one must keep stability (i.e. conservation of energy), accuracy, speed and computational efficiency in mind.

The value of the time step determines the stability of any algorithm used. Therefore, it must be an order of magnitude smaller than the fastest motions of the system. Typically, this would be the vibration of a bond that involves a hydrogen atom with a period of the order of 10 fs, and consequently the time step should be $1/10^{th}$ i.e. atleast of the order of 1fs in case of explicit solvent. When implicit solvent is used, the time step can be larger, from 2 to 5 fs. Small time steps would lead to insufficient sampling of the phase space whereas large time

steps would lead to instabilities in the integration algorithm and high energy overlaps. So, there is always a trade-off between accuracy and economy.

In order to achieve effective increase of the timescale, several algorithms such as SHAKE, RATTLE, and LINCS exist which constrain the valence geometry of the solvent molecules and then there are others using torsional-angle dynamics and rigid-body dynamics in which elements of structure (e.g., α-helical segments) are considered fixed. The use of simplified protein models enables one to increase the timescale further because of averaging out fast motions that are not present at the coarse-grained level.

In order to compare the simulations to the physical observables, they need to mimic the experimental conditions to some extent. Some of these conditions are temperature, pressure, boundary, solvent models and so on.

a. Initial conditions

A particular thermodynamic state is characterised by an ensemble. An ensemble is a collection of systems belonging to a single macroscopic state with differing microscopic states. The widely used ensembles are NVT (fixed number of atoms, N, fixed volume, V, and fixed temperature, T) , NPH (N,V and enthalpy, H is fixed), NPT (N, P and Temperature is fixed), íVT (fixed chemical potential, volume and temperature), NVE (N,V and energy is fixed).

When simulating the ensemble, different approaches are available for temperature control and energy removal in a realistic way. Some of the most common temperature scaling methods are Berendsen thermostat, Nosé-Hoover thermostat, Andersen thermostat and Langevin dynamics. In Berendsen's thermostat, velocities are rescaled to adjust the kinetic energy of the system. Nose-Hoover is a more sophisticated method, in which the Hamiltonian of the system is modified to correspond to temperature, and not total-energy, conservation. The Andersen includes random collisions of molecules with an imaginary heat bath at desired temperature and is thus a stochastic method. The

Langevin thermostat does not follow the Newton's equations of motion but rather Langevin dynamics. In this method, random force is given to all particles at each step to reduce the velocities using a constant friction. This frictional force decreases the temperature. The random force is selected from a Gaussian distribution and adds kinetic energy into the particles of the system, with its variance being the function of the selected temperature and time step. Thus random force is balanced with the frictional force which in turn maintains the temperature of the system. Likewise, there are numerous barostats available which keep pressure constant.

b. ***Force-fields***

In MD, force fields define the potential energy function V which governs the interactions between atoms. This V consists of terms characterizing different interactions of the system. The functional forms of the force fields are a trade-off between accuracy in representing forces acting on atoms and low computational cost or ease of parameterization. Potential terms for both bonded and non-bonded interactions are defined and contained in these force-fields where the former includes the bond, angle, dihedral, and improper interaction terms, while the latter includes the Van der Waals (vdW) and electrostatic interaction terms.

A typical MD potential which is used in GROMACS software package is of the following form:

$$V = V_{bond} + V_{angle} + V_{dihedral} + V_{improper} + V_{vdw} + V_{elec}$$

i.e.

$$\boldsymbol{V = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} \sum_{n=1}^{N} K_\varphi^n[1 + cos(n\varphi - \varphi_0] + \sum_{impropers} K_\omega(\omega - \omega_0)^2 + \sum_{i,j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i,j}\left[\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right]} \quad (2.13)$$

Figure 2.9 describes the potential energy terms graphically in detail.

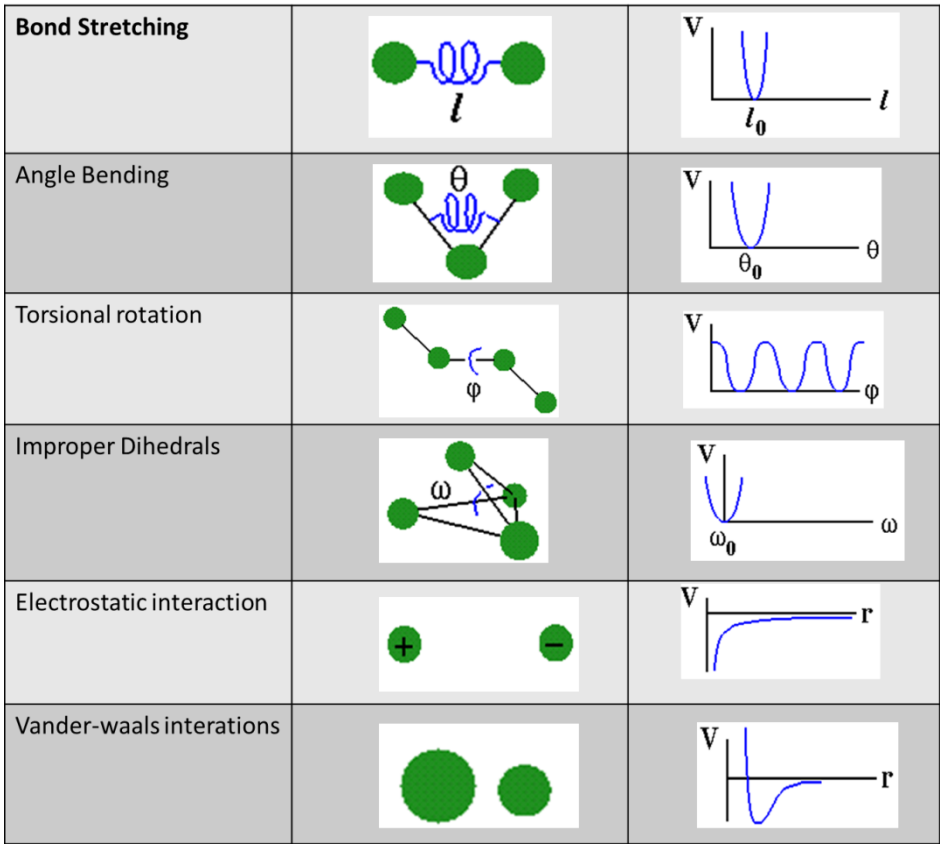


Figure 2.9: An illustration of the potential energy function.

Where bond stretching represents the energy required to stretch or compress a covalent bond, angle bending is the energy required to bend a bond from its equilibrium angle, θ, improper dihedral, is the energy required to deform a planar group of atoms from its equilibrium angle, ω_o, usually equal to zero, torsional rotation is the energy of torsion needed to rotate about bonds, van-der-waals interactions is the steric exclusion and long-range attraction energy and electrostatic interactions is the Coulomb potential function for electrostatic interactions of charges.

There are several force fields established like AMBER, GROMOS, OPLS, CHARMM, etc. All of these derive force constants from quantum-mechanical

calculations and from experimental measurements like vibrational bond-spectra or melting points of solvents and differ in the derivation of the parameters.

A dihedral angle is the angle between two plane surfaces which controls the rotation about the bond while an improper controls the "planarity" of the four atoms. In case of nonbonded interactions, the electrostatic term describes Coulombic interactions between two charged atoms, while the vdW term includes interactions due to induced dipoles and excluded volumes of pairs of atoms.

The vdW potential is attractive at long distance, but quickly becomes repulsive and prevents atoms from overlapping with each other. Among the bonded interactions, the bond, angle and improper terms all have the form of a harmonic potential and are very close to their equilibrium. In comparison, the dihedral potential is much softer which allows the backbone dihedral to adopt a broad range of values thereby giving biomolecules the flexibility to undergo large conformational changes,

For this research, we have utilised the AMBER 99sb-ildn force field with leap-frog algorithm. The reason for choosing this force field was that it was the updated and current force-field generated and accounted for several information regarding the angles and bonds. Also, in comparison to force-fields such as GROMOS and CHARMM, it was much more adaptable and easily applied for a system as huge as ours.

c. Long-range interactions

Calculation of non-bonded forces like Van der Waals interactions and electrostatics is very time-consuming. Various cut-offs are applied in order to deal with these calculations which otherwise are of biological interest especially the electrostatic interactions. However, these cut-offs might introduce significant

nonphysical effects and also increase the cost of computation. Therefore, methods like Ewald summation, the reaction field or cell multipole are applied.

In this work, we have used particle-mesh Ewald (PME) (a variant of Ewald sum method) to treat the long-range interactions. In PME the reciprocal space Ewald sums are B-spline interpolated on a grid and the convolutions necessary to evaluate the sums are calculated via fast Fourier transformations.

d. Boundary conditions and solvent treatment

It is imperative to determine the correct boundaries during simulation as it ensures the macroscopic properties to be calculated using relatively small number of particles. Periodic Boundary Conditions allow simulating bulk solid and liquid properties with a small number of atoms by eliminating surface effects. It replicates the cubic box that contains the protein and solvent through space.

In this, particles crossing a boundary of the simulation cell emerge back from the opposite side.

Essentially we try to model an infinite (approximately macroscopic) amount of matter with a small, finite simulation cell. It follows a minimum image convention in which each particle interacts only once with a given particle; by the particle proper or its periodic image, whichever one is closer.

Other important factor during simulation is the treatment of solvents which screen the electrostatic interactions. There are two solvent models available for including the solvent effects within the system; implicit and explicit solvent models. In case of implicit models, mostly homogenous medium is used to represent the solvent and the effect is mimicked by using the corresponding solvent dielectric constant.

The explicit model on the other hand, represents the molecules as atoms and therefore is more physically realistic. It relies on using hundreds or thousands of

discrete solvent molecules and is a widely used method for carrying out simulations in solvent. Such calculations converge only slowly because of the large number of particles involved.

Implicit models treat the solvent as a continuous medium having the average properties of the real solvent, and surrounding the solute beginning at the van der Waals surface. A variety of continuum models have been described including the generalized Born, Surface Area model, where the total solvation free energy is given as the sum of a solvent-solvent cavity term, a solute-solvent van der Waals term and a solute - solvent electrostatic polarisation term. The most widely used models are the rigid type TIP (Transferable Interaction Potential) models and SPC (Simple Point Charge) models.

MD simulations in this thesis have utilised both the TIP3P and SPC water models.

2.3.3 A typical MD run

Before we start with the actual run, it is imperative to answer a couple of questions pertaining to the research and technique. In short, we outline the following steps in order to begin and eventually conceptualise a typical md simulation.

1. What is our scientific question or aim of performing this simulation?

We want to understand the allosteric mechanism of our protein and the communication involved between the allosteric site and active site. We also want to monitor and observe the structural changes in the protein.

2. Can our results be verified? Do we have any record or previous facts to support our hypothesis?

There is published literature available and the data is available from in-house resources.

3. How much time is required to achieve the desired results?

We successfully completed our simulations within this academic tenure in order to quantify the published results and record our observations.

4. Do we have resources at hand?

Our work was supported by the British Council for funding to EPCC (Edinburgh Parallel Computing Centre) and UK's national supercomputer HecToR which was also supported generously by the ECDF for another year. Darwin Trust of Edinburgh provided funding for the academic years.

5. Now, we need an initial model of protein.

We begin with the crystal structure of Leishmania mexicana pyruvate kinase, an enzyme which has been determined crystallography within our research group.

6. What is the environment of our simulation?

We simulate our protein in solvent and further the charge is neutralized by addition of counter ions.

7. What are the conditions for our simulations? Temperature and pressure?

We maintain a standard/benchmark temperature of 318 K and pressure of 1 bar.

8. What is the shape of our simulation box and boundary conditions?

Since our protein is a tetramer and the system is fairly large, we use a dodecahedron box which is deemed suitable for such large systems. We use periodic boundary conditions and particle mesh Ewald summation to properly calculate our electrostatic interactions.

9. How long do we intend to perform a simulation for ?

Since we have been interested in conformational changes and transitions, we started with simulations ranging between 50-100ns keeping in mind with our available resources and time constraints.

10. What is the timestep of our calculation? Does our system require gradual heating?

We have used timestep between 2-5fs depending upon the size and complexity of our simulating molecule. 5fs is mainly used for large molecule and bigger system.

11. Once we have set out these initial commandments, we now optimize our protein structure which begins with fixing the protein coordinates and energy minimization with steepest descents method.

12. Then, MD simulation of a solvent is done under NVT conditions using a berendsen thermostat in order to maintain the temperature of 318 K.

13. This is shortly followed by NPT equilibration by V-rescale thermostat, a modified berendsen method in which the protein is relaxed from constraints. Around 500-3000 steepest descent steps are sufficient to transfer the protein (+water) from the "experimental" minimum to a "local force field related" minimum.

14. Now, since our system is equilibrated, we can perform the final production run (50-100ns, in our case).

15. Then, finally we obtain our trajectories and manipulate them in order to test our hypothesis. The structures at selected time point (frames) should be stored. These structures are further analyzed using computer graphics and specialized software analysis tools like VMD, Pymol, etc.

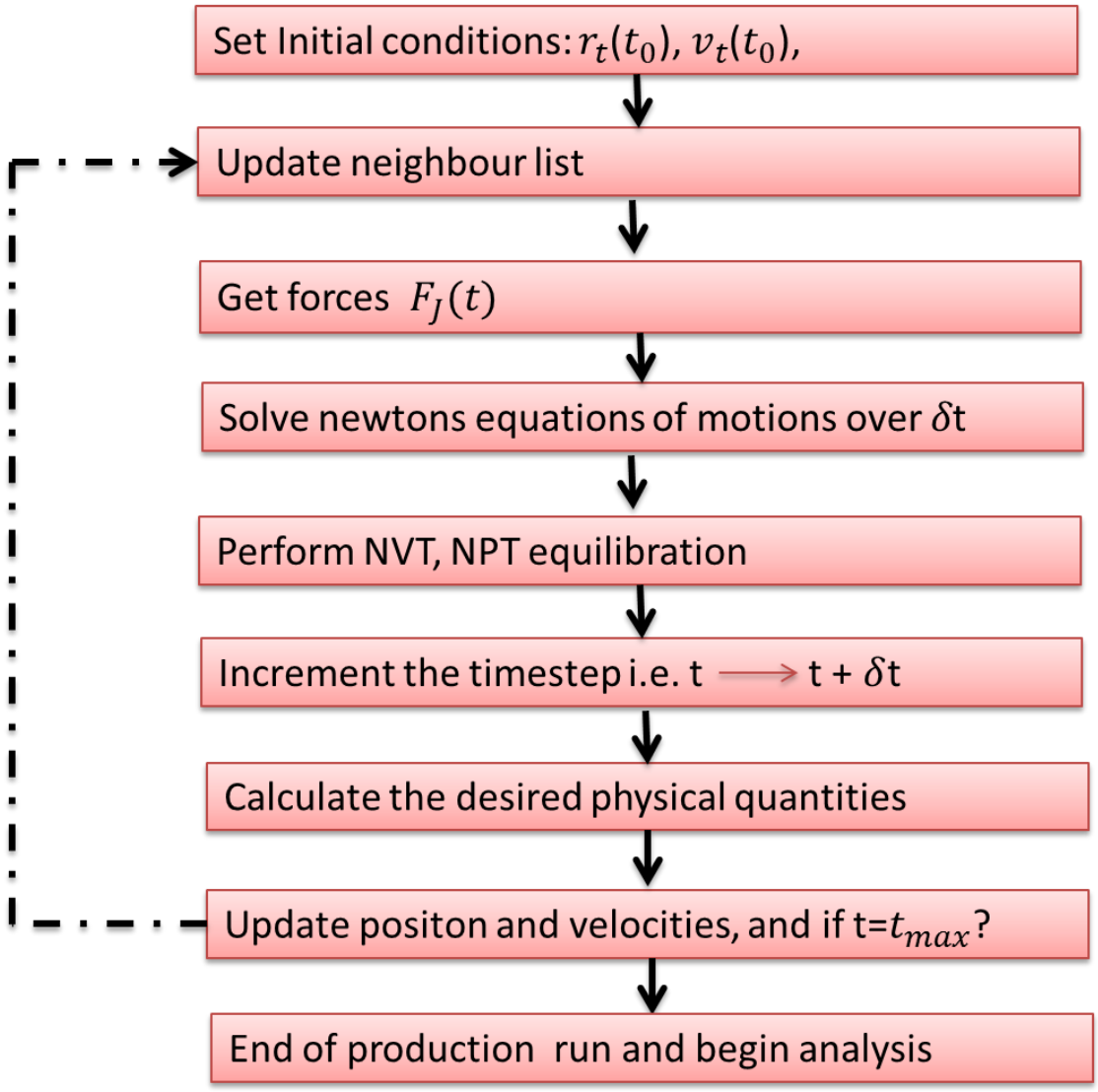


Figure 2.10: Steps involved in a typical MD simulation

Some of the most common softwares used for Molecular Dynamics are Amber, CHARMM, Desmond, Gromacs, NAMD/VMD, Tinker, Discovery Studio (Accelrys), HyperChem, Yassara, Desmond/Maestro (Schroedinger Inc.).

In this work, we have used Gromacs 4.5.5 to perform all of our MD simulations.

2.3.4 Variations of MD simulations

One of the limiting factors of MD simulations is the timescales (Fig. 2.7), which might not be sufficient enough to sample large conformational changes due to the limited availability of computing resources. In order to observe the most significant and intricate details of a system, one would have to consider a timescale of the order of ns to microsecond. There have been several implementations and variations to the current MD schemes to extend the rate of sampling by introduction of external force or biasing the degrees of freedom. Some of the most common methods are:

1. Coarse-grained molecular dynamics

In a standard all atom simulation, there is a tradeoff between the replication of experimental conditions and timescales of simulations to track the system. Coarse-grained method (CG-MD) is based on reducing the degress of freedom addressed in the potential form and the proteins are most often represented as collections of beads, where each bead represents an amino acid (known as Gō-like models) is represented by two, four, or more beads, depending on the level of structural accuracy needed for a given simulation. This approach has been successfully applied to observe the transition rates and track the protein-ligand binding[159].

2. Steered molecular dynamics

Steered molecular dynamics (SMD) relies on introduction of an external force to probe the mechanical properties of proteins which extends the scope of MD simulation. It follows either a time-dependent force or constant velocity protocol and restricts the degrees of freedom which steers the system along a

prescribed path. This method accelerates the process that would take long enough in a conventional MD and thus has been successfully implemented to replicate the AFM studies of immunoglobulin domains of titin and fibronectin besides many other significant observations[160, 161].

3. Targeted molecular dynamics

In Targeted molecular dynamics (TMD), harmonic constraints are used on root mean square deviation (RMSD) between the two structures i.e. starting and target, by reducing the sum over the distances of all atoms in every step of the simulation. A system is guided from a starting conformation to a desired "final or target" conformation. This approach has been successfully implemented to observe the conformational transitions in various proteins like insulin[162], glukokinase, etc.

4. Normal mode analysis

Normal mode analysis (NMA) decomposes the structural fluctuations of the minimized conformation of a protein into harmonic orthogonal modes. The low frequency modes are examined to extract the most frequent structural transitions. The two characteristics of these low frequency modes i.e. accessibility at lower energy and cooperativity amongst the atomic motions, make this method suitable to observe the signal propagation events[163].

5. Other approaches

There are several other methods to extend the capabilities of an all-atom MD simulation like **Biased molecular dynamics (BMD)[164],** which overcomes the potential energy barriers using a biased potential that applies an external perturbation to the system's reaction coordinate on an as-needed basis through the

course of the simulation. Another technique called **REMD i.e. Replica exchange molecular dynamics** is based on temperature[165]. This technique was first developed by Sugita and Okamoto to overcome the problems of protein folding with multiple minima. used to enhance the sampling of conformations from a standard molecular dynamics simulations. This is done by alternating the temperatures of systems with similar potential energies to sample conformations. This is turn reduces the potential energy barriers and thus allows sampling and surface might be overcome, allowing for the exploration of new conformational space[166].

In case of **Accelerated molecular dynamics (aMD)** which is an enhanced sampling method, the potential energy landscape is modified by raising the energy wells that are below a certain threshold level. This improves the conformational space sampling by reducing energy barriers separating different states of a system. This results in reduction of adjacent energy basin barriers and allows the system to sample conformational space that cannot be easily accessed in a classical MD simulation[167]. Another technique aimed at improving the sampling is **umbrella sampling** technique. Here, instead of constraining the reaction coordiantes, these are restrained and pulled to a target value by a bias potential[168]. This allows sampling of full momentum space. This technique is implemented in combination with either weighted histogram analysis method(WHAM)[169, 170] or umbrella integration[171].

Similarly, **conformational flooding** allows the system to explore new regions of phase space by lowering the free energy barriers using a multi-variate Gaussian potential[172].

In case of essential dynamics, geometric constraints are applied along selected principal modes to extrapolate the essential subspaces from short MD simulations.

2.3.5 Recent Applications and benchmarks

MD simulations have been successfully employed to understand several complex events like allosteric mechanisms, protein folding, energy transfers, protein-ligand interactions etc.

Most popular objects of MD simulation are the heme proteins[173] which serve as a base for new methods testing. Several studies have been performed to understand the protein transport phenomena like the diffusion paths of neuroglobin[174], free energy landscapes of cytoglobin[175]. Water transport and ion channels have been monitored quite remarkably for aquaporins and also the potassium channels[176].

Protein-DNA interactions have also been determined by application of MD for proteins like p53 where the binding modes to DNA quadruplexes were investigated to reveal details of Lac repression[177].

Protein folding events have been studied for villin[178] where the simulations were done for a microsecond becoming a benchmarks or record for length of simulation. Several studies have been done targeting the interactions of neuraminidases with antiviral drugs to probe the binding sites, flexibility and receptivity of a protein-drug interaction[179, 180].

Large scale motions of biosensors have been studied using essential dynamics which revealed the functionally important motions. Several crucial motions for ligases, Tobacco Mosaic Virus , ATPase[181] have been investigated by MD simulations or extensions of MD.

The most common example for study of allosteric transitions between the two T and R states has been in human haemoglobin which revealed many transient effects that would otherwise be difficult to observe in experimental conditions[182].

In terms of application in medical problems, MD simulations have been extensively applied to study Alzheimer's disease[183], thyroid hormone transport protein, transthyterin (TTR)[184], HIV virus[185] and osteoporosis.

Despite a huge array of successfully simulations and 30 years of rapid development, MD simulations still face problems in the areas of computing power [186]. One of the key challenges of computational biophysics is to simulate an effectively long timescale particularly of an order of a millisecond or even second to quantify the biological events discreetly and high computational demands prohibit routine simulations greater than a microsecond in length, leading in many cases to an inadequate sampling of conformational states.

Another limitation is the approximations of potential energy surfaces which might need superposition of quantum corrections. Availability of good potential functions is one of the main conditions for expansion of the area of applicability of the MD simulations to the realistic quantitative analysis of the behavior and properties of real materials.

In addition to ignoring the quantum-mechanical effects, molecular dynamics studies are also limited by the short time scales. In order to reproduce thermodynamic properties and fully elucidate all binding-pocket configurations relevant to drug design, all the possible conformational states of the protein must be explored by the simulation. This occurs on time scales that are much longer than those amenable to simulation. Using modern computers it is possible to calculate 10^6– 10^8 timesteps. Therefore we can only simulate processes that occur within 1 – 100 ns . This is a serious limitation for many problems that involve thermally-activated processes, cluster/vapor film deposition, annealing of irradiation damage, etc.

However, keeping all shortcomings aside, there have been rapid improvements over the years to augment the timescales and sample a wider conformational space by use of alternative methods as discussed in previous sections, application of much enhanced and sophisticated force-fields, novel

hardware designs[187], design of processors with enormous computing powers, the most recent and famous one been developed codenamed, Anton, a supercomputer capable of performing microseconds of simulation per day. With Anton, simulations longer than one millisecond[188] have successfully captured protein folding and unfolding as well as drug-binding events[189].

2.4 ANALYSIS METHODS

2.4.1 Principal Component Analysis

As described in the previous section, a typical MD trajectory consists of the information of time-evolution of the coordinates of all the constituent atoms forming the system being studied. Commonly used MD timesteps are on the order of 1 fs while the simulation time may range from a few to tens of nanoseconds, in any moderately sized configuration. A single resultant trajectory can thus easily contain a huge amount of data. For an N-atom system, the input dataset for PCA can be constructed as a trajectory matrix in which each column contains a Cartesian coordinate for a given atom at each output timestep. The major challenge for a dynamics simulation analysis lies in extracting the significant and corresponding motions from the data. Principal component Analysis[190] (PCA) is a common statistical tool that helps in identifying similarities and dissimilarities in a particular dataset and thus detects patterns in a particular dataset. Originally, the technique was introduced by Pearson for multi-dimensional least squares fitting[191] and later in 1933, Hotelling introduced PCA for analysing correlations within multi-dimensional data[192].

The central idea lies in finding a coordinate transformation which describes the major structural fluctuations in a set of new, collective coordinates[193]. This is done by expressing the fluctuations in terms of covariance and the transformation is brought by the diagonalization of the matrix. This diagonalization of the positional covariance matrix yields a new set of

orthonormal vectors which are called as eigenvectors. These orthogonal set of eigenvectors are also known as modes and describe the maximum variation in the observed conformational distribution [194, 195] Therefore for every eigenvector we have a corresponding eigenvalue describing magnitude of the motion along the eigenvector. Normally the eigenvectors are arranged in a decreasing order and thus the first eigenvector (principal component) describe the majority of covariance and thereby important structural fluctuations of the system[196].

Broadly speaking, there there are two main steps in PCA[165, 197, 198] technique:

1. The calculation of the covariance matrix, C, of the positional deviations.

The elements of the matrix, C are defined as

$$\boldsymbol{C_{ij}} = \langle (\boldsymbol{x_i} - \langle \boldsymbol{x_i} \rangle)(\boldsymbol{x_j} - \langle \boldsymbol{x_j} \rangle) \rangle \qquad (2.14)$$

2. The Diagonalization of this matrix

Where, x_i and x_j are the atomic coordinates of residues i and j respectively, and the brackets denote the ensemble average. Being symmetric in nature, C is diagonalized by an orthogonal coordinate transformation also known as eigenvector decomposition method[199].

$$\boldsymbol{C} = \boldsymbol{T} \cup \boldsymbol{T^T} \qquad (2.15)$$

Where T is a matrix of column eigenvectors and ∪ is a diagonal matrix containing the corresponding eigenvalues (λ). Together, the normalized eigenvectors represent the intrinsic collective coordinates of the protein in configuration space and the eigenvalues λ correspond to the mean square eigenvector coordinate fluctuation, and therefore, contain the contribution of each principal component to the total fluctuation. The eigenvectors (or modes or

collective degrees of freedom) are usually sorted according the decreasing eigenvalues. Therefore, for a system of N atoms; C is a 3N × 3N matrix. If at least 3N configurations are used to construct C, then 3N − 6 eigenvectors with nonzero eigenvalues are obtained. Six eigenvalues are exactly zero, which correspond to the eigenvectors describing the overall rotation and translation. Each of these 3D vectors describes the magnitude and direction of the RMS fluctuations of a given atom, within a given principal component. The MD trajectory can be projected onto each eigenvector by forming the dot product of atomic displacements with each eigenvector for all time steps. The resulting distribution of each projection would have a variance (and standard deviation); this is the physical meaning of the eigenvalues, which measures the spatial amplitude of each PC across the full trajectory.

The plot of eigenvalues against the index of the corresponding eigenvector shows that only a few first eigenvectors possess large eigenvalues with the higher indexed vectors having eigenvalues many orders of magnitude smaller. This visual inspection of the individual eigenvectors is helpful to estimate the nature of eigenmodes. As most of the variance in the original data is contained and described by only a few first modes, it is then imperative to presume that the motions along these 'essential eigenmodes' dominate the dynamics of the systems and contain the most important global information. Followed by identification of a subset of important eigenmodes, further analysis detailing each mode can be undertaken by projecting the original trajectory along a given (or a set of) eigenvector.

If μ_i is the i^{th} eigenvector of C (the i^{th} column of T), then the original configurations can be projected onto each of the principal components to yield the principal coordinates, $p_i(t)$ as follows:

$$\boldsymbol{p_i(t) = \mu_i \,.(x(t) - \langle x \rangle)} \qquad (2.16)$$

The variance i.e. $\langle p_i^2 \rangle$ equals to the eigenvalue, λ_i .

For visualization, these projections can be transformed back to their Cartesian coordinates:

$$\boldsymbol{x_i'(t) = p_i(t).\mu_i + \langle x \rangle} \quad (2.17)$$

Also, we can compare two sets of eigenvectors μ and ν with each other by taking their inner products:

$$\boldsymbol{I_{ij} = \mu_i . \nu_j} \quad (2.18)$$

Another factor called subspace overlaps i.e. summed squared inner products is often calculated which expresses the presence of dimensional subspace of set μ (n) within the dimensional subspace of set ν (m). For a full overlap, m should be larger (O=1).[200, 201]

$$\boldsymbol{O_n^m = \sum_{i=1}^{n} \sum_{j=1}^{m} (\mu_i . \nu_j)^2} \quad (2.19)$$

In this thesis, PCA is carried out using g_covar and g_anaeig modules of GROMACS 4.5.5[202].

As the large eigenvalues correspond to large fluctuations i.e. low frequency correlated motions, these are important in terms of studying enzymatic catalysis. With the large set of eigenvectors obtained after PCA, only a few are chosen for analysis. The choice may depend upon either visualization of the eigenvalue spectrum or a guess estimation to choose depending upon the study and requirements. There are several reasons stated by various authors emphasizing the importance of deciding the number of eigenvectors retained for analysis[203,

204]. To summarize those, under extraction can lead to the loss of relevant information and over extraction can be difficult to interpret and/or replicate[205].

A number of criteria have been proposed which guide the retention of eigenvectors. For instance, the latent root criterion, also known as the eigenvalue-one criterion or the Kaiser criterion[206] retains eigenvectors with eigenvalues greater than 1. The Cattell Scree test[207], which uses simple line plot representations of the eigenvalues displaying the relative importance of each component in fitting the data to major conformational reorganisation and the point of flattening of the line is selected as the cut-off for the eigenvector choice. Horn in 1965 proposed a method based on the generation of random variables which involves the comparison of the observed eigenvalues extracted from the correlation matrix with uncorrelated normal variables. It follows a Monte Carlo simulation process and is known as Parallel Analysis or PA[205]. Another method called the MAP test (Minimum Average Partial) was proposed by Velicer in 1976[208] , based on the application of PCA and in the subsequent analysis of partial correlation matrices. The method seeks to determine what components are common, and is proposed as a rule to find the best factor solution rather than cut off points. However, in practice mostly Kaiser criterion is used.

APPLICATIONS OF PCA

The applications of PCA are varied and diverse with the technique being implemented in all the fields. For the purpose of this thesis, we will highlight some of the applications in structural biology and Molecular Dynamics Simulations.

A PCA on a given MD trajectory gives insight into the collective dynamics of the system, revealing the dominant global motions e.g. a conformational change upon ligand binding or the motion of two domains connected by a hinge region.

In MD simulations, a protein can diffuse through the solvent and this motion contributes the most to the coordinate changes, and is often not of interest. In order to prevent detection of these degrees of freedom, each structure of the simulation is structurally aligned to a reference structure prior to calculating the covariance matrix. That way, the six global degrees of freedom are removed from the system as described previously. Nevertheless, this fitting procedure can be ambiguous in flexible systems and thus produce artefacts [209, 210]. The large number of steric hindrances and constraints including bonds, angles greatly reduces the degrees of freedom actually available for a protein. With the ordering of PCA eigenvalues in decreasing variances, the first few eigenvectors of a PCA describe anharmonic large-scale motions and together form the so called essential subspace[211]. The new coordinates given by the PCA eigenvectors are often called collective in the sense that in general all atoms contribute to each individual eigenvector. This way we can explore the high-dimensional data set. Most often, three-dimensional visualizations can be used to plot correlations within the components. [212, 213] As the principal components are uncorrelated, they may represent different aspects of the samples. This suggests that PCA can serve as a useful first step before clustering or classification of samples. However, deciding how many and which components to use in the subsequent analysis is a major challenge that can be addressed in several ways[214]. For example, one can use components that correlate with a phenotype of interest or use enough components to include most of the variation in the data[215]. PCA results depend critically on pre-processing of the data and on selection of variables. Thus, inspecting PCA plots can potentially provide insights into different choices of pre-processing and variable selection.

PCA is often implemented using the singular value decomposition (SVD) of the data matrix[212]. The sample-like Eigen array and the gene-like eigengene patterns are both uncovered simultaneously by SVD[213, 216]. Many applications beyond dimensional reduction, classification and clustering have taken advantage of global representations of expression profiles generated by this decomposition. Applications include identifying patterns that correlate with experimental artefacts and filtering them out[217], estimating missing data,

associating genes and expression patterns with activities of regulators and helping to uncover the dynamic architecture of cellular phenotypes[213, 216, 218].

Strengths and Weaknesses

There are several strengths and weakness of PCA which makes it a robust yet *interpret-at-your-own-risk analysis method.*

With the generation of new set of vectors equivalent to Cartesian coordinates with the focus being primarily on the first few vectors, this reduction of dimensionality means a loss of information and desired features in some way.

For example, in the case of a ligand binding to a protein, the fluctuation of a side chain in the binding pocket may be uncorrelated with the large global motions of the whole protein.

Nevertheless, PCA has proven extremely useful to identify large motions that are often related with the protein's function.

In PCA, a linear coordinate transformation is performed. If the relation between the individual components is not linear – think of a curved point cloud –, PCA will not be able to fully detect the underlying relation and will result in a higher dimensionality than required in case of nonlinearity. Several techniques have been adopted to non-linear cases, e.g. Kernel-PCA[219].

2.4.2 Root-mean-squared Displacement

Root-mean-squared displacement, commonly referred to as RMSD, is a numerical measure of the difference between the two structures. (normally calpha atoms or backone atoms. (Equation 2.20)

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{atoms}} [r_i(t_1) - r_i(t_2)]^2}{N_{atoms}}} \qquad (2.20)$$

N_{atoms} is the number of atoms whose positions are being compared, $r_i(t)$ refers to the position of atom i at a time t. Therefore for each timestep (frame), there is one RMSD value. While fitting, the aim is to reduce the difference between the two structures by finding relative orientations of the two molecules where there is minimal distance[193]. Its applications are diverse and include monitoring structural changes in simulations of protein folding and dynamics[154, 221-226], evaluating the quality of structure prediction schemes[227-229], comparing the diversity of model structures derived from experiments[230, 231], assessing the properties of modelling approaches at different levels of resolution[232], and defining high-resolution shapes of polymers[233].

2.4.3 Root-mean-squared fluctuation

The variability in the conformation of trajectories can be monitored by calculating the root mean square fluctuations (RMSF) for individual atoms. The main difference between RMSD and RMSF is that RMSF is calculated over time. RMSF for a specific number of structures is defined as a root mean-square-average distance between an atom and its average position in a given set of structures and characterizes local changes along the protein chain. The RMSF captures, for each atom, the fluctuation about its average position. This gives insight into the flexibility of regions of the protein and corresponds to the crystallographic B-factors (temperature factors) and thus is a measure of the deviation between the positions of particle i and some reference position.

The RMSF for residue i is computed as,

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{j=1}^{T}\left[r_i(t_j) - r_i^{ref}\right]^2} \qquad (2.21)$$

where T is the trajectory time over which the RMSF is calculated, r_i^{ref} is the reference position of residue i, which is time-averaged position of the particle i.e. $r_i^{ref} = \hat{r_i}$. For every residue we have one RMSF value.

These values can also be compared to the isotropic atomic crystallographic B-factors, which are related by[234]

$$B_i = \frac{8\pi^2}{3} RMSF_i^2 \qquad (2.22)$$

2.4.4 Correlation matrices

As is the case with covariance matrices described in the previous section, the correlation matrix is also one of the important things to examine when looking at data. It is the starting point for the study of principal components and factor analysis. The benefit of the correlation matrix is that it is simple, and depicts the nature of correlation between the variables.

MATHEMATICAL DERIVATION

Correlation matrix is derived from variance-covariance matrix.
So, if we consider the sample variance as follows:

$$s^2 = \frac{\sum_{i \in s}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i \in s}(x_i - \bar{x})(x_i - \bar{x})}{n-1} \qquad (2.23)$$

Here, i∈s means the number of observations to be calculated i.e. for example in case of protein structures, this is the number of atoms for which the variance has been calculated and thus the value of i would

range from 1 to say for instance 1992 (no. of calpha atoms of our studied protein structure i.e. lmpyk).

And further,

$$\boldsymbol{SS_{xy}} = \frac{\sum_{i \in s}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})}{\boldsymbol{n-1}} \qquad (2.24)$$

Being a sum of squares, it can be generalized for a three variate system with x,y,z co-ordinates (for studying protein molecules) into a compact form using matrix notation as follows:

$$\sum = \frac{\mathbf{1}}{\boldsymbol{n-1}} \begin{bmatrix} \boldsymbol{SS_{xx}} & \boldsymbol{SS_{xy}} & \boldsymbol{SS_{xz}} \\ \boldsymbol{SS_{yx}} & \boldsymbol{SS_{yy}} & \boldsymbol{SS_{yz}} \\ \boldsymbol{SS_{zx}} & \boldsymbol{SS_{zy}} & \boldsymbol{SS_{zz}} \end{bmatrix} \qquad (2.25)$$

This is the variance-covariance matrix and a slight modification in this matrix yields correlation matrix with the latter relating the variates directly. Thus the diagonal values in correlation matrix will be 1.

The element of correlation matrix is defined as follows:

$$\boldsymbol{\rho_{xy}} = \frac{\boldsymbol{SS_{xy}}}{\sqrt{\boldsymbol{SS_{xx}}}\sqrt{\boldsymbol{SS_{yy}}}} \qquad (2.26)$$

And the correlation matrix[125] is defined as

$$c_{ij} = \frac{\langle x_i . x_j \rangle}{\sqrt{\langle x_i^2 \rangle \langle x_j^2 \rangle}} \qquad (2.27)$$

x_i and x_j are difference vectors between the $i^{th}$and j^{th} Ca atom, respectively, and their average positions in the molecule-fixed frame.

One of the important questions which concern a particular MD simulation is the accuracy of determining the atomic motions of the studied system. It is bound to include artefacts from inaccuracies in force-fields, insufficient sampling, and convergence problems and so on. Keeping these artefacts in mind, the data obtained in MD simulations can be plotted to quantify and examine some of the structurally relevant and important motions governing a particular system of interest. The functions of proteins are linked to structural and molecular dynamics[29] and the enzymatic activities and ligand binding events are realized through conformational motions[235]. Thus, conformational motions determine the energetic, kinetics and strength of thermodynamic driving forces [236-238]. As in the case of allostery and domain motions, these are typically collective in nature[239].

Thus, correlations measured between fluctuating Calpha carbon atoms would provide insight about protein function and pairwise correlation of residues [124, 240, 241]. As the residue clusters are likely to involve inter-domain communication, they do not show individual residue interaction that could be important for the protein functional mechanism. This provides a starting point for experimental and further computational studies designed to model conformational changes in the protein.

The measure of correlation between the fluctuations ΔR_i and $\Delta R_j$of ith and jth alpha carbons can be assessed by finding the projection of one on the other, i.e., $\Delta R_i . \Delta R_j$at every instant and averaging over the full trajectory. The average, $\langle \Delta R_i . \Delta R_j \rangle$,if positive, indicates that the two residues move, on the average, in the

same direction. A negative correlation, or anticorrelation, indicates that the two atoms move in opposite directions. If two residues are displaced equally along the same direction, then their motions will be positively correlated and the distance between them will not change. If, on the other hand, they move in opposite directions, their motion will be negatively correlated and the distance between them will either increase or decrease. Positively correlated motions represent rigid body motions. Negatively correlated motions, resulting from distance changes between two residues, represent the interactions between them.

The covariance matrix,$C_{ij} = \langle \Delta R_i \Delta R_j \rangle$ given in the previous section is a 3Nx3N matrix where N is the number of atoms. The motion of each atom appears as erratic fluctuations in the molecular dynamics trajectory. However, there are strong correlations between the motions of atoms. PCA allows for organizing the motions into organized patterns. It is a multivariate statistical technique to find atomic correlations in the Cartesian coordinate space [242, 243]. The first step in PCA after performing the simulation is generation of a covariance matrix, C from a trajectory, where the element C_{ij} is obtained as:

$$\boldsymbol{C_{ij} = \langle \Delta R_i \Delta R_j \rangle} \qquad (2.28)$$

where, ΔR_i is the instantaneous fluctuation, $\Delta R_i = (R_i - \langle R_i \rangle)$, of the position $R_i = \{x_i y_i z_i\}$, of the ith atom from the time average, $\langle R_i \rangle$, x_i, y_i, z_i denote the Cartesian coordinates of the ith atom. The covariance matrix is then written as

$$\boldsymbol{C_{ij} = \langle (R_i - \langle R_i \rangle)(R_j - \langle R_j \rangle) \rangle} \qquad (2.29)$$

This is a product of all coordinates and thus yields $3N - 6$ vectors from those Cartesians coordinates. Here, N is the number of atoms considered in the analysis.

In order to simplify the analysis, we consider only the alpha carbons of the protein and consider only the components of the inner products $\Delta R_i . \Delta R_j$ of the fluctuations of the i[th] and the j[th] alpha carbons. With this simplification, the covariance matrix, which we now call the correlation matrix, reduces to an NxN matrix

$$\boldsymbol{Corr_{ij}} = \langle \Delta \boldsymbol{R_i} . \Delta \boldsymbol{R_j} \rangle \qquad (2.30)$$

A scalar correlation matrix can then be defined in Cartesian space by the following equation

$$\boldsymbol{Corr_{ij}} = \frac{\langle (\boldsymbol{r_i} - \langle \boldsymbol{r_i} \rangle)(\boldsymbol{r_j} - \langle \boldsymbol{r_j} \rangle) \rangle}{\sqrt{\langle (\boldsymbol{r_i} - \langle \boldsymbol{r_i} \rangle)(\boldsymbol{r_i} - \langle \boldsymbol{r_i} \rangle) \rangle \langle (\boldsymbol{r_j} - \langle \boldsymbol{r_j} \rangle)(\boldsymbol{r_j} - \langle \boldsymbol{r_j} \rangle) \rangle}} \qquad (2.31)$$

to denote the correlation between atoms i and j[244]. The value of $Corr_{ij}$ ranges from -1.0 to 1.0, with positive and negative values indicating correlated and anticorrelated motion. r_i and r_j are the position of atoms i and j and brackets denotes the trajectory average over snapshots. The matrix is calculated across all a carbon atoms of the protein.

The diagonal elements of the correlation matrix are the mean squared fluctuations of the alpha carbons which are related to the experimentally determined B-factors as

$$\langle \Delta \boldsymbol{R_i}^2 \rangle = \frac{3}{8\pi^2} \boldsymbol{B_i} \qquad (2.32)$$

RMSF[125] for each ensemble can be calculated as:

$$\boldsymbol{RMSF_i} = \sqrt{\langle \boldsymbol{r_i^2} \rangle} \qquad (2.33)$$

2.4.5 Distance Fluctuation analysis

Molecular dynamics trajectories of the apo and the holo structures give information on how much time the system spends in a given state. Structural and dynamic differences in the apo and holo structures are analysed here using a novel plot of pair-wise distance distributions. In order to track the interaction between a pair of residues over the course of an MD simulation the inter-atom distance is calculated at each time point and the frequency distribution is plotted. In many cases, as expected, the inter-residue distances show a normal distribution though frequently bimodal distributions or skewed distributions are observed. This graphical summary provides a convenient way of comparing the behaviour of holo and apo forms of LmPYK .

The dispersions are obtained from the relationship

$$\left\langle (\Delta R)^2 \right\rangle = \left\langle R^2 \right\rangle - \bar{R}^2 . \qquad (2.34)$$

2.5 High Performance Computing

High Performance computing refers to harnessing the power of supercomputers and parallel processing techniques to handle the complex computational processes. With the capability to handle and analyse enormous data at high speeds, it reduces the time from months in standard computers to days or minutes. In this research, two of UK's high end computing resources were used i.e. HECToR and ARCHER. HECToR[245] was the UK's high-end computing resource, funded by the UK Research Councils. The service ran from 2007 to early 2014 and has now been superseded by ARCHER[246]. These are part of a part of the PRACE initiative to provide access to a European pool of supercomputers.

The stated aim of the high performance computer project was a world-class supercomputer located and run in the UK, and provides an invaluable resource for researchers who study problems with a global impact.

	HECToR	ARCHER
Type of Supercomputer	Cray XE6	Cray XC30
Number of compute nodes	2816	3008
Processor per node	2	2
Number of compute processor cores	90,112	72,192
Processor core type	AMD 2.3 GHz 16-core processors	2.7 GHz, 12-core E5-2697 v2 (Ivy Bridge) series processors
Processor core theoretical peak	800 Tflops	1.65 Pflops
Main memory per node	32 GB	64 GB
Number of login nodes	2816	2632 standard, 376 high memory
Switch interconnect	Cray Gemini communication chips	Aries interconnect
Interconnect arrangement	3D torus arrangement with 10 links per router	2D all-to-all electric connections with 84 optical links per group
MPI bandwidth	Peak bi-directional bandwidth 8 GB/s (per link)	peak bisection bandwidth 7200 GB/a
MPI latency	1-1.5µs	~1.3µs
File system	Lustre	3 classes: the high performance, parallel /work filesystems, the NFS /home filesystems, and the Research Data Facility (RDF) filesystems for long term data storage.
Back up disk	168 TB with additional 1.02 PB on esFS archive	RDF; 7.8PB disk, with additional 19.5 PB
Batch system	PBSpro	PBS

Table 2.1: HECToR and ARCHER specifications/ hardware details.

CHAPTER 3:

CASE STUDY I

PYRUVATE KINASE

3 PYRUVATE KINASE

3.1 INTRODUCTION

3.1.1 Architecture of Pyruvate Kinase

Pyruvate kinases catalyze the rate-limiting, irreversible reaction of glycolysis (Figure 3.1) in which phosphoenolpyruvate (PEP) and ADP are converted into pyruvate and ATP (Figure 3.2)[247]. The enzyme, which is primely allosteric, plays an important role in metabolic intersection with the substrates and the product being involved in a number of metabolic pathway[248]. PYK has been identified from various eukaryotes and prokaryotes and has been found to exist as a homotetramer in most of the cases with each chain consisting of 500 residues. Most bacterial PYKs and mammalian isoenzymes, R (expressed in erythrocytes), L (in liver), and M2 (in kidney and lung) are regulated by fructose 1,6-bisphosphate (FBP)[249] while fructose 2,6-bisphosphate is the allosteric effector in trypanosomatid protozoans. PYK along with hexokinase and PFK, constitutes a prime target for chemotherapy and to design drugs against cancer.

There are crystal structures of PYK available from cat muscle, E. coli, rabbit muscle[250, 251], S.cerevisiae[252], Escherichia coli[253], and Leishmania mexicana[254], all of which share somewhat similar architecture. The protein is composed of four subunits, each of which consists of three distinct domains: the A domain with the classic TIM barrel [$(\alpha/\beta)_8$] topology, the mobile, β-stranded B domain inserted between the α- helix and β-strand of the A domain with a somewhat irregular fold, and the C domain with an α+β organization. Also, there is an additional small N-terminal domain, which is absent in the prokaryotes. The four identical subunits are assembled to form a tetramer with D2 symmetry (222 symmetry i.e three, twofold rotation axes intersecting each other at right angles). The inter-subunit interactions are through the A and C domains, defining two interfaces. A long A-A interface related by the

vertical twofold axis and a short C-C interface along the horizontal axis(Figure 3.3).

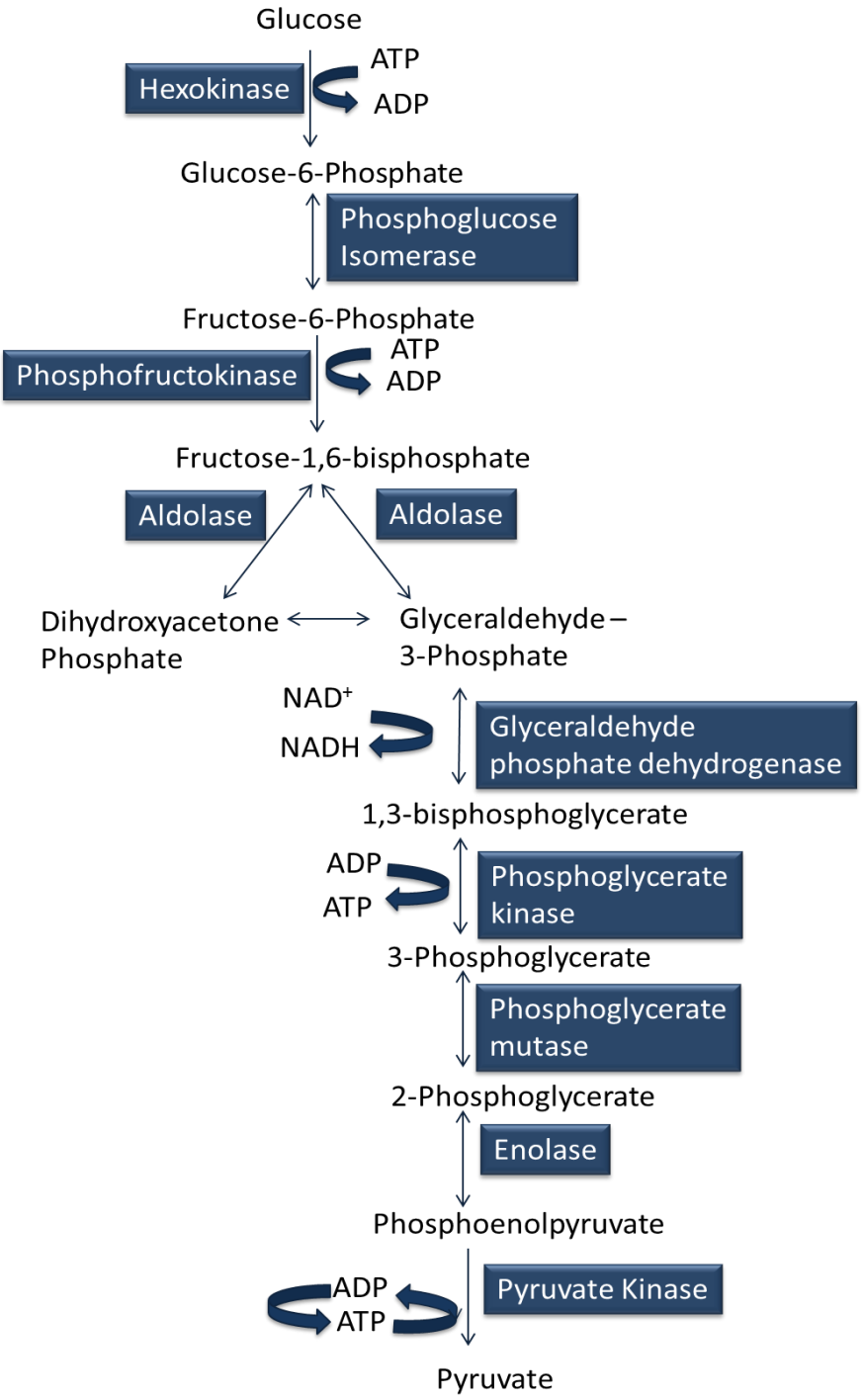


Figure 3.1: The ten enzymes involved in glycolysis

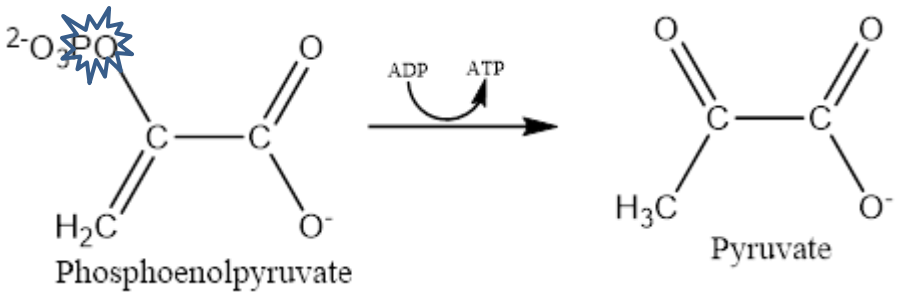


Figure 3.2: Pyruvate kinase converts Phosphoenol pyruvate to pyruvate where the phosphoryl group of phosphoenolpyruvate (PEP) is transferred to ADP to form pyruvate and ATP.

Active site

The active site faces the cleft between and A and B domains, lying closely towards the C-terminal position of the A domain (α/β)8 barrel[255] . The ligands for the binding sites and the cations are present in the active site[251]. The ADP/ATP binding site lies closer to the centre of the molecule[256]. The residues associated with the binding of the substrate, PEP and ADP, are found to be conserved for most of the PYKs[257].

Effector site and activity

PYK displays sigmoidal kinetics towards the substrate PEP and the activity is controlled by a number of physiological effectors like H+, Mg^{2+}, Mn^{2+}, and K^{+} [258]. The heterotropic control is established by allosteric effector molecules which differ in various organisms[259]. The allosteric control is exerted at the C-terminal region where intersubunit contacts between the domains are formed. As the allosteric effectors differ in PYK, the residues of effector site do not exihibt the same degree of conservation as observed in the active site[260-262]. The mammalian M2, R and L isozymes are activated heterotropically by F1,6BP with the exception of M1, where the enzyme

shows hyperbolic kinetics. With the exception of B. licheniformis, B. psychrophilus, B. stearothermophilus which are activated by adenosine monophosphate (AMP) and pentose monophosphates, including ribose 5-phosphate, almost all bacteria have FBP as the allosteric effector molecule[263-265]. In the case of protozoans i.e. Trypanosomatids and leishmania species, fructose 2,6-bisphosphate is the allosteric effector molecule. Also, it has been observed that the binding of F-2,6BP is tighter in the case of T. Brucei in comparison to leishmania PYK[261, 266, 267]. These regulatory differences between parasites and humans can provide a probable opportunity for drug development[268].

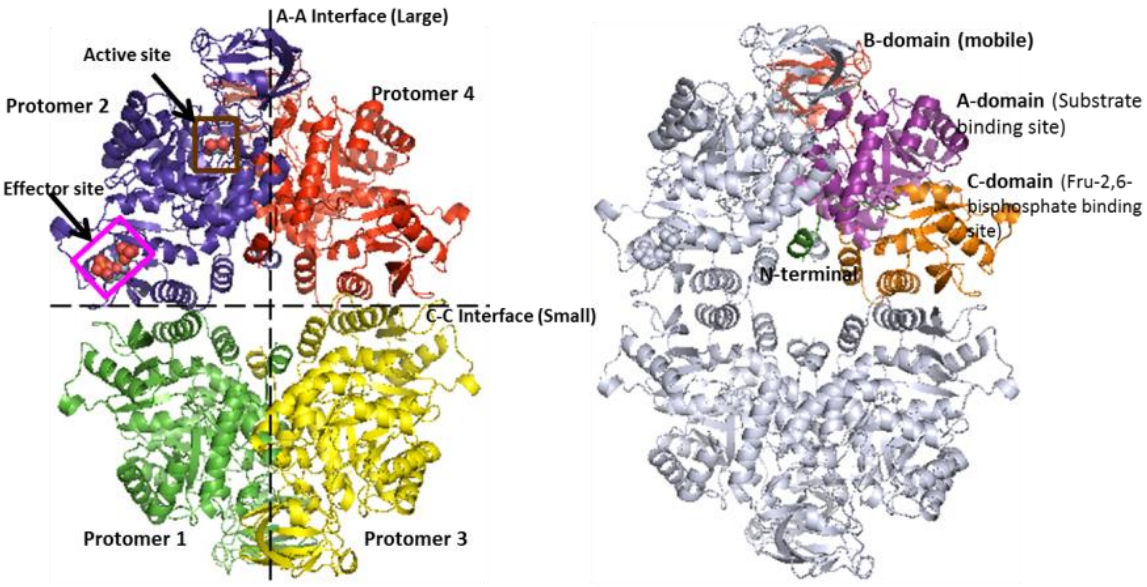


Figure 3.3: The crystal structure of LmPYK colored according to the chains (left) highlighting the interfaces and domain (right).

3.1.2 Scope of this study

In this research, we have employed molecular dynamics (MD) to understand the allosteric mechanism and conformational changes that follow upon binding of the substrates on pyruvate kinase from Leishmania mexicana.

We have also tried to answer as to how binding of FBP at a site over 40Å from the active site enhances the activity of the enzyme.

LmPYK was chosen as a model to study the allosteric activation of PYK as it is the only example where X-ray structures for all relevant liganded states are available; namely a T-state apo structure with no ligands bound; a structure in which only the effector molecule (F26BP) is bound; a substrate-only structure and a fully liganded structure with both substrate analogue and effector bound. Overlays of the tetramers for each of these structures showed that only the unliganded T-state tetramer adopted a significantly different conformation from the other three R-state structures which had very similar (RMSD < 1 Å) conformations. The conformational transitions observed on going from the T to R state preserved the 222 symmetry of the tetramer but involved changes in the relative orientations of the protomer subunits and also changes in the interprotomer interactions across the A-A' and C-C' interfaces. The major differences between the T and R state structures can be described as rigid-body rotations of the protomer domains and led to a description of the a 'rock-and-lock' mechanism for allosteric activation of LmPYK.

The 'rock-and-lock' mechanism fits with the original Monod-Wyman-Changeux (MWC) ideas of an essentially 2 state equilibrium between inactive T-states and active R-states. Despite the insight obtained from the R and T state LmPYK X-ray structures there are still some unanswered questions about the nature of the mechanism by which the enzyme shifts between the "off" and "on" states and in particular the manner by which the activity of the enzyme is altered by the binding of regulatory molecules. In this regard it is intriguing to note from the X-ray structures that the bound F26BP effector molecules do not make interprotomer contacts. Instead they bind to a rather disordered loop region. Experimentally it was found using thermal denaturation studies that binding of the effector to LmPYK significantly increased the thermal melting temperature of the protein and the suggestion has been made that the approximately 7-fold increase of the catalytic rate of LmPYK upon F26BP binding is caused by a

rigidification of the enzyme. Molecular dynamics provides an ideal tool to explore this possible mechanism [220].

The MD simulations for the complete tetramer in the absence of F26BP (the Apo structure), and with F26BP bound, (the Holo structure) are compared with MD simulations of the isolated monomer. The comparison identifies allosterically regulated movements of the tetramer as induced by FBP binding to be distinguished from movements of isolated monomer modes. Comparison of correlations of residue fluctuations from the tetramer and monomer trajectories show the role of thermal fluctuations on the individual domains of the protomers as well as providing insight into conformational transitions of the tetramer which are relevant to the T to R transition. The 'cooling' effect of FBP binding which causes a general reduction in protein flexibility is also analysed. These thermodynamic effects along with specific structural rearrangements, particularly across the tetramer interfaces, provide a detailed picture of the allosteric mechanism of LmPYK.

3.1.3 Dynamic investigation of pyruvate kinase

This section entails the details of molecular dynamics simulation designed to investigate the conformational changes and allosteric movement of pyruvate kinase. This case study helps us to understand the function of protein by local conformational flexibility and provides the quantification of the evaluation of modes/Principal components. The fluctuations and mode analysis produce results which complements the experimental data.

The use of PCA on MD simulations data is computationally expensive due to the complexity and size of the proteins and enormous sets of snapshots generated. Earlier studies on proteins, and indeed even many modern studies focus more on alpha carbons rather than all atom analysis to reduce the complexity of PCA and also to ease the process of data handling which otherwise

might crash the system with overload. We have performed PCA on alpha carbon atoms to enhance our understanding of protein motion.

There have always been arguments regarding the computational and experimental coherence. Also, there is a question of whether the PCA analysis yields accurate protein conformational equilibrium fluctuations. Analysis of multiple properties of the system, e.g. eigenvalues, overlaps, correlations and distance fluctuations increase confidence in the reliability of this technique. The following sections will present the results of this analysis.

3.2 MOLECULAR DYNAMICS SIMULATIONS

3.2.1 System preparation

SIMULATION SET A – 3HQQ structure

Starting Structure:

The structure of Leishmania mexicana pyruvate kinase was modeled from the crystallographic coordinates as obtained from the Protein Data Bank Entry 3HQQ (resolution 5.07 Å) with the allosteric ligand Fructose-2,6-bisphosphate[220] (FBP). This structure was taken as a tetramer for simulation and consists of four identical subunits. Each subunit consists of four distinctive domains: the N-terminal, A-, B- and C-domains. The N terminal is composed of residues 1-17, A-domain has residues 18-88 and 187-356, B-domain which forms the mobile lid has residues 89-186 and the C-domain which incorporates the effector site stretches from residues 357-498.

Four separate MD simulations were carried out using the 3HQQ crystal structure(Table 3.1). The 'apo' and 'holo' tetrameric structures were created from identical crystal coordinates apart from removing the FBP molecule to provide the apo structure. The same procedure was followed for the monomer simulations in which the protomer was extracted from the tetramer (Figure 3.4). All the MD simulations were carried out using GROMACS (GROningen MAchine for Chemical Simulations) package version 4.5[269] with AMBER99sb-ildn[270] forcefield parameter set. The starting structures were solvated in a dodecahedron box placed at a distance of 0.9 nm from the box boundary. Simple point charge (SPC) water molecules were used to fill the box, followed by the addition of sodium (Na+) and chloride (Cl-) ions to neutralize the system. The final systems contained 84000 (monomer) and 300000 atoms (tetramer) with 76000 and 272000 water molecules respectively.

In order to maintain a constant temperature of 318 K, the protein and non-protein atoms were coupled to their own temperature baths using the V-rescale thermostat and a time step of 5fs. This was followed by an NPT equilibration, in which the pressure was maintained isotropically at 1 bar using the Berendsen thermostat[194] with a coupling constant of 0.1 ps. The water molecules and bond lengths were restrained using the SETTLE[271] and LINCS[272] algorithm respectively. A single cut off of 1 nm was used for the treatment of Van der-Waals interactions. Long-range electrostatics were treated using the Particle-Mesh Ewald (PME) method with 0.16 FF grid spacing and 4^{th} order B-spline interpolation for the reciprocal sum space. The systems were also relaxed by 1000 steps of steepest descent energy minimization procedure prior to the simulations. The snapshots were saved every 2ps, thereby yielding 10000-20000 frames (Table2). Periodic Boundary conditions were applied in all the directions. The cumulative simulation time for all four trajectories was ~ 244 ns.

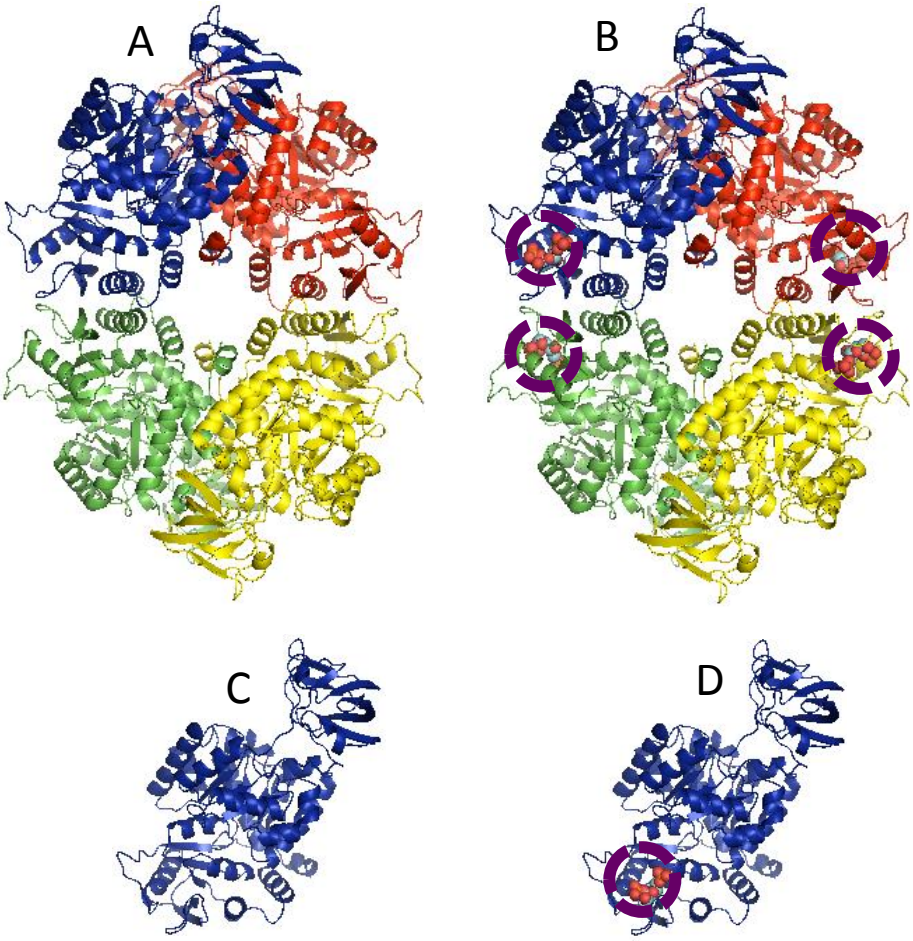


Figure 3.4: The four structures used for simulation derived from 3HQQ crystal structure. (A) Apo Tetramer, (B) Holo Tetramer, (C) Apo Monomer, (D) Holo Monomer. The ligands are highlighted by purple circles in the holo structures.

SIMULATION B – 3HQP set-up

Starting structure:

Another simulation was performed on the fully liganded crystal structure of LmPYK from the crystallographic coordinates as obtained from the Protein Data Bank Entry 3HQP (resolution 2.30 Å). The structure which is the fully ligated R state

of LmPYK contains the allosteric activator; FBP bound at the C-C interfaces, ATP and oxalate (substrate analogue of PEP) at the active site and magnesium ion. (Figures 3.5 & 3.6).

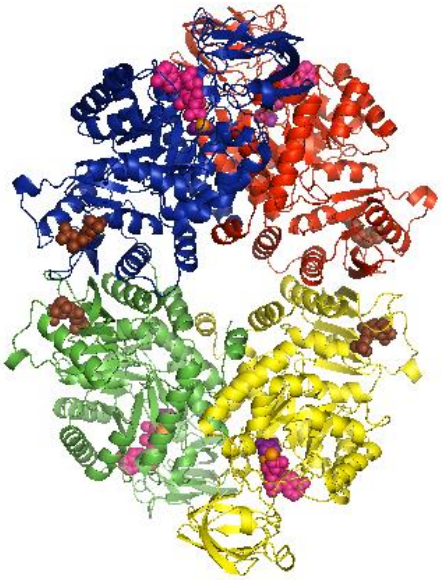


Figure 3.5: 3HQP crystal structure colored according to the different protomers. Also highlighted are the ligands in sphere representation; the allosteric activator FBP (brown), susbtrate oxalate (purple), ATP (magenta) and magnesium ions (orange).

The MD simulation of the fully ligated R state structure was carried out using GROMACS[197] (GROningen MAchine for Chemical Simulations) package version 4.6[165] with AMBER99sb-ildn[270] forcefield parameter set. The system was prepared in a similar way as stated above for the 3HQQ simulation. The final system has 312969 atoms (30408 protein atoms) and 282189 water molecules. Constant temperature of 318 K was maintained by coupling to the V-rescale thermostat[194] using a time step of 2fs. This was followed by an NPT equilibration, in which the pressure was maintained isotropically at 1 bar using the Parrinello-Rahman thermostat [273]with a coupling constant of 0.1 ps. The water molecules and bond lengths were restrained using the SETTLE[271] and LINCS[272] algorithm respectively. A single cut off of 1 nm was used for the treatment of Van der-Waals interactions. Long-range electrostatics were treated using the Particle-Mesh Ewald (PME) method with 0.16 FF

grid spacing and 4th order B-spline interpolation for the reciprocal sum space. The system was relaxed by 1000 steps of steepest descent energy minimization procedure prior to the simulation. The final production simulation was run for 82 ns with the snapshot saved every 2ps, thereby.

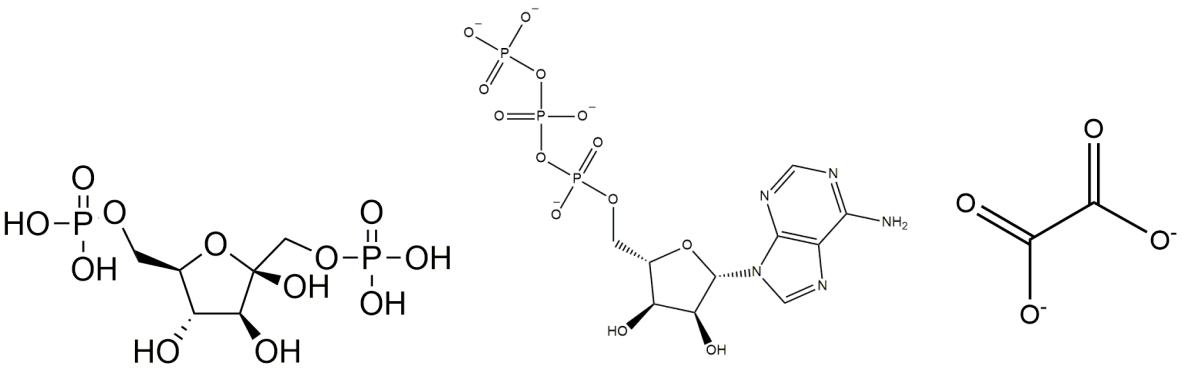


Figure 3.6: Chemical structures of the ligands used in LmPYK simulations. Starting from left, fructose-2,6-bisphosphate (FBP), adenosine tri phosphate (ATP) and oxalate (OXL).

3.2.2 MD simulation workflow

The standard procedure for MD simulation (see theory and methods) was carried out for both the sets. The schematic figure below illustrates the workflow in brief. (Figure 3.7)

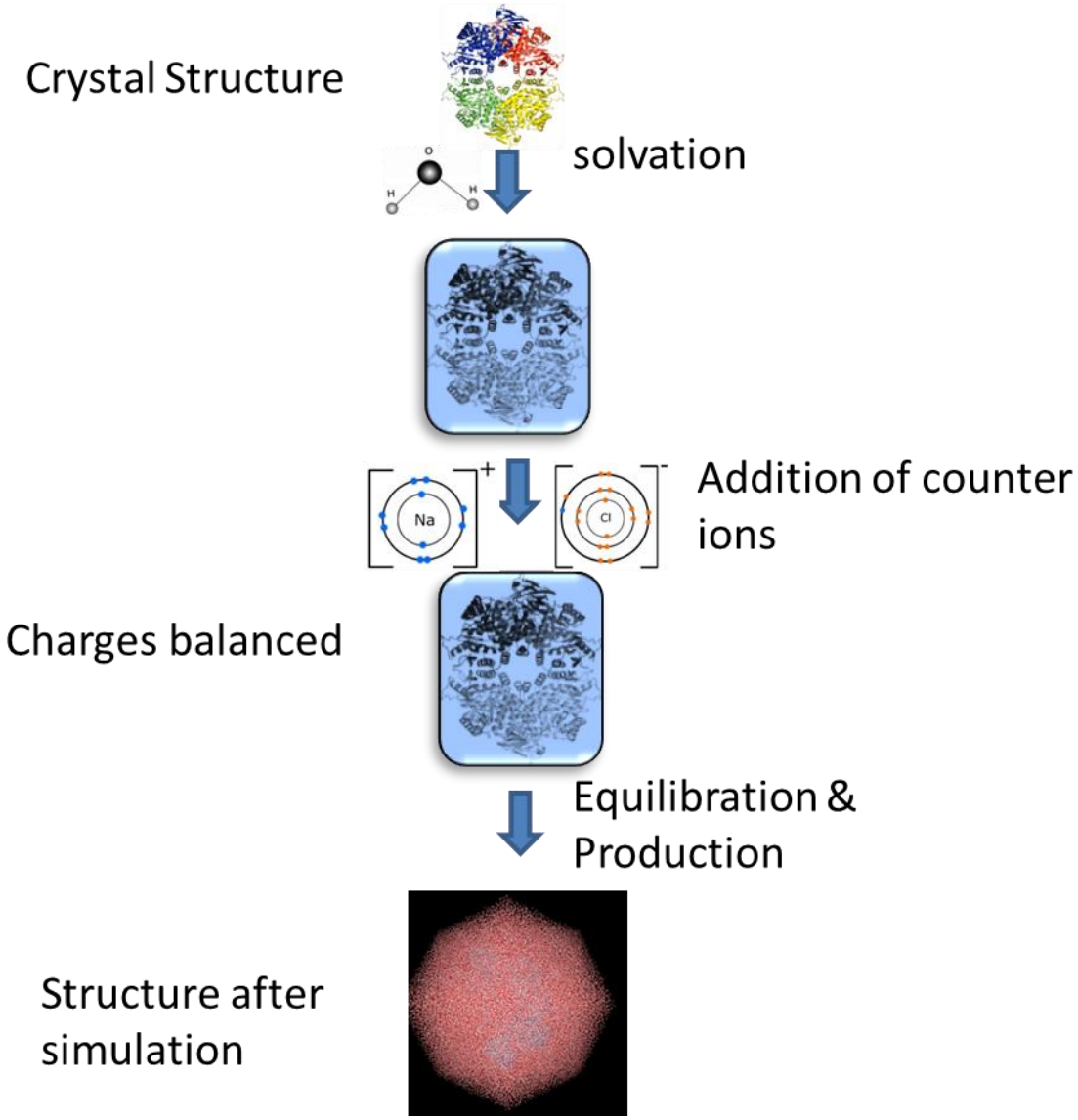


Figure 3.7: An illustration of the simulation steps wherein we start from the co-ordinates of the crystal structure and finally from solvation, neutralization and equilibration, we obtain our trajectories.

3.2.3 MD Parameters

A total of 5 independent simulations were run for LmPYK. Out of these, 4 simulations belong to the crystal structure 3HQQ and would be referred as part of simulation set A and the fifth simulation derived from 3HQP crystal structure will be referred to as simulation B. The following two tables

highlight the main parameters and structural design of the simulations. (Tables 3.1 and 3.2)

Structure	Ligands	No.of chains	Proposed state	SUPERCOMPUTER
SIMULATION SET A – PDB ID: 3HQQ				
Holo Tetramer	FBP	4	Active; R	HECToR
Apo Tetramer	None	4	Inactive	HECToR
Holo Monomer	FBP	1	Active; R	HECToR
Apo Monomer	None	1	Inactive	HECToR
SIMULATION B – PDB ID: 3HQP				
Tetramer	FBP, OXL, ATP, MG	4	Fully active R state	ARCHER

Table 3.1: Structural details of the simulation

Structure	Water atoms	Calpha atoms	System atoms	Time step (fs)	Number of frames	Simulation time (ns)
SIMULATION SET A – PDB ID: 3HQQ						
Holo Tetramer	272010	1992	305354	3	15642	48
Apo Tetramer	272043	1992	305251	5	16114	78
Holo Monomer	76353	498	84689	3	21764	65
Apo Monomer	76374	498	84676	5	10766	53
SIMULATION B – PDB ID: 3HQP						
Tetramer	282189	1992	312969	2	40833	82

Table 3.2: Description of the simulation parameters and trajectory details

3.2.4 HIGH PERFORMANCE COMPUTING

As stated earlier, the MD simulations were run on UK's national Supercomputers managed by the Edinburgh Parallel Computing Centre (EPCC). The reason for running on two supercomputers was that HECToR was superseded and replaced by ARCHER in 2014 (Figure 3.8). Technical specifications of these systems are described in the theory and methods section.

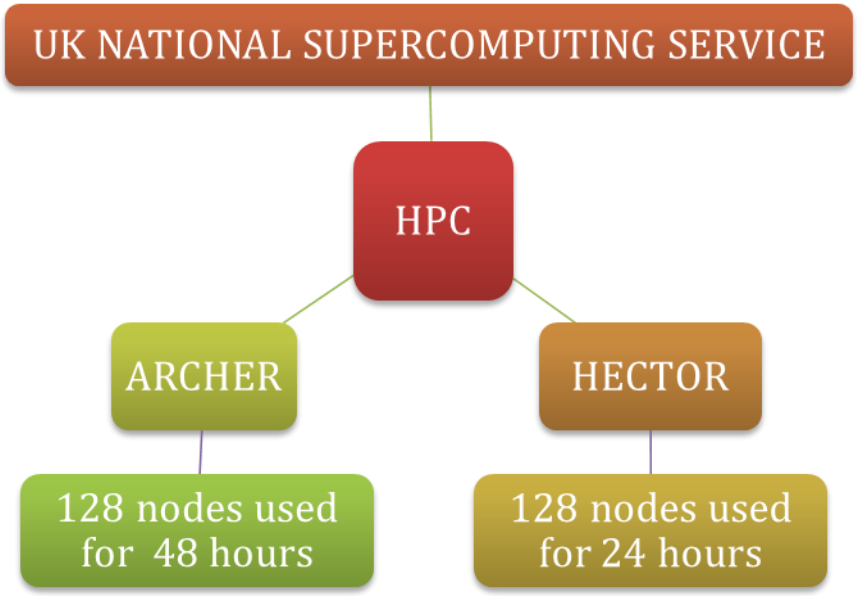


Figure 3.8: Flowchart of the Supercomputer used

3.3 Results

3.3.1 Leishmania mexicana crystal structure comparison

Selection of a starting model for MD simulations

The original paper that hypothesised a 'rock and lock' allosteric mechanism for LmPYK[220] has the following crystal structures:

3HQN = LmPYK	(T-state 2.0 Å)
3HQQ = LmPYK + FBP	(R'-state5.1 Å)
3HQP = LmPYK + FBP + ATP + oxalate	(R-state 2.7Å)

Figure 3.9 shows how the tetramer structures of these three different liganded states are related to each other by concerted rigid body rotations of the core A-C domains by between 1.5^0 and 8^0. The RMS fit for the individual

chains (summarised in Table 3.3.) are however very close with values between 0.6 Å and 0.3Å .

Although the resolution limit of 3HQQ structure is 5 Å, the structural information is both reliable and informative. This effector-only structure has very large cell dimensions of a= 243 Å. b = 254 Å and c = 892 Å. It diffracted to 2.75 Å, but due to the closeness of Bragg spots it was only possible to collect processible data at 5 Å. There are six tetramers in the asymmetric unit and rigid-body refinement of the AC domains for each of the 24 crystallographically independent chains was used to provide strong experimental evidence that the tetramers adopted a conformation close to the fully-liganded R-state structure (3HQP).

In selecting 3HQQ as the starting model for MD simulations we argued that the structures of the individual chains in the tetramer are reliable as they have an RMSfit of ~0.3 Å to the relatively high resolution structure 3HQP, however the small rigid body conformational change between the tetramers in 3HQQ , 3HQP and 3HQN are likely to provide information about the T to R conformational transition (governed by binding of the FBP effector molecule). Furthermore, using 3HQQ with the bound FBP ligand provided a straightforward way of generating a pseudo APO structure for a parallel MD run i.e. preferably starting with either the fully liganded 3HQP structure which would require the active site to be emptied to simulate an apo structure or using 3HQN would require an initial modelling /docking study to try and fit the F26BP molecule into a distorted effector binding pocket.

	3HQN (T-state) 3HQP (R state)	3HQQ (FBP-bound) 3HQN (T-state)	3HQQ (FBP-bound) 3HQP (R-state)
Tetramer RMS fit	3.6Å	2.5Å	2.5Å
AC core RMS fit	0.62Å	0.62Å	0.26Å
Relative rotation of AC core	5.2°	8°	1.4°

Table 3.3: Summary of the crystal structures of lmPYK

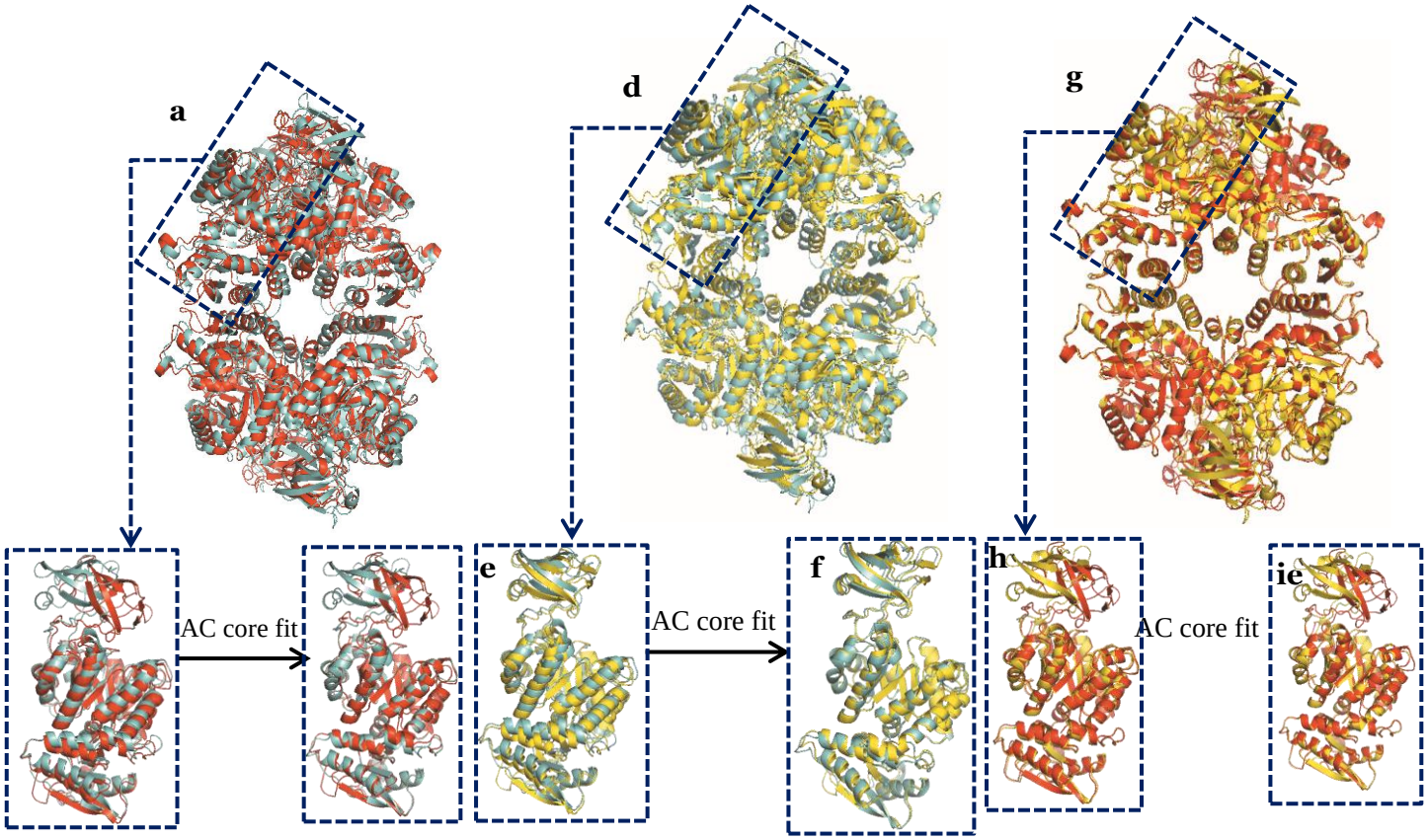


Figure 3.9: Structural overlays of the T and R state Leishmania mexicana pyruvate kinase crystal structure with 3HQQ (FBP-LmPYK)

a) superposition of 498 Cα atoms of all 4 tetramer chains of 3HQN (T state, cyan) and 3HQP (R state, red) gives an rms fit of 3.6 Å

b) detail from a) showing superposition of one of the tetramer chains

c) superposition of 18-88 and 187-498 Cα atoms for the A and C domains of one chain (rmsfit 0.6Å) achieved by a rigid body rotation of 5.2^0.

d) superposition of 498 Cα atoms of all 4 tetramer chains of 3HQQ (simulation structure, yellow) and 3HQN (T state, cyan) give an rms fit of 2.5 Å

e) detail from d) showing superposition of one of the tetramer chains

f) superposition of Cα 18-88 and 187-498 atoms for the A and C domains of one chain (rmsfit 0.6Å) achieved by a rigid body rotation of 8^0

g) superposition of 498 Cα atoms of all 4 tetramer chains of 3HQQ (simulation structure, yellow) and 3HQP (R state, red) give an rms fit of 2.5 Å

h) detail from g) showing superposition of one of the tetramer chains

i) superposition of 18-88 and 187-498 Cα atoms for the A and C domains of one chain (rmsfit 0.26Å) achieved by a rigid body rotation of 1.4^0

Calculation of the angle of rotation

The superposition of the three crystal structures was done in two steps in order to calculate the degree of rotation about the AC core region.

Step 1: First the tetramers of the two crystal structures were superposed onto each other.

Step2: Secondly, the superimposed coordinates of the structure were selected and only the A and C domains were fitted onto the monomer i.e. residues (1-88 and 187-357 of the A domain and residues 357-498 of the C domain) to obtain the following rotation matrices.

Step 3: Considering the rotation matrix is of the following form:

$$\begin{matrix} a1 & a2 & a3 \\ b1 & b2 & b3 \\ c1 & c2 & c3 \end{matrix}$$

Rotation angle θ = arcos[0.5x(a1+b2+c3-1)]

1. T and R state

$$\begin{matrix} 0.99176 & -0.09162 & 0.08951 \\ 0.09277 & 0.99565 & -0.00876 \\ -0.08832 & 0.01699 & 0.99595 \end{matrix}$$

Rotation angle θ = arcos[0.5x(a1+b2+c3-1)]

Rotation angle θ =arccos[0.5x(0.99176+0.99565+0.99595-1)]

=arccos(0.99168)

Rotation angle θ = 7.3°

2. Simulation Model with R state

$$\begin{matrix} 0.99968 & 0.02396 & 0.00808 \\ -0.02394 & 0.99971 & -0.00160 \\ -0.00812 & 0.00141 & 0.99997 \end{matrix}$$

Rotation angle θ = arcos[0.5x(a1+b2+c3-1)]

Rotation angle θ =arccos[0.5x(0.99968+0.99971+0.99997-1)]

=arccos(0.99968)

Rotation angle θ = 1.4°

3. Simulation Model with T state

0.99010	− 0.13995	− 0.01062
0.13958	0.98976	− 0.02996
0.01470	0.02818	0.99949

Rotation angle θ = arcos[0.5x(a1+b2+c3-1)]

Rotation angle θ =arccos[0.5x(0.99010+0.98976+0.99949-1)]

=arccos(0.989675)

Rotation angle θ = 8.2°

3.3.2 Convergence of the parameters

The root-mean-square deviation (RMSD) values of all the Cα atoms relative to the corresponding starting structures for three trajectories (3HQQ Apo tetramer, 3HQQ holo tetramer and 3HQP tetramer simulation) were examined to determine the system equilibrium (Figures 3.10-12). It is often considered that small RMSD values of one simulation indicate a stable state of the system and also suggest that the newly constructed models satisfactorily reproduced the experimental structures. However, the large RMSD values suggest large conformational changes of the investigated system. As is evident from the figures, there is an initial rapid rise which levels off after few nanoseconds and we do not observe a major drift from the starting structure. All the three trajectories reach a stable and equilibrated state quickly with the average RMSD values for simulations A and B being 2.5 and 1.9 Å, respectively (refer section 3.3.3). Also, the low RMSD value for simulation B provides the first evidence to

support our hypothesis of the allosteric ligand F2,6BP having a restraining effect on the enzyme.

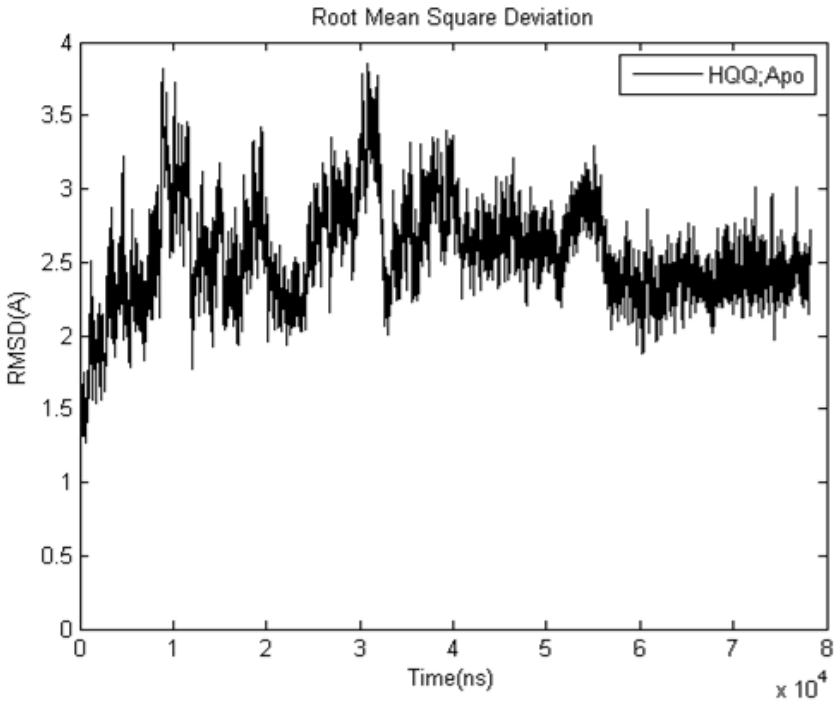


Figure 3.10: RMSD of 3HQQ apo tetramer simulation

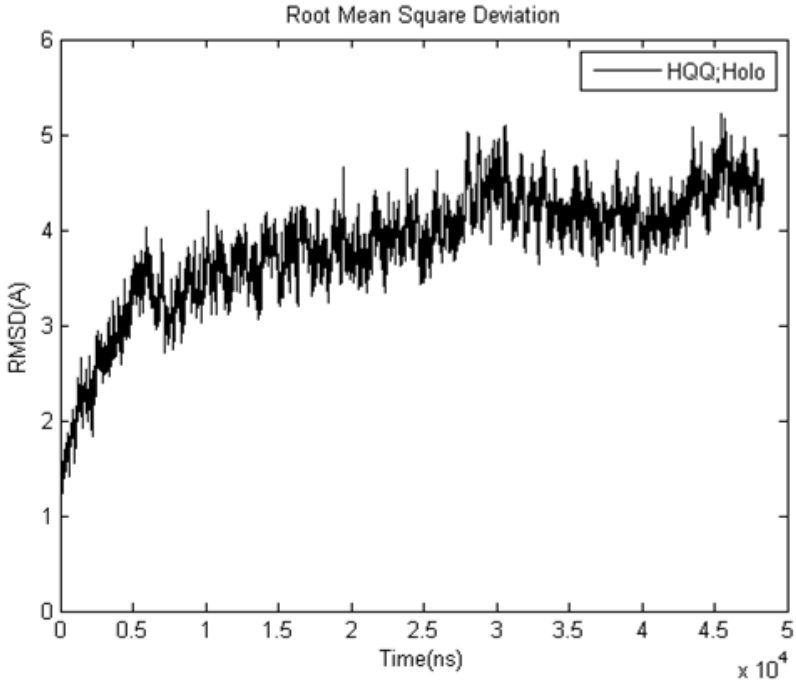


Figure 3.11: RMSD of 3HQQ holo tetramer simulation

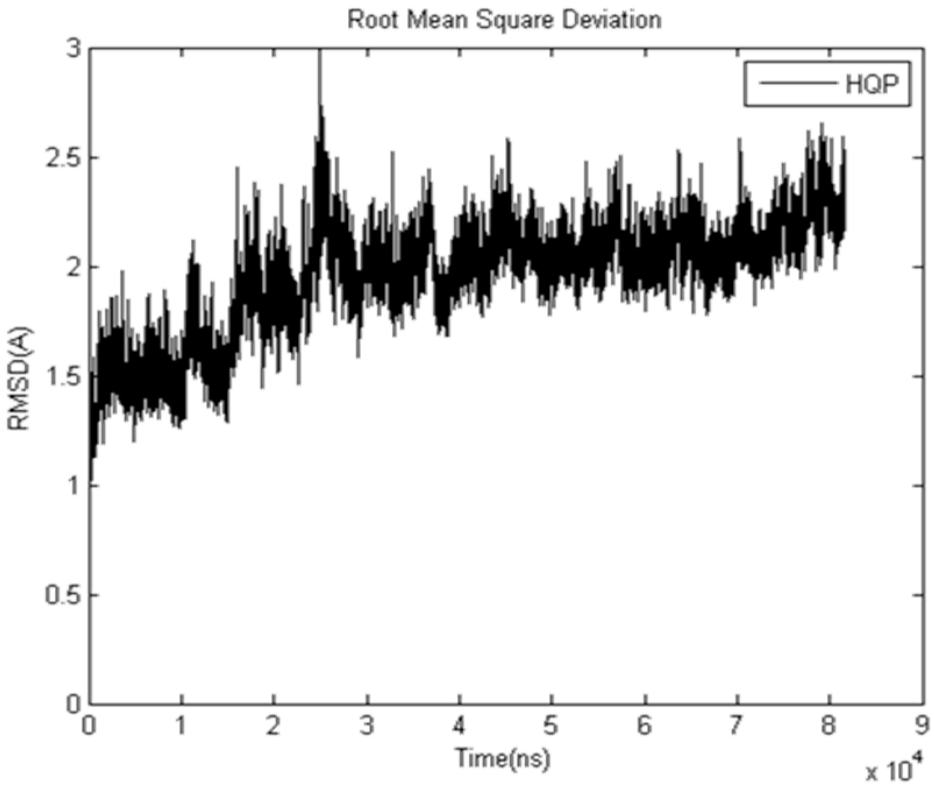


Figure 3.12: RMSD of 3HQP tetramer simulation

3.3.3 Analysis of B-factors

Thermal fluctuations of residues help to understand the protein function by local conformational flexibility. In order to investigate and explore the conformational variability of each trajectory, mean square fluctuations (MSF) of alpha carbon atoms was plotted with respect to the residue number to show the local conformational changes for all the three systems. Mean-square fluctuation about the average position is related to the B factors of crystallography and is also measurable by neutron scattering a[274] and by Mössbauer spectroscopy b[275]. The motion of each atom appears as erratic fluctuations in the molecular dynamics trajectory. However, there are strong correlations between the motions of subsets of atoms and these collective protein motions may be extracted from the elements of the correlation matrix. Each element of the matrix C_{ij} is obtained from the dot product $\langle \Delta R_i . \Delta R_j \rangle$ where ΔR_i and ΔR_j are the instantaneous fluctuation vectors of atoms i and j from their respective average positions over

the time of the simulation. Only the C_α atoms are considered in this analysis. When the index i is equal to j, we obtain the mean squared fluctuations, $\left\langle (\Delta R_i)^2 \right\rangle$, of the residues, which may be compared with fluctuations obtained from the experimental B-factors determined by X-ray crystallographic refinement (refer chapter2). Figure 3.12 shows the comparison between the measured B-factors from the apo X-ray structure (3HQN) and the calculated B-factors from the MD simulation of the apo tetramer. The experimental B-factors from the x-ray structure of apo-LmPYK (PDB code 3HQN) were transformed to mean-square displacement values using the expression: $\left\langle (\Delta R_i)^2 \right\rangle = \frac{3}{8\pi^2} B_i$.

Comparison of the x-ray data and simulation results shows that the latter agrees quantitatively with experiment except the B-domain cap region, residues 89-186 (Figures 3.3 and 3.13), where simulation results show a much larger amplitude of motion compared to the x-ray structure. The larger fluctuation of the B-domain may plausibly be attributed to the aqueous environment in simulations in contrast to a more restrictive crystal environment. Indeed in a number of different crystal forms of PYK the B-domain is found to be totally disordered showing no interpretable electron density while the rest of the structure is well defined[220] suggesting that the B-domain can adopt multiple conformations even in crystal form.

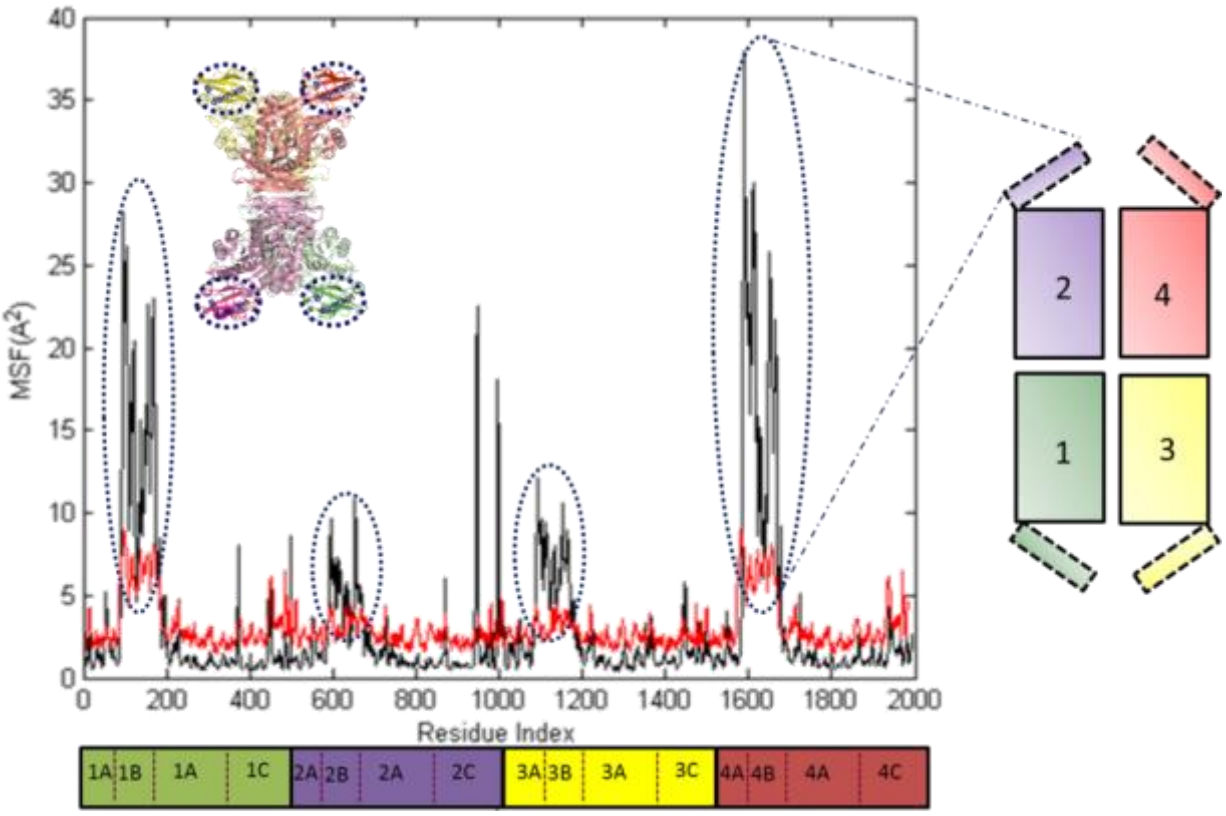


Fig. 3.13 Comparison of experimental B-factors (red) from apoPYK (3HQN) with the mean square fluctuations of the Cα atoms of the LmPYK tetramer calculated from MD simulation (black). Both the representation of the tetramer (right) and the simulation molecular cartoon (inset) highlight the mobile B-domains with dotted lines.

What effect does constraining the PYK as a symmetrical tetramer have on the molecular flexibility? To answer this question we carried out MD simulations on an isolated monomer immersed in a water bath. A comparison of mean square fluctuation values with the corresponding protomer chain locked in as part of a tetramer highlights the highly mobile B-domain movement in the tetramer simulation (Figure 3.14). The enhanced mobility of the B-domain (residues 89-186) is also present in the monomer simulation but is much-reduced with average RMSF values for the B-domains alone of 13.8 $Å^2$ and 6.4 $Å^2$ for tetramer and monomer respectively. This rather unexpected result suggests that the B-domains behave differently (and are more mobile) as part of a tetramer.

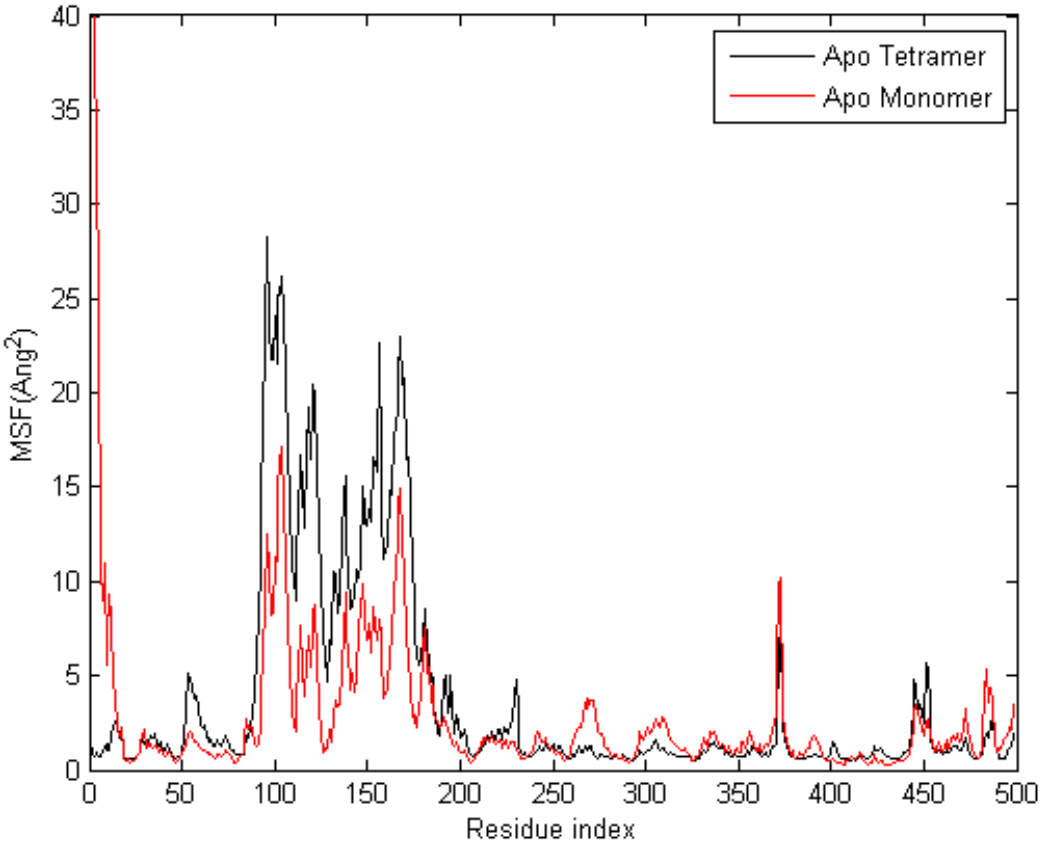


Figure 3.14: Comparison of mean square fluctuations of the Cα atoms of the apo LmPYK tetramer (black) with the apo isolated monomer (red) calculated from MD simulation showing the enhanced mobility of the B-domain (residues 89-186) is even more exaggerated in the tetramer than the isolated monomer.

A comparison of molecular motion of the tetramer with and without the FBP bound (Figure 3.15) which shows the apo structure vibrates slightly more than the FBP-bound holo tetramer with RMS values of 3.2 Å^2 and 2.8 Å^2 respectively. The simulation also shows an interesting asymmetry of movement of the B-domains which breaks the D2symmetry of the tetramer; the B-domains (residues 89-186) of protomers 1, 2, 3, and 4 have MSF values of 13.8 Å^2, 4.7 Å^2, 6.6 Å^2, 16.5 Å^2 respectively. On binding the F26BP effector the MSF values change marginally to 16.7 Å^2, 5.3 Å^2, 5,6 Å^2, 9.4 Å^2 respectively. However as observed experimentally in the X-ray structure (Figure 3.12), the asymmetry of movement of the B-domains, with protomers 1 and 4 vibrating more than protomers 2 and 3, is preserved in the apo and holo MD simulations.

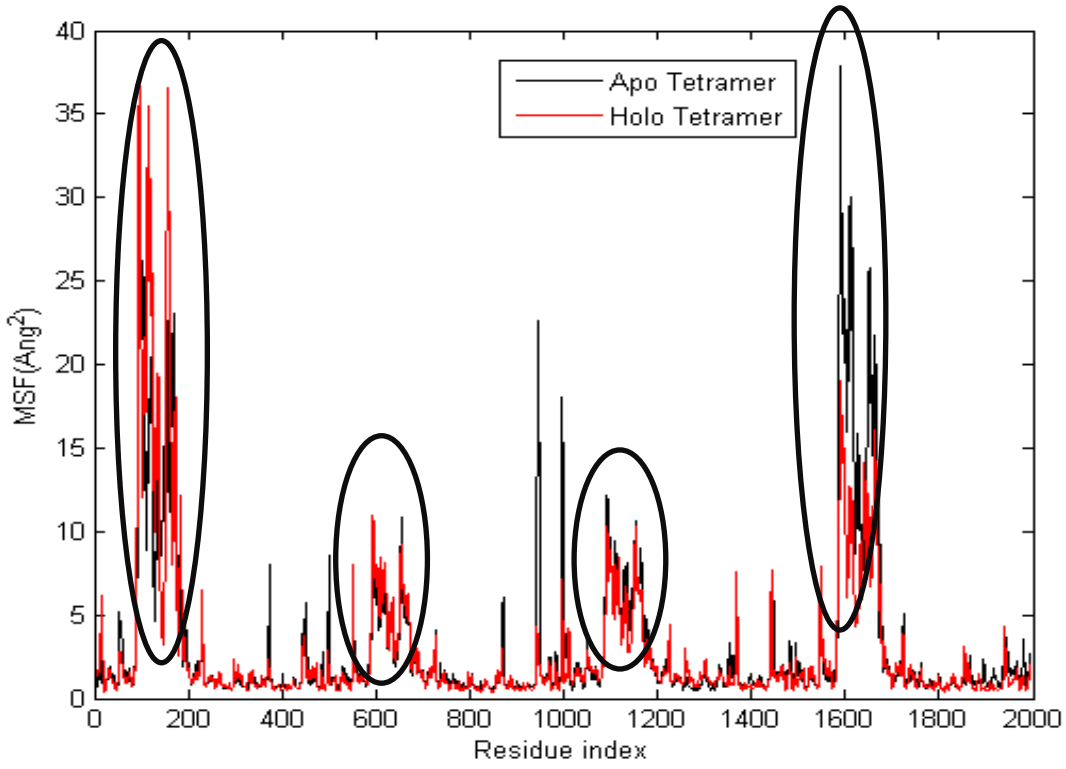


Figure 3.15: Mean square fluctuation comparison of apo LmPYK tetramer (average value 3.182 Å^2) and holo (F26BP)-bound tetramer (average value 2.813 Å^2)

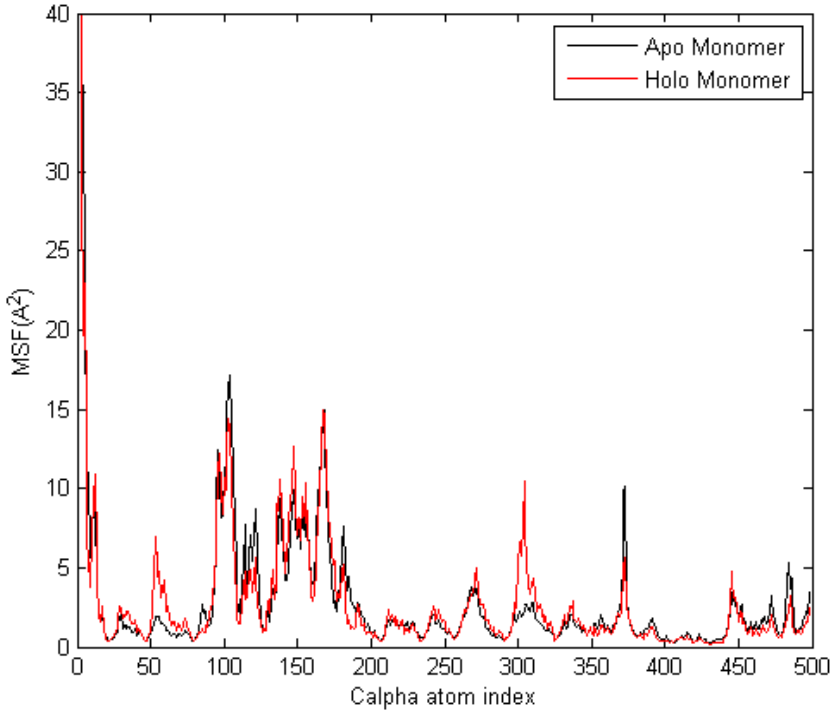


Figure 3.16: Differences in mean fluctuations for the isolated apo monomer (no FBP bound; black) and the isolated holo monomer (with FBP bound; red) .

Comparable simulations showing the effect of FBP on binding to the isolated monomer were also carried out (Figure 3.16). MD simulations (~70ns) of the isolated monomer (residues 1-498) showed that the holo (FBPbound) isolated monomer had small reduction of 0.08 $Å^2$ over the 498 Cα atoms in rms fluctuations compared with the isolated apo monomer.

Figure 3.17 is a plot between the MSF of simulation A and B. It is observed that the b-domains have much reduced fluctuations for simulation B with the average being ~1.5 $Å^2$ as opposed to 2.8 for simulation A (Holo). This reduction can be accounted for by the fact that the presence of active site ligands, i.e. ATP, oxalate and ions have resulted in further restriction of the B-domain movements, thereby stabilizing it into a fully active R-state conformation. Quantification of these thermal fluctuations have provided a solid affirmation to the fact that the allosteric ligand results in the stabilization of the mobile B-domain, thereby reducing the flexibility of the enzyme as whole and providing rigidity for enzyme mechanism. Furthermore, as ligands bind to different locations on the enzyme, they induce intermediate states of closure around the active site (Tables 3.4 and 3.5).

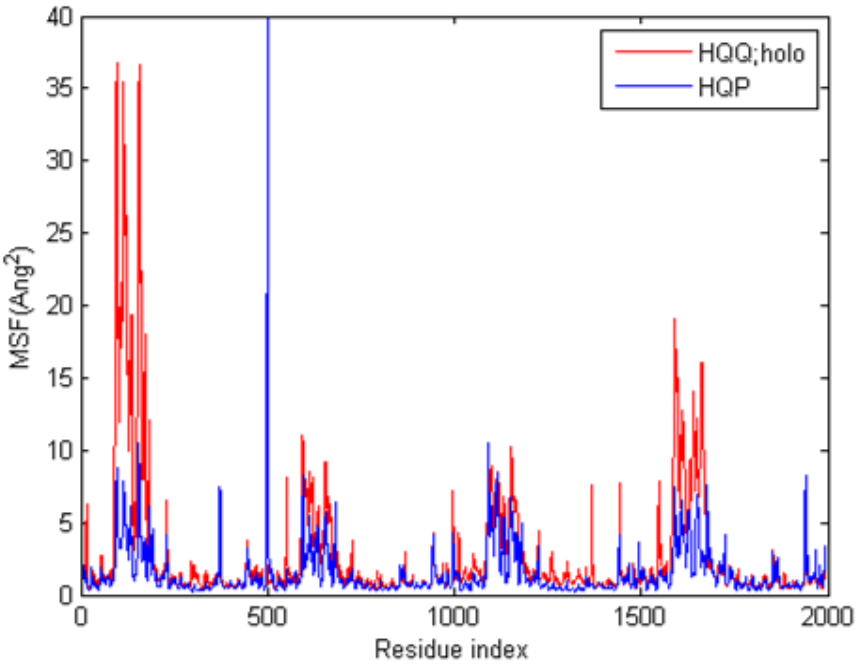


Figure 3.17: Mean square fluctuation comparison of 3HQQ holo tetramer (average value 3.182 $Å^2$) and 3HQP tetramer (average value 2.813 $Å^2$)

	Apo Tetramer				Tet	Holo Tetramer				Tet	Apo Monomer	Holo Monomer
Protomer	1	2	3	4		1	2	3	4			
Whole	3.78	2.11	2.51	4.31	3.18	4.31	1.901	2.21	2.821	2.81	2.982	2.900
Only B	13.78	4.76	6.611	16.52	10.42	16.67	5.33	5.62	9.43	9.26	6.409	6.1335
Rest	1.33	1.46	1.50	1.32	1.40	1.28	1.06	1.38	1.20	1.23	2.142	2.108

Table 3.4 : Mean Square Fluctuations for 3HQQ simulation. The data has been calculated for the whole tetramer and the individual protomers. Also, MSF has been calculated for only B-domains and the rest of the protein without B-domains

	HQP Tetramer				Tet
Protomer	1	2	3	4	
Whole	1.521	1.424	1.408	1.628	1.495
Only B	3.855	2.882	3.8175	3.644	3.550
Rest	0.949	1.067	0.81	1.134	0.987

Table 3.5: Mean Square Fluctuations for 3HQP simulation. The data has been calculated for the whole tetramer and the individual protomers. Also, MSF has been calculated for only B-domains and the rest of the protein without B-domains.

3.3.4 Contact Maps

Contact maps represent the three-dimensional protein structure in a two-dimensional form. It is generated from a binary symmetric matrix. These matrices are composed of an array of the pairwise distances (distance geometry) done for all pairs of atoms, for selected types of atoms (eg, Cα atoms), for groups of atoms (eg, side-chain centres of mass), or for entire amino-acid residues. By taking a certain cut-off value, usually in the range of 6-16Å for the pairwise distances, contact maps are then generated. Usually Cα atoms are used for the

construction of these maps[276]. Residue pairs that are located within a predetermined cut-off in a 3-D structure of the protein will have a value of 1 in the matrix and those further than the cut-off will have a value of 0. The cut-off distance can take a value ranging from 6-16Å and is usually measured between Cα of residue pairs The Cα atoms that are closer to each other in the protein structure than the chosen cutoff distance are considered to be "in contact", thereby producing a binary matrix translated into the contact map [277]. These contact maps reflect well the overall topology of the protein fold.

In order to understand the structural and conformational differences between the two simulated states of pyruvate kinase, contact maps were calculated of the average structures and compared to highlight the regions that have changed upon binding of the allosteric activator. There were a total of 2328 contacts in the Apo tetramer whereas the Holo tetramer had 2333 contacts. As can be seen from the figure, there are several new contacts established between the Cα atoms. Most of the new contacts are within the protomer, suggesting that the effector molecule upon binding results in additional contacts formed within the protomer. These additional contacts would then establish the rocking motion as observed in the crystal structure. Also, the Holo tetramer and HQP structure share some common contacts which are otherwise new in comparison to the Apo tetramer. Several of these contacts are intra-chain with few of them being between the sub-units. The presence of unique intra-chain contacts is consistent with the fact that the binding of ligand stabilises the protein by forming some additional contacts. Also, the new contacts would be required for the binding of ligand which then introduces some conformational changes in the individual chains. However, in case of comparison between HT and HQP, many of the unique contacts are across the A-A interface which is consistent with the observation that this structure has additional active site ligands along with the effector molecule (Figure 3.18-20). For example, we notice a contact established between S314 of chain A and V267 of chain C i.e. adjacent chain. This is to incorporate the ATP and the substrate into the protein. Similarly, there is a contact between Glutamine 297 of chain A and Methionine 302 of chain C.

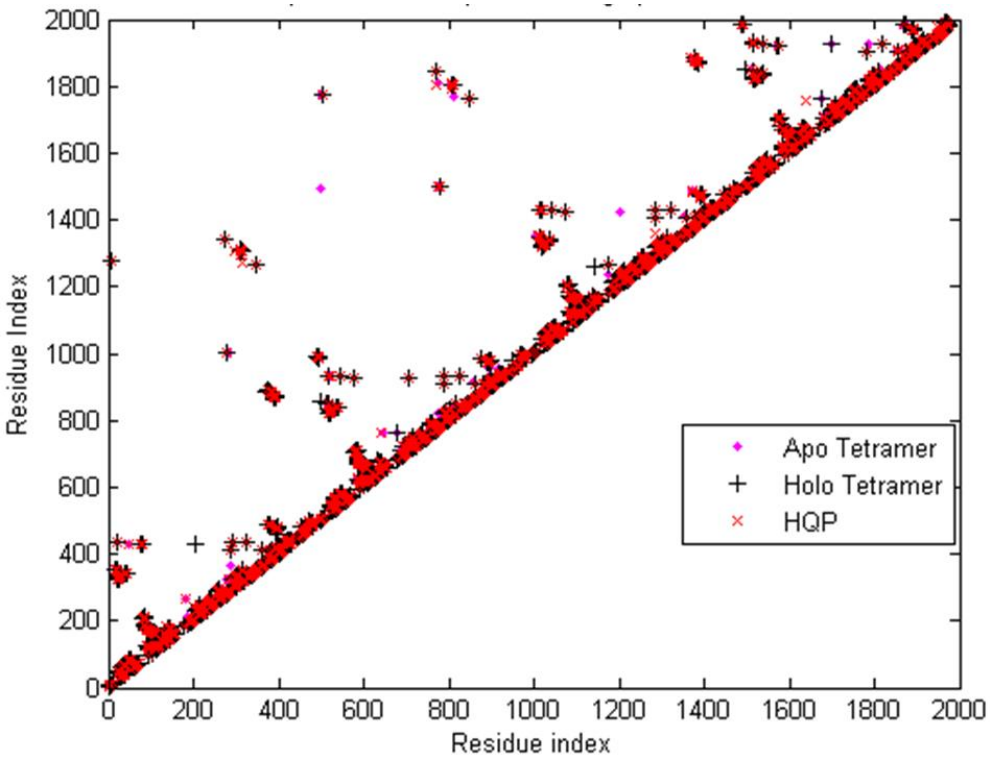


Figure 3.18: Contact map for all the three simulations of the Cα atoms. The magenta dots represent apo tetramer, the black '+' represent the holo tetramer and HQP is in red crosses .

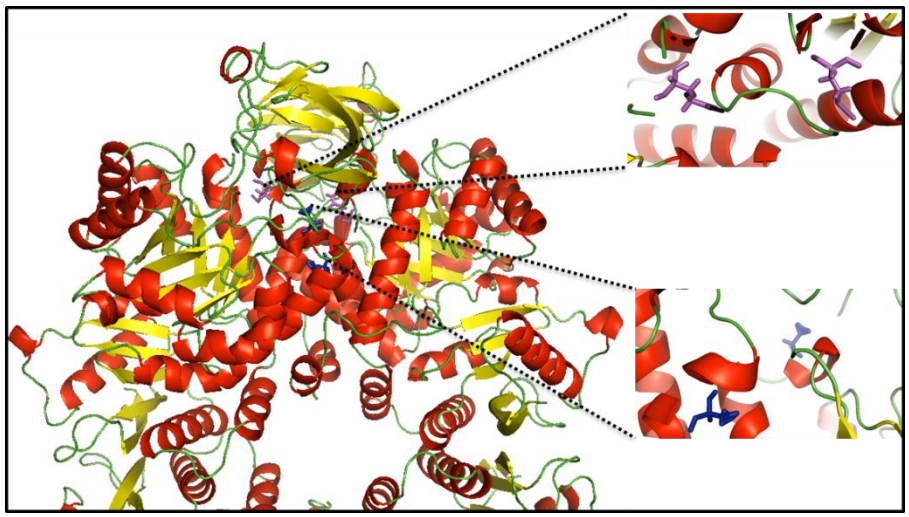


Figure 3.19 : Average protein structure of 3HQP from simulation showing the two additional contacts across the A-A interface (Left) between Glutamine and methionine (shown in purple sticks; top right and between serine and valine (shown in blue sticks; bottom right). The protein is coloured according to secondary structure elements.

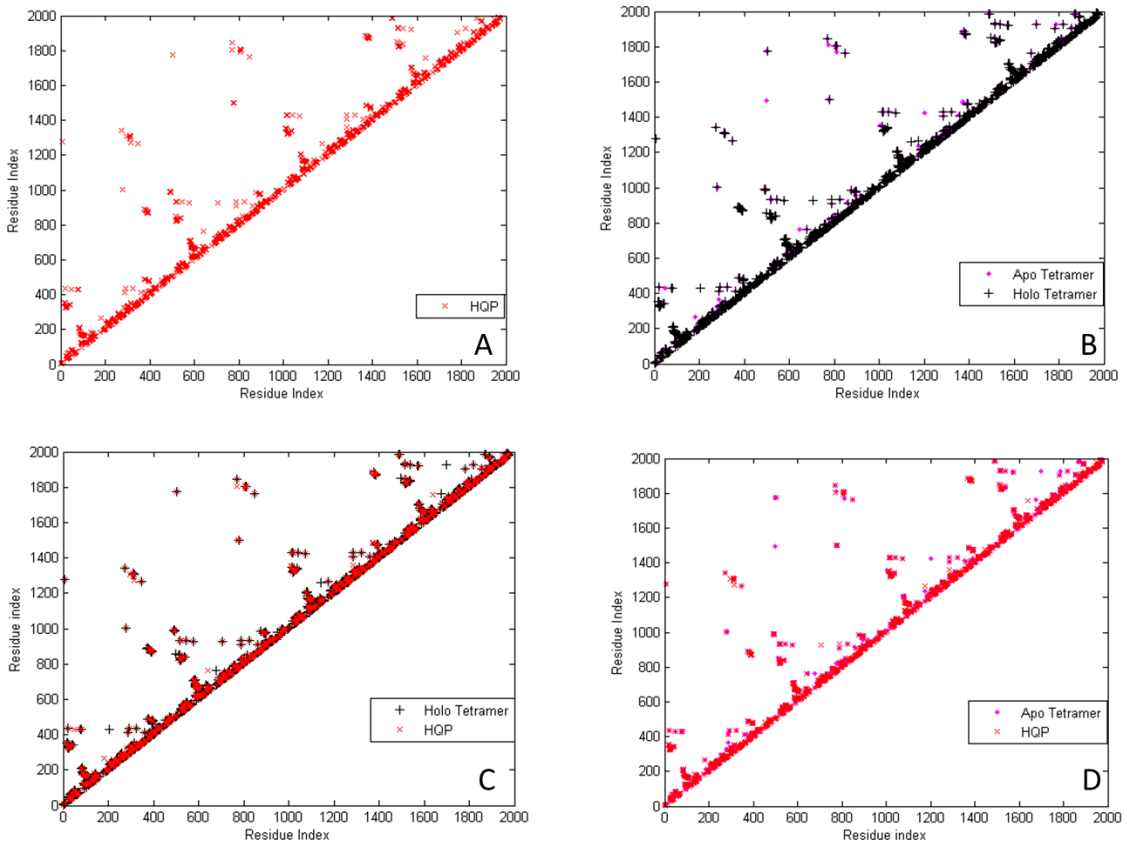


Figure 3.20 : Contact map of the average structure of the protein from simulation. (A) Contact map for HQP, (B) Between Apo Tetramer (magenta) and Holo Tetramer (black), (C) Between Holo Tetramer (black) and HQP (red) and (D) Between Apo Tetramer (magenta) and HQP (red).

3.3.5 Principal Component Analysis

In order to determine the concerted conformational motions of protein, the statistical technique called PCA was utilised. This helped in the identification of collective motions that are relevant to the protein structure i.e. PYK and also reflected the concerted transitions and the important changes brought about by insertion of active site and effector site ligands. PCA was carried out using GROMACS automation software. This was carried out on the data obtained after MD simulation of the crystal structures. One of the important measurements from the covariance matrix is the 'trace of the matrix' and it is actually the sum of all the eigenvalues. This sum can be used to describe the total motility of the system. The values of the trace of covariance matrix for the three systems AT, HT, HQP are 63.39, 56.043 and 29.7951 nm^2, respectively,

demonstrating that binding of active site ligands and effector molecule clearly influences the motional strength of the protein which is also corroborated by the observation of reduced RMSF for holo and HQP as opposed to AT. The motional strength of the protein is more stable and constricted in HQP as seen in RMSF graphs earlier in the section and now reflected by the trace of the matrix.

There are some rules for excluding principal components [described in detail in the materials and methods section]. One of them says to include just enough components to explain 90% of total motility. A second called Kaiser's criterion excludes those PC whose eigenvalues are less than average, i.e. less than one if a correlation matrix has been used. In practice often compromise is used, thus Figure 3.21 presents the percentage and cumulative percentage of variance explained by the first 200 from 5976 eigenvalues. The plot has been obtained by diagonalization of the covariance matrix of atomic fluctuations and is plotted in decreasing order with respect to the corresponding eigenvector indices for all the three simulations. It is shown in the figure that first 10 from 5976 eigenvalues can describe approximately 70-80% of total variation of the system.

The Root Mean Square Inner Product (RMSIP) is used to calculate the overlap between the subspaces. As the RMSIP reflects the similarity of motional directions while the trace of covariance matrix describes the strength of the motions, and so RMSIP calculation was done for the first 200 eigenvectors out of the total 5976 eigenvectors. This provides a quantitative measure for the conformational overlap between different simulations and would also help in comparing the similarity of protein motions for the three systems. The RMSIP measures the degree of overlap or similarity of eigenvector sets (see chapter2). A RMSIP of 1 indicates the sets are identical, while a value of 0 indicates that the eigenvectors are orthogonal[200]. As expected, the highest degree of overlap occurs within the individual trajectories for each protein. The table enlists the calculated values. The RMSIP values between HQP and HT, HQP-AT and HT-AT are 0.262, 0.274 and 0.444, respectively. This suggests that the similarities of the motions for the same protein i.e. PYK under different conditions are low. This further suggests that allosteric effector alone affects the internal motions of the

protein largely and further addition of active site ligands i.e. ATP, oxalate then does affect the major internal motion of the protein (as seen in RMSIP between AT-HT which is 0.4 and then between HQP-HT is 0.2; Table 3.6).

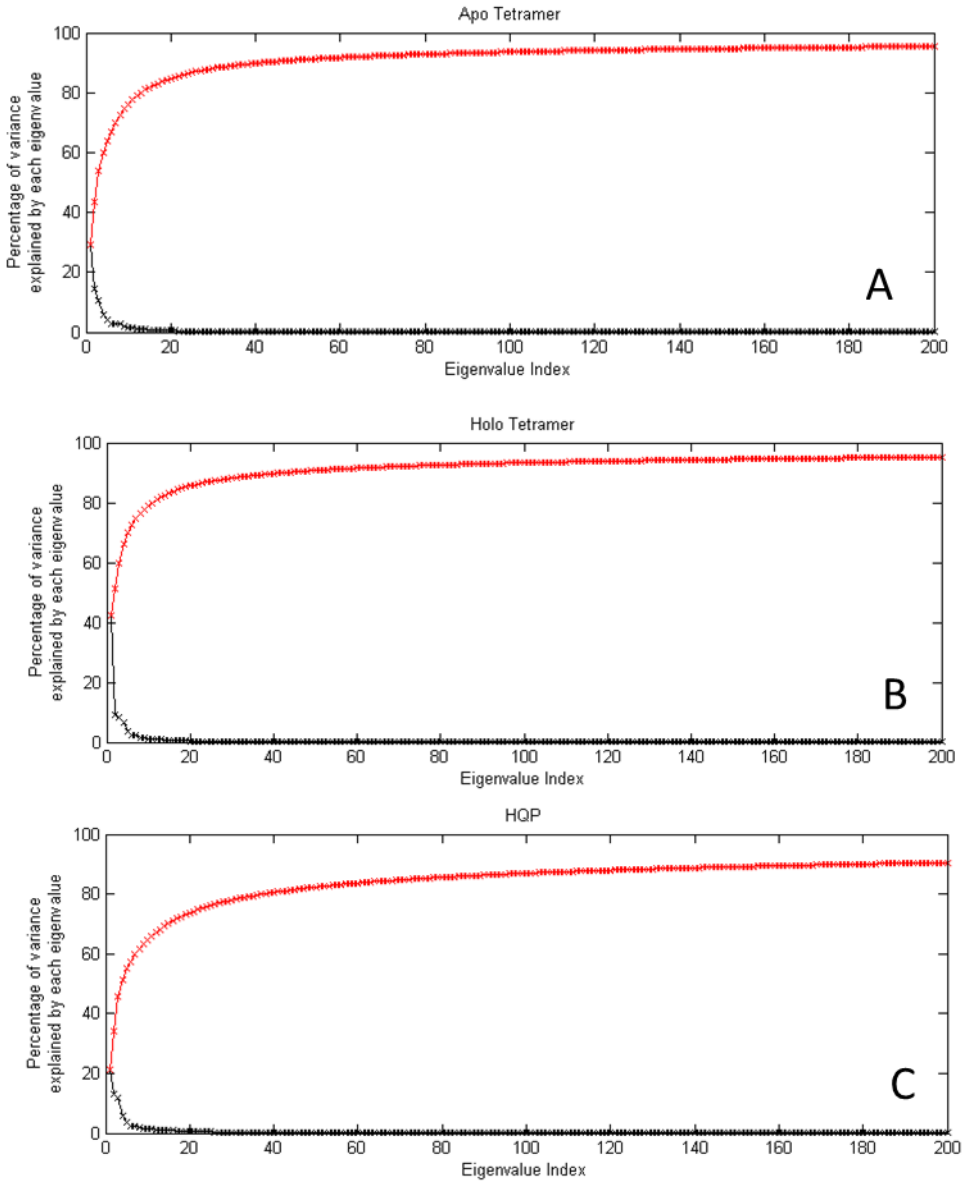


Figure 3.21: Percentage (black) and cumulative percentage (red) of variance for first 200 eigenvalues for Apo tetramer (A), Holo Tetramer (B) and HQP (C) simulation.

	HQP-HT	HQP-AT	HT-AT
Trace	29.79;56.04	29.79;63.39	56.04;63.39
Square root of trace	5.45;7.48	5.45;7.96	7.48;7.96
RMSIP	0.247	0.252	0.443

Table 3.6: RMSIP values for the three simulations.

From the eigenvalue figure (Figure 3.21), it can be seen that only the first few values correspond to the concerted motions which quickly decrease in the amplitude to reach more constrained and localized fluctuations. Also, the comparison of the plots shows that the properties of the motions as described by the first few eigenvectors i.e. principal components are not the same. Such is the case between Apo and Holo tetramer, where the magnitude of eigenvalues are higher for the Holo tetramer. The first principal component/eigenvector in AT represents 29.2 % of the total variance, HT represents 42.1836 % and HQP represents only 21.3044 % of the total variance i.e. total motility in the simulation set. Since the first 10 eigenvectors contribute to a total variance of 76.15%, 79.02% and 64.62% for AT, HT and HQP, respectively, these have been used further to analyse the protein motions. As is quite evident from the cumulative percentage table (Table 3.7), the data is quite dispersed amongst the eigenvectors i.e. the first 3 eigenvectors together show around 50% motions as opposed to the general trend of the first mode being sufficient to show more than half of the variance. This suggests that the conformational transitions in pyruvate kinase are dispersed amongst a set of modes with distinct representation in terms of motions of the protein.

Mode	Apo Tetramer		Holo Tetramer		HQP	
	Eigenvalue (nm^2)	Cumulative Percentage	Eigenvalue (nm^2)	Cumulative Percentage	Eigenvalue (nm^2)	Cumulative Percentage
1	18.5265	29.22622	23.6411	42.18364	6.34658	21.3044
2	9.08084	43.55157	5.18212	51.43033	3.86209	34.26878
3	6.62805	54.00756	4.64238	59.71389	3.45698	45.87328
4	3.75484	59.93095	3.71857	66.3491	1.6687	51.47482
5	2.62194	64.06716	2.093	70.08371	1.08529	55.11796
6	1.90599	67.07392	1.36087	72.51197	0.707453	57.49276
7	1.79126	69.8997	1.26116	74.7623	0.648482	59.6696
8	1.69225	72.56929	0.925659	76.414	0.568455	61.57781
9	1.26244	74.56083	0.798464	77.83872	0.484956	63.20573
10	1.00923	76.15293	0.664495	79.02441	0.422629	64.62442

Table 3.7: Eigenvalues and cumulative percentage for the first 10 principal components (eigenvectors/modes) of Apo, Holo and 3HQP simulation.

The reduction in dimensions as obtained by principal component analysis can be used graphically. Therefore, as the first two components explain most of motility, then a plot showing the distribution of the objects on these two dimensions reflects the overall distribution of the data.

Thus, PCA is helpful in extracting the motion along the eigenvectors. This is obtained by the projection of all trajectory frames across the simulation time on any particular eigenvector. This in turn leads to the generation of a new trajectory which contains motion in the direction of that particular eigenvector. This provides an estimate about the width of the essential space explored by the system as a function of time [211, 278, 279].

Next, the progression of the reaction coordinates also known as projection over time was monitored. Figure 3.22 is a plot of the trajectories which have been projected on the first ten eigenvectors. These plots reflect the degree of anharmonicity of the motions and also show that the probability distributions for the first few principal components are away from Gaussian for all the three systems. We see typically large amplitudes and slow motion of essential coordinates[280] which have been characterized by a slow diffusive kinetics[200]. The systems thus show characteristic multiple-minima protein energy landscape features [211, 281, 282]. As the relaxation and convergence for the essential coordinates are typically beyond the nanosecond timescales, these slow diffusions provide perhaps only partial sampling of the subspace defined by the first 2-3 eigenvectors. However, the convergence of the essential subspace (usually the first 10-20 eigenvectors) is reasonably achieved within a few nanoseconds[156]. Although, the single essential eigenvectors are likely to be not the equilibrium ones, i.e., the eigenvectors obtained by a completely converged statistics, the corresponding atomic collective motions are probably significant as they belong to a good approximation of the equilibrium essential subspace. From these figures we observe that that the system moves to a new equilibrium, which seems to have been reached at about 60 ns. Considering the size of the system and the fact that the relaxation times adhere to much longer timescales, the shifting to a new equilibrium confirms the sampling of a larger conformational space. The next step followed projecting histograms of the probability distributions for the first four eigenvectors (Figure 3.23). These histograms are plots that show the distribution of data. The overall range of a given set of data points is divided to smaller subranges (bins), and the histogram shows how many data points are in each bin. The height of the bar corresponds to the number of data points in the bin. These probability plots give an idea about the degree of anharmonicities of the motions and it is observed that the first 3 eigenvectors are not Gaussian. However, as we move down the spectrum they become more Gaussian or harmonic as we could also say.

The fluctuations are comparatively less for HQP while we see high fluctuations for the Apo tetramer. Also, the fluctuations in HQP are marked by subtle transitions. From these plots, while focussing on the first few eigenvectors, one can see that they describe slow motions arising from slow diffusion kinetics with a large amplitude, thus indicating the sampling of subspace over the simulation time[283]. Convergence is almost achieved for the HQP simulation, which clearly shows the presence of ligands affecting the principal components. It has already been shown that a reliable and statistically significant description of essential subspace on the nanosecond timescale can be shown during the simulation[200]. Considering the fact that the present analysis is focussed on well-equilibrated trajectories, we are reasonable enough in assuming the motions described by our sampling belonging to an essential subspace.

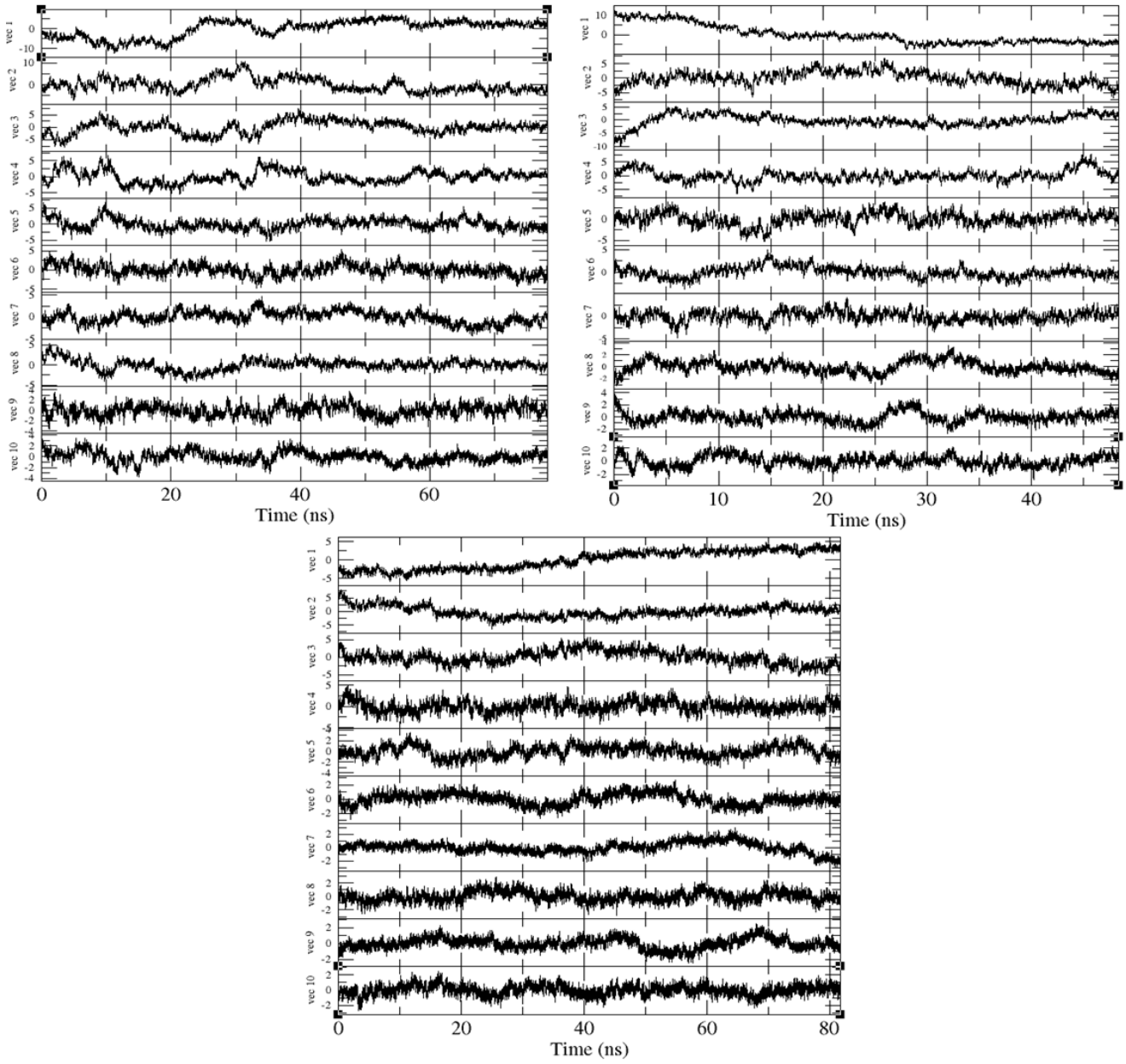


Figure 3.22: Projections on the first 10 eigenvectors for the trajectories along the length of the simulations. X axis respresents the time period in nanosecond and Y axis depicts the projections in nm for each eigenvector. Starting from top left, Projections for Apo tetramer, Projections for Holo tetramer and Projections for HQP tetramer simulation

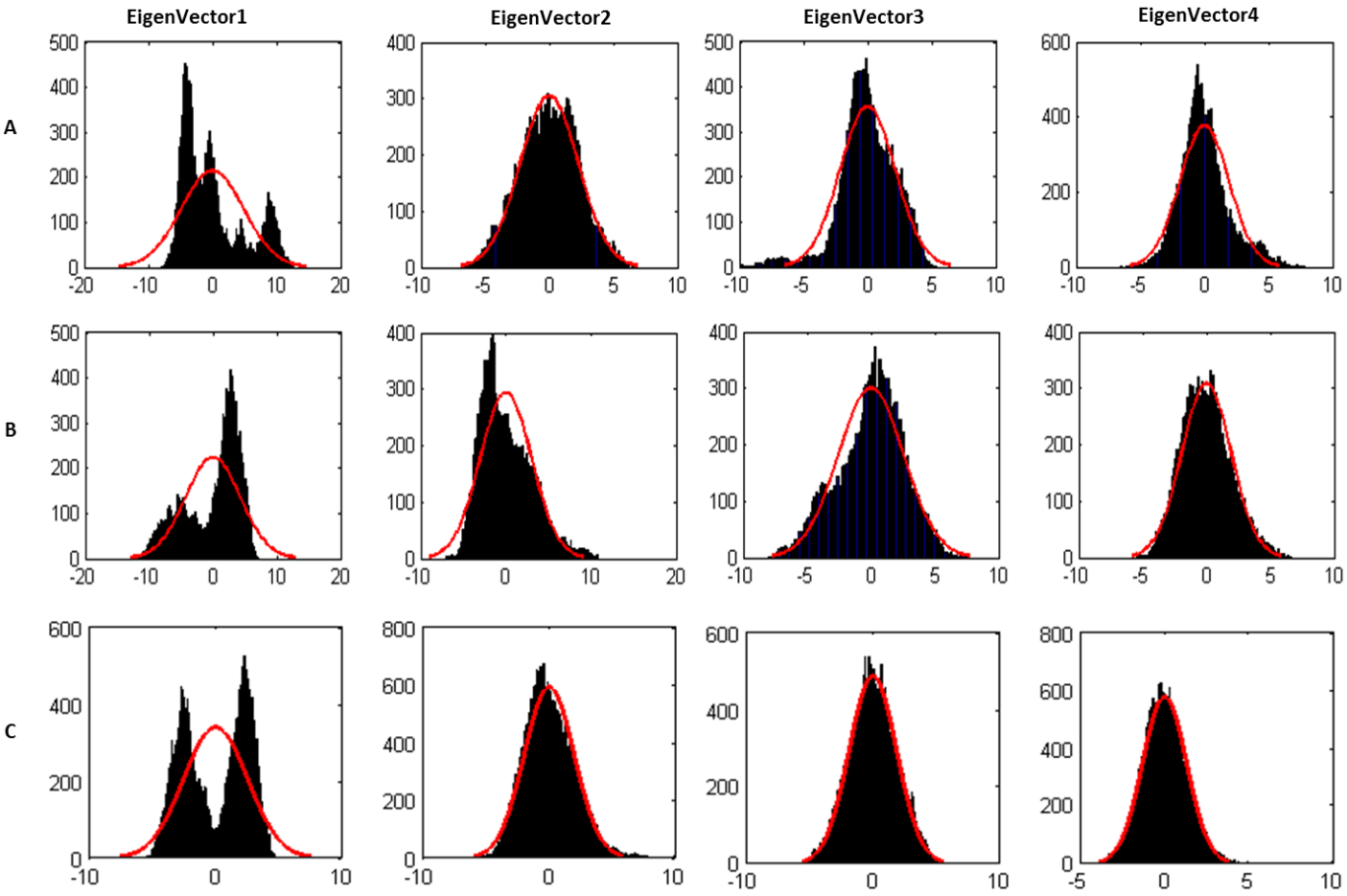


Figure 3.23: Cumulative Histograms for the first four eigenvectors; A: Apo tetramer, B: Holo tetramer, C: HQP tetramer. The red lines depict the data fitted into the Gaussian curve. The data is fitted into the Gaussian curve. The X axis is the range/magnitude of the eigenvectors and Y axis is the frequency.

For the first eigenvector in case of AT, we observe that there is a presence of multiple conformational states which are not distinctly marked. Without the presence of the stabilizing ligand, the molecule could not undergo the conformational transition. However, for the HT and HQP structure, we observe a bimodal distribution i.e. two clear conformational states for the first eigenvector. In the HT, the first state has lower density but since the HQP structure is R state structure, we see the presence of two distinct conformational states with a clear transition. This is indicative of the jumps between two isomeric states which is in agreement with the experimental observation. In HQP, distinct energy wells separated by energy barriers are visited. These type of motions are non-Gaussian.

However, the vectors tend to adopt a more Gaussian behaviour as we move down the spectrum from eigenvectors 1 and 10. But still, we can see that the motion along eigenvectors 2-4 obeys a unimodal albeit a non-Gaussian distribution. Upon further inspection of these eigenvectors, we see that LmPYK explores more than one conformation within the three structures which suggests that instead of a specific structure, LmPYK tetramer state behaves as an ensemble of tertiary conformations undergoing structural transitions.

In case of the motion described by the first principal component, we observe that one conformation predominates in LmPYK between 10-30ns, after which it evolves to a different conformation. However, these conformations are distinctly visible in the Holo tetramer and HQP structure owing to the presence of effector molecule. The presence of ligands also affects the distribution of conformations. In case of Apo Tetramer, the distributions are not clearly marked but the presence of allosteric ligands, i.e. F-2,6BP results in two distinct conformations. In the holo tetramer, however, inequality in the distribution of the conformations is observed but for the fully ligated HQP structure in the presence of substrate and allosteric effector, there is a presence of two conformations with equal distribution.

In order to probe the presence of conformational states and transitions between them, we projected the various eigenvector pairs (Figures 3.24 & 3.25). These plots provide a measure of the mobility of the protein in the essential subspace, thereby showing the clusters representative of explored tertiary conformation that differs amongst the three simulated structures.

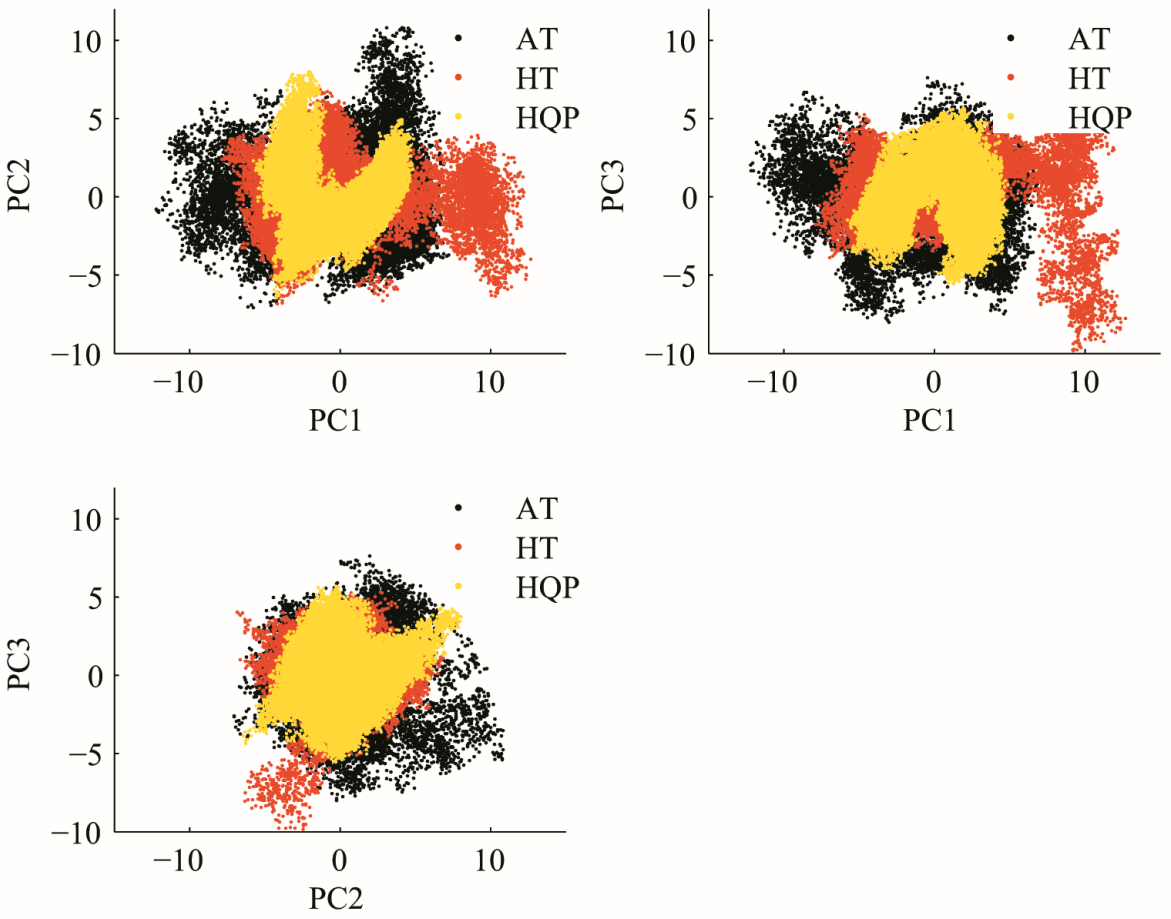


Figure 3.24: Dimensional projection of eigenvectors for all the three simulations. A: Projection on first and second eigenvector. B: Projection on first and third eigenvector. C: Projection on second and third eigenvector. The data is obtained after running PCA on the trajectories. [Starting from top left, the data is plotted on the first two eigenvectors of the covariance matrix (Principal Components 1 and 2) with each point corresponding to the one trajectory frame].

The mobility plots describe the representative clusters explored in the protein conformation and shows whether there are further non-linear correlations between collective motions.

For the apo and holo tetramer, it is noted that the variation along eigenvector 1 axis is greater than along the eigenvector 2 axis unlike the HQP

structure where we see the data is more correlated between the two eigenvectors. The second plot shows the variance along eigenvectors 1 and 3, where again we see the variation along eigenvector 1 being larger compared to eigenvector 3. This means that eigenvector 3 reveals less about the behaviour of the protein in comparison to the first eigenvector. Lastly, the third plot presents projection on eigenvector 2 and showing that their contribution to total motility is much smaller than the one presented on first plot, but correlation between data is much stronger.

Also, can be seen is the possible occurrence of three conformational states for HQP as opposed to AT. These conformational jumps essentially consist of the motion of the cap and also transition of the structure into the T-states as caused due to the binding fo the effector molecule F-2,6BP anwhich locks the tetramer into a stable conformation. These transitions are not localized tobut rather dispersed into distinct basins thus supporting the idea of multilple hierarchy[284].

For each of the eigenvectors, it is possible to extract the displacements of the residues. For a given model, a rich picture emerges of the regions of specific displacement for each of the first few principal eigenvectors. This in turn highlights the specific displacement regions. We observe a pattern amongst the subunits. It can be seen that some of the motions described by one eigenvector are also described by the second and third eigenvectors, (concerted motions) thus so thereafter in similar pattern. The motions as described by the eigenvectors are repeated in a similar fashion thereby reinforcing the fact of concerted motions (Figure 3.26). For instance, the first and second eigenvector appear to capture the same collective motion, but in two different subunits. Similar relationships appear to exist between other sets of eigenvectors. The eigenvectors for the individual trajectories have been plotted as a function of time which shows how the simulation progresses towards the new state. For the apo tetramer we see a very strong one single cluster while HQP shows the presence of two conformational basins which again correlates with the projections in Fig 3.23 of a more Gaussian motion of the protein and it being a concerted ensemble. From these it has also been observed that HQP shows the smallest conformational space followed by

HT as compared to AT which shows larger conformational space. This suggests that HQP is most stable during simulation due to the presence of ligands and the structural changes inflicted upon binding of the effector molecule.

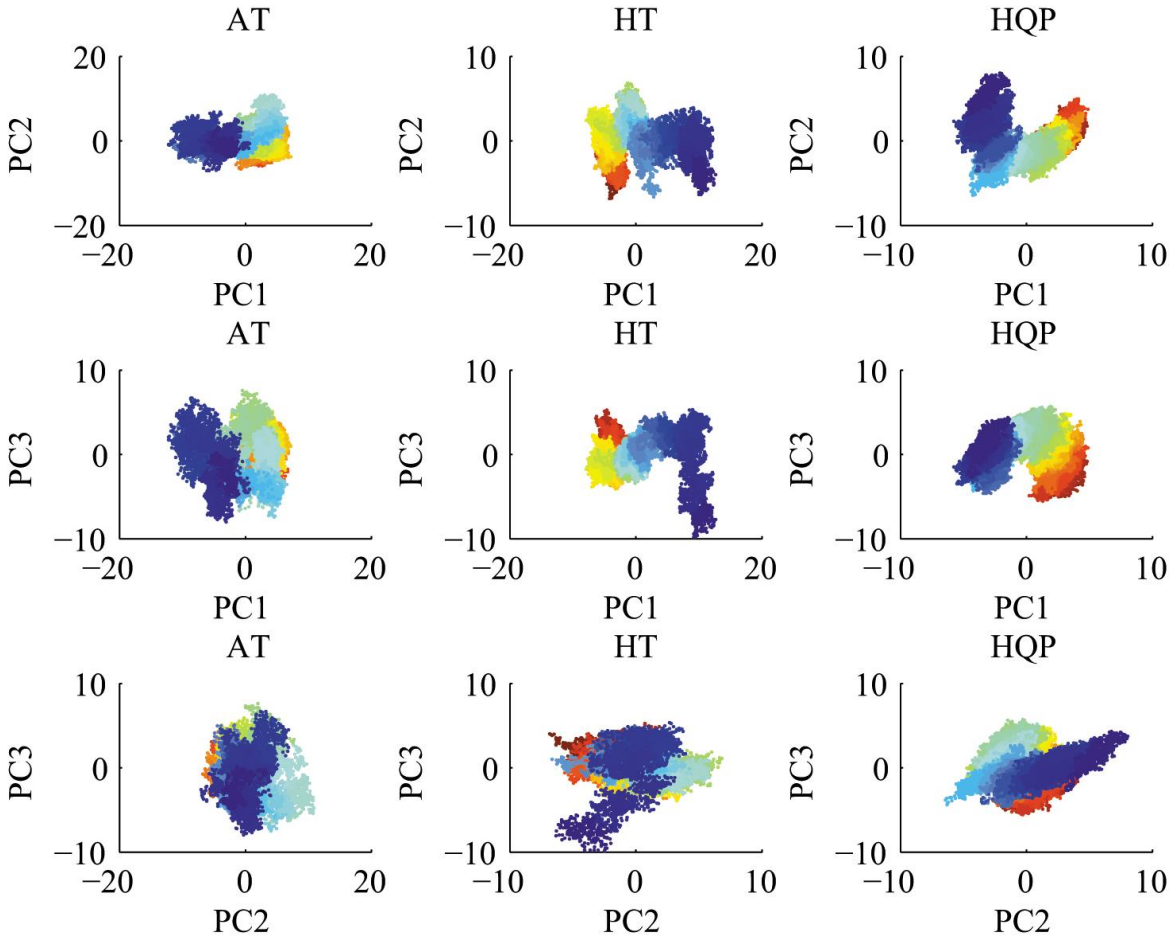


Figure 3.25: The 2-Dimensional projection of eigenvectors for all the three simulations. The data is plotted for the first three eigenvectors with the first column being for the apo tetramer, second column for the holo tetramer and the third column for HQP. The projections are colored for the time period of the simulation ranging from blue to red, with blue being the starting point of the simulation and red denoting the end of simulation and thus the frames towards the end of the simulation.

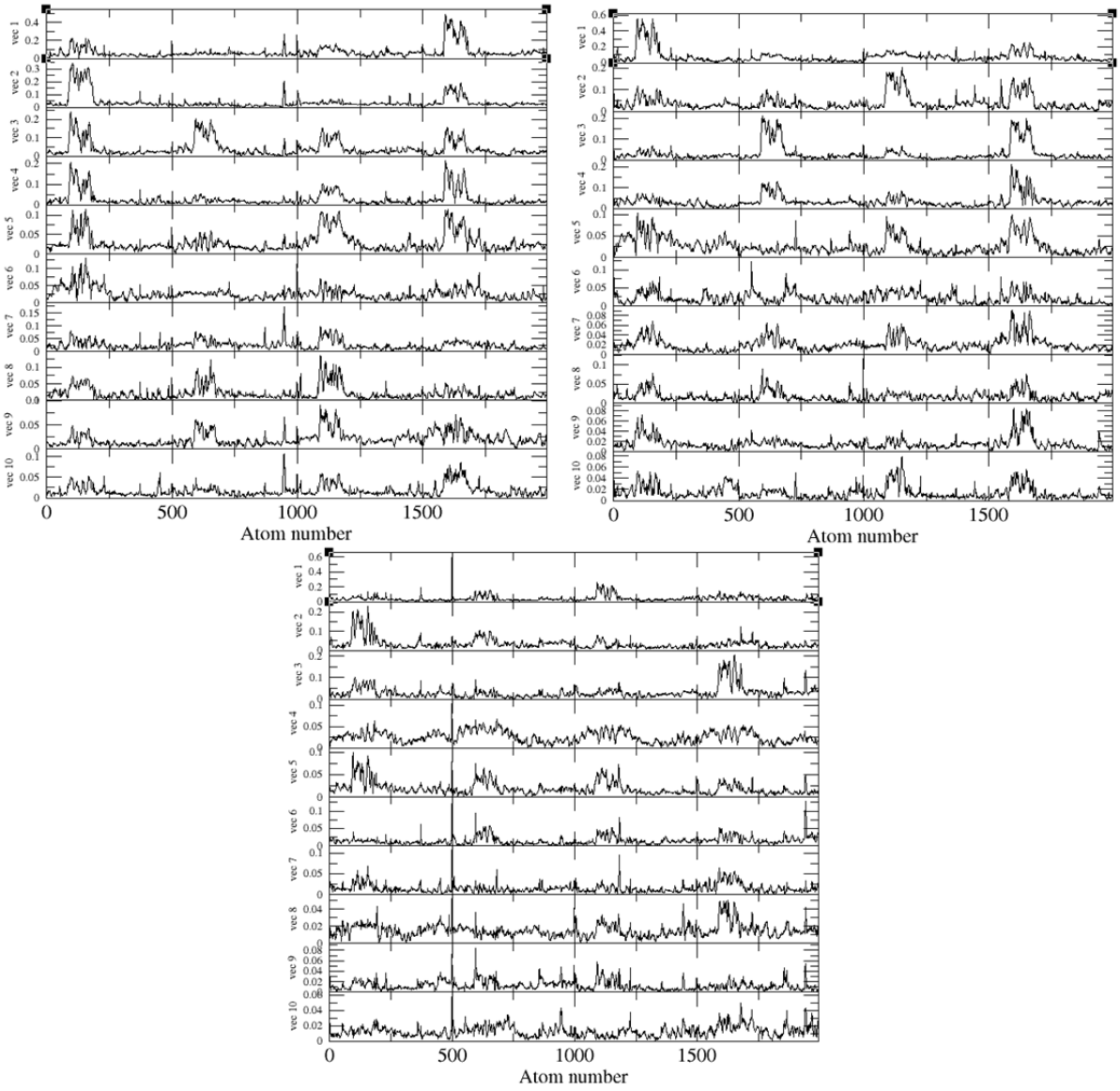


Figure 3.26: Residue displacements in the subspaces spanned by the first 10 eigenvectors for Apo tetramer, Holo Tetramer and HQP tetramer (Starting from Top left).

These results are in agreement with the fact of our allostery model which is concerted and verifies the motional direction and strength of the protein are restricted upon ligand binding. As the allosteric ligand F2, 6BP locks the tetramer into rigid conformation, we observe the decrease in conformational space. The projection of the Cα trajectories on the plane defined by the first and second

eigenvectors indicates that the Apo tetramer has a wider conformation basin as compared to the holo tetramer and the HQP fully ligated structure. Also, the displacement along the second eigenvector is more confined for HQP trajectory as compared to the other two which is also observable in the projection on the plane formed by the second and third eigenvectors. In fact there are two basins of comparable density for the HQP trajectory which are not quite distinct for the apo and holo tetramer trajectories.

We also inspected the projection along the first three eigenvectors for the three simulations to better understand the conformational sampling of the three systems. As is evident from the first glance, the major contributions towards the eigenvectors for all three systems is in the B-domains (Figures 3.27 & 3.28) with the highest contribution from the two adjacent subunits i.e. subunit A and D. The B-domain dominates the correlated motions in the first three eigenvectors for all the systems. For the first eigenvector, the contribution by the residues of the B-domain for subunits A and D is more prominent in the apo and holo tetramer while it is more uniformly distributed for the HQP trajectory. However, the magnitude of fluctuations is quite large for the apo tetramer. These displacements are indicative of the wide conformational space explored by the residues of the proteins which is also seen in the case of projections along PC2 and PC3 where the amplitudes are large. Taken together, these results indicate that the presence of allosteric ligand F2,6BP gives rise to an overall reduction of the conformational space of the enzyme which is mainly due to the reduction of the flexibility in the constant flipping of the B-domain. There is also a reduced mobility observed for the residues of the C-domain as is observed in eigenvector 2 and 3 for the HQP simulations in comparison to the apo trajectory.

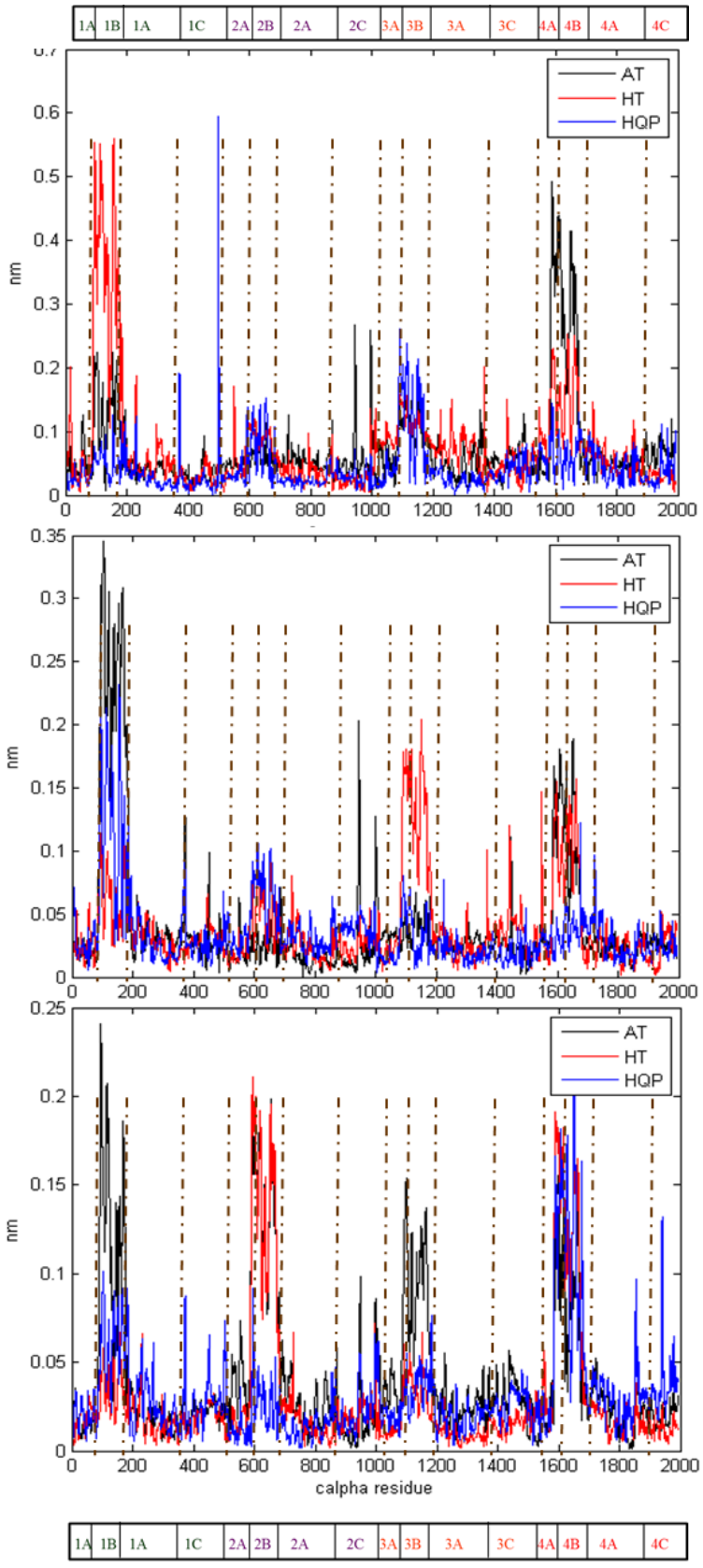


Figure 3.27: Cα root mean square fluctuations projected along the first three eigenvectors for the three trajectories. Top: Projection on first eigenvector, Middle: Projection on second eigenvector, Bottom: Projection on third eigenvector.

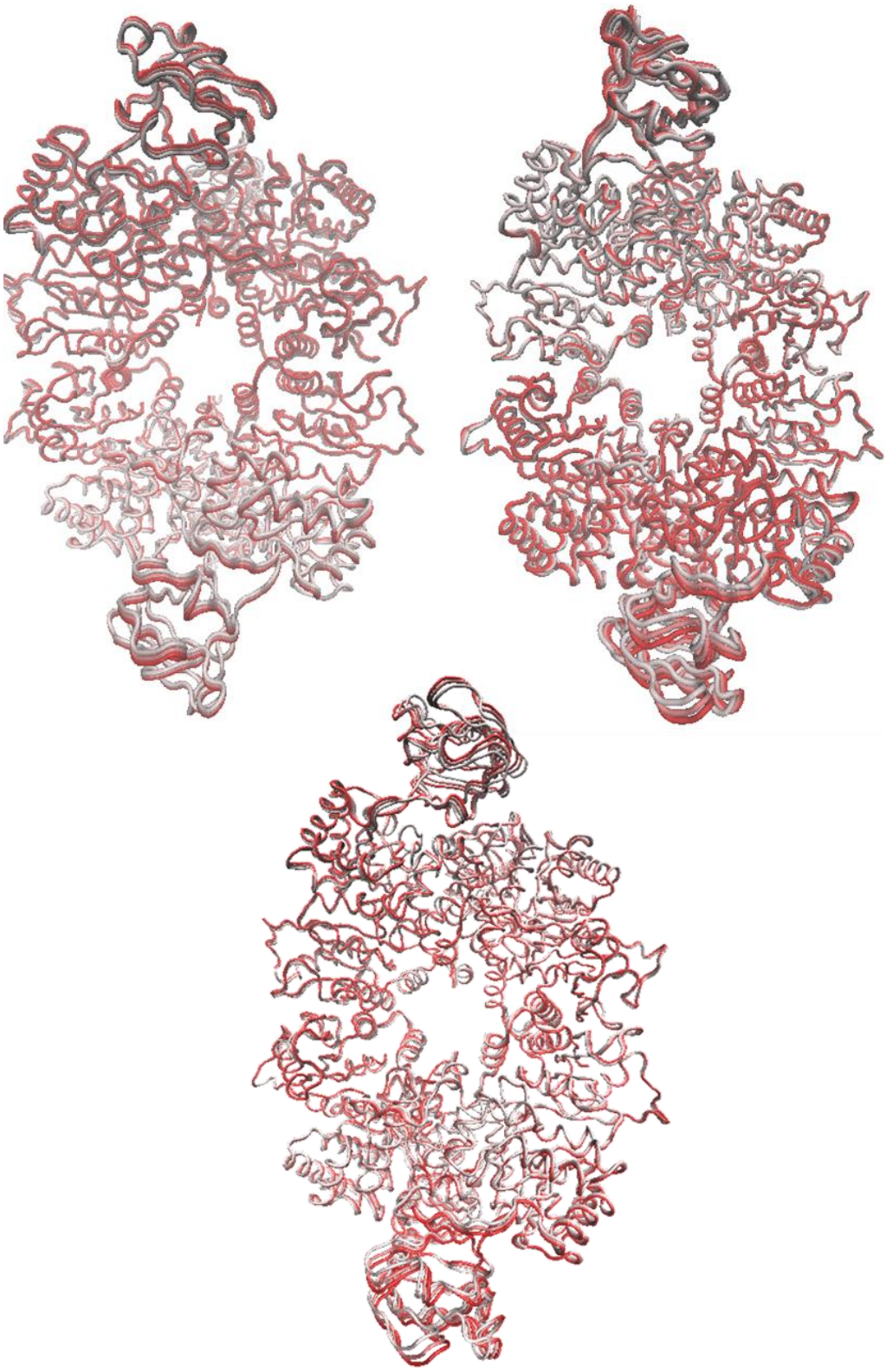


Figure: 3.28 (Top Left) Motion described by the first eigenvector (PC 1) for the Apo tetramer, Holo Tetramer and HQP (Top Left, Right and Centre). The data is projected along the whole simulation time i.e. for 78 ns, 48 ns and 80 ns respectively. The color red indicates the initial structure and the structures in white represent the final structures after simulation.

From the displacement graph and the projection along the first eigenvector, we see that the effector loop (residues 481-487; 979-985; 1476-1483; 1975-1981for first, second, third and fourth subunit respectively) has a high fluctuation in Apo tetramer while it is considerably reduced in Holo tetramer and HQP simulation. This wide displacement is also reflected in the pair projections graphs. The AC core of the protein comprising residues 18-86 and 188-480; 516-584 and 686-978; 1014-1084 and 1184-1476; 1512-1582 and 1680-1974 for the first, second, third and fourth subunit respectively, shows greater stabilization in HT and HQP simulation as opposed to the AT tetramer where the level of displacement is high.

Overall for all the three structures, the structural rearrangements occur mainly in the B-domains and C domains. The simulations show that the presence of allosteric activator F-2,6BP reduces the fluctuations in these regions which is in agreement with the stabilizing effect of the ligand on the protein. Also, the simulations reveal that the amplitude of concerted motions is affected by the ligand.

3.3.6 DISTANCE FLUCTUATION ANALYSIS

Molecular dynamics trajectories of the apo and the holo structures give information on how much time the system spends in a given conformational state[285]. Structural and dynamic differences in the apo and holo structures are analysed here using a novel plot of pair-wise distance distributions. In order to track the interaction between a pair of residues over the course of an MD simulation the inter-atom distance is calculated at each time point and the frequency distribution of the distance is plotted over the time-course of the simulation. In many cases, as expected, the inter-residue distances show a normal distribution though frequently bimodal distributions or skewed distributions are observed. This graphical summary provides a convenient way of comparing the behaviour of holo and apo forms of LmPYK (Figures 3.29, 3.30).

Activator binding changes residue fluctuations around the effector binding region: The ligand F26BP sits between two turn regions and interacts with Glu451 and Gly487. The crystal structure distance between these two residues with F26BP bound is 12.0 Å. The MD simulation with F26BP removed (apo) shows a relaxation of the loop as shown by the increased inter-Cα distance between these two residues compared with the holo-structure (Figure 3.29). Bound F26BP in the holo structure keeps the two residues close to the experimental X-ray distance with a mean distance of 12.5 Å while in the more flexible apo structure there is a broader, almost bimodal distribution with preferred distances of 14.2Å and 17.5 Å (Figure 3.29).

B-domain motion closes the active site. The largest movements in the enzyme activity of PYK is the opening and closing of the B-domain. This has been monitored by the inter-residue distance changes between residues of A (His 57; red) and B (Gln 93; green) domains for each protomer in the tetramer (Figure 3.29; top right). The mean distance in the apo tetramer is 30Å consistent with an open conformation of the B-domain(Figure 3.29; top right). However, for the holo tetramer we see a bimodal distribution with peaks at 23Å and 28 Å which suggests that the bound F26BP is affecting the motion and open or closed state of the B-domain.

Monitoring distortions in the active site α6 helix: Asp 261 and Ala 264 are one helical turn apart in the short α6 helix which provides a crucial geometrical template for binding the substrate. The switch from the T to R states in LmPYK involves the rigid body rotation of the AC core[220] which enables Arg310 on one protomer to form specific hydrogen bonds to the backbone carbonyls of arginine and glycine residues located in the α6' helix of adjacent protomer). The short α6' helix (260-VAR**G**DL**G**VEIP-270) is unusual in that it contains two glycines. The MD simulation supports the idea that the allosteric mechanism may involve a transition between an ordered (R-state) helix able to bind substrate and a disordered (T-state) conformer. We observe a single, well-defined peak for the holo tetramer which corresponds with the ordered (R-state)

helix. The apo tetramer however shows a bimodal distribution for the apo tetramer with peaks at 5.5 Å and 7.8Å (Figure 3.29). These two peaks in apo tetramer suggest that the helix unwinds in the absence of F26BP.

Ligand-Induced Domain Closure: The B-domain is mobile in nature and undergoes a bending motion either towards or away from the A-domain. It has been seen that the presence of ATP in the active site induces closure of the active site and the B-domain moves towards the A-domain thereby reducing the volume of the active site. To analyze and quantitate the degree of ligand-induced closure of the active site, two distances within the cleft were measured for the open and closed conformation. First, Cα carbons (residues 57 and 93) from both the A- and B-domain at the edge of the cleft furthest from the hinge axis were used to indicate the greatest distance between the A- and B-domain. Second, the Cα atoms for a pair of residues (104 and 267) close to the hinge axis were selected to more fully describe the overall movement of the B-domain. The distances in the open conformation is larger as compared to the one in holo conformation A sharp peak in this is clearly visible. The peak is not only sharp; it is almost bimodal which suggests the partial opening of the B-domain for the holo tetramer. This corroborates with the fact that the enzyme is in a equilibrium between the two R-states as reported in the originally published paper, wherein we have partially opened B-domains in the R-state induced by F26BP but the fully active R-state has fully closed B-domains when the enzyme is fully ligated. However, for the apo tetramer, the distance between the residues is high as can be seen with the wide peak observed for apo and thus the B-domain is open. The graphs between various residue pairs point to the the fact that lmpyk structure shows the population shift model of allostery[159]. This is in agreement with the observations from the frequency distribution graphs (Table 3.8).

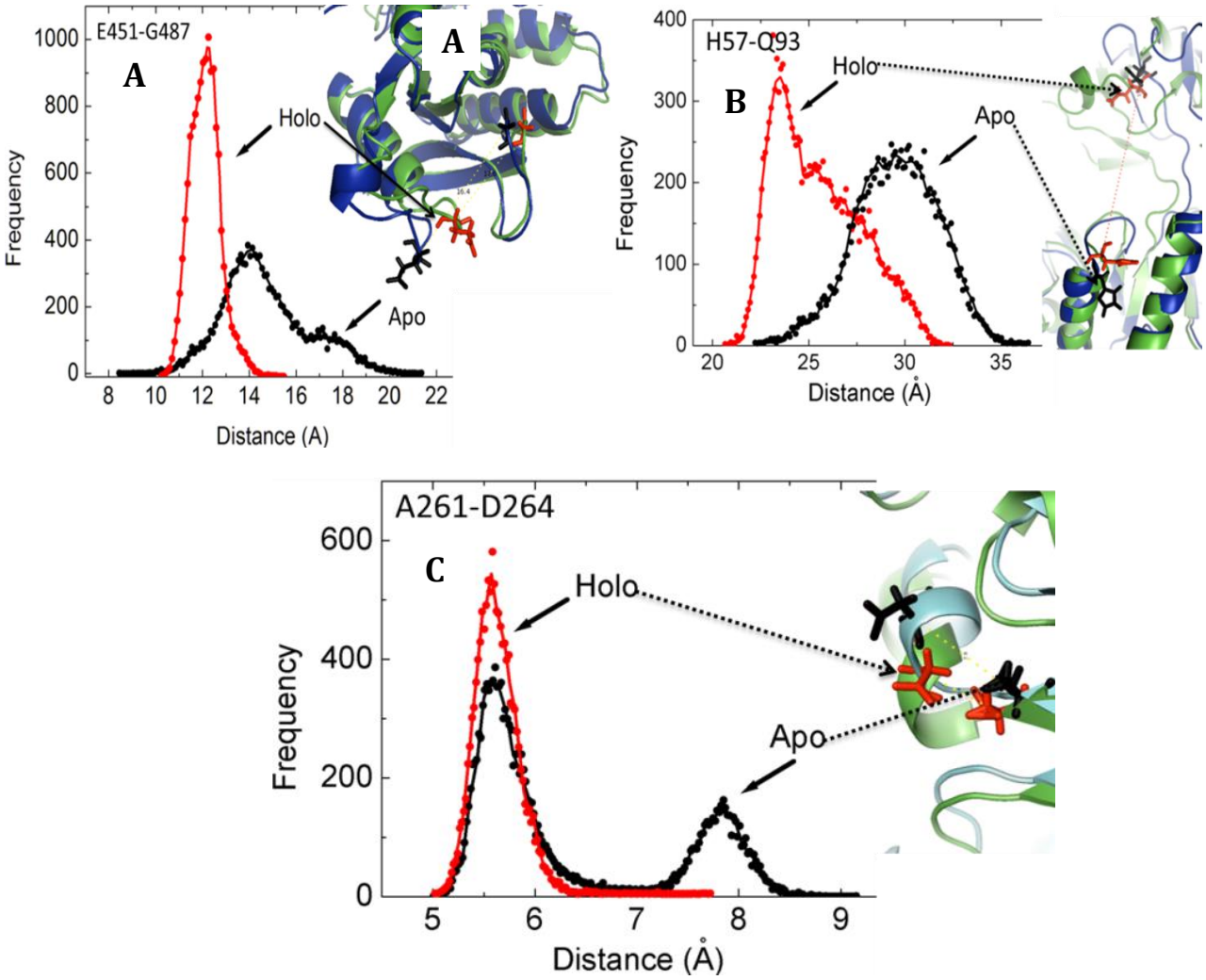


Figure 3.29 (Top Left to Right) : A: Distribution of the distance between GLU451-GLY487 for the apo (black dots) and the holo (red dots) obtained from the MD trajectory, B: Distribution of the distance between His57-Gln93 for apo (black dots) and holo (red dots) obtained from the MD trajectory, C: Distribution of the distance between Asp261-Ala264 for apo (black dots) and holo (red dots) obtained from the MD trajectory.

Inset figure shows conformations of the apo and holo tetramer . Apo is shown as a blue cartoon and the two residues are in stick representation in black. Holo is shown as a green cartoon with residues shown as red sticks .

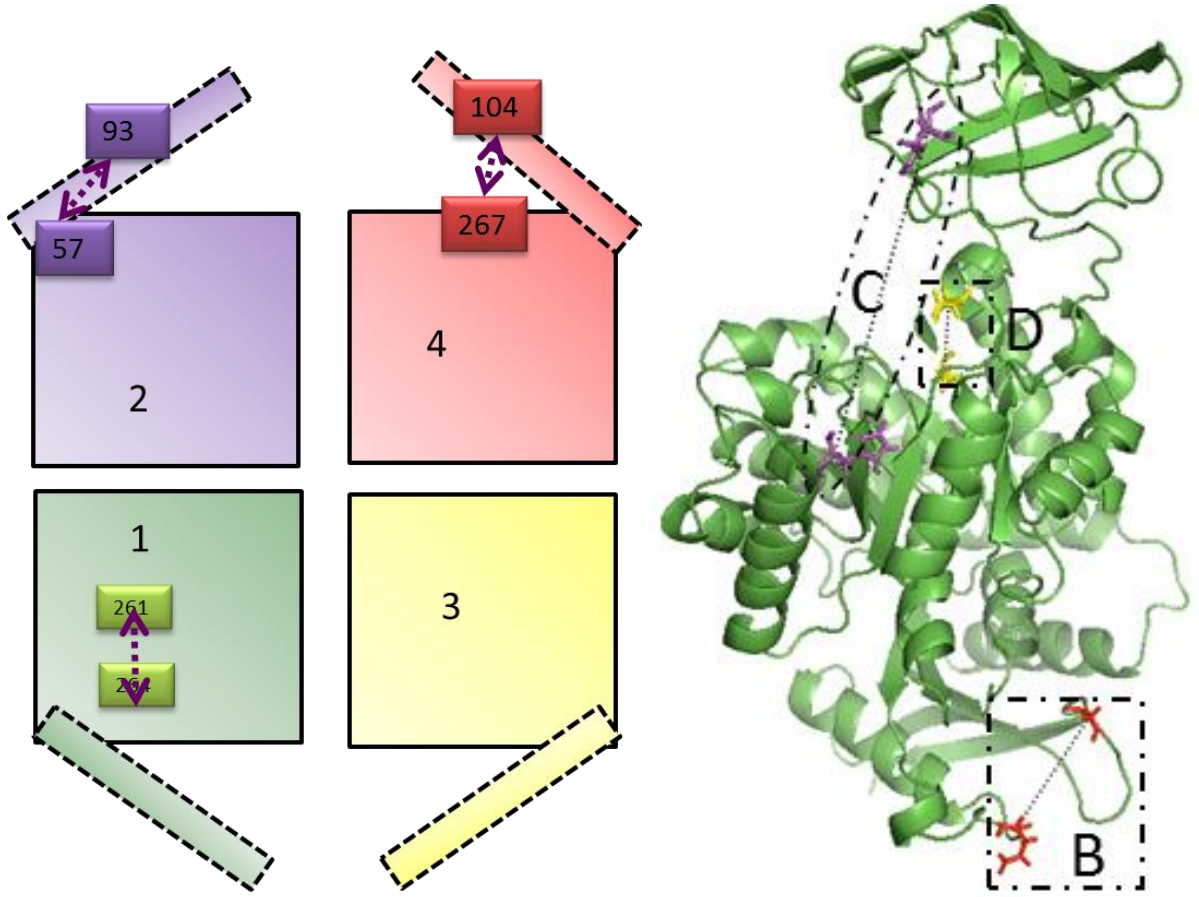


Figure 3.30: Left to Right): A: Schematic representation of the selected pairwise distance distributions within a pyruvate kinase protomer showing conformational changes induced by FBP binding, B: Cartoon of a pyruvate kinase protomer highlighting the residue pairs. 451-487;red, 261-264;yellow, 57-93; purple.

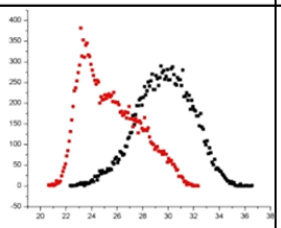
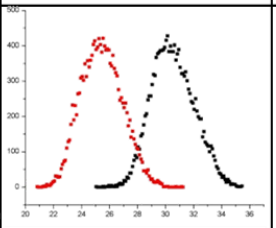
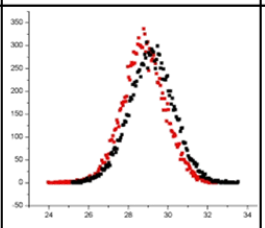
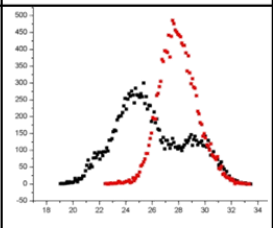
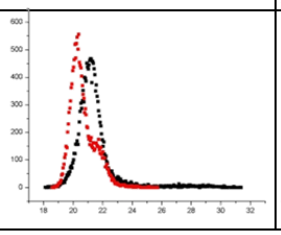
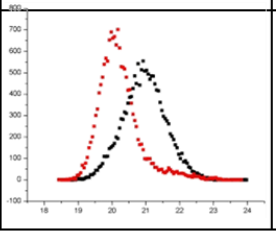
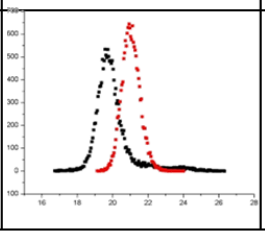
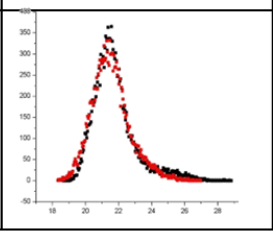
PAIR	A	B	C	D
57-93				
104-267				

Table 3.8: Distances between residue pairs 57-93 and 104-267 during the whole simulation for individual protomers. Apo tetramer is shown in black and holo tetramer is shown in red. X axis denotes the distance in angstrom and Y axis is frequency.

3.3.7 CORRELATION MATRICES

In order to further probe the dynamic behaviour of PYK and understand the fluctuations caused by B-domains within the tetramer, we performed a correlation analysis. This helped to understand the dynamic characteristics of the protein. The correlations were obtained from the atomic fluctuations as described in materials and methods section. Correlated motions can occur among proximal residues and also between regions as in domain–domain communication.

The B-factors yield information on the fluctuations of the individual residues but do not contain information about the correlations between fluctuations of two different residues. The measure of correlation between the fluctuations ΔR_i and ΔR_j of i^{th} and j^{th} alpha carbons can be assessed by calculating the projection of one on the other, i.e.,$\langle \Delta R_i . \Delta R_j \rangle$ at every instant and averaging over the full trajectory. The average,$\langle \Delta R_i . \Delta R_j \rangle$**,** if positive, indicates that the two residues move, on the average, in the same direction. A negative correlation, or anti-correlation, indicates that the two atoms move in opposite directions. If two residues are displaced equally along the same direction, then their motions will be positively correlated and the distance between them will not change. Positively correlated motions represent rigid body motions. If, on the other hand, two atoms move in opposite directions, their motion will be negatively correlated and the distance between them will either increase or decrease[286]. Most generally, positive correlations involve neighboring groups, which move together. Particularly interesting are the anticorrelated motions, especially those that involve entities at long- and midrange across the active sites. However, a completely correlated or anticorrelated motion, $C(i, j) = 1$ or $C(i, j) = -1$, means that the motions have the same phase and period.

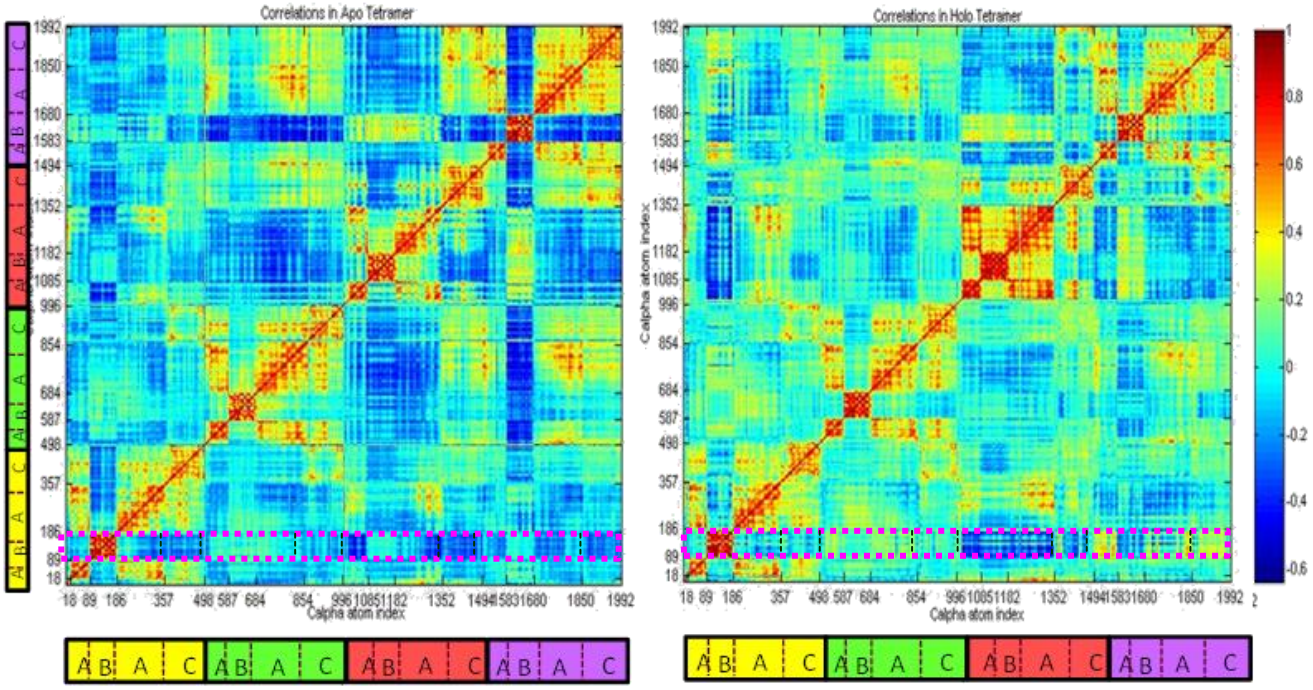


Figure 3.31: Correlation maps for the Apo (Left) and Holo Tetramer (Right). The X and Y axis indicate the calpha residue number and the correlation color ranges from blue (anti-correlated) to red (positively correlated or total correlation).

In Figure 3.31, the diagonal represents the correlations of the residues with themselves hence, the color red implying total correlation ($C(i, j) = 1$). We observe an overall decrease in negative correlations for the Holo Tetramer most notably in the regions involving the C-domains. For instance, the B-domain (89-186) of the first subunit show negative correlations with the residues of the C-domains (357-498; 854-996; 1352-1494; 1850-1992) across the four subunits in the apo tetramer. However, for the holo tetramer, we see that there is a decrease in the extent of negative correlations between the B and C-domains. This suggests that in the presence of the ligand (F2,6BP), the domains show rigid body motions. Also, the correlations become positive for the A-domain in the presence of F-2,6BP.

Anti-correlated residue motions and their changes upon effector binding provide significant information on the allosteric behavior of the LmPYK tetramer. The negative correlations of the tetramer are visualised in Figure 3.32a by shading each anti-correlated residue pair. Shading is proportional to the

magnitude of the anti-correlation and darker shaded regions correspond to groups or domains of anti-correlated residues. The white regions in Figure 3.32a correspond to protomer pairs with positive correlations. The numbering and the pairwise motions of the protomers for each shaded region are shown in Figure 3.32d. The tetramer consists of protomer 1 (residues 1-498), protomer 2 (499-997), protomer 3 (998-1496) and protomer 4 (1497- 1995). The clear block like pattern in Figure 3.32a shows that the protomers are moving in an anticorrelated fashion which is consistent with a concerted rocking motion depicted in Figure 3.32d. Also, the dark regions result from fluctuations of the system with fixed centroid of the tetramer. Fluctuations of a protomer obtained by fixing its centroid shows correlations internal to the protomer.

In terms of negative correlations, if we observe between residues 56-118 (from the B and A domain of first subunit), they are negatively correlated implying that the B and A domain are closer. This is due to the fact that the movement of B domain towards A domain and A domain towards B domain is a negative correlation with a change in the distance between the two. Now that we observe in different chains the pattern, the correlations become positive for the same pair in second and third subunit and again negative for fourth subunit. Subunits 1 and 4 have negative correlations for the same pair in tetramer in comparison to 2 and 3. This also explains the difference in peaks between the subunits for the B-domain as observed in figure 3.12 previously. In Figure 3.12, we see a reduced effect of fluctuations for both apo and holo tetramer i.e. we observe that subunits 1 and 4 have higher peaks as opposed to subunits 2 and 3. This suggests that PYK shows a 'coupled-dimer' behaviour i.e. as dimer of dimers. The two subunits (2 & 3) have low amplitude peaks because the correlations between the residues become positive or/are less negative (-0.04, 0.19, 0.33 and -0.18, respectively). Furthermore, it can be seen that for this particular residue pair, the holo has more negative correlation than apo. That means in Holo, the B-domain comes closer to the A-domain. This is to accentuate the fact the in holo state, the B-domain is closed and hence when it comes towards the apo form, it is more stable. This stabilisation is brought about by the incorporation of FBP.

A more detailed analysis of the matrix showing anti-correlations in a single protomer within the tetramer is presented in Figure 3.32b, where the dark regions show the negatively correlated regions. The fluctuations of the centroid of the protomer have been removed from the trajectory of the monomer, thus the calculated fluctuations are internal to the monomer. Strongest anti-correlations are between the B-domain (residues 89-186) and the N-terminal (1-89) and the C-domains (340-498). The change of correlations upon binding of the effector is of interest. In Figure 3.32c, the shaded regions identify those residues of the protomer where anti-correlations are either weakened or became positive upon binding of the effector. The weakened anti-correlated movements between the B-domain and the N-terminal residues and between the B and C domain residues are consistent with the interpretation from the temperature-factor analysis showing that effector binding dampens the motion of the B-domain. As an example, the negative correlation between residues Asn17 in the N-terminal and Lys118 in the B-domain is -15.0 $Å^2$ in the apo structure and reduces to -4.4 $Å^2$ upon effector binding. As another example, the negative correlation of -15.4 $Å^2$ between Glu156 in the B domain and Asn340 in the C domain of the apo structure reduces to -8.8 $Å^2$ upon effector binding. Since negative correlations result from the opening and closing of the B-domain cap, their decrease suggests that binding of the effector reduces the amplitude of the B-domain oscillations. Notably, the dark regions show the decrease of negative correlations upon binding of the effector. This is consistent with our previous observation of effector binding dampening the motion of the B-domains.

The anti-correlations of the apo structure are stronger than those of holo showing that binding of the effector decreases the negative correlations of the cap region and the effector binding region (Figure 3.31). The differences between Figures 3.14 and 3.15 may be quantified by considering the same residue pairs, i.e., the correlation between residues Asn17 in the N-terminal and Lys118 in the B-domain is -8.5 $Å^2$ in the apo structure and reduces to -5.7 $Å^2$ upon effector binding. As the other example, the negative correlation of -4.4 $Å^2$ between Glu156 in the B domain and Asn340 in the C domain of the apo structure reduces

to -1.8 $Å^2$ upon effector binding. We conclude from this observation that allosteric activity is influenced by the behaviour of the B-domain cap which shows anti-correlated behaviour.

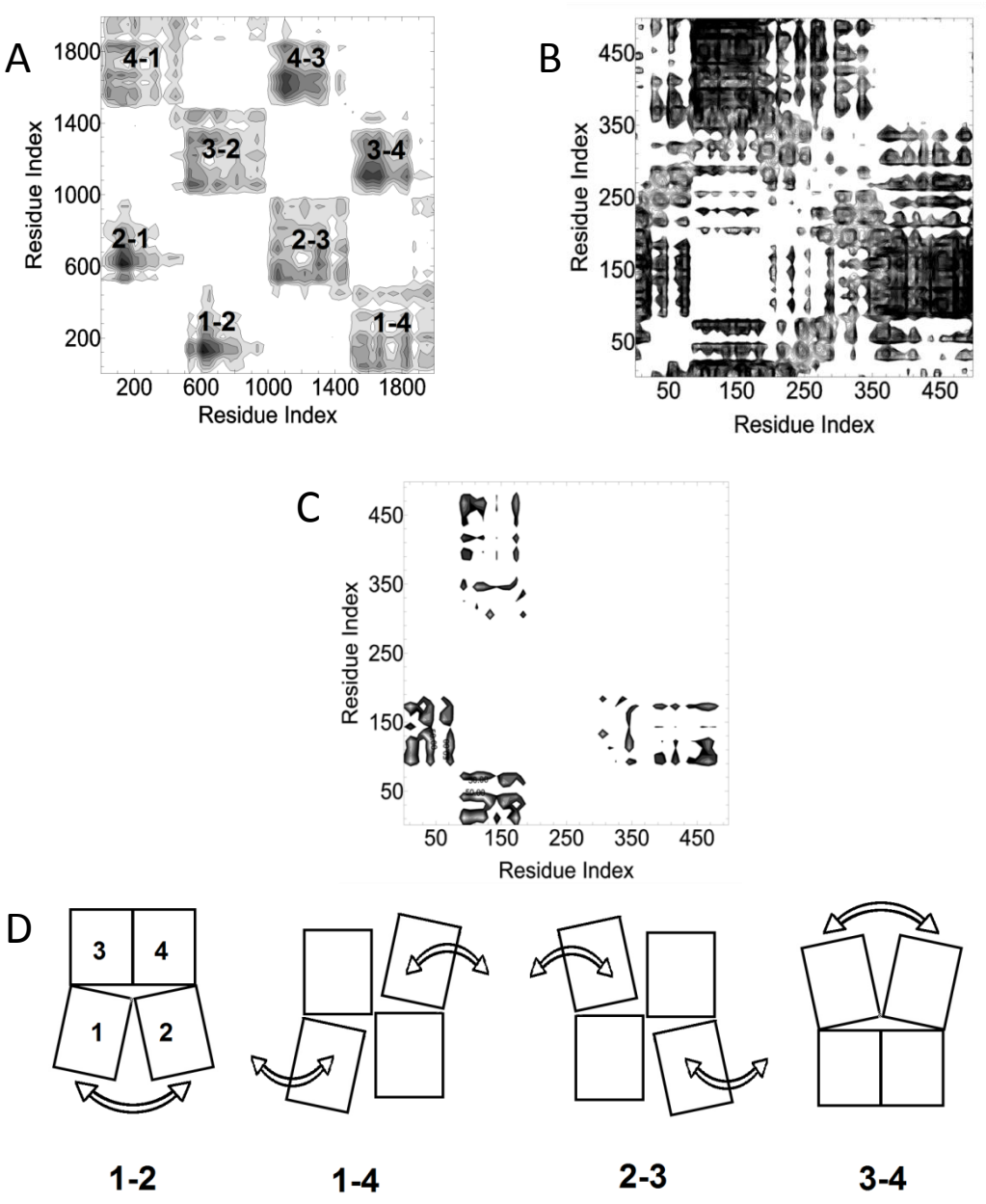


Figure 3.32: Correlated atomic motions in pyruvate kinase; a: Negative correlations for the apo tetramer, b: Negatively correlated regions of the apo protomer. c: Differences in negative correlations of apo and holo protomer, d: Cartoon of the anti-correlated rocking motions of the 4 pyruvate kinase protomers.

3.4 DISCUSSION

There are been some success in reproducing the microscopic observables in our macroscopic ensemble of simulations. The RMSD and RMSF graphs capture the crystallographic B-factors considerably well which increases the confidence in our simulation design. Through these graphs we have shown the stability and convergence of the simulation. We have been able to gain a considerable insight into the conformational dynamics of PYK, specifically with regards to the envisaged "rock and lock mechanism". The simulation protocol followed has also been successful enough to capture the sensitivity of the protein to the presence of allosteric effector, F-2,6-BP and the substrate molecule at the allosteric and active site, respectively.

The monitoring of the various salt bridges and contacts between the residues through our distance fluctuation analysis support our rock and lock mechanism. The opening and closing of the cap residues and the contribution of domains in the overall fluctuation of the enzyme strongly supports our view of contribution of B-domains to the stability and flexibility of the tetramer. The binding of allosteric effector F-2,6BP at a site which is considerably far enough from the active site and also the B-domains has been quite intriguing in respect to the response of B-domain fluctuations. Despite being more than 40 Å away from the cap region, it effectively stabilizes the movements of the B-domain by reducing the fluctuations of the cap residues. This reduction in fluctuation thereby makes the molecule more rigid and compact. The loop connecting the two domains adopts a fairly loose conformation in the absence of an allosteric activator.

One of the unsolved mysteries of allostery is how the overall dynamics is coupled to the local dynamics, e.g., of active site residues. This can be reformulated to the question which correlations exist between overall collective motions and local collective motions, or side chain rotamer states. An additional question pertains to the coupling time. The relevant local motions may not have an instantaneous effect, so the analysis may require taking time shifts into account. These have been probed with the help of mode analysis. Our Principal

Component Analysis thereby confirms the concerted motions observation. It is however interesting to note the assymetrical distribution of B-domain movements. The molecule despite showing concerted tetramer movements, displays a slight drift from the classical model of allostery with assymetrical distribution of the fluctuations between the diagonals where the opposite diagonals seem to behave in similar manner and amplitude of fluctuations. Overall we notice the allosteric control exerted by assymetrical reduction in flexibility of the enzyme.

There is observable asymmetry in the dimers with the tetramer behaving as a dimer of dimers which was seen from the mean square fluctuation analysis of LmPYK. There seem to be a dynamic asymmetry between the dimers of the same tetramer which are diagonally positioned. This could well be a consequence of the subunits sequentially binding and releasing the effector molecule in a diagonally symmetric way and thus leading to stabilization of the B-domain in the bound state.

3.5 SUMMARY

The MD simulations of tetrameric LmPYK in the apo and F26BP-bound states have been used to provide insight into the allosteric effector mechanism triggered by F26BP. By studying the dynamic behaviour of the tetramer compared with isolated monomer chains the simulations also show how the major allosteric effect is linked to a concerted motion of the complete tetramer. The analysis of correlated motion is consistent with a synchronised rocking of the protomers. The differences in the magnitudes of the B-factors on effector binding suggests that the F26BP binding dampens molecular motion throughout the tetramer and effectively cools down the tetramer and traps the tetramer in the active R-state conformation. This hypothesis is also consistent with the difference correlation plots which show that the magnitude of the anti-correlated movements, particularly involving the B-domain, are reduced on F26BP binding. Interestingly the behaviour of an isolated monomer PYK chain shows

almost no effect on the overall B-factor on F26BP binding suggesting that the coupled movement of the protomers is required for this damping effect.

These observations fit well with the classical Monod-Wyman-Changeux (MWC) model of allostery which suggests that oligomeric enzymes undergo symmetrical transitions (classically between the T- and R-states) that can be stabilized by ligand binding [44]. Our observations showing a lowering of B-factors on effector binding might further suggest the effector is not only locking the R-state using local van der Waals and hydrogen bond interactions (as suggested previously [1]), but is also acting as a general heat-sink to cool down the whole tetramer. More recent descriptions of allostery highlight the role of entropy changes on effector binding[19, 25] as well as the importance of the intrinsic dynamic nature of the protein[19, 34]. Our observations on the effect of F26BP binding on ligand flexibility suggest that protein rigidity (related to melting temperature) and intrinsic heat capacity are important factors in stabilizing the R-state of PYK, and other allosteric proteins.

Another insight about the flexibility of PYK comes from a comparison of tetramer with the isolated monomer which shows that the B-domain in the tetramer is more mobile than in the isolated monomer. Interestingly, the B-domain movements in the tetramer have been frequently found to be asymmetric in a large number of PYK X-ray structures (e.g. PDB codes 1PKY, 1PKN, 1AQF, 1F3W)[268]. The asymmetry in these structures is such that B-domains of protomer pair 1 and 4 have similar B-factors and relative orientations which are frequently different from those of protomers 2 and 3. The asymmetry is sometimes imposed by the tetramer lying on a crystallographic 2-fold axis, but there are also many examples showing the same behaviour when there is no crystallographic constraint. The enhanced B-domain movement in the tetramer compared with the isolated monomer again suggests that interprotomer interactions regulate the pairwise B-domain movements. The strong anti-corrleation signals within each protomer (Figure 3.32) highlights the flapping of the B-domains though it is not clear mechanistically how the binding of the

effector dampens the motion or how the synchronous movement of the B-domain pairs is regulated.

The analysis of pairwise distance distributions is consistent with the results from the analysis of correlated motions in the tetramer and shows that F26BP binding not only restricts the mobility of the B-domain (as shown in Figure 3.13) but also keeps the active site helix tight and the enzyme in an active R-state conformation (Figure 3.17). This helical order-disorder transition between the T and R-states of LmPYK is key to explaining the on-off state of the enzyme as the phosphenoyl pyruvate substrate can only bind in the active site when the short α6'-helix (260-VAR**G**DL**G**VEIP-270) adopts an ordered conformation. That these MD simulations have captured this transition for the unliganded apo form within the 80ns simulation indicates the relative instability of this glycine-rich helix. It is also significant in the simulations that the holo state is in the ordered (active R) state over the complete simulation while the apo (unliganded) can sample both disordered and ordered conformations (Figure 3.23). These results suggest that the apo-enzyme can sample a range of conformational states, some of which would be similar to the active R-state. This dynamic description of PYK fits with the 'ensemble representation' and 'population shift' ideas of allosteric activation [19, 287].

The MD analysis presented in this work therefore supports a model in which concerted domain movements as described in the MYC model dominate the allosteric mechanism of LmPYK. However the analysis of specific order-disorder transitions and the 'heat-sink' effect of the bound effector highlight the importance of entropic and vibrational movements in regulating the allosteric effect of effector binding on the enzyme activity of leishmania pyruvate kinase.

CHAPTER 4:

CASE STUDY II

PHOSPHOFRUCTOKINASE

4 PHOSPHOFRUCTOKINASE

4.1 Introduction

4.1.1 Background

Phosphofructokinase (PFK) is another important enzyme in all living cells, which performs the irreversible conversion of fructose-6-phosphate to fructose-1,6-bisphosphate, a committed step in glycolysis(Figure 4.1). As mentioned previously in chapter 1 in detail, glycolysis is one of the major carbohydrates producing reaction pathway besides citric acid cycle and electron transport chain. It usually takes glucose as a starting point and by a series of reactions, converts it into pyruvate. Thereby, maintaining a steady amount of ATP, this acts as a major energy carrier in the living cells.

Phosphofructokinase, being a multisubunit protein, exhibits cooperative kinetics, thus displaying allosteric behaviour in which the activity of subunits is inter-dependent. As mentioned previously, allosteric enzymes can be regulated by binding of an effector molecule which acts as an activator or inhibitor. The cooperative nature of subunits results in activation or inhibition of subsequent subunits depending upon the effector molecule.

The main aim of this study is to computationally investigate the behaviour of phosphofructokinase upon addition of ATP. For exploration of the behaviour of PFK, we have run three broad simulations on the apo and holo forms i.e. for monomer in two conformations namely, apo monomer conformation and holo monomer conformation and on the holo tetramer. Both the crystal structures have been determined by the Walkinshaw group members previously [288, 289], thus providing us with much more authentic data to further discuss the results of the simulation. In this chapter, we present a case study designed to observe and analyse the structural behaviour and conformational fluctuations of phosphofructokinase. We have used the concepts and application of molecular

dynamics simulations to understand the protein function by observing the local conformational changes and flexibility.

The present analysis of exploration of the dominant motions and fluctuations of PFK in this research produce comparable results with the already published data thus increasing the reliability of our simulations. The results of these simulations are summarised below in the following manner; first we describe the starting structures for our simulations, check the convergence of parameters and probe the conformational flexibility of the enzyme. Following this, we run Principal Component Analysis and inspect the eigenvectors and projections and finally analyse the correlation matrices.

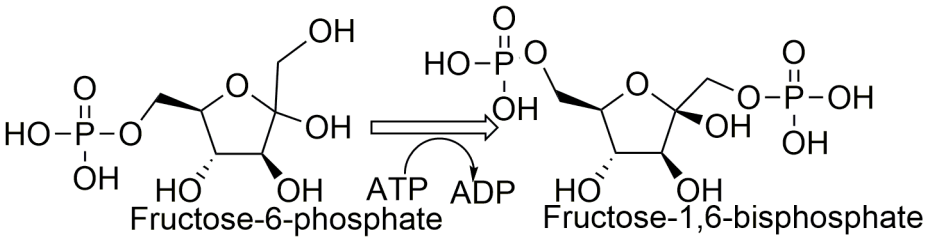


Figure 4.1: Phosphofructokinase converts Fructose-6-phosphate to Fructose-1,6-bisphosphate using ATP for the phosphate transfer.

4.1.2 Trypanosoma brucei Phosphofructokinase (Tb. PFK)

Phosphofructokinases are categorised into two broad groups i.e. ATP-dependent (ATP-PFK) and inorganic pyrophosphate dependent PFKs (PPi-PFK) dependent. This is based upon the type of phosphate source. The classification of PFKs is affected by a single gene mutation with the size of eukaryotic PFKs almost twice in comparison to prokaryotic ATP-PFKs. However, the PPi-PFKs are not monophyletic in origin as observed for most of the ATP-PFKs. The ability to use ATP or PPi is mainly dependent on the presence of one invariant glycine

residue. The PFKs of Trypanosomes are ATP-dependent with Tb. PFK regarded as a chimera of ATP-dependent and PPi-dependent PFKs. This is due to the fact that although it uses ATP as the substrate, but it still possesses structural motifs found in PPi-dependent PFKs[288] [290, 291].

The Walkinshaw group together with other researchers has been extensively investigating the glycolytic pathway enzymes as therapeutic drug targets. One of the major interests has been to study and characterize the role of PFKs in trypanosomes. These are responsible for some of the major neglected tropical diseases like sleeping sickness, chagas disease, leishmaniasis, etc., and are broadly classified into three species, namely, Trypanosoma *brucei*, *Trypanosoma cruzi* and *Leishmania* species.

Glycolysis regulation in Trypanosomes: Trypanosomes show a high degree of dependence on glycolysis for the production of ATP. This pathway serves as the sole catabolic source of energy i.e. ATP at the infective stage of T. brucei. Several glycolytic enzymes like, Phosphofructokinase, Hexokinase, Enolase and Pyruvate kinase affect the life cycle of these parasites and the depletion of either of these enzymes could result in the death of these parasites[266, 267]. Hexokinases and phosphofructokinases, which hydrolyse ATP are tightly regulated in cells where glycolysis occurs in the cytosol. However, in case of trypanosomatids, the first seven enzymes of glycolysis are sequestered inside a peroxisome like compartment, glycosome[292] (Figure 4.2). Thus, there is no feedback control by the products of hexokinase and also the regulation of PFK is somewhat simplified in contrast to its bacterial or mammalian counterparts. The essential role of the glycosomes as indicated in various studies is to prevent hexose kinase (HK) and phosphofructokinase (PFK) from being exposed to the high ATP concentrations of the cytosol as there is no net production of ATP. The ATP used by these two enzymes is balanced by the two ATP molecules produced by phosphoglycerate kinase (PGK) and hence prevents the accumulation of metabolites [293, 294].

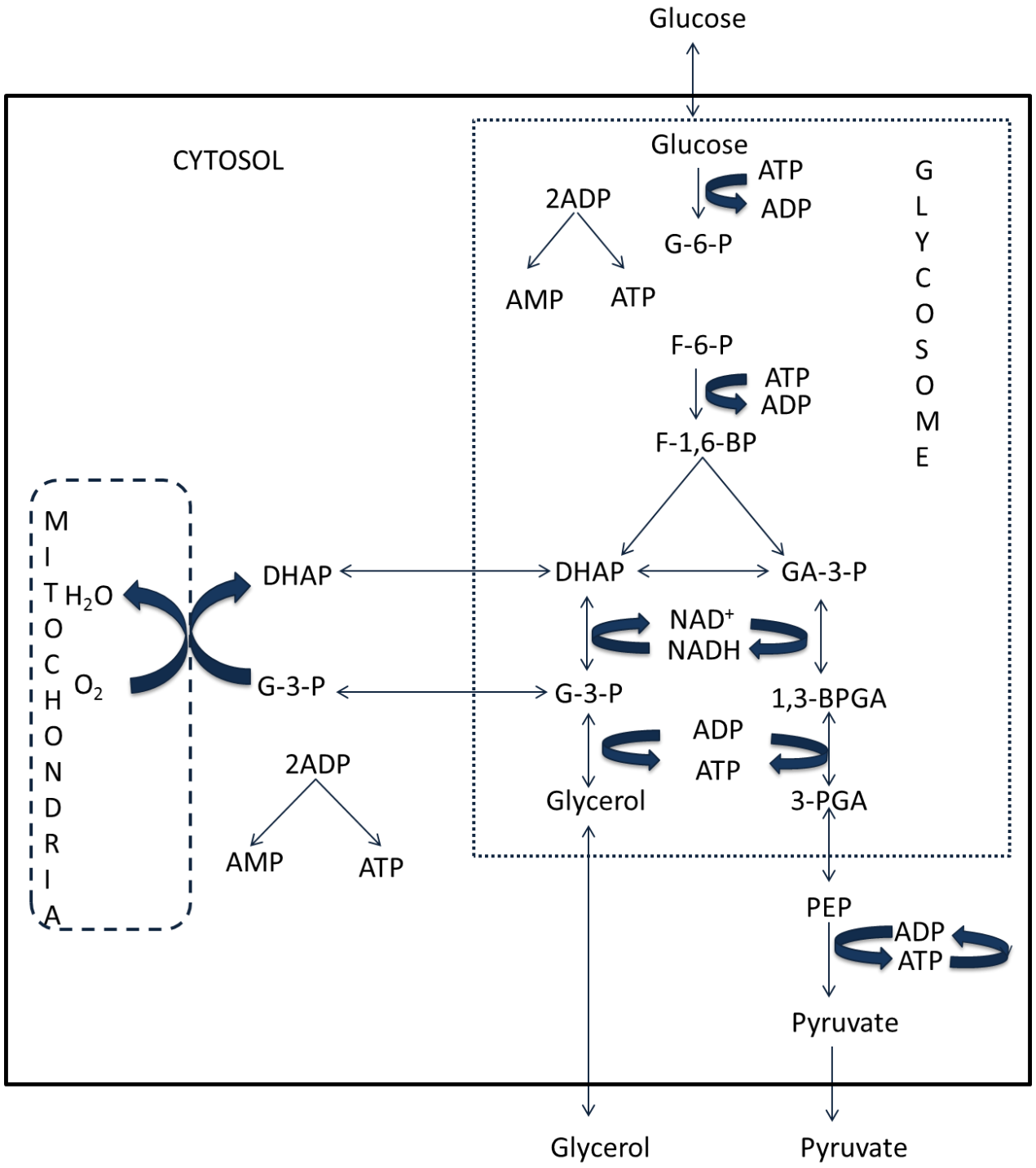


Figure 4.2: Illustration of the glycolysis event in T. brucei.(Adapted [2])

4.1.3 Structure of phosphofructokinase

T. brucei PFK (TbPFK) consists of four identical subunits containing 487 residues and weighing 50 kDa each. AMP is the sole allosteric activator of TbPFK unlike other organisms which are regulated by a number of modulators [295, 296]. Several bacterial PFKs have been studied and crystal structures of ATP-dependent PFKs from *Escherichia coli*, *Bacillus stearothermophilus[297]*, and *Lactobacillus bulgaricus* [298] and the PPi-dependent ATP from *Borrelia burgdorferi* are readily available[299]. TbPFK consists of 3 domains, the two compact B (residues 95‑233 and 386‑409) and C (residues 234‑385 and 442‑453) domains and a third loosely packed domain A (residues 8‑94 and 410‑441). This domain is unique to trypanosomatid PFKs as it possesses the embracing arm (residues 62‑81)[289] that links with the corresponding arm of the adjacent subunit. Also, an additional loop (residues 329–348) which is adjacent to the active site is present in trypanosomes unlike the bacterial or human counterparts (Figure 4.3).

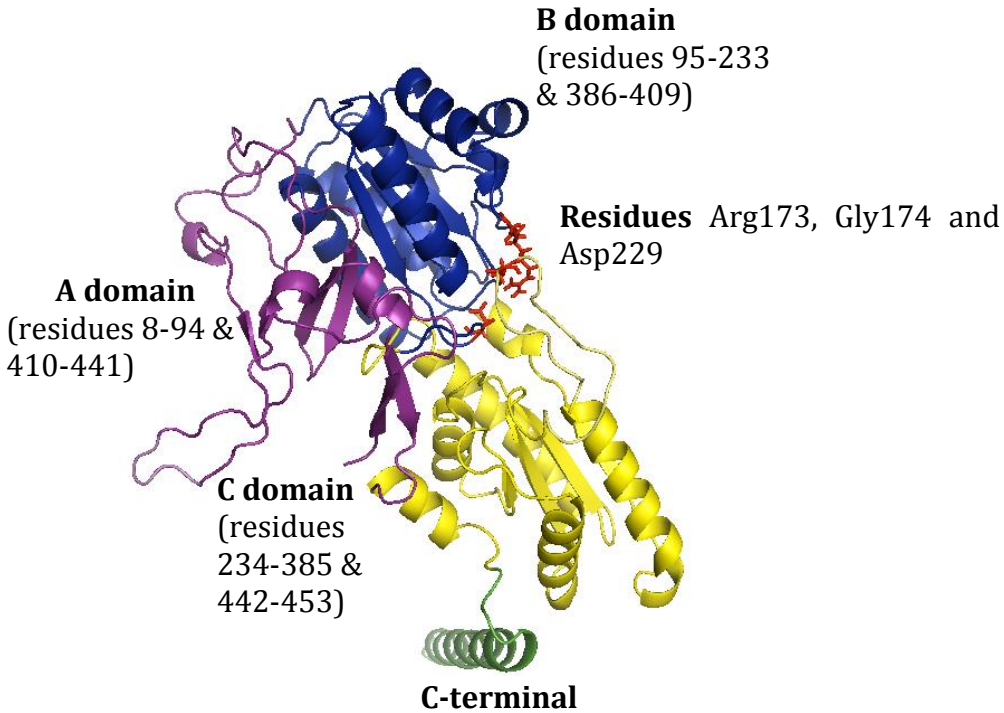


Figure 4.3: Structure of *T.brucei* PFK apoenzyme subunit. The three domains are colored differently: A domain; purple, B domain; blue, C domain; yellow and the C-terminal in green. Residues Arg173, Gly174 and Asp229 are shown in red as sticks.

Active site of TbPFK

The active site is present at the boundary of B and C domains, with the B domain housing the ATP molecule. The ATP-binding site is sandwiched between the N-terminal portion of residues 199–204 and the loop containing Arg173 is found at the boundary of domains B and C, with the ATP bound primarily to the B domain. The large inserted 329-348 loop coordinates the Mg ion which interacts with the α-, β- and γ- phospho groups.

The crystal structure of TbPFK apoenzyme (PDB ID: 2HIG) has two distinct alternative conformations of the active site. These are mainly in the loop regions i.e. the catalytic Asp229 loop, Arg173 loop and the inserted 329-348 loop. One conformation is open (APO-1) where Arg173 loop facilitates the binding of ATP while the second conformation (APO-2) is closed where the loops block the binding of ATP by moving into the active site (Figure 4.4). The hydroxyl of Ser341 from the inserted loop interacts with the side chain of Arg173 which blocks the site. However, in case of holoenzyme, Ser341 interacts with the backbone of Gly174, thereby making way for the ATP to bind at the active site. The catalytic Asp229 loop is positioned away from the active site unlike the APO-2 conformer where it is found to block the phospho group. In general, the APO-1 conformation of the apoenzyme shows similarities with the structure of holoenzyme (PDB ID: 3F5M).

Effector site of TbPFK

The only known allosteric activator of TbPFK is AMP which differs from PFK in other organisms which are allosterically activated by a number of molecules. Also, in comparison to other ATP-dependent PFKs, TbPFK has no known allosteric inhibitors. The binding of effector is mainly controlled by the C-terminal region where the reaching arm forms a lid over the effector site. The effector site, however remains unperturbed by the binding of ATP molecule.

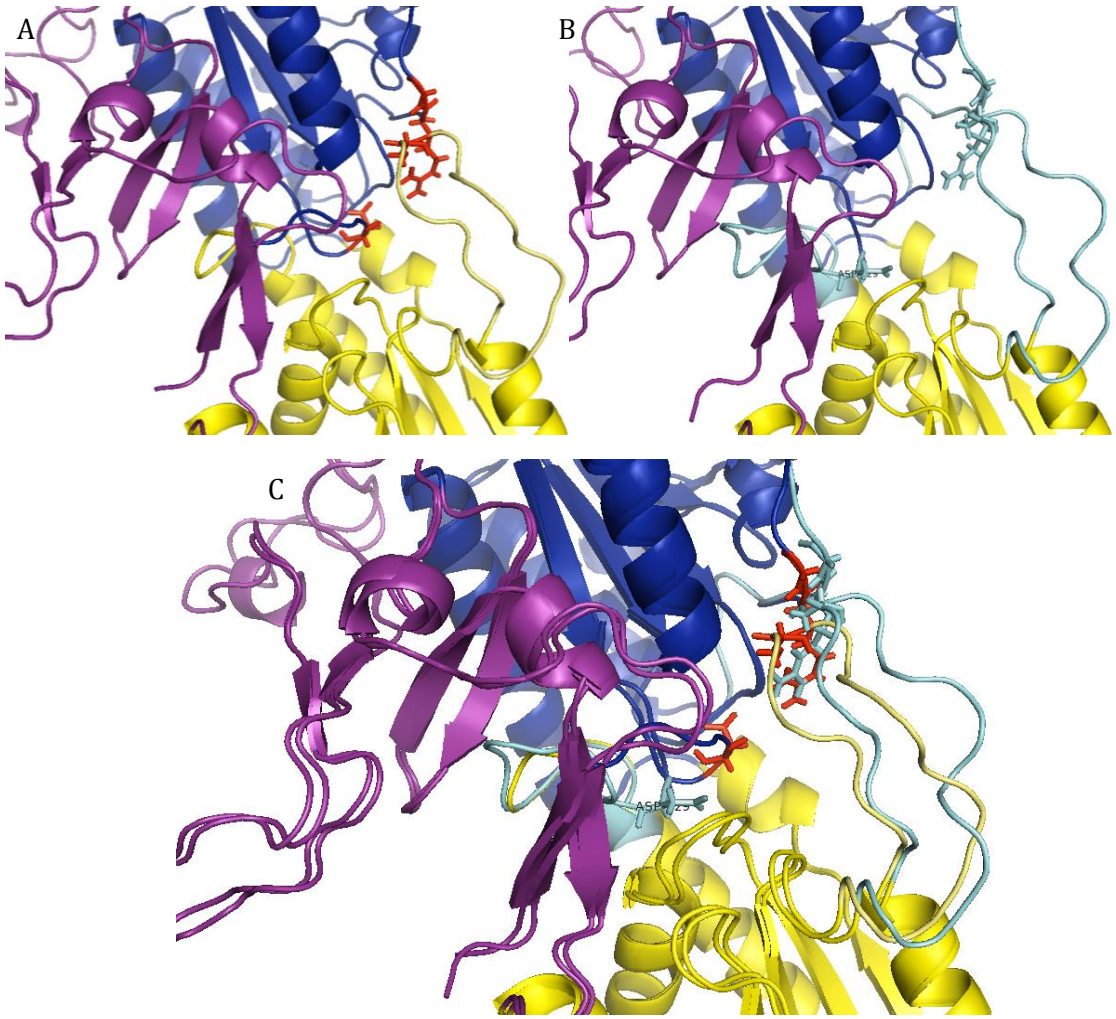


Figure 4.4: Structure of *T.brucei* PFK apoenzyme subunit showing alternative conformations of the loops at the active site of TbPK. The crystal structures have two subunits that differ in the region of the active site. A: Active site region of Apo conformation, B: Active site region of Holo conformation and C: Superpositions of the active site of both the subunits. The loops 329-348, Asp229 loop are colored as cyan for the holo conformation.

4.2 Molecular Dynamics Simulations

This section describes the details of molecular dynamics simulation designed to study the structural changes and monitor the dynamic behaviour of phosphofructokinase. We have analysed the simulation data by employing several techniques namely, Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), Principal Component Analysis (PCA) and correlation matrices as described previously in the materials and methods section of this thesis. The fluctuations and mode analysis produce results which complement the experimental data.

The following sections will present the results of this analysis.

4.2.1 System preparation

MONOMER SIMULATION

Starting Structure:

The structure of ATP-dependent phosphofructokinase from Trypanosoma brucei (TbPFK) was modelled from the crystallographic coordinates as obtained from the Protein Data Bank Entry 2HIG (resolution 2.4 Å). This structure consists of two subunits (A and B) and is a 50 kDa protein with 487 residues in each chain. T.Brucei PFK apoenzyme is folded into two compact domains (domain B residues 95–233 and 386–409; and domain C residues 234–385 and 442–453). The remainder of the subunit is loosely packed domain A with residues 8–94 and 410–441 (Figure 4.3).

Two separate MD simulations were carried out using the 2HIG crystal structure (Table 4.1). As mentioned previously and reported in the literature, this structure consists of two subunits. Although, these have identical sequences and indistinguishable structures, there are two alternative conformations of two loops at the active site (Figure 4.4): the tip of the 329-348 loop and the 229 loop that contains Asp229. Subunit A adopts a relatively closed conformation with well-

defined density for the entire 329-348 loop and is stabilized by interactions between the Asp231 and Asn343, and between the side-chain of Asp231 and Lys344. It is stated in the literature that the tip of the loops approach each other in a similar way as observed in bacterial PFK holoenzyme. This provided us with a putative holo monomer structure for simulation. Subunit B was taken for the apo monomer simulation which has loops pointing out of the ATP-binding pocket. The enzyme thus adopts an inactive and more open conformation.

The simulations were carried out using the standard procedure as adopted in GROMACS software package. The forcefield parameter set of AMBER99sb-ildn was used to describe the structure of the protein. The starting structures of the protein were solvated in a dodecahedron box placed at a distance of 0.9 nm from the box boundary. The simulations were run in explicit solvent and periodic boundary conditions. The apo monomer system consisted of a total of 30507 molecules (95558) whereas the holo monomer system consisted of a total of 30428 molecules (95381 atoms). The Simple Point Charge (SPC) model was used to describe the water molecules while the system was neutralized with the addition of sodium (Na+) and chloride (Cl-) ions. The non-bonded interactions were evaluated using a twin range cutoff of 0.9 and 1.4 nm. A single cut off of 1 nm was used for the treatment of Van der-Waals interactions. Long-range electrostatics were treated using the Particle-Mesh Ewald (PME) method with 0.16 FF grid spacing and 4^{th} order B-spline interpolation for the reciprocal sum space. The reciprocal grid of 72 x 72 x 72 cells was used. The systems were also relaxed by 1000 steps of steepest descent energy minimization procedure prior to the simulations. In order to maintain a constant temperature of 318 K, V-rescale thermostat was applied which is a modified Berendsen thermostat. The protein and non-protein atoms were coupled independently to temperature baths with a coupling time of 0.1 ps. Inhibitor and the solvent were each independently coupled to a temperature bath. The pressure was maintained by weak coupling to a reference pressure of 1 bar, with a coupling time of 1.0 ps and an isothermal compressibility of 4.6×10^{-5} bar^{-1} [32]. The bond lengths and angle of the water molecules were constrained using the SETTLE algorithm [53] while the bond lengths within the protein were constrained using the LINCS algorithm [33]. The

final production run was carried out for 100ns for both the apo and holo monomer while the timestep used for integrating the equations of motion was 0.003 fs for the holo monomer and 0.005 fs for the apo monomer.

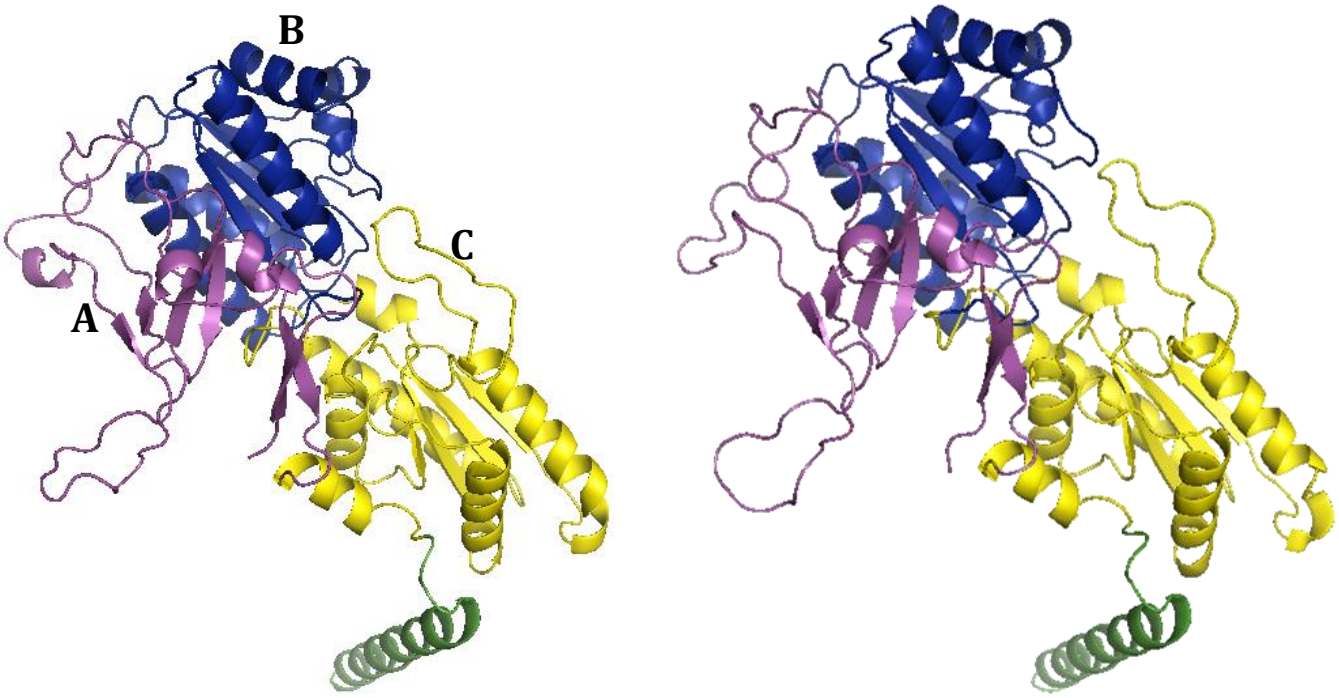


Figure 4.5: (L-R) The starting structures used for simulation derived from 2HIG crystal structure of Apo Monomer and Holo Monomer, respectively. The monomer is colored according to the domains. A domain;purple (residues 8-94 & 410-441), B domain; blue(residues 95-233 & 386-409), C domain; yellow (residues 234-385 & 442-453) and the C-terminal in green.

TETRAMER SIMULATION

Starting structure:

The Holo tetramer simulations was performed on the crystal structure of ATP bound TbPFK as obtained from the crystallographic coordinates of Protein Data Bank Entry 3F5M with a resolution of 2.70 Å.

The structure has similar subunit architecture as that of the apoenzyme with the main differences being at the active site, near the C-terminus region and the subunit contacts. As is the case with apoenezyme, the holoenzyme also folded into three domains; loosely packed domain A and the two compact domains B and C. The unique feature of ATP dependent PFK from trypanosomatids is the presence of embracing arm (residues 62-81) linking with the corresponding arm of the adjacent subunit. Another unique feature is the presence of inserted loop (residues 329-348) in the domain C which forms an important part of the active site.

The MD simulation of the active ATP bound holo enzyme structure was carried out using GROMACS[197] (GROningen MAchine for Chemical Simulations) package version 4.5.5[165] with AMBER99sb-ildn[270] forecfield parameter set. The system was prepared in a similar way as stated above for the monomer simulation. The final system has 170995 atoms (29639 protein atoms) and 141135 water molecules. Constant temperature of 318 K was maintained by coupling to the V-rescale thermostat[194] using a time step of 3fs. This was followed by an NPT equilibration, in which the pressure was maintained isotropically at 1 bar using the Berendsen barostat with a coupling constant of 0.1 ps. The water molecules and bond lengths were restrained using the SETTLE[271] and LINCS[272] algorithm respectively. A single cut off of 1 nm was used for the treatment of Van der-Waals interactions. Long-range electrostatics were treated using the Particle-Mesh Ewald (PME) method with 0.16 FF grid spacing and 4^{th} order B-spline interpolation for the reciprocal sum space. The system was relaxed by 1000 steps of steepest descent energy minimization procedure prior to the simulation. The final production simulation was run for 52 ns with the equations of motions integrated at the rate of 3fs. The snapshots were saved every 3ps, thereby generating 40833 frames 17562.

All the simulations were run the software GROMACS, versions 4.5.5 and 4.6 on HecToR amd ARCHER respectively as summarised in Tables 4.1 & 4.2

SOFTWARE	HECTOR	ARCHER
GROMACS	4.5.5	4.6.3
pRGeNV-CRAY	4.0.46	5.0.41
CRAY-MPICH	5.6.5	1.6.1

Table 4.1: Specifications of the supercomputing facility and gromacs software versions employed for PFK molecular dynamics simulations.

Structure	Machine	Time	Timestep	Snapshots frequency	Frames
Apo Monomer	HecToR	100ns	5fs	5ps	20000
Holo Monomer	ARCHER	100ns	3fs	3ps	33334
Holo Tetramer	HecToR	52ns	3fs	3ps	17562

Table 4.2: Time scales and the trajectory output for the three independent molecular dynamics simulations of PFK.

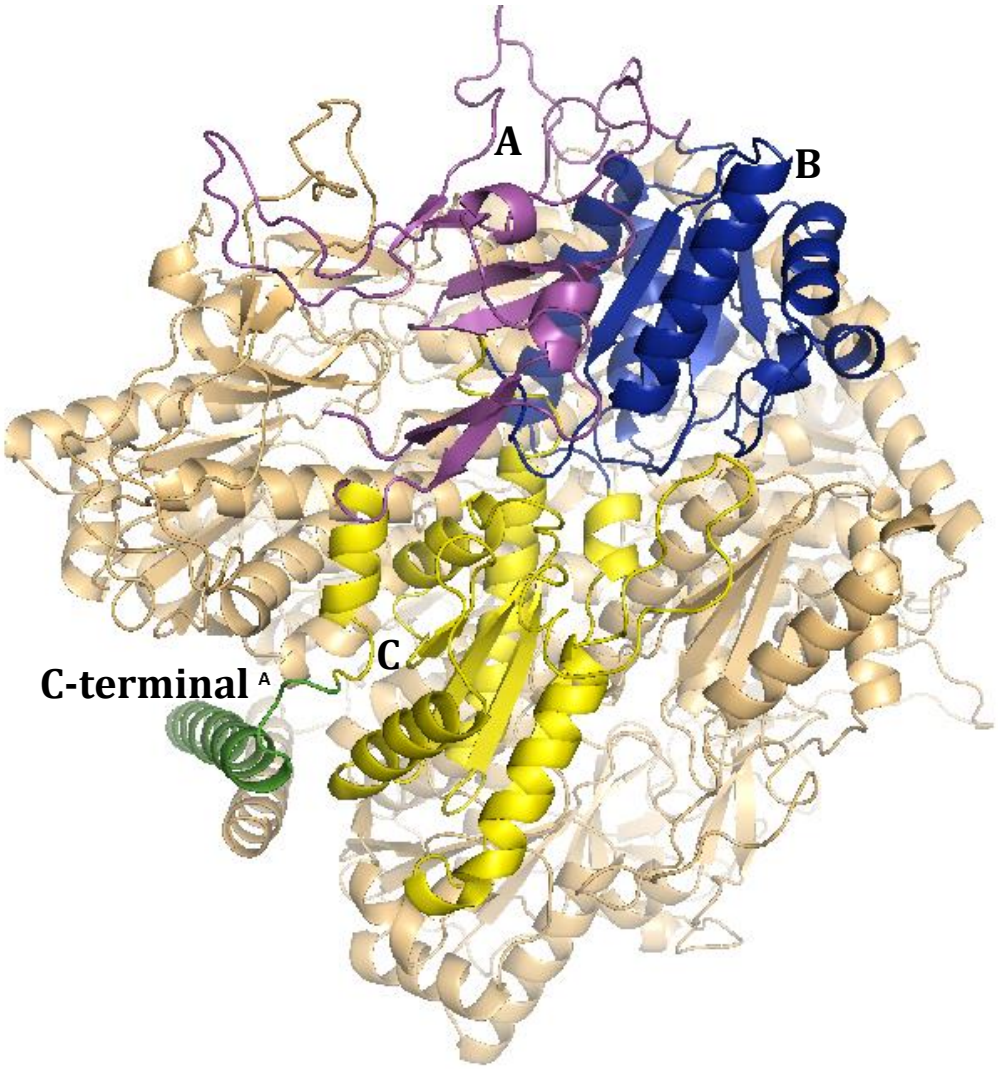


Figure 4.6: 3F5M crystal structure used for the Holo Tetramer simulation of PFK. The tetramer is in background while the subunit is colored according to the different domains in the foreground. The highlighted subunit is colored according to the domains. A domain;purple (residues 8-94 & 410-441), B domain; blue(residues 95-233 & 386-409), C domain; yellow (residues 234-385 & 442-453) and the C-terminal in green.

4.3 RESULTS

The total simulation time for phosphofructokinase was 252 ns. The simulations are summarized in Tables 4.2 & 4.3. From each trajectory, a set of instantaneous properties were derived to ensure the stability and convergence of the simulations. Following this, the analysis was carried out in terms of conformational flexibility i.e. Root Mean Square Fluctuations and Principal Component Analysis.

	Apo Monomer	Holo Monomer	Holo Tetramer
Protein size(atoms)	8028	8051	29639
Box Volume (nm^3)	997.60	989.76	1705
SOL	91473	91245	141135
Calpha atoms	475	476	1905
Solvent density (g/l)	1308.8	1318.8	1723.76
Total atoms	99516	99351	170995
Speed (hr/ns)	0.192	0.164	0.455
Relative speed (ns/day)	125.240	146.453	52.695
Machine	HECTOR	ARCHER	HECTOR

Table 4.3: Overview of the simulated systems.

4.3.1 System stability and conformational flexibility

Three independent simulations were run. The apo state of a protein is a representation of the native state containing the necessary information to perform the natural functions. Two simulations were run for the monomer and one for the tetramer. The holo monomer was modelled from the apo monomer which was possibly present in two conformations. One of the conformations mimics the ATP bound molecule. Table 4.3 provides an overview of the system size.

Energy Terms and Temperature

The stability of the simulation was first analysed in terms of the energy and temperature values. The energy is evaluated in terms of kinetic, potential and total energy of the system from the trajectory. Figure 4.7 shows that all the energies remained constant throughout the simulation time indicating the correct working of the thermostats. The stability of potential energy also demonstrates that the system relaxed and reached equilibrium. As the total energy comprises of the sum of kinetic and potential energy, it also shows conservation throughout the simulation time. The temperature for all the three systems remained around 318 K throughout the simulation yet again reinforcing the correct working of the Berendsen thermostat applied (Figure 4.8).

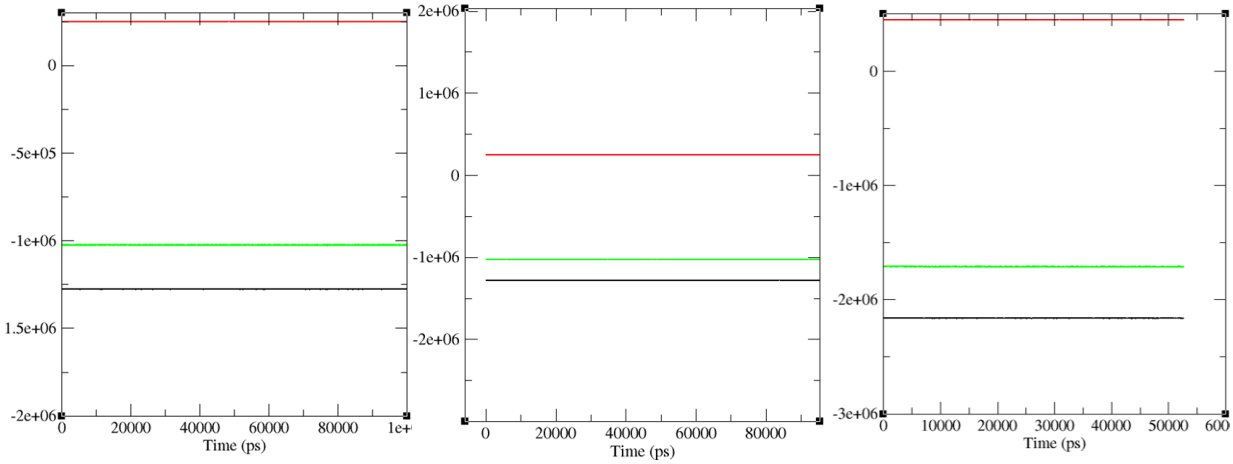


Figure 4.7: Left to Right: The Energy terms for Apo monomer, Holo monomer and Holo tetramer simulations. Potential energy is in black, kinetic energy is in red and green line represents the total energy. X-axis denotes the time in picoseconds and Y-axis denotes the energy terms in KJ/mol.

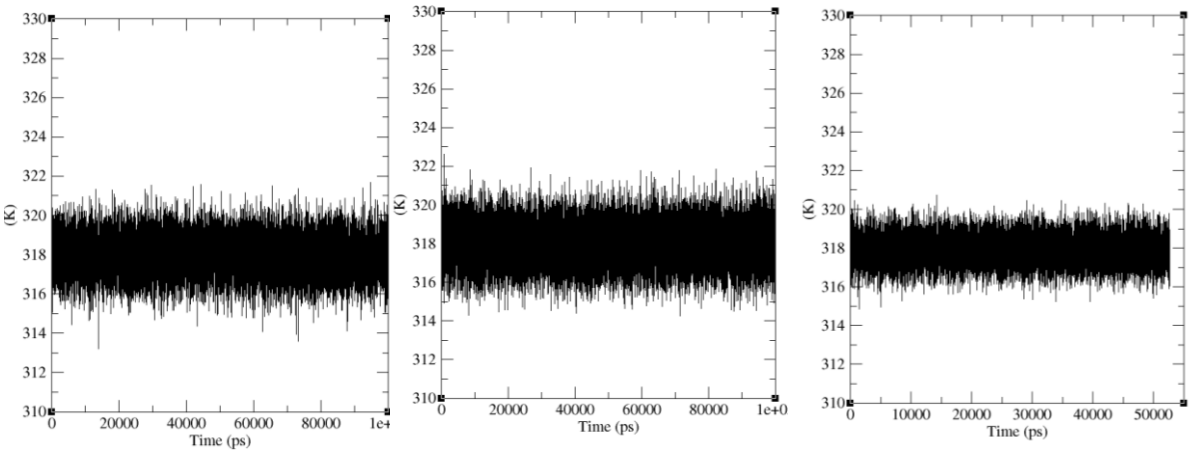


Figure 4.8 : Left to Right: Monitoring the temperature for Apo monomer, Holo monomer and Holo tetramer simulations. X-axis is the time in picoseconds and Y-axis represents the Temperature in Kelvin

4.3.2 ROOT MEAN SQUARE DEVIATION

The stability of the simulation can further be evaluated by monitoring the root mean square deviation of the atoms from the starting structure. We have examined the RMSD of the Cα atoms for all three simulations from their starting structure and plotted it versus time. Figures 4.9-11 show the results of the RMSD calculation. There is an initial rapid rise in the RMSD for the first 20ns which can be accounted for the system to be reaching convergence and relaxing within the solvent. Thereafter, the simulations remained stable without any major drift. The apo monomer converged around an average of 6.9 Å, holo monomer around 6.24 Å and holo tetramer around 2.67 Å. Both the graphs show that the simulations have reached a stable and equilibrated dynamical state. After the initial rapid fluctuations, the simulations remained stable around an average of under 7 Å for the monomers and under 3 Å for the tetramer. From the monomer graph, it is also evident that the conformational change experienced by the holo monomer is less as compared to the apo monomer by nearly 0.7 Å.

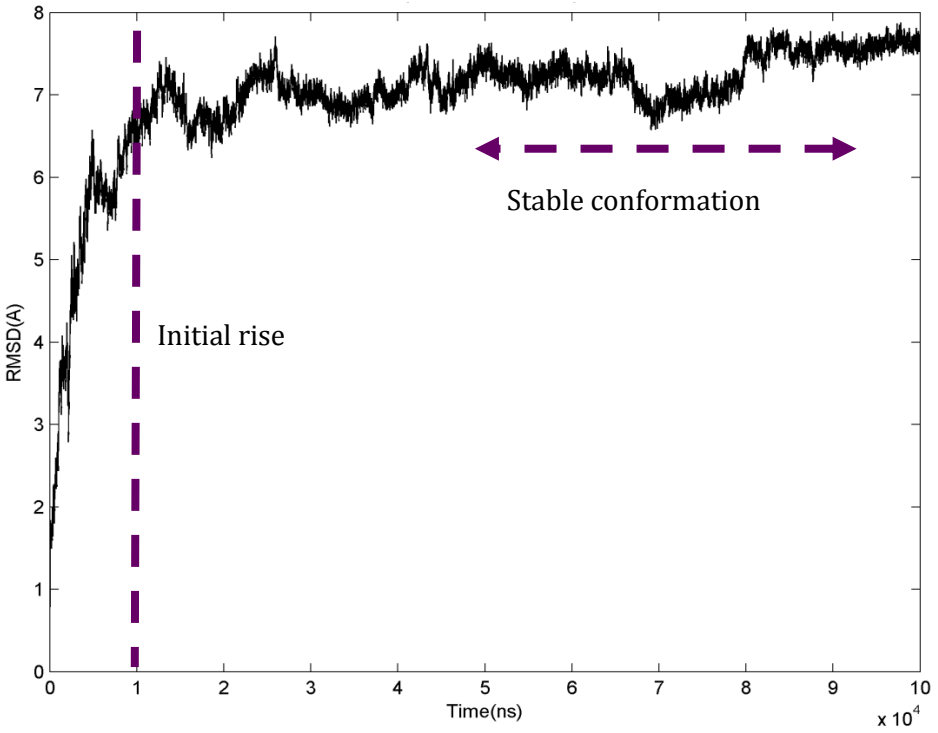


Figure 4.9: RMSD of Apo Monomer. There is a sudden rise in the root mean square deviation upto 8 Å. However, the monomer converged to a stable conformation at 6.9 Å.

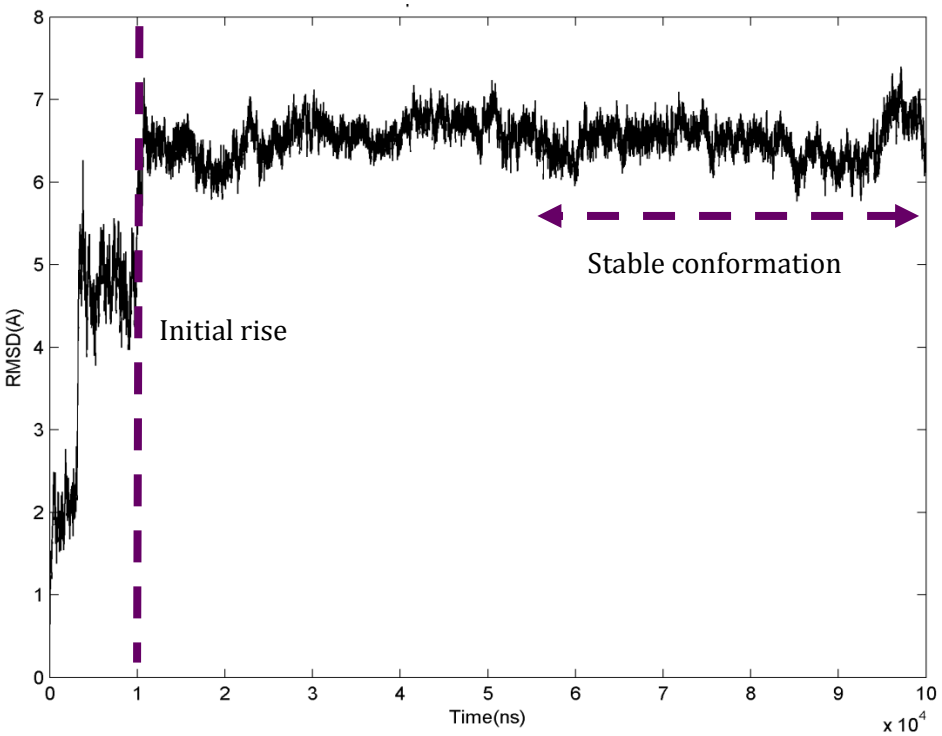


Figure 4.10: RMSD of Holo Monomer. The holo monomer in contrast to the apo monomer shows two sudden peaks of deviation at 2 Å and 7 Å. However, it converges at 6.24 Å.

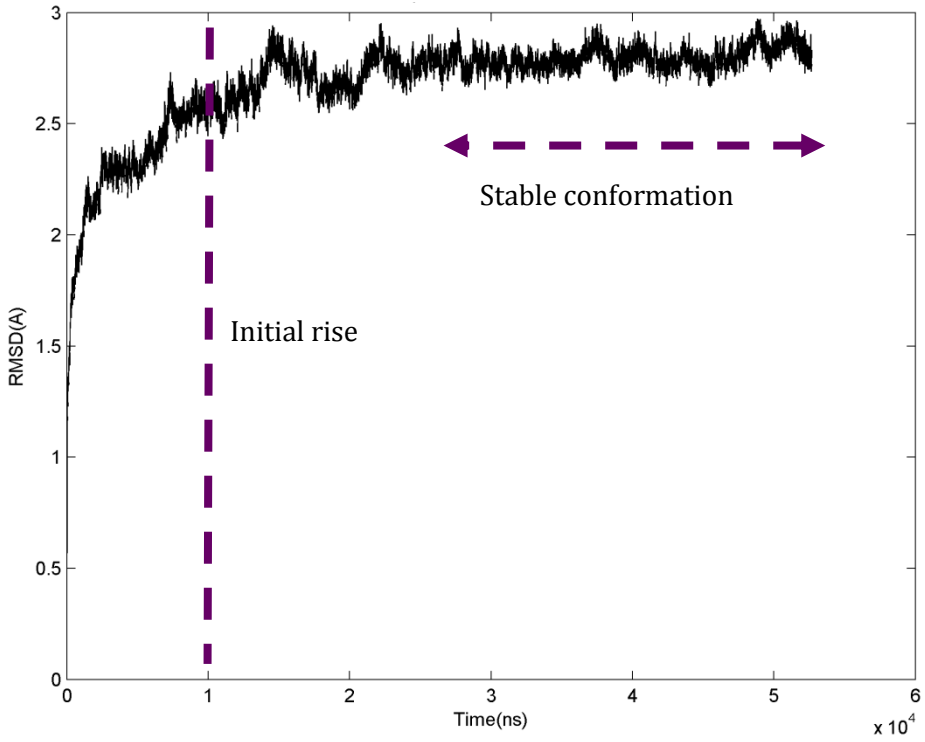


Figure 4.11: RMSD of Holo Tetramer. The tetramer in contrast to the monomers shows the deviation at a lower amplitude of 2 Å with the mean deviation being around 2.4 Å.

AVERAGE STRUCTURES:

After analysing the RMSD of the simulated systems, we calculated the average structures of each trajectory and compared it with the corresponding starting structure. This average structure was evaluated for the equilibrated part of the trajectory which provided an estimate of the overall structural and conformational changes in the apo and holo structures. For the monomers, the first 20 ns were excluded as part of the equilibration and the last 80 ns of the trajectory was taken to compute the average structure (Figure 4.12). In case of the tetramer, the first 15ns were excluded from the average structure.

Figure 4.13 shows the comparison between the average structure and the starting structure used for the simulation. The extended C-terminal for both the monomers curls up during the simulation. Also, the average structure of apo monomer becomes more compact with the curling up of C-terminal and A-domain. There are clear conformational changes in the apo structure in the embracing arm

and the reaching arm and the inserted loop regions. The holo monomer on the other hand shows less conformational changes from the original crystal structure. It maintains its open conformation despite the folding up of the reaching arm towards the embracing arm. There is a noticeable reduction in the distance between the N-terminal and the C-terminal in case of apo monomer. The same is not observed for the holo monomer with bound ATP molecule. The inserted loop (in domain C, residues 329-348) also doesn't show much deviation from the starting structure. The tip of the loops approach each other which is in agreement with the observations for holo tetramer.

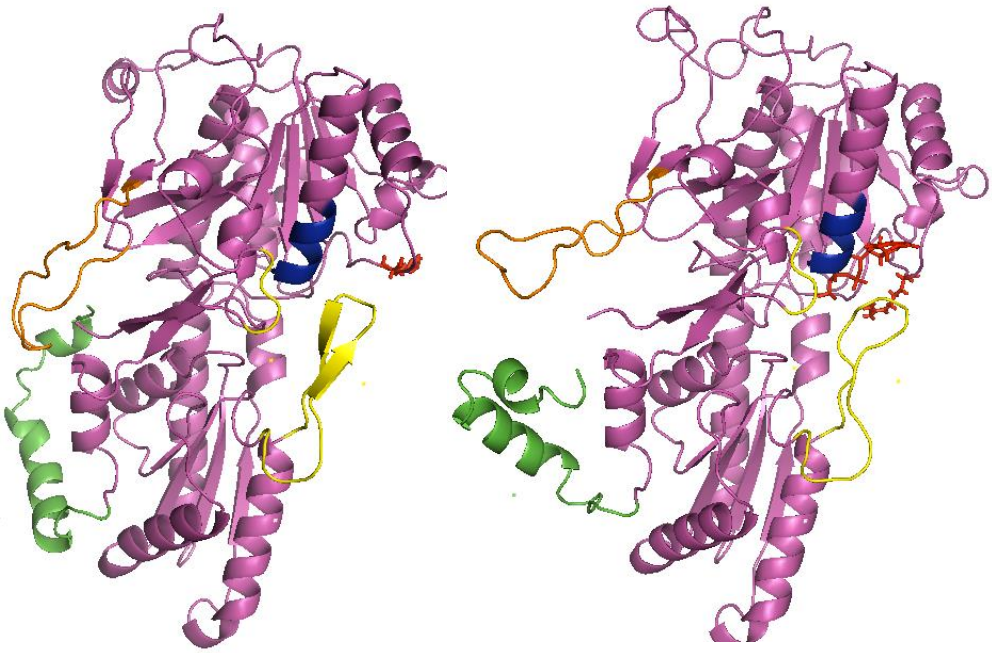


Figure 4.12 : Average structure for the monomer simulations computed from the equilibrated part of the trajectory i.e. last 80ns. Left is the apo monomer and right is the holo monomer. The reaching arm is shown in green;residues 454-485, embracing arm in orange (residues 62-81), insertion loops in yellow (residues 17-20 & 329-348) and the active site region is in blue (residues 199-204) ATP is shown as sticks in the holo monomer and also shown is arginine 173 in sticks which forms a part of the active site region.

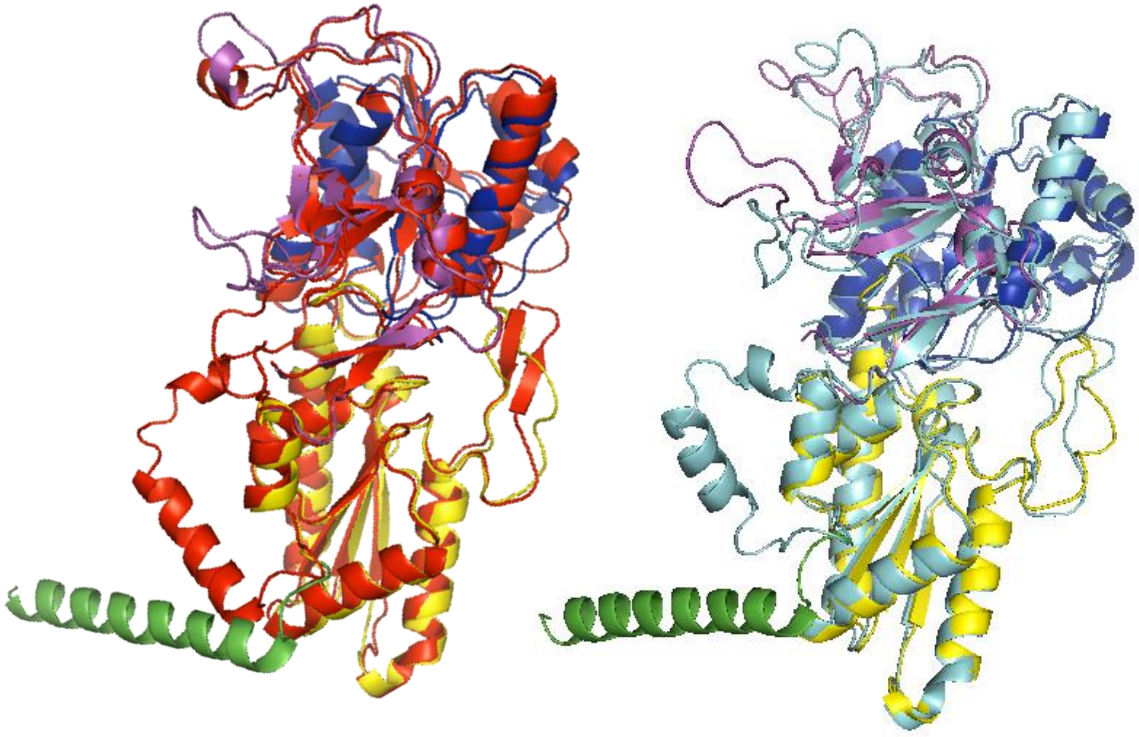


Figure 4.13 : Comparison between the starting and average structure of the two monomers. The starting structure is colored according to the domains as described previously and the apo monomer average structure is in red while the average holo monomer structure is in cyan.

Fig. 4.14 shows the comparison between the average structure (red) and the starting crystal structure for the holo tetramer. There is no disrupting conformational change in the structure besides a slightly different orientation of the reaching arm. The insertion loops and the embracing arm maintain their position. This is also evident in the RMSD graph where we observe a sudden rise which might be due to the fluctuation in C-terminal residues but this rise soon levels off and maintains equilibrium.

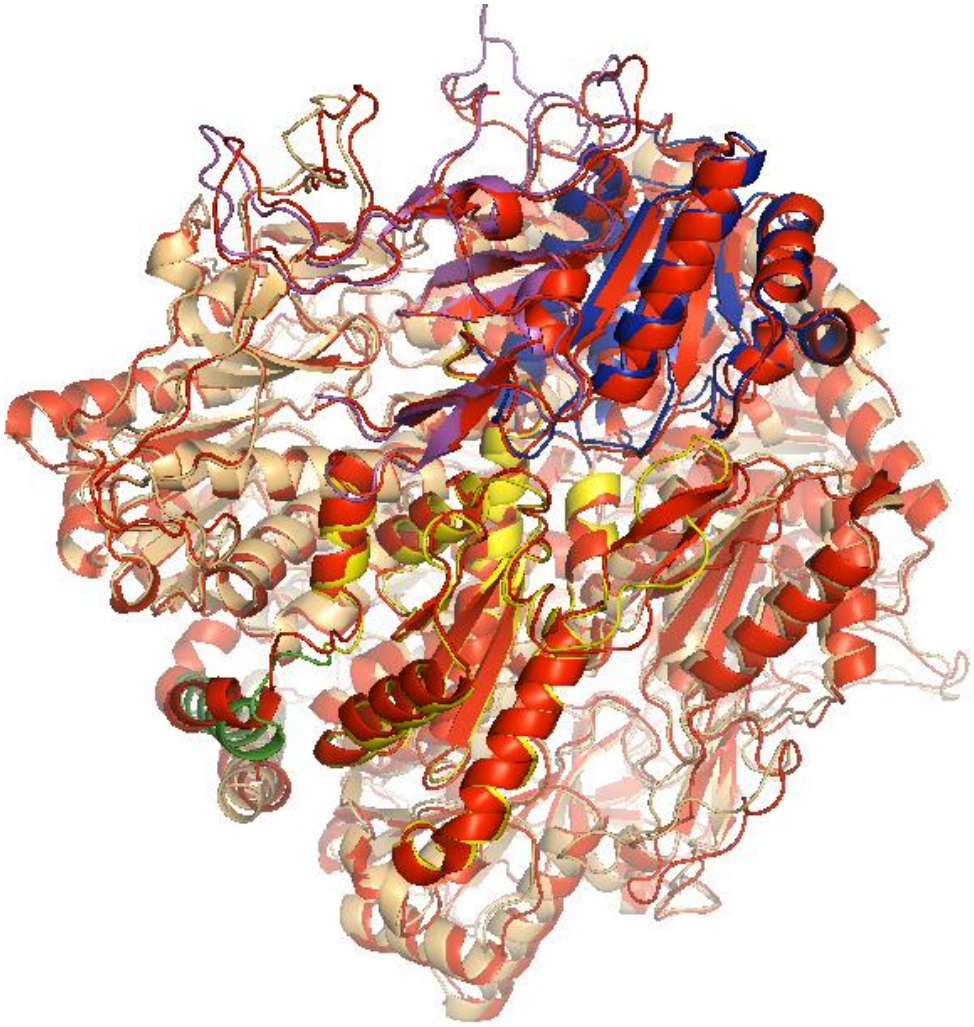


Figure 4.14: Comparison between the starting and average structure of Holo tetramer. The starting structure is colored according to the domains and one single subunit is highlighted in the foreground. The average structure is superimposed in red.

4.3.3 ROOT MEAN SQUARE FLUCTUATION

Thermal fluctuations of residues help to understand the protein function by local conformational flexibility. In order to investigate and explore the conformational variability of each trajectory, the root mean square fluctuation

(RMSF) of alpha carbon atoms was plotted with respect to the residue number to show the local conformational changes for all the three systems.

The root-mean square fluctuations (RMSF) of apo residues compared to the holo are shown (Figures 4.15-17). The average residual fluctuation values of apo monomer, holo monomer and holo tetramer are 1.23, 1.45, 1.02 Å respectively. Although the overall differences between the apo and holo structure are small, it stull indicates that binding of the ligand as a restraining effect on the overall movement of the protein.

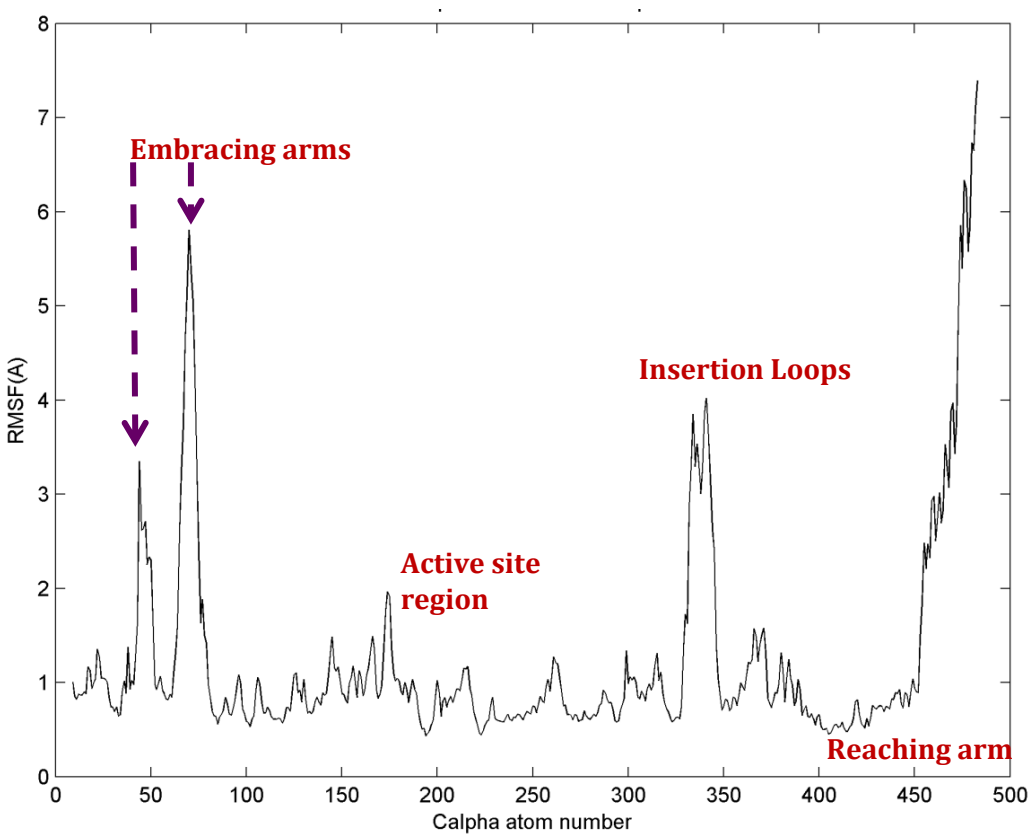


Figure 4.15: RMSF of Apo Monomer

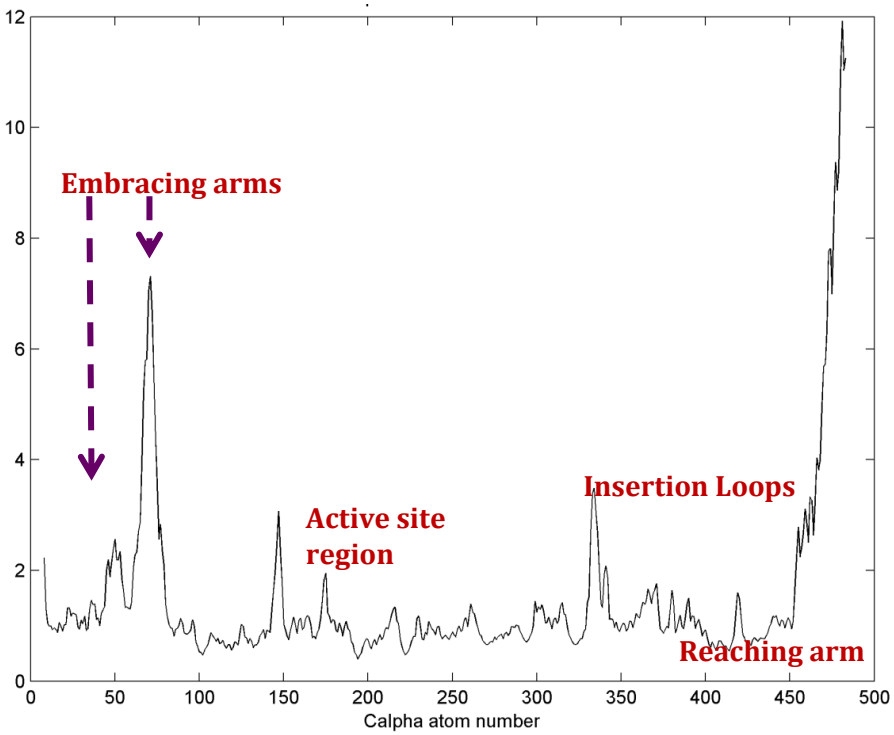


Figure 4.16: RMSF of Holo Monomer

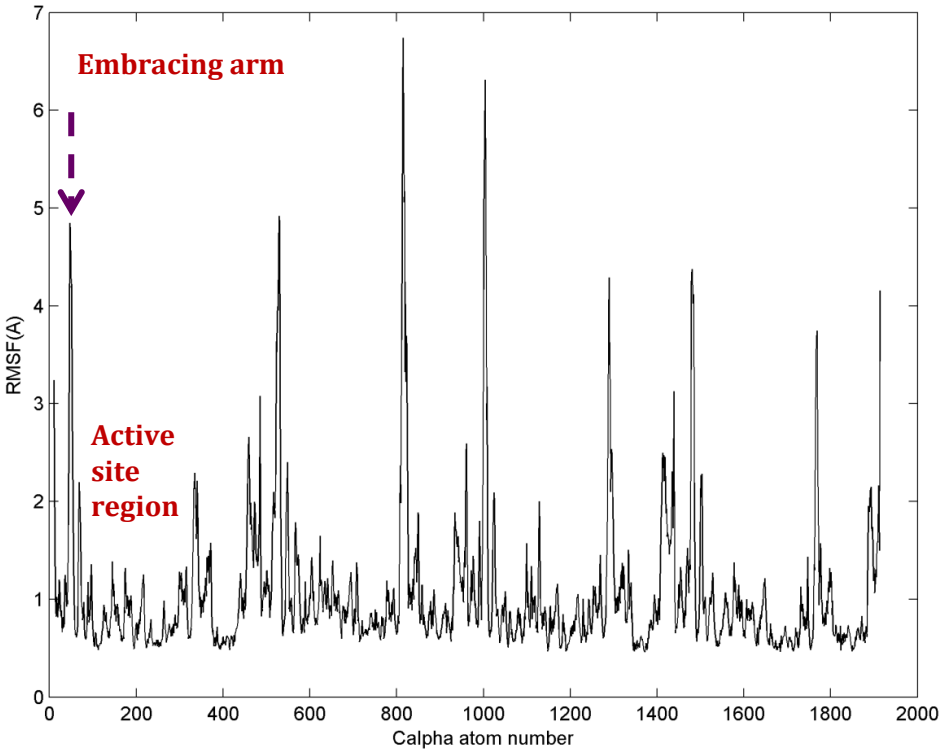


Figure 4.17: RMSF of Holo Tetramer

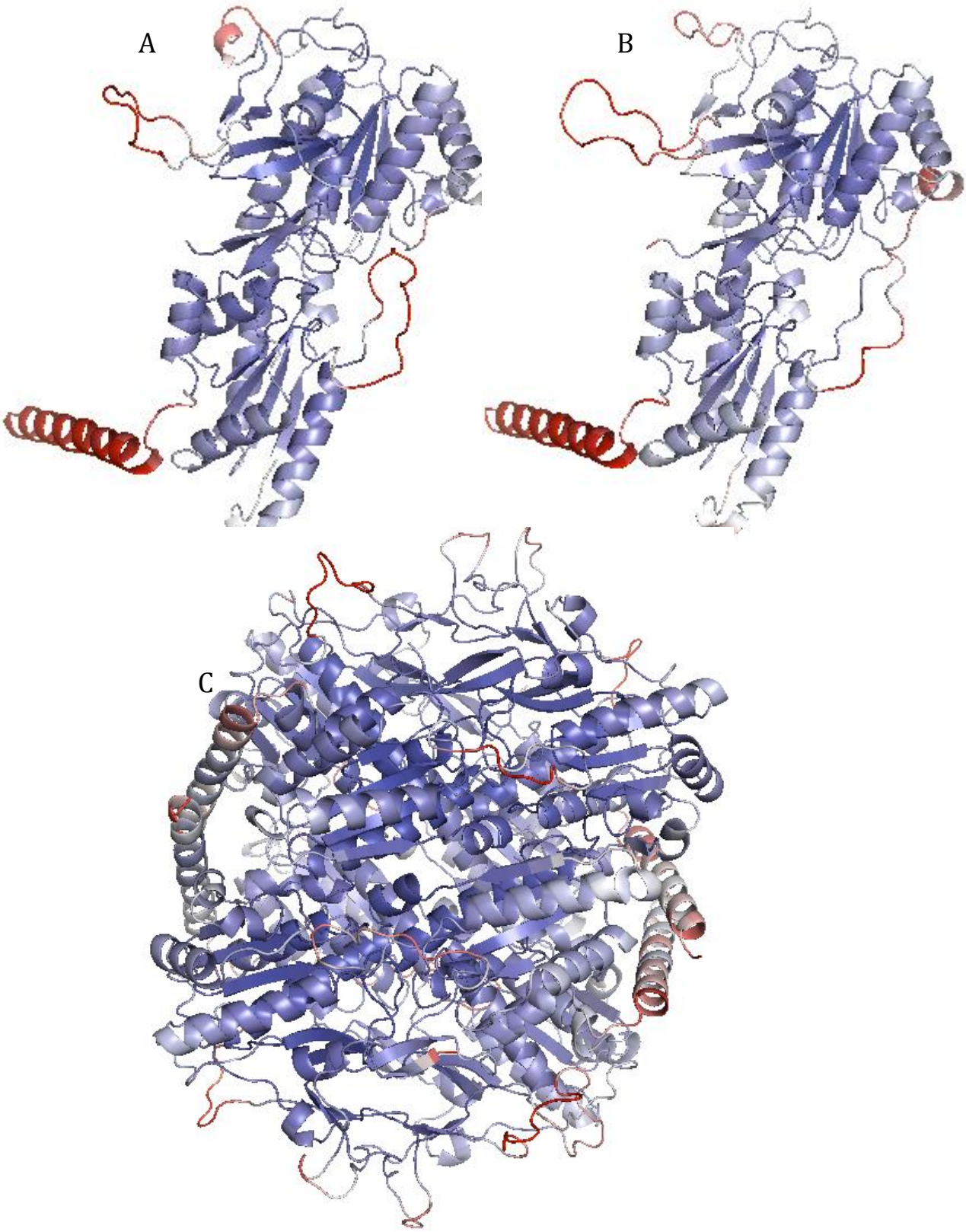


Figure 4.18: Color coded representation of the root mean square fluctuation values of the simulated systems of phosphofructokinase. (A) Represents the fluctuations of apo monomer state; (B) Holo Monomer and (C) Holo tetramer. The highest fluctuations are colored red while the cool regions of the protein are colored blue. The RMSF values have been normalized between 0-3 Angstrom for color coding.

Residues	Apo Monomer	Holo Monomer	Holo Tetramer	Holo Tetramer			
				A	B	C	D
Embracing arm (62-81)	2.88	3.89	1.166	1.11	1.23	1.11	1.20
Reaching arm (454-485)	4.13	5.70	1.75	1.79	1.46	1.96	1.74
Insertion loop (17-20 and 329-348)	1.05;2.71	0.99;1.80	1.04;2.63	0.92;1.59	1.1;3.65	1.1;2.28	1.02;1.90
Active site (R173, 199-204)	1.75;0.83	1.47;0.69	1.12;0.74	0.85;0.68	1.34;0.86	1.46;0.78	0.83;0.66

Table 4.4: Root Mean Square Fluctuations for the different loops of phosphofructokinase. The data is in units of Å.

Phosphofructokinase shows flexibility around residues 50-100, 150, 330-350(329-348 inserted loop) and the regions involving the C-domains (420-430). In order to better understand the overall fluctuation profiles of the two different systems i.e. apo and holo forms of phosphofructokinase, the fluctuation values were converted to a colour code and then structurally mapped onto the corresponding structure of each state (Figure 4.18). Inspection of these figures revealed that the regions of the protein with highest fluctuation values corresponded mainly to the insertion loops, active site regions where ATP binds, embracing and reaching arms (Figure 4.19 & Table 4.4).

We observe that the fluctuations of these regions are very high for the apo monomer but upon inclusion of the ligand i.e. ATP, these values are reduced considerably. For instance, the insertion loops which hold ATP together show marked reduction in fluctuation for the holo monomer (1.80 Å) as opposed to the apo monomer where the value is around 2.7 Å. Same is the case in various subunits of the holo tetramer which shows reduction in the fluctuation. This attenuates the fact that upon binding of ATP, these loops adopt a relatively closed conformation

and are stabilized by the interactions between the neighbouring residues. Also, we see the loops approach each other.

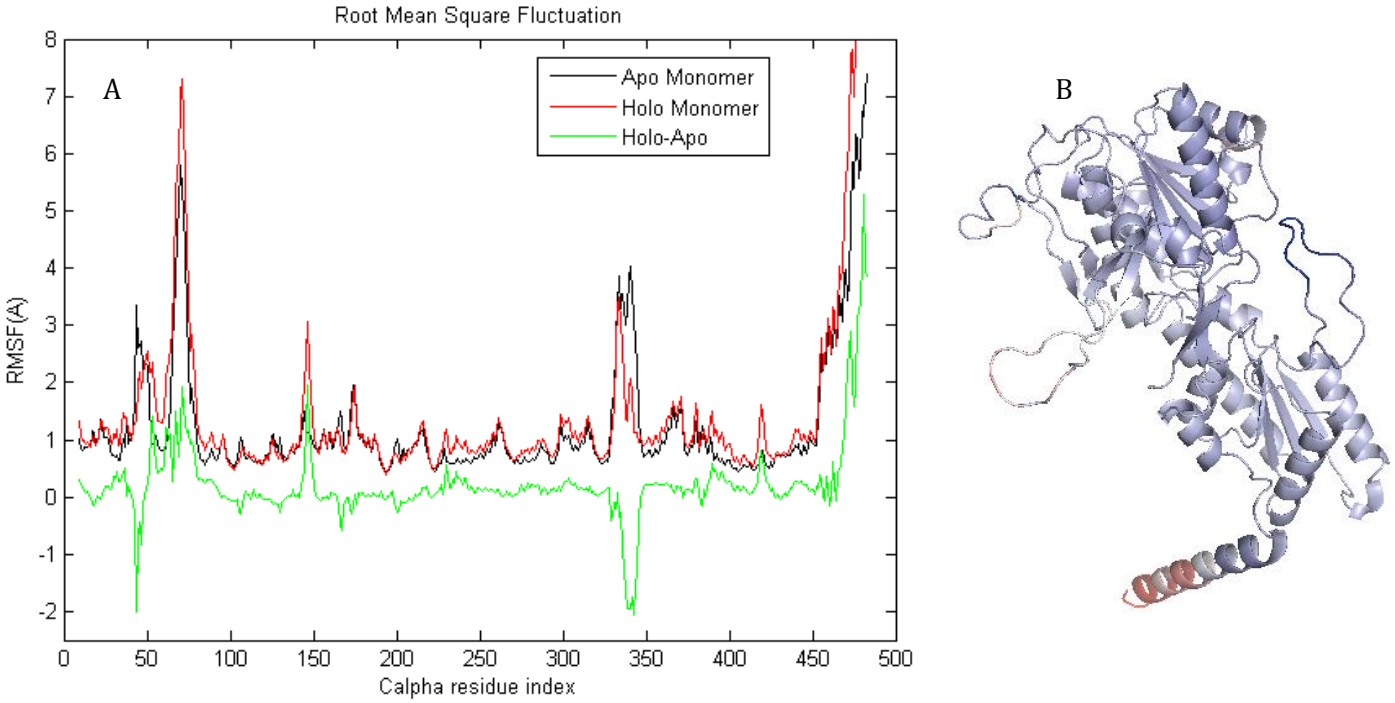


Figure 4.19: Root mean square fluctuation difference between the two different states of phosphofructokinase. (A) Graph obtained by subtraction of rmsf value of Apo Monomer (black line) from the holo monomer (red line). It revealed the regions which show reduction in the fluctuation values upon binding of ATP (green line) in the holo monomer.

(B) These regions (green lines) were then mapped upon the structure of phosphofructokinase which provided a visual inspection of the regions of the protein after stabilisation. The structure is normalized between -2 to 4 Å and the color ranges from blue to red.

4.3.4 PRINCIPAL COMPONENT ANALYSIS

In order to analyse the motions and structural conformations of phosphofructokinase, Principal Component Analysis was performed. As stated previously (refer materials and methods), PCA is one of the effective techniques to analyse the enormous data from MD simulations. Apart from reduction of the dataset to a set of few informative modes, it provides a qualitative picture of the

collective motions relevant to protein structure and is thereby helpful in analysing the transitions and the important changes brought about by insertion of ligands.

Typically, three main steps are involved in any principal component analysis. The first step consists of filtering the internal motions from overall rotation and translation by superposition of the configurations from the ensemble. (i.e. least-square-fitting of each of the configurations onto a reference structure). Second, this "fitted" trajectory is used to construct a variance–covariance matrix that is subsequently diagonalized. The off-diagonal elements of this symmetric matrix represent the co- variances of the atomic displacements relative to their respective averages for each pair of atoms while the diagonal represents the variances of each atomic displacements. Concerted motions of atoms give rise to positive covariance, whereas non-concerted movements give rise to negative covariances. Non-correlated movements give rise to zero covariances. Upon diagonalization, the matrix yields a set of eigenvalues and eigenvectors arranged in decreasing order of the values. These eigenvalues are a representation of the variances along the corresponding eigenvector (or mode). Finally these Principal components can be used to analyse the original trajectory. This projection further yields the time behaviour and distribution of each of the principal coordinates. We sampled in our principal component analysis only the positions of the 475 Cα carbons for apo monomer, 476 Cα carbons for Holo monomer and 1905 Cα carbons for the Holo tetramer, thereby resulting in 3N = 1425, 1428 and 5715 degrees of freedom, respectively.

One of the important measurements from covariance matrix is a trace and it is a sum of the eigenvalues. This sum can be used to describe total motility of the system. The values of the trace of covariance matrix for the three systems i.e. Apo monomer, Holo monomer and Holo tetramer is 13.36, 21.86 and 29.27 nm^2, respectively. This clearly demonstrates the influence of the ligands on the motional strength of the protein. The motional strength of the protein is more stable and constricted in Holo monomer as observed previously in the mean square fluctuation graphs and now has been corroborated with the trace of the matrix obtained after principal component analysis.

One of the principal issues in principal component analysis is deciding where the principal components cease to become significant, and there are well established tests some rules for excluding principal components [refer materials and methods]. One of the most obvious is to include just enough components to explain 90% of total motility. A second method called Kaiser's criterion excludes those PC whose eigenvalues are less than average, i.e. less than one if a correlation matrix has been used. Most often, the screening tests are used to determine the number of eigenmodes/eigenvectors/principal components to focus on for further analysis. It emphasizes finding a region where the smooth decrease of eigenvalues appears to level off to the right of the eigenvalue plot. Following this criterion, the first 10 eigenvectors have been selected in all the three systems. Figure 4.21 presents the percentage and cumulative percentage of variance explained by the first 10 eigenvectors from a total of 1425, 1428 and 5715 eigenvectors for the apo monomer, holo monomer and holo tetramer, respectively. The plot is arranged in a decreasing order with respect to the corresponding eigenvector indices for all the three simulations. It is seen from this plot that the first 10 eigenvectors describe approximately 70-80% of total variation of the system. The following section describes the dominant modes of motion for all the three simulated systems of phosphofructokinase.

From the eigenvalue figure, one can see that only first few values correspond to the concerted motions which quickly decrease in the amplitude to reach more constrained and localized fluctuations. Also, comparing the plot we can see that the properties of the motions as described by the first few eigenvectors i.e. principal components are not the same. Such is the case between Apo and Holo monomer, where the magnitude of eigenvalues are higher for the Holo monomer and tetramer. The first principal component i.e. eigenvector in the Apo monomer represents 33.5 % of the total variance, in the Holo monomer it represents 45.1 % whereas the in Holo tetramer it represents only 32.6 % of the total variance i.e. total motility in the simulation set. The contribution of the data set used for further eigenvector analysis i.e. the first 10 eigenvectors is quite different amongst three simulations. For the apo monomer, the first 10

components contribute to a total variance of 76.95% while for the holo monomer and holo tetramer, the contribution is 86.08% and 62.81%, respectively. As is quite evident from the cumulative percentage table (Table 4.5), the data is quite dispersed amongst the eigenvectors i.e. the first 3 eigenvectors together show around 50% motions as opposed to the general trend of the first mode being sufficient to show more than half of the variance. This suggests that the structural arrangements and the conformational transitions in phosphofructokinase also follow a trend similar to pyruvate kinase, where the data is dispersed amongst a set of modes with distinct motional representation.

Mode	Apo Monomer		Holo Monomer		Holo Tetramer	
	Eigenvalue (nm^2)	Cumulative Percentage	Eigenvalue (nm^2)	Cumulative Percentage	Eigenvalue (nm^2)	Cumulative Percentage
1	4.477	33.508	9.860	45.109	9.567	32.686
2	2.104	49.257	2.734	57.618	2.635	41.689
3	0.923	56.172	2.128	67.353	1.536	46.938
4	0.690	61.34	1.534	74.374	1.243	51.186
5	0.628	66.043	0.769	77.894	0.889	54.225
6	0.457	69.469	0.571	80.509	0.683	56.561
7	0.317	71.849	0.418	82.425	0.559	58.471
8	0.285	73.988	0.353	84.040	0.482	60.121
9	0.217	75.612	0.254	85.206	0.405	61.506
10	0.178	76.949	0.191	86.081	0.382	62.813

Table 4.5: Eigenvalues and cumulative percentage for the first 10 principal components (eigenvectors/modes) of Apo monomer, Holo monomer and Holo tetramer simulation.

In order to have a closer look at the motion along the eigenvector directions, one can project the trajectory onto these individual eigenvectors. A two or three dimensional projection along the major principal components gives a representation of the sampled distribution in configuration space. This in turn also helps to compare multiple ensembles along the principal modes of collective

fluctuations. Therefore, we inspected the progression of the reaction coordinate, also called the projection, over time.

Figure 4.21 is a plot of the trajectories which have been projected on the first ten eigenvectors. These plots reflect the degree of anharmonicity of the motions and also show that the probability distributions for the first few principal components are non-Gaussian for all the three systems. Although, the single essential eigenvectors are likely to be not the equilibrium ones, i.e., the eigenvectors obtained by completely converged statistics, the corresponding atomic collective motions are probably significant as they belong to a good approximation of the equilibrium essential subspace. We projected histograms of the probability distributions for the first four eigenvector (Figure 4.22). These histograms are plots that show the distribution of data. The overall range of a given set of data points is divided to smaller subranges (bins), and the histogram shows how many data points are in each bin. The height of the bar corresponds to the number of data points in the bin. These probability plots give an idea about the degree of anharmonicities of the motions and it is clear that the first 3 eigenvectors are not Gaussian. However, as we move down the spectrum they become more Gaussian or harmonic in nature.

For the holo tetramer, there are large amplitudes and slow motion of essential coordinates[280] which have been characterized by a slow diffusive kinetics[200]. The systems thus show characteristic multiple-minima protein energy landscape features[211, 281, 282]. As the relaxation and convergence for the essential coordinates are typically beyond the nanosecond timescales, these slow diffusions provide perhaps only partial sampling of the subspace defined by the first 2-3 eigenvectors. However, the convergence of the essential subspace (usually the first 10-20 eigenvectors) is reasonably achieved within a few nanoseconds[156]. Convergence is almost achieved for the holo tetramer simulation, which clearly shows the presence of MgATP affecting the principal components. It has already been shown that a reliable and statistically significant description of essential subspace on the nanosecond timescale can be shown during the simulation[200]. Figure 4.21 further depicts the Cα residue contributions to the first 10 principal components.

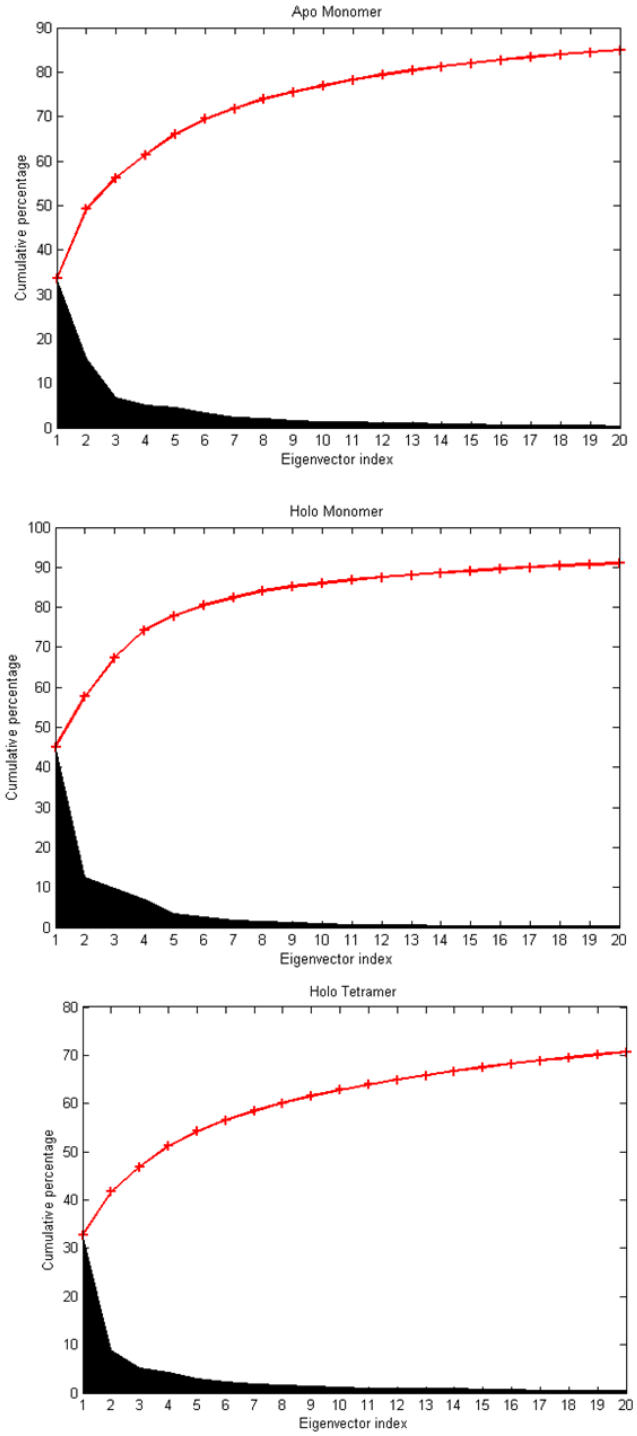


Figure 4.20: Percentage (black) and cumulative percentage (red) of variance for first 20 Eigenvectors for Apo Monomer, Holo Monomer and Holo tetramer simulation.

Also, from these projections and histograms it is observed that the first four eigenvectors show high projections and then converge quickly. The protein seems to sample multiple configuration space as can be seen from the histograms but from 5th

Principal component onwards, we observe a Gaussian behaviour. That is, the positional fluctuations are concentrated in correlated motions in a subspace of only a few degrees of freedom (<10% of the original configurational space), while the other degrees of freedom represent small, independent Gaussian fluctuations. In case of the monomer simulation, we observe that the monomers tend to sample two conformations. But, the holo tetramer seems to sample multiple small conformations i.e. show multiple energy peaks. However, there does seem to exist two clear conformations for the fourth eigenvector after which the components approach a Gaussian behaviour.

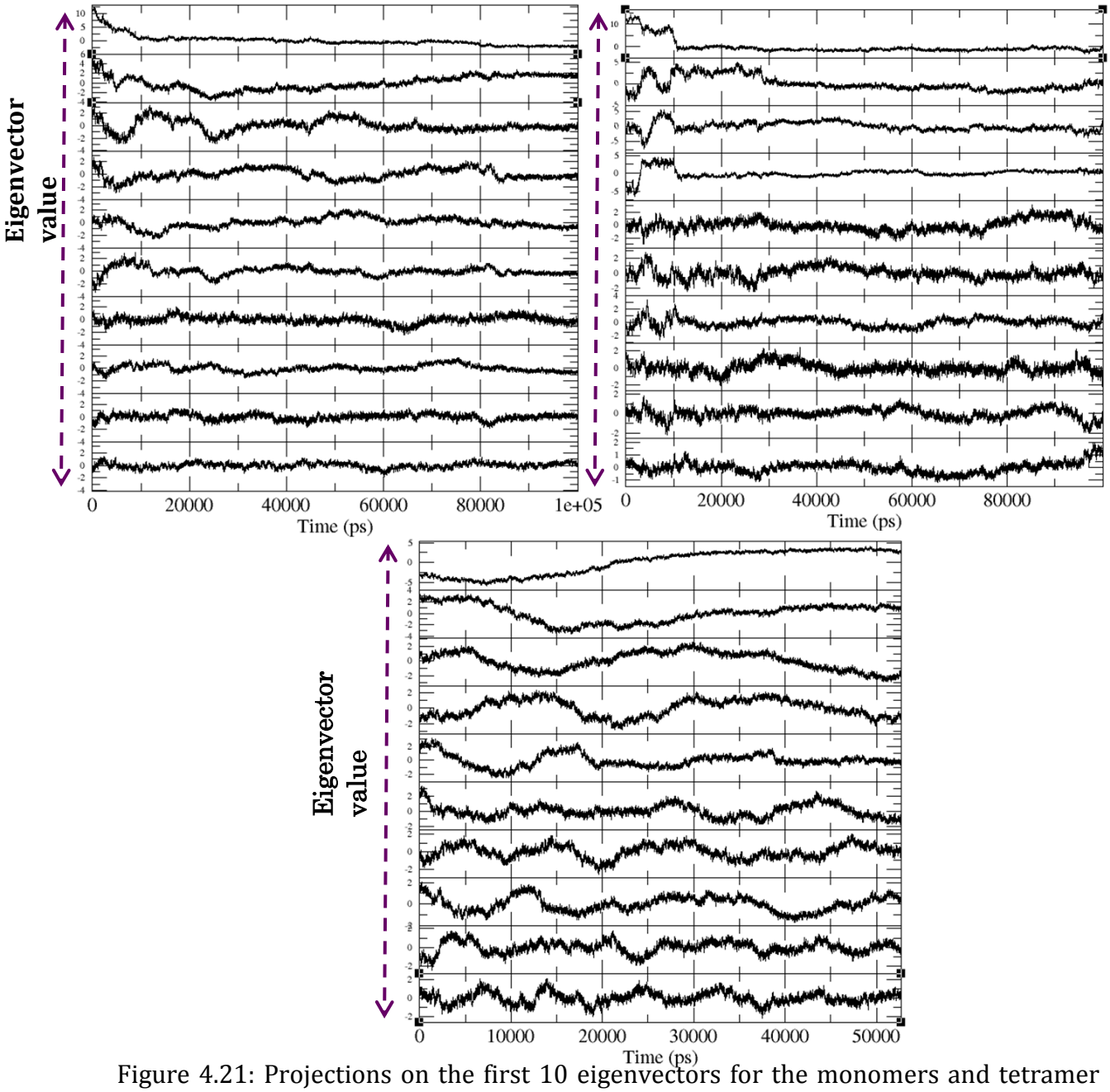


Figure 4.21: Projections on the first 10 eigenvectors for the monomers and tetramer along 100 ns and 52 ns of simulation time, respectively. X axis represents the time period in picosecond and Y axis depicts the projections in nm for each eigenvector. From top left; Apo monomer, Holo monomer and holo tetramer projections.

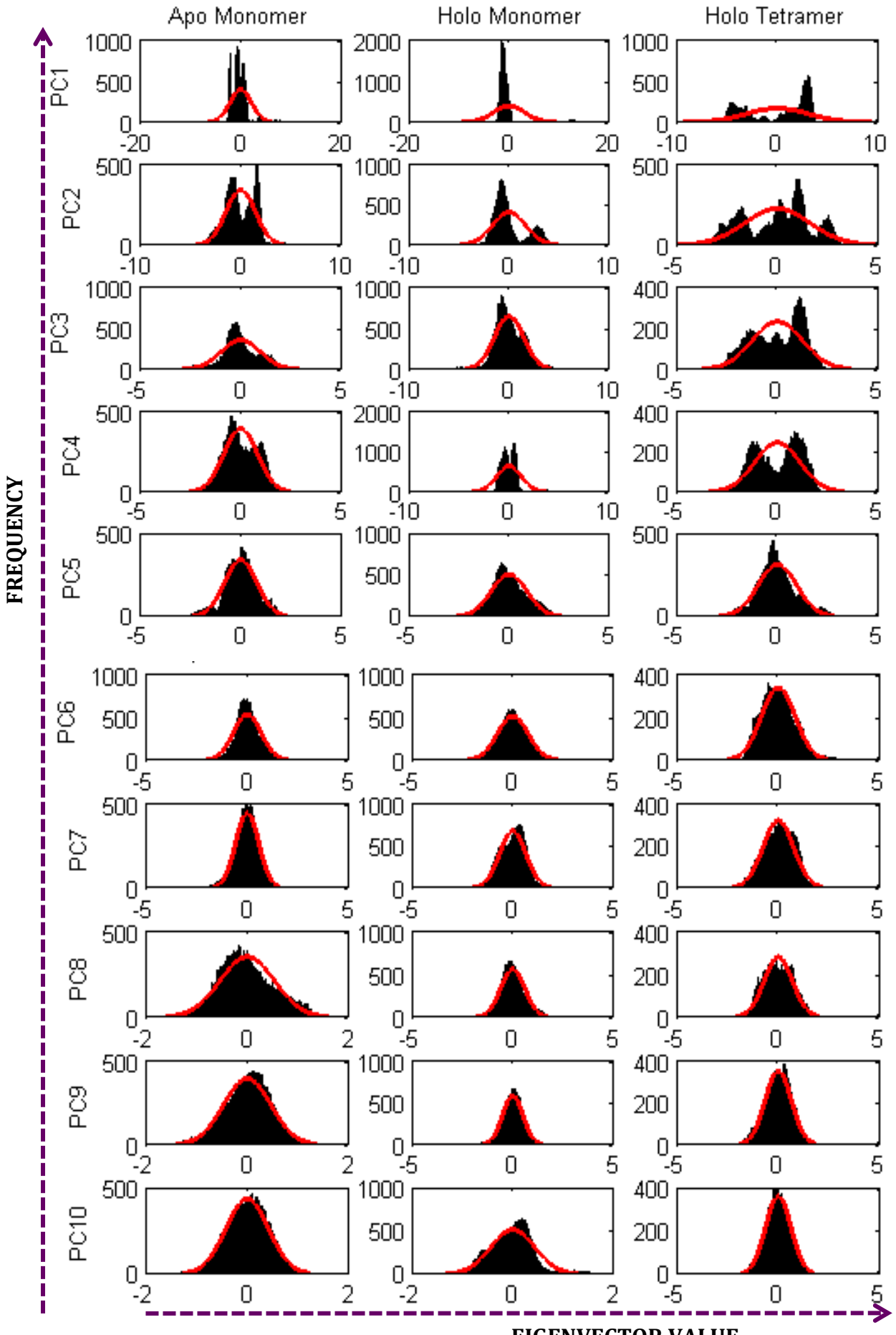

Apo Monomer
Holo Monomer
Holo Tetramer
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
FREQUENCY
EIGENVECTOR VALUE

Figure 4.22: Probability distributions for the displacements along the first 10 eigenvectors obtained from the Cα coordinates; (L-R) A: Apo Monomer, B: Holo Monomer, C: Holo tetramer. The red lines depict the data fitted into the Gaussian curve i.e. Gaussian distributions derived from the eigenvalues of the corresponding eigenvectors. The *x*-axis reflects the range of values in data. The *y*-axis ranges from 0 to the greatest number of elements deposited in any bin.

In order to monitor the motions described by the various eigenvectors, residue displacements were calculated. For a given model, a rich picture emerges of the regions of specific displacement for each of the first few principal eigenvectors. This in turn highlights the specific displacement regions. The plots reveal a pattern amongst the subunits. It can be seen that some of the motions described by one eigenvector are also described by the second and third eigenvectors, (concerted motions) thus so on and so forth. The motions as described by the eigenvectors are repeated in a similar fashion thereby reinforcing the fact of concerted motions. The eigenvectors for the individual trajectories have been plotted as a function of time which shows how the simulation progresses towards the new state. Figure 4.23 shows that for the apo monomer, the large fluctuations along the first few eigenvectors are mainly seen in the insertion loops of residues 17-20, 329-348, embracing arm (62-81) and the reaching arm region (454-485). These fluctuations also arise owing to the flexibility nature of these regions. The collective fluctuations around the substrate-binding regions originate primarily from the fluctuations of the bound ATP substrate, resulting in the dynamic variation of the substrate-binding subsites/pockets which has been observed in the apoenzyme where the loops are in closed conformation and hence deters the binding of ATP.

Figure 4.24 shows the detailed displacements of residues for the first three principal components for all the three simulated systems. In comparison to the apoenzyme, the region involving the embracing arm and insertion loops show high fluctuations in the first and third principal components but these fluctuations are reduced in case of holenzyme. Also, for the apo monomer, PC2 shows reduced fluctuations in the embracing arm region but slightly high fluctuation in the active site region and also in the insertion loops as compared to PC1 and PC3. In the

holoenzyme, the amplitude of fluctuation for embracing arm region is low as compared to PC2 and PC3. Also, PC1 shows less fluctuation overall, but, PC2 and PC3 reveal more concerted motions with the amplitude signals arising in the substrate binding regions and the embracing arms which keep the tetramer together and stabilize the structure.

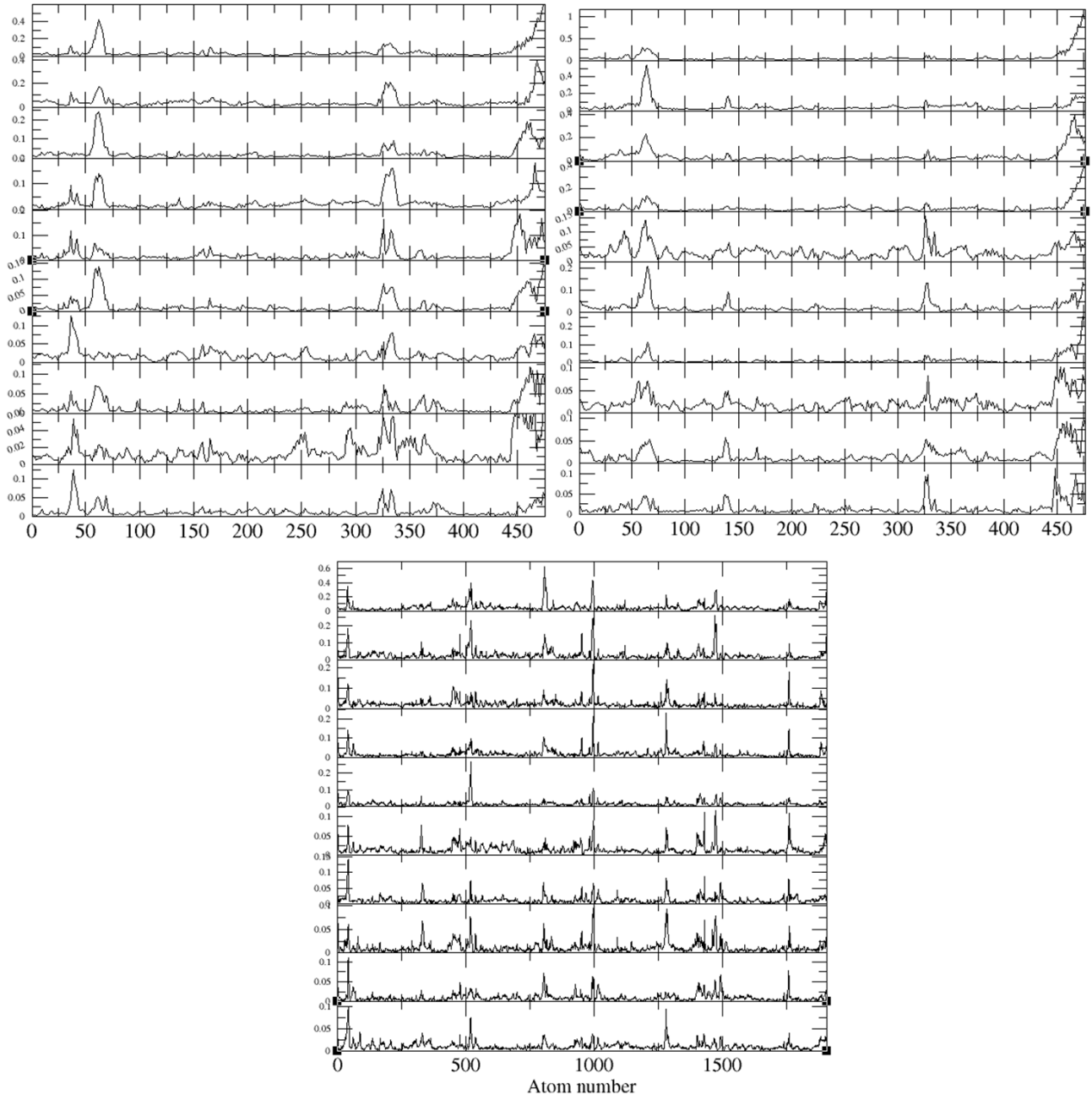


Figure 4.23: (Left to right): Residue displacements in the subspaces spanned by the first 10 eigenvectors for the monomers and tetramer along 100 ns and 52 ns of simulation time, respectively. X axis represents the calpha atom and Y axis depicts the rmsf in nm for each eigenvector. Displacements for Apo monomer, Holo monomer and holo tetramer simulation, respectively.

Furthermore, to probe the presence of conformational states and transitions between them, we calculated the two dimensional projections of the various eigenvector pairs. These plots provide a measure of the mobility of the protein in the essential subspace, thereby showing the clusters representative of explored tertiary conformation that differs amongst the three simulated structures. Figure 4.25 shows the trajectory projected on the first three planes, each defined by two all atom matrix eigenvectors. In the planes of eigenvectors 1 and 2 and eigenvectors 2 and 3, we see that the trajectories are confined within narrower ranges for PC1 and 2 and PC 1 and 3, suggesting the presence of a coupled force field. However, the projections between PC 2 and PC3, the trajectories occupy a wider range which suggests the presence of independent motions between PC2 and PC3.

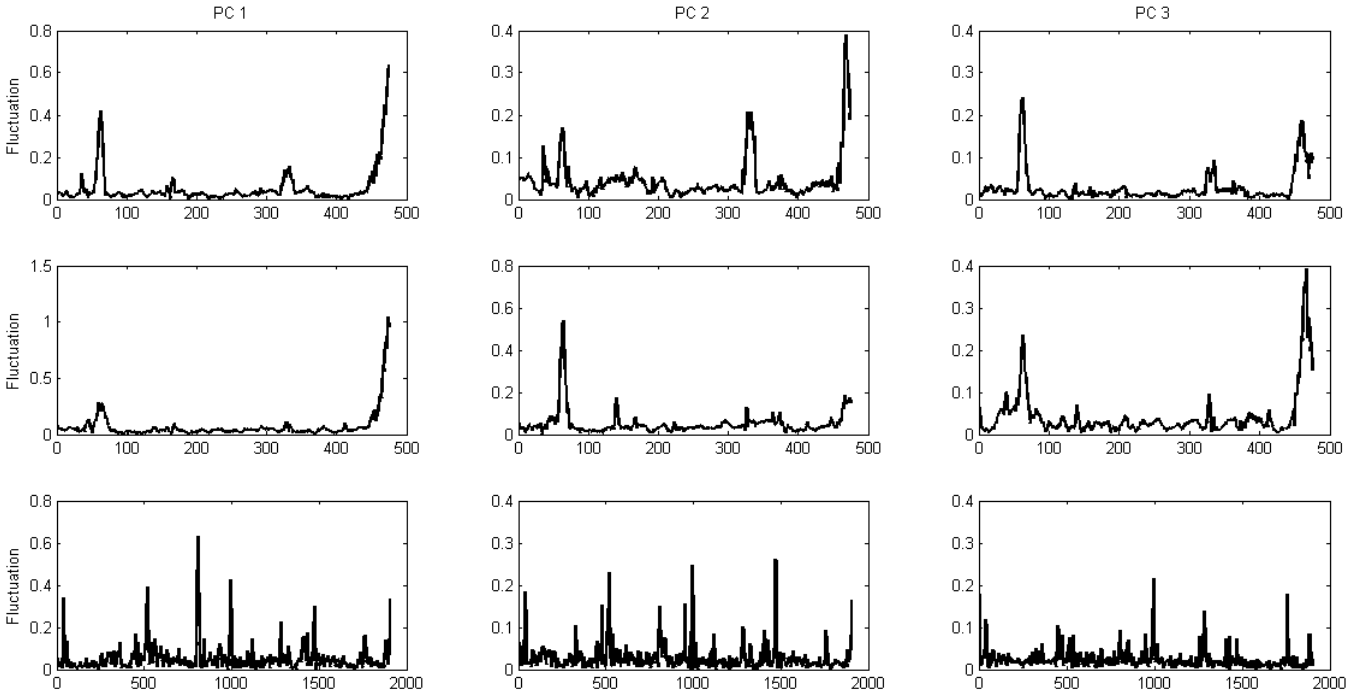


Figure 4.24: Residue displacements i.e. fluctuations of the first 3 Principal Components (PCs) in each dataset i.e. Apo Monomer (Top), Holo Monomer (middle) and Holo Tetramer (Bottom). X axis represents the calpha atom number and Y axis depicts the rmsf in nm for each eigenvector.

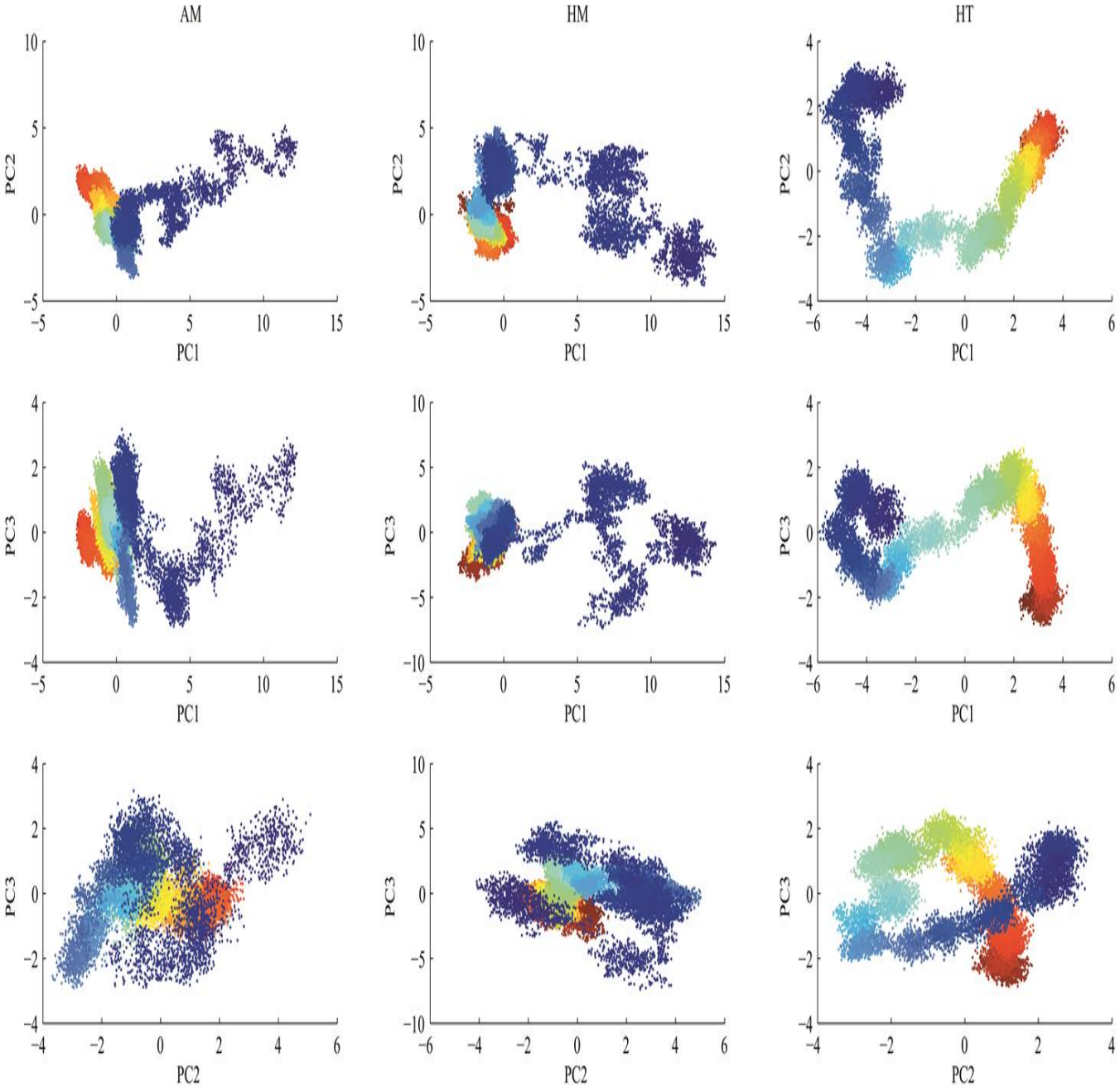


Figure 4.25: The 2-Dimensional projection of eigenvectors for all the three simulations. The data is plotted for the first three eigenvectors with the first column being for the Apo Monomer, second column for the Holo Monomer and the third column for Holo Tetramer. The projections are colored for the time period of the simulation ranging from blue to red, with blue being the starting point of the simulation and red denoting the end of simulation and thus the configurations towards the end of the simulation.

Besides projecting the displacements of the eigenvectors in a 2 dimensional structure we also probed the individual projections along the first three eigenvectors for the simulated systems. This provided us with information about the contribution of each residue to the first three principal componenets. For the the apo and holomonomer, the displacements were projected for the whole structure, but for the holo tetramer, we projected the data on only on one of the four subunits in order to maintain homogenity and compare with the behaviour of monomer. Also, in order to further aid our interpretation, we projected the trajectory by interpolating between the most dissimilar structures in the distribution along the first given principal component. This provided us with the projection along the first principal component for our structures. As is evident from the figures, all the three structures show marked fluctuations in 4 different regions as flanked by the highest peak(Figures 4.26 & 4.27). The highest fluctuation is seen in the end terminal region (resiudes beyond 450). This is quite reasonable though as we have the reaching arm positioned around the same region(residues 454-485). For the first eigenvector, we observe a similar rise in fluctuation in this region. However, in the subsequent eigenvectors, the fluctuations for holo monomer and holo tetramer are quite low in comparison to apo monomer. The second region of fluctuation is around the area flanked by residues 50-90. This also houses the embracing arm (residues 62-81). As can be observed for this region, the first eigenvector has sharp peaks but suqbsequently the fluctuation minimes. However, in comparison to the monomers, it should be noted that the tetramer shows minimal fluctuation in the embracing arm for the eigenvectors 2 & 3. This could be attributed to the fact that the movement of the embracing arm during substrate binding might be reflected upon by the first eigenvector and since it generates a stable conformation, we see the magnitude reduced for the subsequent eigenvectors. The third region to show marked fluctuation is the region involved between residues 320-350. In this region, lies the insertion loop (residues 329-348) and thus we see that except apo monomer in the first eigenvector, the holo monomer and the holo tetramer show decreased degree of fluctuations owing to the rigidity of the insertion loop after ligand binding. This could possibly attributed that initially the amplitude is high in holo form to include the substrate but after inclusion of the substrate, a much rigid conformation is obtained. Another, marked fluctuation is around residues 170-200. Co-incidentally,

this is the area involving the active site residues (residue 199-204) and thus we observe a similar fluctuation profile as for the insertion loop residues. These observations are in support of the data obtained from the crystal structure of PFK[288].

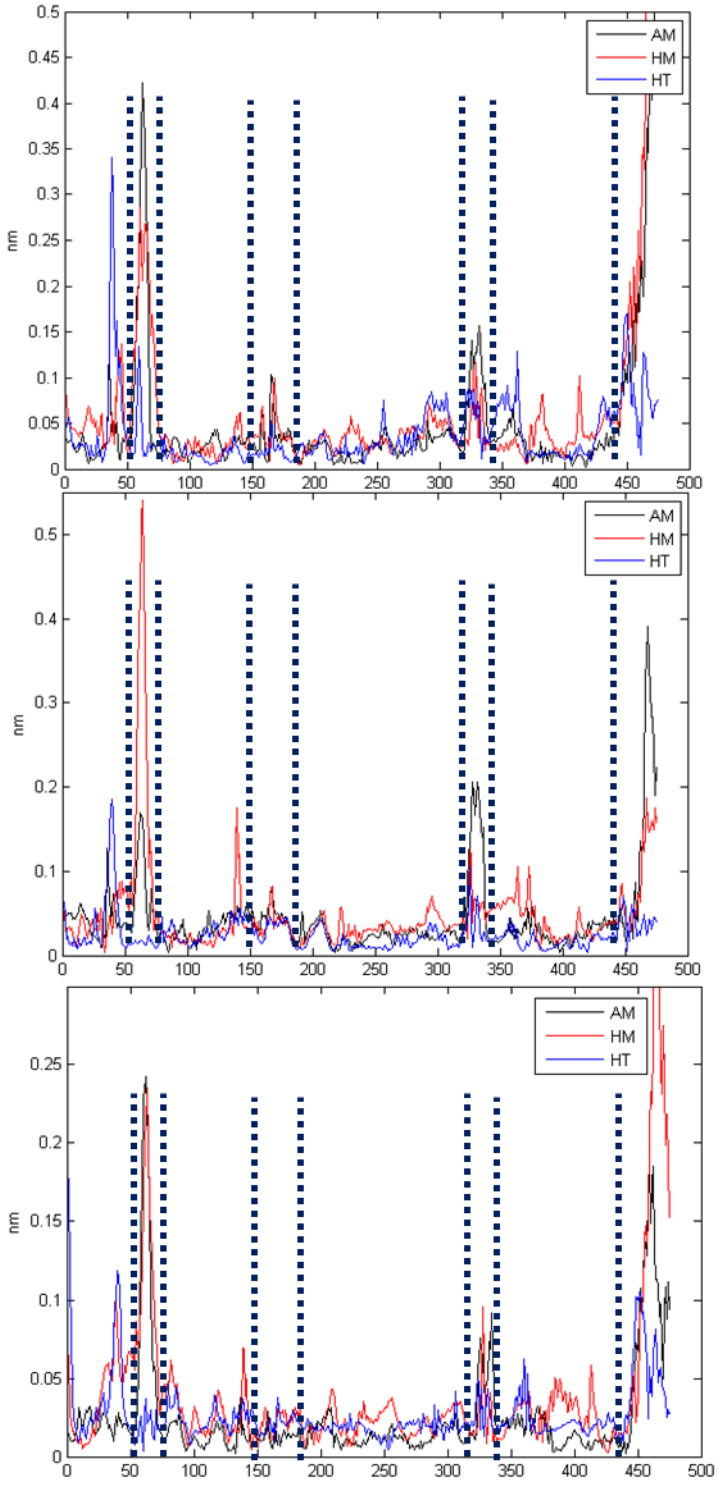


Figure 4.26: Cα root mean square fluctuations projected along the first three eigenvectors for the first subunit of the three generated trajectories. Top: Projection on first eigenvector, Middle: Projection on second eigenvector, Bottom: Projection on third eigenvector.

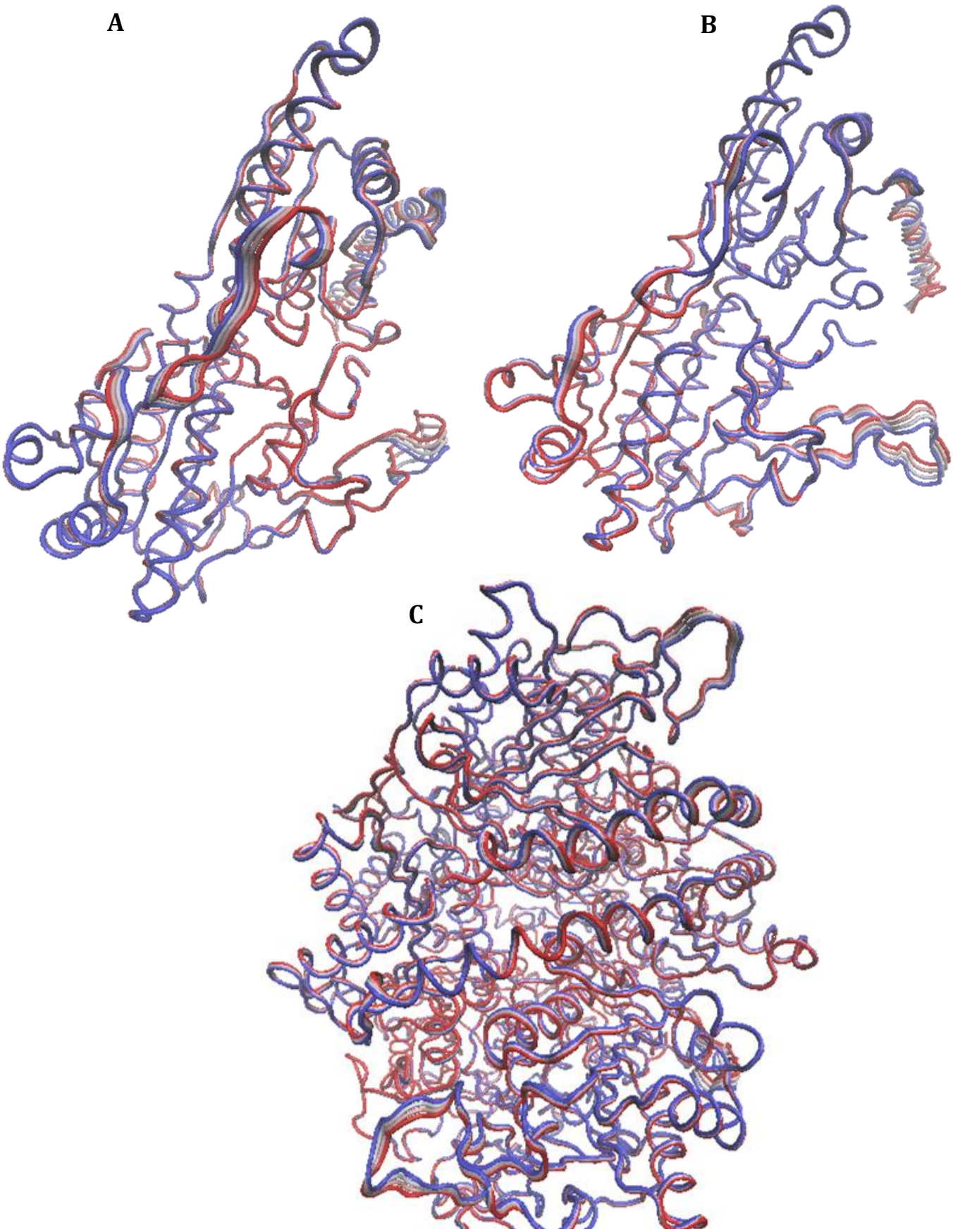


Figure 4.27: Motion described by the first eigenvector (PC 1) for the Apo Monomer (A), Holo Monomer (B) and Holo Tetramer (C). The data is projected along the whole simulation time ith an interval of 20ns each. The color ranges from red to blue with red indicating the beginning of the simulation.

4.3.5 CORRELATION MATRICES

From the above representations of the eigenvectors, it is evident that our principal components contain the essential information pertaining to the behaviour of the protein. To elucidate the relationship between the residues of the protein, we analysed the correlations matrices as obtained from PCA (see materials and methods). As the motion of the protein is constrained to an "essential"subspace, the projections of the principal components show the diagonal and certain non random patterns. So, we analysed the independent as well as correlated inter-residue motions based on cross-correlation matrix and the principal components obtained from them.

The extent of correlations is measured in three ranges i.e. -1, 0, +1. A negative value or negative correlations depict the opposite movements of the residues i.e. the residues move away from each other in opposite directions. 0 would mean non-correlation i.e. the residues are not related to each other and thereby a change or movement in one residue would not alter the conformation of the other. A positive value indicates a positive correlation and thereby depict the concerted motions of the protein in which the residues move in the same direction. We would expect such a behaviour in case of neighbouring atoms which show rigid body motions. However, a completely correlated or anticorrelated motion i.e. a value of 1 or −1, means that the motions have the same phase and period.

In Figures 4.28 & 4.29, we have shown the correlation matrices for the monomer and tetramer, respectively. The matrix is arranged between blue to red from anti-correlation to total correlation. The diagonal represents the correlations of the residues with themselves and thus, the color red implying total correlation ($C(i, j) = 1$).

Overall we see a functional shift towards positive correlation upon addition of the substrate MgATP. This suggests the bound substrate induces the conformational changes in the protein structure which result in rigid body movement of the residues and hence, help in keeping the tetramer together.

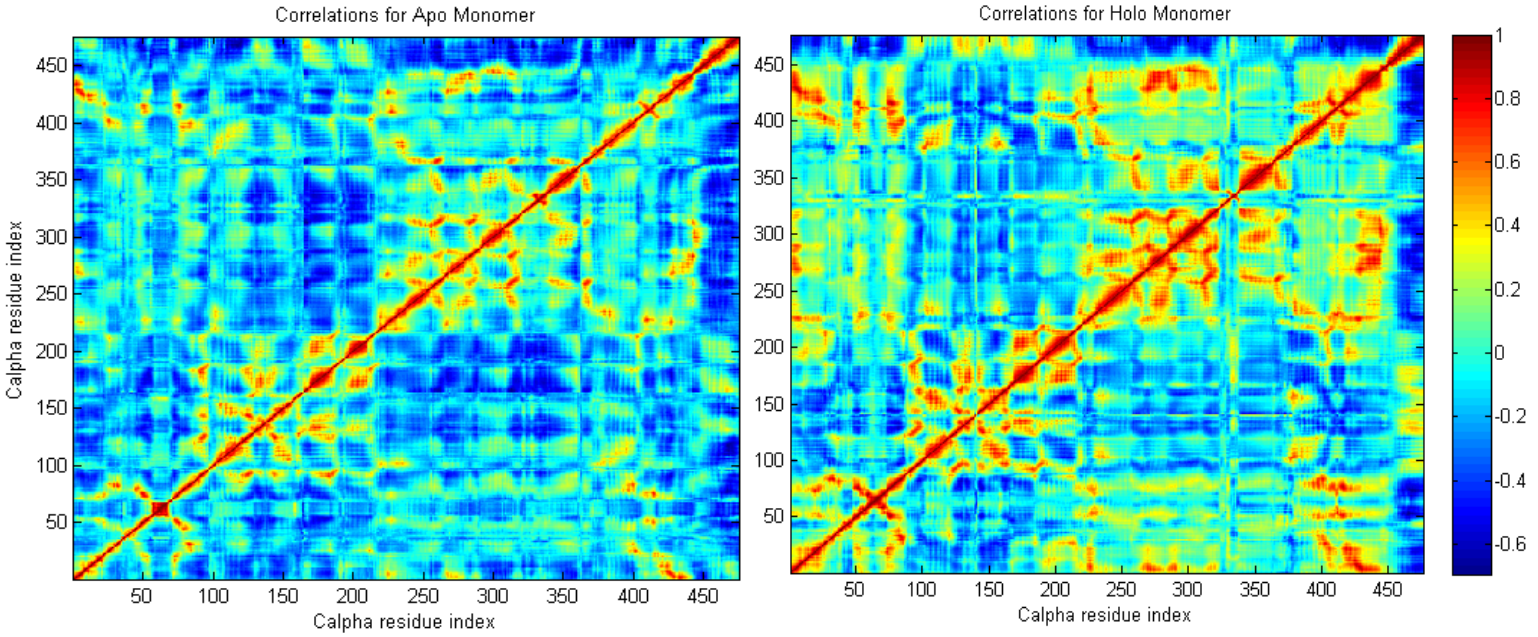


Figure 4.28 : Correlation maps for the Apo (Left) and Holo Tetramer (Right). The X and Y axis indicate the calpha residue number and the correlation color ranges from blue (anti-correlated) to red (positively correlated or total correlation).

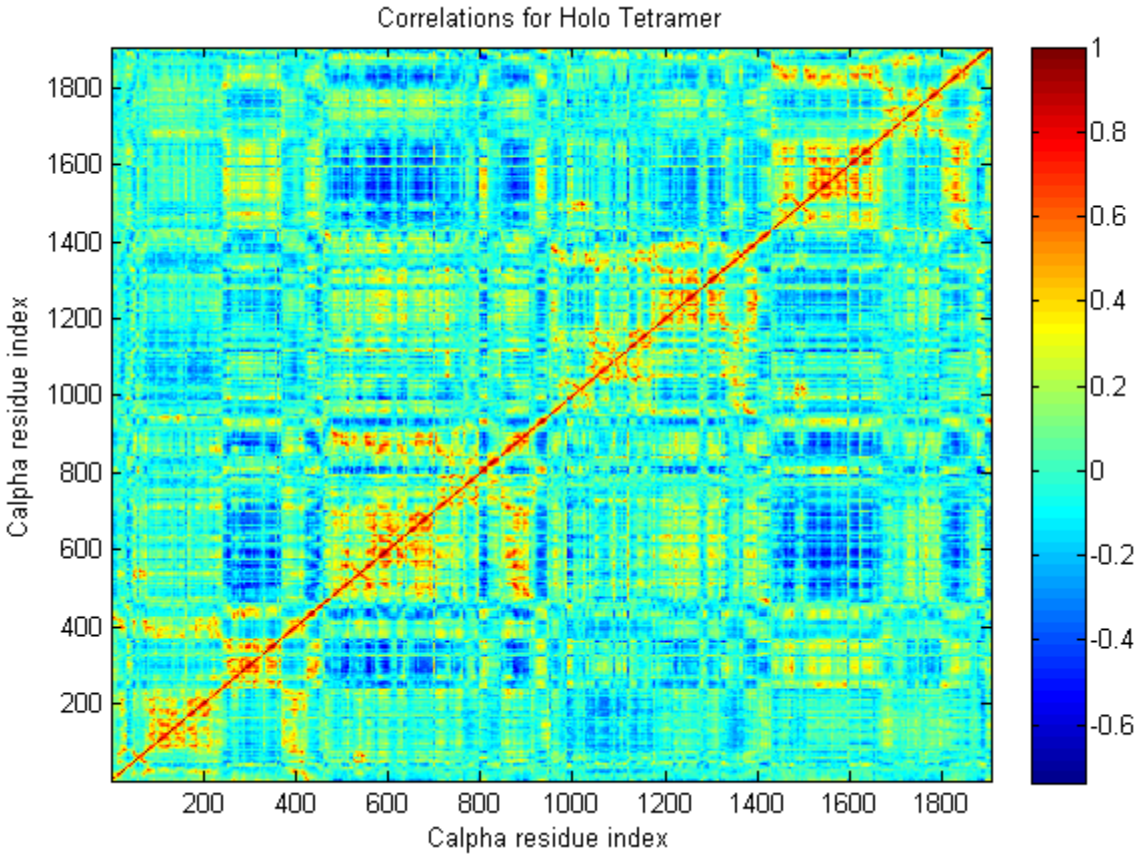


Figure 4.29 : Correlation maps for the Apo (Left) and Holo Tetramer (Right). The X and Y axis indicate the calpha residue number and the correlation color ranges from blue (anti-correlated) to red (positively correlated or total correlation).

Upon closer inspection we observe that there seems to be an increase in correlation between the active site region residues in the holo structure. For instance, we observe that in case of holo, the regions involving the insertion loops (17-20 and 329-348; domain C) and active site residues (199-204; domain B) show a marked reduction in anti-correlation with respect to the apo structure. This suggests and also is in agreement with the observation that upon binding of ATP, these loops adopt an open conformation which would favour the binding of Fructose-6-phosphate in the binding site. Also, an increase in positive correlation amongst the residues confirms that the T-state of phosphofructokinase has a closed active site with the loops being disordered.

Also, the domains B and C show increased correlation suggesting the hinge bending motion and incorporation of ATP in the holo state. Further, the region involving embracing arm and reaching arm depict positive correlations which are necessary to keep the tetramer together.

Of particular interest is also the area involving embracing arm flanked by residues 62-81 and insertion loops i.e. 17-20 & 329-348. This area has a negative correlation for the apo form but in the holo monomer and holo tetramer, we see a marked correlation between these two regions. This suggests that the binding of ATP which is supported by the insertion loops also concertedly manages the movement of the embracing arm. For instance, the correlation between residues 62 and 17 is -0.0883 in apo monomer but for the holo monomer and holo tetramer, this value is 0.0265 and 0.1665, respectively.

Also, as observed previously, the inserted loop is moved out of the ATP-binding pocket for the holoenzyme structure and thus we notice a shift in the position of serine 341 residue to interact with the backbone of glycine 174. This is evident in the correlation matrix where the value of correlation between residues Arginin173 and Serine 341 is -0.36, -0.29, -0.01 and between, Glycine 174 and Serine 341 is -0.37, -0.27 and -0.0025 for the apo monomer, holo monomer and holo tetramer respectively. This reduction in the negative correlation suggests the hydrogen bonds being formed in the holoenzyme.

4.4 CONCLUSION

The prime aim to conduct PFK simulations was to observe the structural behaviour of the protein in terms of domain movement and in the loop regions. Unlike Pyruvate kinase where we also targeted the allosteric site, in case of phosphofructokinase, we adopted the similar approach of combination of MD-PCA but to inspect the dynamic motions only which could later on be extended to study in detail about allosteric nature of PFK. Due to computing time constraints and missing loop structures for PFK, we limited ourselves to the preliminary analyses. However, despite that we still managed to observe few motions pertaining to the active site region and the embracing arms which is also observed experimentally. The principal component analysis of phosphofructokinase can be described in a subspace of very small dimension (less than 1% of the original Cartesian space) consisting of linear combinations of Cartesian degrees of freedom defined in a molecule coordinate system. As we observed for the various projections onto the first few principal components, after the first 4-5 components, the subsequent degrees of freedom can be considered as corresponding to irrelevant Gaussian fluctuations, behaving like near-constraints. The essential subspace itself is defined by the near-constraints, which are related to the mechanical structure of the molecule in a given conformation. These motions are related to the functional behaviour of the enzyme which involves the opening up of the loops to incorporate the substrate and also the hinge-bending movements of B and C domains which enclose the active site.

In the initial stages we concentrated mostly on the global dynamical properties of PFK and thus to grasp the types of motion that follow, we splitted our dynamics into different modes of motion and analysed these modes individually. Such splitting into individual modes is quite meaningful when the fluctuations are not equally spread over all the modes, but rather, localized to few modes.

With application of PCA on our data, we did derive the essential modes that contribute to the overall dynamics of the protein. With our time evolution of these collective coordinates during the simulation, we projected our trajectory onto these coordinates and found out that the first mode of PFK is not only the one with the largest (mean) amplitude, but also that most others tend to describe fluctuations with higher frequencies(Figure 4.21). Since we focussed on first 10 modes of collective fluctuation which contribute significantly to the total fluctuation, we viewed the type of fluctuations of these modes by filtering our trajectory along the first most dominant mode (Figures 4.26 & 4.27). This provided a very nice view of the overall motion of the protein depicted by the first mode.

Also, projecting the trajectory snapshots onto the plane formed by the first two principle components reveals a semicircle, or U/W-shape, relationship (Figure 4.25). This U-shaped pattern probably indicates the random diffusion within the simulation and we can interpret it as a thermal motion along a shallow free-energy landscape[7]. Although this projection does not directly reveal the dominating large-scale conformational changes within the system, it does inform us of the more accessible degrees-of-freedom for thermal motion along our investigated time scale. A PC analysis on our trajectory data revealed that the majority of the motions during Tb. PFK simulations are dominated by the loop regions and the enclosing arms which play an important role in protein activity and substrate binding.

The principal component analysis of phosphofructokinase can be described in a subspace of very small dimension (less than 1% of the original Cartesian space) consisting of linear combinations of Cartesian degrees of freedom defined in a molecule coordinate system. As we observed for the various projections onto the first few principal components, after the first 4-5 components, the subsequent degrees of freedom can be considered as corresponding to irrelevant Gaussian fluctuations, behaving like near-constraints. The essential subspace itself is defined by the near-constraints, which are related to the mechanical structure of the molecule

in a given conformation. These motions are related to the functional behaviour of the enzyme which involves the opening up of the loops to incorporate the substrate and also the hinge-bending movements of B and C domains which enclose the active site.

The 2D and 3D plots of our first two and three PCs and the RMSD versus time plot has also further enhanced our understanding of the conformational space available and sampled by the protein. Since, at present we have run preliminary analysis on PFK and examined the monomer and tetramer single structures only, generalizing the concepts of dynamic motion and the key residues for PFK would be quite premature at this time. This, however, will be extended in the future to investigate deeply the effects of the arms and loops in both the apo and holo forms.

.

CONCLUDING REMARKS

5 CONCLUDING REMARKS

Allostery is a universal phenomenon by virtue of which activity of proteins can be regulated by sites distinct from the active site. Due to the numerous advantages offered by allosteric drugs, identification and modulation of allosteric binding sites is gaining immense interest. Understanding allosteric regulation on a molecular level is important for pharmaceutical research. Although numerous studies are being conducted in the field of allostery and the associated interactions, there is still no concrete theory for atomic level manifestation of allosteric effect. Also, there still seems to be a debate to answer the coupling of allosteric changes with local conformational perturbations.

In the present work, two different enzymes of glycolytic pathway have been studied: pyruvate kinase and phosphofructokinase. In both the cases, we have used Molecular Dynamics Simulations as the prime method to yield dynamic and structural information. In case of former enzyme, we have explored the allosteric regulation and interactions brought about by the effector molecule while in the latter case we have tried to understand the general conformational changes as incurred upon ligand binding.

PYK from Leishmania mexicana catalyzes the final reaction of glycolysis and is allosterically activated by Fructose-2,6-bisphosphate (FBP). The presence of allosteric site 40 Å away from the active site makes it interesting to understand the signal cascade and communication network across the protein. The dynamics of pyruvate kinase have been investigated and the changes of collective motions throughout the tetramer in the presence and absence of FBP have been observed. From these simulations, it is quite evident that the four subunits show a deviation from the symmetrical assumptions of MWC model with their dynamics being slightly independent. The results presented in this work suggest that the highly mobile B-domains which otherwise are not known to actively participate in allosteric movements do play a substantial role in overall fluctuation of the enzyme. We have shown that in the presence of FBP these B-

domains show a decrease in fluctuations which cools down the enzyme into an active conformation.

The second enzyme studied in this thesis is phosphofructokinase. PFK performs the committed step of glycolysis i.e. conversion of fructose-6-phosphate to fructose-1,6-bisphosphate. This is interesting to understand the mechanisms and conformational changes in the tetramer in the presence of ATP. We have carried out Molecular Dynamics Simulations on this protein to further enhance our knowledge of allostery and also gain insight into the structural and dynamical properties at the atomic level. The preliminary results from the simulations are in agreement with the published data.

In this study, MD simulations were exploited to understand the conformational changes of pyruvate kinase and also phosphofructokinase. Many analytical methods were used to handle the data obtained from MD simulations with the prime focus on PCA and cross-correlation analysis. Residue fluctuation graphs complemented well with the previously published data which increased the validity of our work. This analysis also identified the highly mobile B-domains being the prime contributor to the overall fluctuation profile of pyruvate kinase. Cross-correlation analysis helped in identifying the coordinated and collective motions of the proteins and complemented the 'rock and lock' hypothesis. The two dynamic and significant properties like RMSF and cross correlation analysis were used to compare the two distinct states of the same protein i.e. apo and holo forms (apo as in ligand free and holo as in ligand bound or complexed state). Rather than just focussing on highly fluctuating residues, we have related the collective dynamics to specific motions of the protein to identify the underlying mechanism of action. This gave insights into the possible mechanism of regulation by ligands whether allosteric as in case of pyruvate kinase or the conformational changes as in case of phosphofructokinase. Most importantly the results of the large simulations presented make perfect biochemical sense and the observed changes in correlated motions and localised active site helical melting with and without FBP provide important support for a 'rock and lock' model of allostery.

The main aim of our work was to try to answer how the binding of effector molecule affects the dynamics of the tetramer through MD simulation studies. The results presented in this work have provided new and exciting insights into the allosteric mechanism. Also, it is observed that the old view of proteins existing in equilibrium between discrete conformational states is rather obsolete and population of conformations is favoured. It is quite surprising and intriguing to observe that pyruvate kinase despite being a huge tetramer and having a millisecond turnover is accessible at an atomistic time scale.

It is also evident from this study that the combination of Molecular Dynamics simulations and Principal Component Analysis can be well exploited to understand the structural dynamics of a protein which is otherwise restricted in static representation of protein structures. This also provides an excellent opportunity to track the conformational changes following a ligand perturbation and also understand the communication between the two distinct sites i.e. allosteric and active site. From a thermodynamic point of view, the major quest is to understand the relationship between functional properties of protein and distribution of states within an ensemble and most importantly how environmental changes such as binding events and external stimuli alter the function. This work by means of atomistic molecular dynamics simulations has tried to move a step further in understanding the response of the proteins on ligand binding (by means of case studies on pyruvate kinase and phosphofructokinase). Our simulations provide new insights into protein regulation at an atomistic scale and significantly contribute towards understanding the underlying dynamics and functions of the proteins. The main points to note in our work is that the analysis of the MD simulations makes excellent physical sense and the differences between the simulations with and without FBP match with the 'rock-and-lock' allosteric mechanism. In particular we see changes in the anti-correlative motion between the four protomers that matches the removal of FBP, we see a dampening of the B-domain movements and a general cooling of the FBP-bound structure and we see assigns of a melting of the active site helix. The MD simulations are quite expensive experiments to study in atomic detail the effects of FBP binding and to be able to pick out these

effects from 50 to 80 ns simulations of such a large system provides strong support that these simulations are physically reasonable and are providing insight into how binding of FBP affects the allosteric mechanism.

This thesis presented investigation of regulation of PYK and PFK but not detailed accounts into the network of allostery and communication pathways. A PC analysis on an MD trajectory data revealed that the major motions during the simulation are dominated by the B-domains in PYK whose dynamic behavior is known but it not believed to be related to the allosteric movements. However, since we strongly show the role of B-domains in stability, we can possibly work towards developing an allosteric network between the B-domain and the allosteric site. In case of PFK, the major motions are present in the flexible arms and active site region, which reiterates the experimental structure data. However, this can serve as a seed for further simulations on PFK by taking the complete set of protein i.e. Apo form and the Holo form to further inspect the movements of the loops. The exploration of allosteric networks and mapping of residues involved in communication cascades is limited by time scale and analysis detail of the simulations. With increase in computing power and exploration of higher order correlation functions and theoretical models can be helpful in addressing the issue. Further work needs to be done in order to identify the novel residues involved in the allosteric communication which could serve as putative drug targets to design isoform specific allosteric modulators.

Despite the fact that the allosteric signals are subtle to track with the possibilities of an array of communication pathways being followed, we have still been able to complement our results with the structural information provided by experiments. The B-factors in case of PYK and PFK are very much in sync with the experimental data. Besides this, they also reveal an alternating profile with the diagonally related chains in case of PYK being symmetrical. This can be taken as a working hypothesis to inspect further as to why the chains behave as a dimer of dimers. We could do so by either extending the simulations by another 50-60 ns to check if the profile remains consistent or we could run the simulations in GPUs to track whether there is any deviation or further enhancement to our results. Also, since we know the behaviour of the first few Principal components, we

could further focus on the next set of components to detect the allostery network or the residues which might be different from the previously described residues in this thesis. Principal components contain vast amount of information about protein regulation and motion which could be exhaustive to work upon but certainly we could limit the first set of investigating PCs as the first 50 and perhaps random 50 in between. This would reveal alternating conformations of the protein. Principally, we have laid the groundwork with our current research which would serve as benchmark for further simulations when extended.

Since the basic principles of MD simulations and the analytical methods have been studied in detail, we could now apply this to other proteins of the glycolytic pathway one by one. This would provide an excellent opportunity to complement and enhance the structural information with our simulations and then track the possible outcomes for protein modulators. The important role of molecular dynamics simulation stimulates from the fact that biomacromolecules such as protein exist in a dynamic state of motion. This has been seen in case of PYK and PFK. These dynamic motions are directly related to the function of protein like protein-binding interactions or signalling, self-assembly and conformational change. Experimental techniques aid the understanding of the structural features but are constrained in their approach to characterize the dynamic motions. MD simulations presented here in conjugation with the PCA technique provides a new means to model the flexibility and conformational changes at an atomic level and also explore novel areas which are otherwise difficult to characterize experimentally.

In order to be able to progress towards designing allosteric modulators, we could take preliminary steps towards Structure based drug design focussing on the structure of protein targets. Since the technique is based on designing molecules specific to the protein target or structural domains, we could implement our observations from MD simulations. For instance, the first step would be to target the residues of B-domain area which are quite evident and then probably mutate the residues to see if the effect remains the same. This would confirm the role of the B-domains in intrinsic allosteric nature of the protein. Further, mutating the active site residues or the allosteric site residues would be

beneficial. This could be then served as a target to design allosteric modulators for our protein.

In conclusion, we found that using Molecular dynamics simulation alongwith PCA analysis provides a rich array of structural information which needs to be extensively analysed in order to develop potential modulators. The analyses of different PC subspaces allowed us to form a coherent conclusion concerning involvement of B-domains in protein functionality. Also, by mapping these PCs onto 2D and 3D plots for the first two and three PCs, and onto an RMSD versus time plot, allowed us to understand the conformational space relationship better. There are a number of possible outcomes and exciting opportunities ahead for us to extend our work in terms of allosteric network and also to begin designing modulators for our protein. Another extension would be to focus on the flexible regions of the protein primarily, the target receptor and design improved modulators by following the lock and key concept of the static receptor. Since, experimental structures are limited in terms of exploring additional cryptic or allosteric sites; this can be done by extending our work. Similar work has been done in HIV-integrase where a novel binding site was identified through simulations[300, 301]. Also, another example is where a novel binding site was identified for p53 protein and also for human β1 & β2 adrenergic receptors [179, 302-304].

ACKNOWLEDGEMENTS

6 ACKNOWLEDGEMENTS

As I finish writing this thesis, I get nostalgic and suddenly surrounded with flashbacks of the summer of 2011 when I first came to Edinburgh. This was my first step away from a tightly pampered cocoon into a more realistic and independent world. So, here I want to take this opportunity to thank everyone who have helped me along the educational journey and transformed me from a naïve and spoilt brat to a much mature and sensible individual.

First and foremost, I am grateful to my supervisor Prof. Malcolm D. Walkinshaw for providing me with a platform to pursue this research. Without his never-ending and ever encouraging & morale boosting support, discussions and guidance, it would not have been possible to finish this work. I am also thankful to him for bestowing me with his pearls of wisdom and trusting me with my work. It has always been a great sense of relief to have such a cool and wonderful supervisor around whose office was always open for umpteen discussions and comical reliefs (with remarkable witty comments from his side). Very special thanks to my co-supervisor Dr. Paul Taylor, for his valuable suggestions and discussions throughout the course of this work. I would also like to thank him for being so prompt and patient with my silly mails and most importantly being receptive to them. I would also like to thank my other committee members Dr. Martin Wear and Dr. Julien Michel for their advice and comments.

A special thanks to Darwin Trust of Edinburgh for providing me with the funding for my research. I would additionally like to extend my thanks to Prof. Burak Erman from Koc University, Turkey, who collaborated in this research work. Thank you for your insightful comments and valuable discussions. My thanks also extends to all the Walkinshaw group members for their welcoming and helpful nature. Also, thanks to all my friends with whom I shared my office. It was great to be surrounded by such a fine bunch of jovial and interactive people who made the office a cheerful place even through hair splitting times.

And last but not the least, my heartfelt gratitude to my parents for their never-ending support and care which has been and is invaluable in my life. Thank you Dad for spoiling me but keeping my head above water with your daily testimonials albeit in a light hearted manner. You have been my mentor and buddy since I first stepped into this world and continue to do so. Thank you Mom for your warmth and positivity through turbulent times. I hope to make you guys proud and live upto the expectations.

APPENDICES

7 APPENDIX

7.1 INPUT FILES FOR PYK SIMULATIONS

A. HQQ Apo tetramer simulation

NPT Equilibration

```
title                = NPT equilibration
;define             = -DPOSRES  ; position restrain the protein
; Run parameters
integrator          = md            ; leap-frog integrator
nsteps              = 50000       ; 250 ps
dt                  = 0.005                 ; 5 fs
; Output control
nstxout             = 0             ; save coordinates every 5 ps
nstvout             = 0             ; save velocities every 5 ps
nstenergy           = 10            ; save energies every 5 ps
nstlog              = 1000          ; update log file every 5 ps
; Bond parameters
continuation        = yes           ; first dynamics run
constraint_algorithm = lincs        ; holonomic constraints
constraints         = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter          = 1             ; accuracy of LINCS
lincs_order         = 4             ; also related to accuracy
; Neighborsearching
ns_type                     = grid          ; search neighboring grid cells
nstlist             = 10            ; 50 fs
rlist               = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb            = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw                = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong           = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype         = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order           = 4             ; cubic interpolation
fourierspacing      = 0.16          ; grid spacing for FFT
; Temperature coupling is on
tcoupl              = V-rescale     ; modified Berendsen thermostat
tc-grps             = Protein Non-Protein; two coupling groups - more accurate
tau_t               = 0.1   0.1     ; time constant, in ps
ref_t               = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl          = berendsen     ; Berendsen thermostat
pcoupltype          = isotropic     ; uniform scaling of box vectors
tau_p               = 1.0           ; time constant, in ps
ref_p               = 1.0           ; reference pressure, in bar
```

```
compressibility = 4.5e-5        ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = no            ; assign velocities from Maxwell distribution
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

NVT Equilibration

```
title           = NVT equilibration
define          = -DPOSRES  ; position restrain the protein
; Run parameters
integrator      = md            ; leap-frog integrator
nsteps          = 50000                 ; 1000 ps
dt              = 0.005                 ; 5 fs
; Output control
nstxout         = 0             ; save coordinates every 50 fs
nstvout         = 0             ; save velocities every 50 fs
nstenergy       = 10            ; save energies every 50 fs
nstlog          = 1000          ; update log file every 5 ps
; Bond parameters
continuation    = no            ; first dynamics run
constraint_algorithm = lincs    ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 10            ; 50 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
; Temperature coupling is on
tcoupl          = V-rescale     ; Berendsen thermostat
tc-grps         = Protein Non-Protein; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl          = no            ; no pressure coupling in NVT
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
```

```
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; assign velocities from Maxwell distribution
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

MD Simulation

```
title           = production MD
; Run parameters
integrator      = md            ; leap-frog algorithm
; Output control
nstxout         = 0             ; save coordinates every 2 ps
nstvout         = 0             ; save velocities every 2 ps
nstxtcout       = 1000          ; xtc compressed trajectory output every 5 ps
nstenergy       = 1000        ; save energies every 5 ps
nstlog          = 1000        ; update log file every 5 ps
; Bond parameters
constraint_algorithm = lincs   ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 5             ; 25 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
nstcomm = 10                  ; remove com every 10 steps
; Temperature coupling is on
tcoupl          = V-rescale     ; modified Berendsen thermostat
tc-grps         = Protein Non-Protein; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl        = berendsen       ; Berendsen thermostat
pcoupltype      = isotropic     ; uniform scaling of box vectors
tau_p           = 1.0           ; time constant, in ps
ref_p           = 1.0           ; reference pressure, in bar
compressibility = 4.5e-5        ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
```

```
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; Velocity generation is on
gen_temp        = 318           ; reference temperature, for protein in K
```

B. HQQ Holo tetramer simulation

NPT Equilibration

```
title           = NPT equilibration
;define        = -DPOSRES  ; position restrain the protein
; Run parameters
integrator      = md            ; leap-frog integrator
nsteps          = 62500       ; 250 ps
dt              = 0.004                 ; 4 fs
; Output control
nstxout         = 0             ; save coordinates every 5 ps
nstvout         = 0             ; save velocities every 5 ps
nstenergy       = 10            ; save energies every 5 ps
nstlog          = 1000          ; update log file every 5 ps
; Bond parameters
continuation    = yes           ; first dynamics run
constraint_algorithm = lincs  ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 10            ; 50 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
; Temperature coupling is on
tcoupl          = V-rescale     ; modified Berendsen thermostat
tc-grps         = Protein Non-Protein; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl        = berendsen       ; Berendsen thermostat
```

```
pcoupltype      = isotropic     ; uniform scaling of box vectors
tau_p           = 1.0           ; time constant, in ps
ref_p           = 1.0           ; reference pressure, in bar
compressibility = 4.5e-5        ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = no            ; assign velocities from Maxwell distribution
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

NVT Equilibration

```
title           = NVT equilibration
define          = -DPOSRES  ; position restrain the protein
; Run parameters
integrator      = md            ; leap-frog integrator
;nsteps         = 50000                 ; 1000 ps
nsteps      = 62500     ; 250 ps
dt              = 0.004                 ; 4 fs
; Output control
nstxout         = 0             ; save coordinates every 50 fs
nstvout         = 0             ; save velocities every 50 fs
nstenergy       = 10            ; save energies every 50 fs
nstlog          = 1000          ; update log file every 5 ps
; Bond parameters
continuation    = no            ; first dynamics run
constraint_algorithm = lincs    ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 10            ; 50 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing  = 0.16          ; grid spacing for FFT
; Temperature coupling is on
tcoupl          = V-rescale     ; Berendsen thermostat
tc-grps         = Protein Non-Protein   ; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
```

```
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl          = no            ; no pressure coupling in NVT
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; assign velocities from Maxwell distribution
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

MD Simulation

```
title           = production MD
; Run parameters
integrator      = md            ; leap-frog algorithm
dt              = 0.003                 ; 3 fs
; Output control
nstxout         = 0             ; save coordinates every 2 ps
nstvout         = 0             ; save velocities every 2 ps
nstxtcout       = 1000          ; xtc compressed trajectory output every 5 ps
nstenergy       = 1000      ; save energies every 5 ps
nstlog          = 1000      ; update log file every 5 ps
; Bond parameters
constraint_algorithm = lincs  ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 5             ; 25 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
nstcomm = 10                ; remove com every 10 steps
; Temperature coupling is on
tcoupl          = V-rescale     ; modified Berendsen thermostat
tc-grps         = Protein Non-Protein; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl        = berendsen       ; Berendsen thermostat
```

```
pcoupltype      = isotropic     ; uniform scaling of box vectors
tau_p           = 1.0           ; time constant, in ps
ref_p           = 1.0           ; reference pressure, in bar
compressibility = 4.5e-5        ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; Velocity generation is on
gen_temp        = 318           ; reference temperature, for protein in K
```

C. HQP Holo tetramer simulation

NPT Equilibration

```
title           = NPT equilibration
;define         = -DPOSRES  ; position restrain the protein
; Run parameters
integrator      = md            ; leap-frog integrator
nsteps          = 125000            ; 250 ps
dt              = 0.002                 ; 2 fs
; Output control
nstxout         = 0             ; save coordinates every 5 ps
nstvout         = 0             ; save velocities every 5 ps
nstenergy       = 10            ; save energies every 5 ps
nstlog          = 1000          ; update log file every 5 ps
; Bond parameters
continuation    = yes           ; first dynamics run
constraint_algorithm = lincs    ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 10            ; 50 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.4           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
```

```
pme_order      = 4              ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
; Temperature coupling is on
tcoupl         = V-rescale      ; Berendsen thermostat
tc-grps        = Protein Non-Protein; two coupling groups - more accurate
tau_t          = 0.1    0.1     ; time constant, in ps
ref_t          = 318    318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl      = Parrinello-Rahman          ; pressure coupling is on for NPT
pcoupltype  = isotropic                  ; uniform scaling of box vectors
tau_p       = 2.0                        ; time constant, in ps
ref_p       = 1.0                        ; reference pressure, in bar
compressibility = 4.5e-5                 ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc            = xyz            ; 3-D PBC
; Dispersion correction
DispCorr       = EnerPres       ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = no             ; assign velocities from Maxwell distribution
gen_temp       = 300            ; temperature for Maxwell distribution
gen_seed       = -1             ; generate a random seed
```

NVT Equilibration

```
title          = NVT equilibration
define         = -DPOSRES  ; position restrain the protein
; Run parameters
integrator     = md             ; leap-frog integrator
nsteps         = 125000         ; 250 ps
dt             = 0.002                  ; 2 fs
; Output control
nstxout        = 0              ; save coordinates every 50 fs
nstvout        = 0              ; save velocities every 50 fs
nstenergy      = 10             ; save energies every 50 fs
nstlog         = 1000           ; update log file every 5 ps
; Bond parameters
continuation   = no             ; first dynamics run
constraint_algorithm = lincs    ; holonomic constraints
constraints    = all-bonds      ; all bonds (even heavy atom-H bonds) constrained
lincs_iter     = 1              ; accuracy of LINCS
lincs_order    = 4              ; also related to accuracy
; Neighborsearching
ns_type                 = grid           ; search neighboring grid cells
nstlist        = 10             ; 50 fs
rlist          = 1.0            ; short-range neighborlist cutoff (in nm)
rcoulomb       = 1.0            ; short-range electrostatic cutoff (in nm)
rvdw           = 1.0            ; short-range van der Waals cutoff (in nm)
rlistlong      = 1.0            ; long-range neighborlist cutoff (in nm)
```

```
; Electrostatics
coulombtype   = PME          ; Particle Mesh Ewald for long-range electrostatics
pme_order     = 4            ; cubic interpolation
fourierspacing = 0.16        ; grid spacing for FFT
; Temperature coupling is on
tcoupl        = V-rescale    ; Berendsen thermostat
tc-grps       = Protein Non-Protein; two coupling groups - more accurate
tau_t         = 0.1   0.1    ; time constant, in ps
ref_t         = 318   318    ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl        = no           ; no pressure coupling in NVT
; Periodic boundary conditions
pbc           = xyz          ; 3-D PBC
; Dispersion correction
DispCorr      = EnerPres     ; account for cut-off vdW scheme
; Velocity generation
gen_vel               = yes          ; assign velocities from Maxwell distribution
gen_temp      = 300          ; temperature for Maxwell distribution
gen_seed      = -1           ; generate a random seed
```

MD Simulation

```
title         = production MD
; Run parameters
integrator    = md           ; leap-frog algorithm
dt          = 0.002     ; 2 fs
;nsteps       = 20000000     ; 0.005 * 20000000 = 100000 ps or 100 ns
;nsteps       = 200000       ; 0.005 * 200000 = 1 ns
;dt           = 0.005                ; 5 fs
;dt           = 0.003                ; 3 fs
; Output control
nstxout       = 0            ; save coordinates every 2 ps
nstvout       = 0            ; save velocities every 2 ps
nstxtcout     = 1000         ; xtc compressed trajectory output every 5 ps
nstenergy     = 1000       ; save energies every 5 ps
nstlog        = 1000       ; update log file every 5 ps
; Bond parameters
constraint_algorithm = lincs  ; holonomic constraints
constraints   = all-bonds    ; all bonds (even heavy atom-H bonds) constrained
lincs_iter    = 1            ; accuracy of LINCS
lincs_order   = 4            ; also related to accuracy
; Neighborsearching
ns_type               = grid         ; search neighboring grid cells
nstlist       = 5            ; 25 fs
rlist         = 1.0          ; short-range neighborlist cutoff (in nm)
rcoulomb      = 1.0          ; short-range electrostatic cutoff (in nm)
rvdw          = 1.4          ; short-range van der Waals cutoff (in nm)
rlistlong     = 1.0          ; long-range neighborlist cutoff (in nm)
```

```
cutoff-scheme   = Verlet
; Electrostatics
coulombtype   = PME          ; Particle Mesh Ewald for long-range electrostatics
pme_order     = 4            ; cubic interpolation
fourierspacing = 0.16        ; grid spacing for FFT
nstcomm = 10                 ; remove com every 10 steps
; Temperature coupling is on
tcoupl        = V-rescale    ; modified Berendsen thermostat
tc-grps       = Protein Non-Protein; two coupling groups - more accurate
tau_t         = 0.1   0.1    ; time constant, in ps
ref_t         = 318   318    ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl        = Parrinello-Rahman          ; pressure coupling is on for NPT
pcoupltype    = isotropic    ; uniform scaling of box vectors
tau_p         = 2.0                ; time constant, in ps
ref_p         = 1.0          ; reference pressure, in bar
compressibility = 4.5e-5     ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc           = xyz          ; 3-D PBC
; Dispersion correction
DispCorr      = EnerPres     ; account for cut-off vdW scheme
; Velocity generation
gen_vel               = yes          ; Velocity generation is on
gen_temp      = 318          ; reference temperature, for protein in K
```

7.2 INPUT FILES FOR PFK SIMULATIONS

A. PFK Apo Monomer simulation

NPT Equilibration

```
title                    = NPT equilibration
;define                 = -DPOSRES  ; position restrain the protein
; Run parameters
integrator              = md            ; leap-frog integrator
nsteps                  = 50000      ; 250 ps
dt                      = 0.005                 ; 5 fs
; Output control
nstxout                 = 0             ; save coordinates every 5 ps
nstvout                 = 0             ; save velocities every 5 ps
nstenergy               = 10            ; save energies every 5 ps
nstlog                  = 1000          ; update log file every 5 ps
; Bond parameters
continuation            = yes           ; first dynamics run
constraint_algorithm = lincs  ; holonomic constraints
constraints             = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter              = 1             ; accuracy of LINCS
lincs_order             = 4             ; also related to accuracy
; Neighborsearching
ns_type                         = grid          ; search neighboring grid cells
nstlist                 = 10            ; 50 fs
rlist                   = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb                = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw                    = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong               = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype             = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order               = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
; Temperature coupling is on
tcoupl                  = V-rescale     ; modified Berendsen thermostat
tc-grps                 = Protein Non-Protein; two coupling groups - more accurate
tau_t                   = 0.1   0.1     ; time constant, in ps
ref_t                   = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl          = berendsen     ; Berendsen thermostat
pcoupltype      = isotropic     ; uniform scaling of box vectors
tau_p                   = 1.0           ; time constant, in ps
ref_p                   = 1.0           ; reference pressure, in bar
compressibility = 4.5e-5        ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc                     = xyz           ; 3-D PBC
; Dispersion correction
```

```
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = no            ; assign velocities from Maxwell distribution
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

NVT Equilibration

```
title           = NVT equilibration
define          = -DPOSRES  ; position restrain the protein
; Run parameters
integrator      = md            ; leap-frog integrator
nsteps          = 50000                 ; 1000 ps
dt              = 0.005                 ; 5 fs
; Output control
nstxout         = 0             ; save coordinates every 50 fs
nstvout         = 0             ; save velocities every 50 fs
nstenergy       = 10            ; save energies every 50 fs
nstlog          = 1000          ; update log file every 5 ps
; Bond parameters
continuation    = no            ; first dynamics run
constraint_algorithm = lincs    ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 10            ; 50 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
; Temperature coupling is on
tcoupl          = V-rescale     ; Berendsen thermostat
tc-grps         = Protein Non-Protein; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl          = no            ; no pressure coupling in NVT
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; assign velocities from Maxwell distribution
```

```
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

MD Simulation

```
title           = production MD
; Run parameters
integrator      = md            ; leap-frog algorithm
dt              = 0.005                 ; 5 fs
; Output control
nstxout         = 0             ; save coordinates every 2 ps
nstvout         = 0             ; save velocities every 2 ps
nstxtcout       = 1000          ; xtc compressed trajectory output every 5 ps
nstenergy       = 1000      ; save energies every 5 ps
nstlog          = 1000      ; update log file every 5 ps
; Bond parameters
constraint_algorithm = lincs   ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 5             ; 25 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
nstcomm = 10                ; remove com every 10 steps
; Temperature coupling is on
tcoupl          = V-rescale     ; modified Berendsen thermostat
tc-grps         = Protein Non-Protein; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl        = berendsen       ; Berendsen thermostat
pcoupltype      = isotropic     ; uniform scaling of box vectors
tau_p           = 1.0           ; time constant, in ps
ref_p           = 1.0           ; reference pressure, in bar
compressibility = 4.5e-5        ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; Velocity generation is on
gen_temp        = 318           ; reference temperature, for protein in K
```

B. PFK Holo Monomer simulation

NPT Equilibration

```
title                    = NPT equilibration
;define                 = -DPOSRES  ; position restrain the protein
; Run parameters
integrator               = md        ; leap-frog integrator
nsteps                   = 50000     ; 250 ps
dt                       = 0.005     ; 5 fs
; Output control
nstxout                  = 0         ; save coordinates every 5 ps
nstvout                  = 0         ; save velocities every 5 ps
nstenergy                = 10        ; save energies every 5 ps
nstlog                   = 1000      ; update log file every 5 ps
; Bond parameters
continuation             = yes       ; first dynamics run
constraint_algorithm     = lincs     ; holonomic constraints
constraints              = all-bonds ; all bonds (even heavy atom-H bonds) constrained
lincs_iter               = 1         ; accuracy of LINCS
lincs_order              = 4         ; also related to accuracy
; Neighborsearching
ns_type                  = grid      ; search neighboring grid cells
nstlist                  = 10        ; 50 fs
rlist                    = 1.0       ; short-range neighborlist cutoff (in nm)
rcoulomb                 = 1.0       ; short-range electrostatic cutoff (in nm)
rvdw                     = 1.0       ; short-range van der Waals cutoff (in nm)
rlistlong                = 1.0       ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype              = PME       ; Particle Mesh Ewald for long-range electrostatics
pme_order                = 4         ; cubic interpolation
fourierspacing           = 0.16      ; grid spacing for FFT
; Temperature coupling is on
tcoupl                   = V-rescale ; modified Berendsen thermostat
tc-grps                  = Protein Non-Protein; two coupling groups - more accurate
tau_t                    = 0.1   0.1 ; time constant, in ps
ref_t                    = 318   318 ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl                   = berendsen ; Berendsen thermostat
pcoupltype               = isotropic ; uniform scaling of box vectors
tau_p                    = 1.0       ; time constant, in ps
ref_p                    = 1.0       ; reference pressure, in bar
compressibility          = 4.5e-5    ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc                      = xyz       ; 3-D PBC
; Dispersion correction
DispCorr                 = EnerPres  ; account for cut-off vdW scheme
```

```
; Velocity generation
gen_vel                 = no            ; assign velocities from Maxwell distribution
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

NVT Equilibration

```
title           = NVT equilibration
define          = -DPOSRES  ; position restrain the protein
; Run parameters
integrator      = md            ; leap-frog integrator
nsteps          = 50000                 ; 1000 ps
dt              = 0.005                 ; 5 fs
; Output control
nstxout         = 0             ; save coordinates every 50 fs
nstvout         = 0             ; save velocities every 50 fs
nstenergy       = 10            ; save energies every 50 fs
nstlog          = 1000          ; update log file every 5 ps
; Bond parameters
continuation    = no            ; first dynamics run
constraint_algorithm = lincs    ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 10            ; 50 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
; Temperature coupling is on
tcoupl          = V-rescale     ; Berendsen thermostat
tc-grps         = Protein Non-Protein; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl          = no            ; no pressure coupling in NVT
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; assign velocities from Maxwell distribution
```

```
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

MD Simulation

```
title           = production MD
; Run parameters
integrator      = md            ; leap-frog algorithm
dt              = 0.005                 ; 5 fs
; Output control
nstxout         = 0             ; save coordinates every 2 ps
nstvout         = 0             ; save velocities every 2 ps
nstxtcout       = 1000          ; xtc compressed trajectory output every 5 ps
nstenergy       = 1000      ; save energies every 5 ps
nstlog          = 1000      ; update log file every 5 ps
; Bond parameters
constraint_algorithm = lincs  ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 5             ; 25 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
nstcomm = 10                ; remove com every 10 steps
; Temperature coupling is on
tcoupl          = V-rescale     ; modified Berendsen thermostat
tc-grps         = Protein Non-Protein; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl        = berendsen       ; Berendsen thermostat
pcoupltype      = isotropic     ; uniform scaling of box vectors
tau_p           = 1.0           ; time constant, in ps
ref_p           = 1.0           ; reference pressure, in bar
compressibility = 4.5e-5        ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; Velocity generation is on
gen_temp        = 318           ; reference temperature, for protein in K
```

C. PFK Holo Tetramer simulation

NPT Equilibration

```
title                    = NPT equilibration
;define                 = -DPOSRES  ; position restrain the protein
; Run parameters
integrator               = md        ; leap-frog integrator
;nsteps                  = 62500     ; 250 ps
;dt                      = 0.004     ; 4 fs
nsteps                   = 83333     ; 250 ps
dt                       = 0.003     ; 3 fs
; Output control
nstxout                  = 0         ; save coordinates every 5 ps
nstvout                  = 0         ; save velocities every 5 ps
nstenergy                = 10        ; save energies every 5 ps
nstlog                   = 1000      ; update log file every 5 ps
; Bond parameters
continuation             = yes       ; first dynamics run
constraint_algorithm     = lincs     ; holonomic constraints
constraints              = all-bonds ; all bonds (even heavy atom-H bonds) constrained
lincs_iter               = 1         ; accuracy of LINCS
lincs_order              = 4         ; also related to accuracy
; Neighborsearching
ns_type                  = grid      ; search neighboring grid cells
nstlist                  = 10        ; 50 fs
rlist                    = 1.0       ; short-range neighborlist cutoff (in nm)
rcoulomb                 = 1.0       ; short-range electrostatic cutoff (in nm)
rvdw                     = 1.0       ; short-range van der Waals cutoff (in nm)
rlistlong                = 1.0       ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype              = PME       ; Particle Mesh Ewald for long-range electrostatics
pme_order                = 4         ; cubic interpolation
fourierspacing           = 0.16      ; grid spacing for FFT
; Temperature coupling is on
tcoupl                   = V-rescale ; modified Berendsen thermostat
tc-grps                  = Protein Non-Protein; two coupling groups - more accurate
tau_t                    = 0.1   0.1 ; time constant, in ps
ref_t                    = 318   318 ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl                   = berendsen ; Berendsen thermostat
pcoupltype               = isotropic ; uniform scaling of box vectors
tau_p                    = 1.0       ; time constant, in ps
ref_p                    = 1.0       ; reference pressure, in bar
compressibility          = 4.5e-5    ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
```

```
pbc             = xyz           ; 3-D PBC
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = no            ; assign velocities from Maxwell distribution
gen_temp        = 300           ; temperature for Maxwell distribution
gen_seed        = -1            ; generate a random seed
```

NVT Equilibration

```
title           = NVT equilibration
define          = -DPOSRES  ; position restrain the protein
; Run parameters
integrator      = md            ; leap-frog integrator
;nsteps         = 50000                 ; 1000 ps
;dt             = 0.005                 ; 5 fs
nsteps          = 83333                 ; 250 ps
dt              = 0.003                 ; 3fs
; Output control
nstxout         = 0             ; save coordinates every 50 fs
nstvout         = 0             ; save velocities every 50 fs
nstenergy       = 10            ; save energies every 50 fs
nstlog          = 1000          ; update log file every 5 ps
; Bond parameters
continuation    = no            ; first dynamics run
constraint_algorithm = lincs    ; holonomic constraints
constraints     = all-bonds     ; all bonds (even heavy atom-H bonds) constrained
lincs_iter      = 1             ; accuracy of LINCS
lincs_order     = 4             ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist         = 10            ; 50 fs
rlist           = 1.0           ; short-range neighborlist cutoff (in nm)
rcoulomb        = 1.0           ; short-range electrostatic cutoff (in nm)
rvdw            = 1.0           ; short-range van der Waals cutoff (in nm)
rlistlong       = 1.0           ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype     = PME           ; Particle Mesh Ewald for long-range electrostatics
pme_order       = 4             ; cubic interpolation
fourierspacing  = 0.16          ; grid spacing for FFT
; Temperature coupling is on
tcoupl          = V-rescale     ; Berendsen thermostat
tc-grps         = Protein Non-Protein   ; two coupling groups - more accurate
tau_t           = 0.1   0.1     ; time constant, in ps
ref_t           = 318   318     ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl          = no            ; no pressure coupling in NVT
; Periodic boundary conditions
```

```
pbc              = xyz          ; 3-D PBC
; Dispersion correction
DispCorr         = EnerPres     ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; assign velocities from Maxwell distribution
gen_temp         = 300          ; temperature for Maxwell distribution
gen_seed         = -1           ; generate a random seed
```

MD Simulation

```
title            = production MD
; Run parameters
integrator       = md           ; leap-frog algorithm
; Output control
nstxout          = 0            ; save coordinates every 2 ps
nstvout          = 0            ; save velocities every 2 ps
nstxtcout        = 1000         ; xtc compressed trajectory output every 5 ps
nstenergy        = 1000       ; save energies every 5 ps
nstlog           = 1000       ; update log file every 5 ps
; Bond parameters
constraint_algorithm = lincs   ; holonomic constraints
constraints      = all-bonds    ; all bonds (even heavy atom-H bonds) constrained
lincs_iter       = 1            ; accuracy of LINCS
lincs_order      = 4            ; also related to accuracy
; Neighborsearching
ns_type                 = grid          ; search neighboring grid cells
nstlist          = 5            ; 25 fs
rlist            = 1.0          ; short-range neighborlist cutoff (in nm)
rcoulomb         = 1.0          ; short-range electrostatic cutoff (in nm)
rvdw             = 1.0          ; short-range van der Waals cutoff (in nm)
rlistlong        = 1.0          ; long-range neighborlist cutoff (in nm)
; Electrostatics
coulombtype      = PME          ; Particle Mesh Ewald for long-range electrostatics
pme_order        = 4            ; cubic interpolation
fourierspacing = 0.16           ; grid spacing for FFT
nstcomm = 10                  ; remove com every 10 steps
; Temperature coupling is on
tcoupl           = V-rescale    ; modified Berendsen thermostat
tc-grps          = Protein Non-Protein; two coupling groups - more accurate
tau_t            = 0.1   0.1    ; time constant, in ps
ref_t            = 318   318    ; reference temperature, one for each group, in K
; Pressure coupling is off
pcoupl         = berendsen     ; Berendsen thermostat
pcoupltype       = isotropic    ; uniform scaling of box vectors
tau_p            = 1.0          ; time constant, in ps
ref_p            = 1.0          ; reference pressure, in bar
compressibility = 4.5e-5        ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc              = xyz          ; 3-D PBC
```

```
; Dispersion correction
DispCorr        = EnerPres      ; account for cut-off vdW scheme
; Velocity generation
gen_vel                 = yes           ; Velocity generation is on
gen_temp        = 318           ; reference temperature, for protein in K
```

SEMINARS & COURSES

8 PRESENTATIONS/SEMINARS/COURSES

1. Protein Purification : seminar-cum-course-cum-workshop.
2. Chaired a sesion on 'Kinetoplastid Meeting-1'.
3. Presented a talk in 'Kinetoplastid Meeting-2' on , 24th may 2013.
5. Structural Biology workshop-cum-course.
6. Bioinformatics short workshop.
7. First year Biology Program course 'Structure & Function of Proteins'.
8. Workshop on Eddie.
9. Publishing your work - Training, poster presentation on 5th June 2013
10. Event asistant-1, 2012 April
11. Event asistant-2, 2013, April Science fest
12. 'Alostery Workshop' in Turkey - Poster presentation
13. 'Alostery Workshop' in Edinburgh-poster presentation
14. Mathematical Course - Molecular Dynamics Seminar-NAIS
15. Java Workshop - 2012
16. Sequencing and Gene course – December 2011
17. Chaired a sesion at ISMB 3rd year talk in December 2011
18. 'Introductory Statistics' - An interactive statistics course
19. Atended Crystalography Summer School at University of St. Andrews
20. Seminar 'Multicore Programming with Open MP'.
21. International Scientific Advisory Board Meeting 30/31 August 2012
22. Proteomics Course
23. EPCC Mesage Pasing Programming Seminar
24. Laser Gene Workshop
25. Postgraduate Poster Workshop
26. Laboratory Demonstrator Induction Course

27. Ensemble Workshop

28. Demonstration SFP course October-November 2013

29. Poster Presentation – Edinburgh University ; Second Year Poster

30. Carlisle Meeting 4 - 7th September 2013 – Talk

31. Carlisle Meeting 4 - 7th September 2013 Poster

32. December 2013 : Talk at the ISMB/ICB Symposium

33. Facilitator for First Year PhD students (Writing 10 Week Report)

34. Facilitator for Second Year PhD students (Writing 1st Year review; 10 Month report)

35. Tour guide for the Postgraduate open day

36. Demonstrator for the Drug Discovery course

37. Tutorial on Molecular Docking : May 2014

38. 10 – 14 September 2014, Modeling of Biomolecular Systems Interactions, Dynamics, and Allostery: Bridging Experiments and Computations Koc University , Istanbul, Turkey : Poster presentation

BIBLIOGRAPHY

9 BIBLIOGRAPHY

1. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules.* Nat Struct Mol Biol, 2002. **9**(9): p. 646-652.
2. Bakker, B.M., et al., *Glycolysis in Bloodstream Form Trypanosoma brucei Can Be Understood in Terms of the Kinetics of the Glycolytic Enzymes.* Journal of Biological Chemistry, 1997. **272**(6): p. 3207-3215.
3. Lu, S., S. Li, and J. Zhang, *Harnessing allostery: A novel approach to drug discovery.* Medicinal Research Reviews, 2014.
4. Lesk, A.M., *Introduction to Protein Architecture: The Structural Biology of Proteins*. 2001: Oxford University Press.
5. Smith, C., *Nature.* 2003. **422**(6929): p. 341.
6. Poehlsgaard, J. and S. Douthwaite, *Nature Reviews Microbiology.* 2005. **3**(11): p. 870-881.
7. Cho, Y., et al., *Science.* 1994. **265**(5170): p. 346-355.
8. Dobson, C.M., *Nature.* 2002. **418**(6899): p. 729-730.
9. Kelly, J.W., *Nature Structural Biology.* 2002. **9**(5): p. 323-325.
10. Koo, E.H., Lansbury P.T, Jr., and J.W. Kelly, *Proceedings of the National Academy of Sciences of the United States of America.* 1999. **96**(18): p. 9989-9990.
11. Nielsen, M.H., F.S. Pedersen, and J. Kjems, *Retrovirology.* 2005. **2**.
12. Ohtaka, H. and E. Freire, *Progress in Biophysics and Molecular Biology.* 2005. **88**(2): p. 193-208.
13. Bacha, U., et al., *Biochemistry.* 2004. **43**(17): p. 4906-4912.
14. Venkatraman, S., et al., *Journal of Medicinal Chemistry.* 2005. **48**(16): p. 5088-5091.
15. Wootten, D., A. Christopoulos, and P.M. Sexton, *Emerging paradigms in GPCR allostery: implications for drug discovery.* Nat Rev Drug Discov, 2013. **12**(8): p. 630-644.
16. Fenton, A.W., *Allostery: an illustrated definition for the 'second secret of life'.* Trends Biochem Sci, 2008. **33**(9): p. 420-5.
17. Goodey, N.M. and S.J. Benkovic, *Allosteric regulation and catalysis emerge via a common route.* Nat Chem Biol, 2008. **4**(8): p. 474-82.
18. Formaneck, M.S. and Q. Cui, *The use of a generalized born model for the analysis of protein conformational transitions: a comparative study with explicit solvent simulations for chemotaxis Y protein (CheY).* J Comput Chem, 2006. **27**(16): p. 1923-43.
19. Cui, Q. and M. Karplus, *Allostery and cooperativity revisited.* Protein Science, 2008. **17**(8): p. 1295-1307.
20. Csermely, P., R. Palotai, and R. Nussinov, *Induced fit, conformational selection and independent dynamic segments: an extended view of binding events.* Trends in Biochemical Sciences. **35**(10): p. 539-546.
21. Pan, Y., et al., *Mechanisms of transcription factor selectivity.* Trends in Genetics. **26**(2): p. 75-83.
22. Sinha, N. and R. Nussinov, *Point mutations and sequence variability in proteins: Redistributions of preexisting populations.* Proceedings of the National Academy of Sciences, 2001. **98**(6): p. 3139-3144.
23. Kuriyan, J. and D. Eisenberg, *The origin of protein interactions and allostery in colocalization.* Nature, 2007. **450**(7172): p. 983-90.
24. Changeux, J.P. and S.J. Edelstein, *Allosteric mechanisms of signal transduction.* Science, 2005. **308**(5727): p. 1424-8.
25. Cooper, A. and D. Dryden, *Allostery without conformational change- A plausible Model* European Biophysics Journal, 1984(11): p. 103-109.
26. Andersen, O.S. and R.E. Koeppe, 2nd, *Bilayer thickness and membrane protein function: an energetic perspective.* Annu Rev Biophys Biomol Struct, 2007. **36**: p. 107-30.
27. Popovych, N., et al., *Dynamically driven protein allostery.* Nat Struct Mol Biol, 2006. **13**(9): p. 831-8.

28. Homans, S.W., *Probing the binding entropy of ligand-protein interactions by NMR.* Chembiochem, 2005. **6**(9): p. 1585-91.
29. Wand, A.J., *Dynamic activation of protein function: A view emerging from NMR spectroscopy.* Nat Struct Mol Biol, 2001. **8**(11): p. 926-931.
30. Kar, G., et al., *Allostery and population shift in drug discovery.* Curr Opin Pharmacol, 2010. **10**(6): p. 715-22.
31. Kenakin, T. and L.J. Miller, *Seven transmembrane receptors as shapeshifting proteins: the impact of allosteric modulation and functional selectivity on new drug discovery.* Pharmacol Rev, 2010. **62**(2): p. 265-304.
32. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Biol, 2003. **10**(1): p. 59-69.
33. Datta, D., et al., *An allosteric circuit in caspase-1.* J Mol Biol, 2008. **381**(5): p. 1157-67.
34. Motlagh, H.N., et al., *The ensemble nature of allostery.* Nature, 2014. **508**(7496): p. 331-339.
35. Nussinov, R., et al., *Allosteric Conformational Barcodes Direct Signaling in the Cell.* Structure. **21**(9): p. 1509-1521.
36. Huang, M., et al., *Conformational Transition Pathway in the Activation Process of Allosteric Glucokinase.* PLoS ONE, 2013. **8**(2): p. e55857.
37. Bell, J.E., J.K. Bell, and D.J. Abraham, *Allosteric Proteins and Drug Discovery*, in *Burger's Medicinal Chemistry and Drug Discovery*. 2003, John Wiley & Sons, Inc.
38. Fang, Z., C. Grütter, and D. Rauh, *Strategies for the Selective Regulation of Kinases with Allosteric Modulators: Exploiting Exclusive Structural Features.* ACS Chemical Biology, 2012. **8**(1): p. 58-70.
39. Palmieri, L. and G. Rastelli, *αC helix displacement as a general approach for allosteric modulation of protein kinases.* Drug Discovery Today, 2013. **18**(7–8): p. 407-414.
40. Taly, A., et al., *Nicotinic receptors: allosteric transitions and therapeutic targets in the nervous system.* Nat Rev Drug Discov, 2009. **8**(9): p. 733-750.
41. Christopoulos, A., *Allosteric binding sites on cell-surface receptors: novel targets for drug discovery.* Nat Rev Drug Discov, 2002. **1**(3): p. 198-210.
42. Kamata, K., et al., *Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase.* Structure, 2004. **12**(3): p. 429-38.
43. Villaverde, A., *Allosteric enzymes as biosensors for molecular diagnosis.* FEBS Letters. **554**(1): p. 169-172.
44. Monod, J., J. Wyman, and J.-P. Changeux, *On the nature of allosteric transitions: A plausible model.* Journal of Molecular Biology, 1965. **12**(1): p. 88-118.
45. Changeux, J.P., *Allostery and the Monod-Wyman-Changeux model after 50 years.* Annu Rev Biophys, 2012. **41**: p. 103-33.
46. Horovitz, A. and K.R. Willison, *Allosteric regulation of chaperonins.* Current Opinion in Structural Biology, 2005. **15**(6): p. 646-651.
47. Shimizu, T.S., et al., *Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis.* Nat Cell Biol, 2000. **2**(11): p. 792-796.
48. Henry, E.R., et al., *A tertiary two-state allosteric model for hemoglobin.* Biophysical Chemistry, 2002. **98**(1–2): p. 149-164.
49. Koshland, D., *Application of a theory of enzyme specificity to protein synthesis.* Proc Natl Acad Sci USA, 1958. **44**: p. 98 - 104.
50. Koshland, D.E., *Correlation of structure and function in enzyme action.* Science, 1963. **142**: p. 1533-1541.
51. Tzeng, S.-R. and C.G. Kalodimos, *Allosteric inhibition through suppression of transient conformational states.* Nat Chem Biol, 2013. **9**(7): p. 462-465.
52. Boehr, D.D., R. Nussinov, and P.E. Wright, *The role of dynamic conformational ensembles in biomolecular recognition.* Nat Chem Biol, 2009. **5**(11): p. 789-796.

53. Jaffe, E. and S. Lawrence, *The Morpheein Model of Allostery: Evaluating Proteins as Potential Morpheeins*, in *Allostery*, A.W. Fenton, Editor. 2012, Springer New York. p. 217-231.
54. Jaffe, E.K., *Morpheeins – a new structural paradigm for allosteric regulation.* Trends in Biochemical Sciences. **30**(9): p. 490-497.
55. Tzeng, S.R. and C.G. Kalodimos, *Dynamic activation of an allosteric regulatory protein.* Nature, 2009. **462**(7271): p. 368-72.
56. Tzeng, S.-R. and C.G. Kalodimos, *Protein activity regulation by conformational entropy.* Nature, 2012. **488**(7410): p. 236-240.
57. Daily, M.D. and J.J. Gray, *Local motions in a benchmark of allosteric proteins.* Proteins, 2007. **67**(2): p. 385-99.
58. Daily, M.D., T.J. Upadhyaya, and J.J. Gray, *Contact rearrangements form coupled networks from local motions in allosteric proteins.* Proteins, 2008. **71**(1): p. 455-66.
59. Namboodiri, S., et al., *Looking for a sequence based allostery definition: A statistical journey at different resolution scales.* Journal of Theoretical Biology, 2012. **304**(0): p. 211-218.
60. Li, X., et al., *Toward an understanding of the sequence and structural basis of allosteric proteins.* Journal of Molecular Graphics and Modelling, 2013. **40**(0): p. 30-39.
61. Perutz, M.F., et al., *Nature.* 1960. **185**(4711): p. 416-422.
62. Wang, Q., et al., *Toward understanding the molecular basis for chemical allosteric modulator design.* Journal of Molecular Graphics and Modelling, 2012. **38**(0): p. 324-333.
63. Jeremy M. Berg, J.L. Tymoczko, and L. Stryer, *Biochemistry*, 2012, W.H. Freeman: New York. p. 1120.
64. Scheer, J.M., M.J. Romanowski, and J.A. Wells, *A common allosteric site and mechanism in caspases.* Proceedings of the National Academy of Sciences, 2006. **103**(20): p. 7595-7600.
65. Erlanson, D.A., J.A. Wells, and A.C. Braisted, *TETHERING: Fragment-Based Drug Discovery.* Annual Review of Biophysics and Biomolecular Structure, 2004. **33**(1): p. 199-223.
66. Modesti, M., *Fluorescent Labeling of Proteins*, in *Single Molecule Analysis*, E.J.G. Peterman and G.J.L. Wuite, Editors. 2011, Humana Press. p. 101-120.
67. Link, H., K. Kochanowski, and U. Sauer, *Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo.* Nat Biotech, 2013. **31**(4): p. 357-361.
68. Wenthur, C.J., et al., *Drugs for Allosteric Sites on Receptors.* Annual Review of Pharmacology and Toxicology, 2014. **54**(1): p. 165-184.
69. Macarron, R., et al., *Impact of high-throughput screening in biomedical research.* Nat Rev Drug Discov, 2011. **10**(3): p. 188-195.
70. Bembenek, S.D., B.A. Tounge, and C.H. Reynolds, *Ligand efficiency and fragment-based drug discovery.* Drug Discovery Today, 2009. **14**(5–6): p. 278-283.
71. Hajduk, P.J. and J. Greer, *A decade of fragment-based drug design: strategic advances and lessons learned.* Nat Rev Drug Discov, 2007. **6**(3): p. 211-219.
72. Dror, R.O., et al., *Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs.* Nature, 2013. **503**(7475): p. 295-299.
73. Huang, W., et al., *Allosite: a method for predicting allosteric sites.* Bioinformatics, 2013. **29**(18): p. 2357-2359.
74. Mitternacht, S. and I.N. Berezovsky, *Binding Leverage as a Molecular Basis for Allosteric Regulation.* PLoS Comput Biol, 2011. **7**(9): p. e1002148.
75. Mitternacht, S. and I.N. Berezovsky, *Coherent Conformational Degrees of Freedom as a Structural Basis for Allosteric Communication.* PLoS Comput Biol, 2011. **7**(12): p. e1002301.
76. Qi, Y., et al., *Identifying Allosteric Binding Sites in Proteins with a Two-State Gō Model for Novel Allosteric Effector Discovery.* Journal of Chemical Theory and Computation, 2012. **8**(8): p. 2962-2971.

77. Panjkovich, A. and X. Daura, *Exploiting protein flexibility to predict the location of allosteric sites.* BMC Bioinformatics, 2012. **13**(1): p. 273.
78. Pritchard, L., et al., *Simple intrasequence difference (SID) analysis: an original method to highlight and rank sub-structural interfaces in protein folds. Application to the folds of bovine pancreatic trypsin inhibitor, phospholipase A2, chymotrypsin and carboxypeptidase A.* Protein Engineering, 2003. **16**(2 %U http://peds.oxfordjournals.org/content/16/2/87.abstract): p. 87-101.
79. Scott, D.L., et al., *Structures of free and inhibited human secretory phospholipase A2 from inflammatory exudate.* Science, 1991. **254**(5034 %U http://www.sciencemag.org/content/254/5034/1007.abstract): p. 1007-1010.
80. Lehninger, A., D. Nelson, and M. Cox, *Lehninger Principles of Biochemistry*. 2008: W. H. Freeman.
81. Chennubhotla, C., Z. Yang, and I. Bahar, *Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL.* Molecular BioSystems, 2008. **4**(4): p. 287-292.
82. Bryn Fenwick, R., et al., *Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(4): p. E445-E454.
83. Boehr, D.D., et al., *The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis.* Science, 2006. **313**(5793): p. 1638-1642.
84. Gunasekaran, K., B. Ma, and R. Nussinov, *Is allostery an intrinsic property of all dynamic proteins?* Proteins, 2004. **57**(3): p. 433-43.
85. Lange, O.F., et al., *Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution.* Science, 2008. **320**(5882): p. 1471-1475.
86. Poulsen, H., *An introduction to three-dimensional X-ray diffraction microscopyThis article forms part of a special issue dedicated to advanced diffraction imaging methods of materials, which will be published as a virtual special issue of the journal in 2013.* Journal of Applied Crystallography, 2012. **45**(6): p. 1084-1097.
87. Wlodawer, A., et al., *Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination.* FEBS Journal, 2013. **280**(22): p. 5705-5736.
88. Smyth, M.S. and J.H.J. Martin, *x Ray crystallography.* Molecular Pathology, 2000. **53**(1): p. 8-14.
89. Evans, P. and A. McCoy, *An introduction to molecular replacement.* Acta Crystallogr D Biol Crystallogr, 2008. **64**(Pt 1): p. 1-10.
90. Bauman, J.D., et al., *Detecting Allosteric Sites of HIV-1 Reverse Transcriptase by X-ray Crystallographic Fragment Screening.* Journal of Medicinal Chemistry, 2013. **56**(7): p. 2738-2746.
91. Safo, M.K., et al., *X-ray crystallographic analyses of symmetrical allosteric effectors of hemoglobin: compounds designed to link primary and secondary binding sites.* Acta Crystallographica Section D, 2002. **58**(4): p. 634-644.
92. Gronwald, W. and H.R. Kalbitzer, *Automated structure determination of proteins by NMR spectroscopy.* Progress in Nuclear Magnetic Resonance Spectroscopy, 2004. **44**(1): p. 33-96.
93. Skrisovska, L., M. Schubert, and F.T. Allain, *Recent advances in segmental isotope labeling of proteins: NMR applications to large proteins and glycoproteins.* Journal of Biomolecular NMR, 2010. **46**(1): p. 51-65.
94. Rieko, I. and A.T. Dennis, *Protein dynamics from NMR.* Nature Structural & Molecular Biology, 2000. **7**(9): p. 740-743.
95. Pervushin, K., *Impact of Transverse Relaxation Optimized Spectroscopy (TROSY) on NMR as a technique in structural biology.* Quarterly Reviews of Biophysics, 2000. **33**(02): p. 161-197.

96. Pervushin, K., et al., *Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution.* Proceedings of the National Academy of Sciences, 1997. **94**(23): p. 12366-12371.
97. Raman, S., et al., *Accurate Automated Protein NMR Structure Determination Using Unassigned NOESY Data.* Journal of the American Chemical Society, 2009. **132**(1): p. 202-207.
98. Jahnke, W., et al., *Strategies for the NMR-Based Identification and Optimization of Allosteric Protein Kinase Inhibitors.* ChemBioChem, 2005. **6**(9): p. 1607-1610.
99. Palmer, A.G., *Probing molecular motion by NMR.* Curr. Opin. Struct. Biol., 1997. **7**: p. 732-737.
100. Bastiaens, P.I.H. and R. Pepperkok, *Observing proteins in their natural habitat: the living cell.* Trends in Biochemical Sciences, 2000. **25**(12): p. 631-637.
101. Johnson, C.K., *Calmodulin, conformational states, and calcium signaling. A single-molecule perspective.* Biochemistry, 2006. **45**(48): p. 14233-14246.
102. Gaczynska, M. and P.A. Osmulski, *AFM of biological complexes: What can we learn?* Current Opinion in Colloid and Interface Science, 2008. **13**(5): p. 351-367.
103. Mou, J., et al., *Chaperonins GroEL and GroES: Views from atomic force microscopy.* Biophysical Journal, 1996. **71**(4): p. 2213-2221.
104. Hansma, P.K., et al., *Tapping mode atomic force microscopy in liquids.* Applied Physics Letters, 1994. **64**(13): p. 1738-1740.
105. Osmulski, P.A. and M. Gaczynska, *Nanoenzymology of the 20S proteasome: Proteasomal actions are controlled by the allosteric transition.* Biochemistry, 2002. **41**(22): p. 7047-7053.
106. Junker, J.P., F. Ziegler, and M. Rief, *Ligand-dependent equilibrium fluctuations of single calmodulin molecules.* Science, 2009. **323**(5914): p. 633-637.
107. Vertrees, J., et al., *COREX/BEST server: a web browser-based program that calculates regional stability variations within protein structures.* Bioinformatics, 2005. **21**(15): p. 3318-3319.
108. Hilser, V.J. and E. Freire, *Structure-based Calculation of the Equilibrium Folding Pathway of Proteins. Correlation with Hydrogen Exchange Protection Factors.* Journal of Molecular Biology, 1996. **262**(5): p. 756-772.
109. Freire, E., *Can allosteric regulation be predicted from structure?* Proceedings of the National Academy of Sciences, 2000. **97**(22): p. 11680-11682.
110. Pan, H., J.C. Lee, and V.J. Hilser, *Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble.* Proceedings of the National Academy of Sciences, 2000. **97**(22): p. 12020-12025.
111. Shulman, A.I., et al., *Structural Determinants of Allosteric Ligand Activation in RXR Heterodimers.* Cell, 2004. **116**(3): p. 417-429.
112. Hardy, J.A. and J.A. Wells, *Searching for new allosteric sites in enzymes.* Curr Opin Struct Biol, 2004. **14**(6): p. 706-15.
113. Lichtarge, O., H. Bourne, and F. Cohen, *An evolutionary trace method defines binding surfaces common to protein families.* J Mol Biol, 1996. **257**: p. 342 - 358.
114. Collier, G. and V. Ortiz, *Emerging computational approaches for the study of protein allostery.* Archives of Biochemistry and Biophysics, 2013. **538**(1): p. 6-15.
115. Brinda, K. and S. Vishveshwara, *Oligomeric protein structure networks: insights into protein-protein interactions.* BMC Bioinformatics, 2005. **6**(1): p. 296.
116. Bader, G.D., et al., *BIND - The Biomolecular Interaction Network Database.* Nucleic Acids Research, 2001. **29**(1): p. 242-245.
117. Zanzoni, A., et al., *MINT: A Molecular INTeraction database.* FEBS Letters, 2002. **513**(1): p. 135-140.
118. Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces.* Journal of Molecular Biology, 1998. **280**(1): p. 1-9.

119. Keskin, O., B. Ma, and R. Nussinov, *Hot regions in protein-protein interactions: The organization and contribution of structurally conserved hot spot residues.* Journal of Molecular Biology, 2005. **345**(5): p. 1281-1294.
120. Ofran, Y. and B. Rost, *Protein-protein interaction hotspots carved into sequences.* PLoS Computational Biology, 2007. **3**(7): p. 1169-1176.
121. Cho, K.I., D. Kim, and D. Lee, *A feature-based approach to modeling protein-protein interaction hot spots.* Nucleic Acids Research, 2009. **37**(8): p. 2672-2687.
122. Li, L., et al., *A computational investigation of allostery in the catabolite activator protein.* J Am Chem Soc, 2007. **129**(50): p. 15668-15676.
123. Bradley, M.J., P.T. Chivers, and N.A. Baker, *Molecular Dynamics Simulation of the Escherichia coli NikR Protein: Equilibrium Conformational Fluctuations Reveal Interdomain Allosteric Communication Pathways.* Journal of Molecular Biology, 2008. **378**(5): p. 1155-1173.
124. Lange, O.F. and H. Grubmüller, *Generalized correlation for biomolecular dynamics.* Proteins: Structure, Function, and Bioinformatics, 2006. **62**(4): p. 1053-1061.
125. Lange, O.F., H. Grubmüller, and B.L. de Groot, *Molecular Dynamics Simulations of Protein G Challenge NMR-Derived Correlated Backbone Motions.* Angewandte Chemie International Edition, 2005. **44**(22): p. 3394-3399.
126. Zhang, D. and J.A. McCammon, *The Association of Tetrameric Acetylcholinesterase with ColQ Tail: A Block Normal Mode Analysis.* PLoS Comput Biol, 2005. **1**(6): p. e62.
127. Feher, V.A., et al., *Computational approaches to mapping allosteric pathways.* Current Opinion in Structural Biology, 2014. **25**: p. 98-103.
128. Lenaerts, T., et al., *Quantifying information transfer by protein domains: Analysis of the Fyn SH2 domain structure.* 2008. **8**: p. 43.
129. Dennis, S., T. Kortvelyesi, and S. Vajda, *Computational mapping identifies the binding sites of organic solvents on proteins.* Proceedings of the National Academy of Sciences, 2002. **99**(7): p. 4290-4295.
130. Kortvelyesi, T., et al., *Improved mapping of protein binding sites.* Journal of Computer-Aided Molecular Design, 2003. **17**(2-4): p. 173-186.
131. Verdonk, M.L., et al., *Superstar: improved knowledge-based interaction fields for protein binding sites.* Journal of Molecular Biology, 2001. **307**(3): p. 841-859.
132. Ruppert, J., W. Welch, and A.N. Jain, *Automatic identification and representation of protein binding sites for molecular docking.* Protein Science, 1997. **6**(3): p. 524-533.
133. Laskowski, R.A., *SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions.* Journal of Molecular Graphics, 1995. **13**(5): p. 323-330.
134. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* Journal of Molecular Graphics and Modelling, 1997. **15**(6): p. 359-363.
135. Brady, G.P., Jr. and P.W. Stouten, *Fast prediction and visualization of protein binding pockets with PASS.* Journal of Computer-Aided Molecular Design, 2000. **14**(4): p. 383-401.
136. Binkowski, T.A., S. Naghibzadeh, and J. Liang, *CASTp: Computed Atlas of Surface Topography of proteins.* Nucleic Acids Research, 2003. **31**(13): p. 3352-3355.
137. Weisel, M., E. Proschak, and G. Schneider, *PocketPicker: analysis of ligand binding-sites with shape descriptors.* Chemistry Central Journal, 2007. **1**(1): p. 1-17.
138. Mehio, W., et al., *Identification of protein binding surfaces using surface triplet propensities.* Bioinformatics, 2010. **26**: p. 2549 - 2555.
139. An, J., M. Totrov, and R. Abagyan, *Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes.* Molecular & Cellular Proteomics, 2005. **4**(6): p. 752-761.
140. Laurie, A.T.R. and R.M. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites.* Bioinformatics, 2005. **21**(9): p. 1908-1916.

141. van Gunsteren, W.F., et al., *Biomolecular Modeling: Goals, Problems, Perspectives.* Angewandte Chemie International Edition, 2006. **45**(25): p. 4064-4092.
142. Scheraga, H.A., M. Khalili, and A. Liwo, *Protein-folding dynamics: overview of molecular simulation techniques.* Annu Rev Phys Chem, 2007. **58**: p. 57-83.
143. Car, R. and M. Parrinello, *Unified Approach for Molecular Dynamics and Density-Functional Theory.* Physical Review Letters, 1985. **55**(22): p. 2471-2474.
144. Alder, B.J. and T. Wainwright, *Proc. Int. Symp. Transp. Process. Stat. Mech*. 1958. 97.
145. Alder, B.J. and T.E. Wainwright, *Phase Transition for a Hard Sphere System.* The Journal of Chemical Physics, 1957. **27**(5): p. 1208-1209.
146. Rahman, A., *Correlations in the Motion of Atoms in Liquid Argon.* Physical Review, 1964. **136**(2A): p. A405-A411.
147. Rahman, A. and F.H. Stillinger, *Molecular Dynamics Study of Liquid Water.* The Journal of Chemical Physics, 1971. **55**(7): p. 3336-3359.
148. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* Nature, 1977. **267**(5612): p. 585-590.
149. Brooks Iii, C.L., *Characterization of "native" apomyoglobin by molecular dynamics simulation.* Journal of Molecular Biology, 1992. **227**(2): p. 375-380.
150. AE, M. and v.G. WF, Biochemistry, 1992. **31**: p. 7745.
151. Daggett, V. and M. Levitt, *Protein Unfolding Pathways Explored Through Molecular Dynamics Simulations.* Journal of Molecular Biology, 1993. **232**(2): p. 600-619.
152. J, B., G. T, and O. H, J. Physique II, 1993. **3**: p. 245.
153. Daggett, V. and A.R. Fersht, Trends Biochem. Sci., 2003. **28**: p. 18.
154. Duan, Y. and P.A. Kollman, *Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution.* Science, 1998. **282**(5389): p. 740-744.
155. Pande, V.S. and D.S. Rokhsar, *Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein G.* Proceedings of the National Academy of Sciences, 1999. **96**(16): p. 9062-9067.
156. Roccatano, D., et al., *A molecular dynamics study of the 41-56 β-hairpin from B1 domain of protein G.* Protein Science, 1999. **8**(10): p. 2130-2143.
157. Tsai, J., M. Levitt, and D. Baker, *Hierarchy of structure loss in MD simulations of src SH3 domain unfolding.* Journal of Molecular Biology, 1999. **291**(1): p. 215-225.
158. Adcock, S.A. and J.A. McCammon, *Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins.* Chemical Reviews, 2006. **106**(5): p. 1589-1615.
159. Okazaki, K.I. and S. Takada, *Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(32): p. 11182-11187.
160. Lu, H. and K. Schulten, *Steered molecular dynamics simulation of conformational changes of immunoglobulin domain I27 interprete atomic force microscopy observations.* Chemical Physics, 1999. **247**(1): p. 141-153.
161. Olsen, O.H., et al., *A combined structural dynamics approach identifies a putative switch in factor VIIa employed by tissue factor to initiate blood coagulation.* Protein Science, 2007. **16**(4): p. 671-682.
162. Schlitter, J., et al., *Targeted molecular dynamics simulation of conformational change - application to the T ↔ R transition in insulin.* Molecular Simulation, 1993. **10**(2-6): p. 291-308.
163. Whitford, P.C., S. Gosavi, and J.N. Onuchic, *Conformational transitions in adenylate kinase: Allosteric communication reduces misligation.* Journal of Biological Chemistry, 2008. **283**(4): p. 2042-2048.
164. Marchi, M. and P. Ballone, *Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems.* Journal of Chemical Physics, 1999. **110**(8): p. 3697-3702.

165. Lindahl, E., B. Hess, and D. van der Spoel, *GROMACS 3.0: a package for molecular simulation and trajectory analysis.* Molecular modeling annual, 2001. **7**(8): p. 306-317.
166. Sugita, Y. and Y. Okamoto, *Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape.* Chemical Physics Letters, 2000. **329**(3–4): p. 261-270.
167. Hamelberg, D., J. Mongan, and J. McCammon, *Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules.* J Chem Phys, 2004. **120**: p. 11919 - 11929.
168. Torrie, G.M. and J.P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.* Journal of Computational Physics, 1977. **23**(2): p. 187-199.
169. Kumar, S., et al., *THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.* Journal of Computational Chemistry, 1992. **13**(8): p. 1011-1021.
170. Souaille, M. and B.t. Roux, *Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations.* Computer Physics Communications, 2001. **135**(1): p. 40-57.
171. Kästner, J. and W. Thiel, *Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "Umbrella integration".* The Journal of Chemical Physics, 2005. **123**(14): p. 144104.
172. Grubmüller, H., *Predicting slow structural transitions in macromolecular systems: Conformational flooding.* Physical Review E, 1995. **52**(3): p. 2893-2906.
173. Banci, L., *Molecular dynamics simulations of metalloproteins.* Current Opinion in Chemical Biology, 2003. **7**(1): p. 143-149.
174. Orlowski, S. and W. Nowak, *Topology and thermodynamics of gaseous ligands diffusion paths in human neuroglobin.* Biosystems, 2008. **94**(3): p. 263-266.
175. Orlowski, S. and W. Nowak, *Locally enhanced sampling molecular dynamics study of the dioxygen transport in human cytoglobin.* Journal of Molecular Modeling, 2007. **13**(6-7): p. 715-723.
176. Boiteux, C., et al., *Ion conductance vs. pore gating and selectivity in KcsA channel: Modeling achievements and perspectives.* Journal of Molecular Modeling, 2007. **13**(6-7): p. 699-713.
177. Buyong, M. and J.L. Arnold, *Probing potential binding modes of the p53 tetramer to DNA based on the symmetries encoded in p53 response elements.* Nucleic Acids Research, 2007. **35**(22): p. 7733-7747.
178. Freddolino, P.L. and K. Schulten, *Common Structural Transitions in Explicit-Solvent Simulations of Villin Headpiece Folding.* Biophysical Journal. **97**(8): p. 2338-2347.
179. Feixas, F., et al., *Exploring the role of receptor flexibility in structure-based drug discovery.* Biophysical Chemistry, 2014. **186**: p. 31-45.
180. Le, L., et al., *Molecular Dynamics Simulations Suggest that Electrostatic Funnel Directs Binding of Tamiflu to Influenza N1 Neuraminidases.* PLoS Comput Biol, 2010. **6**(9): p. e1000939.
181. Aksimentiev, A., et al., *Insights into the Molecular Mechanism of Rotation in the Fo Sector of ATP Synthase.* Biophysical Journal. **86**(3): p. 1332-1344.
182. Vesper, M.D. and B.L. de Groot, *Collective Dynamics Underlying Allosteric Transitions in Hemoglobin.* PLoS Comput Biol, 2013. **9**(9): p. e1003232.
183. Kassler, K., A.C. Horn, and H. Sticht, *Effect of pathogenic mutations on the structure and dynamics of Alzheimer's Aβ42-amyloid oligomers.* Journal of Molecular Modeling, 2010. **16**(5): p. 1011-1020.
184. Rodrigues, J.R., et al., *Potentially amyloidogenic conformational intermediates populate the unfolding landscape of transthyretin: Insights from molecular dynamics simulations.* Protein Science, 2010. **19**(2): p. 202-219.

185. Trylska, J., et al., *HIV-1 Protease Substrate Binding and Product Release Pathways Explored with Coarse-Grained Molecular Dynamics.* Biophysical Journal. **92**(12): p. 4179-4187.
186. Durrant, J. and J.A. McCammon, *Molecular dynamics simulations and drug discovery.* BMC Biology, 2011. **9**(1): p. 71.
187. Liu, W., et al., *Accelerating molecular dynamics simulations using Graphics Processing Units with CUDA.* Comput Phys Commun, 2008. **179**: p. 634 - 641.
188. Shaw, D., et al., *Atomic-level characterization of the structural dynamics of proteins.* Science, 2010. **330**: p. 341 - 346.
189. Shan, Y., et al., *How does a drug molecule find its target binding site?* J Am Chem Soc, 2011. **133**: p. 9181 - 9183.
190. Van Aalten, D.M.F., et al., *A comparison of techniques for calculating protein essential dynamics.* Journal of Computational Chemistry, 1997. **18**(2): p. 169-181.
191. Pearson, K., *{On lines and planes of closest fit to systems of points in space}.* Philosophical Magazine, 1901. **2**(6): p. 559-572.
192. *Analysis of a complex of statistical variables into principal components*, 1933, Warwick & York: US. p. 417-441.
193. Leach, A.R., *Molecular Modelling*, in *Molecular Modelling, principles and applications*. 2001, Pearson Education Limited: Dorchester, Dorset.
194. Berendsen, H.J.C., Postma, J. P. M. , van Gunsteren, W. F. , Dinola, A. , Haak, J. R., *Molecular dynamics with coupling to an external bath.* Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
195. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
196. Doruker, P., A.R. Atilgan, and I. Bahar, *Dynamics of proteins predicted by molecular simulations and analytical approaches: Application to α-amylase inhibitor.* Proteins: Structure, Function and Genetics, 2000. **40**(3): p. 512-524.
197. Berendsen, H.J.C., D. van der Spoel, and R. van Drunen, *GROMACS: A message-passing parallel molecular dynamics implementation.* Computer Physics Communications, 1995. **91**(1–3): p. 43-56.
198. van der Spoel, D., et al., *Gromacs User Manual*. 2010.
199. Hayward, S., A. Kitao, and N. Go, *Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis.* Proteins: Structure, Function and Genetics, 1995. **23**(2): p. 177-186.
200. Amadei, A., M.A. Ceruso, and A. Di Nola, *On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations.* Proteins: Structure, Function, and Bioinformatics, 1999. **36**(4): p. 419-424.
201. Hess, B., *Convergence of sampling in protein simulations.* Physical Review E, 2002. **65**(3): p. 031910.
202. *<GROMACS=manual-4.5.4.pdf>.*
203. Fabrigar, L.R., et al., *Evaluating the use of exploratory factor analysis in psychological research.* Psychological Methods, 1999. **4**(3): p. 272-299.
204. Hayton, J.C., D.G. Allen, and V. Scarpello, *Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis.* Organizational Research Methods, 2004. **7**(2): p. 191-205.
205. *Comparison of five rules for determining the number of components to retain*, 1986, American Psychological Association: US. p. 432-442.
206. Yeomans, K.A. and P.A. Golder, *The Guttman-Kaiser Criterion as a Predictor of the Number of Common Factors.* Journal of the Royal Statistical Society. Series D (The Statistician), 1982. **31**(3): p. 221-229.
207. Cattell, R.B., *The Scree Test For The Number Of Factors.* Multivariate Behavioral Research, 1966. **1**(2): p. 245-276.

208. Velicer, W., *Determining the number of components from the matrix of partial correlations.* Psychometrika, 1976. **41**(3): p. 321-327.
209. Liu, Y.-S., Y. Fang, and K. Ramani, *Using least median of squares for structural superposition of flexible proteins.* BMC Bioinformatics, 2009. **10**(1): p. 29.
210. Gapsys, V., D. Seeliger, and B.L. de Groot, *New Soft-Core Potential Function for Molecular Dynamics Based Alchemical Free Energy Calculations.* Journal of Chemical Theory and Computation, 2012. **8**(7): p. 2373-2382.
211. Amadei, A., A.B.M. Linssen, and H.J.C. Berendsen, *Essential dynamics of proteins.* Proteins: Structure, Function, and Bioinformatics, 1993. **17**(4): p. 412-425.
212. Ringnér, M., *What is principal component analysis?* Nature biotechnology, 2008. **26**(3): p. 303-304.
213. Alter, O., P.O. Brown, and D. Botstein, Proc. Natl. Acad. Sci. USA, 2000. **97**: p. 10101-10106.
214. Jolliffe, I.T., *Principal Component Analysis*, 2002, Nature Publishing Group.
215. Khan, J., Nat. Med., 2001. **7**: p. 673-679.
216. Holter, N.S., Proc. Natl. Acad. Sci. USA, 2000. **97**: p. 8409-8414.
217. Nielsen, T.O., Lancet, 2002. **359**: p. 1301-1307.
218. Li, C.M. and R.R. Klevecz, Proc. Natl. Acad. Sci. USA, 2006. **103**: p. 16254-16259.
219. Mika, S., et al., *Kernel PCA and de-noising in feature spaces*, in *Proceedings of the 1998 conference on Advances in neural information processing systems II*1999, MIT Press. p. 536-542.
220. Morgan, H.P., et al., *Allosteric mechanism of pyruvate kinase from Leishmania mexicana uses a rock and lock model.* J Biol Chem, 2010. **285**(17): p. 12892-8.
221. Daura, X., et al., *Reversible peptide folding in solution by molecular dynamics simulation.* Journal of Molecular Biology, 1998. **280**(5): p. 925-932.
222. Daura, X., W.F. van Gunsteren, and A.E. Mark, *Folding–unfolding thermodynamics of a β-heptapeptide from equilibrium simulations.* Proteins: Structure, Function, and Bioinformatics, 1999. **34**(3): p. 269-280.
223. Zagrovic, B., E.J. Sorin, and V. Pande, *β-hairpin folding simulations in atomistic detail using an implicit solvent model.* Journal of Molecular Biology, 2001. **313**(1): p. 151-169.
224. Zagrovic, B., et al., *Simulation of Folding of a Small Alpha-helical Protein in Atomistic Detail using Worldwide-distributed Computing.* Journal of Molecular Biology, 2002. **323**(5): p. 927-937.
225. Yang, J.S., et al., *All-Atom Ab Initio Folding of a Diverse Set of Proteins.* Structure, 2007. **15**(1): p. 53-63.
226. Verma, A. and W. Wenzel, *A Free-Energy Approach for All-Atom Protein Simulation.* Biophysical Journal, 2009. **96**(9): p. 3483-3494.
227. Schueler-Furman, O., et al., *Progress in Modeling of Protein Structures and Interactions.* Science, 2005. **310**(5748): p. 638-642.
228. Rangwala, H. and G. Karypis, *fRMSDPred: Predicting local RMSD between structural fragments using sequence information.* Proteins: Structure, Function, and Bioinformatics, 2008. **72**(3): p. 1005-1018.
229. Zhang, Y., *Progress and challenges in protein structure prediction.* Current Opinion in Structural Biology, 2008. **18**(3): p. 342-348.
230. Andrec, M., et al., *A large data set comparison of protein structures determined by crystallography and NMR: Statistical test for structural differences and the effect of crystal packing.* Proteins: Structure, Function, and Bioinformatics, 2007. **69**(3): p. 449-465.
231. Saccenti, E. and A. Rosato, *The war of tools: how can NMR spectroscopists detect errors in their structures?* Journal of Biomolecular NMR, 2008. **40**(4): p. 251-261.
232. Sullivan, D.C. and I.D. Kuntz, *Conformation spaces of proteins.* Proteins: Structure, Function, and Bioinformatics, 2001. **42**(4): p. 495-511.

233. Müller, C.L., et al., *In the eye of the beholder: Inhomogeneous distribution of high-resolution shapes within the random-walk ensemble.* The Journal of Chemical Physics, 2009. **130**(21): p. -.
234. B.T.M. Willis, A.W.P., *Thermal Vibrations in Crystallography* 1ed. 1975, Cambridge: CAMBRIDGE UNIVERSITY PRESS.
235. Benkovic, S.J. and S. Hammes-Schiffer, *A Perspective on Enzyme Catalysis.* Science, 2003. **301**(5637): p. 1196-1202.
236. Onuchic, J.N., Z. Luthey-Schulten, and P.G. Wolynes, *THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective.* Annual Review of Physical Chemistry, 1997. **48**(1): p. 545-600.
237. Lee, A.L. and A.J. Wand, *Microscopic origins of entropy, heat capacity and the glass transition in proteins.* Nature, 2001. **411**(6836): p. 501-504.
238. Forman-Kay, J.D., *The 'dynamics' in the thermodynamics of binding.* Nat Struct Mol Biol, 1999. **6**(12): p. 1086-1087.
239. Agarwal, P.K., et al., *Network of coupled promoting motions in enzyme catalysis.* Proceedings of the National Academy of Sciences, 2002. **99**(5): p. 2794-2799.
240. Brooks, B.R., et al., *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.
241. Waight, A.B., J. Love, and D.-N. Wang, *Structure and mechanism of a pentameric formate channel.* Nat Struct Mol Biol, 2010. **17**(1): p. 31-37.
242. Ichiye, T. and M. Karplus, *Collective Motions in Proteins - a Covariance Analysis of Atomic Fluctuations in Molecular-Dynamics and Normal Mode Simulations.* Proteins-Structure Function and Genetics, 1991. **11**(3): p. 205-217.
243. Lindahl, E., B. Hess, and D. van der Spoel, *GROMACS 3.0: a package for molecular simulation and trajectory analysis.* Journal of Molecular Modeling, 2001. **7**(8): p. 306-317.
244. Lange, O.F. and H. Grubmuller, *Full correlation analysis of conformational protein dynamics.* Proteins, 2008. **70**(4): p. 1294-312.
245. Edinburgh Parallel Computing Centre. *HECToR: UK National Supercomputing Service*. 2007-2014 22-05-2014]; Available from: http://www.hector.ac.uk/.
246. Edinburgh Parallel Computing Centre. *ARCHER*. 2014 22-05-2014]; Available from: http://www.archer.ac.uk/.
247. Gupta, V. and R. Bamezai, *Human pyruvate kinase M2: a multifunctional protein.* Protein Sci, 2010. **19**: p. 2031 - 2044.
248. Fothergill-Gilmore, L.A. and P.A.M. Michels, *Evolution of glycolysis.* Progress in Biophysics and Molecular Biology, 1993. **59**(2): p. 105-235.
249. Ikeda, Y., T. Tanaka, and T. Noguchi, *Conversion of Non-allosteric Pyruvate Kinase Isozyme into an Allosteric Enzyme by a Single Amino Acid Substitution.* Journal of Biological Chemistry, 1997. **272**(33): p. 20495-20501.
250. S, W., et al., Biochemistry, 1996. **35**: p. 691.
251. Larsen, T.M., et al., Biochemistry, 1994. **33**: p. 6301.
252. Jurica, M.S., et al., *The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate.* Structure, 1998. **6**(2): p. 195-210.
253. Mattevi, A., et al., *Crystal structure of Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition.* Structure, 1995. **3**(7): p. 729-741.
254. Rigden, D.J., et al., *The structure of pyruvate kinase from Leishmania mexicana reveals details of the allosteric transition and unusual effector specificity.* Journal of Molecular Biology, 1999. **291**(3): p. 615-635.
255. Stuart, D.I., et al., *Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 Å.* Journal of Molecular Biology, 1979. **134**(1): p. 109-142.
256. Bezares, G., et al., *Isolation and sequence determination of an active site peptide of rabbit muscle pyruvate kinase.* Archives of Biochemistry and Biophysics, 1987. **253**(1): p. 133-137.

257. Enriqueta Muñoz, M. and E. Ponce, *Pyruvate kinase: current status of regulatory and functional properties.* Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, 2003. **135**(2): p. 197-218.
258. Mattevi, A., M. Bolognesi, and G. Valentini, *The allosteric regulation of pyruvate kinase.* FEBS Letters, 1996. **389**(1): p. 15-19.
259. Anastasiou, D., et al., *Inhibition of pyruvate kinase M2 by reactive oxygen species contributes to cellular antioxidant responses.* Science, 2011. **334**: p. 1278 - 1283.
260. Valentini, G., et al., *The Allosteric Regulation of Pyruvate Kinase: A SITE-DIRECTED MUTAGENESIS STUDY.* Journal of Biological Chemistry, 2000. **275**(24): p. 18145-18152.
261. Van Schaftingen, E., F.R. Opperdoes, and H.-G. Hers, *Stimulation of Trypanosoma brucei pyruvate kinase by fructose 2,6-bisphosphate.* European Journal of Biochemistry, 1985. **153**(2): p. 403-406.
262. Zanella, A., et al., *Red cell pyruvate kinase deficiency: molecular and clinical aspects.* Br J Haematol, 2005. **130**(1): p. 11-25.
263. Tanaka, K., et al., *Molecular Cloning of the Genes for Pyruvate Kinase of Two Bacilli, Bacillus psychrophilus and Bacillus licheniformis, and Comparison of the Properties of the Enzymes Produced in Escherichia coli.* Bioscience, Biotechnology, and Biochemistry, 1995. **59**(8): p. 1536-1542.
264. Noguchi, T., H. Inoue, and T. Tanaka, *The M1- and M2-type isozymes of rat pyruvate kinase are produced from the same gene by alternative RNA splicing.* Journal of Biological Chemistry, 1986. **261**(29): p. 13807-12.
265. Waygood, E.B., M.K. Rayman, and B.D. Sanwal, *The Control of Pyruvate Kinases of Escherichia coli. II. Effectors and Regulatory Properties of the Enzyme Activated by Ribose 5-Phosphate.* Canadian Journal of Biochemistry, 1975. **53**(4): p. 444-454.
266. Albert, M.-A., et al., *Experimental and in Silico Analyses of Glycolytic Flux Control in Bloodstream Form Trypanosoma brucei.* Journal of Biological Chemistry, 2005. **280**(31): p. 28306-28315.
267. Bakker, B.M., et al., *What Controls Glycolysis in Bloodstream Form Trypanosoma brucei?* Journal of Biological Chemistry, 1999. **274**(21): p. 14551-14559.
268. Wooll, J.O., et al., *Structural and Functional Linkages Between Subunit Interfaces in Mammalian Pyruvate Kinase.* Journal of Molecular Biology, 2001. **312**(3): p. 525-540.
269. Hess, B., et al., *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 435-447.
270. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(8): p. 1950-1958.
271. Miyamoto, S. and P.A. Kollman, *Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models.* Journal of Computational Chemistry, 1992. **13**(8): p. 952-962.
272. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry, 1997. **18**(12): p. 1463-1472.
273. Parrinello, M. and A. Rahman, *Polymorphic transitions in single crystals: A new molecular dynamics method.* Journal of Applied Physics, 1981. **52**(12): p. 7182-7190.
274. Doster, W., S. Cusack, and W. Perry, *Dynamical transition of myoglobin revealed by inelastic neutron scattering.* Nature, 1989. **337**: p. 754-756.
275. Goldanskii, V.I. and Y.F. Krupyanskii, *Protein and protein-bound water dynamics studied by Rayleigh scattering of Mossbauer radiation (RSMR).* Quarterly Reviews of Biophysics, 1989. **22**(1): p. 39-92.
276. Crippen, G.M. and I.D. Kuntz, *Directional structural features of globular proteins.* Journal of Theoretical Biology, 1977. **66**(1): p. 47-61.
277. Levitt, M., J. Mol. Biol., 1976. **104**: p. 59.

278. Levy, R.M., et al., *Quasi-harmonic method for studying very low frequency modes in proteins.* Biopolymers, 1984. **23**(6): p. 1099-1112.
279. van Aalten, D.M.F., et al., *The essential dynamics of thermolysin: Confirmation of the hinge-bending motion and comparison of simulations in vacuum and water.* Proteins: Structure, Function, and Bioinformatics, 1995. **22**(1): p. 45-54.
280. Amadei, A., A. Linssen, and H. Berendsen, Proteins, 1993. **17**: p. 412.
281. Wlodek, S.T., et al., *Molecular Dynamics of Acetylcholinesterase Dimer Complexed with Tacrine.* Journal of the American Chemical Society, 1997. **119**(40): p. 9513-9522.
282. Arora, K. and T. Schlick, *In Silico Evidence for DNA Polymerase-β's Substrate-Induced Conformational Change.* Biophysical Journal. **87**(5): p. 3088-3099.
283. Amadei, A., et al., *A kinetic model for the internal motions of proteins: Diffusion between multiple harmonic wells.* Proteins: Structure, Function, and Bioinformatics, 1999. **35**(3): p. 283-292.
284. Kitao, A., S. Hayward, and N. Go, *Energy landscape of a native protein: Jumping-among-minima model.* Proteins: Structure, Function, and Bioinformatics, 1998. **33**(4): p. 496-517.
285. Wieligmann, K., L.F.P. De Castro, and M. Zacharias, *Molecular Dynamics Simulations on the Free and Complexed N-Terminal SH2 Domain of SHP-2.* In Silico Biology, 2002. **2**(3): p. 305-311.
286. Solène Grosdidier, L.R.C., Víctor Buzón, Greg Brooke, Phuong Nguyen, John D. Baxter, Charlotte Bevan, Paul Webb, Eva Estébanez-Perpiñá, and Juan Fernández-Recio, *Allosteric Conversation in the Androgen Receptor Ligand-Binding Domain Surfaces.* Molecular Endocrinology, 2012. **26**(7): p. 1078-1090.
287. Chung-Jung, T. and N. Ruth, *A Unified View of "How Allostery Works".* PLoS Computational Biology, 2014. **10**(2).
288. McNae, I.W., et al., *The Crystal Structure of ATP-bound Phosphofructokinase from Trypanosoma brucei Reveals Conformational Transitions Different from those of Other Phosphofructokinases.* Journal of Molecular Biology, 2009. **385**(5): p. 1519-1533.
289. Martinez-Oyanedel, J., et al., *The First Crystal Structure of Phosphofructokinase from a Eukaryote: Trypanosoma brucei.* Journal of Molecular Biology, 2007. **366**(4): p. 1185-1198.
290. ter Kuile, B.H. and H.V. Westerhoff, *Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway.* FEBS Letters. **500**(3): p. 169-171.
291. Alves, A.M.C.R., et al., *Different Physiological Roles of ATP- and PPi-Dependent Phosphofructokinase Isoenzymes in the Methylotrophic Actinomycete Amycolatopsis methanolica.* Journal of Bacteriology, 2001. **183**(24): p. 7231-7240.
292. Bakker, B.M., et al., *Compartmentation protects trypanosomes from the dangerous design of glycolysis.* Proceedings of the National Academy of Sciences, 2000. **97**(5): p. 2087-2092.
293. Michels, P.A.M., et al., *Metabolic functions of glycosomes in trypanosomatids.* Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 2006. **1763**(12): p. 1463-1477.
294. Michels, P.A.M., et al., *The Glycosomal ATP-Dependent Phosphofructokinase of Trypanosoma Brucei must have Evolved from an Ancestral Pyrophosphate-Dependent Enzyme.* European Journal of Biochemistry, 1997. **250**(3): p. 698-704.
295. Nwagwu, M. and F.R. Opperdoes, *Regulation of glycolysis in Trypanosoma brucei: hexokinase and phosphofructokinase activity.* Acta Tropica, 1982. **39**(1): p. 61-72.
296. Cronin, C.N. and K.F. Tipton, *Purification and regulatory properties of phosphofructokinase from Trypanosoma (Trypanozoon) brucei brucei.* Biochemical Journal, 1985. **227**(1): p. 113-124.
297. Evans, P.R., G.W. Farrants, and P.J. Hudson, *Phosphofructokinase: structure and control.* Philosophical transactions of the Royal Society of London. Series B: Biological sciences, 1981. **293**(1063): p. 53-62.

298. Paricharttanakul, N.M., et al., *Kinetic and structural characterization of phosphofructokinase from Lactobacillus bulgaricus.* Biochemistry, 2005. **44**(46): p. 15280-15286.
299. Moore, S.A., et al., *The structure of a pyrophosphate-dependent phosphofructokinase from the lyme disease spirochete Borrelia burgdorferi.* Structure, 2002. **10**(5): p. 659-671.
300. Frembgen-Kesner, T. and A.H. Elcock, *Computational Sampling of a Cryptic Drug Binding Site in a Protein Receptor: Explicit Solvent Molecular Dynamics and Inhibitor Docking to p38 MAP Kinase.* Journal of Molecular Biology, 2006. **359**(1): p. 202-214.
301. Schames, J., et al., *Discovery of a novel binding trench in HIV integrase.* J Med Chem, 2004. **47**: p. 1879 - 1881.
302. Summa, V., et al., *Discovery of Raltegravir, a Potent, Selective Orally Bioavailable HIV-Integrase Inhibitor for the Treatment of HIV-AIDS Infection.* Journal of Medicinal Chemistry, 2008. **51**(18): p. 5843-5855.
303. Wassman, C.D., et al., *Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53.* Nat Commun, 2013. **4**: p. 1407.
304. A., I. and J.A. McCammon, *Mapping the Druggable Allosteric Space of G-Protein Coupled Receptors: a Fragment-Based Molecular Dynamics Approach.* Chem. Biol. Drug Design, 2010. **76**.