TARGETED CONFORMATIONAL TRANSITION PATHWAYS OF LARGE PROTEINS BY INTEGRATING ELASTIC NETWORK MODELING WITH MONTE CARLO SIMULATIONS

by

Yasemin Yeşiltepe

B.S., Chemical Engineering, Boğaziçi University, 2013

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering

Boğaziçi University

2017

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisors, Prof. Pemra Doruker, Prof. Turkan Haliloğlu, and Prof. Rahmi Özışık for their aspiring guidance, everlasting support, invaluably constructive criticism, patience and encouragement on uncountable occasions during this thesis and my entire graduate education years. As my mentors, I hope they will always be present in my future work.

I would like to acknowledge the members of the thesis committee, Assoc. Prof. Burak Alakent, Assist. Assoc. Prof. Demet Akten and Assoc. Prof. Sezen Soyer Uzun for their times, valuable discussions and productive comments on this work.

I would also like to take this opportunity to express my warm thanks to Deniz Turgut, Arzu Uyar, Burak Kaynak for their help in my thesis.

Many thanks to Doğa Fındık, Muhammet Memiç, Nihat Baysal and Deniz Rende for their great friendship and making the past years enjoyable.

I would like to extend my thanks to Çağrı Yanık for his invaluable support, patience and, as in all things, his unconditional love, and always being by my side with his incredible heart, infinite warmth and understanding.

Last but not least, I would like to express my deepest love and dedicate this thesis to my family, who supported me without any doubt my whole life and assured me that I can achieve anything.

ABSTRACT

TARGETED CONFORMATIONAL TRANSITION PATHWAYS OF LARGE PROTEINS BY INTEGRATING ELASTIC NETWORK MODELING WITH MONTE CARLO SIMULATIONS

ANM-MC is a coarse-grained simulation technique, in which the anisotropic network model (ANM) generates the collective modes for deforming the protein structure towards target direction, and Monte Carlo (MC) algorithm minimizes the energy of the deformed structure. ANM-MC was modified in this thesis for the exploration of conformational transition pathways of large protein systems. The analyzed dataset consists of 13 large proteins that present either hinge-bending, DNA-binding or shear-type conformational transitions. All proteins contain multiple chains with 600 to 8000 residues in total. The distance (RMSD) between initial and final conformations spans a broad range of 2.1 - 20.6 Å. Initially, the adjustment of ANM-MC parameters to large proteins was performed using adenylate kinase and calmodulin. Later, detection of local implausible geometries for some large proteins lead to a modified version of ANM-MC with variable ANM deformations. As a result, all the proteins in the dataset approach the target structures successfully following suitable potential energy paths. Some of the proteins need more than the slowest 10 modes to approach the target better, specifically 50 to 150 modes may be necessary. For such proteins, relatively lower indexed modes still play a dominant role during initial stages of the simulation with higher indexed modes chosen in later stages. Furthermore, coarse-grained intermediates along the transition pathway are reverse-mapped to full-atomistic structures by energy minimization. ANM-MC program is highly efficient based on the reported computational cost for each protein in the dataset.

ÖZET

KOLLEKTİF AĞ MODELLEMESİ VE MONTE CARLO SİMULASYONLARI İLE BUYUK PROTEİNLERİN HEDEFLİ KONFORMASYONEL GEÇİŞLERİ

Kaba ölçekli bir simülasyon tekniği olan ANM-MC'de anizotropik ağ yapı modeli (ANM) protein yapılarını hedefe doğru deforme etmek için gerekli kolektif modları sağlar ve Monte Carlo (MC) tekniği ise deforme edilen yapının enerjisini azaltır. ANM-MC bu tezde büyük proteinlerin kon-formasyonel geçiş yollarının incelenmesi amacı ile düzenlenmiştir. İncelenen veri kümesi, DNA'ya bağlanan, menteşe ve kayma türü hareket yapan 13 büyük proteinlerden oluşmaktadır. Bütün proteinler çoklu zincirli olup toplam rezidü sayıları 600 ile 8000 arasında değişmektedir. Başlangıç ve hedef deneysel yapıların arasındaki kök ortalama kare sapma değerleri 2.1 - 20.6 Å aralığındadır. ANM-MC parametrelerinin büyük proteinlere uygulanabilmesi için, adenilat kinaz and kalmodulin üzerinde bir ön calışma yapılmıştır. Daha sonra, bazı büyük proteinlerin geome-trilerinde makul olmayan yerel deformasyonların saptanmasıyla, değişken ANM deformasyonlarının kullanıldığı yeni bir ANM-MC versiyonu önerilmiştir. Bunun sonucunda verisetindeki bütün proteinler hedef yapılarına başarılı bir şekilde ulaşmıştır. Bazı proteinler hedefe ulaşmak için en yavaş 10 moddan fazlasına, özellikle 50 ile 150 arasındaki modlara ihtiyaç duymuştur. Bu proteinler için, nispeten daha düşük endeksli olan modlar simulasyonun başlangıc aşamasında baskın bir rol oynamış, ilerleyen aşamada ise yerini yüksek indeksli modlara bırakmıştır. Ayrıca, geçiş yolu üzerindeki kaba ölçekli yapılar, enerji minimizasyonu ile atomistik detayda yapılara geri dönüştürülmüştür. Verisetindeki her bir proteinin hesaplama maliyeti rapor edilerek, ANM-MC programının verimliliği gösterilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Å	Angstrom
C^{α}	Alpha-carbon atom
$\mathbf{H}$	Hessian matrix
$\mathbf{H}_{ij}$	Super element matrix of the hessian matrix
$h(x)$	Heavyside step function
k_B	Boltzmann's constant
l_i	Bond length connecting atoms i-1 and I
l_i^S	Bond length connecting backbone and sidechain
N	Residue number
N	Total number of residues
r_c	Cutoff radius
$\boldsymbol{R}$	Rotation matrix
$\mathbf{R}_i$	Position vector of site I
$\mathbf{R}_{new}$	New coordinate matrix of the generated conformation
R_{ij}	Distance between site i and j
$R_{required}$	Desired approach value
T	Absolute temperature
$\mathbf{U}$	Eigenvector
V	Potential energy

γ	Force constant
ξ	Random number
$\Delta\mathbf{R}_i$	Fluctuation of vector of site i
$<\Delta\mathbf{R}_i^2>$	Mean-square fluctuation of site i
$\Delta\mathbf{x}_i$	Change in the x coordinate of the position vector of site i
$\Delta\mathbf{y}_i$	Change in the y coordinate of the position vector of site i
$\Delta\mathbf{z}_i$	Change in the z coordinate of the position vector of site i
θ_i	bond angle between bond l_i and l_{i+1}
θ_i^S	bond angle between l_i and l_i^S, and φ_i^S
φ_i	torsional rotation of the bond l_i
φ_i^S	torsion angle defined by l_{i-1}, l_i and l_i^S
φ_i^-	torsional angle of backbone bond preceding i^{th} alpha carbon
φ_i^+	torsional angle of backbone bond succeeding i^{th} alpha carbon
Φ	New conformation
Φ_o	Original conformation

LIST OF ACRONYMS/ABBREVIATIONS

AK	Adenylate Kinase
AMP	Adenosine monophosphate
APS	ATP Sulfurylase
ATCase	Aspartate Transcarbamylase
ATP	Adenosine triphosphate
ANM	Anisotropic Network Model
BB	Backbone-backbone
Cpn	Chaperonin Lidless Mm-cpn
CS	Citrate synthase
DF	Deformation factor
DNA	Deoxyribonucleic acid
E	Energy
ENM	Elastic Network Model
EPSPS	5-enolpyruvylshikimate-3-phosphate synthase
GluT	Glutamate Transporter
GNM	Gaussian Network Model
Hb	Hemoglobin
k	Spring constant
Lacl	Lac Repressor
LR	Long-range

MD	Molecular Dynamics
MC	Monte Carlo
MCs	Monte Carlo steps
MYO	Myosin
NMA	Normal Mode Analysis
PDB	Protein Data Bank
PS	Strength of perturbation
QacR	Hypothetical Transcriptional Regulator in QACA
R	Relaxed state
RMSD	Root-mean-square deviation
RnaP	Ribonucleic acid Polymerase
SB	Sidechain-backbone
SC	Side-chain
SR	Short-range
SS	Sidechain-sidechain
T	Tense form
UPRTase	Uracil Phosphoribosyltransferase

1. INTRODUCTION

Proteins are generally not static systems and populate ensembles of conformations. Understanding of protein dynamics requires the study of transitions between different conformational states occurring on a variety of length and time scales. Highly populated states and the transitions between them can be described by the phenomena of energy landscape, which provides the characterization and classification of protein motions. Basically, an individual protein molecule explores the energy landscape [1, 2].

Many biomolecules visit more than one conformation in order to perform their biological function. Some dynamic processes such as enzyme activity, ligand binding, and signal transduction involve essential conformational changes of entire domains. Information on conformational changes, by which the protein travels from one state to another, is essential for a complete molecular understanding of a biological process. Identification of intermediate conformations along the pathways as a function of time may elucidate the molecular mechanism of a specific biological process. Experimental techniques have capability to provide the average structures of important macromolecular systems [3, 4]. Many structures have been resolved based on the long-lived stable functional states under different conditions or with different ligands. However, experimental data mostly do not exist for the intermediate structures along the conformational transition pathway. Namely, it may not be possible to obtain a full mechanistic description to see both structures and energy at the atomic level of interactions by experiment alone even if atomic force microscopy (AFM) has a capability for such information to some extent. Moreover, the disproportion between the total number of known structures in Protein Data Bank (120 thousand proteins) [5] and in the RefSeq database (84 million sequences) [6] clearly demonstrates that experimental procedures for protein structure determination are very expensive and time-consuming. On the other hand, computational techniques are able to elucidate the intermediate transient or short-lived conformational states [7]. In other words, current experimental methods quite often seek support from accurate theoretical model predictions.

The limits of in-silica methods for determining pathways in large complex molecules could be overcome with the development of efficient methods such as path sampling; namely

the sequence of the conformational states within double-ended paths [8-11]. Simulation of some particular dynamical events can also be described by some statistical mechanical methods [12-15]. On the other hand, some other sampling methods are not utilized two known endpoints for the study of the transition pathways [16-18]. In particular, development of efficient, coarse-grained approaches is necessary for the simulation of large biological systems and/or long-time scale phenomena, such as conformational transitions.

The anisotropic network model (ANM) [19, 20] is a coarse-grained normal mode approach, in which the protein molecule is represented by a set of nodes placed at the positions of alpha carbon atoms. The energy of the protein system is described a harmonic potential calculated by summing over harmonic interactions between close-neighboring residues in the structure. ANM has been reported successful by some computational studies in reproducing collective modes that conform with the conformational transition directions [21] and predicting atomic fluctuations for even large biological systems. Doruker *et al.* indicated that the equivalent cooperative motions can be obtained upon further coarse-graining of the protein structure along the backbone, which reveals that hierarchical levels of coarse-graining considerably reduce the computational time [22]. Accordingly, very large biological systems such as GroEL-GroES complex, restriction endonuclease EcoRI-DNA complex, triosephosphate isomerase, ribosome complex, RNA polymerase have been analyzed [23-29].

A computationally efficient algorithm, ANM-MC [30], has been proposed to investigate the conformational transition pathways of proteins; which incorporates the collective deformation directions obtained from ANM into an MC simulation technique, based on the knowledge-based potentials of proteins [31-33]. ANM-MC is an iterative methodology composed of ANM and MC cycles to approach the target conformations. The normal modes are updated iteratively based on the deformed structure. At each iteration the structure is deformed along a slowest mode that overlaps best with the desired transition. Then, energy minimization is performed on the deformed structure by the MC technique. This approach can generate a feasible targeted pathway between two conformations along collective modes within short computational times. The method has been validated with its application to adenylate kinase and hemoglobin, for which a large variety of studies are available in the literature [30]. Later, ANM-MC algorithm has been used for the construction of successful path-

ways for a diverse set of proteins consisting of hinge-bending proteins, DNA-binding proteins, enzymes with loop closure, and proteins with shear-type conformational transitions [34]. Closer predictions between experimental and computational targeted closed states using the initial structure have been achieved efficiently with the improvement of ANM-MC parameters.

The construction of targeted pathways between open and closed states by ANM-MC has shown success. However, the current method needs further developments for the efficient calculations of pathways for large systems by addressing the scalability problem. In the current thesis, the ANM-MC parameters have been analyzed and further optimized to obtain plausible intermediate coarse-grained protein structures along the conformational transition pathway. Accordingly, the new efficient and optimized ANM-MC method is applied to large molecules. Finally, reverse-mapping of all intermediate/snapshots back to an atomistic system is employed for the construction of the full atomistic 3-D structure of large proteins.

In the following chapter, the experimental and computational methods are provided for conformational transition pathways of large proteins. The detail explanation of the optimization of ANM-MC method and its applications to a set of large molecules is provided in Chapter 4 and 5, and reverse mapping of their trajectories are given in Chapter 6, respectively. The conclusions and recommendations are summarized in Chapter 7.

2. LITERATURE REVIEW

2.1. Protein Structure Determination

Macromolecules are constructed by linear sequences of amino acids linked via peptide bonds by forming a polypeptide chain. Amino acids are naturally occurring with different side chains, sizes, shapes, charge distributions and hydrogen bonding capabilities which enable proteins perform many functions essential to life. The discovery of the tertiary structure of a protein chain, or the quaternary structure of its complexes give an insight about its structure, dynamics and function relationship, which may be examined in vitro, in vivo, and in silico.

2.1.1. Experimental techniques

The major experimental methods used in tertiary protein structure determination are X-ray crystallography and nuclear magnetic resonance. X-ray crystallography has made the largest contribution to the protein structure determination, which can produce information at atomic resolution. The rapid growth of protein structure data can be attributed to the technological advances in X-ray crystallography, although the solution of crystal structures of multicomponent systems and membrane proteins have been remained still challenging. Formation of crystals to collect high-resolution data for structure determination may not have sufficient quality due to the poor protein quality [35]. However, X-ray crystallography is still the most powerful technique for structure determination and analysis of proteins. In particular, globular proteins are comparatively easier to crystallize in preparation for X-ray crystallography while crystallization of membrane proteins are difficult [36]. Up to current date, nearly 75000 X-ray crystal structures of proteins have been deposited in Protein Data Bank (PDB) which are consisted of 70% of all X-ray structure of all compounds in PDB.

Alternatively, Nuclear magnetic resonance (NMR) spectroscopy is a powerful technique to determine the three-dimensional (3D) structures of biological macromolecules at atomic resolution. NMR is currently the only method that can accurately define atomic structures in solution by estimating more accurate information between pairs of atoms and providing the

final possible conformations of solvated proteins. Solved NMR structures of proteins deposited in the Protein Data Bank (PDB) obtained in the form of Cartesian coordinates for each atom in the protein [37] have reached over 10,000 protein structures containing 350 or more residues whereas the first deposited structures date back only to 1989. Another advantage of NMR spectrum is these structures have been mostly determined in solution, similar to their natural conditions. However, the degree of sensitivity of all the parameters measured by NMR spectroscopy to molecular structure and dynamics can restrain to construct models of the three dimensional structures of proteins. Besides, the NMR spectra are highly congested with the size of the protein, particularly a small protein of 100 residues containing approximately 800 protons, which makes the spectral assignment easily deformed or structure determination impracticable [38].

The other methods could be cryo-electron microscopy and small-angle X-ray scattering (SAXS). The former one is suitable to large and stable proteins like ribosomes which are braced by rigid twists of RNA Ribosomania has been focused by cryo-EM researchers in the past couple of years [39]. Dozens of cryo-EM structures of ribosomes from a multitude of organisms have been quickly published. Especially, the first high-resolution models of human ribosomes have been determined [40, 41]. As both ordered and disordered proteins in a broad range of molecular sizes are the case, SAXS comes a powerful method for the structural characterization in solution under various experimental conditions varying from extreme to nearly native. It helps to determine the sizes and shapes of proteins and complexes [42].

2.1.2. Computational techniques

Protein structure determination from amino acid sequence has been regarded as a fundamental scientific grand challenge in computational biology and chemistry. Computational techniques can attempt a prediction of a protein structure by providing a means of generating a plausible structure for proteins whose structures have not been experimentally determined [43]. Even though, over the last decades, many experimental techniques have been applied for determination of the protein structures, and consequently, the determined structures of several proteins have been added to libraries; accessing the ligand binding and conformational changes of proteins are still limited both for technical and economic reasons. In this

regard, protein libraries, i.e. Protein Data Bank, constructed by experimental/lab data are too limited, expensive, and slow to build libraries of even thousands of conformations or ligand binding proteins, which makes the understanding functions of proteins still largely unknown.

2.1.2.1. MD Based Studies. The most realistic and plausible structures of biomolecular systems are represented by full atomistic molecular dynamics (MD) simulation which could be challenging and daunting task for the exploration of large scale conformational transitions in biological macromolecules. Many methods including transition pathway sampling and the original spring method have been proposed to obtain the plausible transition pathways between stable states [9, 10, 12, 44-51]. Specifically, this string strategy has been based on the full atomistic MD simulations and applied successfully on many biological systems to explore only their particular conformational transitions [11, 52-59].

2.1.2.2. Reduced Models. The reduced or coarse-grained models are designed to study biomolecule behaviors, such as large internal motions or conformational changes in order to reduce the number of degrees of freedom by building models in which some of the atoms are ignored or groups of them are treated as united pseudo-atoms without losing its important features, such as structural details, internal dynamics and characteristic interactions. Coarse-grained model simplifies the atomistic structure and adopts a simple potential function that uses knowledge of resolved structures which makes it an uncontroversial one to escape the computational cost in protein structures. The Cα atom is explicitly defined and serves as the most important point within a residue in most cases.

2.1.2.3. Monte Carlo Energy Minimization Method. Optimization procedures are required for a comprehensive understanding of interatomic interactions leading to the most-stable conformations of a protein from a linear polypeptide chain. The energy surface is so strongly anharmonic that there are many local minima within the range of thermal fluctuations although proteins appear to have apparent dual harmonic and anharmonic aspects while they behave in a way where the energy surface are harmonic [60, 61] The existence of conformations of a protein in many local minima in the multidimensional energy surface is challenging in locating the global minimum of the empirical potential function and becomes aggravated as the size of the protein increases [62]. MD calculations are free from the assumption of harmonicity and very challenging whenever complexity appears. An efficient

global optimization procedure which is capable of finding an appropriate local minimum is required for a successful prediction of the three-dimensional structure of proteins while it is computationally inexpensive enough to be used in the search procedure, yet sufficiently accurate to ensure the uniqueness of the conformation. In this perspective, Monte Carlo method is a cost alternative method to simulate the protein dynamics free from the assumption of harmonicity. The existence of corresponding multiple conformational substates experimentally [63] shows the validation of motions associated with normal modes with medium and high frequencies with the assumption of harmonicity of the energy surface [64]. The Monte Carlo approach basically combines the power of conventional energy minimization and the Metropolis criteria [65] to find local minima.

2.1.2.4. Large Proteins. The interest in studying large proteins in detail are still a missing part in literature due to the lack of enough structure information. It is not feasible to explore the molecular motions of proteins by fully atomistic MD simulations and normal mode analysis (NMA). It appears important in order to classify repeating structural patterns of proteins and to understand the relations between structure and functions. As the coarse grained approaches have been useful in elucidating the functionally important collective motions of large proteins [66, 67], the investigation of large quaternary structures and biomolecular complexes have become possible computationally with the representation of protein structures as elastic networks [68-70]. Doruker *et al.* investigated the collective motions extracted by the coarse-grained ANM calculations performed on influenza virus hemagglutinin, a homotrimeric enzyme composed of 1509 residues as a large molecule by eliminating the particular Cα atoms along the backbone structure repeatedly on the protein pattern and successfully obtained the equivalent cooperative motions [22]. Hinsen *et al.* analyzed the domain motions of large proteins with around 2700 residues and indicated that the advantage of Normal mode analysis as providing the same domain identification and a more detailed picture of domain movements with one experimental structure.

2.1.2.5. Reverse-Mapping Studies. Thanks to the coarse grained methods, the simulation of molecular systems, and especially the exploration of the behavior of larger systems have offered steps forward in future studies; which also allows beyond the time scale readily accessible at atomistic level [71]. Although the ideal method for protein structure prediction is to search through the conformational space and identify the lowest free-energy states under

given force fields, it is usually not feasible to search even medium size proteins in full atomic representation due to the astronomical number of possible conformational states. Reconstructing 3D coordinates of every atom in protein molecules from coarse grained structures is the central theme in structural biology. However, the sufficient investigation of properties of even large length of proteins requires the preparation of full atomistic well structures obtained by computer simulations as the information on some of the atomic positions are missed depending on the degree of coarse-graining. Nevertheless, most of the protein structure prediction methods provide protein chains in a reduced model by different representations. While UNRES [72] represents a residue by three points of C-alpha (Cα), side-chain center, and a virtual peptide group, ROSETTA [73] uses backbone heavy atoms and C-beta (Cβ). TASSER [74] and I-TASSER [75] represent a residue by two points of Cα and the side-chain center of mass. In these models, the coarse-grained structural implications are lack of the properties of absolute configurations or orientations of atoms which demonstrates the dynamics of proteins in detail. Consequently, the coarse grained structures needs to be transformed into an appropriate full atomistic representation which is denoted as back mapping, inverse mapping or reverse mapping. The usual method in reverse mapping procedure is first to relax and equilibrate the coarse grained structure before performing the transformation of the structure back. Secondly, the resulting full atomistic structure is relaxed by short molecular dynamics runs or other alternative methods to correct the new structure. Needless to mention that the coarse grained schemes refer the same properties in the reverse mapping structures.

Doruker *et al.* demonstrated the successful application of reverse mapping method on specific snapshots of coarse-grained PE melt over two different lattices to obtain the fully atomistic representation [76]. There are also some online programs capable of transferring reduced models to full atomic coordinates such as PULCHRA [77], NEST [78], MAXSPROUT [79], MODELLER [80] and REMO [81].

2.1.2.6. NMA Based Techniques. A better suited technique to study large structural rearrangements than MD is Normal mode analysis (NMA). Although it is more expensive in terms of memory, it takes much less demanding in terms of CPU time. NMA is designed to investigate the vibrational motion of a harmonic oscillating system which are of small am-

plitude in a potential energy in the immediate vicinity of its equilibrium [82]. The fundamental hypothesis is that the largest and functionally relevant movements in a protein are described by the vibrational normal modes exhibiting the lowest frequencies. A variety of successful techniques have been developed with the help of NMA on proteins for the prediction of collective, large amplitude motions proteins and protein assemblies [64, 66, 83-91]. The study of the slow dynamics of biomolecules with the use of normal mode analysis have been prevalent. The applications of normal modes have been utilized on a range of small proteins to molecular machines like lysozyme, HIV1-protease, aspartate transcarbamylase, myosin, integrins, Ca-ATPase, F1-ATPase, Chaperonin GroEL, ribosome [92-105].

2.1.2.7. ENM Based Studies. The elastic network model (ENM) [106-108] has been employed broadly as a coarse grained structure based model in conjunction with NMA [85, 91, 109]. It is a simple model and method resulted which provides a satisfactory description of the correlations between atomic fluctuations within computation times at least one degree of order less than commonly used molecular approaches. ENM models basically focus on the deformation of the system along the low frequency normal modes by utilizing the Hessian matrix of the potential built with the experimental structure.

The network model is a major determinant of accessible modes of function could be concluded by the utilization of collective motions relevant to function occurring along the low energy normal modes of motions predicted by ENMs [21, 100, 102, 110-113]. ENMs coupled with NMA have explored successful conformational transitions by searching an energy minima on conformational energy landscape in a given neighborhood. Indeed, they are very precursor in the advances of proceeding methods providing plausible intermediate structures along a transition whereas they are insufficient in building a transition pathway between two states [7]. Kim and Jernigan *et al*. developed a computationally efficient method based on a coarse grained ENM for searching of the transition of a protein between two conformations very successfully and reliably; which interpolates the distances between amino acids within the range of two end structures within the defined cutoff distance [98, 114]. This study particularly indicates the application of NMA into the large systems by dividing the system into pieces and carrying out NMA calculations for each of them [115-117].

Maragakis employed the plastic network model (PNM), an extension of ENM, which uses the driven molecular dynamics methods in conjunction the minimum energy pathway within two known end structures and the intermediates structures if available. ENM is applied to the each end side of the pathway and the intermediate structures are determined for the construction of the pathway of *Escherichia coli* adenylate kinase with an energy minimization method [118]. Two ENM surfaces by using the minimum potential search on the surface to construct a pathway have also been investigated by the other studies [53, 119-121].

Bahar and Haliloglu *et al.* introduced the Gaussian Network Model (GNM) [85, 109] for investigating the dynamics of proteins and their complexes. A broad range of applications of GNM validates that it successfully provides information on the size of fluctuations away from the mean positions and on the structural elements providing adequate flexibility to enable conformational changes [122-126]. Li *et al.* serves the latest version of GNM which covers more than 95% of the structures currently available in the Protein Data Bank (PDB) with the ability of being a useful resource on their biological assemblies for establishing the bridge between structure, dynamics and function [127].

2.1.2.8. ANM Based Studies. ANM is a promising tool as a simple ENM for describing the collective dynamics of a broad range of biomolecular systems. Eyal and Bahar *et al.* examined the performance of ANM on a large set of proteins as a function of its optimal model parameters with a highest agreement with experimental data and provided and its limits of accuracy and applicability of prediction for different residue types and secondary structures. As a conclusion, better correlation is observed with increase in the resolution of the structure of interest. Additionally, residue fluctuations in globular proteins are more accurately predicted with ANM than those in nonglobular proteins [128].

Another approach, *adaptive* ANM (*a*ANM), based on ANM has been proposed and applied to the allosteric transitions of the GroEL-GroES complex. It is based on the simultaneous generation of pairs of intermediate conformations, starting from the known endpoints to create a series of intermediate conformers iteratively until the two intermediates are close enough within a predefined root-mean-square-deviation (RMSD) [28].

Das *et al.* has proposed a simple and efficient method, ANMPathway, to investigate gives physically meaningful conformational transitions of globular and membrane proteins. Similarly, each of the end-states are represented by ENM to construct two-state potential by mixing these two ENMs. The transition state is searched as a minimum energy structure and two separate steepest descent minimizations are performed to connect the end-states along with the transition state or a constructed pathway [7].

Eyal *et al.* introduced a user friendly ANM server with an indentation to serve the researchers with little background in computational biology which requires only the PDB file of the structure of interest as an input and gives the output structures with collective modes with the option of controlling model parameters [128]. The same group presented an improved version including upgraded functionals and capabilities by providing protein structures with the other extension parts or namely biomolecular complexes and assemblies with DNA, RNA and ligands [129].

Kantarci *et al.* introduced ANM-MC methodology in which collective modes are obtained from ANM and combined with MC simulations for energy minimization and produce conformational transition pathways as an output. This methodology was carried out on AK and hemoglobin structures successfully. Uyar *et al.* developed the ANM-MC algorithm for the successful transition pathways for a set of 26 proteins with different protein dynamics features. Then they applied ANM-MC technique to a membrane protein with set of modeled structures of human beta-2-adrenergic receptor for the exploration of conformational transition between active and inactive states [130].

Additionally, Uyar *et al.* developed a further technique Rg-ANM-MC to predict possible unknown conformations of closed/bound structures [34] and even utilize Rg as a constraint in the ANM-MC technique to predict apo/unbound form of the protein of homodimer p50 nuclear factor-kappa B (NF-κB) [131].

2.1.2.9. Other Studies. One of the different performance of successive normal mode calculations by moving a stable initial structure along the intermediate structures with the overlap best with the target structure with the help of reparameterization of the network along the pathway [132].

One of the different approaches termed as tCONCOORD method, a reimplementation of the CONCOORD approach, is proposed which provides fast and accurate prediction of protein flexibility on experimentally known conformational transitions by allowing a computationally efficient sampling of conformational transitions of a protein based on geometrical considerations [133].

Additional methods such as the weighted ensemble method [103, 134-136] and the dynamic importance sampling [137-139] could be noteworthy for the determination of conformational transitions. Huber *et al.* [136] introduced a path sampling method guaranteed to controlled binding of a ligand to a receptor statistically, which was developed further for the configurations coarse-grained folded proteins [140]. Zhang *et al.* demonstrated later that the method based on standard Brownian dynamics can also statistically comprise the Markovian and non-Markovian dynamics. The application of the similar strategy is later preferred in the "forward flux" sampling [141, 142].

3. MATERIALS AND METHODS

3.1. Anisotropic Network Model

The Anisotropic Network Model (ANM) is a simple yet powerful tool introduced by Atılgan *et al.* [19] and Doruker *et al.* [22]. ANM provides the relation between function and intrinsic dynamics of biological macromolecules rooted in Normal Mode Analysis. It is basically an Elastic Network Model in which the protein is represented as an elastic mass-and-spring network, where the interaction among the Cα atoms of residues are described by harmonic springs.

The network of Cα atom interactions is generated within a predetermined cutoff distance and defined with the Force constant matrix, Hessian matrix (*H*). *H* is 3Nx3N matrix composing of super elements represented as

$$H = \begin{bmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,N} \\ H_{2,1} & H_{2,2} & & H_{2,N} \\ \vdots & & \ddots & \vdots \\ H_{N,1} & H_{N,2} & \cdots & H_{N,N} \end{bmatrix} \tag{3.1}$$

where the off-diagonal super elements are,

$$H_{i,j} = \begin{bmatrix} \frac{\partial^2 V}{\partial X_i \partial X_j} & \frac{\partial^2 V}{\partial X_i \partial Y_j} & \frac{\partial^2 V}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Y_i \partial X_j} & \frac{\partial^2 V}{\partial Y_i \partial Y_j} & \frac{\partial^2 V}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Z_i \partial X_j} & \frac{\partial^2 V}{\partial Z_i \partial Y_j} & \frac{\partial^2 V}{\partial Z_i \partial Z_j} \end{bmatrix} \tag{3.2}$$

In practice the off-diagonal super elements are calculated by

$$H_{i,j} = \begin{bmatrix} -\gamma\frac{(X_j - X_i)(X_j - X_\iota)}{R_{ij}^2} & -\gamma\frac{(X_j - X_i)(Y_j - Y_\iota)}{R_{ij}^2} & -\gamma\frac{(X_j - X_i)(Z_j - Z_\iota)}{R_{ij}^2} \\ -\gamma\frac{(Y_j - Y_i)(X_j - X_\iota)}{R_{ij}^2} & -\gamma\frac{(Y_j - Y_i)(Y_j - Y_\iota)}{R_{ij}^2} & -\gamma\frac{(Y_j - Y_i)(Z_j - Z_\iota)}{R_{ij}^2} \\ -\gamma\frac{(Z_j - Z_i)(X_j - X_\iota)}{R_{ij}^2} & -\gamma\frac{(Z_j - Z_i)(Y_j - Y_\iota)}{R_{ij}^2} & -\gamma\frac{(Z_j - Z_i)(Z_j - Z_\iota)}{R_{ij}^2} \end{bmatrix} \quad (3.3)$$

where R_{ij} is the distance between sites i and j, γ is the universal force constant. H is the second derivate matrix of the overall potential function which is given as

$$V = \left(\frac{\gamma}{2}\right)\sum_i\sum_j h(r_c - R_{ij})(\Delta\boldsymbol{R}_j - \Delta\boldsymbol{R}_i)^2 \quad (3.4)$$

where h(x) is the heavyside step function (h(x)=1 if x ≥ 0, and zero otherwise) and ΔR_i is the fluctuation in the position vector R_i of site i ($1 \leq i \leq$ N). It can also be represented in a matrix form as:

$$V = (^1/_2)\Delta\boldsymbol{R}^T\boldsymbol{H}\Delta\boldsymbol{R} \quad (3.5)$$

where $\Delta\boldsymbol{R}$ is position fluctuations vector.

Eigenvalues and eigenvectors of the Hessian matrix are extracted to obtain modes, namely motions of the protein. The decomposition of Hessian matrix yields 3N-6 non-zero eigenvalues where the 6 zero eigenvalues indicate three translation and three rotation motions of the protein (i.e. rigid body motion). After the 6 zero eigenvalues are discarded, the corresponding eigenvectors of remaining (non-zero) eigenvalues determine the respective frequencies and deformation directions of the motions, called modes.

The cutoff distance defines the maximum pairwise interaction range between Cα atoms. It is the only predetermined parameter in the model and usually differs from 5 to 30 Å. Atilgan *et al.* [19] shows that ANM uses a cutoff value of 18.0 Å after calibration with experimental fluctuations. Eyal *et al.* examined the performance of the model on a single pa-

rameter (i.e. cut-off) and concluded that the large cutoff in ANM does not mask the topological differences between residues [128]. More realistic anisotropic displacement parameters are obtained using lower cut-offs in another study of Eyal *et al*. [143]. Globular proteins are defined best for cut-offs of 15-21 Å on average whereas individual proteins are better with smaller cut-offs of 12-15 Å [144]. Uyar *et al.* benchmarked cutoff values on the dataset diverse in terms of folds and number of domains and recommended the cut-off value of 10 Å for ANM [34]. Besides, the sparsity of Hessian matrix increases with smaller cut-offs which offers higher performance and lower memory usage. Considering both theoretical and computational aspects, the value of 10 Å for distance cut-off is chosen in this thesis.

3.2. Monte Carlo (MC) Simulation

The Monte Carlo simulation/Metropolis technique is a simplified and an efficient yet realistic simulation method developed to simulate biological macromolecules [32]. It is an energy function based on a reduced model for protein structures. Its predictions has been shown to be in agreement with residue order parameters [31] and hydrogen exchange data [33] from NMR measurements.

In this present study, this coarse grained MC technique is utilized to minimize the energy of the protein structure after the perturbations along the normal modes. In this reduced coarse-grained model, the protein structure is represented by two interaction sites, Cα atom and side chain centroid (SC) for each residue (i.e. only Cα atom is taken into account for Glycine amino acid). A schematic representation of a protein segment between C_{i-2}^{α} and C_{i+1}^{α} is given in Figure 3.1. Needless to mention, for a protein with N residues, the conformation of the backbone has 3N-6 variables including N-1 backbone virtual bonds l_i, N-2 bond angles θ_i, and N-3 torsional angles ϕ_i. Sidechain attached to the C_i^{α} is designated as Si. l_i is the ith virtual bond extending from C_{i-1}^{α} to C_i^{α}. ϕ_i is the rotational angle of the bond l_i defined by the respective locations of the four backbone units C_{i-2}^{α}, C_{i-1}^{α}, C_i^{α} and C_{i+1}^{α}. θ_i is the bond angle between bonds l_i and l_{i+1}.

The conformational space of a residue in the available template coarse grained structure is predicted by the energy of two effective interaction sites (Cα atom and SC) of a residue evaluated depending on the distance in between, and the type, of amino acid that the

sites belong to. The interresidue interaction energy based on distance-dependent potentials are taken from Bahar and Jernigan *et al.* [145] obtained by the derivation of 302 reference structures at different environments [146, 147]. These knowledge based potentials are not fit to functions, and are discrete instead, at 0.4 Å resolution (i.e: mesh sizes are taken into account for the determination of background probabilities in the discrete state formalism). The set of parameters are driven from the potentials with these reference states. Besides Bahar and Jernigan *et al.* presents effective contact potentials obtained from the integration of radial distributions over different distance ranges [148]. The overall interaction energy is obtained by the summation of the interaction energies over N residues as

$$E_{LR}(\Phi) = \sum_{i=1}^{N-3} \sum_{j=1+3}^{N} E_{SS}(r_{ij}) + \sum_{i=1}^{N-4} \sum_{j=1+4}^{N} E_{SB}(r_{ij}) + \sum_{i=1}^{N-5} \sum_{j=1+5}^{N} E_{BB}(r_{ij}) \qquad (3.6)$$

where SS, SB, and BB are the potentials between side-chain sites, side-chain and backbone sites, and backbone sites and r_{ij} is the distance between the residues i and j in conformation Φ, respectively.

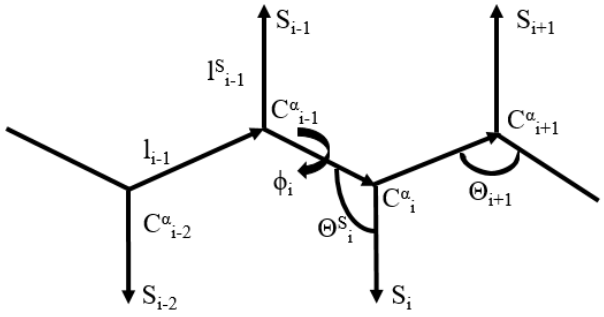


Figure 3.1. Schematic representation of the virtual bond model.

The conformations of the individual residues, both Cα atoms and SCs, contribute to the total energy of the virtual model system. Formerly, the conformational energy of Cα atoms are formed due to the changes in bond angle and bond torsions, and expressed as statistical potential [149]. Specifically, the bending of bond angles and rotational angles Φ_i- and Φi^+ of the virtual backbone bonds and the pairwise interdependence of these torsion

and/or bond angle bending. Latterly, the short-range potential energy of backbone and side chains is expressed as based on the virtual bond model given by Bahar *et al.* [149] as:

$$E_{SR}(\Phi) = \sum_{i=2}^{N} E(l_i) + \sum_{i=2}^{N-1} E(\theta_i) + \sum_{i=3}^{N-1} \left[\frac{E(\phi_i^-)}{2} + \frac{E(\phi_i^+)}{2} + \Delta E(\phi_i^-, \phi_i^+) \right] + \sum_{i=3}^{N-1} [\Delta E(\theta_i, \phi_i^-) + \Delta E(\theta_i, \phi_i^+)] \tag{3.7}$$

The potential energy for the bond length (l_i^S), bond angle (θ_i^S), and torsion angle (ϕ_i^S) of side chain *i* is as below:

$$E_{SR}^S(\Phi) = \sum_{i=1}^{N} E(l_i^S) + \sum_{i=2}^{N} E(\theta_i^S) + \sum_{i=3}^{N} E(\phi_i^S) \tag{3.8}$$

The distance dependent potential of the protein structure is simply summation of short-range backbone energies (E_{SR}), side-chain conformational energies (E_{SR}^S) and long-range interaction energies (E_{LR}) of its individual residues from Eqns. 3.7, 3.8, and 3.6, respectively.

$$E(\Phi) = \sum_{i=2}^{N} E_{SR}(\Phi) + E_{SR}^S(\Phi) + E_{LR}(\Phi) \tag{3.9}$$

The Monte Carlo minimization process is used in various studies [32, 33, 130, 131, 149, 150]. Individual residues are moved by randomly choosing a Cα atom or SC interaction site and iteratively generated random conformation is accepted or rejected according to the Metropolis criterion [65]. At each iteration, the Cartesian coordinates of each residue are perturbed N times for a protein structure with N residues by a random generated number, ξ, between 0 and 1 as:

$$\begin{aligned} x_{new} &= x_{old} + (2\xi - 1)\delta r_{max} \\ y_{new} &= y_{old} + (2\xi - 1)\delta r_{max} \\ z_{new} &= z_{old} + (2\xi - 1)\delta r_{max} \end{aligned} \tag{3.10}$$

where δr_{max} is a proportionality factor controlling the limits or strength of perturbation. The Metropolis criterion simply controls the acceptance of each move in accordance with the overall knowledge based potential of the system. If the new configuration has lower energy than its predecessor, then the starting point of the next iteration is the configuration already on hand. Otherwise, if the Boltzmann factor, $exp\ (\frac{-\Delta V}{k_B T})$ is larger than the random number between 0 and 1, then the new configuration is accepted; if not it is simply rejected, and the initial configuration is examined further.

3.3. ANM-MC Technique

Transition conformations are deformed by ANM along the collective deformation directions within a cutoff distance iteratively followed by MC runs to correct the deformed conformations by minimizing the overall knowledge based potentials of the system. Coarse grained biomolecular structures are employed and each residue has two nodes: the Cα atom and SC while only Cα atom is taken into account for Glycine amino acid which is a very unique amino acid as it contains a hydrogen as its side chain rather than a carbon as is the case in all other amino acids. Schematic flowchart of the method developed is provided in Figure 3.2.

The algorithm [30] is as follows: It takes crystal structures of initial and final/target states as inputs and reduces them to coarse-grained structures in Step 1. Deformations are given to the initial/apo/open structure along the slow mode direction of its aligned target/bound/closed structure by superimposed technique [151] within a predetermined cutoff distance in Step 2. Lowest number of eigenvalues and corresponding eigenvectors are extracted, which gives frequency eigenmodes for each generated structure in Step 3. Dot product of the eigenvectors provides the highest overlap in the target direction in Step 4. Deformation is given along the selected mode direction with highest overlap in Step 5. The deformed structure is the generated structure with the equation below

$$R_{new} = R \pm U \times DF \tag{3.11}$$

where *DF* is a specified deformation factor and *R* and R_{new} are the previous and new position

vectors, respectively in Step 5 which is the last step of ANM. The same deformation determined for the Cα atoms by ANM is then applied to the corresponding SC atoms in which MC takes into account. The deformed conformation is allowed to relax by efficient MC energy minimization runs utilizing the knowledge based potentials, in Step 6. The target vector is updated in Step 7 based on the modification of the previous state. This procedure is iterated until the very initial structure overlaps the target structure most.

RMSD is used in this study as a parameter to determine the extent of convergence of the initial structure over the target structure. ANM-MC trajectories for the dataset are presented in Section 5 for the following parameters: DF = 0.3 Å, R_c = 10 Å, MCs = 15 and MC PS = 0.15 Å.

3.4. Reverse-Mapping Technique

The coarse-grained snapshots coming from ANM-MC simulations are reverse mapped and subjected to energy minimization (Figure 3.3). In this particular study, the coarse grained structures are constructed of the Cα atoms and side chains as representing each residue of protein chains. The present contribution of atomistic details into a coarse grained scheme in the framework of Monte Carlo (MC) simulations utilizes a convenient reverse mapping procedure, specifically by introducing the side chains with the same amount of position shifts of the Cα atoms. For mapping at each iteration, the reference full-atomistic structure is the full atomistic structure minimized at previous iteration. In detail, the position changes in Cα atoms of residues are determined and following the side chain atoms of each corresponding residue are relocated with the same amount of position shifts of Carbon alpha atoms. Then, the obtained full atomistic scheme is corrected by minimizing energy of the system with CHARMM22 all-hydrogen protein force field [152], one of the widely used force fields. The process of energy minimization is performed explicitly with non-periodic boundary conditions. Structure is first immersed in a sphere containing water molecules using VMD software [153]. Protonation states for histidine residues are assigned based on hydrogen bonding opportunities. NAMD software [154] is used to minimize the system energy along CHARMM22 all-hydrogen protein force field. Energy minimization is carried out using steepest descent method by 10000 steps.

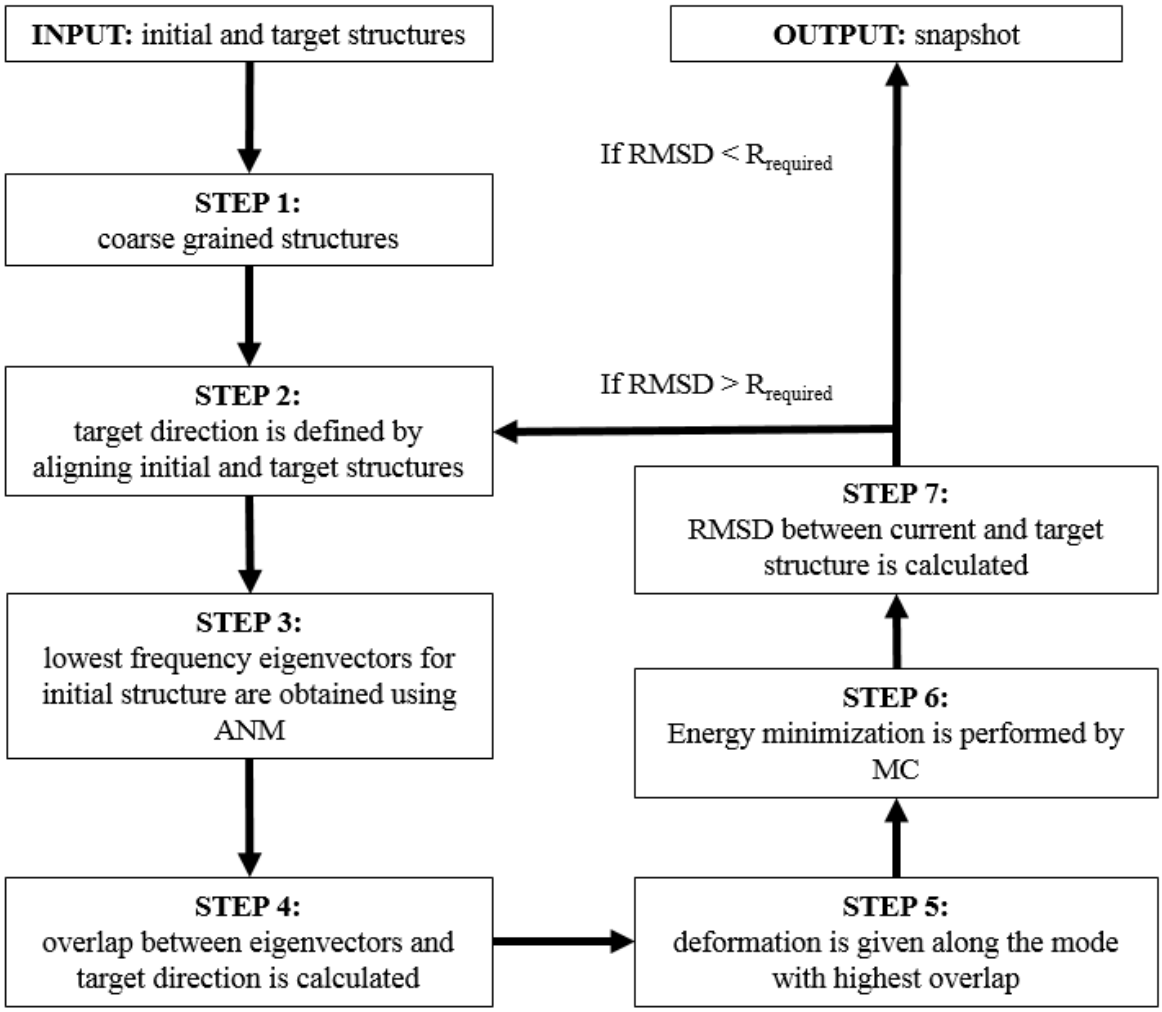


Figure 3.2. Flowchart of ANM-MC algorithm.

Figure 3.4 depicts the procedure on a coarse-grained snapshot of a protein in the dataset. In Figure 3.4a, the coarse grained segment of 5-enolpyruvylshikimate-3-phosphate synthase (ESPSPS) is shown, taken from the 1st snapshot in its ANM-MC trajectory (Chain A, residue 3-10). The reference full-atomistic structure is the structure minimized at previous iteration. Figure 3.4b illustrates its reverse-mapped version, where red circles indicate the Cα atoms. As a preparation to energy minimization step, water molecules are placed by VMD (Figure 3.4c), then it is optimized (Figure 3.4d). This mechanism is repeated along the 327 iterations of ANM-MC trajectory of EPSPS.

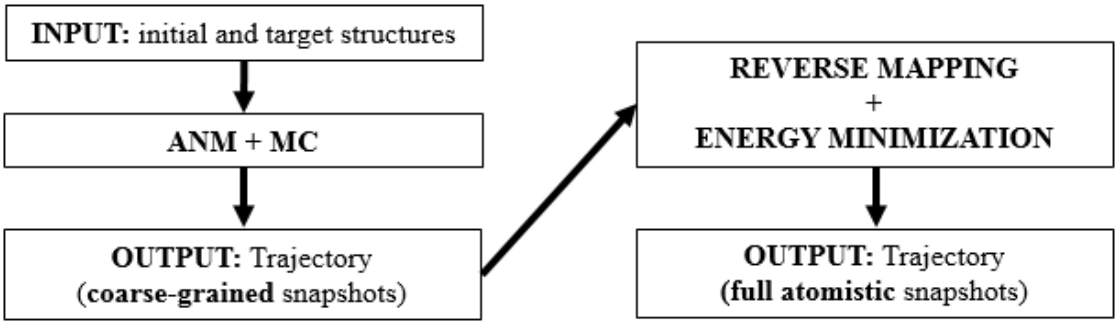


Figure 3.3. Flowchart of reverse-mapping procedure.

Several issues need to be addressed in the full atomic model reconstruction such as plausibility of geometries. Physical full atomic models with regular bond-length and bond-angle need to be constructed. What is a significant challenge is to move the Cα atoms in a correct direction. Most of the virtual points (Cα atom and SC) are moved around in the conformations generated in reduced representation by MC method so that the global topology of full atomic models is at least not worse than the initial reduced model. Additionally, the local structures of the models of the full atomic models should be constructed in a way that the correct hydrogen-bonding networks and the native-like secondary structures are kept optimal.

The coarse grained trajectories obtained by ANM-MC program are investigated whether the generated coarse grained protein structure have plausible geometry in Section 5. The reverse mapped models at each iteration are built with the geometry at previous iteration in order to keep the regular molecular geometry. The three-dimensional constructed conformations are followed by energy minimization runs as to relax both the global body and the local parts of the reverse mapped structures in Section 6.

3.5. Computational Details

In this section, some practical information on the use of the ANM-MC program will be provided. The algorithm is implemented in Python 2.7 and can be run on any computer system. All ANM-MC runs are performed on a local server with an Intel (R) Xeon (R) Model X5670 12-core CPU, clocked at 2.93 GHz with 48 GB RAM.

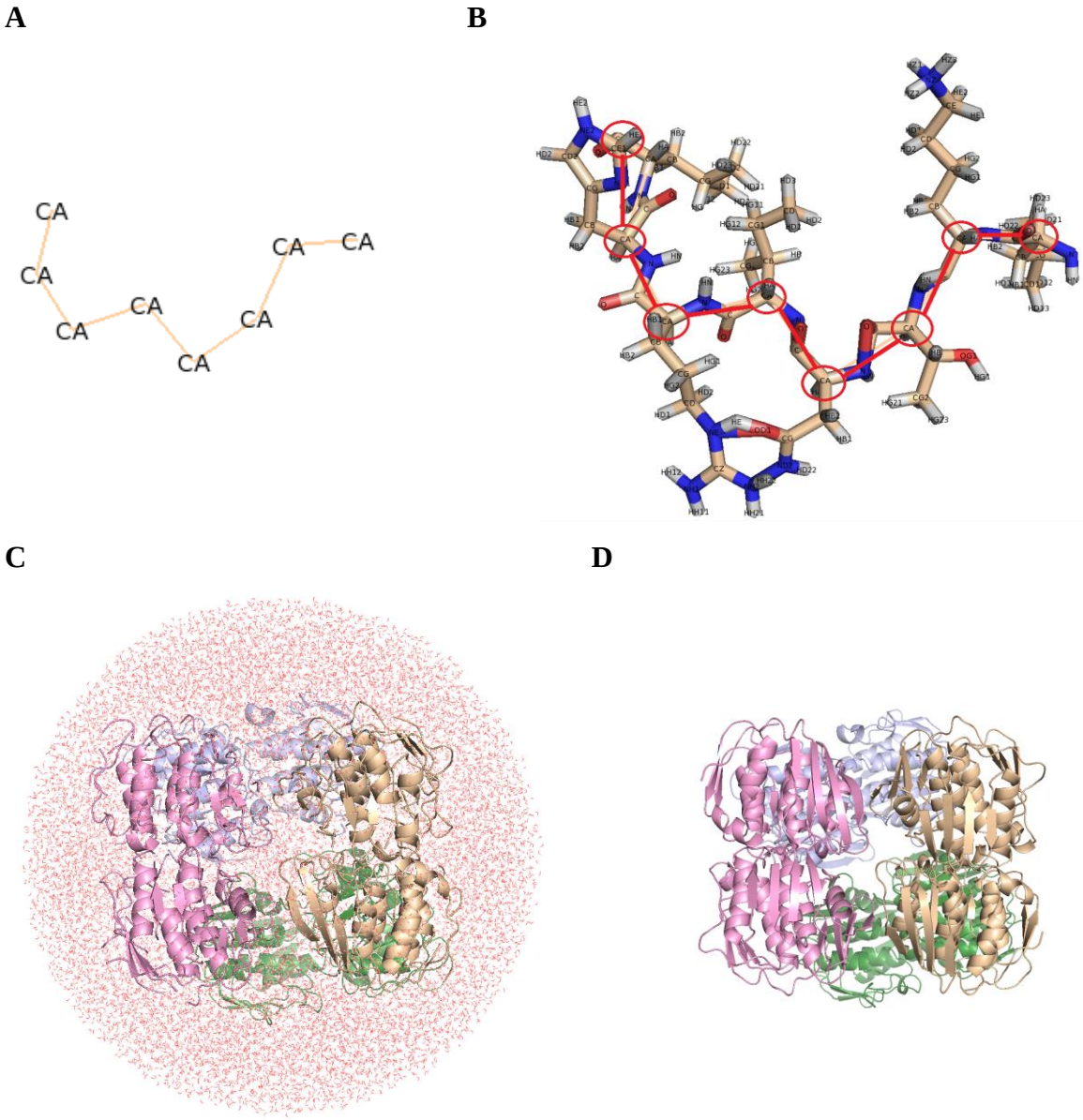


Figure 3.4. Reverse mapping procedure of 1st snapshot of EPSPS. A) Coarse-grained (ribbon) structure of A3-10. B) Atomistic structure of A3-10. C) Atomistic structure of whole enzyme with water molecules around. D) Atomistic structure of whole enzyme after optimized.

The biggest challenge in ANM is obtaining the most significant eigenvalues within applications to large proteins. For a large protein with N residues, Hessian becomes a large 3N-by-3N matrix with a majority of zero-valued elements. To take advantage of this, Py-

thon's sparse-type matrix is used in the construction of the Hessian. Sparse data-type in Python (and similarly in other programming languages) are designed to efficiently store data, save a significant amount of memory and enable speed up in the data processing of matrices that contain a large number of zero-valued elements. In ANM with the use of sparse data-type, Hessian matrices consist of only the nonzero-elements and their indices are stored by compressed sparse row matrices [155], which significantly reduces the amount of memory required for data storage.

A widely-used eigenvalue package, *eigsh* [156] from SciPy library is used in ANM, which is a wrapper around ARPACK [81], SSEUPD and DSEUPD libraries that utilizes the Implicitly Restarted Lanczos Method in eigenvalue and eigenvector calculations. It offers user the advantage of computing only a few of eigenvalues with user specified features without requiring auxiliary storage while providing them precisely.

Another significant challenge in ANM-MC program is the computational overhead of MC part in a single ANM-MC iteration. In initial step of MC, the total amount of energy of the supramolecular system is calculated only once to initiate the interactions. At every step, a randomly chosen residue is relocated by a random distance. Then, the energy of that residue and the residues around its neighborhood are calculated and the difference is added to the previously calculated total energy. In this way, energy calculations are not performed for each single residue, which results in saving the computational time for the residues that are not in the neighborhood of a relocated residue and not affected by them at all.

3.6. Protein Set

The proteins are listed in Table 3.1, together with the total number of residues and chains. The PDB codes of open and closed structures that correspond to the respective initial and target conformations are also given with their mutual (initial) RMSD. The dataset is composed of multi-chain and relatively large proteins undergoing large conformational changes between open and closed structures. Both structures contain same number of residues, same sequence and no missing residue and any binding ligand. For the proteins whose a number of residues is not resolved due to their structural disorder in their x-ray structures are constructed with the MODELLER [157]. Open (skyblue) and closed (sand) structures

are shown in A.1 by cartoon representation via Pymol [158-160]. The cartoon representations of open (green) structure aligned to closed (red) crystal structure, and vector representations of conformational directions are drawn in Figure 3.5.

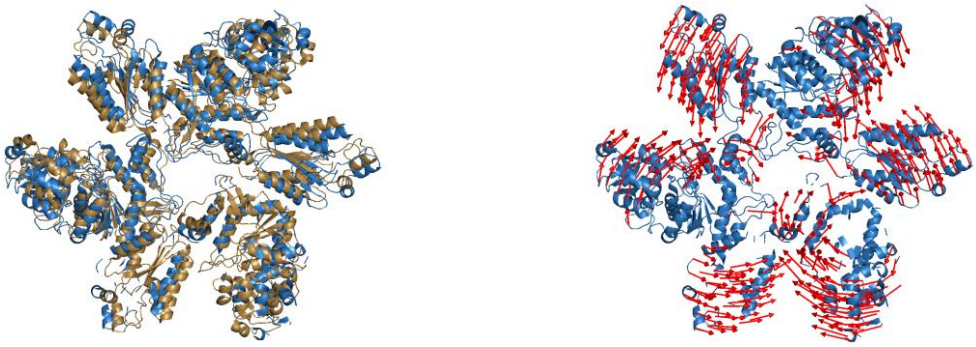
Table 3.1. Protein dataset used in the ANM-MC program.

Proteins	**Symbol**	**Number of Residues**	**Chains**	**PDB IDs open/closed**
ATP Sulfurylase	APS	1716	3	1M8P / 1I2D [161]
Aspartate Transcarbamylase	ATCase I	926	4	1D09 / 1RAC [162]
Aspartate Transcarbamylase	ATCase II	924	4	1Q95 / 1ZA1 [163]
Chaperonin Lidless Mm-cpn	Cpn I	7856	16	3IYF / 3J03 [164]
Chaperonin Lidless Mm-cpn	Cpn II	7856	16	3IYF / 3LOS [165]
Citrate synthase	CS	858	2	5CSC / 6CSC [166]
5-enolpyruvylshikimate-3-phosphate synthase	EPSPS	1708	4	1RF4 / 1RF5 [167]
Glutamate Transporter	GluT I	1215	3	1XFH / 3KBC [168]
Glutamate Transporter	GluT II	1215	3	1XFH / 3V8G [168]
Glutamate Transporter	GluT III	1215	3	3V8G / 3KBC [168]
Chaperonin GroEL (R/T→R’/T)	GroEL I	7322	14	2C7E / 2C7C [28]
Chaperonin GroEL (R”/T→R’/T)	GroEL II	7966	21	1AON / 2C7C [28]
Chaperonin GroEL (R”/R→R”/T)	GroEL III	8015	21	1GRU / 1AON [28]
Chaperonin GroEL (T/R→R”/R)	GroEL IV	7336	14	2C7E / 1GRU [28]
Chaperonin GroEL (R/T→T/T)	GroEL V	7336	14	2C7E / 1GR5 [28]
Hemoglobin	Hb	576	4	1BBB / 1A3N [162, 169]
Lac Repressor	Lacl	944	3	1TLF/ 1EFA [170]

Table 3.1. Protein dataset used in the ANM-MC program (cont.)

Proteins	**Symbol**	**Number of Residues**	**Chains**	**PDB IDs open/closed**
Myosin	MYO	719	2	1VOM / 1MMA [171]
Hypothetical Transcriptional Regulator in QACA	QacR	744	4	1JT0 / 1JUS [172]
RNA Polymerase II	RnaP I	3666	10	1I50 [173] / 1WCM [174]
RNA Polymerase II	RnaP II	3666	10	2E2J / 2E2H [175]
Uracil Phosphoribosyltransferase	UPRTase	846	4	1XTU / 1XTT [176]

ATP Sulfurylase (1M8P/1I2D)



Aspartate Transcarbamoylase (1D09/1RAC)

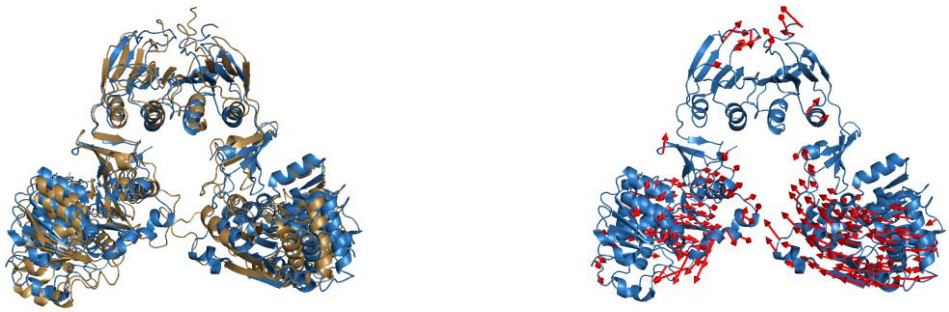
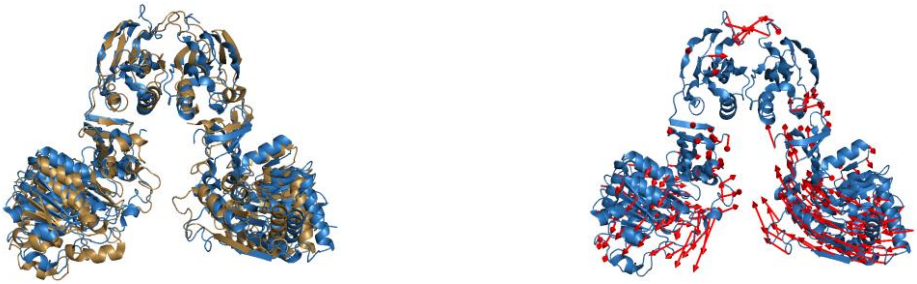
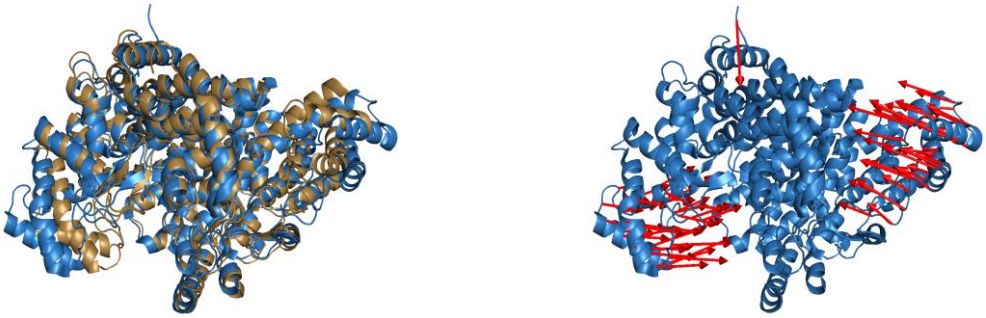


Figure 3.5. Open (skyblue) structure aligned to closed (sand) crystal structure and vectoral representation of conformational directions from open structure to close structure.

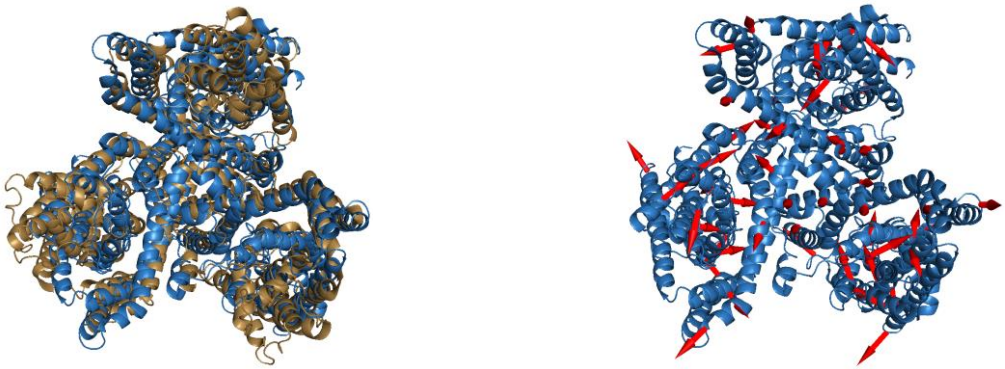
Aspartate Transcarbamoylase (1Q95/1ZA1)



Citrate synthase (5CSC/6CSC)



Glutamate Transporter (1XFH/3KBC)



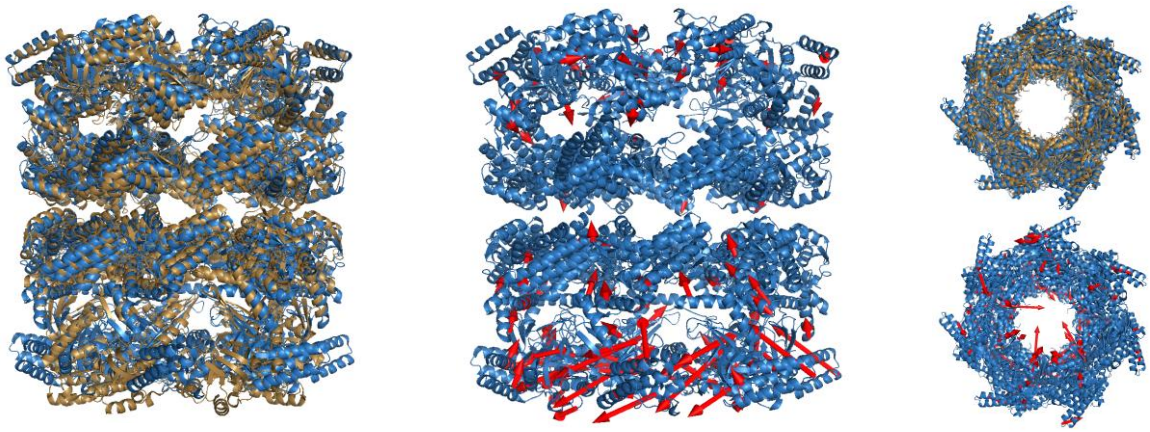
Glutamate Transporter (1XFH/3V8G)

Figure 3.5. Open (skyblue) structure aligned to closed (sand) crystal structure and vectoral representation of conformational directions from open structure to close structure (cont.)

Glutamate Transporter (3KBC/3V8G)

Hemoglobin (1BBB/1A3N)

GroEL (2C7E/2C7C)



GroEL (1AON/2C7C)

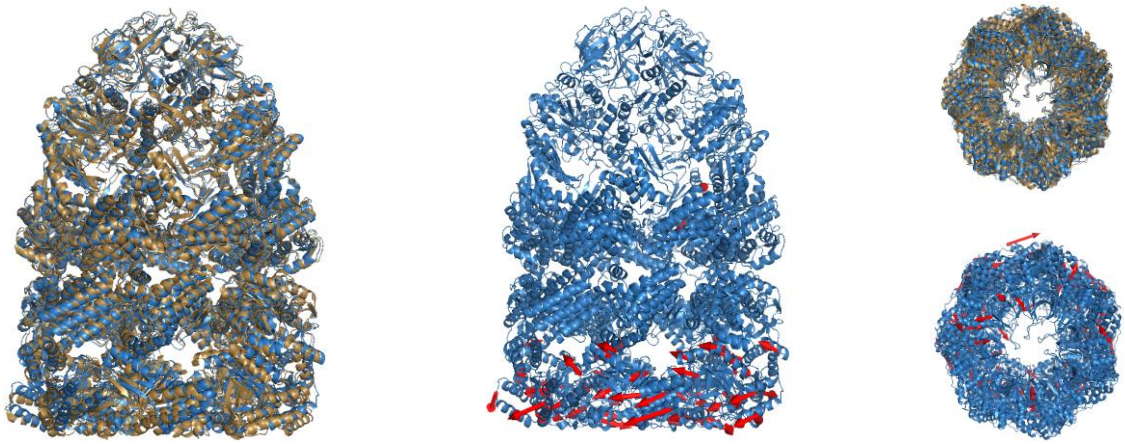
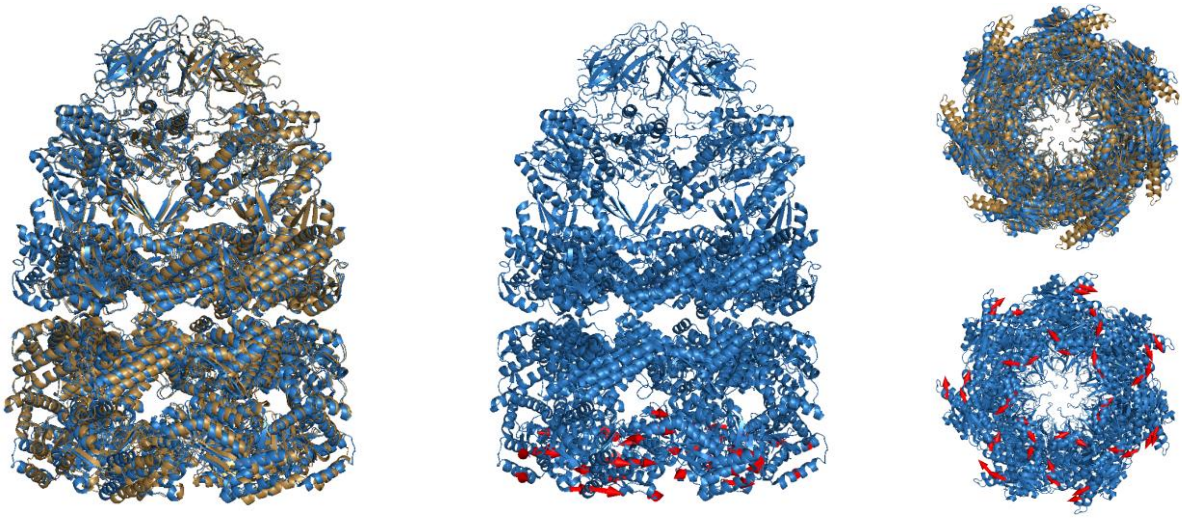
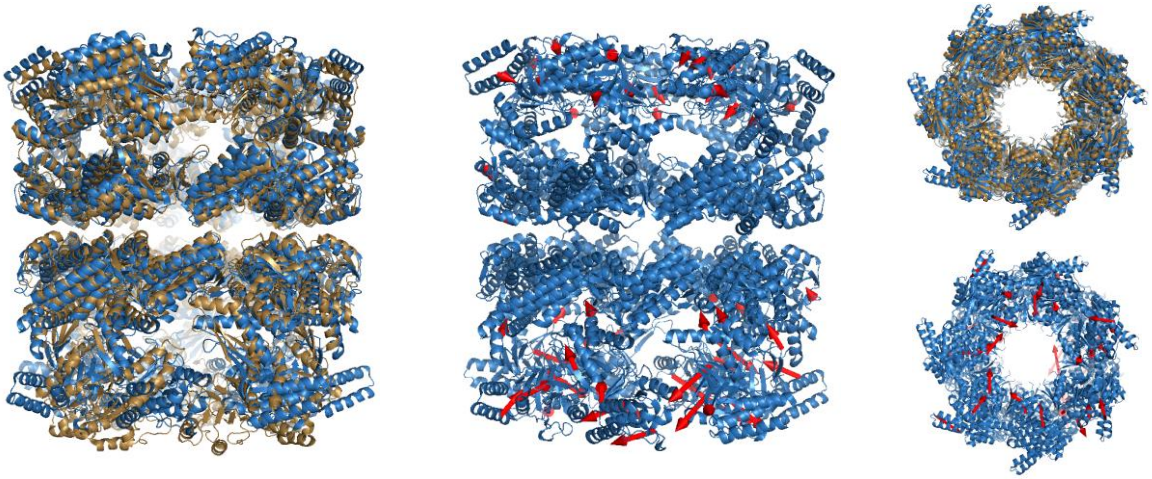


Figure 3.5. Open (skyblue) structure aligned to closed (sand) crystal structure and vectoral representation of conformational directions from open structure to close structure (cont.)

GroEL (1GRU/1AON)



GroEL (2C7E/1GRU)



GroEL (2C7E/1GR5)

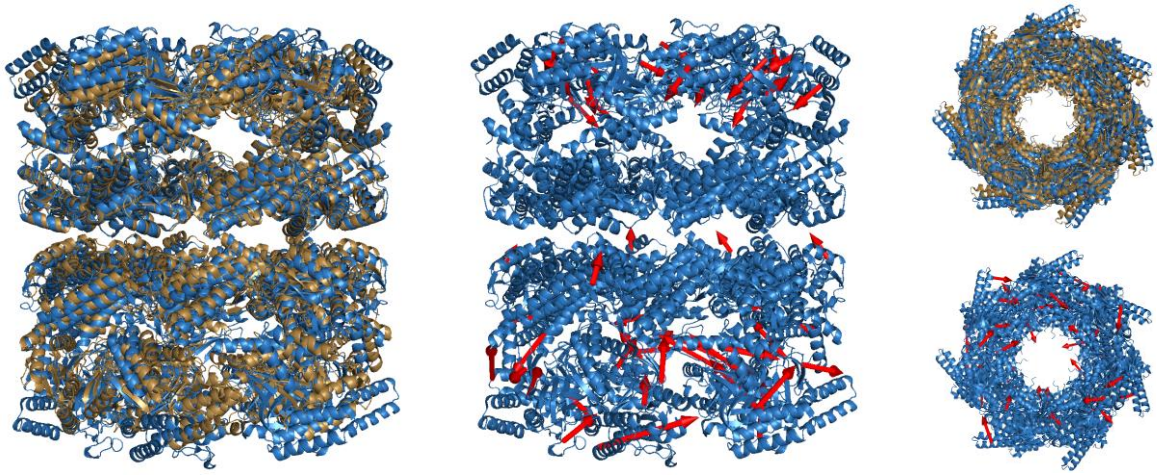
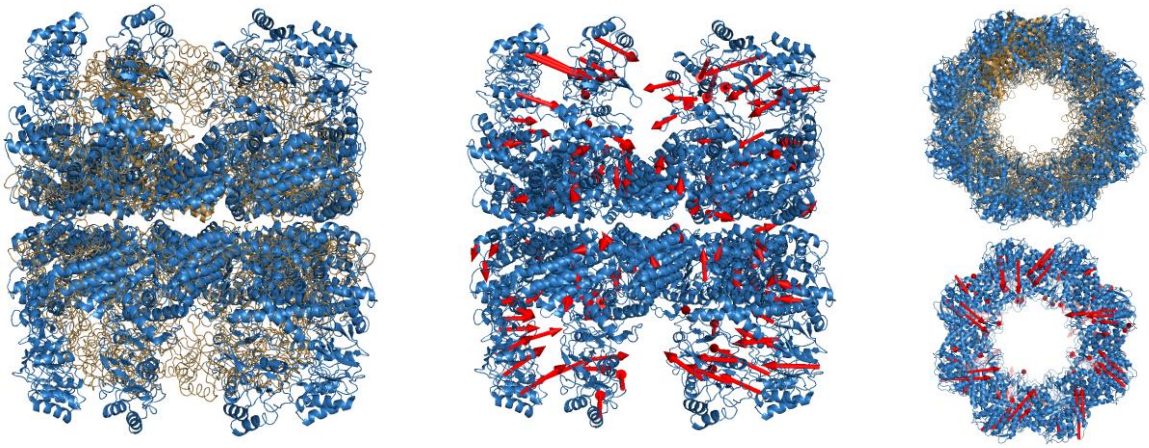
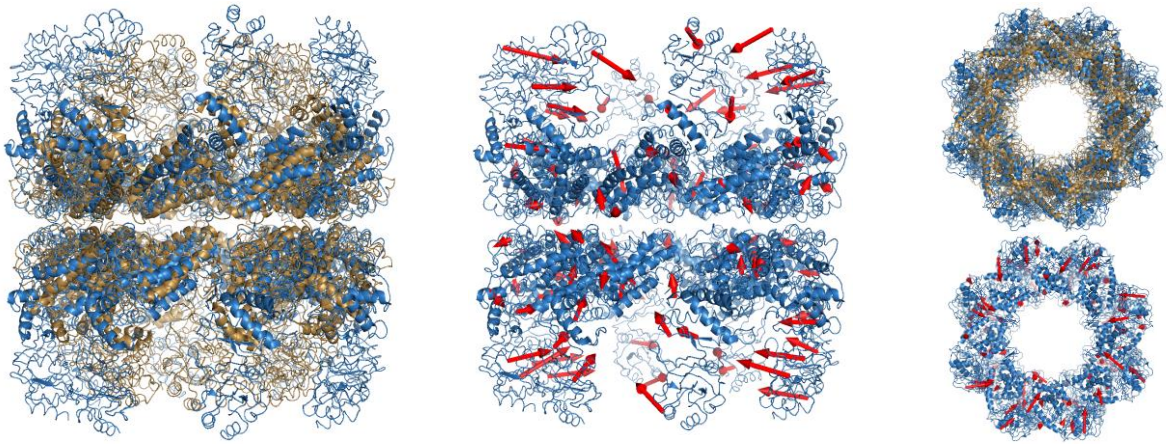


Figure 3.5. Open (skyblue) structure aligned to closed (sand) crystal structure and vectoral representation of conformational directions from open structure to close structure (cont.)

Lidless Mm-cpn (3IYF/3J03)



Lidless Mm-cpn (3IYF/3LOS)

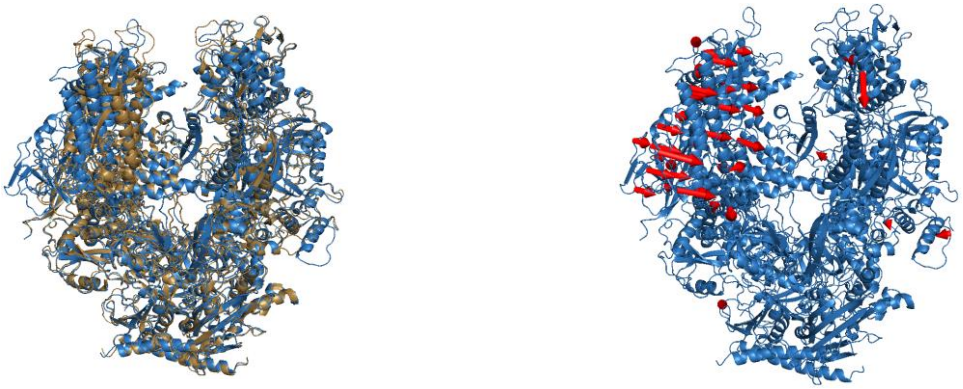


Hypothetical Transcriptional Regulator in QACA (1JT0/1JUS)

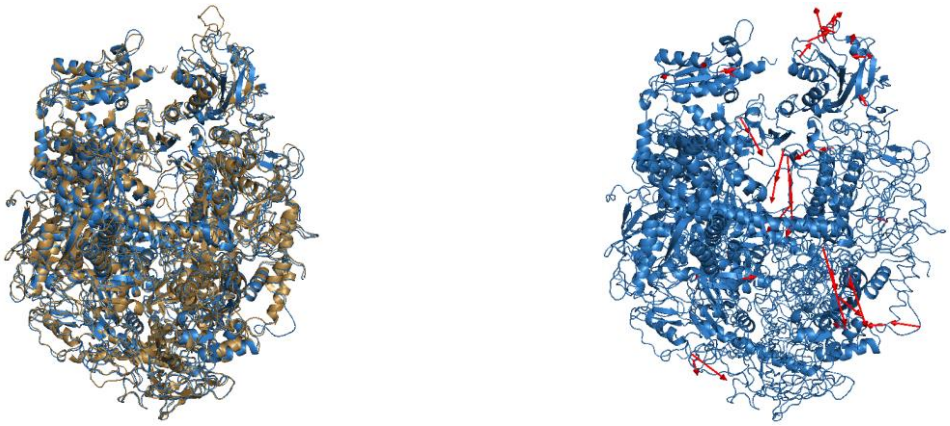
Lac Repressor (1TLF/1EFA)

Figure 3.5. Open (skyblue) structure aligned to closed (sand) crystal structure and vectoral representation of conformational directions from open structure to close structure (cont.)

RNA Polymerase II (1I50/1WCM)



RNA Polymerase II (2E2J/2E2H)



Myosin (1MMA/1VOM)

5-enolpyruvylshikimate-3-phosphate synthase (1RF4/1RF5)

Figure 3.5. Open (skyblue) structure aligned to closed (sand) crystal structure and vectoral representation of conformational directions from open structure to close structure (cont.)

Uracil Phosphoribosyltransferase (1XTT/1XTU)

Figure 3.5. Open (skyblue) structure aligned to closed (sand) crystal structure and vectoral representation of conformational directions from open structure to close structure (cont.)

4. OPTIMIZATION OF ANM-MC SIMULATION PARAMETERS

4.1. Determination of the Spring Constant for Virtual Bonds

ANM explains the internal motions of a protein subject to a harmonic potential by representing the biological macromolecule as an elastic mass-and-spring network. When a spring is stretched or compressed by a mass, the spring develops a restoring force resulting from elastic potential energy. In the network, springs represent the interactions between close-neighboring Cα atoms. After the protein structures is deformed along a collective ANM mode at each iteration, the energy of the deformed structure is minimized by MC simulation based on a coarse-grained potential combining both short range and long range potentials summed over the entire structure.

In knowledge-based potential utilized for MC simulation, the harmonic potential of a backbone virtual bond connecting Cα atoms i and $i+1$ is defined using the instantaneous positions of these atoms as $r_i = (x_i\ y_i\ z_i)$ and $r_{i+1} = (x_{i+1}\ y_{i+1}\ z_{i+1})$, respectively.

$$E(l_i) = \gamma(r_i - r_{i+1})^2 \tag{4.1}$$

This term contributes to the overall energy of the protein structure as given in Equation 3.9 in Section 3.2.

The value of the force constant γ determines the scale of changes in the virtual bond lengths of the backbone. Kurt *et al.* chose the force constant of 10 J/Å/mol in structure-based prediction of potential binding and nonbinding peptides to HIV-1 Protease [177], whereas Kantarci and Uyar *et al.* preferred the force constant of 30 J/Å/mol in investigation of conformational transitions of proteins using multi-scale modeling approaches [30, 34, 131]. Still unrealistically large extensions/contractions of a limited number of backbone bonds have been observed during some conformational transitions. To overcome such artifacts, AK and calmodulin are utilized to adjust the value of the force constant in the MC minimization process and thereby obtain strictly proper backbone geometry of structures along the transition pathway. AK is a 214-residue enzyme that catalyzes transfer of the terminal phosphoryl

group from ATP to AMP. Calmodulin is a multifunctional intermediate calcium-binding messenger protein expressed in all eukaryotic cells [178].

Calmodulin is a small, highly conserved protein that is 148 amino acids long. It has two approximately symmetrical globular domains separated by a flexible linker region. It is such a flexible protein that binds a wide variety of target proteins with various shapes. Particularly the linker region of calmodulin is found flexible in NMR studies even when it is not bound to a target protein. Its flexibility causes a larger conformational change in N-domain (Thr 5- Arg 74) than in the C-domain (Glu 82 – Thr 146) [179].

The coarse grained geometry is the three-dimensional arrangement of the Cα atoms that constitute the protein molecule. Protein geometries can be specified in terms of bond lengths, bond angles and torsional angles. Bond length or bond distance between two bonded Cα atoms is the average distance in a protein molecule. Bahar and Jerningan *et al.* reported backbone virtual bond lengths as 3.81 ± 0.02 Å in their statistical analysis of known structures made for an assessment of the utility of short range energy considerations [149]. Another statistical analysis shows a narrow distribution of Cα - Cα bond lengths centered at 3.78 Å (range: 3.65 Å - 3.90 Å) in a virtual model of protein structure. As an exception, cis peptide bonds before Proline residues adopt a different conformation, where the Cα–Cα distance is 2.97 Å on average [180].

When open (PDB ID: 4AKE) and closed (PDB ID: 1AKE) forms and some experimentally known intermediate structures (PDB IDs: 2BBW, 1DVR, 1ANK, 1E4Y, 2ECK, 1E4V, 3HPQ, 3HPR, 1S3G, 3DKV, 3FB4, 3AKY, 2AKY, 1AKY [146]) of AK are analyzed, the bond length distribution is between 3.65 Å - 3.90 Å in their virtual model structures as shown in Figure 4.1a. Proline amino acid bond lengths are around 2.95 Å in each AK structure and discarded from Figure 4.1a. Alternatively, the distribution of bond lengths of calmodulin with solution NMR data consisting of 160 conformers (PDB ID: 2K0E) are given in Fig 4.1b to provide a better illustration about the distribution of distances Cα–Cα in solvated conformations along a transition pathway. Bond length distributions of crystal structures are observed between 3.75 Å - 3.90 Å with mean of 3.82 Å and standard deviation of 0.04 Å. The minimum value of the distribution of crystal Calmodulin structures is 3.75 Å whereas it is 3.65 Å for NMR models.

In the previous version of ANM-MC program [34] a force constant of 30 J/Å/mol was adopted that generates a bond length distribution around 3.80 Å with standard deviation of 0.07 Å during the transition from open (PDB: 4AKE [181]) to closed (PDB: 1AKE [182]) forms of AK. Bond lengths increased leading to higher variance towards the end of the simulation. Some bonds lengths reached to lower and upper values of 3.4 Å and 4.2 Å, respectively, which questions the plausibility of the intermediate structures.

In order to obtain intermediates consistent with these distributions of experimental bond lengths between 3.65 Å - 3.90 Å, the force constant is modified here to a larger value to make the bonds stiffer so that the intermediate structures adopt more realistic backbone bond lengths.

Different force constants in the range of 30 – 1000 J/Å/mol are tested to find the closest overlap between the predicted and experimental structures. The bond length distributions of both NMR and predicted model of calmodulin are demonstrated in Figure 4.2 for the force constants of 300, 500 and 800 J/Å/mol. The closest agreement between the NMR and predicted model is achieved with the force constant of 500 J/Å/mol. Also, the bond lengths of calmodulin obtained with 500 J/Å/mol shows a more consistent distribution with crystal structures than NMR ones. By comparison of the one in the previous study (30 J/Å/mol) [34, 131], it gives the distributed around 3.8 Å with less standard deviation of 13% and variance %32 as in shown in Figure 4.2b.

AK is a flexible and dynamic protein, which can easily change its shape in response to changes in its environment or other factors. It undergoes a big conformational change with RMSD of 7.1 from open to closed states. Although those large conformational transitions are expected to reflect certain changes in virtual bond angles and dihedral angles as in the nature of conformational transitions, both bond angles and dihedral angles are clustered at the same residues in the intermediate structures of AK shown in Figure 4.3a.

New adjusted force constant (500 J/Å/mol) achieves closer shape to the distribution of crystal AK structures than the previously adopted force constant does in Figure 4.4b. Since the proteins are much more flexible in solution, in their nature environment, NMR data of calmodulin are expected to reflect much more variance by referenced to its initial model.

However, even solvated calmodulin's spectra has the clustering of rotational angles around the same residues when different force constant are applied as in the Figure 4.4. Hence, predicted structures do not undergo high changes in bond angles and dihedral angles as it does so in bond lengths.

On the other hand, a degree of stiffness is reflected in energy profiles comparatively. As the harmonic potential gets stiff more, bonds are literally gets tighter and contributes more to the total energy of protein structure. In Figure 4.5, the amount of contribution of force constant when it gets 500 J/Å/mol gets quite high, compared to that of 30 J/Å/mol. This dominant contribution to the total energy is not seen when the force constant in increased from 300 to 500 J/Å/mol and/or 500 to 800 J/Å/mol in Figure 4.6a for AK whereas the energy profiles shift higher each time when the force constant is adopted higher in Figure 4.5b for calmodulin.

The reason behind the increase in energy profiles needs to be examined by observing the individual contributions of energy terms in MC in detail. First of all, one of the most distant energy profiles occurs for AK between the force constants of 30 to 500 J/Å/mol. In Figure 4.6, the groups of individual energy terms and their changing profiles with the force constants of 30 and 500 J/Å/mol are demonstrated. The largest proportional difference among those energy couples is observed in short-range backbone interaction energies due to the dominant coupling energies of between theta-phi-right. The Cα–Cα interactions in the long-range interaction energies are also affected comparatively by the change of stiffness factor. On the other side, potentials among the short-range interaction energies for sidechains contribute to the overall energy equivalently by the bond stretching, bond angle bending, and torsional angle changes. Nonetheless, bonds are stretched with much higher variance as the potential strength of the spring increases. A similar scenario is also observed in the stretching of bonds individually effecting the amount of energy: the contributions of bond stretches are not influential to the overall energy much when a higher amount of stiffness factor is applied. It is not surprised since the shift value depends on the force constant multiplied by the square difference in bond lengths which is indeed relatively small in the overall system. Additionally, energy value coming from bond stretches with the force constant of 500 J/Å/mol takes steps closer to straight shape. It is the positive effect directly coming from

the tighter bond length distributions and also validates that the force constant of 500 J/Å/mol gives more plausible conformational geometries.

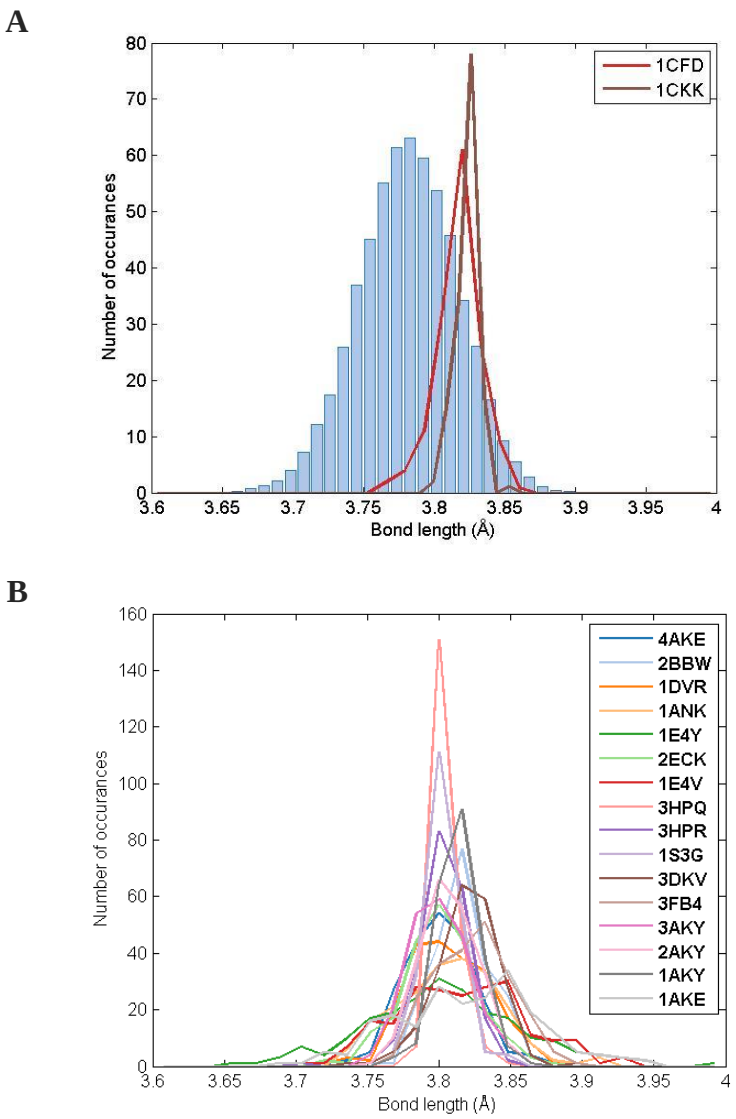


Figure 4.1. Distribution of bond lengths for A) different crystal structures of AK. B) 160 NMR models (PDB ID: 2K0E) and two crystal structures of calmodulin (PDB IDs: 1CFD, 1CKK).

Additionally, the individual energy profiles and their percentages along the transition with the force constant of 500 J/Å/mol are observed for calmodulin in Figure 4.7. It is concluded that any of the individual energy terms does not fluctuate much along the trajectory. Instead, all follow a straight pattern, which indicates that members of the energy groups (i.e. long-range interaction energies) do not compensate each other.

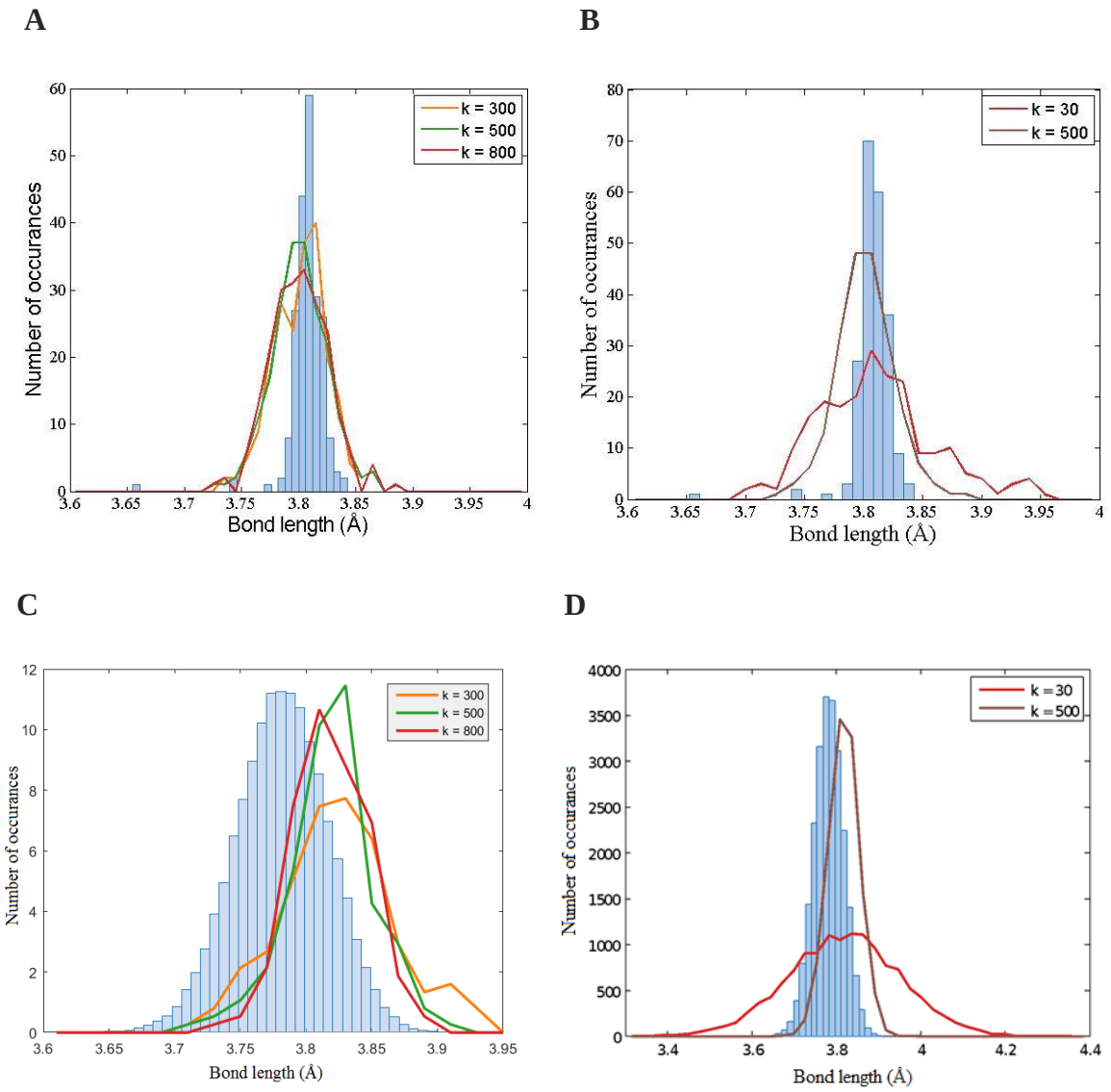


Figure 4.2. Distribution of bond lengths of AK for force constants of A) 300, 500 and 800 J/Å/mol, B) 30 and 500 J/Å/mol and calmodulin NMR models for force constants of C) 300, 500 and 800 J/Å/mol and D) 30 and 500 J/Å/mol.

What is interesting is that the RMSD profiles follow very similar patterns for different force constants adopted although the total energy of the systems shift remarkably up as the network becomes stiffer (Figure 4.5) except for the force constant of 30 J/Å/mol in the trajectory of calmodulin (Figure 4.5b). To illustrate that, the RMSD values among the final structures are listed in Table 4.1 in order to see how similar the final structures of AK actually are. It is encountered that the final structures of the trajectories obtained with high force

constants (i.e. above 500 J/Å/mol) are more similar each other. This might be explained with the tighter networks obtained with stiff force constants.

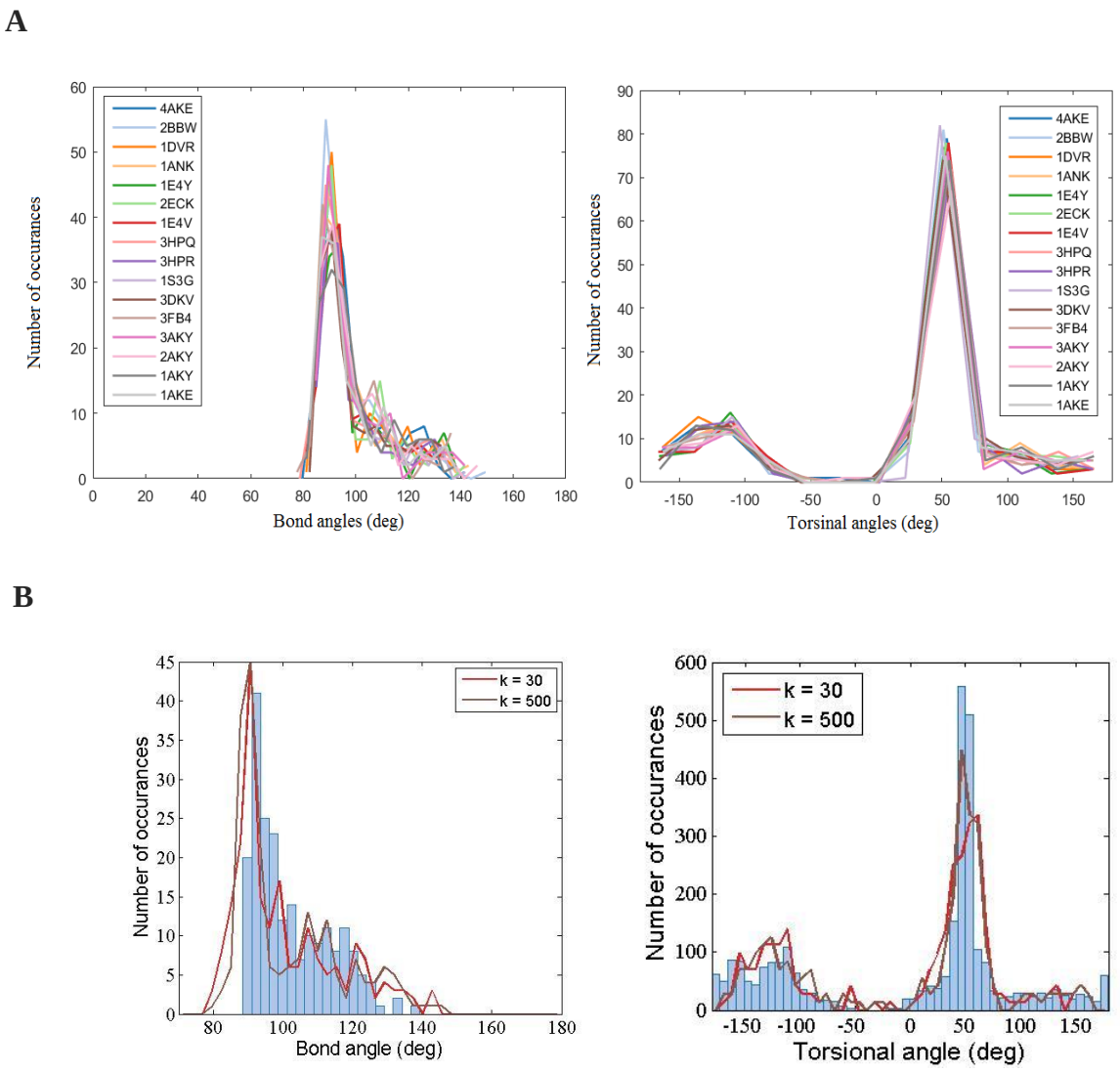


Figure 4.3. A) Bond angle and torsional angles of different intermediate AK structures and B) their distributions with the force constant of 30 and 500 J/Å/mol.

Finally, angle profiles among the domains of AK are drawn in Figure 4.9 to monitor the similarities/dissimilarities occurring in the geometry of conformers during transitions, particularly around the intermediate region between the two end points (open and closed structures). AK contains three well-defined domains: the rigid, predominantly β sheet CORE domain (gray, AK residues 1-29, 68-117, 161-214); the ATP-binding or LID domain (green,

AK residues 118-160) and NMP or AMP binding domain (yellow, AK residues 30-67). The inter domain angles of AK protein are formed based on Beckstein *et al.* [183] between the couples of LID – CORE domains and NMP – CORE domains. The NMP-CORE angle θ_{NMP} is formed by the centers of geometry of the backbone and C^{β} atoms in CORE-LID, CORE, and NMP of AK. θ_{LID} is defined equivalently as the angle between CORE and CORE-LID, and LID. LID dominates large scale motions and its closure precedes the bending motion in the NMP domain [184]. Beckstein *et al.* also reported that AK transition paths follow a path in which NMP movement needs to cross a moderate free-energy barrier while LID one is barrier-less for apoenzyme. [183] Therefore, transitions from open to closed states comprise larger domain movements in LID-CORE than those in NMP-CORE [185]. In Figure 4.8b, ANM-MC trajectories obtained with a range of force constant of 30 - 800 J/Å/mol follow similar paths with a slight kick. When side chains are complemented to coarse grained conformational structures in Figure 4.8d, both domain movements around NMP and LID require crossing a slight barrier as compared to those in Figure 4.8c.

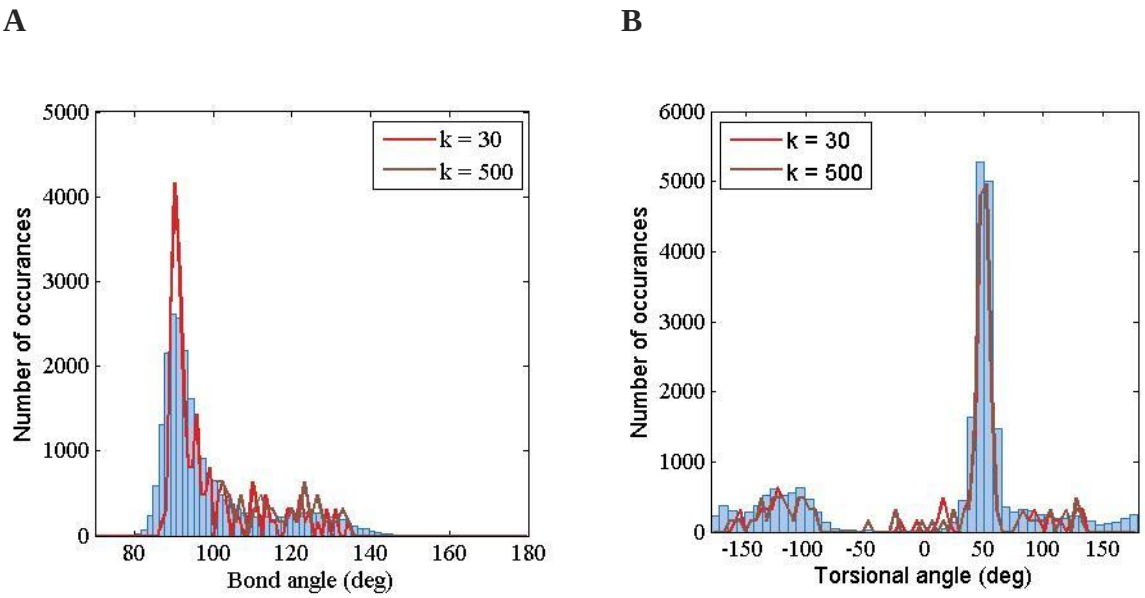


Figure 4.4. Distribution of A) bond angles and B) torsional angles for force constants of 30 and 500 J/Å/mol for calmodulin structures.

4.2. Adjustment of the Magnitude of MC Perturbation Strength

The existence of many local deformations in generated structures, resulting from ANM at each iteration, makes relaxing the conformations of a protein challenging [186]. This problem becomes aggravated as the size of the system increases. The approach here is taking the advantage of the power of conventional Metropolis Monte Carlo method to relax the system in local combinatorial optimization. Monte Carlo-minimization method overcomes energy barriers by random changes in stepping downward through potential energy effectively and let the system undergoes relaxation sufficiently. Traveling effectively over the head of potential energy barriers requires the factor that control the efficiency of MC method, which is the size of each Monte Carlo step.

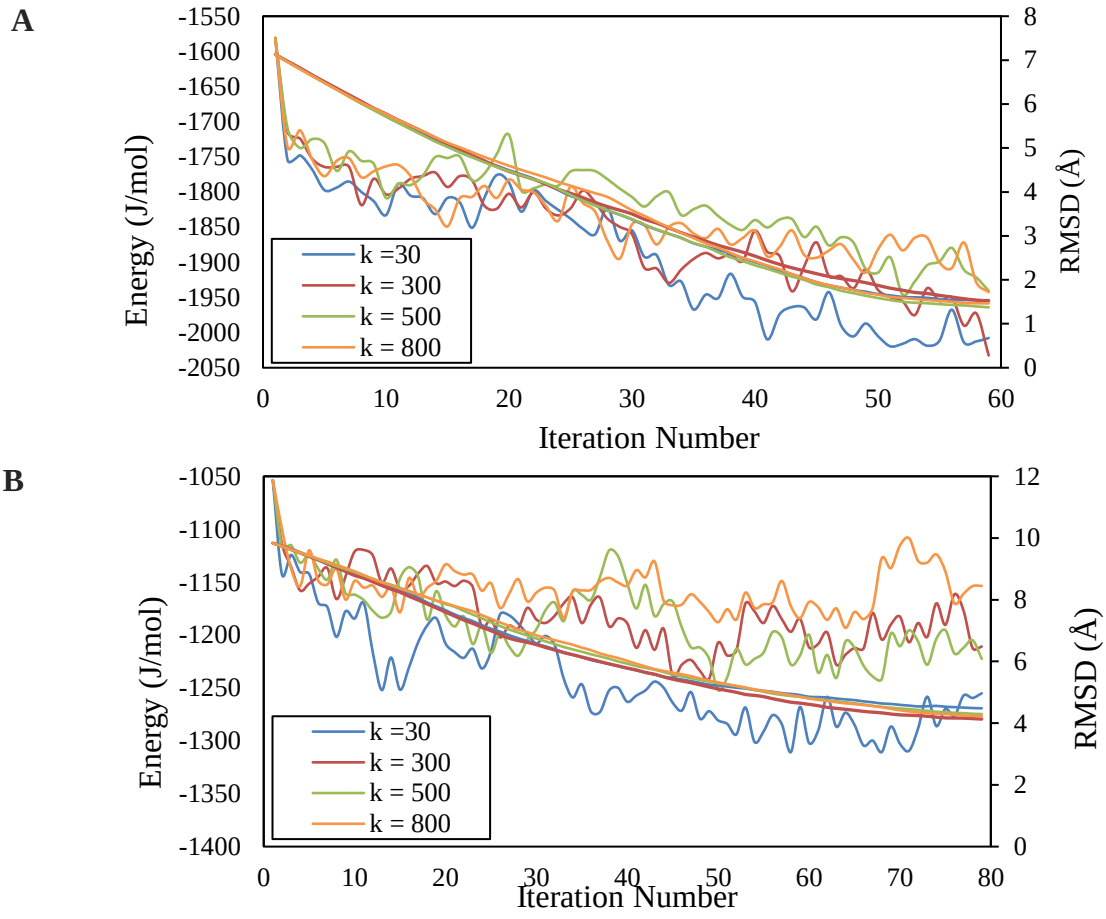


Figure 4.5. Energy profiles of A) AK and B) calmodulin along the ANM-MC trajectory with different force constants.

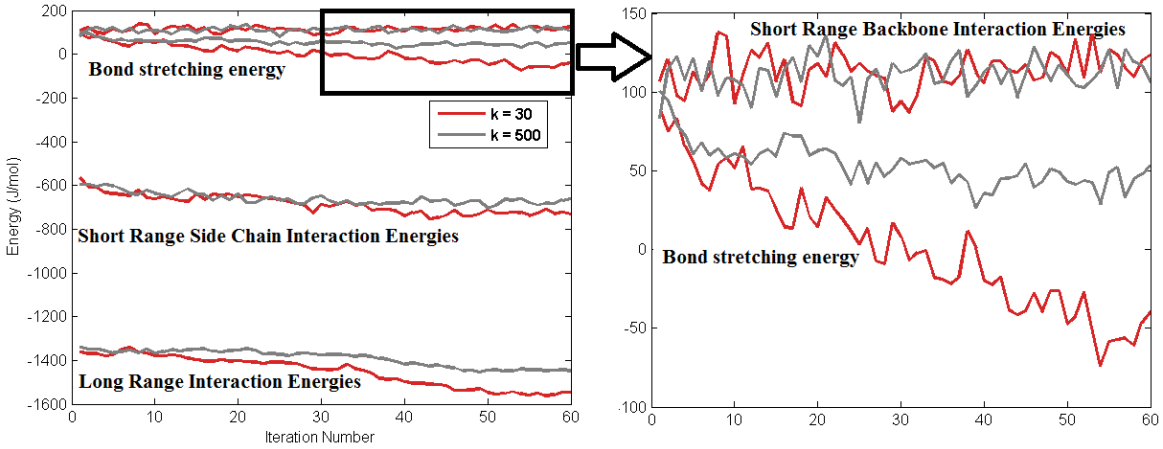


Figure 4.6. Energy profiles of AK along the ANM-MC trajectory with two different spring constants of 30 J/Å/mol and 500 J/Å/mol.

The choice of the parameter δr_{max}, which is proportionality factor controlling the strength of perturbation, determines the size of the trial moves in MC simulations leading to the acceptance rate with a certain probability. In the cases where it is too large, the resulting configuration will have very high energy and virtually all of the attempts will be probabilistically rejected. Conversely, if it is too small, the protein system may have a hard time sampling all of the available search space. In other words, the change in potential energy is not sufficient to explore the potential surface and most moves will be accepted. A good average step size depends upon the system being modeled. One of the usual methods in determination of the perturbation strength is monitoring the energy profiles by keeping the number of accepted Monte Carlo steps fixed as a guide while the proportionality factor changes.

Gur *et al.* [184] tested the ability of MC/Metropolis acceptance ratio to determine the energy barriers circumvent while taking detours on energy profile of AK. When the accepted ratio is 1, it means all the moves are accepted and/or significant increases in energy ascent is avoided. This leads protein samples along the softest modes with the lower tendency to proceed in the direction of the investigated transition which usually requires surmounting an energy barrier. When a relatively large acceptance ratio (0.90) is selected, protein has minimal interference for with the structure-encoded (intrinsic) dynamics manifested by softest ANM modes as the system progresses toward the target. When the acceptance ratio is 0.5, transition trajectories on free energy surface are spanned by inter domain angles of AK.

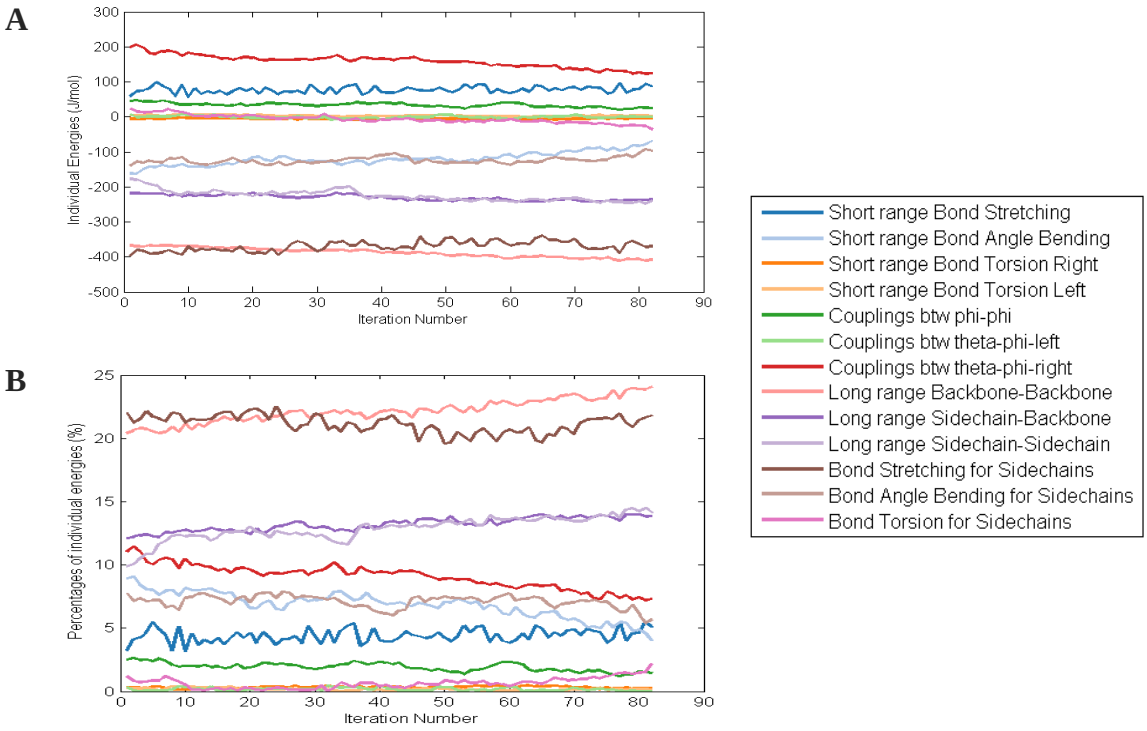


Figure 4.7. A) Individual energy profiles and B) their percentages along the transition for calmodulin with the force constant of 500 J/Å/mol.

The criteria of this study here is to obtain the optimum acceptance ratio that leads consistently energetically most favorable and/or least expensive paths located in configuration space by testing whether a trial move is accepted or not depending on the magnitude of the move. Energy profiles of the ensemble of different runs confirm that independent runs invariably sampled a well-defined similar low-energy region of the energy landscape.

Kantarci-Carsibasi *et al.* reported MC perturbation strength as 0.01 Å with acceptance factor of 1, which simply accepts all the moves and detour a randomly chosen local region by small steps [30]. By comparison, Uyar *et al.* chose a much larger perturbation strength, 0.10 Å, which permits efficiently map the transition pathways when the spring constant 30 J/Å/mol is used [34].

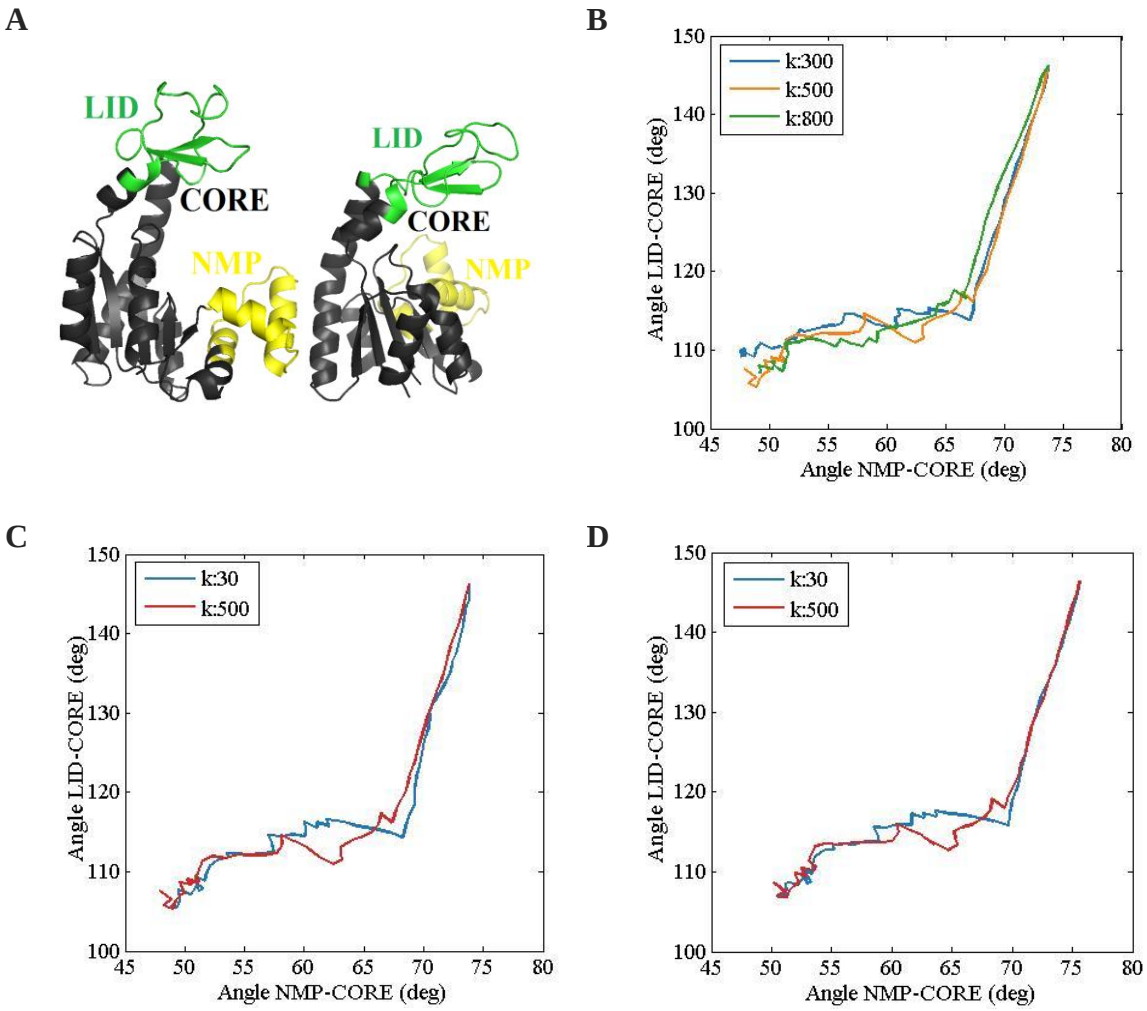


Figure 4.8. A) Open (PDB ID: 4AKE) and closed (PDB ID: 1AKE) structures of AK. Its angle pathways between LID-CORE and NMP-CORE of B) Cα atoms with 300, 500 and 800 J/Å/mol, C) Cα atoms and D) Cα and SC atoms with 30 and 500 J/Å/mol.

In this section, all AK runs are performed with the force constant of 500 J/Å/mol as adjusted in Section 4.1. In Figure 4.9, the acceptance rates for AK protein are controlled with the different perturbation strengths. Parameters that are recommended by Uyar *et al*. [34] gives an acceptance factor of about 0.62. However, even the acceptance rate about 40-45% is observed really high, where *PS* is 0.10 Å indicating that the trial moves stuck on a particular local minima in search space and search energetically unfavorable regions more frequently. Therefore, decrease in the acceptance rate to 0.3 helps visiting more local minima while searching for global optimization, no doubt. However, as the perturbation length in MC runs is taken longer, the protein structure is observed to be able to climb over the local energy barriers and reach the conformations with lower energy to an extent. In other words,

the expected decrease in potential energy is not observed as the acceptance rates become considerably lower than 0.3. This inefficiency arises as MC perturbation strength becomes higher than 0.20 Å leading exceedingly random jumps over the overall energy surface rather than exploring it sufficiently. In overall, the desired decrease in the potential is achieved when the MC acceptance ratio is between 0.3 and 0.4 or the MC *PS* is 0.15 Å (*DF*: 0.2, *MCs:* 20). Thus, this optimized MC step size can give an indication of whether the full search space is explored, or if the simulation hops between a few states in a localized region.

Table 4.1. RMSDs between final structures obtained with ANM-MC program for AK.

	***k*: 30**	***k*: 150**	***k*: 300**	***k*: 500**	***k*: 800**	***k*: 1000**
***k*: 30**	0					
***k*: 150**	1.17	0				
***k*: 300**	1.21	1.18	0			
***k*: 500**	1.22	1.13	1.08	0		
***k*: 800**	1.23	1.15	1.19	1.03	0	
***k*: 1000**	1.20	1.06	1.00	0.96	0.91	0

In contrast the study of Gur *et al*. where two different MC/Metropolis acceptance ratio (0.5 and 0.9) are tested on AK, RMSD between the updated endpoints as a function of MC cycle number has sharper decrease with acceptance ratios of 0.5 than that of 0.9 [184]. However, in Figure 4.9, RMSD values between undated conformation and target surprisingly go with the almost same profiles along the trajectories although the energy of the system varies with each different MC perturbation strength value. This is expected being coming from ANM which is indeed the dominating part in ANM-MC simulations, especially when considering the selected mode profiles in each run having the same distribution in Figure 4.10.

4.3. Determination of the Number of MC steps

The optimum MC acceptance rate with corresponding MC *PS* already helps reaching the lower potential with the same amount of computation time. The next factor that controls the efficiency of MC simulations is the number of independent MC steps attempted in each cycle.

Gur *et al.* examined transition trajectories on free energy surface spanned by AK interdomain angles with three different acceptance ratios (50%, 90% and 100%) by ANM-MC starting from two ends independently. Trajectories starting from the open and closed structures merged on average in 11 and 15 steps for the acceptance ratio of 50% and 90%, respectively. In contrast, convergence could not be observed in independent runs performed for the case in which acceptance ratio is 100% unless these runs are expended to 150 and 200 cycles [184]. In overall, less MC cycles can be sufficient to overlap on the similar energy surface by just adjusting the acceptance ratios.

In this section, AK and calmodulin runs are performed with force constant of 500 J/Å/mol and MC perturbation strength of 0.15 Å and as adjusted in Section 4.1 and Section 4.2, respectively. Different number of MC steps (5, 10, 15 and 20) are applied for AK and calmodulin right after the protein structure is deformed by ANM at each iteration, as in Figure 4.11. As the perturbation strength is increased to an extent, it is expected to have jumps on energy surface more freely yet on well-settled points, as shown in Section 4.2. Instead of having 20 MC steps in previous ANM-MC studies [34, 131], less MC steps (10 and/or 15) are concluded as sufficient to approximate the structures in low energy enough to be considered as stable in its nature.

Even when MC energy minimizations are also carried out over the experimentally known closed AK structures (PDBs: 1ANK, 2ECK, 1E4V) in equivalent conditions, their reduced energy values are between -1900 to -1800 J/mol corresponding to the energy profiles obtained with 10 – 15 *MCs*.

4.4. Jumping in Certain Iterations of MC Technique

The main reason of the application of MC energy minimization method right after ANM at each iteration is to relax the deformed structure to an extent. It is well known that MC is an efficient yet powerful method as applied to the reduced models. However, MC becomes challenging as size of the system increases. Here the MC technique is employed at conformations of interest as well as simply discarded at the particular ones. In Figure 4.12, different energy profile is observed when MC is applied at one iteration out of two or three iteratively. In other words, for the red line, MC is applied at one iteration and skipped the next one out of two iterations. This procedure repeats at every two iterations along the trajectory in order to get a balance over the structure and trajectory.

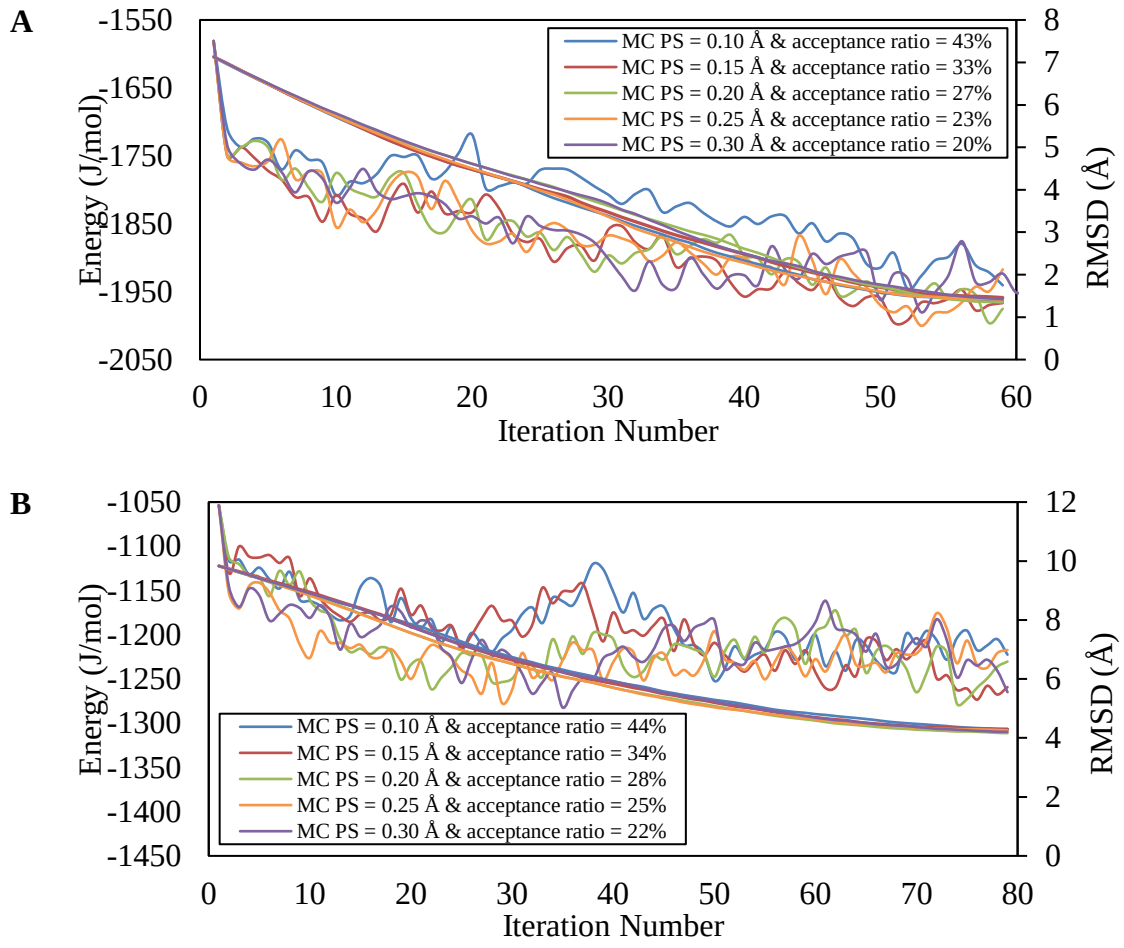
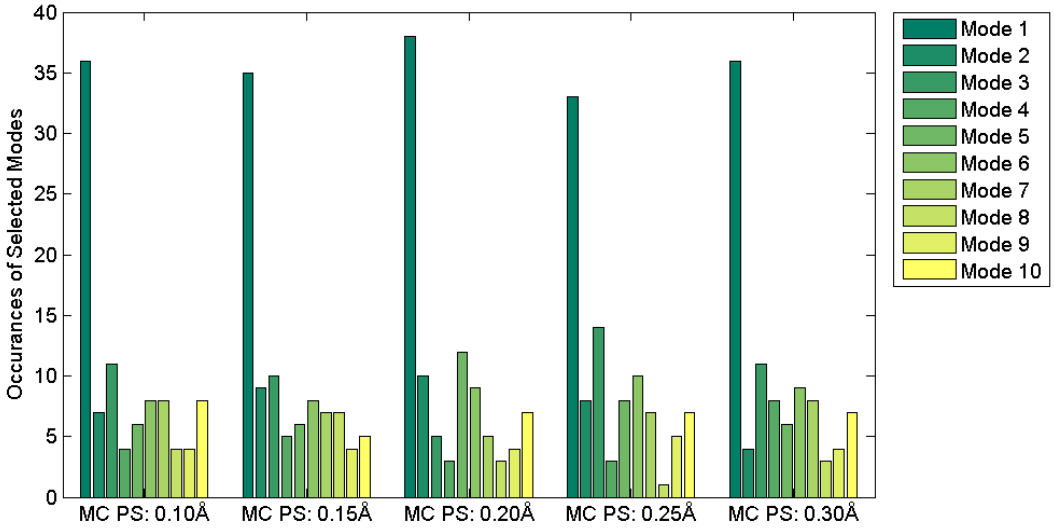


Figure 4.9. The energy profiles of A) AK and B) calmodulin with different perturbation strengths and corresponding acceptance ratios in MC tool (k = 500 J/Å/mol, DF = 0.2, MCs = 20).

For the case where the MC is employed at once in every three steps, the energy profile of the system does not seem feasible during its conformational transitions. On the other hand, the energy profile obtained in which MC technique is applied once in every other step exhibits fluctuations yet stable profile, comparatively to the performing MC at every iteration.

A



B

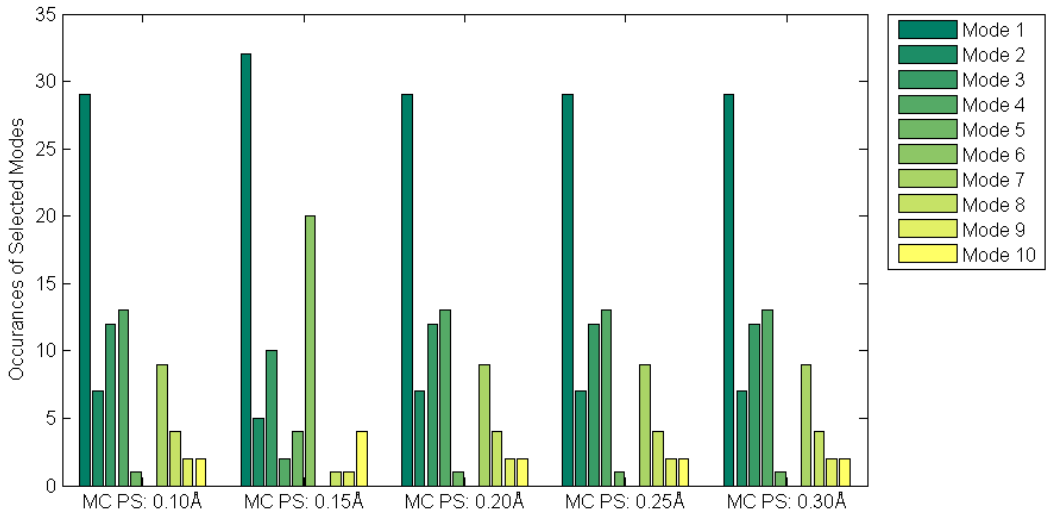


Figure 4.10. Occurrences of selected modes for the runs with different MC perturbation strength for A) AK and B) calmodulin (k = 500 J/Å/mol, DF = 0.2, MCs = 20).

This approach is not applied in this thesis to the protein data set in neither Section 5 nor 6 since the plausibility of the generated structures is one of the main points emphasized in the rest of the thesis. Nevertheless, it is only showed here that MC energy minimization method being skipped at certain iterations could be an alternative choice for the challenging runs (i.e. large systems).

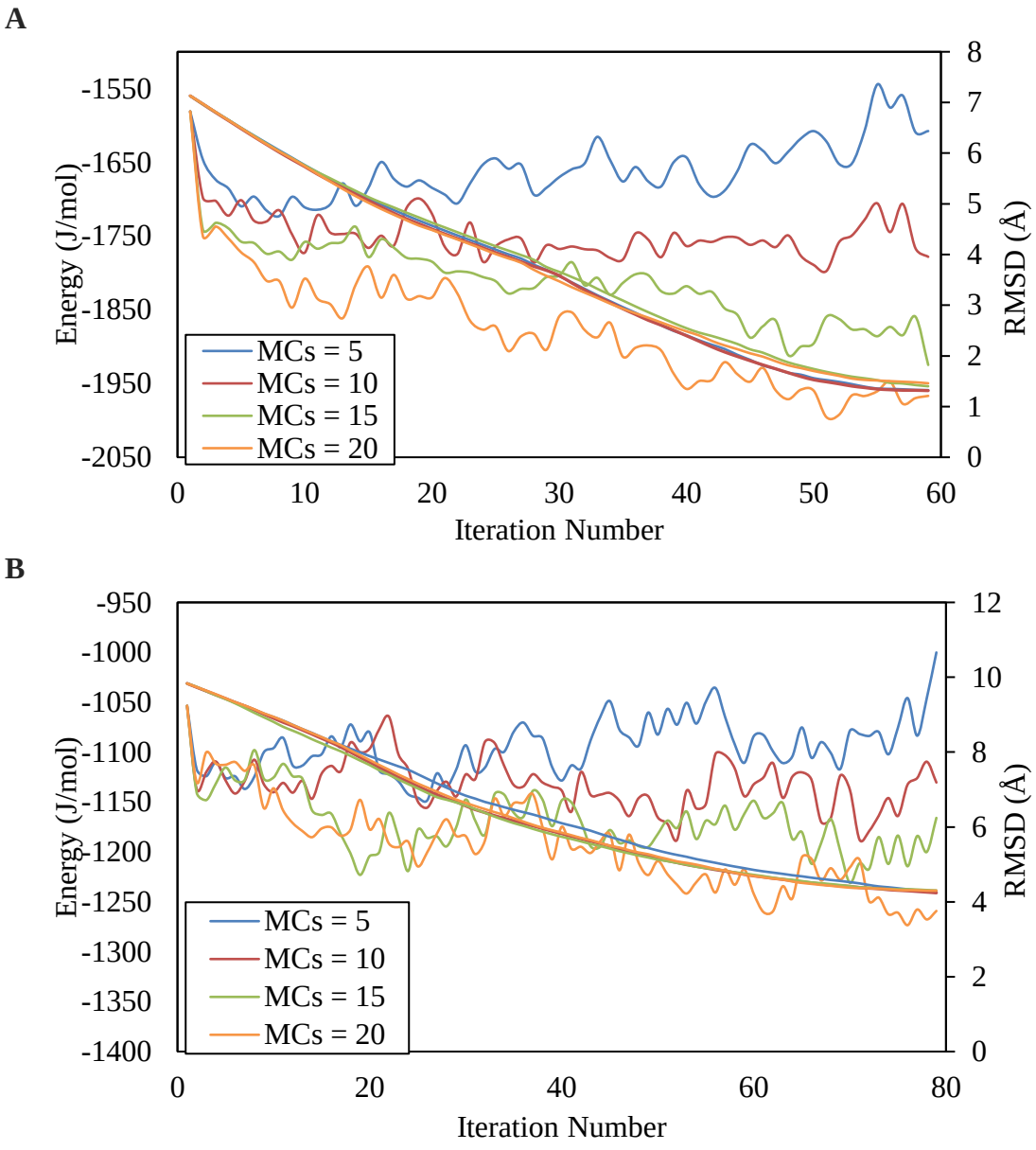


Figure 4.11. The energy landscapes of a) AK and b) calmodulin obtained with different number of MC steps in MC tool (k = 500 J/Å/mol, DF = 0.2, MC PS = 0.15 Å).

4.5. Adjustment of ANM Deformation Factor for Large Systems

ANM-MC methodology generates new conformations by deforming the protein structure along the collective mode directions. Deformation is given in the direction of the selected mode which overlaps with the target conformation most. In other words, deformation helps the protein structure iteratively move straight along a specified pathway. The pre-specified deformation factor in ANM-MC program determines the convergence rate of the initial structure over the target one. If it is too small, the initial structure cannot achieve sufficient conformational transitions to reach the target. If it is too large, the conformations are deformed excessively and seek the support of other methods like MD or much longer MC simulations. The desired conformational transition pathway is safely achieved by *DF* of 0.2 Å [30, 34, 131].

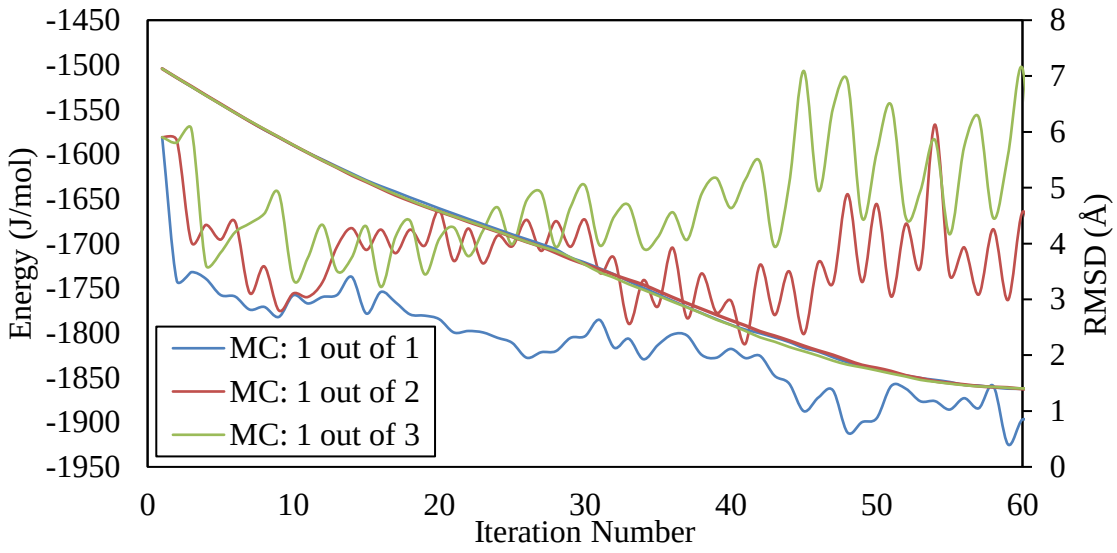


Figure 4.12. Energy profiles for two cases: i) MC is applied at each iteration, ii) MC is applied once at two iterations and iii) MC is applied once at three iterations. (k = 500 J/Å/mol, *DF* = 0.2, MC *PS* = 0.15 Å, *MCs* = 15).

Two different larger *DF* (0.3, 0.4 Å) constants are also tested to determine the limits of the deformation given to the structures for the sake of computational cost to large proteins in Figure 4.13 (k = 500 J/Å/mol, MC *PS* = 0.15 Å, *MCs* = 15). The higher deformation given to the AK in ANM, the faster the protein reaches the target. As the ANM deformation factor

is increased by 50% (from 0.2 to 0.3), it is observed that protein follows the same energy pathway within 33% less iteration number or computation time. As it gets higher (0.4), the energies of the system starts to fluctuate irrepressibly around undesired potential values.

Although different ANM deformation factors are tested, there is no change observed in final RMSD profiles between the end structure at each iteration and target. However, Uyar *et al.* concluded improvement in final RMSDs for AK (1.9 (*MCs*: 100), 1.6 (*MCs*: 50) and 1.4 (*MCs*: 20)) and calmodulin (4.6 (*MCs*: 100), 4.4 (*MCs*: 50), 4.1 (*MCs*: 20)) as less *MCs* are applied [34]. It is suspected to have similar RMSD profiles coming from adopting close numbers of MC steps (5 - 20) in this study for AK and calmodulin. Nevertheless, higher number of *MCs* (50 and 100) are tested to assure that final RMSD values differ due to the number of MC steps adopted and difference (from 5 to 10, 10 to 15, 15 to 20) is too small to reflect it over RMSD. However, the expected change is not observed in Figure 4.14 as higher number of MC steps are taken (k = 500 J/Å/mol, DF = 0.2, MC PS = 0.15 Å, MCs = 10, 15, 20, 50 and 100). Instead, it leaves another question such that if the force constant itself can be the predominant factor on RMSD profiles in ANM-MC runs.

RMSD and energy profiles are monitored with previous adopted force constant [34] in order to see if the reported differences in final RMSD values come with the number of MC steps taken. Figure 4.15 relieves the case where final RMSD values indeed differ due to the force constant taken. As MC makes virtual bonds tighter, structures are not affected much no matter how many MC steps are taken or how much energy gets lower. It makes sense when considering the definition of RMSD such that it gives deviation of distances between virtual bonds and higher force constant in MC restrains bonds deviate. Besides, improvements in final RMSDs as less *MCs* are adopted shows that AK and calmodulin do not always seek conformations whose energies are so low. That also confirms the conclusion made in Section 4.3 that less MC steps are sufficient to have the conformational structures in low energy enough to be considered as stable in its nature.

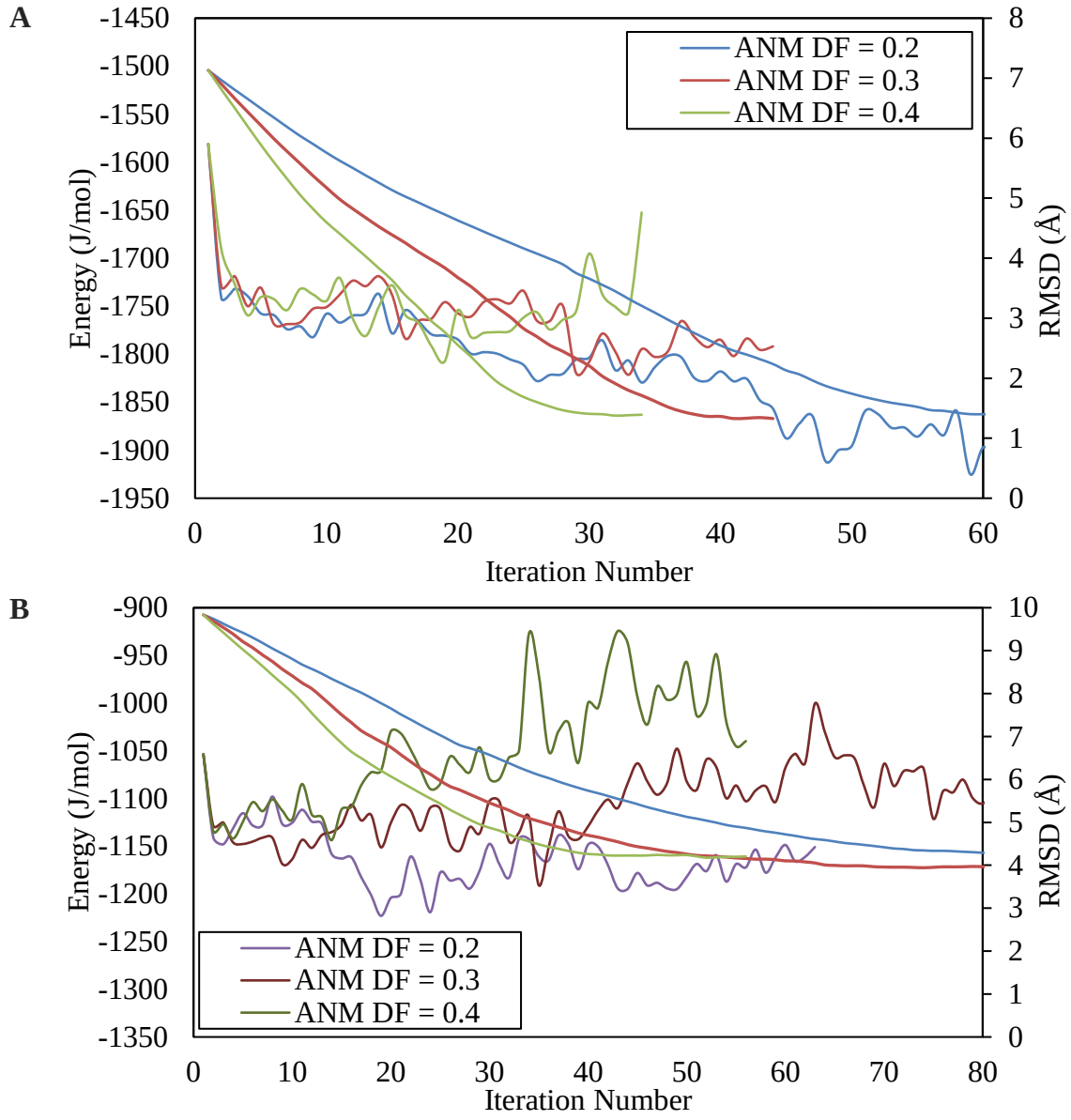


Figure 4.13. Energy profiles of a) AK and b) calmodulin with different ANM deformations along the trajectory (k = 500 J/Å/mol, MC PS = 0.15 Å, MCs = 15) (cont.)

4.6. Adjustment of ANM-MC Parameters to the Efficient Applications of Large Proteins

Another aspect is being able to perform successful ANM-MC simulations efficiently before it is applied to a large system. Computation times are proportional to number of residues and iteration numbers and inversely proportional to ANM deformation factors. It is desired to have the optimum performance with the new adjusted ANM-MC parameters over

their computational cost.

Table 4.2 and 4.3 lists the computation times changing with different ANM-MC parameters for AK and calmodulin, respectively. The runs with adjusted parameters (*DF* = 0.3, MC *PS* = 0.15 Å, *MCs* = 15) are given in bold as recommended runs. Their simulation times are on a fair average and seem to reflect the optimum cost over a successful performance. Keep in mind that these performance evaluations are only for AK and calmodulin, but it is hoped to have a similar performance/cost profile with a large system in the next section.

Last but not least, the RMSD values between experimentally available AK structures and the snapshots with adjusted ANM-MC parameters are listed in Table 4.4. All the RMSD values are below 3.0 Å and even 2.0 Å after the 50th iteration. Note that these experimental AK structures are not necessarily found on the ANM-MC trajectory from open state (PDB ID: 4AKE) to closed (PDB ID: 1AKE) state, which also indicates that the RMSD values do not show how much the generated conformers close to structures that they were meant to be. It only shows how successfully conformational transition pathways are predicted and similar to the experimentally known AK structures.

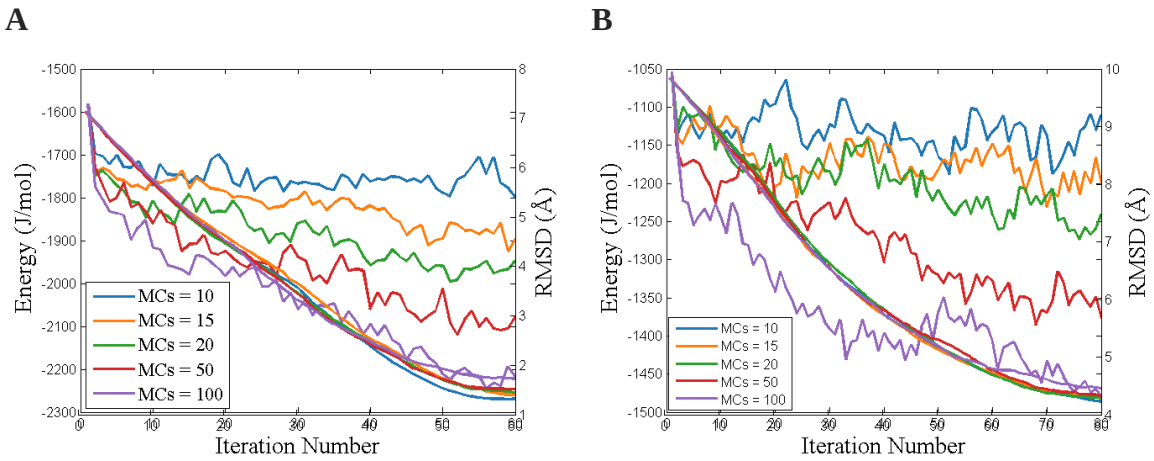


Figure 4.14. Energy profiles of a) AK and b) calmodulin with different number of MC steps along ANM-MC trajectories (*k* = 500 J/Å/mol, *DF* = 0.2, MC *PS* = 0.15 Å).

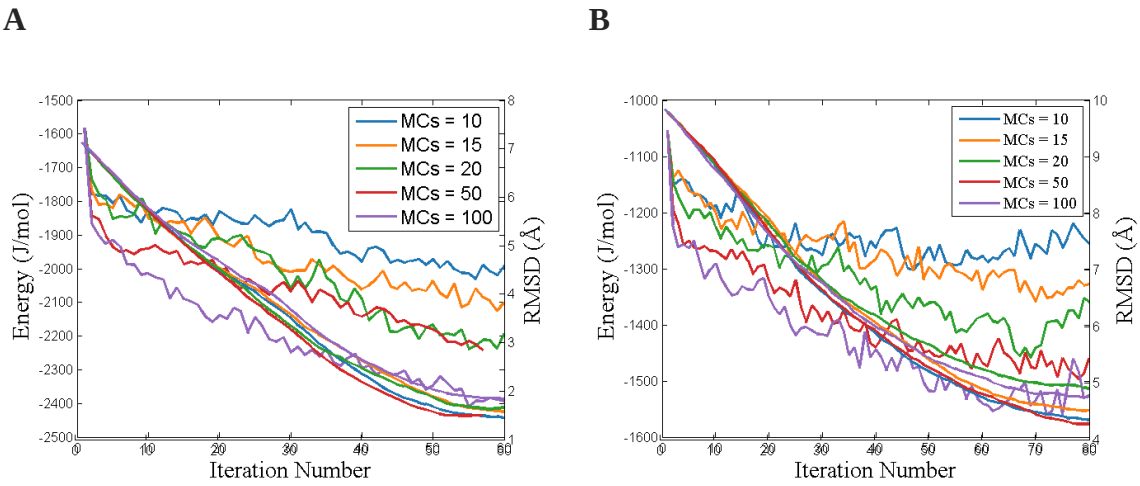


Figure 4.15. Energy profiles of a) AK and b) calmodulin with different number of MC steps along ANM-MC trajectories (k = 30 J/Å/mol, DF = 0.2, MC PS = 0.15 Å).

Table 4.2. Computation times with different ANM-MC parameters of AK with 214 residues.

ANM *DF*	**MC step number**	**Initial RMSD**	**Final RMSD**	**ANM (s)**	**MC (s)**	**Iteration number**	**Duration (min)**
0.2	20	7.13	1.447	0.8	23.8	60	24
0.3			1.394			43	17
0.4			1.318			33	13
0.2	**15**		1.393		15.3	63	17
0.3			**1.327**			**44**	**12**
0.4			1.376			32	9
0.2	10		1.315		9.3	59	10
0.3			1.346			41	7
0.4			1.360			34	6

Table 4.3. Computation times with different ANM-MC parameters of calmodulin with 148 residues.

ANM DF	MC step number	Initial RMSD	Final RMSD	ANM (s)	MC (s)	Iteration number	Duration (min)
0.2	**20**	**9.83**	**4.285**	**0.41**	**12.9**	**81**	**18**
0.3			**3.971**			**94**	**21**
0.4			**4.137**			**52**	**12**
0.2	**15**		4.173		10.1	112	20
0.3			**3.941**			**85**	**15**
0.4			4.175			52	9
0.2	10		4.219		6.3	85	9
0.3			4.182			62	7
0.4			4.310			46	5

Table 4.4. RMSD values between crystal structures and ANM-MC snapshots for AK

Snapshot	PDBs (RMSD (Å))				
5	**1ake** (5.99)	**4ake** (1.50)			
10	**1ake** (4.98)	2bbw (3.13) 1dvr (3.0)	**4ake** (2.96)		
15	1ank (4.27) **1ake** (4.26)	**4ake** (3.88)	2bbw (2.73)	1dvr (1.33)	
20	**4ake** (4.69)	2ayk (3.81) 3ayk (2.74) 1ayk (3.73)	3dkv (3.64) 3hpr (3.57) 3hpq (3.57)	**1ake** (3.53) 1ank (3.53) 1e4v (3.52)	1dvr (2.64) 2bbw (2.60)
25	**4ake** (5.41)	2aky (3.06) 3aky (3.00) 1aky (2.99) 3fb4 (2.84)	3dkv (2.77) 3hpr (2.75) 1s3g (2.71) 3hpq (2.71)	1e4v (1.69) 2eck (2.69) 1e4y (2.65) 1ank (2.63)	1dvr (2.53) **1ake** (1.71)

Table 4.4. RMSD values between crystal structures and ANM-MC snapshots for AK (cont.)

Snapshot	**PDBs (RMSD (Å))**				
30	**4ake** (6.07)	1dvr (3.05) 2aky (2.48) 1aky (2.46) 3aky (2.44)	3fb4 (2.35) 3dvk (2.23) 1s3g (2.16) 3hpr (2.07)	1e4y (2.04) 3hpq (2.03) 1e4v (2.03) **1ake** (2.03) 2eck (2.02)	1ank (1.99)
35	**4ake** (6.35)	1dvr (3.26) 1aky (2.06) 2aky (2.05) 3aky (2.03)	3fb4 (1.96) 3dvk (1.78) 1s3g (1.73)	3hpq (1.56) 3hpr (1.54) 2eck (1.53)	1ank (1.52) **1ake** (1.52) 1e4y (1.52) 1e4v (1.51)
40	**4ake** (6.81) 1dvr (3.59)	1aky (2.01) 3aky (1.97) 2aky (1.95) 3fb4 (1.95)	3dvk (1.69) 1s3g (1.68) 1e4y (1.47)	3hpq (1.38) 1ank (1.36) 3hpr (1.34)	2eck (1.33) **1ake** (1.33) 1e4v (1.32)

5. ANM-MC SIMULATIONS OF CONFORMATIONAL TRANSITION PATHWAYS FOR LARGE PROTEINS

ANM-MC parameters (number of MC moves and their perturbation strength, ANM deformation factor in each cycle and force constant for backbone bonds) have been adjusted in the previous section to increase the efficiency of ANM-MC technique and at the same time produce reliable pathways. This technique is now ready for application to much larger system sizes and/or to systems with large conformational changes through efficient usage of memory and computation speed.

Central to the examination of the ANM-MC methodology here is approximations and parameters that needed to be critically tested in Section 4 for the further applications to large proteins. With the extensive examinations of the transitions of AK and Calmodulin and the comparisons with experimental data and results from other studies [30, 34], it is conceivable that multiple conformations are accessible within much less computation time, given the plausible character of RMSD profiles. In this section, it is of interest to identify the common features, if any, of the ANM-MC trajectories with the adjusted parameters (i.e. efficient parameters for the application of ANM-MC to large proteins).

5.1. Application of ANM-MC to the Dataset of Proteins

Cut-off values may vary in exploration of proteins with different 3D geometries or folds. It is important to determine the proper cut off value in an ANM-MC run is accounting pairwise interactions. The cut off value of 10 Å has been followed in the Section 4 with reference to recent ANM-MC work [34] on hinge-bending proteins on account of the fact that initial structures approach their targets closer. However, the protein set investigated in that reference contains relatively small and compact proteins. When this cut-off value (10 Å) is applied to the protein set in this thesis, additional zero eigenvalue(s), corresponding to the 7^{th} mode, is observed in some of the proteins in the dataset. Note that there are six zero eigenvalues corresponding to three translation and three rotation motions/modes of the whole protein in normal mode analysis. However, the additional 7^{th} zero eigenvalue indicates that the cut-off value (10 Å) used is not sufficient to produce a proper network and some

solvent-exposed residue(s) on loops or chain ends may not have proper connections to the protein and thereby act as independent entities. Due to this observation, a cutoff value of 15 Å is used for such proteins, namely Citrate synthase, Glutamate Transporter, myosin, RNA Polymerase II and GroEL. Table 5.1 lists the cut-off values for each protein adopted in ANM in this thesis.

Table 5.1 Cut-off values and mode numbers used in ANM-MC runs for the protein dataset.

Proteins	**Cut-off (Å)**	**Mode Number**	**Proteins**	**Cut-off (Å)**	**Mode Number**
APS	10	10	GroEL II	15	10
ATCase I	10	10	GroEL III	15	10
ATCase II	10	10	GroEL IV	15	100
Cpn I	15	50	GroEL V	15	150
Cpn II	15	50	Hb	10	10
CS	15	10	LacI	10	10
EPSPS	10	100	MYO	15	50
GluT I	15	150	QacR	10	10
GluT II	15	150	RnaP I	15	10
GluT III	15	50	RnaP II	15	10
GroEL I	15	150	UPRTase	10	10

RMSD profiles indicate the RMSD between each intermediate/snapshot along ANM-MC trajectory and the target structure. Generated pathways invariably exhibit some common features such as having a similar profile in RMSD and energy. Each RMSD profile has a sharp decrease along slow modes (Table 5.2), then follows a relatively smooth decrease along relatively higher modes, and finally reaching a plateau where the simulation is ended.

For the protein dataset involving 22 distinct conformational transitions, runs are performed with the parameters determined in Section 4 (k = 500 J/Å/mol, *MCs* = 15, MC *PS* = 0.15 Å, *DF* = 0.3) and the adopted cut-off values. Initial and final RMSDs to target are listed

in Table 5.2 for ANM-MC runs including 10, 50, 100 or 150 slowest ANM modes. Half of the runs successfully approach the target state within 3.0 Å using the lowest 10 modes. For the unsuccessful runs that cannot reach the target within slowest 10 modes, the number of low-frequency modes chosen for deformation is increased to the first 50, 100 or 150 modes. As a result, the final RMSDs fall into the range of 1.3 – 4.3 Å.

5.2. Analysis of Selected Modes in Conformational Transitions in ANM-MC Runs

Tama *et al.* found that there exists a single low-frequency normal mode that overlaps well with the conformational change for most proteins [21]. Additionally, Krebs *et al.* found that most of known protein motions can be described well by a few low-frequency normal modes [187]. In many cases, only one or two low-frequency normal modes are sufficient to capture the protein motions well [188].

The softest modes predicted by ANM have been recognized to have biological functional significance [189]. The slowest modes, in particular the first and second modes, are found to drive the transition strongly at early stages, succeeded by gradually increasing modes during the second half of the trajectory when the structure attempts to have local changes more. Each mode represents a path away from the previous conformation by moving along these paths together undergoing collective changes. Low frequency modes are known to be robustly determined by the overall shape and contact topology of the examined structure, whereas higher modes describe local dynamics.

In half of the runs in the dataset, a large portion of the reconfiguration occurs via the first 10 global modes (Figure 5.1), where the final RMSD falls below 2.8 Å (Table 5.2). In Figure 5.1, for Citrate synthase and Hemoglobin, third mode is chosen more rather than the first one, while second mode is chosen predominantly in the case of QacR. Nevertheless, the first 10 modes are chosen to different extents in those proteins and a predominant mode during the entire ANM-MC trajectories is observed in ATP Sulfurylase, Citrate synthase, Hemoglobin and QacR.

For other proteins such as Chaperonin GroEL, Glutamate Transporter, Chaperonin Lidless Mm-cpn and 5-enolpyruvylshikimate-3-phosphate synthase, the first 10 modes are

Table 5.2. Initial and Final RMSD values for ANM-MC runs including 10, 50, 100 and 150 lowest-frequency modes.

Proteins	**Initial RMSD (Å)**	**Final RMSD (Å) Mode 10**	**Final RMSD (Å) Mode 50**	**Final RMSD (Å) Mode 100**	**Final RMSD (Å) Mode 150**
APS	4.4	**2.7**			
ATCase I	5.0	**2.1**			
ATCase II	4.9	**1.9**			
Cpn I	16.3	4.4	**2.5**		
Cpn II	16.3	5.6	**4.2**	4.1	4.0
CS	2.8	**1.2**			
EPSPS	26.0	6.2	3.4	**3.0**	
GluT I	9.7	8.9	4.2	3.7	**2.7**
GluT II	8.5	7.5	4.2	3.1	**2.6**
GluT III	4.9	3.6	**2.4**		
GroEL I	10.8	7.1	4.7	4.2	**3.2**
GroEL II	3.1	**2.5**			
GroEL III	2.6	**2.2**			
GroEL IV	6.7	4.9	3.4	**2.8**	
GroEL V	10.9	8.0	4.6	4.0	**3.0**
Hb	3.5	**1.9**			
LacI	14.6	**2.3**			
MYO	4.7	2.4	**2.0**		
QacR	20.4	**2.8**			
RnaP I	4.6	**2.8**			
RnaP II	2.4	**2.2**	2.0	1.9	1.9
UPRTase	2.1	**1.5**			

apparently not sufficient for a completely successful ANM-MC run (Table 5.1). Therefore, the number of modes is increased to 50, 100 or 150 in order to obtain a satisfactory approach to target. It is discussed in the following case study of GroEL how higher number of modes affects the trajectory in ANM-MC simulations.

5.2.1 Previous findings on GroEL

The bacterial Chaperonin GroEL is a supramolecular machine, whose hydrophobic residues located at the entrance to the central cavity of its heptameric ring [190], captures non-native proteins [191, 192]. Each ring accepts a single molecule of unfolded polypeptide by enclosing a central cavity [193]. It regulates adenosine triphosphate (ATP) by assisting in folding and refolding misfolded or partially folded proteins [194-196]. It exhibits more than one functional conformation mostly stabilized by ligand binding. It binds ATP and the GroES co-chaperonin, then turns a massive structure [197] by doubling the GroEL cavity volume and occluding its hydrophobic binding surface [198, 199].

GroEL is a cylindrical protein composed of two complex rings stacked back-to-back, each containing seven subunits of 525 residues [200]. Each unit is three-domain structure composed of equatorial (residues 6−133 and 409−523), apical (residues 191−376), and intermediate hinge-like (residues 134−190 and 377−408) domains ATP-binding site of GroEL corresponds to the equatorial domain, where most of the inter-subunit contacts within and between heptameric rings occur. Binding of unfolded polypeptide occurs after the apical domain forms the opening of the cylinder and a number of hydrophobic residues are exposed towards the central cavity. The apical and equatorial domains are connected by the intermediate hinge-like domain [201, 202].

GroEL undergoes structural transitions between multiple forms (Figure 5.2). Allosteric transitions in GroEL-GroES complex are triggered by ATP binding and hydrolysis. The series of these transitions are the T state, with high affinity for substrate proteins, the ATP-bound R state, and the R″ (GroEL–ADP–GroES) complex [203]. Yang and Bahar *et al.* [28] explored allosteric transitions of GroEL by efficient aANM toll for unraveling potential transition pathways sampled by large complexes/assemblies. aANM is applied to both intrinsic and intact GroEL-GroES system starting from either end, or proceeding simultaneously from

both ends to obtain the closest RMSDs, which is achieved upon moving along low frequency modes. Inter-residue interactions were mostly found in conserved residues near transition states. Alternatively, Hyeon, Lorimer and Thirumalai [204] examined GroEL allosteric dynamics by Brownian dynamics (BD) simulations using a state-dependent self-organized polymer model.

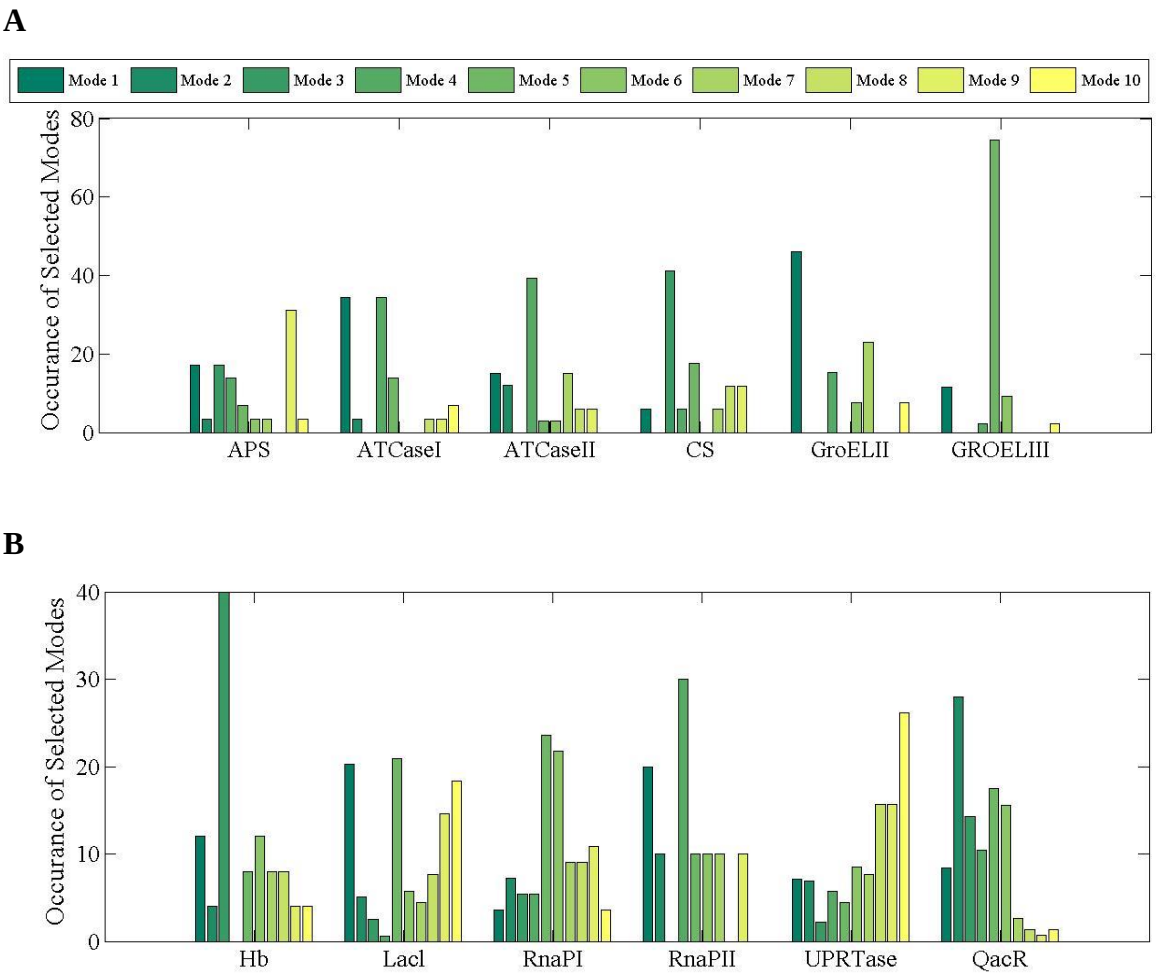


Figure 5.1. Histograms of the selected modes throughout the trajectories including first 10 modes.

The allosteric cycle is summarized in Figure 5.2 with the following conformational transitions as: T/T (PDB ID: 1GR5) → R/T (PDB ID: 2C7E) → R'/T (PDB ID: 2C7C) → R''/T (PDB ID: 1AON) → R''/R (PDB ID: 1GRU) → T/R (PDB ID: 2C7E). Table 5.2 lists the RMSDs between these alternative states. The RMSD between subunits of R' and R'' is lower than the resolution of these structures, only 1.50 A°, while intact R'/T and R"/T differ

from R”/R by less than 3 A°, which is comparable to the resolution of the structures. So, results for the allosteric cycle of R’/T→R’’/T →R”/R (GroEL III and IV) might be condensed due to the resolution of existing structural data and the similarity of R’ and R’’ although all the pathways are reported. On the other side, T/T, R/T and R”/R differ from R/T, R’T and T/R by 6.5, 10.8, and 10.9 A, respectively.

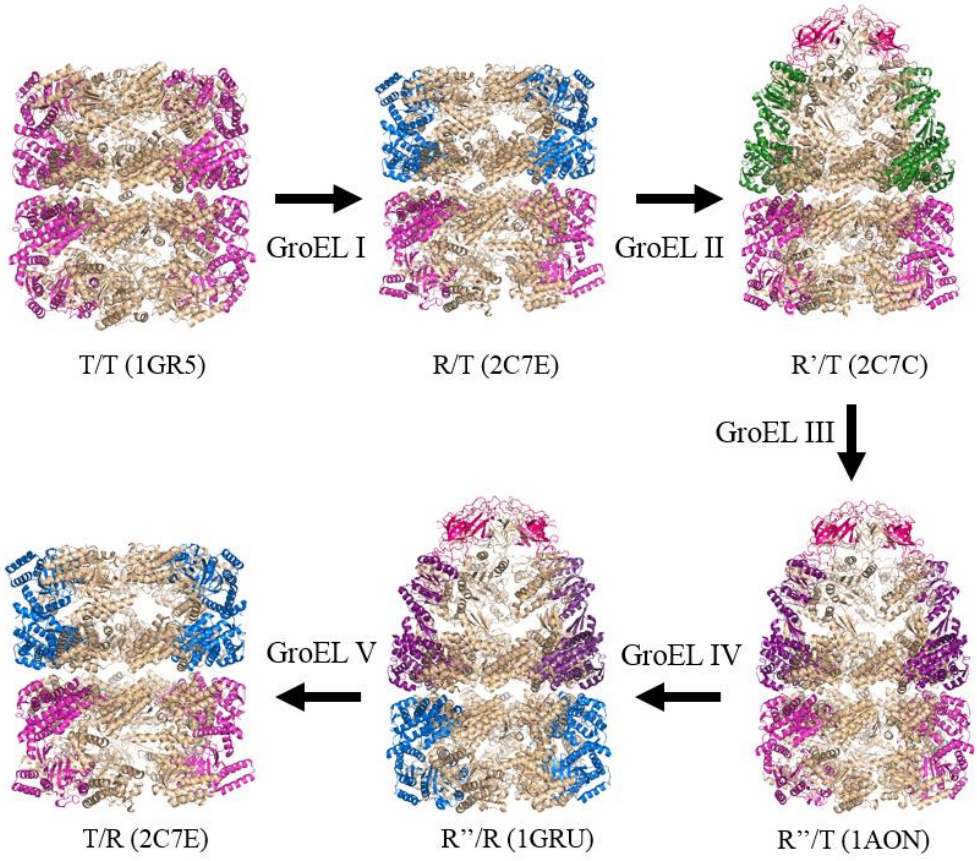


Figure 5.2. The allosteric cycle of GroEL consisting of two rings with the states as follows: T (magenta, ATP-free), R (blue, ATP-bound), R’ (green, ATP-, GroES (pink)- and substrate-bound) and R’’ (purple, ADP-, GroES (pink)- and substrate-bound) [28].

Global modes correspond the lowest frequency modes along entropically preferred configurations [169] and allow substantial decrease in RMSD to the target structure. Collective motions are easily accessed near a given equilibrium state for Groel intrinsically [205]. Yang and Bahar *et al.* concluded that low frequency modes recruit the initial state of deformation and undertake a dominant role for achieving a substantial displacement in the functional direction while higher frequency modes recruit near energy barrier for a single subunit of GroEL [28].

5.2.2. ANM-MC study on GroEL

The set of ANM modes accessible to the GroEL complex is larger than that of the single subunit by a factor of 14 (number of subunits). Thus, it is expected to recruit a larger number of modes in the low frequency regime to achieve the same fractional contribution of single GroEL subunits to observed changes.

In ANM-MC simulations of GroEL, the structural changes are not induced by the slowest modes. Therefore a substantial decrease in RMSD is not observed based on the slowest 10 modes ANM (Table 5.2) for the large conformational changes in the transitions of GroEL I, II, V (T/T → R/T, R/T → R'/T and R''/R → T/R).

Figure 5.3 shows the selected modes among the lowest 10, 50, 100, 150, 200 and 500 modes in ANM-MC simulations of GroEL V (1GRU (R"/R) → 2C7E (T/R)). 10 non-degenerate modes are not sufficient for GroEL to reach the target successfully. Higher number of modes (50, 100 and 150) are tested to obtain lower RMSD to target structure (Table 5.2). As the higher modes become accessible, they are chosen more for deformation in later stages of the simulation. Initial deformations are still observed along slower modes dominantly. In other words, the overall shape is robustly determined by relatively low frequency modes. This dominance of low frequency modes is consistent with the previously studied large scale of motions in GroEL [204].

Figure 5.4 shows how higher modes enable GroEL to undergo local changes and reach target better. At the same time, it clearly demonstrates that incorporation of higher modes are computationally unfavorable in simulations of such a large complex. The RMSD value between open and closed forms of GroEL V (1GRU / 2C7E) starts from 10.9 Å. It drops to the final RMSD, i.e. the RMSD between final snapshot and the target structure, value of 8.0, 4.6, 4.0, 3.0, 2.6, and 2.1 Å with the selected modes among the lowest 10, 50, 100, 150, 200 and 500 modes, respectively. Considering the amount of decrease in RMSD, i.e. how much the final snapshot approach to target, over the computational time spent for each ANM step, the most profitable number of modes seem 100 and 150 for GroEL V (1GRU / 2C7E) in ANM-MC simulations.

Additionally, it is conspicuous that all the modes are selected among the lowest 100 modes even if higher number of modes, i.e. above 100 are sought in Figure 5.3. There is always a chance that the randomness in MC can completely change the direction of mode profiles. However, the similar case is observed in three different simulations when higher number modes, 150, 200 and 500, are sought. So, what could help the final RMSD reach a lower value if the selected modes keep being selected within the lowest 100 modes although higher number of modes are offered? This might be expected to be sourcing from some eigenvalues missed by SciPy eigenvalue package. The best way to testify it is to leave the randomness coming from MC out and iterate along only ANM. As quite expected, some missing eigenvalues show up in the list of lowest eigenvalues when higher number of eigenvalues are asked from the eigenvalue package in SciPy. It might be unfair to evaluate it as a failure, but it is good to keep in mind that eigenvalue packages may fail when a limited number of smallest eigenvalues are sought. Since it is computationally fine to look for the smallest eigenvalues up to 100, it is recommended to let SciPy eigenvalue package approximate by a large list, up to 100, in advance and keep ANM select mode among the first less number of modes (i.e. 10, 50) (Figure 5.4). It is discussed in detail in the following section.

5.3. Computational Efficiency in ANM-MC Applications to Large Proteins

5.3.1. The Eigenvalue Problem

Computational and practical information on the use of ANM-MC program has been already provided in detail, in Section 3.5. The compressed sparse row matrix is utilized together with the eigenvalue package, *eigsh*, from SciPy library in ANM calculations. *eigsh* is a wrapper around ARPACK, SSEUPD and DSEUPD libraries, designed to solve large-scale eigenvalue problems. It is preferred for its efficacy in computing user-selected number of eigenvalues and corresponding eigenvectors, and its applicability on large scale, symmetric, sparse matrices.

Eigenvalue routines by *eigsh* package are tested against for all eigenvalues of Hessian matrix found by *eigh* package in NumPy library, in the context of precision and computational time. Comparing the routines of SciPy and NumPy's inbuilt solvers, a peculiar behavior is encountered as displayed in Figure 5.5.

Figure 5.5a shows the lowest 16 eigenvalues obtained right after 1st ANM iteration with three different ways: *eigsh* package in SciPy for the smallest 16, *eigsh* package in SciPy for the smallest 56 eigenvalues and *eigh* package in NumPy. 6th-zero eigenvalue is missing when only the lowest 10 eigenvalues are requested by *eigsh* (Mode 10 - SciPy). However, *eigsh* is successful in finding the 6th-zero eigenvalue when higher number of eigenvalues are requested (Mode 50 - SciPy). ARPACK fails to operate well at finding small eigenvalues [206]. Starting with 7th eigenvalue, the nonzero eigenvalues (red - asterisk) are found to be shifted from the other numerical (blue – line & blue - circle) ones. Moreover, there is some level of degeneracy in some eigenvalues produced by *eigsh* (Figure 5.5b). Around eigenvalue number 110, 115, 120 and 150 two numerical results start to diverge from each other. It is coming from the breakdown of the Lanczos algorithm which *eigsh* uses through a call to the ARPACK library. It is based on a set of orthogonal vectors and loss of orthogonality results in approximate eigenvalues located as some spurious pairs of nearby real eigenvalues.

Unfortunately, there is not a perfect eigenvalue package that computes a few eigenvalues and eigenvectors. Since it is computationally inefficient to compute the entire spectrum of Hessian for a large matrix, an eigenvalue package is still the only alternative to get the smallest eigenvalues for large proteins with much less computational time. The current package fails to find all the eigenvalues when the number of eigenvalues requested is less than 30. To address this issue in this thesis, at least 30 eigenvalues are calculated whatever the user input is. Then, only the first n-eigenvalues requested by the user are returned.

Table 5.3 lists the CPU times for a single ANM iteration for each protein in the dataset, what uses *eigsh* in SciPy library. ANM-MC program does not run in parallel and there is no file reading (I/O), therefore the total CPU time given here does not differ much from the wall-clock time. ANM has two main successive processes: eigenvalue/eigenvector calculations and mode selections. Process time spent in ANM mostly comes from the former. The elapsed times of all ANM-MC runs are provided in Appendix C. The *time* function of *time* module is utilized in measuring wall-clock times.

Please keep in mind that there are several factors that affect CPU time varies even just a bit between different runs of the same job due to many reasons such as cycle stealing on

systems with integrated channels, buffer interference caused by concurrent tasks, storage access and space allocation and so on [207]. Nonetheless, Table 5.3 is only for one instance of calculations without any explicit performance testing for ANM, showing the relation between process time spent in ANM cycles and different number of eigenvalues and residues.

Figure 5.6 depicts the exponential relation between the number of residues and the process time. Note that the R-squared value is 0.9388, which is an indicator of a good fit of the line to the data. Computation time of eigenvalues depends directly on the size of Hessian matrix, or the number of residues, which specifically increases above 2000 residues.

5.3.2. Notes on Performance and Memory Usage

Memory and CPU usage of ANM is monitored for the calculation of a few eigenvalues to investigate the tradeoff between costs of catching a few smallest eigenvalues previously computed by *eigsh* – SciPy and the costs of finding all eigenvalues recomputed by *eigh* - NumPy. In ANM-MC simulations, ANM part, particularly eigenvalue calculations, requires the most memory. Memory usage of ANM is analyzed up to 150 modes for each protein, by monitoring how much of our server's RAM is being used. The eigenvalue calculations with *eigsh* – SciPy along with the use of sparse Hessian matrix is found to be feasible for all the proteins in the dataset. For the proteins whose number of residues is less than 2000, the allocated memory is below 500 MB and do not vary with the number of eigenvalues sought between 10 and 150. Two of the largest proteins in the dataset with more than 8000 residues, Chaperonin GroEL-GroES and Chaperonin Lidless Mm-cpn, ANM requires 5.9 GB of memory. Nevertheless, neither the memory usage has been monitored for higher number of modes nor a depth analysis of the memory usage in an ANM-MC run has been performed thoroughly.

To sum up, should a small number of eigenvalues requested, less than 30, some eigenvalues might be missing or deviated from the original when calculated with *eigsh* – SciPy. For proteins with residue numbers less than 2000, CPU times and memory allocations do not vary much, as seen in Table 5.4, therefore for the sake of reliability of results and accuracy of eigenvalue calculations, all of eigenvalues of Hessian matrices are computed. For

other proteins, the ones with more than 2000 residues, *eigsh* – SciPy is used in a way that 30 eigenvalues are calculated and only the first 10 are used.

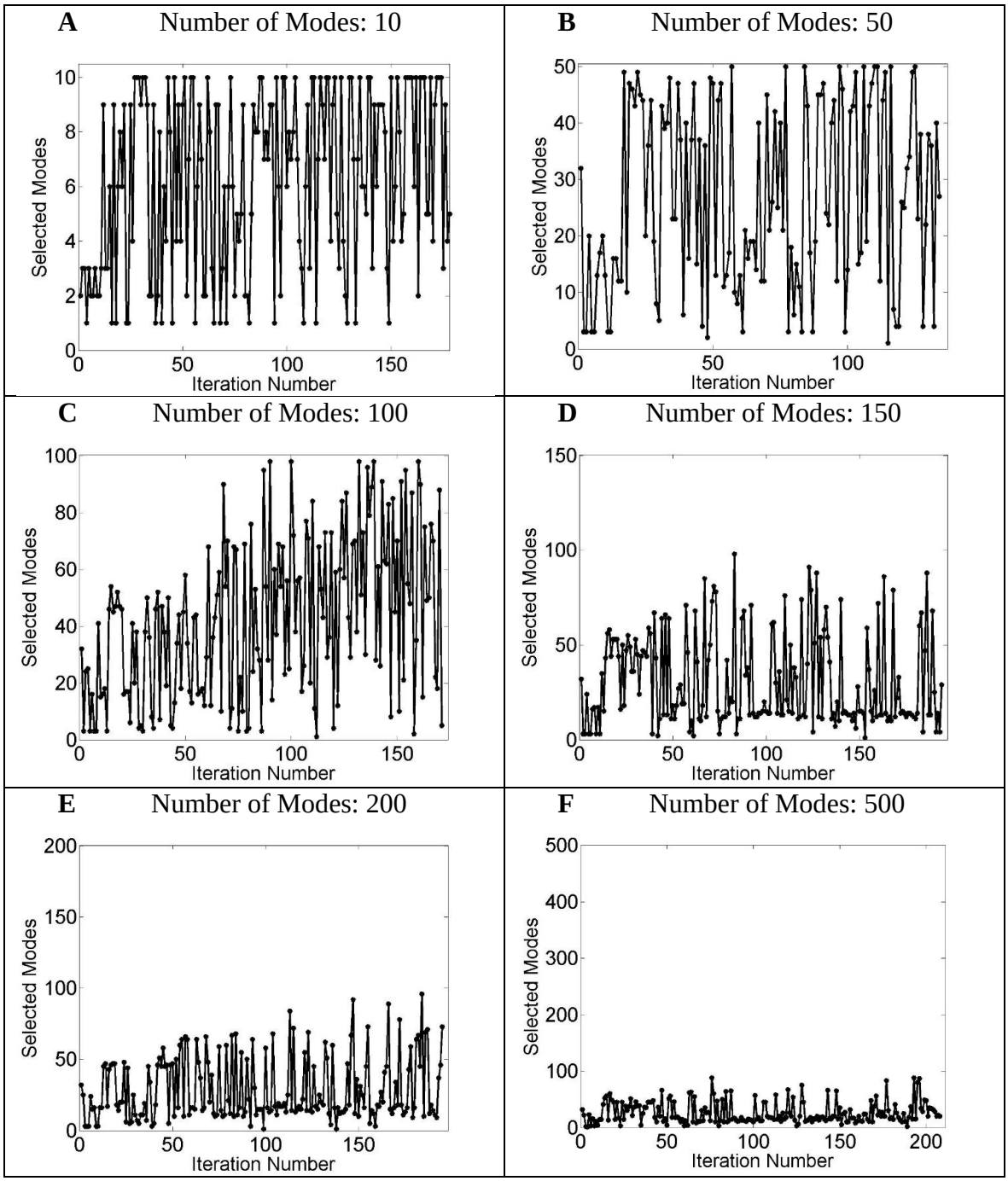


Figure 5.3. Profiles of selected modes among the lowest 10, 50, 100, 150, 200 and 500 modes (written above of each plot) in ANM-MC simulations of GroEL V.

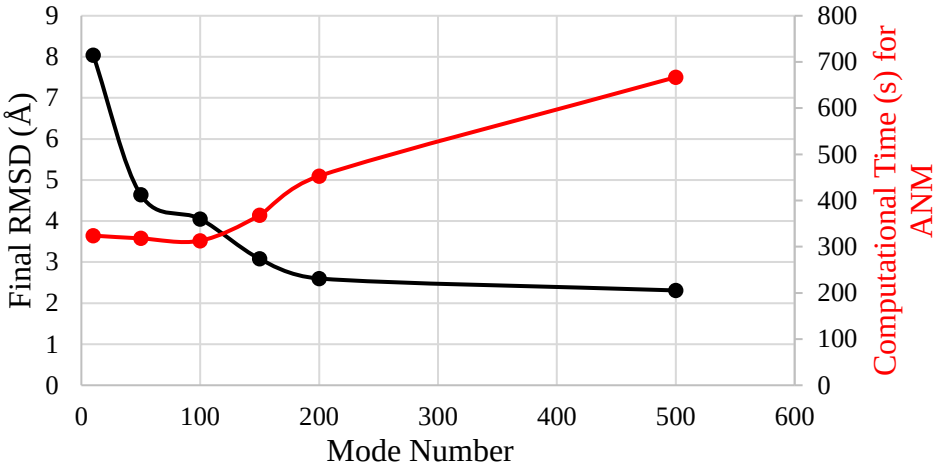


Figure 5.4. Final RMSD (Å) reached in ANM-MC simulations of GroEL V with different number of modes and their computational time spent for a single ANM.

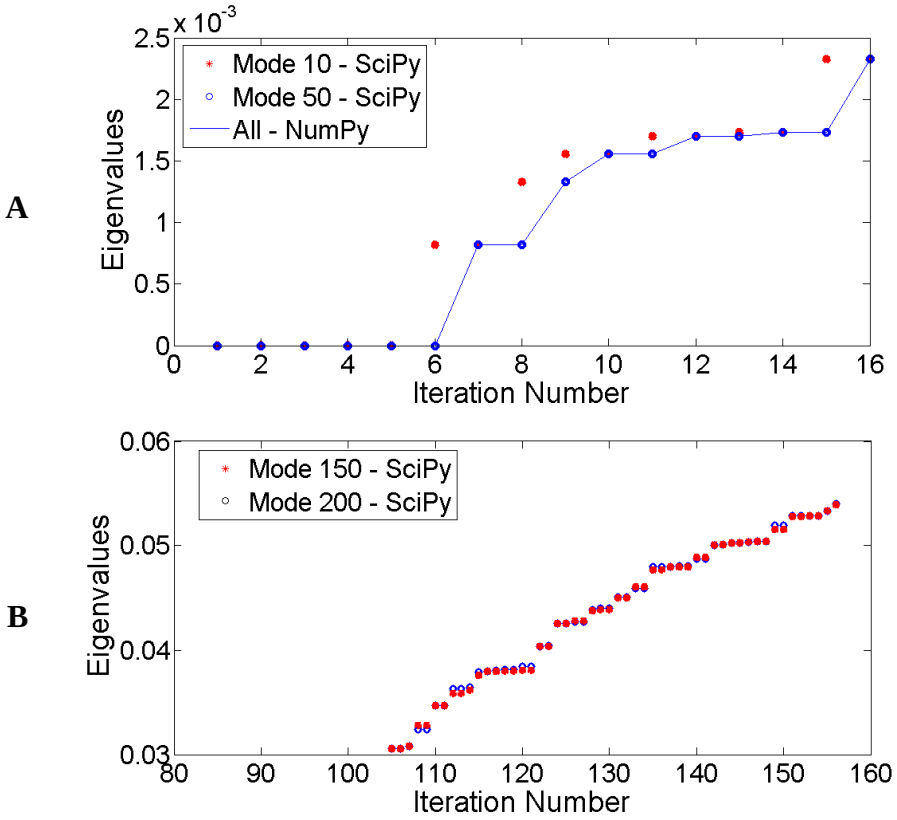


Figure 5.5. Eigenvalue profiles at 1st iteration of GroEL V. A) The first 10 and 50, and all eigenvalues found by *eigsh* in SciPy and *eigh* NumPy, respectively. B) The first 150 and 200 eigenvalues found by *eigsh* in SciPy.

5.4. Modified ANM-MC Methodology: Application of Variable Deformation Factors

Different deformation factors (DF = 0.2, 0.3 and 0.4 Å) in ANM-MC have been applied to AK and calmodulin in Section 4.5, as a result of which 0.3 Å was found optimum, providing both computational efficiency and satisfactory approach to target for large proteins. However, there may be a risk of observing some geometrical failures in the snapshots with this higher DF, especially when more localized deformations are used in the higher modes. On the other hand, DF of 0.2 Å has been regarded relatively safer [34], but not sufficiently fast to be applicable to large systems.

Table 5.3. CPU times spent on ANM.

Proteins	**Number of Residues**	**Time (ANM) Mode 10 (s)**	**Time (ANM) Mode 50 (s)**	**Time (ANM) Mode 100 (s)**	**Time (ANM) Mode 150 (s)**	**Time (ANM) All modes (s)**
APS	1716	20				
ATCase I	926	6				
ATCase II	924	6				
Cpn I	7856	591	351	398	450	-
Cpn II	7856	640	321	377	409	-
CS	858	5				
EPSPS	1708	16	17	19		19
GluT I	1215	12	11	17	26	34
GluT II	1215	12	12	17	21	31
GluT III	1215	10	10	12	13	
GroEL I	7322	281	282	291	320	-
GroEL II	7966	351	335	357	388	-
GroEL III	8015	327	313	383	426	-
GroEL IV	7336	324	318	313	368	-
GroEL V	7336	331	283	290		-
Hb	576	2				

Table 5.3. CPU times spent on ANM (cont.)

Proteins	**Number of Residues**	**Time (ANM) Mode 10 (s)**	**Time (ANM) Mode 50 (s)**	**Time (ANM) Mode 100 (s)**	**Time (ANM) Mode 150 (s)**	**Time (ANM) All modes (s)**
LacI	944	2				
MYO	719	4	4			
QacR	744	4				
RnaP I	3666	78	68	205		
RnaP II	3666	67	67	72		
UPRTase	846	5				7

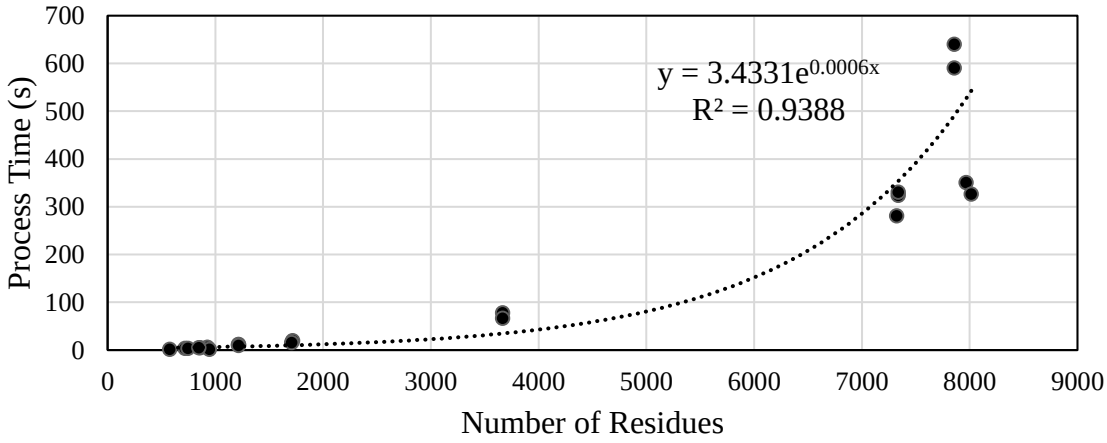


Figure 5.6. ANM process time with 10 modes with respect to number of residues.

It has been underlined in Section 4.2 that ANM is, indeed, the dominating factor in the conformational changes in ANM-MC simulations. It follows that the degree of deformation to which a protein conformation subjected is substantially resulting from ANM collective moves. Besides, exceeding the upper limit of disturbance beyond a protein structure can hold or stand may leave the generated conformational structures exposed to locally unrealistic geometries that cannot be relieved by MC simulation. When *DF* of 0.3 Å is used in the dataset, final snapshots of a few proteins are found to be subjected to such deformations,

which are reflected by outliers in the geometric parameters in virtual bond lengths on the backbone.

The solution would be letting the system have relatively small disturbances via ANM (i.e. *DF* of 0.2) for the cases where structures are exposed to highly changed ups and downs in geometric parameters and local dynamics along their pathways. However, a better way would be preventing conformations being deformed in advance. Needless to emphasize that once a conformation undergoes an excessive regional differentiation geometrically (i.e. with *DF* of 0.3), it loses its chance or ability to recover those regions in next iterations and unfortunately holds it along the rest of the ANM-MC trajectory.

As a solution, the modified version of ANM-MC is introduced (Figure 5.7) where small amount of ANM disturbances will be adapting, as a caution, before a possible and/or inevitable permanent deformation happen, in particular in intra-domain rearrangements. Large amount of *DF*s will remain to be applied for the sake of large steps along the way of target as long as the geometric features of each obtained structures allow. But, steps will be taken carefully this time. The geometric features will be controlled at each iteration before taking the next step. Once a region is found deformed too much, that same step will be retaken by a smaller *DF* and checked again.

DF of 0.3 is accounted as the largest *DF* value. It is the starting point at each iteration and it is lowered to 0.2 whenever needed. However, there are some structures observed of which lowering the *DF* to 0.2 becomes still insufficient and further decrease in *DF* is needed. At those times, *DF* of 0.1 is taken and that is the lowest value permitted. *DF* of 0.1 is the last chance for the structure which cannot escape from deformations allowed by 0.2. The reason of putting a stop at 0.1 is to enable the generated structure have a sufficiently large excursion while preventing it to stray from its intended pathway and eventually start to detour. Thus, the initial structure will be deformed along each tested *DF*s of 0.3, 0.2, and 0.1.

For the cases in which *DF* of 0.1 is not small enough to obtain a fairly acceptable conformation, last scenario for those cases is to reluctantly increase the number of MC steps by 5 each time until the structure starts to keep itself in plausible geometry. MC is computationally expensive for out of question. In Section 4.3, 15 MC steps are decided adequate to

fix/refine the AK and calmodulin structures yet computationally cheap to apply to large proteins. The number of MC steps have to be limited due to computational reasons. It is decided to be up to the number of 50 MC steps. The structure will be checked after each 5 additional MC steps and MC will be stopped if necessary (i.e. geometric features become plausible).

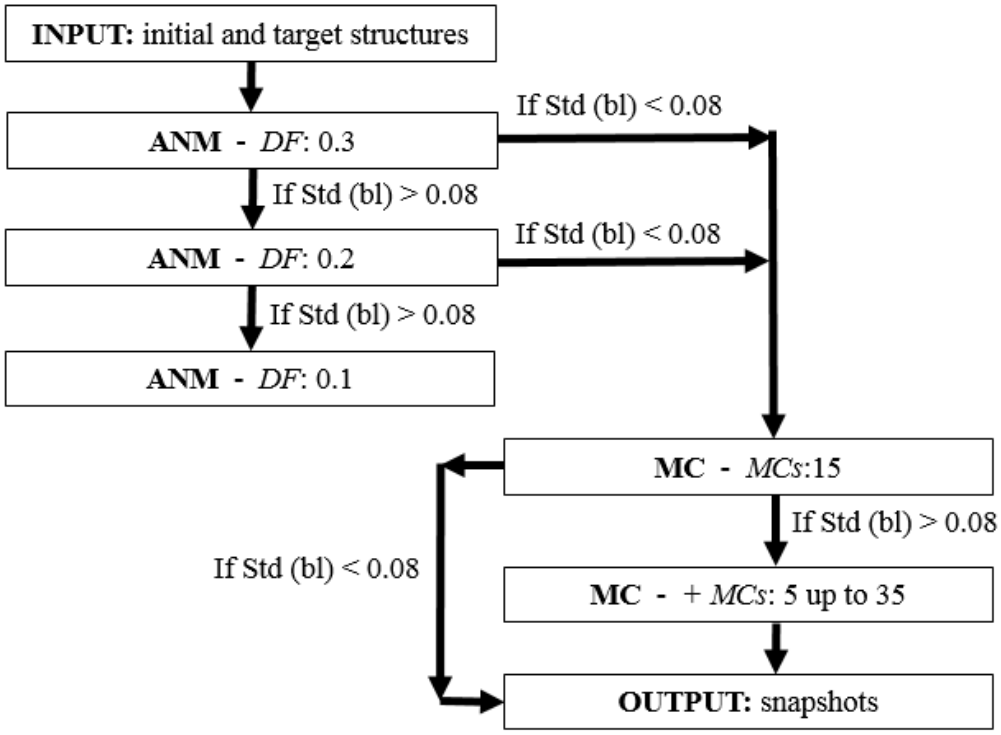


Figure 5.7. Flow chart of modified ANM-MC methodology.

Nevertheless, there should be a criteria to decide up if already ANM-made deformed structures undergo excessive deformations yet to avoid being permanently deformed. In this regard, geometric parameters, especially bond lengths, are always being a loyal determining factor in degree of plausibility of conformations. In Section 4, it has been kept in mind that virtual bonds should yield compression and expansion to an extent to let Cα atoms have harmonic motions sufficiently but avoid substantial motions that leads large differentiations in structures. So, the standard deviation of bond lengths at each iteration (0.08) is taken as a tolerance value and it is discussed in the following case study of myosin to adjust the value of variance that virtual bonds can have in ANM-MC simulations.

Modified version of ANM-MC will be assessed in the next section on myosin, later it will be applied to the all proteins in the dataset.

5.5. Application of Modified ANM-MC to Myosin

Myosin is a well-known ATP-dependent motor protein mostly found in muscle cells and its structure and function being conserved across species [208]. Myosin is subjected to a fluctuating environment. It is involved in transmission of information regarding the presence or absence of a γ-phosphate, coordination of ATP hydrolysis and transformation along a pathway of increasingly larger conformational changes [209].

Myosin is a 728-residue protein containing two domains: a 700-residue CORE domain (blue white), a 28-residue TAIL domain (orange) (Figure 5.8). Two conformations of myosin are used in the analysis of ANM-MC simulation from open to closed structures with respective PDB codes: 1VOM and 1MMA [171, 209]. The small-bottom domain (TAIL-orange) is mobile and exhibits a contractile motion which undergoes a relatively large conformational change by closing towards the CORE domain. The large-upper domain (CORE -blue white) is rather stationary which undergoes an elongational motion along a longitudinal direction [209].

In ANM-MC simulations of myosin, the desired substantial decrease in RMSD is not obtained within lowest 10 ANM modes, either as in the case study of GroEL. Low frequency modes recruit the initial global change of myosin up to iteration of 16^{th} and local structural changes are induced by the relatively higher modes in the rest of the simulation (Figure 5.9). Large conformational change is mainly observed in the initial structure of small-lower domain (orange) towards its final structure (green) (Figure 5.8). Higher modes enable this relatively mobile domain approach target further. The original RMSD of 4.7 Å between initial and target structures falls down to 2.5 Å and 2.0 Å within 10 modes and 50 modes, respectively (Figure 5.9).

Bond length distribution throughout the simulation, and energy profile together with RMSD to target are plotted in Figure 5.10. During ANM-MC simulation of myosin, implausible structures are observed, being generated starting from iteration 26^{th}. It arises from the

high *DF* value negatively affecting geometric features in local regions of myosin. This undesired effect on local geometry becomes double when higher modes start being selected. Some bonds are found over-extending in the mobile domain of myosin (Figure 5.11c). These extensions are considered beyond the limits that a protein structure can hold in reality [149, 180]. It is reflected as suddenly starting subsequent peaks in energy profile (Figure 5.11b), starting from iteration 26^{th}. Such kind of large extensions in some bonds make generated structures being implausible. To illustrate, Figure 5.11c is a screenshot of a small part of myosin which undergoes large extensions observed in virtual bonds between the residues of GLN - 203 and ALN – 205, being 5.9 Å away from each other. This is unquestionably out of normal distribution seen in Section 4.1 and quite beyond being acceptable.

It is presumed that these over-extensions occur due to relatively large *DF* value (0.3) and 15 steps of MC do not sufficiently refine deformed regions. A couple of benchmark tests are performed to score ANM DF and *MCs* effects not only over geometric parameters but also in terms of computational efficiency. Firstly, higher number of MC steps are applied in comparison with routine 15 *MCs* in fixing deformations. Secondly, less amount of deformation is given to monitor the *DF* effect on geometric parameters.

Figure 5.10 shows that energy and RMSD profiles of two different ANM-MC runs of myosin: increased *MCs* to 50 by keeping *DF* same (0.3) on the left, decreased *DF* to 0.2 by keeping *MCs* same (15) on the right. So, both cases are for recovering deformations and both of them achieve progress in energy and geometric profiles. However, higher number of MC steps are not able make myosin have more plausible conformations while it is computationally on the unfavorable side. When considering the process times that ANM takes about 4 min and 50 *MCs* take about 543 min, this is not surprising. So, using *DF* of 0.2 seems more promising, but its drawback is higher number of iterations that need to be taken to approach target. Nevertheless, both runs do not offer the intended results in the bit stream plots of bond lengths (Figure 5.11).

Lower *DF*s (0.1 and 0.05) are alternatively hoped for improvement in plausibility of virtual bond lengths (Figure 5.12). This case satisfies the expectations, in particular with *DF* of 0.05. However, when decreasing *DF* into half (i.e. 0.1 to 0.05), number of iterations becomes double (i.e. 118 to 239) which makes the computational time double (i.e. 5.5 hours to

11 hours), as well. On the other hand, one of the virtual bonds goes out of normal distribution and has a bond length of 4.81 Å with *DF* of 0.1, which is reflected as an undesired peak in energy profile appearing around 70th iteration.

In overall, keeping *DF* values low (0.2, 0.1 and 0.05) are obviously a better choice rather than increasing *MCs*, but higher *DF* is still worth to being utilized not to lose computational time with higher number of iterations.

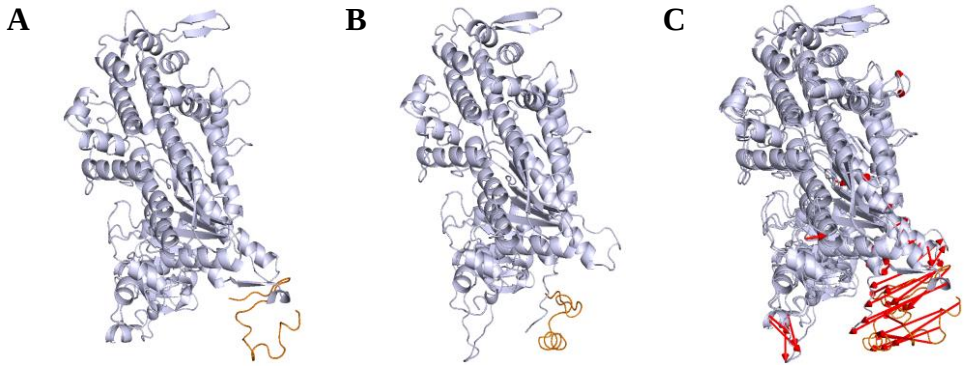


Figure 5.8. A) Open structure of myosin (PDB ID: 1VOM) B) Closed structure of myosin (PDB ID: 1MMA) C) Mode vectors of myosin (1VOM) aligned to closed structure (1MMA).

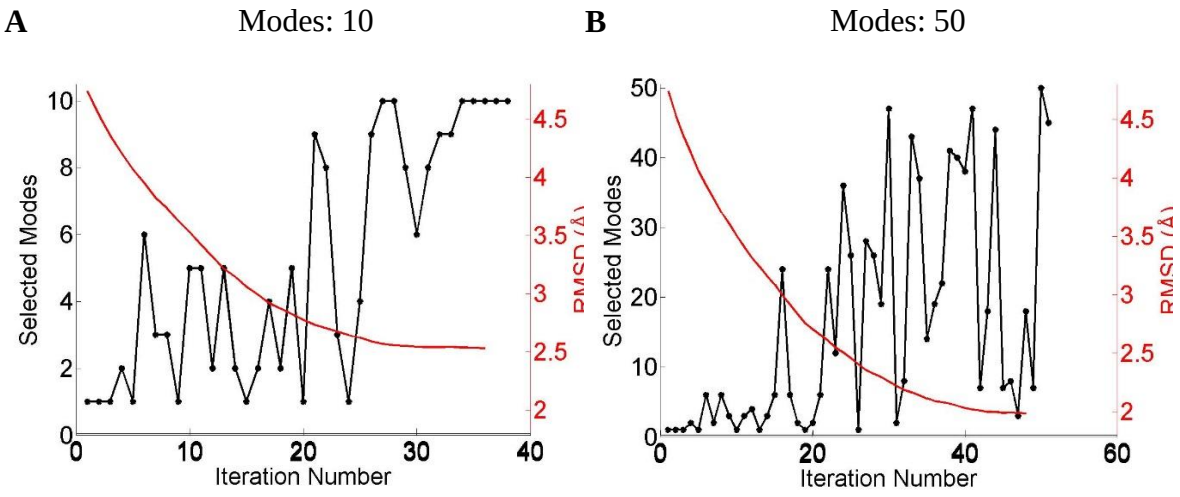


Figure 5.9. Mode profiles of myosin. A) Selected modes within first 10 modes. B) Selected modes within first 50 modes.

As the alternative solution introduced in Section 5.5, the initial myosin structure is deformed along tested *DF*s of 0.3, 0.2 and 0.1, in order. Virtual bond lengths are controlled at every single iteration by monitoring their extendibility margin. Figure 5.14 demonstrates the standard deviation profiles throughout ANM-MC trajectory of myosin, AK and calmodulin before and after being subjected to energy minimization. Standard deviation profile right after ANM, is indeed a reflection of its energy profile. For myosin, standard deviation of bond lengths reaches about 0.20 Å with ANM and can be decreased to about 0.15 Å with the help of MC. For AK and calmodulin, the highest standard deviation is seen as 0.082 and 0.056 and decreased to 0.074 and 0.053, respectively.

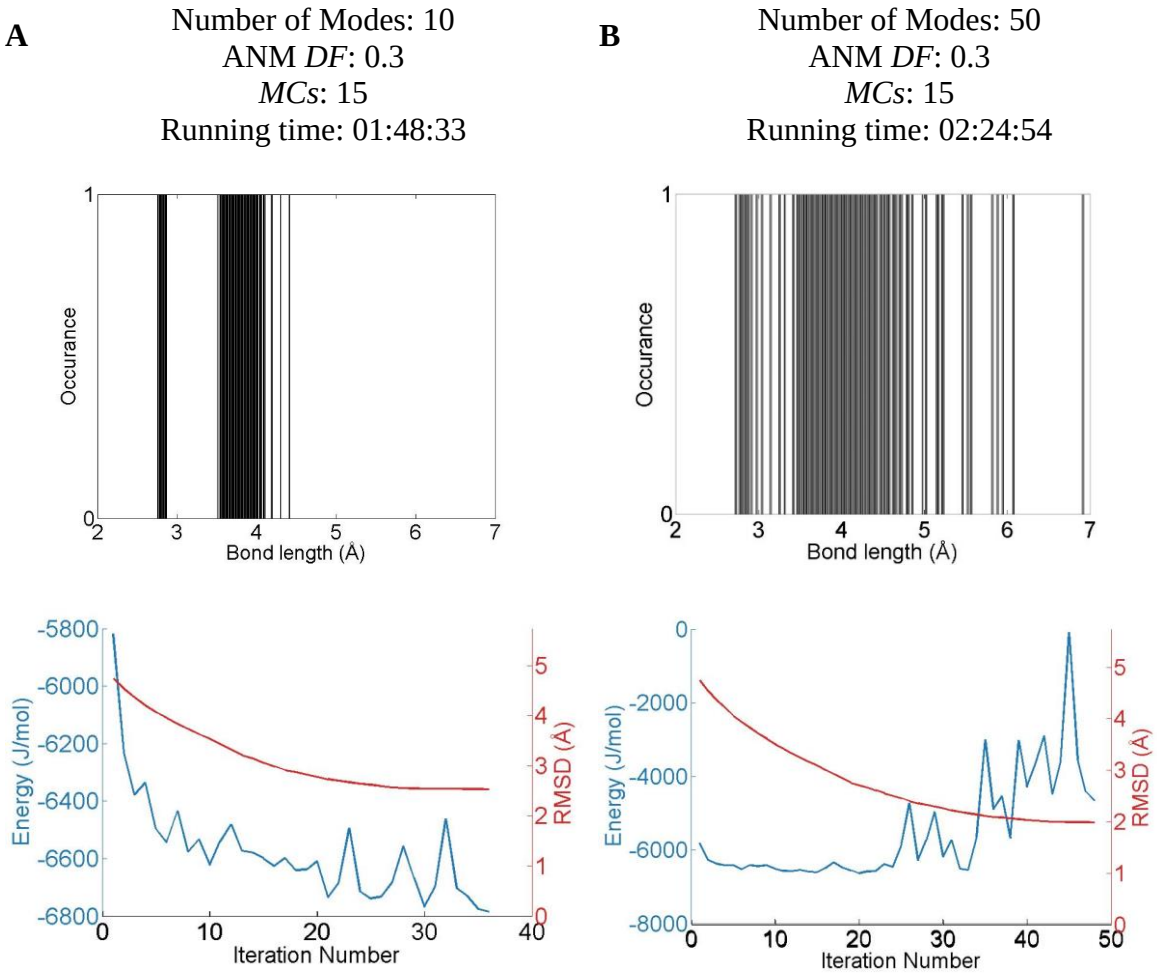


Figure 5.10. Bond length distribution, and energy and RMSD profiles of myosin with A) 10 modes and B) 50 modes (*DF*: 0.3, *MCs*: 15).

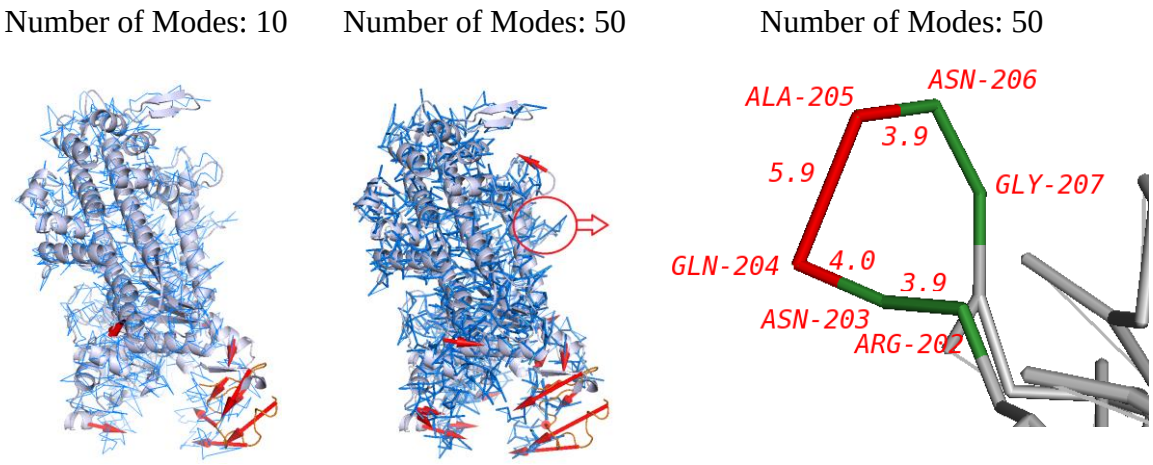


Figure 5.11. Open and closed X-ray structures (cartoon) and final conformation (ribbon) of myosin within 10 and 50 modes, respectively.

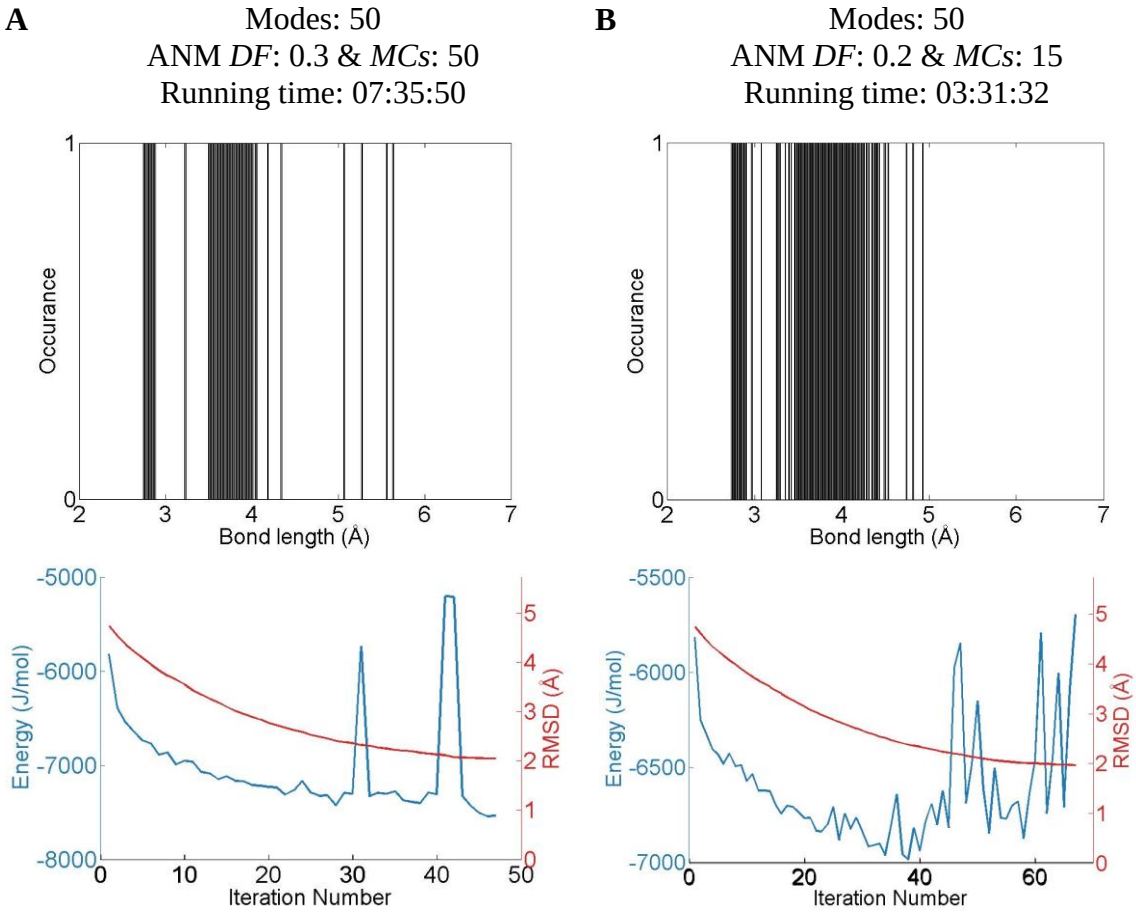


Figure 5.12. Bond length distribution, and energy and RMSD profiles of myosin. A) *DF*: 0.3, *MCs*: 50, Modes: 50. B) *DF*: 0.2, *MCs*: 15, Modes: 50.

A tolerance value needs to be determined here to limit excessive compressions and expansions of virtual bond lengths of myosin. The bond length profiles of AK and calmodulin are monitored again (Figure 5.14) and 0.08 seems an optimum choice as a limit that myosin can take at most.

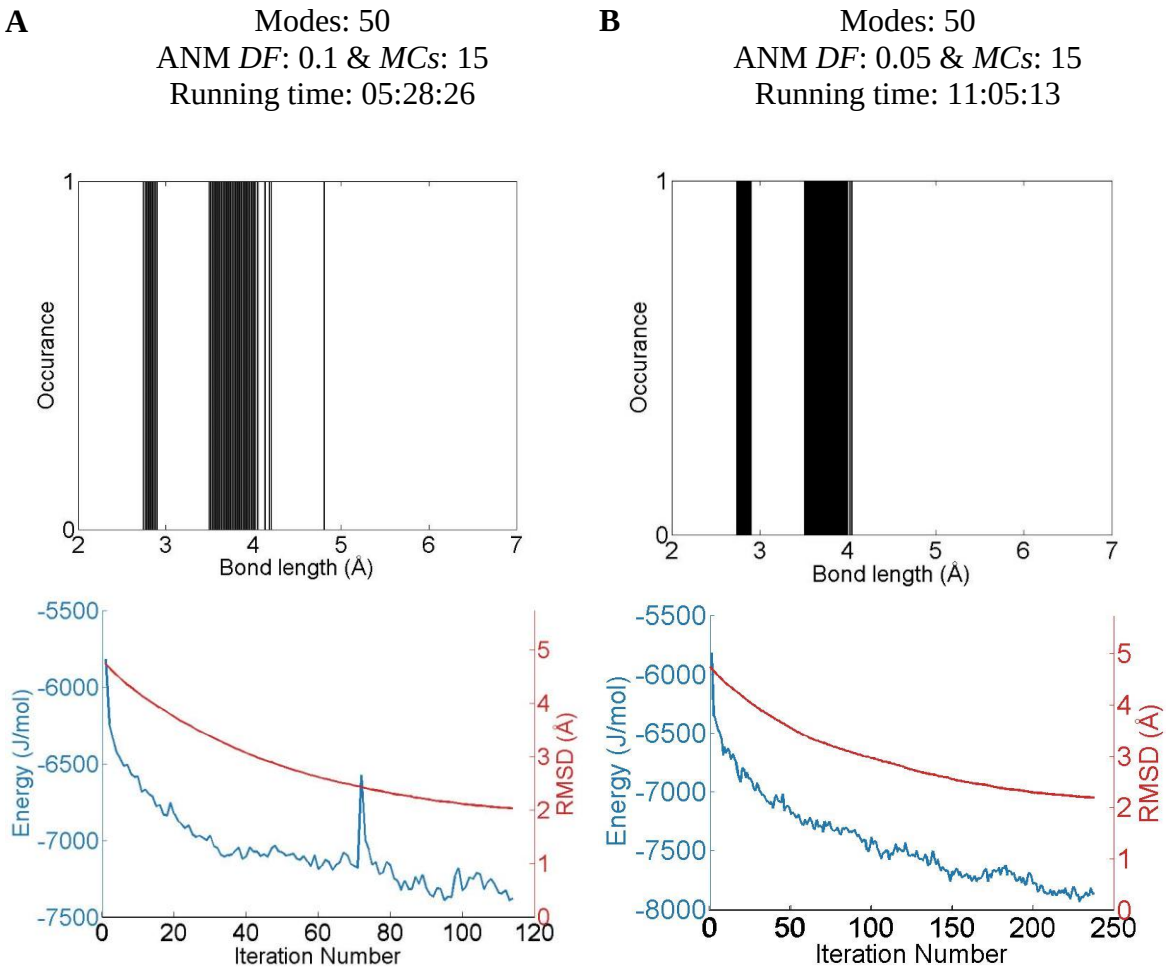
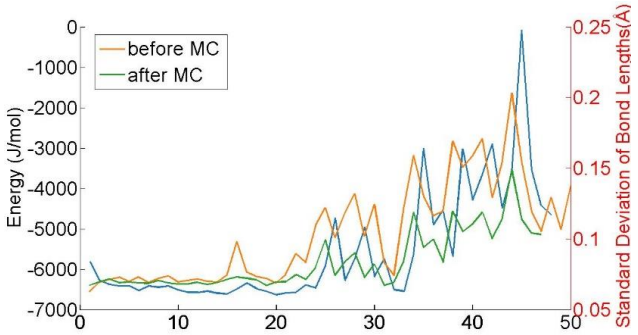


Figure 5.13. Bond length distribution, and energy and RMSD profiles of myosin. a) *DF*: 0.1, *MCs*: 50, Modes: 15. b) *DF*: 0.05, *MCs*: 15, Modes: 50.

Myosin shows a successful ANM-MC performance in terms of plausibility of generated structures and computational efficiency. In Figure 5.15, myosin starts with *DF* of 0.3 and keeps it until 20th iteration. That is expected since bond lengths of myosin has a standard deviation profiles with only little ups and downs (Figure 5.14). *DF* of 0.2 is accepted at some iterations as a supporter to 0.3. Whenever high peaks start to appear after 25th iteration in the

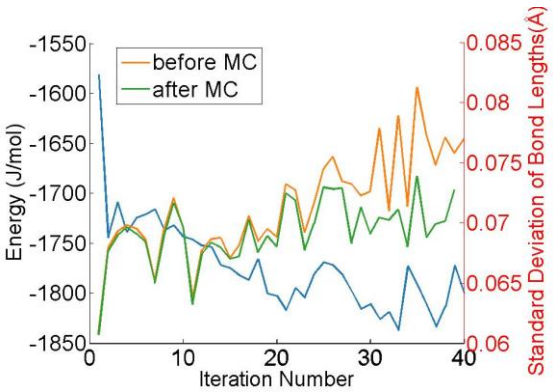
previous run (Figure 5.14), 0.1 is applied. Compared to previous run, lower *DF* value undoubtedly enable myosin decrease standard deviation of bond lengths to 0.085 and keep it within the boundary after the 25th iteration. Also, higher number of *MCs* are applied to drop high standard deviation at the iterations of 44th, 45th, 56th, 67th and 70th. Those iteration numbers directly corresponds to the highest peaks in Figure 5.15a. When examining the standard deviation profile after MC, those peaks seem disappear with the help of additional MC steps and it is directly proportional to energy profile. Besides, the bit stream plot of bond lengths are very similar to the one obtained with *DF* of 0.1 in a more efficient way. The wall time of the modified ANM-MC is 3.5 hours whereas it takes 11 hours for the same plausibility of myosin conformations.

Myosin (*DF*: 0.3, Modes: 50, *MCs*: 15)



AK

(*DF*: 0.3, Modes: 10, *MCs*: 15)



Calmodulin

(*DF*: 0.3, Modes: 10, *MCs*: 15)

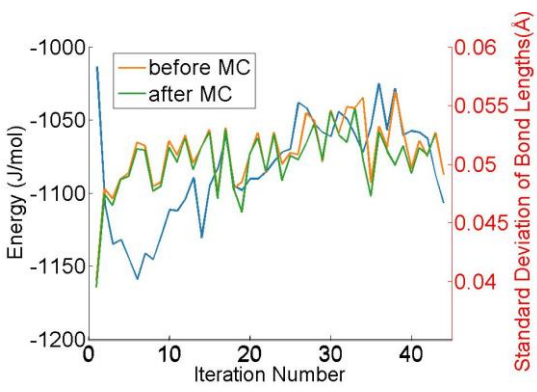


Figure 5.14. Standard deviation of bond length and energy profiles (blue) in ANM-MC runs of a) myosin (*DF*: 0.3, Modes: 50, *MCs*: 15) b) AK (*DF*: 0.3, Modes: 10, *MCs*: 15) and c) calmodulin (*DF*: 0.3, Modes: 10, *MCs*: 15).

5.6. Application of Modified ANM-MC Methodology to Large Proteins

ANM-MC simulations of all the proteins in the dataset along their conformational changes are monitored with the parameters adjusted in previous sections. Independent ANM-MC runs have been performed three times for each protein, but same profiles have been obtained in each one. Pathways are analyzed in detail by monitoring energy, overlap, collectivity and bond length profiles throughout ANM-MC runs. Collectivity and overlap values of selected modes and profiles of modes are plotted in Figure A.1 and A.2, respectively. RMSD and energy profiles are provided in Figure B1. Besides, ANM *DF*, MC steps and bond length profiles are provided in Figure B2 for all proteins in the data set.

5.6.1. Commonly Used Terms in ANM-MC Analysis

The degree of collectivity of a protein motion is calculated with the equation below:

$$\kappa = \frac{1}{N}\exp(-\sum_{i}^{N} \alpha \Delta R_i^2 log \alpha \Delta R_i^2) \tag{5.1}$$

where ΔR_i is the amplitude of the displacement of atom I [21]. In this thesis, it is used to estimate the degree of collectivity of each conformational change reflecting the number of Cα atoms which are significantly affected during ANM-MC trajectories.

Another measure is overlap value which is a determining factor for a successful and fast ANM-MC runs. In general case, the overlap of each slowest mode with the target direction presents a high overlap value particularly chosen after low frequency modes.

5.6.2. ANM-MC Parameters Used in the Rest of the Thesis

Different ANM cutoff (10 and 15 Å) and modes (10, 50, 100 and 150) are used for each protein, as mentioned in Sections 5.1 and 5.3, respectively. Just to state briefly again, cutoff value of 10 Å is used for the proteins with number of modes in parenthesis: ATP Sulfurylase (10), Aspartate Transcarbamoylase (10), Hemoglobin (10), Lac Repressor (10), 5-enolpyruvylshikimate-3-phosphate synthase (100), Uracil Phosphoribosyltransferase (10)

and Hypothetical Transcriptional Regulator in QACA (10). Cutoff value of 15 Å is used for the proteins with number of modes in parenthesis: Citrate synthase (10), Glutamate Transporter (50 and 150), myosin (50), RNA Polymerase II (10) and GroEL (10, 100 and 150) (Table 5.2).

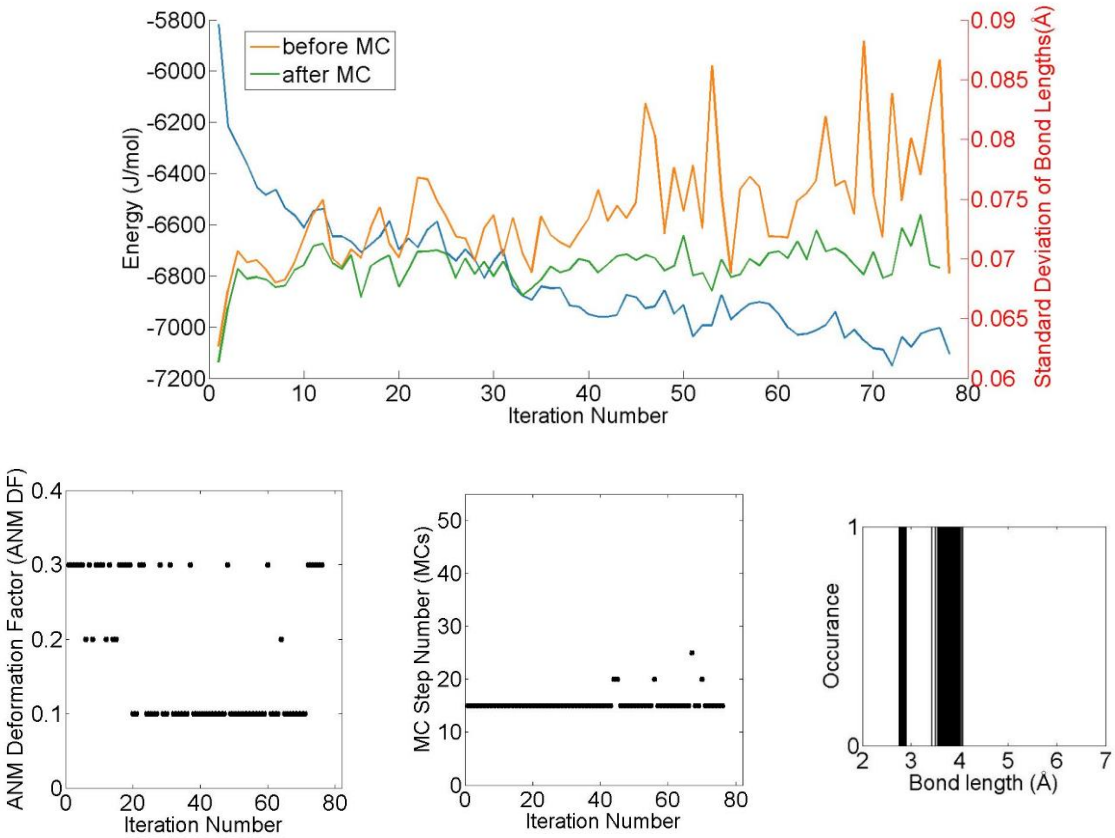


Figure 5.15. Standard deviation of bond length and energy profiles in ANM-MC runs of myosin with the updated version of ANM-MC program (Wall clock time: 03:43:14).

Combination of three trials of *DF*s are used as 0.3, 0.2 and 0.1, in order, by being cut once standard deviation of any bond length exceeds 0.08. Energy minimization parameters (k = 500 J/Å/mol and *PS* = 0.1 Å) are kept constant except *MCs*. It starts with 15 *MCs* and applies more *MCs* up to 50 whenever necessary.

The extent of application of lower *DF*s (0.2 and 0.1) affects final RMSD values reached in modified ANM-MC, as well. Figure 5.16 shows final RMSD values reached in

both ANM-MC and modified ANM-MC runs for each proteins. The pairs of proteins being deformed dominantly along *DF* of 0.3 in modified ANM-MC do not exhibit difference in final RMSD values while the proteins visiting lower *DF* values more lead final conformational structures being somewhat far apart from target structure. All runs reaches final RMSD value below 3.0 Å except Glutamate Transporter which leads higher but still acceptable final RMSD.

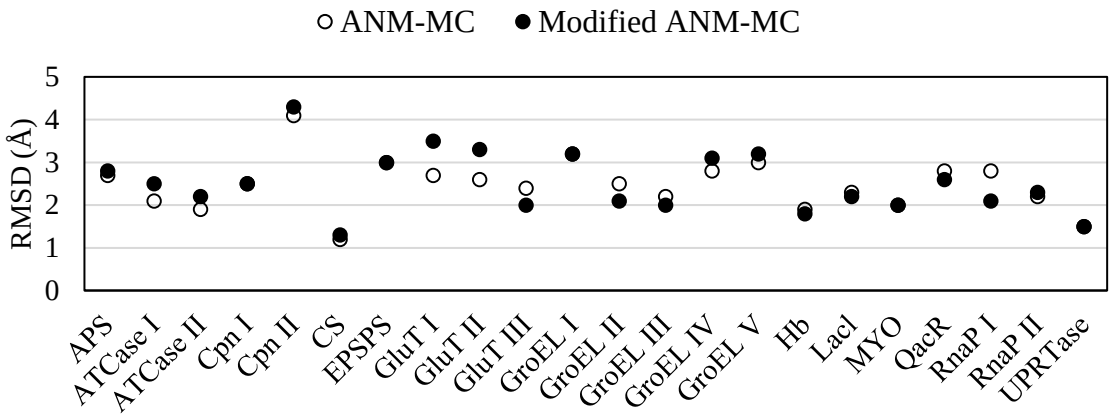


Figure 5.16. Final RMSD values reached in usual ANM-MC and modified ANM-MC.

Table 5.4. Initial and Final RMSD values for selected modes in modified ANM-MC runs. The best results are shown in bold for each protein.

Proteins	**Initial RMSD (Å)**	**Final RMSD (Å) Mode 10**	**Final RMSD (Å) Mode 50**	**Final RMSD (Å) Mode 100**	**Final RMSD (Å) Mode 150**
APS	4.4	**2.8**			
ATCase I	5.0	**2.5**			
ATCase II	4.9	**2.2**			
Cpn I	16.3	8.7	**2.6**	2.5	
Cpn II	16.3	9.4	4.7	**4.3**	
CS	2.8	**1.3**			
EPSPS	26.0	6.2	3.4	**3.0**	
GluT I	9.7	9.1	5.2	4.2	**3.5**

Table 5.4. Initial and Final RMSD values for selected modes in modified ANM-MC runs. The best results are shown in bold for each protein (cont.).

Proteins	**Initial RMSD (Å)**	**Final RMSD (Å) Mode 10**	**Final RMSD (Å) Mode 50**	**Final RMSD (Å) Mode 100**	**Final RMSD (Å) Mode 150**
GluT II	8.5	7.7	4.3	3.7	**3.3**
GluT III	4.9	3.6	**2.6**	2.0	
GroEL I	10.8	10.2	6.0	4.7	**3.2**
GroEL II	3.1	**2.5**	2.1		
GroEL III	2.6	**2.1**	2.0		
GroEL IV	6.7	5.6	3.5	**3.1**	
GroEL V	10.9	10.5	5.5	4.7	**3.2**
Hb	3.5	**1.8**			
LacI	14.6	**2.2**			
MYO	4.7	2.5	**2.0**		
QacR	20.4	**2.6**			
RnaP I	4.6	**3.1**			
RnaPII	2.4	**2.3**	2.1		
UPRTase	2.1	**1.5**			

5.6.3. Features of Conformational Transitions in Modified ANM-MC Runs

5.6.3.1. ATP Sulfurlyase (APS). ATP Sulfurylase is an enzyme that catalyzes the primary step of intracellular sulfate activation. It consists of three stacked rings of identical subunits with 572 residues divided into three distinct domains: N-terminal, catalytic and allosteric. [210, 211]. Rotational rearrangement of domains occurs as a result of the R (PDB ID: 1I2D) to T (PDB ID: 1M8P) transition, revealing a slight expansion and more open conformation in the T-state structure [211]. The crystal structures of the R and T states differ by the rotation of the allosteric domain upon inhibitor binding and a separate loop movement in the catalytic domain [212].

In ANM-MC run of ATP sulfurylase, initial overlap of the selected mode with target direction is 0.6 and decreases to 0.1 where RMSD approaches the plateau. The collectivity starts at 0.7 and decreases slowly to 0.6. It is observed that the catalytic domain shows more rigid and stable moves while three N-terminal and allosteric domains are relatively mobile and undergo both translational and rotational motions collectively, taking large steps (within 30 iterations). ATP sulfurylase is a flexible molecule and approaches target with plausible structures only selecting *DF* of 0.3 along the trajectory.

ATP sulfurylase undergoes conformational transition under low-frequency normal modes regarded as the collective degrees of freedom. It reaches half of its initial RMDS value at the end of its trajectory within first 10 modes. Krillova and Siméon *et al.* also examined ATP Sulfurylase involving large-amplitude domain motions and conformational transitions based on the combination of path planning algorithms and NMA. They concluded in the same way that low-frequency normal modes are able to let ATP Sulfurylase have large conformational transitions and the use of higher number of modes (up to 50) do not show significant performance improvement [213].

5.6.3.2. Aspartate Transcarbamylase (ATCase). Aspartate transcarbamoylase is an enzyme composed of two catalytic trimers and three regulatory dimers. Aspartate transcarbamoylase are found in two distinct conformations with binding of ATP or CTP: ATP favoring the active conformation (R state – PDB ID: 1D09) and CTP favoring inactive conformation (T state - PDB ID: 1RAC), respectively [214, 215]. Trimers favor being closer together in the inactive state. ATCase remains in T state in the absence of substrates. Its transitions from T to R state occur with bindings of substrates on an active site at the interface between adjacent catalytic subunits, which induces conformational changes accompanied by a global conformational change. During the presence of substrates, ATCase remains in the R state [216].

The pathway from R to T state (i.e. when substrates leave ATCase) is simulated via ANM-MC and plausible conformations are generated accordingly. The conformational changes result from the unbinding of substrates and attributed to the global motions [214, 215]. The pairs of R and T states are apart ~ 5 Å and get close ~ 2 Å far away. Its bond lengths have plausible amount of expansion/compressions considering being deformed along *DF* of 0.3 throughout the trajectory that all bond length distributions are observed even

within 4 Å. The conformational intermediate structures are generated with 10 modes, mostly by 1st and 4th modes. This tells that low-frequency normal modes play significant role in large-scale conformational changes of ATCase as indicated in a previous study, as well [217].

5.6.3.3. Citrate synthase (CS). Citrate synthase is an enzyme that consists of two identical subunits with 858 residues, but functions as a homodimer. Large conformational transitions involve relative movement of almost rigid domains of Citrate synthase [218] and hinge type of motions are observed [219, 220].

Citrate synthase exhibits low-frequency normal modes leading essentially large conformational changes. Although open (PDB ID: 5CSC) and closed (PDB ID: 6CSC) structures of Citrate synthase are so close (2.8 Å), it reaches final RMSD of 1.2 Å within 10 modes. The third lowest frequency mode is often selected along the trajectory, as stated in a previous study by Marques and Sanejouand *et al.* [88]. Tama and Sanejouand *et al.* states that a single low-frequency normal mode (i.e. 3rd mode) determines the direction motion and agrees well with the conformational change if the protein structure shows highly collective motions together [21]. However, Citrate synthase starts with a relatively low collectivity (~ 0.4) and travels between 0.3 and 0.5.

5.6.3.4. Chaperonin Lidless Mm-cpn (Cpn). Chaperonins capture non-native proteins and promote folding in an ATP-dependent manner. Group 2 Chaperonins help thousands of different proteins fold into tertiary structures, called hydrolysis of ATP leading the conformational transitions of Group 2 Chaperonin [221]. Group 2 chaperonin is a large cylindrical complex similar to GroEL. ANM-MC simulations are performed open (PDB ID: 3IYF) and closed (PDB ID: 3J03 (*in vitro*), 3LOS (*in vitro*)) forms of Lidless Mm-cpn composed of eight identical rings or subunits face back-to-back, having an overall toroidal-like shape. Each subunit consists of apical, intermediate, and equatorial domains, again, as in GroEL [222].

To the best of our knowledge, the dynamics of Group 2 Chaperonin has not been examined in detail whereas the opening and closing mechanism of GroEL-GroES have been discussed by many studies using a coarse-grained model [24, 98, 223]. Simulating Group 2

Chaperonin is computationally expensive even with coarse grained system. Lee and Kim *et al.* studied Lidless the dynamics of Mm-cpn applying SCENM (symmetry constrained elastic network model) to only 6 subunits by using the advantage of its symmetric molecular structure to gain a computational advantage [222].

Two different ANM-MC runs are performed: i) from nucleotide-free (open – *in vitro*) to nucleotide-induced (closed - *in vitro*) and ii) from open (*in vitro*) to closed (*in vivo*). Both runs need higher indexed modes. However, the second run (from *in vitro* protein to *in vivo* protein) does not yield successful RMSD profile as much as the first run (from *in vitro* protein to *in vitro* protein). Unfortunately, *in vivo* structural information is not available for open state of Mm-cpn.

Chaperonin Lidless Mm-cpn has plausible conformational structures although higher indexed modes are recruited in contrast to Glutamate Transporter and GroEL. Additionally, generated structures are plausible even if higher amount of deformation is being applied along the entire trajectory and higher modes start being involved right at the beginning of pathway (~ 20^{th} iteration). The initial overlap value is ~ 0.7. It starts to decrease once higher modes get being selected. The collectivity profile starts at 0.4 and shows ups and downs between ~ 0.4 and ~ 0.6.

5.6.3.5. Glutamate Transporter (GluT). Glutamate transporter is a neurotransmitter transporter protein that catalyzes glutamate across a membrane. Its biomolecular conformational transitions are essential to biological functions are implicated in stroke, epilepsy and neurodegenerative diseases in the central nervous system [224]. It is a complex macromolecular system and needs to be found in many conformational states to perform such functions. Here the conformational transition pathways of the excitatory amino acid transporters (EAATS) are examined, which are responsible for secondary active transport of amino acids like glutamate and aspartate across the plasma membrane. Aspartate transporter GltPh is a bowl-shaped homotrimer protein resolving functionally in multiple states and used as a structural prototype which consists of three protomers with eight TM helices and two helix-turn-helix motifs at the substrate-binding core [224]. The structural changes in all three protomers enable the transport of substrate via visiting between the outward and inward facing states of

the trimer. Conformational states between the three crystal structures of GltPh (Inward facing state - PDB ID: 1XFH [224], intermediate state - PDB ID: 3V8G [225], outward facing state - PDB ID: 3KBC [226] are analyzed.

Das, Bahar and Roux *et al.* examined the transition between the intermediate and inward facing states constructing a physically reasonable pathway between two endpoints of a conformational transition using ANMPathway. Accordingly, the transport domain does not exhibit change in its internal structure at the beginning of the transition and then starts to have local arrangements. The trimerization domain does not undergoes large changes in its internal structure along the entire pathway while the long flexible loop in the extracellular part performs significant movements. By comparison of the ANMPathway results with a series of conventional MD runs, the data collected from all-atom simulations yielded snapshots in accord with the transition pathway predicted by ANMPathway [7].

Glutamate Transporter shows a collectivity profile with successive ups-and-downs jumping between 0.1 and 0.5. On the other hand, its overlap profile starts around 0.4 and decreases to ~ 0.2 somewhere in the half of pathway and reaches ~ 0.1 at the end.

Glutamate Transporter is one of the proteins in the dataset that was deformed excessively along its trajectory when regular ANM-MC methodology was applied. To illustrate, the virtual bond between 414-B-LYS and 415-B-THR started with 3.81 Å and stretched up to 9.53 Å at an iteration somewhat close to the end of the pathway. When it is simulated with modified version of ANM-MC, that bond is found between 3.76 - 4.11 Å, virtual bond lengths are observed to distribute much plausibly. However, two of virtual bonds show undesired but acceptable expansions to 4.48 Å between 21-A-ILE and 22-A-LEU and 4.58 Å between 126-B-GLN and 137-B-ALA at the 68^{th} and 359^{th} iterations, respectively. It undergoes combination of different amount of ANM deformations (mostly *DF* of 0.1) and higher number of MC steps are taken at some iterations as in 69^{th} and 359^{th} iterations.

Moreover, Glutamate Transporter needs higher indexed ANM modes (50-150) in order to undergo significant inter-domain motions by trimerization domain that closely maintains its internal conformation during the transition while lower indexed ANM modes (10) are

recruited by the transport domain that practically undergoes a downward rigid-body movement relative to the trimerization domain. The pairs of open and closed Glutamate Transporter are far apart (9.7 Å) in both the native states and barely reach RMSD of 9.1 Å within lowest 10 modes, but come close (5.2, 4.2 and 3.5 Å) within 50, 100 and 150 modes, respectively.

Jiang *et al.* also performed an ANM analysis for Glutamate Transporter, which indicates two major types of large-scale motions occur that are energetically more favorable: an asymmetric stretching/contraction and a symmetric opening/closing of the three subunits. Only the domains at top approach each other along both motions while domain at the bottom of the transporter remain rigid within the membrane [227]. In contrast, Kohn *et al.* proposed small-scale molecular motions at the core region are in glutamate transporters [228]. It transforms local motions around a rigid core domain more than global moves, which explains why it requires higher indexed modes to complete its transformation to target.

5.6.3.6. GroEL. Conformational changes of intact GroEL along with mode profiles are already discussed in the Case Study of GroEL in Section 5.3. Only features of conformational transitions of GroEL obtained with new ANM-MC methodology will be provided in this section.

Five different pathways between GroEL pairs are generated with Modified ANM-MC. All the simulations provide smooth energy profiles, which is indeed sign of plausible conformational structures generated along the pathway. However, the bond length profiles tell totally the opposite. On the other side, neither *DF* profiles nor MC step profiles reflect those excessively extended bond lengths. If so, those significant extensions should have been reflected to ANM and MC parameters. In such cases, *DF* would keep the generated structure for the tree trials and MC would minimize its energy up to 50 steps. There is no such case not only in GroEL but also in all other proteins.

Contrary to doubts over the performance of modified ANM-MC, such long virtual bond lengths in generated structures are, in fact, sourcing from the reference X-ray structures (PDB ID: 1GRU, 2C7E). As an illustration, 1GRU has a virtual bond length of 7.28 Å between the residues of GLU-191 and GLU-192 in H, I, M and N chains. It is suspected due

to the level of detail present in the atomic structures of GroEL. It is possible that the quality of the data collected on the crystal containing GroEL might not be fine because GroEL has local flexibility or motion and 1GRU has a really low resolution as 12.5 Å. The atomic details in those residues are still kept in order not to lose the intact shape of GroEL

5.6.3.7. Hemoglobin (Hb). Hemoglobin is and allosteric protein which undergoes conformational transition between the tense (T, unliganded) and relaxed (R or R2, O2 or CO-bound) forms. Xu and Bahar *et al.* examined R2 conformation which is closer to the oxygenated state of Hemoglobin to understand the molecular mechanisms or intersubunit interactions. They showed that transition from T to R2 is driven by global modes in GNM and ANM. Indeed, oxygen binding involves a local structural change, but the region it perturbs corresponds to a global hinge region. This explains the allosteric behavior of Hemoglobin and how global hinge bending motions lead collective changes in the pathway of T to R2 [169].

Conformational transition of Hemoglobin is dominantly driven by the third lowest frequency mode by 40 % as in the case of Citrate Synthase. In contrast, Hemoglobin has a relatively higher collectivity profile (~ 0.6 throughout the pathway) but not high sufficiently to state confidently that the third low-frequency normal mode determines the direction motion and compares well with the conformational change in Hemoglobin.

5.6.3.8. Hypothetical Transcriptional Regulator in QACA (QacR). QacR is a repressor protein that regulates expression of QacA [229], which consists of 2 chains with 744 residues. QacR is one of the two proteins that undergo the largest conformational transition with 10 low frequency modes. The distance between the open and closed structure decreases from 20.4 Å to 2.8 Å along with 150 iterations. The second mode is chosen more among the 10 lowest modes. Its overlap value keeps going around 0.4 until 120^{th} iteration. Besides, it collectivity remains around 0.7 - 0.8, which is one of the highest collectivity profile in the dataset. It always choses the first *DF* trial (0.3) and never fails in the plausibility of intermediate structures.

5.6.3.9. 5-enolpyruvylshikimate-3-phosphate synthase (EPAPS). It is the other molecule performing the most successful ANM-MC simulation with the amount of RMSD change.

EPSPS is an enzyme that consists of 4 subunits and 1708 residues. It exists in an open conformation in the absence of substrates and/or inhibitors. Marques *et al.* [230] identified that the apo EPSPS undergoes extensive conformational changes leading to a closed conformation upon binding to the substrate and/or inhibitor. The closing mechanism of EPSPS is induced by ligands embedded between the domains, where one domain twists against another [230, 231]. It starts with an initial RMSD of 26.0 Å and decreases to 6.2 Å, 3.4 Å and 3.0 Å within 10, 50 and 100 modes, respectively. It reaches ¾ of its pathway to target with the global modes. The structural changes induced by the lowest frequency modes allow for substantial decrease in the distance of 19.8 Å. For further decrease in RMSD, higher indexed modes recruit EPSPS at each step as it travels between in the last ¼ of the pathway. Higher frequency modes involve increasingly as EPSPS is proceeding close to the target. The dominance of soft modes (10) can be roughly considered as 19.8 Å in RMSD change, while it is 2.8 Å and 0.4 Å for the lowest 50 and 100 modes, respectively.

The run with 100 modes is favored with the larger decrease to 3.0 Å in RMSD. Up to iteration of 150th, soft modes are selected and overlap is ranging between 0.4 and 0.6. Relatively higher indexed modes (20-50) start being chosen between the iterations of 150th and 200th where overlap decreases from 0.4 to 0.2. After 200th iteration, higher modes (50-100) are selected as EPSPS rigidity increases [231] and enzyme is in favor of being deformed less and overlap decreases further from 0.2 to 0.1.

5.6.3.10. Lac Repressor (Lacl). The lac repressor is a DNA-binding protein involved in gene regulation in the metabolism of lactose in bacteria. It has two binding sites for lac operator and inducer molecule. The binding of LacL to a promotor on a lac gene inhibits the transcription of that lac gene. An inducer molecule steps into the transcription by binding to Lac repressor, which leads a conformational change with the release of the DNA, as well [232].

A large conformational change of LacL (RMSD: 14.6 to 2.2 Å) between DNA-bounded state (PDB ID: 1EFA) and Inducer-bounded state (PDB ID: 1TLF) is simulated with 10 modes with collectivity of constantly above 0.5. Soft modes are expected as Flynn *et al.* stated the dominant hinge region appeared in conformational motions of LacL [233]. Its initial overlap is ranging between 0.5 and 0.7 at the first 50 iterations where mode highest

indexed mode is 5. It sharply decreases from 0.5 to 0.1 in the following 30 iterations with relatively higher modes: modes of 6^{th}, 7^{th} and 8^{th}.Then, it gets the other half of the pathway with overlap of 0.1 with the selected modes of 9^{th} and 10^{th}.

It completes its conformational transitional pathway always accepting the first trial in *DF* (0.3) without any further MC help, as in usual ANM-MC, with plausible intermediate structures.

5.6.3.11. Myosin (MYO). Conformational transition of myosin is already discussed in the Case Study of myosin in Section 5.5 in detail. The overlap value of the selected modes is initially not high enough for myosin to exhibit a quick heading toward its target direction. The degree of collectivity of motions are also not so high though throughout the entire simulation indicating that the number of atoms affected during conformational change is not significant.

5.6.3.12. RNA Polymerase II (RnaP). Transcription of cellular genomes is carried out by essentially orthologous enzymes, multi-subunit DNA dependent RNA polymerases (RNAPs) [234]. RNA polymerases are supra molecular enzymes which construct RNA chains using DNA genes as templates to catalyze DNA transcriptions. RNA Polymerase II is the key enzyme in the catalysis of DNA-directed synthesis of mRNA in eukaryotic transcription. It is essential to life and found in all organisms and some viruses. It is important to understand the transcription between its folded structures.

RNA polymerase II is a clamp-like enzyme with the active site located near the center of the enzyme. Feig *et al.* described dynamics of RNA polymerase II starting from two crystal structures with open (PDB ID: 2E2J [175]) and closed (PDB ID: 2E2H [175]) trigger loop forms by molecular-dynamics simulations. It binds the L-shaped nucleic acid complex and dynamic arrangements and conformations are mostly observed in the active site of structures. Nucleic acid translocation occurs primarily in the simulations with an open trigger loop structure and active site nucleotide of the closed trigger loop structure shows a dominant movement closing towards the terminal RNA ribose [235].

RNA polymerase II is 10 and/or 12 subunit enzymes that depends on additional factors for transcription initiation, elongation, and termination. Cheung *et al.* provides the key aspects of RNA polymerase II initiation and elongation by providing crystal structures together which substantially advances our understanding of RNA Pol II. [236]. However, conformational transitions are still unclear how the 10 subunit Pol II core before binding to the subcomplex Rpb4/7 [174] generates the complete 12 subunit enzyme [237, 238] and how the complete enzyme has a closed conformation of the Pol II clamp domain, which only permits passage of single-stranded DNA to the active site [239].

Crystallographic structures of the 10-subunit PolII complex from Saccharomyces cerevisiae with open (PDB: 1I50 [173]) and closed (PDB: 1WCM [174]) conformations are used as starting structures for ANM-MC simulations. The transition pathway between apo and DNA-bound conformations of the yeast RNA polymerase, which is a hetero-10-mer with 3666 residues, is presented. The RNAP structures have a crab claw shape, mainly conserved core around the active site and a multifunctional clamp. The transition mainly occurs around clamp side, the mobile domain on one of the pincers, which displays a closing of the cleft due to the rigid body motion of the clamp [29]. Indeed, although large transitional conformations occur around clamp side, RNAPII is a large molecule and the overall RMSD between open and closed structures in the dataset is only 4.6 Å. It takes 35 iterations to complete its transition from its apo state to DNA-bound state.

Overlap value is a determining factor for a successful approach to target. The lowest initial overlap values are seen only in RNAPII and Glutamate Transporters, which are considered least successful runs in approaching to target structure. The overlap with target direction starts around 0.5 and decreases to ~0.1 around 20th iteration where RMD profile reaches a plateau. The energy of the conformations is stable along the transition, as expected and the geometry is kept plausible. Collectivity values are higher than 0.4 on average for the all proteins except myosin and RNA Polymerase II. In overall, ANM-MC simulation is able to reach to target structure with an RMSD of 2.8 A.

5.6.3.13. Uracil Phosphoribosyltransferase (UPRTase). Uracil phosphoribosyltransferase is an enzyme that catalyzes uracil to uridine monophosphate (UMP) and is involved in allo-

steric regulation with cytidine triphosphate (CTP). The conformational transition is simulated between UPRTase in complex with UMP (PDB ID: 1XTT) and UPRTase in complex with UMP and CTP (PDB ID: 1XTU) [240]. The initial RMSD is 2.1 Å, which is lower than the resolution of structures. So, it may not be considered a good case for analysis of an ANM-MC simulation, but Uracil phosphoribosyltransferase is a supra molecule with 4 subunits and 846 residues, which deserves to be in the dataset. Along the pathway, the selected modes exhibit a fluctuating trend while 8^{th}, 9^{th} and 10^{th} modes are chosen more. Its overlap starts from 0.5 and follows a decreasing path to 0.1 whereas the collectivity is above 0.6 and reaches even 0.9 at some iterations.

6. REVERSE MAPPING AND GEOMETRY OPTIMIZATION SIMULATIONS

The approach of modeling large chemical and biological systems at the coarse-grained level relies on the description of molecular interactions in a rather simple and computationally efficient way. However, it is also desirable to obtain atomistic intermediates or conformers by so-called reverse-mapping, which describes the transition from a low-resolution to a fine representation by refinement of the conformations [241]. It was shown in Section 5 that ANM-MC is an efficient algorithm that is applicable to very large systems. The plausibility of its intermediate structures in regard of protein geometry was also discussed in detail. At this point, they are ready to be mapped to full-atomistic structures.

Building the full atomistic model based on a coarse-grained structure requires creation of information, specifically arrangement of atoms in peptide backbones and side-chains. In the given approach here, the atomic details of conformations are introduced by the initial full-atomistic structure (i.e. open, crystal structure) and those details are hold by intermediate structures. Later energy minimization by explicit or implicit solvation method is used to refine the atomistic structure. All the details regarding the technique are provided in Section 3.4.

The reverse-mapping and energy minimization methods are initially tested on a small protein, AK, which undergoes a substantial conformational transition (RMSD: 7.13 Å to 1.35 Å) using first 10 slowest modes. Two independent AK runs are used, which have similar RMSD, energy and mode profiles (Figure 6.2). Snapshots from one run are optimized using implicit solvation, whereas energy minimization with explicit water molecules and periodic boundary conditions is performed for the second run. Computational details are provided in Section 3.4. The internal geometry of reverse-mapped structures is assessed using the MolProbity server [242].

For a physically realistic description of the system, the structures are examined with the objective of finding stable conformations via degree of allowed rotations represented by torsion angles phi and psi. In the Ramachandran diagrams of initial and final structures and

generated intermediates of AK in Figure 6.3, the only outliers are seen in open (PDB ID: 4AKE, Resolution: 2.2 Å, Ser129A) crystal structure and generated structure at 20th iteration with explicit solvation (Ala11A) of AK. The reason of observing an outlier at 20th iteration might be due to the high *DF* value (0.3), which might yield excessive deformation at that particular region. Although both ANM-MC simulations do not necessarily iterate parallel, highly similar Ramachandran plots are observed at the same iterations using implicit and explicit solvation models.

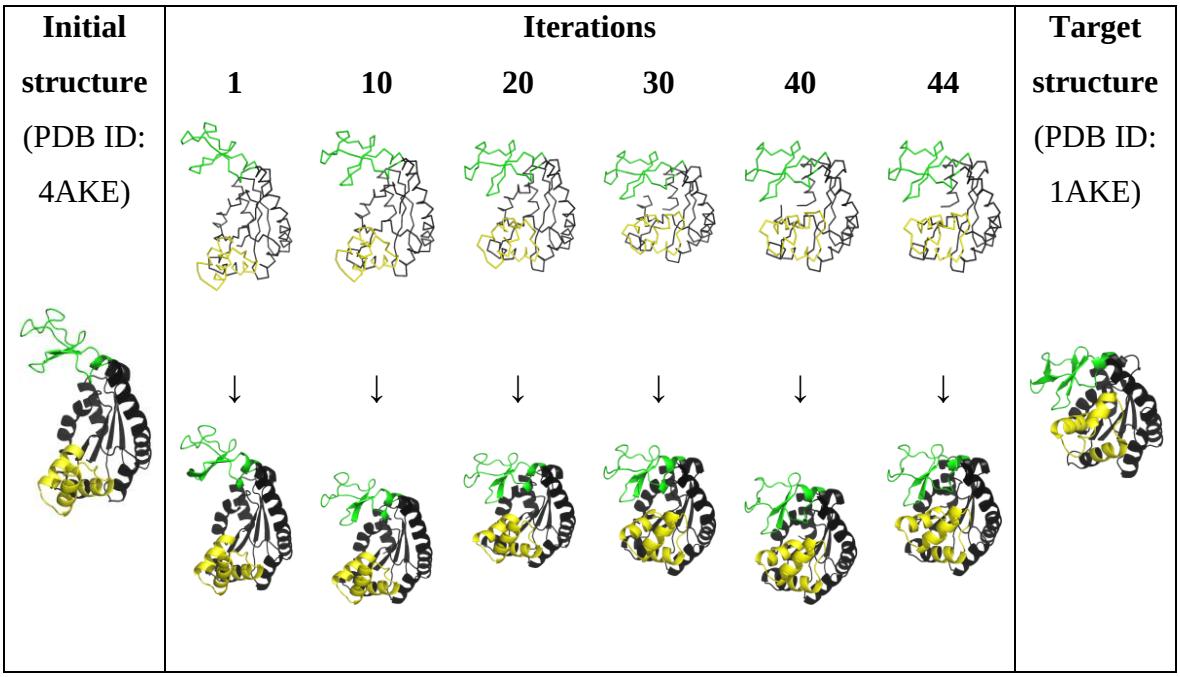


Figure 6.1. Coarse grained (ribbon) structures and full atomistic (cartoon) structures (optimized with explicit solvation model) of AK at different iterations.

As another indicator of geometric errors in the optimized models of AK, some statistics are provided about the favorable and undesired rotations in Figure 6.4. Accordingly, that both implicit and explicit solvation models share a similar profile and yield 95% of backbone rotations geometrically favorable and only 1% of them undesired (Figure 6.4a). On the other side, more rotamer outliers observed in AK sidechains with explicit solvation. However, the crystal AK structure (PDB ID: 4AKE) have much more unusual conformations of backbone and sidechains than the optimized ones. This means that the geometric mistakes in the very initial structure (crystal AK structure) are refined by energy minimization in the following

iterations. In overall, either implicit or explicit solvation model in geometry optimization offers favorable full-atomistic structures.

Moreover, wall-clock times are 60 min and 30 min on average for energy minimization using implicit and explicit solvation models, respectively. Since the elimination of water molecules in implicit solvent does not provide a computational advantage, explicit solvent is used in the rest of the simulations, which is usually considered more accurate than implicit solvent models.

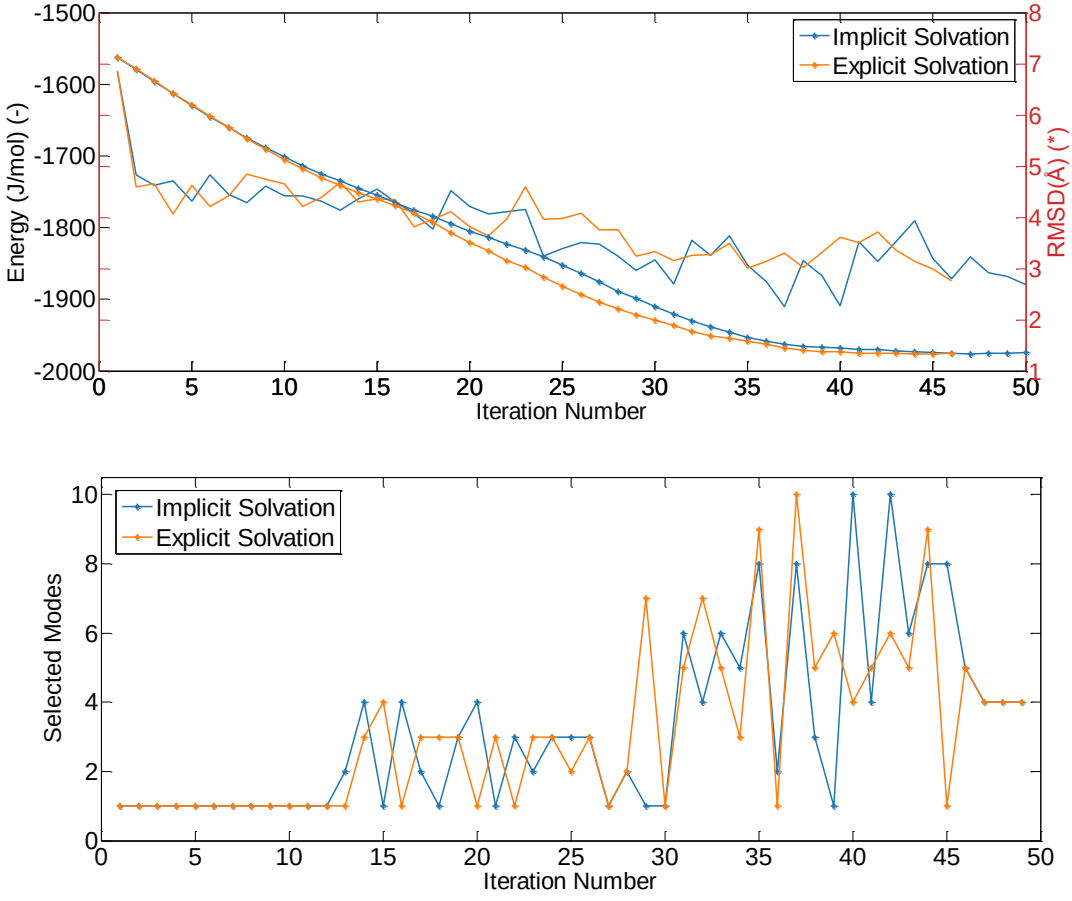


Figure 6.2. Energy, RMSD and mode profiles of two independent runs of AK using implicit and explicit solvation models, respectively.

The reverse mapping and energy minimization procedures are later applied to the trajectories of EPSPS (RMSD: 26.0 Å to 3.0 Å), GluT I (RMSD: 9.7 Å to 2.7 Å) and QacR

(RMSD: 20.4 Å to 2.8 Å), which undergo the largest conformational transitions in the dataset. Those runs are chosen not only to examine a substantial conformational transition of a large protein but also to see the effect of selected modes on ANM-MC snapshots.

The coordinates of unliganded forms of crystallographic structures of EPSPS at 2.2 and 2.3 Å resolution are deposited with PDB IDs of 1RF4 (open) and 1RF5 (closed), respectively. The reverse mapping mechanism is performed iteratively along the trajectory of 327 snapshots of EPSPS. Figure 6.5 demonstrates both coarse grained structures (ribbon) coming from ANM-MC (only minimized with MC) and full atomistic structures (cartoon) minimized with NAMD at 8 different iterations (1^{st}, 50^{th}, 100^{th}, 150^{th}, 200^{th}, 250^{th}, 300^{th} and 327^{th}).

Time evolution of the RMSDs of the EPSPS between relaxed full atomistic conformations and virtual structure at each iteration is shown in Figure 6.6. The reverse mapped form displays some conformational changes during the geometry minimization with RMSD changes between 2.1 to 2.2 Å.

EPSPS has four domains, each of which is composed of three βαβαββ folding units. Figure 6.7a and 6.7b demonstrates two chains of apo enzyme and inhibitor-bound enzyme. β-sheets structures of each unit contain both parallel and antiparallel strands whereas the α-helices are parallel. The order of the principal secondary structural elements are seen both open and closed forms. In the Ramachandran diagram of open and closed X-ray structures of EPSPS in Figure 6.7c and 6.7d, respectively, the outliers are observed in residue atoms of Met76ABC (PDB ID: 1RF4), and , Lys212A, Lys337AB, Ala356A, Thr341BD, Asp312C, Leu373C, Lys221D, Ser227D and Glu335D (PDB ID: 1RF5 - Chain B). All other residues have no steric clashes in their alpha-helical and beta-sheet conformations or are allowed to come a little closer together. All other residues have no error in their alpha-helical and beta-sheet conformations.

In Figure 6.8, the Ramachandran plots of EPSPS at certain iterations demonstrate how dihedral angles ψ against φ of residues distribute favored and energetically allowed regions or map as outliers. . The Ramachandran plots of EPSPS trajectory serve as a useful indicator of the quality of its three-dimensional structures generated by reverse mapping method. The

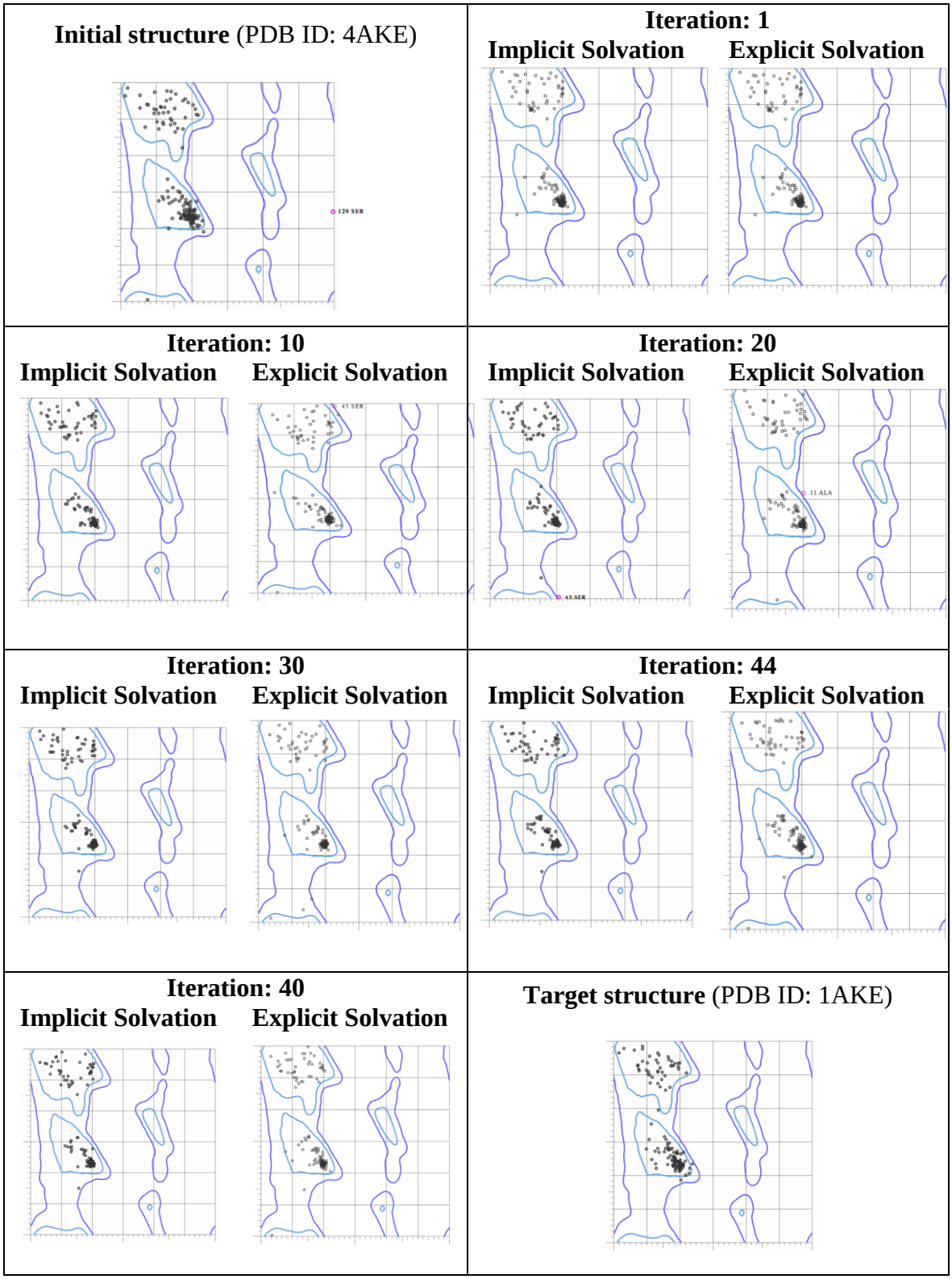


Figure 6.3. Ramachandran plots of crystal AK structures and relaxed AK conformations using implicit and explicit solvation models at different iterations.

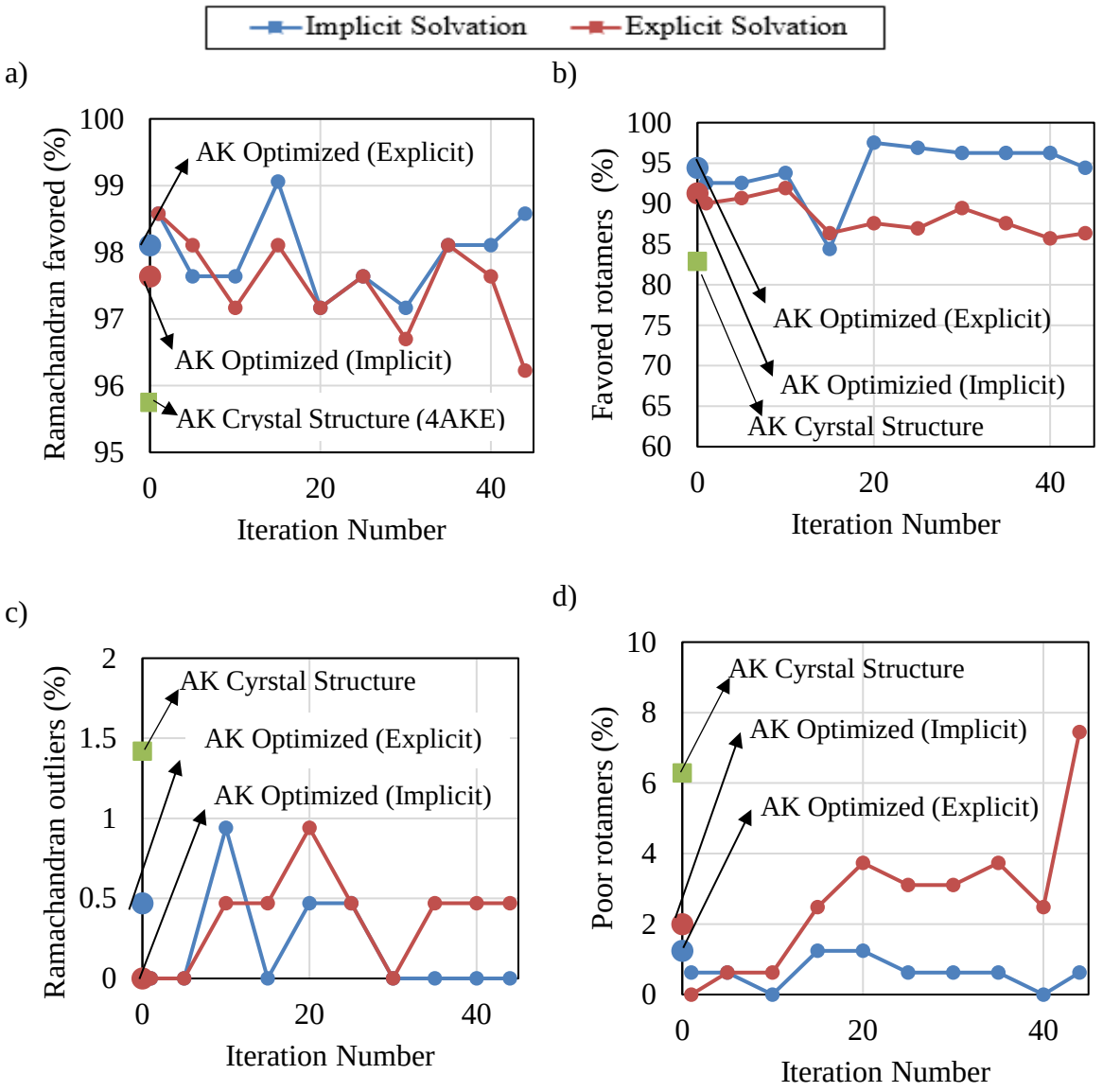


Figure 6.4. Statistics on favorable and undesired rotations around backbones and sidechains of AK at different iterations using implicit and explicit solvation models.

first conformation show regions fully allowed and slightly allowed for all amino acids together with Proline and Glycine. At 50th iteration, a few unfavorable conformations start being seen due to steric hindrance between polypeptide backbones. After 100th iteration, disallowed regions of torsion angle values are increasingly observed. Especially in the second half of the trajectory, incredible amount of sterically disallowed torsional angle rotations are found in the excluded regions.

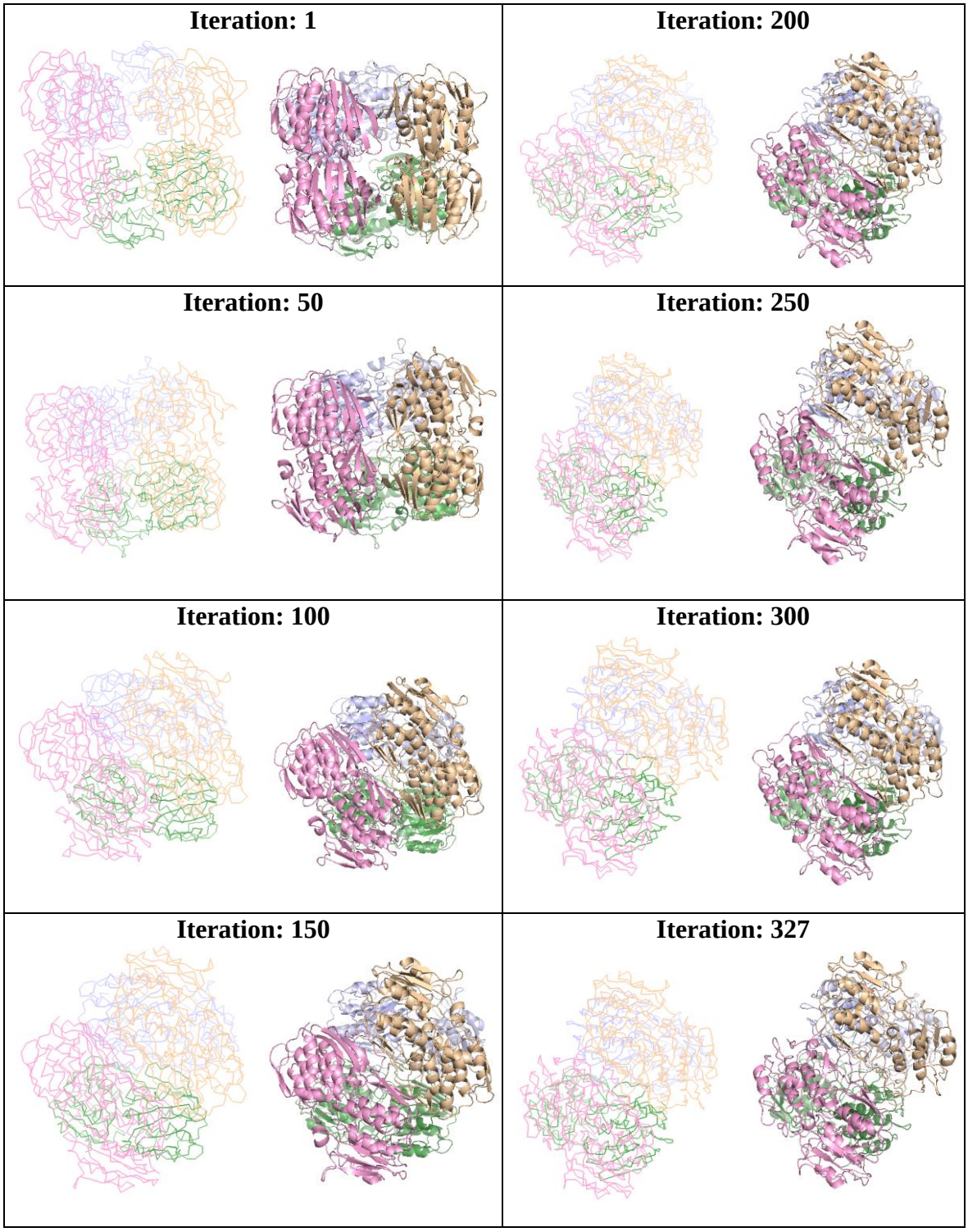


Figure 6.5. Coarse grained (ribbon) structures and full atomistic (cartoon) structures (optimized) of EPSPS found at different iterations.

In the assessment of the quality of protein full-atomistic structures, the overview of allowed and disallowed regions yield excessive amount of torsion angles outside the low-

energy regions as proceeded away from the original structure; which indicate problems in the generated full-atomistic EPSPS structures. This is, in fact, an expected outcome once the full atomistic structure starts not being in agreement with a fully plausible geometry even after it gets stabilized or relaxed. Full atomistic structures are based on previously generated ones. Namely, the plausibility of each generated conformation depend on the degree of plausibility of another computationally derived conformation, except the very first snapshot which is reversely mapped using the experimental coordinates of crystal structure.

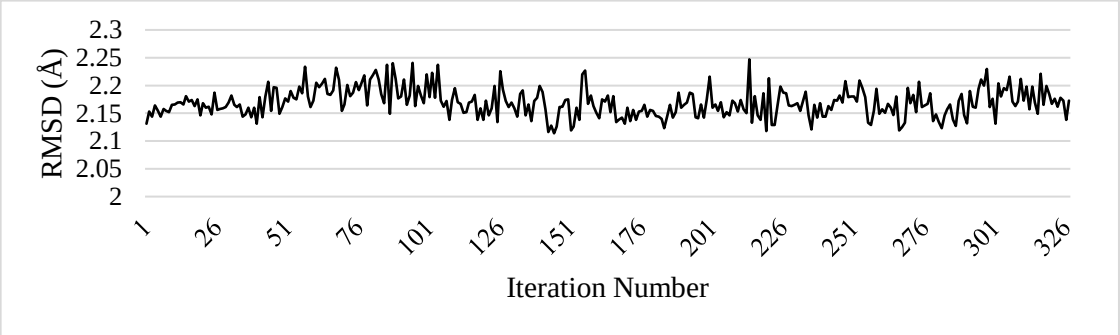


Figure 6.6. RMSD profile between coarse grain and full atomistic structures of EPSPS.

Additionally, the mode profile of EPSPS is examined to observe whether the underlying reason of observing so many outliers in conformations is the higher modes, i.e. more local deformations, being adopted in the second half of the trajectory. Figure 6.9 shows the correlation between the selected modes and the amount of Ramachandran outliers and poor rotamers in the conformations of backbones and sidechains of EPSPS, respectively. Figure 6.10 shows that EPSPS and GluT have a sharper decrease in Ramachandran favored backbones and a sharper increase in the amount of Ramachandran outliers compared to QacR. However, the same situation is not seen in the rotamer profiles. It should be stated that errors in optimized geometries are substantially sourced from the initial crystal structure used in ANM MC. Both GluT and QacR have a considerable number of outliers in their Ramachandran plots (Figure 6.11).

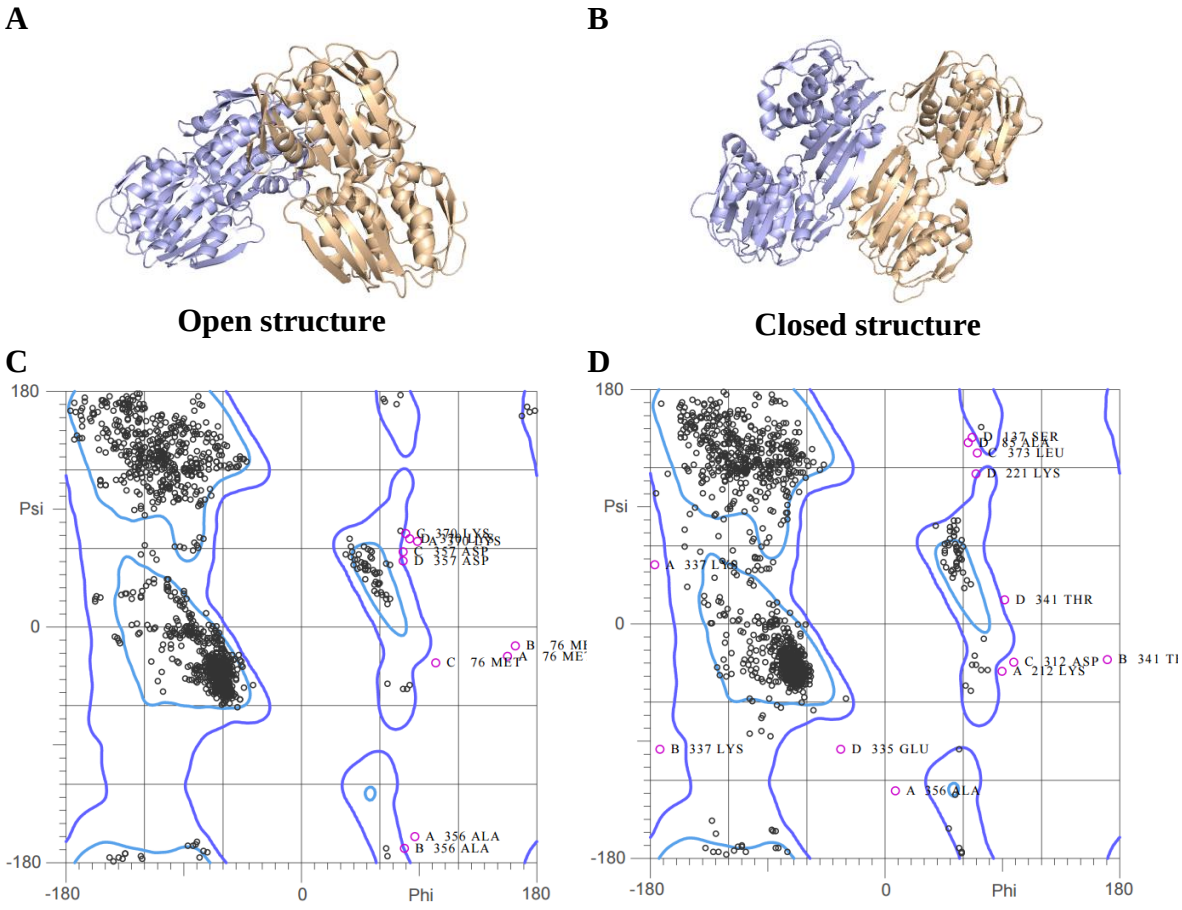


Figure 6.7. EPSPS. A) Open conformation (PDB ID: 1RF4) B) Closed conformation (PDB ID: 1RF5) C) Ramachandran plot of open conformation D) Ramachandran plot of closed conformation.

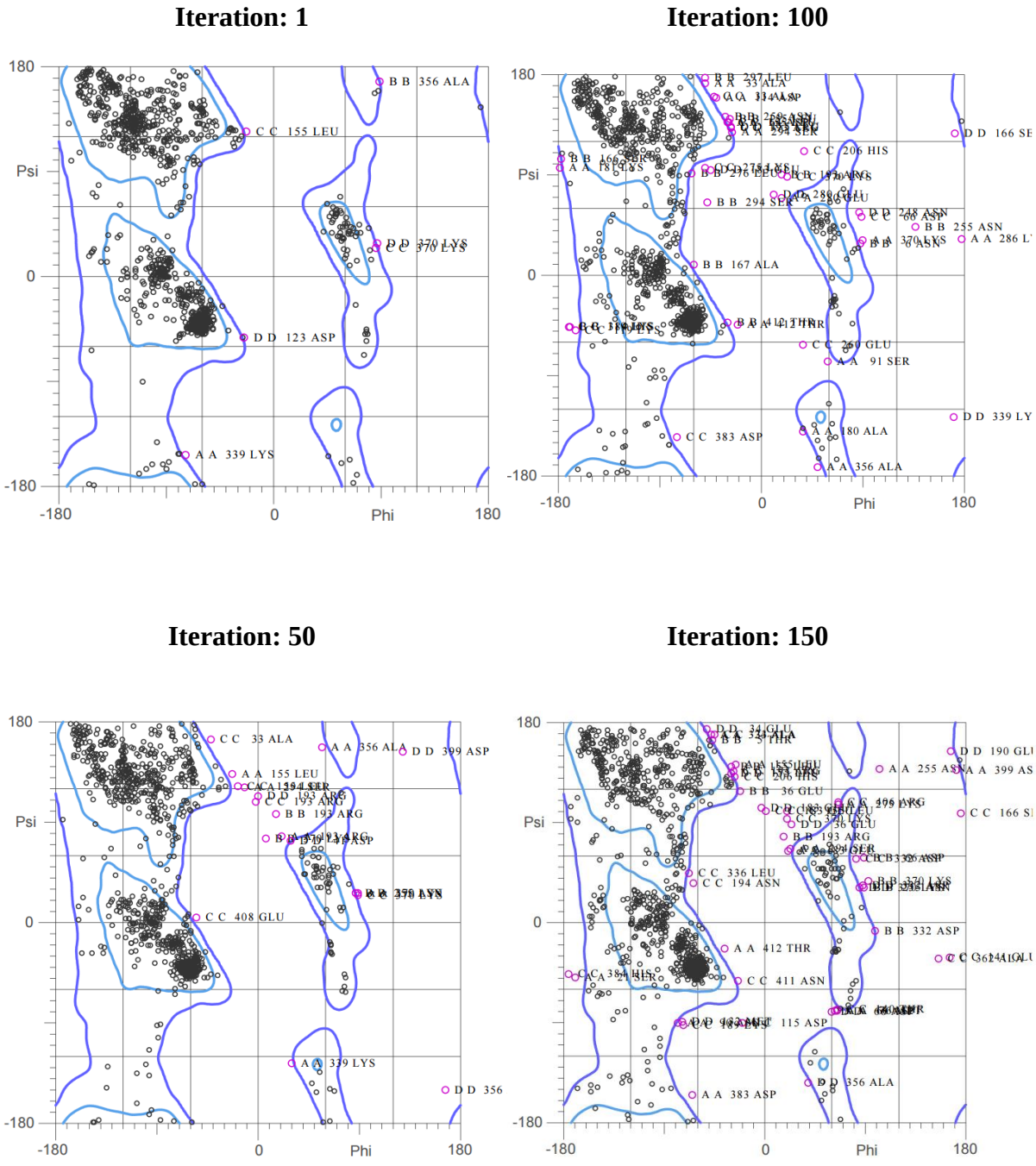
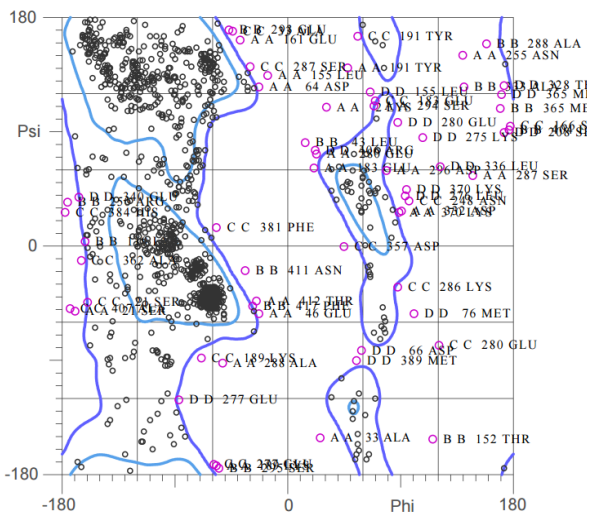
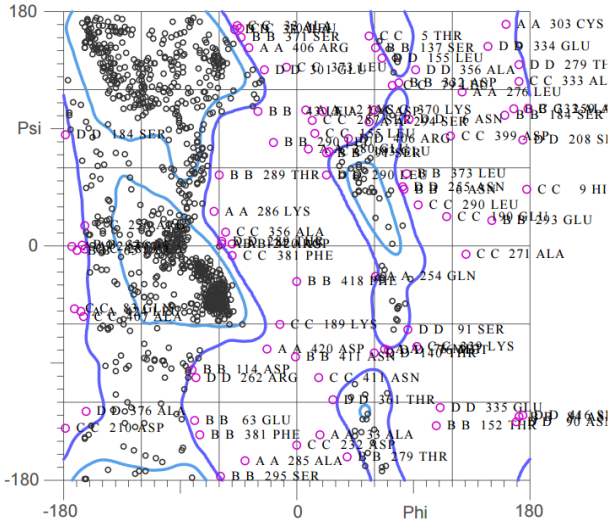


Figure 6.8. Ramachandran plots of relaxed EPSPS conformations at different iterations.

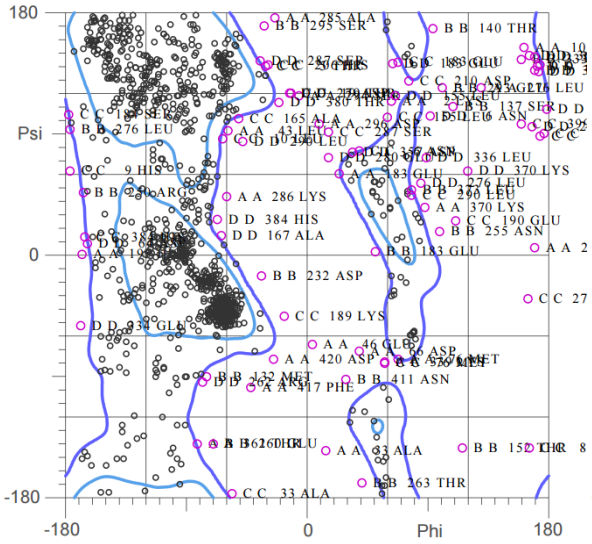
Iteration: 200



Iteration: 300



Iteration: 250



Iteration: 327

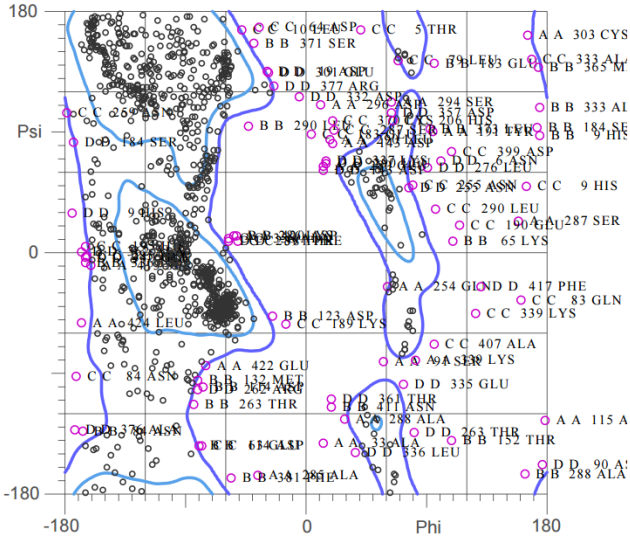


Figure 6.8. Ramachandran plots of relaxed EPSPS conformations at different iterations (cont.)

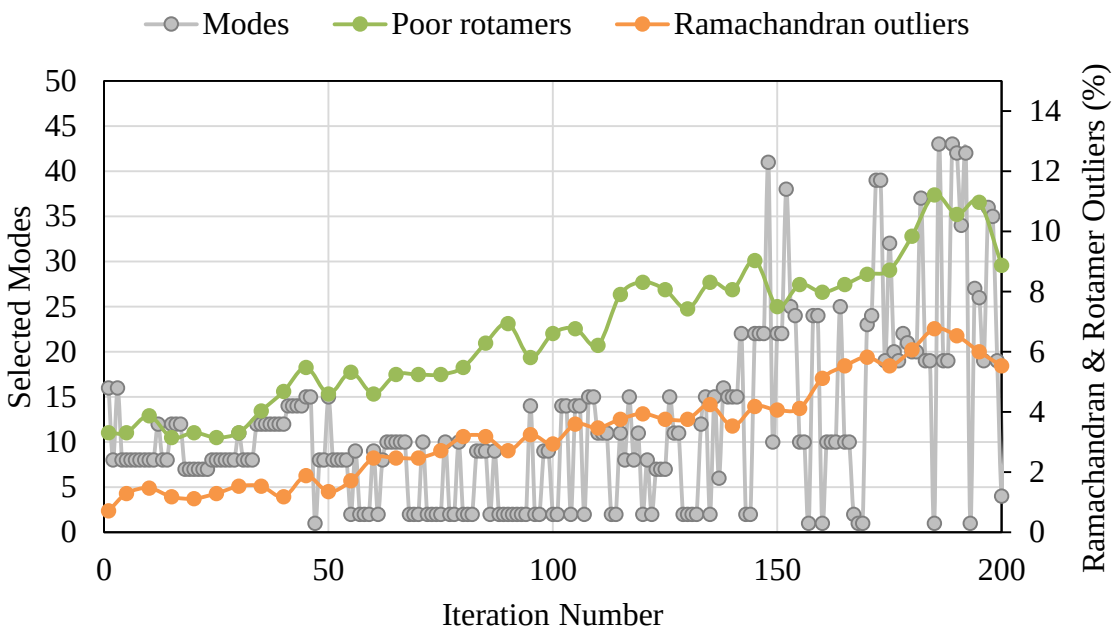


Figure 6.9. Correlation of selected modes of EPSPS with the amount of Ramachandran outliers and poor rotamers observed in optimized atomistic conformations.

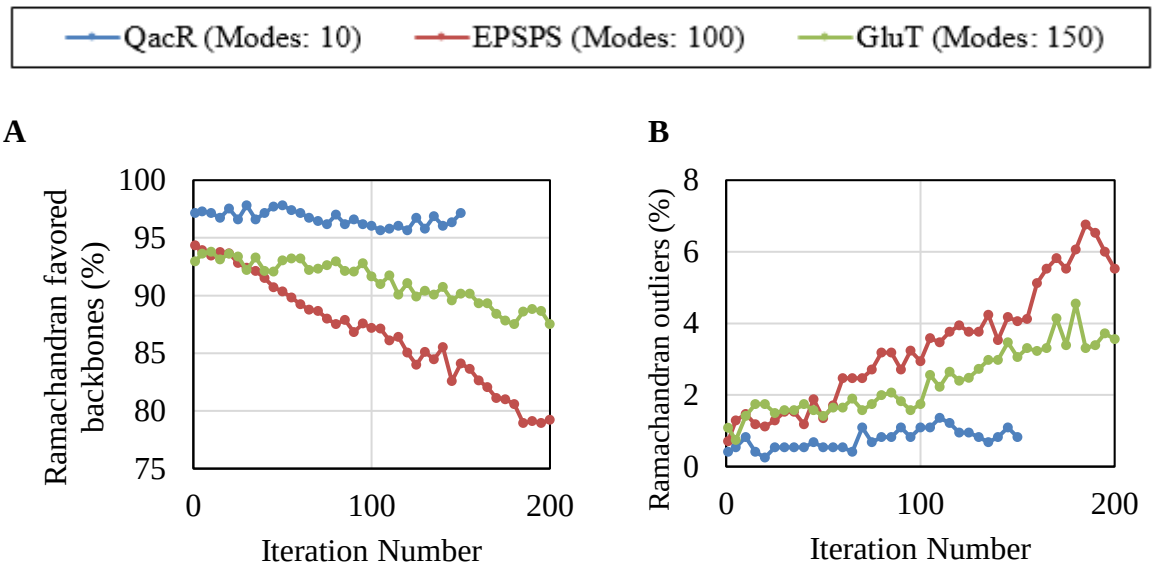


Figure 6.10. Percentages of unusual conformations of backbones and sidechains in the ANM-MC trajectories of QacR (10), EPSPS (100) and GluT (150) within the modes in parenthesis.

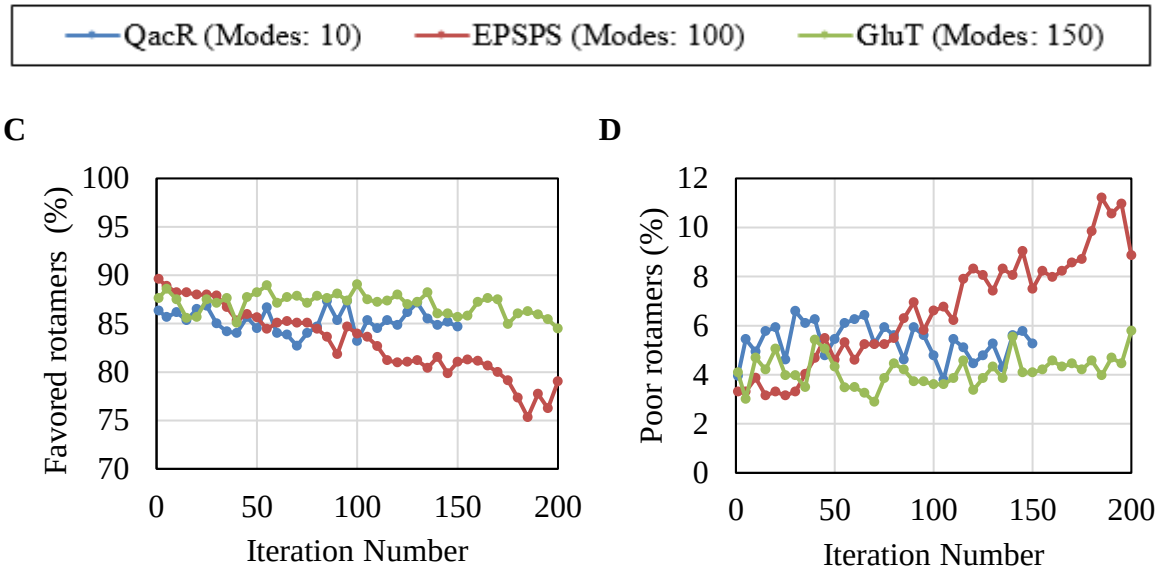


Figure 6.10. Percentages of unusual conformations of backbones and sidechains in the ANM-MC trajectories of QacR (10), EPSPS (100) and GluT (150) within the modes in parenthesis (cont.)

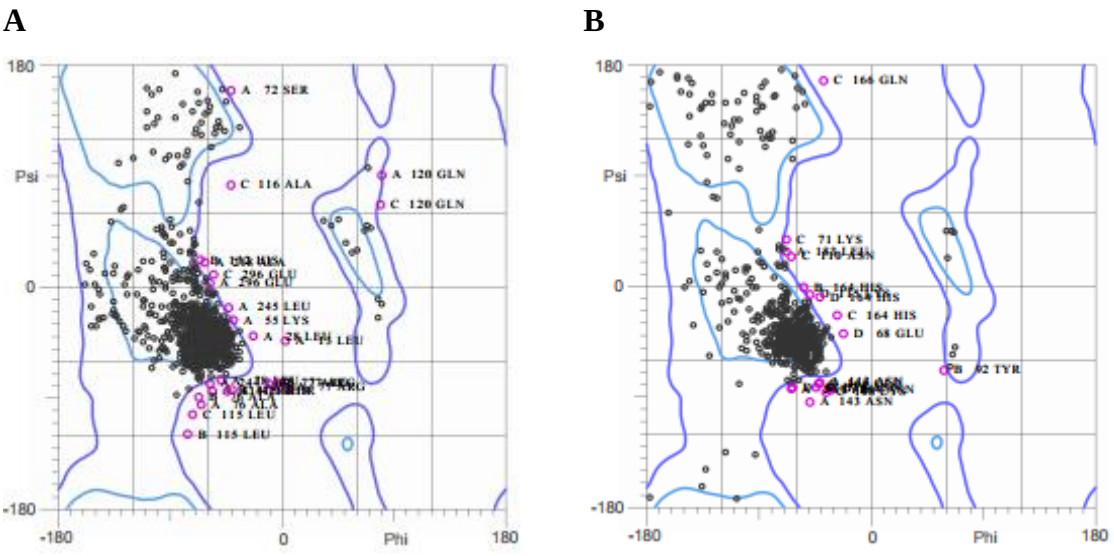


Figure 6.11. Ramachandran plots of A) GluT (PDB ID: 1XFH) and B) QacR (PDB ID: 1JT0).

7. CONCLUSIONS AND RECOMENDATIONS

7.1. Conclusions

Conformational transition pathways of large proteins are efficiently explored by ANM-MC for different proteins such as hinge-bending, small ligand-binding and DNA-binding proteins, shear type proteins, enzymes and receptors. ANM-MC parameters are adjusted by a series of benchmark analyses for the later application to large proteins. The adjusted ANM-MC parameters are applied to the large systems. Some arrangements in ANM-MC methodology besides parameter adjustment are adopted for addressing the structural problems/concerns arising in large systems. A detailed analysis of conformational transitions is provided. Conversion of ANM-MC snapshots to full atomistic conformations is suggested by reverse mapping techniques.

7.1.1. Adjustment of ANM-MC parameters

ANM-MC parameters (i.e. ANM deformation factor, force constant for virtual bonds, perturbation strength, and number of steps in MC) are benchmarked using two proteins, namely AK and calmodulin. Then geometric parameters (i.e. bond lengths, bond angles, dihedral angles) throughout the trajectory are monitored and adjusted with respect to crystal structures of AK and NMR models of calmodulin. The adjustment of force constant in MC minimization method (k = 500 J/Å/mol) yields geometrically plausible conformational structures. MC step numbers (MCs = 15) along with the standardization of characteristic perturbation strength (MC PS = 0.15 Å) are optimized for large proteins. For the sake of computational efficiency, structures are suggested to be deformed via ANM larger (DF = 0.3) to reach target faster. As a result, crystal AK structures are observed lying very close to the computationally obtained pathway intermediates. NMR data of calmodulin also validates the success of the ANM-MC technique.

7.1.2. Application of ANM-MC to large systems

The adjusted ANM-MC parameters are applied to the dataset consisting of 13 large proteins and enzymes of which total number of residues roughly ranges from 600 to 8000. The initial distance (RMSD) between end structures span a broad range of 2.1 – 26.0 Å.

Different cutoff values between 8 and 18 Å have been used in previous elastic network studies, the choice of 10 Å has been found most successful in terms of getting closer to the target state for small and medium sized proteins [19, 21, 28, 34, 128]. However, a higher cutoff value is found necessary here for application of ANM to large proteins. The relatively higher cutoff of 15 Å is especially suitable for protein complexes with cavities and holes.

Most of the proteins in the dataset exhibit successive conformational transitions along the slowest 10 modes that primarily direct the protein from the open to the closed state. The third mode dominates in the runs of aspartate transcarbamylase, citrate synthase and hemoglobin. For other proteins necessitating multiple slow modes for transitions, there is a more gradual approach to closed structure.

For some proteins, higher indexed modes are involved in their pathways, which might be attributed to the compactness of large proteins having a collection of various conformational states [21]. Those proteins are chaperonin GroEL, chaperonin lidless Mm-cpn, glutamate transporter, myosin, RNA Polymerase II and 5-enolpyruvylshikimate-3-phosphate synthase. As in nature of elastic network models, slower modes play a dominant role in the initial stage of deformations, while increased involvement of higher indexed modes gradually complement as protein proceed away from the initial state.

Computational efficiency of the ANM-MC program is discussed in the context of precision and computational time for large proteins. Wall times of ANM-MC runs are reported for each protein. Process time of ANM rises exponentially with respect to the number of residues above 2000. Clock times of the runs depend on the minimum number of modes used (10 - 150). The performance of the program over its computational cost is reported for each protein in the dataset.

A key component to the success of ANM-MC method is the ability of initial structure to approach the target, which is expressed via RMSD. Final RMSDs are in the range of 1.5 - 4.3 Å with the aim of having a final RMSD within 3.0 Å or proceeding at least the half way of the pathway. All the runs start with global modes, which is reflected by a high overlap value, and then sometimes require more local modes accompanied by relatively lower overlap.

Higher indexed modes permit relatively local movements in the intrinsic motions and enable the conformations move toward the target direction. However, such modes may cause excessive compressions and/or expansions in virtual bond lengths. For such cases, *DF* of 0.3 is found too large, consequently, a slight modification in ANM-MC methodology is introduced. The structure is deformed along the *DF*s of 0.3, 0.2, and 0.1 with an upper limit of 0.08 in standard deviation of bond lengths with respect the initial structure. For the cases where bond length variance exceeds the limit of 0.08, higher number of MC steps is utilized up to 50. With the modified ANM-MC methodology, all the initial structures follow plausible pathways and successfully approach the target structures. The energy profiles do not exhibit sudden peaks for any of the proteins in the dataset, which is accompanied by plausible geometric parameters (bond length, bond angle, dihedral angle) throughout the trajectories.

The lowest ten frequency modes play a significant role in successful ANM-MC runs in terms of reflecting intrinsic dynamics property of the proteins by itself. In the runs completed modes within the range of 1 to 10, higher amount of deformation applied by ANM is allowed (*DF* = 0.3) along the trajectories. Virtual bond lengths in every single snapshot do not show a deviation from initial crystal structure, even without help of extra MC steps.

The lowest ten frequency modes play a significant role in successful ANM-MC runs in terms of reflecting intrinsic dynamics property of the proteins by itself. In the runs completed only first 10 modes, higher amount of deformation can be applied in ANM (*DF* = 0.3) along the trajectories. Virtual bond lengths in every single snapshot do not show a deviation from initial crystal structure, even without help of extra MC steps. On the other hand, in the other set of runs incorporating higher indexed modes (50 - 150), lower *DF* and higher *MCs* need to be occasionally applied in ANM (*DF* = 0.2 and 0.1) so that virtual bond lengths do

not deviate too much in the usual ANM-MC approach. One exception is chaperonin lidless Mm-cpn that can tolerate high *DF* of 0.3 in spite of the higher modes chosen throughout its trajectory.

7.1.3. Conversion ANM-MC virtual structures to full atomistic conformations

Reverse-mapping technique is performed on the coarse-grained ANM-MC snapshots successively and then, energy minimization is employed to give relaxation to the obtained full atomistic structures. As the coarse-grained structures of ANM-MC runs exhibit plausible geometric features, the question addressed here is to what extent the reverse-mapped geometries are plausible. This procedure is first tested on a small protein, AK (only first 10 modes) using implicit and explicit solvation. The optimized atomistic conformations of AK show the desired backbone geometry with both solvation models. Because the explicit solvation model provides computational advantage for AK, it is further applied to several proteins in the dataset. These proteins are EPSPS, GluT and QacR that present the largest conformational transitions in the dataset.

QacR (only first 10 modes) shows a similar geometry profile with AK, whereas GluT (150 modes) reflects undesirable effects of higher modes on its backbone geometry. Similarly, EPSPS (100 modes) gives a considerable amount of Ramachandran outliers and poor rotamers, which are correlated with the selected modes in the second part of its pathway.

7.2. Recommendations

The algorithm may be extended for application to systems containing DNA or RNA chains. For this purpose, coarse-graining of such chains and their accompanying knowledge-based potentials need to be developed first. If successful, ANM-MC can be further applied to protein- DNA/RNA complexes or the supramolecule ribosome containing RNA chains.

Collective modes can be randomly chosen at certain stages of the simulations instead of directing them towards the target structure in all iterations. Furthermore, a combination of different indexed modes can be used for deformation in a single iteration.

Reverse-mapping technique can be applied using some constraints during energy minimization step, such as fixing the backbone atoms initially, to minimize undesirable local geometries. Further short MD simulations can be performed on reverse-mapped snapshots to observe sampled conformations.

Lastly, versions of ANM-MC program can be offered to the research community as a user-friendly web server mainly to predict and visualize the collective motions of large complexes.

REFERENCES

1. Okazaki, K., N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes, "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations", *Proceedings of the National Academy of Sciences of the United States of America,* 103.32: 11844-9, (2006).

2. Frauenfelder, H., S. G. Sligar, and P. G. Wolynes, "The energy landscapes and motions of proteins", *Science,* 254.5038: 1598-603, (1991).

3. Lei, M., J. Velos, A. Gardino, A. Kivenson, M. Karplus, and D. Kern, "Segmented transition pathway of the signaling protein nitrogen regulatory protein C", *Journal of Molecular Biology,* 392.3: 823-36, (2009).

4. Henzler-Wildman, K. A., V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hubner, and D. Kern, "Intrinsic motions along an enzymatic reaction trajectory", *Nature,* 450.7171: 838-44, (2007).

5. Berman, H., K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank", *Nature Structural & Molecular Biology,* 10.12: 980, (2003).

6. Pruitt, K. D., T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins", *Nucleic Acids Research,* 35.Database issue: D61-5, (2007).

7. Das, A., M. Gur, M. H. Cheng, S. Jo, and I. Bahar, "Exploring the Conformational Transitions of Biomolecular Systems Using a Simple Two-State Anisotropic Network Model", *Plos Computational Biology,* 10.8, (2014).

8. Pan, A. C., T. M. Weinreich, S. Piana, and D. E. Shaw, "Using Long-Timescale Molecular Dynamics Simulations to Benchmark Enhanced Sampling Methods", *Biophysical Journal* 108.2: 183a-183a, (2015).

9. Elber, R. and M. Karplus, "A Method for Determining Reaction Paths in Large Molecules - Application to Myoglobin", *Chemical Physics Letters,* 139.5: 375-380, (1987).

10. Maragliano, L., A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, "String method in collective variables: Minimum free energy paths and isocommittor surfaces", *Journal of Chemical Physics,* 125.2, (2006).

11. Pan, A. C., D. Sezer, and B. Roux, "Finding transition pathways using the string method with swarms of trajectories", *Journal of Physical Chemistry B,* 112.11: 3432-3440, (2008).

12. Bolhuis, P. G., D. Chandler, C. Dellago, and P. L. Geissler, "Transition path sampling: Throwing ropes over rough mountain passes, in the dark", *Annual Review of Physical Chemistry,* 53: 291-318, (2002).

13. Chandler, D., "Statistical-Mechanics of Isomerization Dynamics in Liquids and Transition-State Approximation", *Journal of Chemical Physics,* 68.6: 2959-2970, (1978).

14. Dellago, C., P. G. Bolhuis, F. S. Csajka, and D. Chandler, "Transition path sampling and the calculation of rate constants", *Journal of Chemical Physics,* 108.5: 1964-1977, (1998).

15. Du, W. N., K. A. Marino, and P. G. Bolhuis, "Multiple state transition interface sampling of alanine dipeptide in explicit solvent", *Journal of Chemical Physics,* 135.14, (2011).

16. Grubmuller, H., "Predicting Slow Structural Transitions in Macromolecular Systems - Conformational Flooding", *Physical Review E,* 52.3: 2893-2906, (1995).

17. Hamelberg, D., J. Mongan, and J. A. McCammon, "Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules", *Journal of Chemical Physics,* 120.24: 11919-11929, (2004).

18. Laio, A. and M. Parrinello, "Escaping free-energy minima", *Proceedings of the National Academy of Sciences of the United States of America,* 99.20: 12562-12566, (2002).

19. Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model", *Biophysical Journal,* 80.1: 505-515, (2001).

20. Doruker, P., A. R. Atilgan, and I. Bahar, "Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to alpha-amylase inhibitor", *Proteins-Structure Function and Genetics,* 40.3: 512-524, (2000).

21. Tama, F. and Y. H. Sanejouand, "Conformational change of proteins arising from normal mode calculations", *Protein Engineering,* 14.1: 1-6, (2001).

22. Doruker, P., R. L. Jernigan, and I. Bahar, "Dynamics of large proteins through hierarchical levels of coarse-grained structures", *Journal of Computational Chemistry,* 23.1: 119-127, (2002).

23. Doruker, P., L. Nilsson, and O. Kurkcuoglu, "Collective dynamics of EcoRI-DNA complex by elastic network model and molecular dynamics simulations", *Journal of Biomolecular Structure & Dynamics,* 24.1: 1-15, (2006).

24. Keskin, O., I. Bahar, D. Flatow, D. G. Covell, and R. L. Jernigan, "Molecular mechanisms of chaperonin GroEL-GroES function", *Biochemistry,* 41.2: 491-501, (2002).

25. Kurkcuoglu, O., R. L. Jernigan, and P. Doruker, "Collective dynamics of large proteins from mixed coarse-grained elastic network model", *Qsar & Combinatorial Science,* 24.4: 443-448, (2005).

26. Kurkcuoglu, O., R. L. Jernigan, and P. Doruker, "Loop motions of triosephosphate isomerase observed with elastic networks", *Biochemistry,* 45.4: 1173-1182, (2006).

27. Wang, Y. M., A. J. Rader, I. Bahar, and R. L. Jernigan, "Global ribosome motions revealed with elastic network model", *Journal of Structural Biology,* 147.3: 302-314, (2004).

28. Yang, Z., P. Majek, and I. Bahar, "Allosteric Transitions of Supramolecular Systems Explored by Network Models: Application to Chaperonin GroEL", *Plos Computational Biology,* 5.4, (2009).

29. Yidirim, Y. and P. Doruker, "Collective motions of RNA polymerases. Analysis of core enzyme, elongation complex and holoenzyme", *Journal of Biomolecular Structure & Dynamics,* 22.3: 267-280, (2004).

30. Kantarci-Carsibasi, N., T. Haliloglu, and P. Doruker, "Conformational Transition Pathways Explored by Monte Carlo Simulation Integrated with Collective Modes", *Biophysical Journal,* 95.12: 5862-5873, (2008).

31. Haliloglu, T., "Coarse-grained simulations of the conformational dynamics of proteins", *Computational and Theoretical Polymer Science,* 9.3-4: 255-260, (1999).

32. Haliloglu, T. and I. Bahar, "Coarse-grained simulations of conformational dynamics of proteins: Application to apomyoglobin", *Proteins-Structure Function and Genetics,* 31.3: 271-281, (1998).

33. Kurt, N. and T. Haliloglu, "Conformational dynamics of chymotrypsin inhibitor 2 by coarse-grained simulations", *Proteins-Structure Function and Genetics,* 37.3: 454-464, (1999).

34. Uyar, A., N. Kantarci-Carsibasi, T. Haliloglu, and P. Doruker, "Features of Large Hinge-Bending Conformational Transitions. Prediction of Closed Structure from Open State", *Biophysical Journal,* 106.12: 2656-2666, (2014).

35. Liu, H. L. and J. P. Hsu, "Recent developments in structural proteomics for protein structure determination", *Proteomics,* 5.8: 2056-2068, (2005).

36. Weinan, E., W. Q. Ren, and E. Vanden-Eijnden, "String method for the study of rare events", *Physical Review B,* 66.5, (2002).

37. Standley, D. M., A. R. Kinjo, K. Kinoshita, and H. Nakamura, "Protein structure databases with new web services for structural biology and biomedical research", *Briefings in Bioinformatics,* 9.4: 276-285, (2008).

38. Mulder, F. A. A. and M. Filatov, "NMR chemical shift data and ab initio shielding calculations: emerging tools for protein structure determination", *Chemical Society Reviews,* 39.2: 578-590, (2010).

39. Callaway, E., "The Revolution Will Not Be Crystallized", *Nature,* 525(7568): 172-174, (2015).

40. Khatter, H., A. G. Myasnikov, S. K. Natchiar, and B. P. Klaholz, "Structure of the human 80S ribosome", *Nature,* 520.7549: 640-U338, (2015).

41. Amunts, A., A. Brown, J. Toots, S. H. W. Scheres, and V. Ramakrishnan, "The structure of the human mitochondrial ribosome", *Science,* 348.6230: 95-98, (2015).

42. Kikhney, A. G. and D. I. Svergun, "A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins", *Febs Letters,* 589.19: 2570-2577, (2015).

43. Floudas, C. A., "Computational methods in protein structure prediction", *Biotechnology and Bioengineering,* 97.2: 207-13, (2007).

44. Elber, R., "Long-timescale simulation methods", *Current Opinion in Structural Biology,* 15.2: 151-156, (2005).

45. Fischer, S. and M. Karplus, "Conjugate Peak Refinement - an Algorithm for Finding Reaction Paths and Accurate Transition-States in Systems with Many Degrees of Freedom", *Chemical Physics Letters,* 194.3: 252-261, (1992).

46. Branduardi, D., F. L. Gervasio, and M. Parrinello, "From A to B in free energy space", *Journal of Chemical Physics,* 126.5, (2007).

47. Isralewitz, B., M. Gao, and K. Schulten, "Steered molecular dynamics and mechanical functions of proteins", *Current Opinion in Structural Biology,* 11.2: 224-230, (2001).

48. Schlitter, J., M. Engels, P. Kruger, E. Jacoby, and A. Wollmer, "Targeted Molecular-Dynamics Simulation of Conformational Change - Application to the T->R Transition in Insulin", *Molecular Simulation,* 10.2-6: 291-&, (1993).

49. van der Vaart, A. and M. Karplus, "Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method", *Journal of Chemical Physics,* 122.11, (2005).

50. van der Vaart, A. and M. Karplus, "Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations", *Journal of Chemical Physics,* 126.16, (2007).

51. Yang, H. J., H. Wu, D. W. Li, L. Han, and S. H. Huo, "Temperature-dependent probabilistic roadmap algorithm for calculating variationally optimized conformational transition pathways", *Journal of Chemical Theory and Computation,* 3.1: 17-25, (2007).

52. Gan, W. X., S. C. Yang, and B. Roux, "Atomistic View of the Conformational Activation of Src Kinase Using the String Method with Swarms-of-Trajectories", *Biophysical Journal,* 97.4: L8-L10, (2009).

53. Jo, S., H. A. Rui, J. B. Lim, J. B. Klauda, and W. Im, "Cholesterol Flip-Flop: Insights from Free Energy Simulation Studies", *Journal of Physical Chemistry B,* 114.42: 13342-13348, (2010).

54. Kirmizialtin, S., V. Nguyen, K. A. Johnson, and R. Elber, "How Conformational Dynamics of DNA Polymerase Select Correct Substrates: Experiments and Simulations", *Structure,* 20.4: 618-627, (2012).

55. Matsunaga, Y., H. Fujisaki, T. Terada, T. Furuta, K. Moritsugu, and A. Kidera, "Minimum Free Energy Path of Ligand-Induced Transition in Adenylate Kinase", *Plos Computational Biology* 8.6, (2012).

56. Ovchinnikov, V., M. Karplus, and E. Vanden-Eijnden, "Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI", *Journal of Chemical Physics,* 134.8, (2011).

57. Stober, S. T. and C. F. Abrams, "Energetics and Mechanism of the Normal-to-Amyloidogenic Isomerization of beta 2-Microglobulin: On-the-Fly String Method Calculations", *Journal of Physical Chemistry B,* 116.31: 9371-9375, (2012).

58. Vashisth, H. and C. F. Abrams, "All-atom structural models of insulin binding to the insulin receptor in the presence of a tandem hormone-binding element", *Proteins-Structure Function and Bioinformatics,* 81.6: 1017-1030, (2013).

59. Vashisth, H., L. Maragliano, and C. F. Abrams, ""DFG-Flip" in the Insulin Receptor Kinase Is Facilitated by a Helical Intermediate State of the Activation Loop", *Biophysical Journal,* 102.8: 1979-1987, (2012).

60. Elber, R. and M. Karplus, "Multiple Conformational States of Proteins - a Molecular-Dynamics Analysis of Myoglobin", *Science,* 235.4786: 318-321, (1987).

61. Go, N. and T. Noguti, "Structural Basis of Hierarchical Multiple Substates of a Protein", *Chemica Scripta,* 29A: 151-164, (1989).

62. Li, Z. Q. and H. A. Scheraga, "Monte-Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding", *Proceedings of the National Academy of Sciences of the United States of America,* 84.19: 6611-6615, (1987).

63. Frauenfelder, H., F. Parak, and R. D. Young, "Conformational Substates in Proteins", *Annual Review of Biophysics and Biophysical Chemistry*, 17: 451-479, (1988).

64. Go, N., T. Noguti, and T. Nishikawa, "Dynamics of a Small Globular Protein in Terms of Low-Frequency Vibrational-Modes", *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, 80.12: 3696-3700, (1983).

65. Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines", *Journal of Chemical Physics*, 21.6: 1087-1092, (1953).

66. Hinsen, K. and G. R. Kneller, "A simplified force field for describing vibrational protein dynamics over the whole frequency range", *Journal of Chemical Physics*, 111.24: 10766-10769, (1999).

67. Hinsen, K., A. Thomas, and M. J. Field, "Analysis of domain motions in large proteins", *Proteins-Structure Function and Genetics*, 34.3: 369-382, (1999).

68. Abseher, R. and M. Nilges, "Efficient sampling in collective coordinate space", *Proteins-Structure Function and Genetics*, 39.1: 82-88, (2000).

69. Bahar, I., B. Erman, R. L. Jernigan, A. R. Atilgan, and D. G. Covell, "Collective motions in HIV-1 reverse transcriptase: Examination of flexibility and enzyme function", *Journal of Molecular Biology*, 285.3: 1023-1037, (1999).

70. Bahar, I. and R. L. Jernigan, "Cooperative fluctuations and subunit communication in tryptophan synthase", *Biochemistry*, 38.12: 3478-3490, (1999).

71. Wassenaar, T. A., K. Pluhackova, R. A. Bockmann, S. J. Marrink, and D. P. Tieleman, "Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models", *Journal of Chemical Theory and Computation*, 10.2: 676-90, (2014).

72. Liwo, A., M. Baranowski, C. Czaplewski, E. Golas, Y. He, D. Jagiela, P. Krupa, M. Maciejczyk, M. Makowski, M. A. Mozolewska, A. Niadzvedtski, S. Oldziej, H. A. Scheraga, A. K. Sieradzan, R. Slusarz, T. Wirecki, Y. P. Yin, and B. Zaborowski, "A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions", *Journal of Molecular Modeling*, 20.8, (2014).

73. Kolinski, A., "Protein modeling and structure prediction with a reduced representation", *Acta Biochimica Polonica*, 51.2: 349-371, (2004).

74. Zhang, Y. and J. Skolnick, "Automated structure prediction of weakly homologous proteins on a genomic scale"', *Proceedings of the National Academy of Sciences of the United States of America*, 101.20: 7594-7599, (2004).

75. Yang, J. Y., R. X. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The I-TASSER Suite: protein structure and function prediction", *Nature, Methods* 12.1: 7-8, (2015).

76. Doruker, P. and W. L. Mattice, "Reverse mapping of coarse-grained polyethylene chains from the second nearest neighbor diamond lattice to an atomistic model in continuous space", *Macromolecules,* 30.18: 5520-5526, (1997).

77. Rotkiewicz, P. and J. Skolnick, "Fast procedure for reconstruction of full-atom protein models from reduced representations", *Journal of Computational Chemistry,* 29.9: 1460-5, (2008).

78. Petrey, D., Z. X. Xiang, C. L. Tang, L. Xie, M. Gimpelev, T. Mitros, C. S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I. Y. Y. Koh, E. Alexov, and B. Honig, "Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling", *Proteins-Structure Function and Genetics,* 53.6: 430-435, (2003).

79. Holm, L. and C. Sander, "Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors", *Journal of Molecular Biology,* 218.1: 183-94, (1991).

80. Yang, Z., K. Lasker, D. Schneidman-Duhovny, B. Webb, C. C. Huang, E. F. Pettersen, T. D. Goddard, E. C. Meng, A. Sali, and T. E. Ferrin, "UCSF Chimera, MODELLER, and IMP: An integrated modeling system", *Journal of Structural Biology,* 179.3: 269-278, (2012).

81. Li, Y. and Y. Zhang, "REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks", *Proteins* 76.3: 665-76, (2009).

82. Skjaerven, L., S. M. Hollup, and N. Reuter, "Normal mode analysis for proteins", *Journal of Molecular Structure-Theochem,* 898.1-3: 42-48, (2009).

83. Hinsen, K., "Analysis of domain motions by approximate normal mode calculations", *Proteins-Structure Function and Genetics,* 33.3: 417-429, (1998).

84. Kurkcuoglu, O., R. L. Jernigan, and P. Doruker, "Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions", *Polymer,* 45.2: 649-657, (2004).

85. Bahar, I., A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential", *Folding & Design,* 2.3: 173-81, (1997).

86. Brooks, B. and M. Karplus, "Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor", *Proceedings of the National Academy of Sciences of the United States of America,* 80.21: 6571-5, (1983).

87. Li, G. and Q. Cui, "A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca(2+)-ATPase", *Biophysical Journal,* 83.5: 2457-74, (2002).

88. Marques, O. and Y. H. Sanejouand, "Hinge-bending motion in citrate synthase arising from normal mode calculations", *Proteins,* 23.4: 557-60, (1995).

89. Mouawad, L. and D. Perahia, "Diagonalization in a Mixed Basis - a Method to Compute Low-Frequency Normal-Modes for Large Macromolecules", *Biopolymers,* 33.4: 599-611, (1993).

90. Tama, F., F. X. Gadea, O. Marques, and Y. H. Sanejouand, "Building-block approach for determining low-frequency normal modes of macromolecules", *Proteins-Structure Function and Genetics,* 41.1: 1-7, (2000).

91. Tirion, M. M., "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis", *Physical Review Letters,* 77.9: 1905-1908, (1996).

92. Adamovic, I., S. M. Mijailovich, and M. Karplus, "The elastic properties of the structurally characterized myosin II s2 subdomain: A molecular dynamics and normal mode analysis", *Biophysical Journal,* 94.10: 3779-3789, (2008).

93. Brooks, B. and M. Karplus, "Normal-Modes for Specific Motions of Macromolecules - Application to the Hinge-Bending Mode of Lysozyme", *Proceedings of the National Academy of Sciences of the United States of America,* 82.15: 4995-4999, (1985).

94. Cui, Q., G. H. Li, J. P. Ma, and M. Karplus, "A normal mode analysis of structural plasticity in the biomolecular motor F-1-ATPase", *Journal of Molecular Biology,* 340.2: 345-372, (2004).

95. Gaillard, T., E. Martin, E. S. Sebastian, F. P. Cossio, X. Lopez, A. Dejaegere, and R. H. Stote, "Comparative normal mode analysis of LFA-1 integrin 1-domains", *Journal of Molecular Biology,* 374.1: 231-249, (2007).

96. Gibrat, J. F. and N. Go, "Normal Mode Analysis of Human Lysozyme - Study of the Relative Motion of the 2 Domains and Characterization of the Harmonic Motion", *Journal De Chimie Physique Et De Physico-Chimie Biologique,* 88.11-12: 2581-2585, (1991).

97. Hayward, S., A. Kitao, and H. J. C. Berendsen, "Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme", *Proteins-Structure Function and Genetics,* 27.3: 425-437, (1997).

98. Kim, M. K., R. L. Jernigan, and G. S. Chirikjian, "An elastic network model of HK97 capsid maturation", *Journal of Structural Biology,* 143.2: 107-117, (2003).

99. Levitt, M., C. Sander, and P. S. Stern, "Protein Normal-Mode Dynamics - Trypsin-Inhibitor, Crambin, Ribonuclease and Lysozyme", *Journal of Molecular Biology,* 181.3: 423-447, (1985).

100. Reuter, N., K. Hinsen, and J. J. Lacapere, "Transconformations of the SERCA1 Ca-ATPase: A normal mode study", *Biophysical Journal*, 85.4: 2186-2197, (2003).

101. Tama, F. and C. L. Brooks, "Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis", *Journal of Molecular Biology*, 345.2: 299-314, (2005).

102. Thomas, A., K. Hinsen, M. J. Field, and D. Perahia, "Tertiary and quaternary conformational changes in aspartate transcarbamylase: A normal mode study", *Proteins-Structure Function and Genetics*, 34.1: 96-112, (1999).

103. Zhang, B. W., D. Jasnow, and D. M. Zuckerman, "The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures", *Journal of Chemical Physics*, 132.5, (2010).

104. Zheng, W. J. and S. Doniach, "A comparative study of motor-protein motions by using a simple elastic-network model", *Proceedings of the National Academy of Sciences of the United States of America*, 100.23: 13253-13258, (2003).

105. Zoete, V., O. Michielin, and M. Karplus, "Relation between sequence and structure of HIV-1 protease inhibitor complexes: A model system for the analysis of protein flexibility", *Journal of Molecular Biology*, 315.1: 21-52, (2002).

106. Flory, P. J., "Statistical Thermodynamics of Random Networks", *Proceedings of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences*, 351.1666: 351-380, (1976).

107. Go, N. and H. A. Scheraga, "Analysis of Contribution of Internal Vibrations to Statistical Weights of Equilibrium Conformations of Macromolecules", *Journal of Chemical Physics*, 51.11: 4751-&, (1969).

108. Go, N. and H. A. Scheraga, "Use of Classical Statistical-Mechanics in Treatment of Polymer-Chain Conformation", *Macromolecules*, 9.4: 535-542, (1976).

109. Haliloglu, T., I. Bahar, and B. Erman, "Gaussian dynamics of folded proteins", *Physical Review Letters*, 79.16: 3090-3093, (1997).

110. Bahar, I., T. R. Lezon, A. Bakan, and I. H. Shrivastava, "Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins", *Chemical Reviews*, 110.3: 1463-97, (2010).

111. Bahar, I., T. R. Lezon, L. W. Yang, and E. Eyal, "Global Dynamics of Proteins: Bridging Between Structure and Function", *Annual Review of Biophysics, Vol 39* 39: 23-42, (2010).

112. Bahar, I. and A. J. Rader, "Coarse-grained normal mode analysis in structural biology", *Current Opinion in Structural Biology*, 15.5: 586-592, (2005).

113. Gur, M., E. Zomot, and I. Bahar, "Global motions exhibited by proteins in micro- to milliseconds simulations concur with anisotropic network model predictions", *Journal of Chemical Physics*, 139.12, (2013).

114. Kim, M. K., R. L. Jernigan, and G. S. Chirikjian, "Efficient generation of feasible pathways for protein conformational transitions", *Biophysical Journal*, 83.3: 1620-1630, (2002).

115. Kim, M. K., R. L. Jernigan, and G. S. Chirikjian, "Rigid-cluster models of conformational transitions in macromolecular machines and assemblies", *Biophysical Journal*, 89.1: 43-55, (2005).

116. Schuyler, A. D. and G. S. Chirikjian, "Efficient determination of low-frequency normal modes of large protein structures by cluster-NMA", *Journal of Molecular Graphics & Modelling*, 24.1: 46-58, (2005).

117. Schuyler, A. D., R. L. Jernigan, P. K. Qasba, B. Ramakrishnan, and G. S. Chirikjian, "Iterative cluster-NMA: A tool for generating conformational transitions in proteins", *Proteins-Structure Function and Bioinformatics*, 74.3: 760-776, (2009).

118. Maragakis, P. and M. Karplus, "Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase", *Journal of Molecular Biology*, 352.4: 807-822, (2005).

119. Best, R. B., Y. G. Chen, and G. Hummer, "Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of arc repressor", *Structure*, 13.12: 1755-1763, (2005).

120. Zheng, W. J., B. R. Brooks, and G. Hummer, "Protein conformational transitions explored by mixed elastic network models", *Proteins-Structure Function and Bioinformatics*, 69.1: 43-57, (2007).

121. Zhu, F. Q. and G. Hummer, "Gating Transition of Pentameric Ligand-Gated Ion Channels", *Biophysical Journal*, 97.9: 2456-2463, (2009).

122. Knowles, T. P., A. W. Fitzpatrick, S. Meehan, H. R. Mott, M. Vendruscolo, C. M. Dobson, and M. E. Welland, "Role of intermolecular forces in defining material properties of protein nanofibrils", *Science*, 318.5858: 1900-1903, (2007).

123. Li, H. C., S. Sakuraba, A. Chandrasekaran, and L. W. Yang, "Molecular Binding Sites Are Located Near the Interface of Intrinsic Dynamics Domains (IDDs)", *Journal of Chemical Information and Modeling*, 54.8: 2275-2285, (2014).

124. Reuveni, S., R. Granek, and J. Klafter, "Proteins: Coexistence of stability and flexibility", *Physical Review Letters*, 100.20, (2008).

125. Yang, L. W. and I. Bahar, "Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes", *Structure*, 13.6: 893-904, (2005).

126. Zimmermann, M. T., S. P. Leelananda, A. Kloczkowski, and R. L. Jernigan, "Combining Statistical Potentials with Dynamics-Based Entropies Improves Selection from Protein Decoys and Docking Poses", *Journal of Physical Chemistry B*, 116.23: 6725-6731, (2012).

127. Li, H. C., Y. Y. Chang, L. W. Yang, and I. Bahar, "iGNM 2.0: the Gaussian network model database for biomolecular structural dynamics", *Nucleic Acids Research*, 44.D1: D415-D422, (2016).

128. Eyal, E., L. W. Yang, and I. Bahar, "Anisotropic network model: systematic evaluation and a new web interface", *Bioinformatics*, 22.21: 2619-2627, (2006).

129. Eyal, E., G. Lum, and I. Bahar, "The anisotropic network model web server at 2015 (ANM 2.0)", *Bioinformatics*, 31.9: 1487-1489, (2015).

130. Carsibasi, Nigar Kantarci. "Conformational transitions of proteins explored by Monte Carlo simulations integrated with collective modes." *Chemical Engineering*: Bogazici University, 2009. 138. Vol. Ph.D. Ed. Pemra Doruker.

131. Uyar, A. "Conformational transitions of proteins using multi-scale modeling approaches." *Chemical Engineering*: Bogazici University, 2015. 151. Vol. Ph.D. Ed. Doruker.

132. Miyashita, O., J. N. Onuchic, and P. G. Wolynes, "Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins", *Proceedings of the National Academy of Sciences of the United States of America*, 100.22: 12570-12575, (2003).

133. Seeliger, D., J. Haas, and B. L. de Groot, "Geometry-based sampling of conformational transitions in proteins", *Structure*, 15.11: 1482-1492, (2007).

134. Bhatt, D. and I. Bahar, "An adaptive weighted ensemble procedure for efficient computation of free energies and first passage rates", *Journal of Chemical Physics*, 137.10, (2012).

135. Bhatt, D., B. W. Zhang, and D. M. Zuckerman, "Steady-state simulations using weighted ensemble path sampling", *Journal of Chemical Physics*, 133.1, (2010).

136. Huber, G. A. and S. Kim, "Weighted-ensemble Brownian dynamics simulations for protein association reactions", *Biophysical Journal*, 70.1: 97-110, (1996).

137. Perilla, J. R., O. Beckstein, E. J. Denning, and T. B. Woolf, "Computing Ensembles of Transitions from Stable States: Dynamic Importance Sampling", *Journal of Computational Chemistry*, 32.2: 196-209, (2011).

138. Woolf, T. B., "Path corrected functionals of stochastic trajectories: towards relative free energy and reaction coordinate calculations", *Chemical Physics Letters*, 289.5-6: 433-441, (1998).

139. Zuckerman, D. M. and T. B. Woolf, "Dynamic reaction paths and rates through importance-sampled stochastic dynamics", *Journal of Chemical Physics,* 111.21: 9475-9484, (1999).

140. Rojnuckarin, A., S. Kim, and S. Subramaniam, "Brownian dynamics simulations of protein folding: Access to milliseconds time scale and beyond", *Proceedings of the National Academy of Sciences of the United States of America,* 95.8: 4288-4292, (1998).

141. Allen, R. J., P. B. Warren, and P. R. ten Wolde, "Sampling rare switching events in biochemical networks", *Physical Review Letters,* 94.1, (2005).

142. Escobedo, F. A., E. E. Borrero, and J. C. Araque, "Transition path sampling and forward flux sampling. Applications to biological systems", *Journal of Physics-Condensed Matter,* 21.33, (2009).

143. Eyal, E., C. Chennubhotla, L. W. Yang, and I. Bahar, "Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models", *Bioinformatics,* 23.13: I175-I184, (2007).

144. *ANM 2.0 site documentation,* http://anm.csb.pitt.edu/anmdocs/Documentation.html#Cutoff

145. Bahar, I. and R. L. Jernigan, "Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms", *Journal of Molecular Biology,* 281.5: 871-84, (1998).

146. Bernstein, F. C., T. F. Koetzle, G. J. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank. A computer-based archival file for macromolecular structures", *European Journal of Biochemistry,* 80.2: 319-24, (1977).

147. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank", *Nucleic Acids Research,* 28.1: 235-42, (2000).

148. Bahar, I. and R. L. Jernigan, "Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation", *Journal of Molecular Biology,* 266.1: 195-214, (1997).

149. Bahar, I., M. Kaplan, and R. L. Jernigan, "Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches", *Proteins,* 29.3: 292-308, (1997).

150. Haliloglu, T., "Characterization of internal motions of Escherichia coli ribonuclease H by Monte Carlo simulation", *Proteins,* 34.4: 533-9, (1999).

151. Umeyama, S., "Least-Squares Estimation of Transformation Parameters between 2 Point Patterns", *Ieee Transactions on Pattern Analysis and Machine Intelligence,* 13.4: 376-380, (1991).

152. MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins", *The Journal of Physical Chemistry B,* 102.18: 3586-616, (1998).

153. Humphrey, W., A. Dalke, and K. Schulten, "VMD: visual molecular dynamics", *Journal of Molecular Graphics,* 14.1: 33-8, 27-8, (1996).

154. Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, "Scalable molecular dynamics with NAMD", *Journal of Computational Chemistry,* 26.16: 1781-802, (2005).

155. Bokovoi, V. A. and Y. V. Morozov, "Ionic and Tautomeric Equilibria of Benzimidazole and Its Pyridoxyl Derivatives - Analysis of Electronic-Spectra with an Oscillating Structure", *Zhurnal Fizicheskoi Khimii,* 62.9: 2372-2380, (1988).

156. *Eigsh Python Package*,2014 https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.sparse.linalg.eigsh.html#r182

157. Pieper, U., N. Eswar, H. Braberg, M. S. Madhusudhan, F. P. Davis, A. C. Stuart, N. Mirkovic, A. Rossi, M. A. Marti-Renom, A. Fiser, B. Webb, D. Greenblatt, C. C. Huang, T. E. Ferrin, and A. Sali, "MODBASE, a database of annotated comparative protein structure models, and associated resources", *Nucleic Acids Research,* 32: D217-D222, (2004).

158. DeLano, W. L., "Use of PYMOL as a communications tool for molecular science.", *Abstracts of Papers of the American Chemical Society,* 228: U313-U314, (2004).

159. DeLano, W. L., "PyMOL molecular viewer: Updates and refinements", *Abstracts of Papers of the American Chemical Society,* 238, (2009).

160. DeLano, W. L. and J. W. Lam, "PyMOL: A communications tool for computational models", *Abstracts of Papers of the American Chemical Society,* 230: U1371-U1372, (2005).

161. Rader, A. J. and S. M. Brown, "Correlating allostery with rigidity", *Molecular Biosystems,* 7.2: 464-471, (2011).

162. Daily, M. D. and J. J. Gray, "Local motions in a benchmark of allosteric proteins", *Proteins-Structure Function and Bioinformatics,* 67.2: 385-399, (2007).

163. Kantrowitz, E. R., "Allostery and cooperativity in Escherichia coli aspartate transcarbamoylase", *Archives of Biochemistry and Biophysics,* 519.2: 81-90, (2012).

164. Zhang, J. J., P. Minary, and M. Levitt, "Multiscale natural moves refine macromolecules using single-particle electron microscopy projection images", *Proceedings of the National Academy of Sciences of the United States of America*, 109.25: 9845-9850, (2012).

165. Baker, M. L., M. R. Baker, C. F. Hryc, T. Ju, and W. Chiu, "Gorgon and pathwalking: Macromolecular modeling tools for subnanometer resolution density maps", *Biopolymers*, 97.9: 655-668, (2012).

166. Delarue, M. and P. Dumas, "On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models", *Proceedings of the National Academy of Sciences of the United States of America*, 101.18: 6957-6962, (2004).

167. Park, H., J. L. Hilsenbeck, H. J. Kim, W. A. Shuttleworth, Y. H. Park, J. N. Evans, and C. H. Kang, "Structural studies of Streptococcus pneumoniae EPSP synthase in unliganded state, tetrahedral intermediate-bound state and S3P-GLP-bound state", *Molecular Microbiology*, 51.4: 963-971, (2004).

168. Hanelt, I., D. Wunnicke, E. Bordignon, H. J. Steinhoff, and D. J. Slotboom, "Conformational heterogeneity of the aspartate transporter Glt(Ph)", *Nature Structural & Molecular Biology*, 20.2: 210-214, (2013).

169. Xu, C. Y., D. Tobi, and I. Bahar, "Allosteric changes in protein structure computed by a simple mechanical model: Hemoglobin T <-> R2 transition", *Journal of Molecular Biology*, 333.1: 153-168, (2003).

170. Saecker, R. M. and M. T. Record, "Protein surface salt bridges and paths for DNA wrapping", *Current Opinion in Structural Biology*, 12.3: 311-319, (2002).

171. Zheng, W. J., B. R. Brooks, and D. Thirumalai, "Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations", *Proceedings of the National Academy of Sciences of the United States of America*, 103.20: 7664-7669, (2006).

172. Ramos, J. L., M. Martinez-Bueno, A. J. Molina-Henares, W. Teran, K. Watanabe, X. D. Zhang, M. T. Gallegos, R. Brennan, and R. Tobes, "The TetR family of transcriptional repressors", *Microbiology and Molecular Biology Reviews*, 69.2: 326-+, (2005).

173. Cramer, P., D. A. Bushnell, and R. D. Kornberg, "Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution", *Science*, 292.5523: 1863-76, (2001).

174. Armache, K. J., S. Mitterweger, A. Meinhart, and P. Cramer, "Structures of complete RNA polymerase II and its subcomplex, Rpb4/7", *The Journal of Biological Chemistry*, 280.8: 7131-4, (2005).

175. Wang, D., D. A. Bushnell, K. D. Westover, C. D. Kaplan, and R. D. Kornberg, "Structural basis of transcription: Role of the trigger loop in substrate specificity and catalysis", *Cell,* 127.5: 941-954, (2006).

176. Daily, M. D., T. J. Upadhyaya, and J. J. Gray, "Contact rearrangements form coupled networks from local motions in allosteric proteins", *Proteins-Structure Function and Bioinformatics,* 71.1: 455-466, (2008).

177. Kurt, N., T. Haliloglu, and C. A. Schiffer, "Structure-based prediction of potential binding and nonbinding peptides to HIV-1 protease", *Biophys J* 85.2: 853-63, (2003).

178. Stevens, F. C., "Calmodulin: an introduction", *Canadian Journal of Biochemistry and Cell Biology,* 61.8: 906-10, (1983).

179. Chou, J. J., S. Li, C. B. Klee, and A. Bax, "Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains", *Nature Structural & Molecular Biology,* 8.11: 990-7, (2001).

180. Korkut, A. and W. A. Hendrickson, "A force field for virtual atom molecular mechanics of proteins", *Proc Natl Acad Sci U S A* 106.37: 15667-72, (2009).

181. Muller, C. W., G. J. Schlauderer, J. Reinstein, and G. E. Schulz, "Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding", *Structure,* 4.2: 147-56, (1996).

182. Muller, C. W. and G. E. Schulz, "Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 A resolution. A model for a catalytic transition state", *Journal of Molecular Biology,* 224.1: 159-77, (1992).

183. Beckstein, O., E. J. Denning, J. R. Perilla, and T. B. Woolf, "Zipping and Unzipping of Adenylate Kinase: Atomistic Insights into the Ensemble of Open <-> Closed Transitions", *Journal of Molecular Biology,* 394.1: 160-176, (2009).

184. Gur, M., J. D. Madura, and I. Bahar, "Global transitions of proteins explored by a multiscale hybrid methodology: application to adenylate kinase", *Biophysical Journal,* 105.7: 1643-52, (2013).

185. Meireles, L., M. Gur, A. Bakan, and I. Bahar, "Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins", *Protein Science,* 20.10: 1645-58, (2011).

186. Gibson, K. D. and H. A. Scheraga, "Revised Algorithms for the Buildup Procedure for Predicting Protein Conformations by Energy Minimization", *Journal of Computational Chemistry,* 8.6: 826-834, (1987).

187. Krebs, W. G., V. Alexandrov, C. A. Wilson, N. Echols, H. Y. Yu, and M. Gerstein, "Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic", *Proteins-Structure Function and Genetics,* 48.4: 682-695, (2002).

188. Yang, L., G. Song, and R. L. Jernigan, "How well can we understand large-scale protein motions using normal modes of elastic network models?", *Biophysical Journal*, 93.3: 920-9, (2007).

189. Bahar, I., "On the functional significance of soft modes predicted by coarse-grained models for membrane proteins", *The Journal of General Physiology*, 135.6: 563-73, (2010).

190. Fenton, W. A., Y. Kashi, K. Furtak, and A. L. Horwich, "Residues in chaperonin GroEL required for polypeptide binding and release", *Nature*, 371.6498: 614-9, (1994).

191. Fenton, W. A. and A. L. Horwich, "GroEL-mediated protein folding", *Protein Science*, 6.4: 743-60, (1997).

192. Sigler, P. B., Z. Xu, H. S. Rye, S. G. Burston, W. A. Fenton, and A. L. Horwich, "Structure and function in GroEL-mediated protein folding", *Annual Review of Biochemistry*, 67: 581-608, (1998).

193. Weber, F., F. Keppel, C. Georgopoulos, M. K. Hayer-Hartl, and F. U. Hartl, "The oligomeric structure of GroEL/GroES is required for biologically significant chaperonin function in protein folding (vol 5, pg 977, 1998)", *Nature Structural Biology*, 6.2: 200-200, (1999).

194. Saibil, H. R. and N. A. Ranson, "The chaperonin folding machine", *Trends in Biochemical Sciences*, 27.12: 627-32, (2002).

195. Thirumalai, D. and G. H. Lorimer, "Chaperonin-mediated protein folding", *Annual Review of Biophysics and Biomolecular Structure*, 30: 245-69, (2001).

196. Horovitz, A. and K. R. Willison, "Allosteric regulation of chaperonins", *Current Opinion in Structural Biology*, 15.6: 646-51, (2005).

197. Shtilerman, M., G. H. Lorimer, and S. W. Englander, "Chaperonin function: folding by forced unfolding", *Science*, 284.5415: 822-5, (1999).

198. Xu, Z., A. L. Horwich, and P. B. Sigler, "The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex", *Nature*, 388.6644: 741-50, (1997).

199. Weissman, J. S., C. M. Hohl, O. Kovalenko, Y. Kashi, S. Chen, K. Braig, H. R. Saibil, W. A. Fenton, and A. L. Horwich, "Mechanism of GroEL action: productive release of polypeptide from a sequestered position under GroES", *Cell*, 83.4: 577-87, (1995).

200. Boisvert, D. C., J. Wang, Z. Otwinowski, A. L. Horwich, and P. B. Sigler, "The 2.4 A crystal structure of the bacterial chaperonin GroEL complexed with ATP gamma S", *Nature Structural Biology*, 3.2: 170-7, (1996).

201. Chen, S., A. M. Roseman, A. S. Hunter, S. P. Wood, S. G. Burston, N. A. Ranson, A. R. Clarke, and H. R. Saibil, "Location of a folding protein and shape changes in GroEL-GroES complexes imaged by cryo-electron microscopy", *Nature,* 371.6494: 261-4, (1994).

202. Mayhew, M., A. C. da Silva, J. Martin, H. Erdjument-Bromage, P. Tempst, and F. U. Hartl, "Protein folding in the central cavity of the GroEL-GroES chaperonin complex", *Nature,* 379.6564: 420-6, (1996).

203. Yifrach, O. and A. Horovitz, "Nested cooperativity in the ATPase activity of the oligomeric chaperonin GroEL", *Biochemistry,* 34.16: 5303-8, (1995).

204. Hyeon, C., G. H. Lorimer, and D. Thirumalai, "Dynamics of allosteric transitions in GroEL", *Proceedings of the National Academy of Sciences of the United States of America,* 103.50: 18939-44, (2006).

205. Bahar, I., C. Chennubhotla, and D. Tobi, "Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation", *Current Opinion in Structural Biology,* 17.6: 633-40, (2007).

206. *Sparse Eigenvalue Problems with ARPACK*,https://docs.scipy.org/doc/scipy-0.18.1/reference/tutorial/arpack.html

207. Jones, Michael B., Daniela Ro, #351, Marcel-C, #259, #259, lin Ro, and #351, "CPU reservations and time constraints: efficient, predictable scheduling of independent activities", *SIGOPS - Operating Systems Review,* 31.5: 198-211, (1997).

208. Pollard, T. D. and E. D. Korn, "Acanthamoeba myosin. I. Isolation from Acanthamoeba castellanii of an enzyme similar to muscle myosin", *The Journal of Biological Chemistry,* 248.13: 4682-90, (1973).

209. Zheng, W. and B. R. Brooks, "Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: myosin versus kinesin", *Biophysical Journal,* 89.1: 167-78, (2005).

210. MacRae, I. J., I. H. Segel, and A. J. Fisher, "Crystal structure of ATP sulfurylase from Penicillium chrysogenum: insights into the allosteric regulation of sulfate assimilation", *Biochemistry,* 40.23: 6795-804, (2001).

211. Ullrich, T. C., M. Blaesse, and R. Huber, "Crystal structure of ATP sulfurylase from Saccharomyces cerevisiae, a key enzyme in sulfate activation", *The EMBO Journal,* 20.3: 316-29, (2001).

212. Gerstein, M. and N. Echols, "Exploring the range of protein flexibility, from a structural proteomics perspective", *Current Opinion in Chemical Biology,* 8.1: 14-9, (2004).

213. Kirillova, S., J. Cortes, A. Stefaniu, and T. Simeon, "An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins", *Proteins,* 70.1: 131-43, (2008).

214. Hammes, G. G., "Multiple conformational changes in enzyme catalysis", *Biochemistry,* 41.26: 8221-8, (2002).

215. Daily, M. D., T. J. Upadhyaya, and J. J. Gray, "Contact rearrangements form coupled networks from local motions in allosteric proteins", *Proteins,* 71.1: 455-66, (2008).

216. Demerdash, O. N. and J. C. Mitchell, "Density-cluster NMA: A new protein decomposition technique for coarse-grained normal mode analysis", *Proteins,* 80.7: 1766-79, (2012).

217. Mahajan, S. and Y. H. Sanejouand, "On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins", *Archives of Biochemistry and Biophysics,* 567: 59-65, (2015).

218. Hayward, S. and H. J. C. Berendsen, "Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme", *Proteins-Structure Function and Genetics,* 30.2: 144-154, (1998).

219. Remington, S., G. Wiegand, and R. Huber, "Crystallographic Refinement and Atomic Models of 2 Different Forms of Citrate Synthase at 2.7-a and 1.7-a Resolution", *Journal of Molecular Biology,* 158.1: 111-152, (1982).

220. Wiegand, G. and S. J. Remington, "Citrate Synthase - Structure, Control, and Mechanism", *Annual Review of Biophysics and Biophysical Chemistry,* 15: 97-117, (1986).

221. Iizuka, R., T. Yoshida, Y. Shomura, K. Miki, T. Maruyama, M. Odaka, and M. Yohda, "ATP binding is critical for the conformational change from an open to closed state in archaeal group II chaperonin", *Journal of Biological Chemistry,* 278.45: 44959-44965, (2003).

222. Lee, H., S. Seo, M. Kim, J. B. Choi, S. M. Kim, T. J. Jeon, and M. K. Kim, "Opening and closing of a toroidal group II chaperonin revealed by a symmetry constrained elastic network model", *Protein Science,* 23.6: 703-713, (2014).

223. Kim, M. K., G. S. Chirikjian, and R. L. Jernigan, "Elastic models of conformational transitions in macromolecules", *Journal of Molecular Graphics & Modelling,* 21.2: 151-160, (2002).

224. Yernool, D., O. Boudker, Y. Jin, and E. Gouaux, "Structure of a glutamate transporter homologue from Pyrococcus horikoshii", *Nature,* 431.7010: 811-8, (2004).

225. Verdon, G. and O. Boudker, "Crystal structure of an asymmetric trimer of a bacterial glutamate transporter homolog", *Nature Structural & Molecular Biology,* 19.3: 355-7, (2012).

226. Reyes, N., C. Ginter, and O. Boudker, "Transport mechanism of a bacterial homologue of glutamate transporters", *Nature*, 462.7275: 880-5, (2009).

227. Jiang, J., I. H. Shrivastava, S. D. Watts, I. Bahar, and S. G. Amara, "Large collective motions regulate the functional properties of glutamate transporter trimers", *Proceedings of the National Academy of Sciences of the United States of America*, 108.37: 15141-6, (2011).

228. Koch, H. P. and H. P. Larsson, "Small-scale molecular motions accomplish glutamate uptake in human glutamate transporters", *Journal of Neuroscience*, 25.7: 1730-6, (2005).

229. Grkovic, S., M. H. Brown, M. J. Roberts, I. T. Paulsen, and R. A. Skurray, "QacR is a repressor protein that regulates expression of the Staphylococcus aureus multidrug efflux pump QacA", *Journal of Biological Chemistry*, 273.29: 18665-18673, (1998).

230. Marques, M. R., A. Vaso, J. Ruggiero, M. A. Fossey, J. S. Oliveira, L. A. Basso, D. S. dos Santos, W. F. de Azevedo, and M. S. Palma, "Dynamics of glyphosate-induced conformational changes of Mycobacterium tuberculosis 5-enolpyruvylshikimate-3-phosphate synthase (EC 2.5.1.19) determined by hydrogen-deuterium exchange and electrospray mass spectrometry", *Biochemistry*, 47.28: 7509-7522, (2008).

231. Borges, J. C., J. H. Pereira, I. B. Vasconcelos, G. C. dos Santos, J. R. Olivieri, C. H. I. Ramos, M. S. Palma, L. A. Basso, D. S. Santos, and W. F. de Azevedo, "Phosphate closes the solution structure of the 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) from Mycobacterium tuberculosis", *Archives of Biochemistry and Biophysics*, 452.2: 156-164, (2006).

232. Lewis, M., G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Z. Lu, "Crystal structure of the lactose operon repressor and its complexes with DNA and inducer", *Science*, 271.5253: 1247-1254, (1996).

233. Flynn, T. C., L. Swint-Kruse, Y. F. Kong, C. Booth, K. S. Matthews, and J. P. Ma, "Allosteric transition pathways in the lactose repressor protein core domains: Asymmetric motions in a homodimer", *Protein Science*, 12.11: 2523-2541, (2003).

234. Svetlov, V. and E. Nudler, "Basic mechanism of transcription by RNA polymerase II", *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms*, 1829.1: 20-28, (2013).

235. Feig, M. and Z. F. Burton, "RNA Polymerase II with Open and Closed Trigger Loops: Active Site Dynamics and Nucleic Acid Translocation", *Biophysical Journal*, 99.8: 2577-2586, (2010).

236. Cheung, A. C. M. and P. Cramer, "A Movie of RNA Polymerase II Transcription", *Cell*, 149.7: 1431-1437, (2012).

237. Armache, K. J., H. Kettenberger, and P. Cramer, "Architecture of initiation-competent 12-subunit RNA polymerase II", *Proceedings of the National Academy of Sciences of the United States of America*, 100.12: 6964-8, (2003).

238. Bushnell, D. A. and R. D. Kornberg, "Complete, 12-subunit RNA polymerase II at 4.1-A resolution: implications for the initiation of transcription", *Proceedings of the National Academy of Sciences of the United States of America*, 100.12: 6969-73, (2003).

239. Edwards, A. M., C. M. Kane, R. A. Young, and R. D. Kornberg, "Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter in vitro", *The Journal of Biological Chemistry*, 266.1: 71-5, (1991).

240. Christoffersen, S., A. Kadziola, E. Johansson, M. Rasmussen, M. Willemoes, and K. F. Jensen, "Structural and Kinetic Studies of the Allosteric Transition in Sulfolobus solfataricus Uracil Phosphoribosyltransferase: Permanent Activation by Engineering of the C-Terminus", *Journal of Molecular Biology*, 393.2: 464-477, (2009).

241. Stansfeld, P. J. and M. S. P. Sansom, "From Coarse Grained to Atomistic: A Serial Multiscale Approach to Membrane Protein Simulations", *Journal of Chemical Theory and Computation*, 7.4: 1157-1166, (2011).

242. Chen, V. B., W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography", *Acta Crystallographica Section D-Biological Crystallography*, 66: 12-21, (2010).

APPENDIX A: MODE, OVERLAP AND COLLECTIVITY PROFILES IN ANM-MC SIMULATIONS

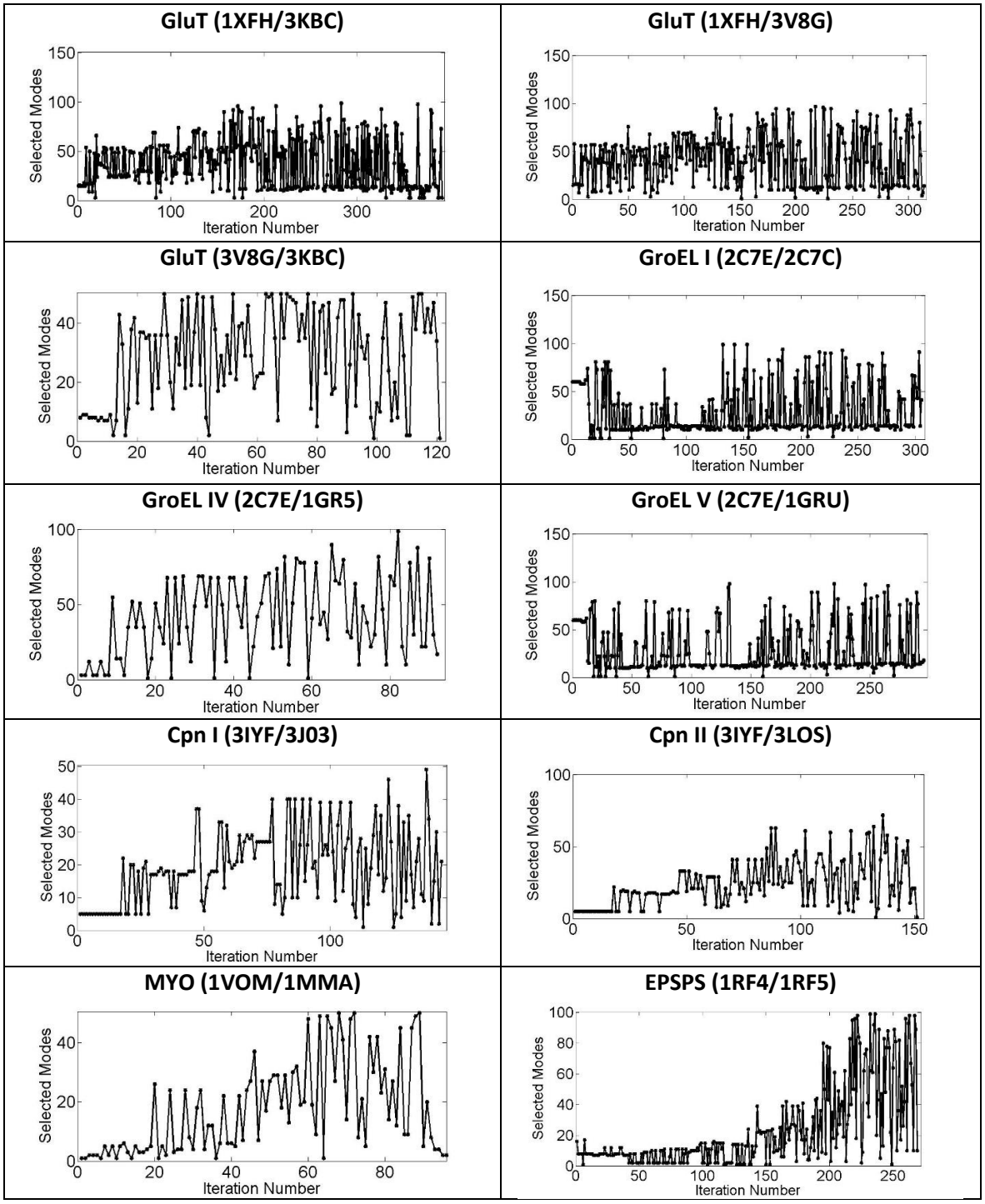


Figure A.1. Mode profiles in modified ANM-MC runs (k = 500 J/Å/mol, MC PS = 0.15 Å).

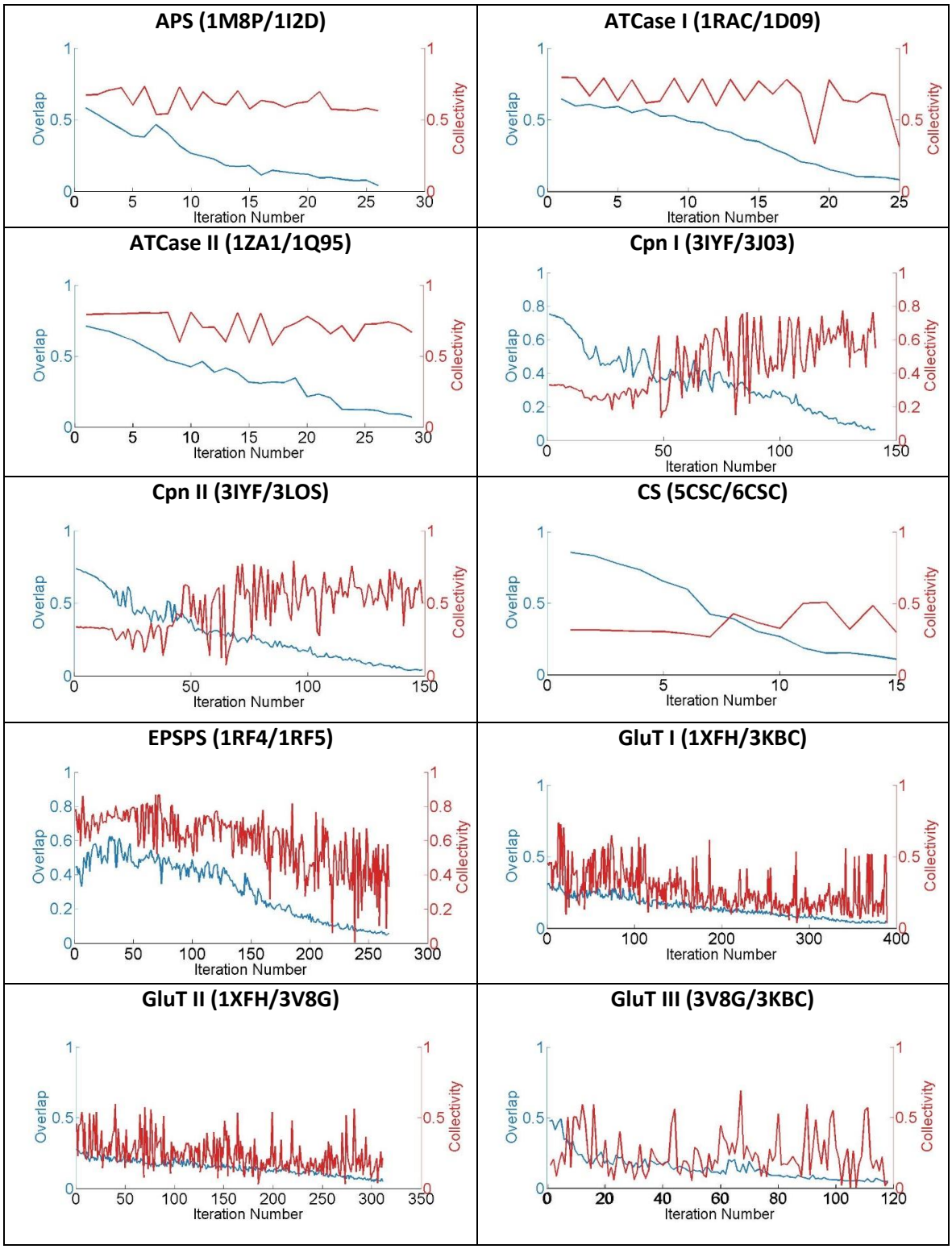


Figure A.2. Overlap and collectivity data in modified ANM-MC runs (k = 500 J/Å/mol, MC PS = 0.15 Å).

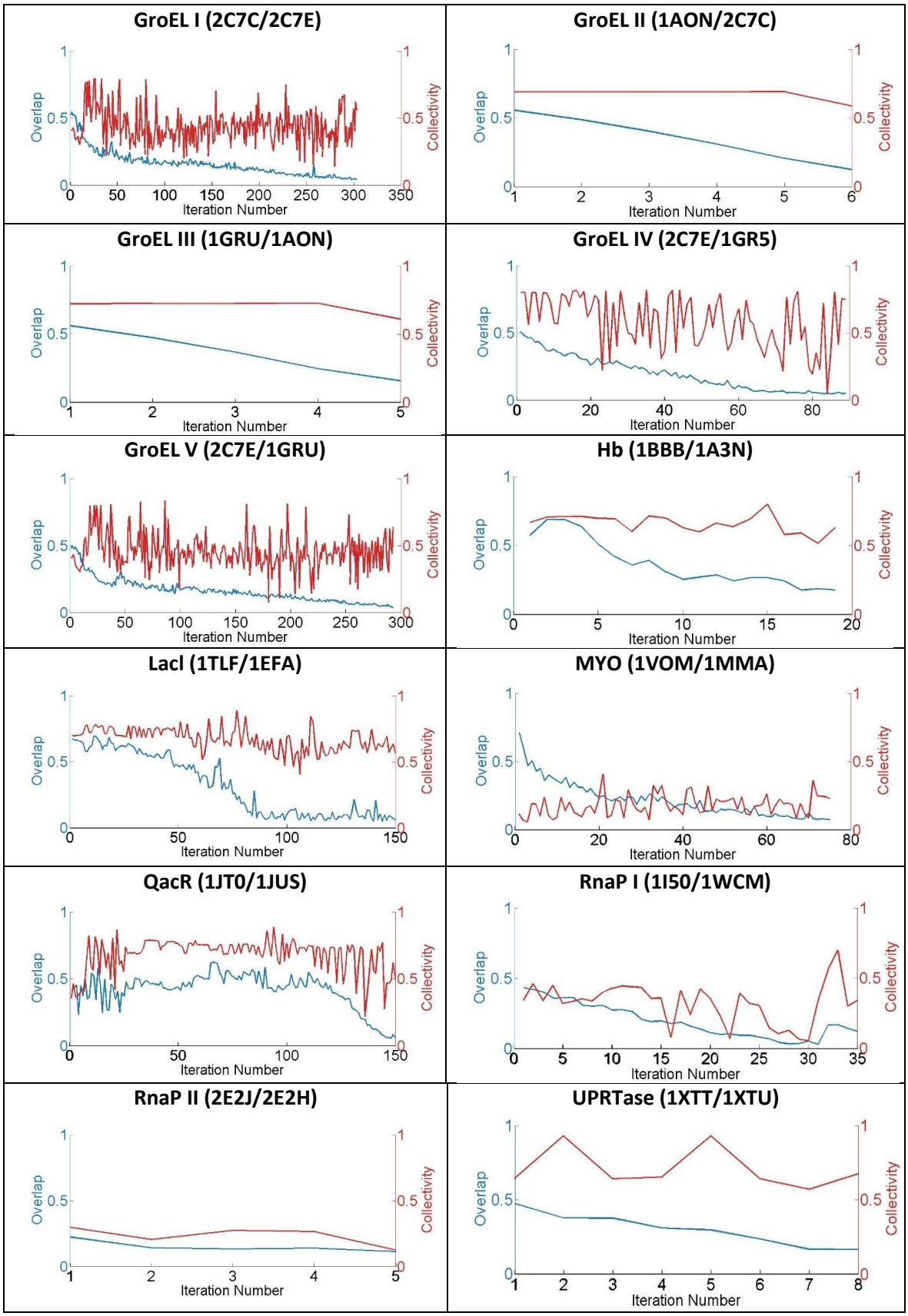


Figure A.2. Overlap and collectivity data in modified ANM-MC runs (k = 500 J/Å/mol, MC PS = 0.15 Å) (cont.)

APPENDIX B: ENERGY AND RMSD PROFILES IN ANM-MC SIMULATIONS

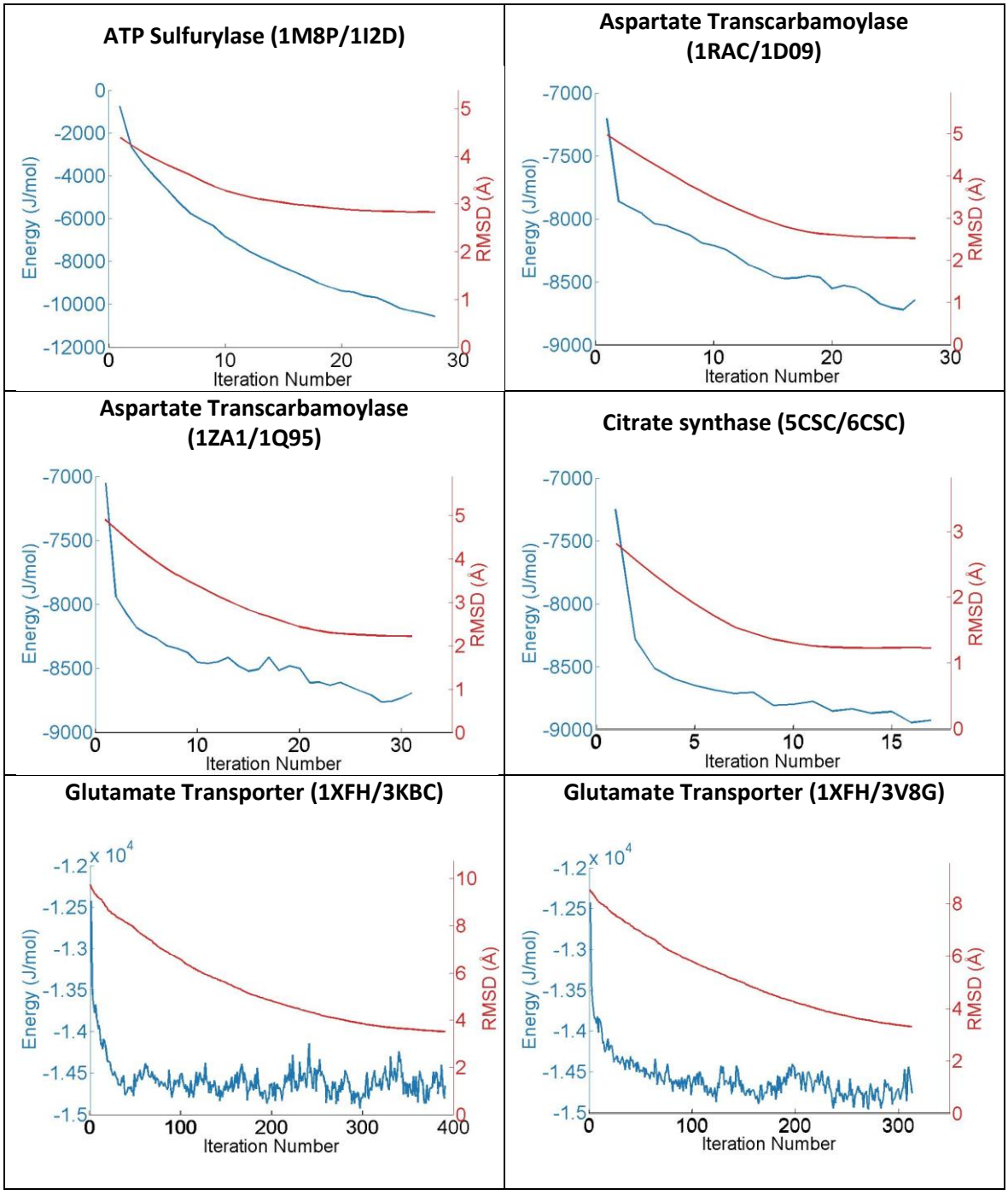


Figure B.1. RMSD and energy profiles of ANM-MC runs.

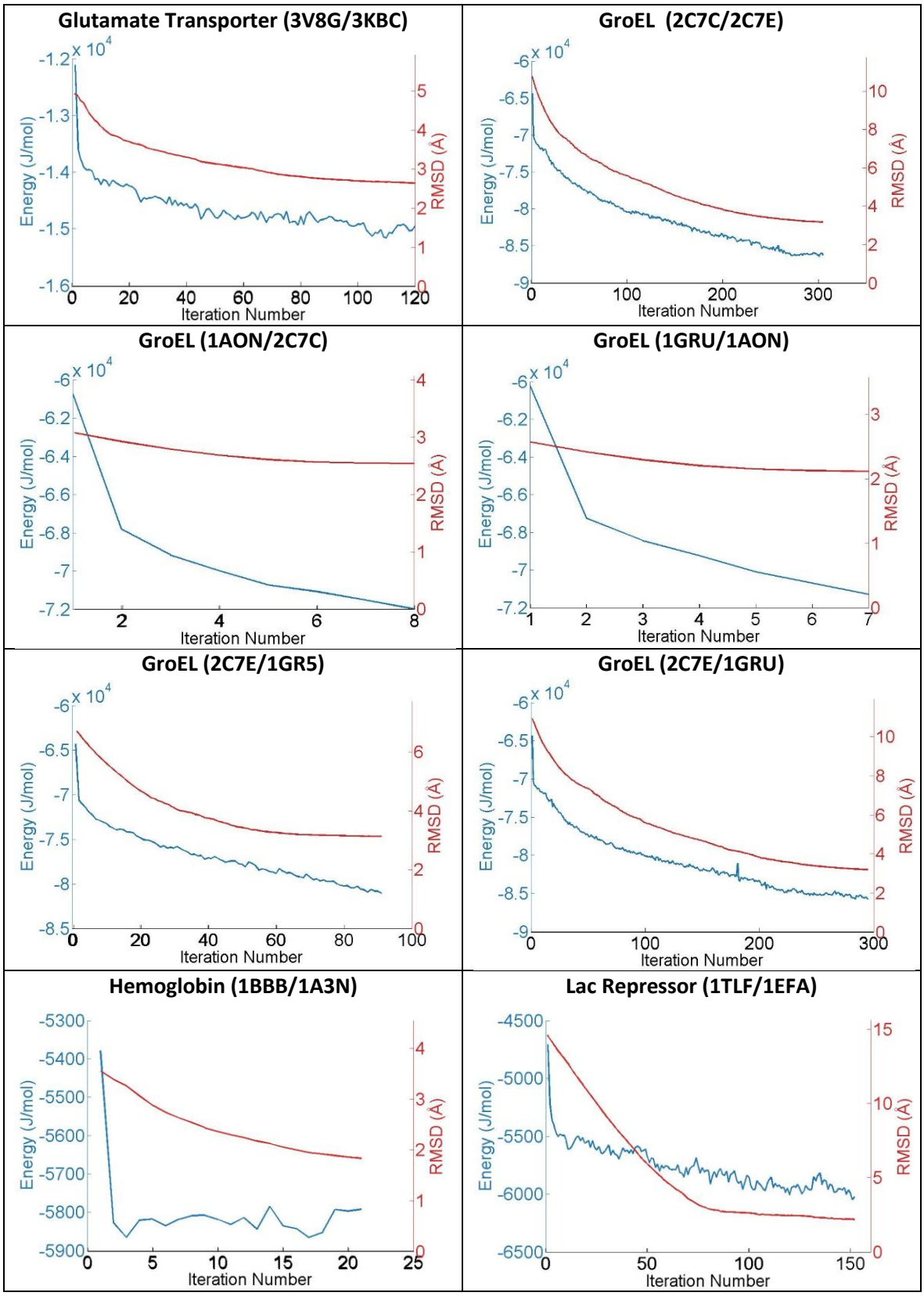


Figure B.1. RMSD and energy profiles of ANM-MC runs (cont.)

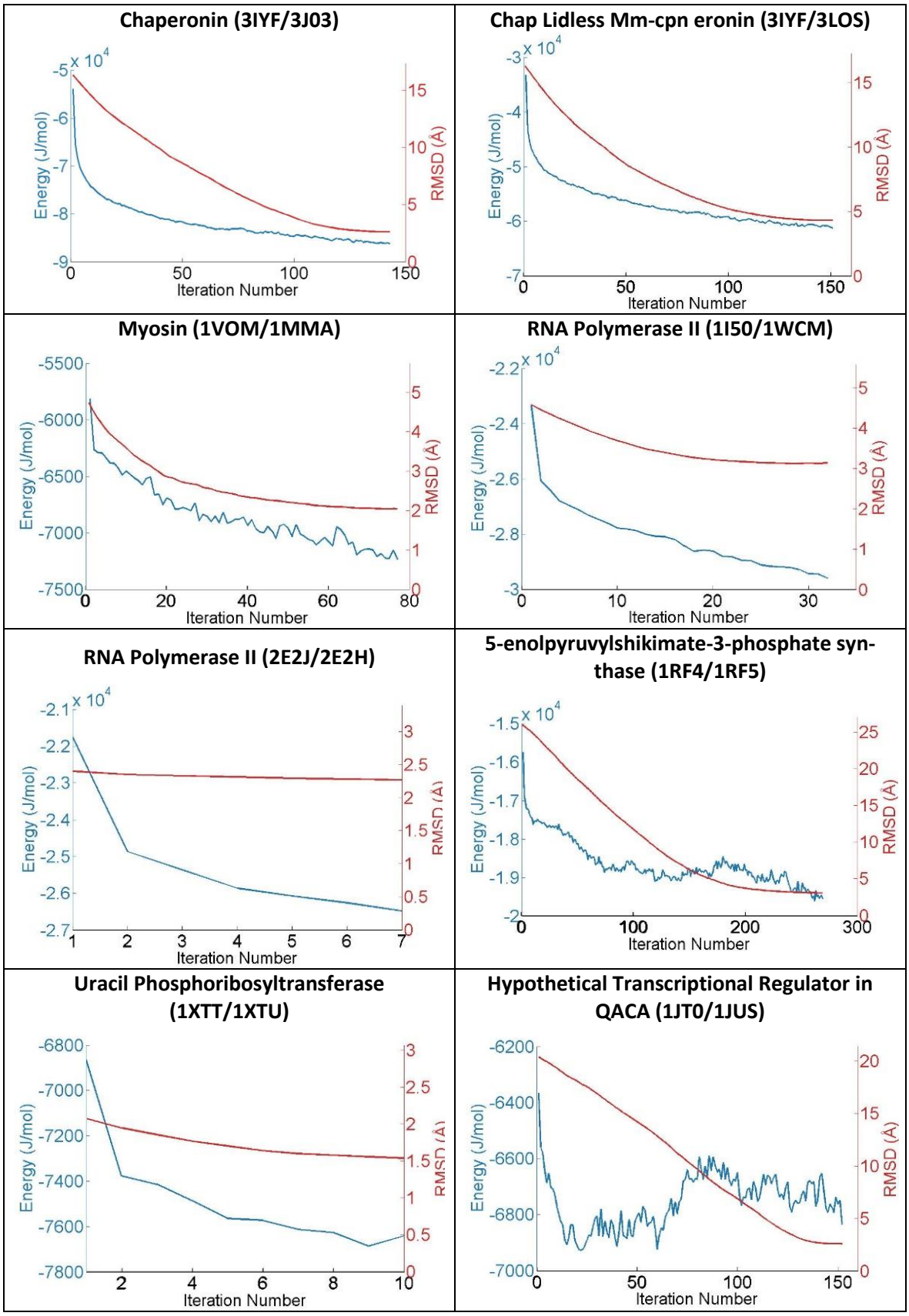


Figure B.1. RMSD and energy profiles of ANM-MC runs (cont.)

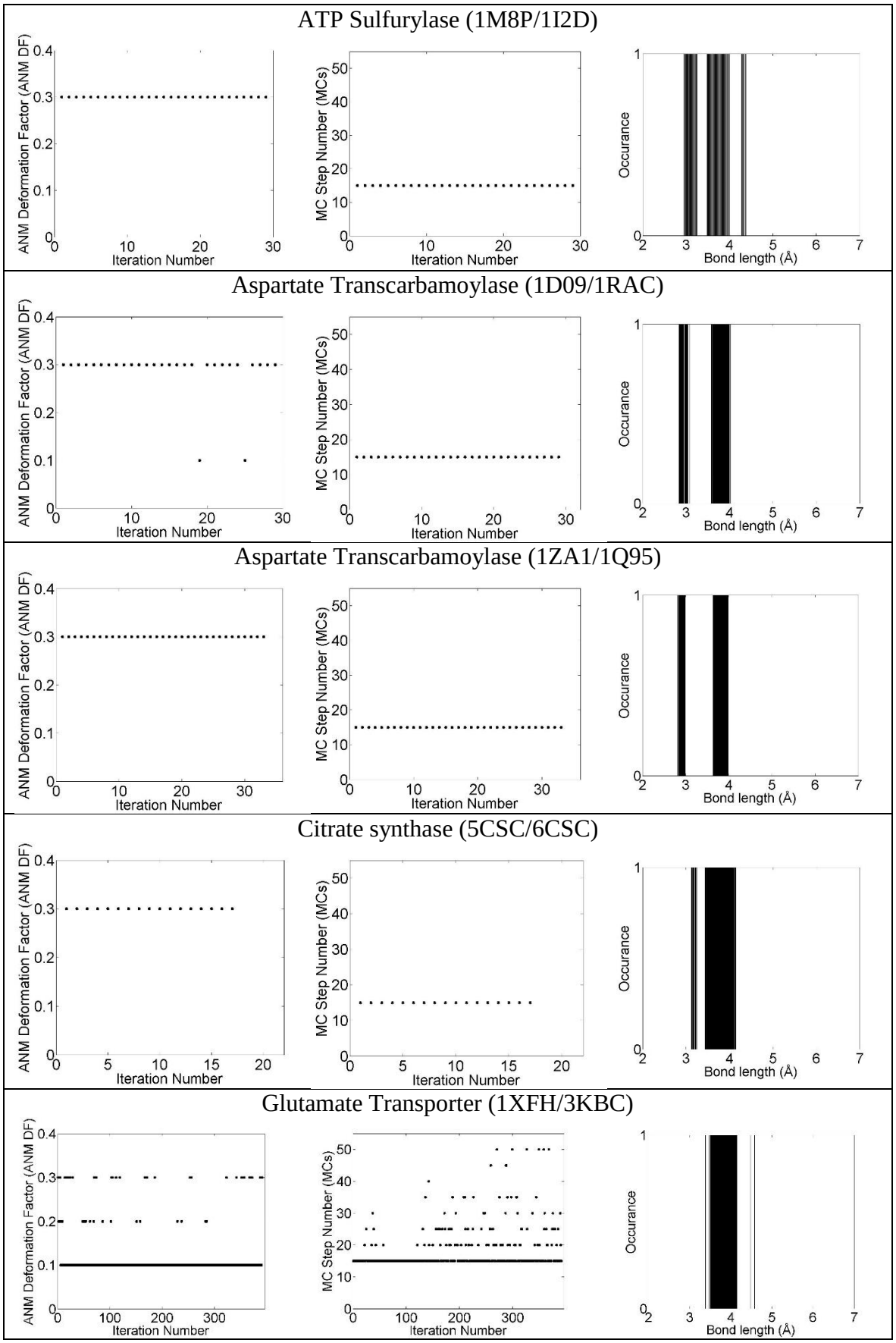


Figure B.2. Profiles of ANM-MC parameters.

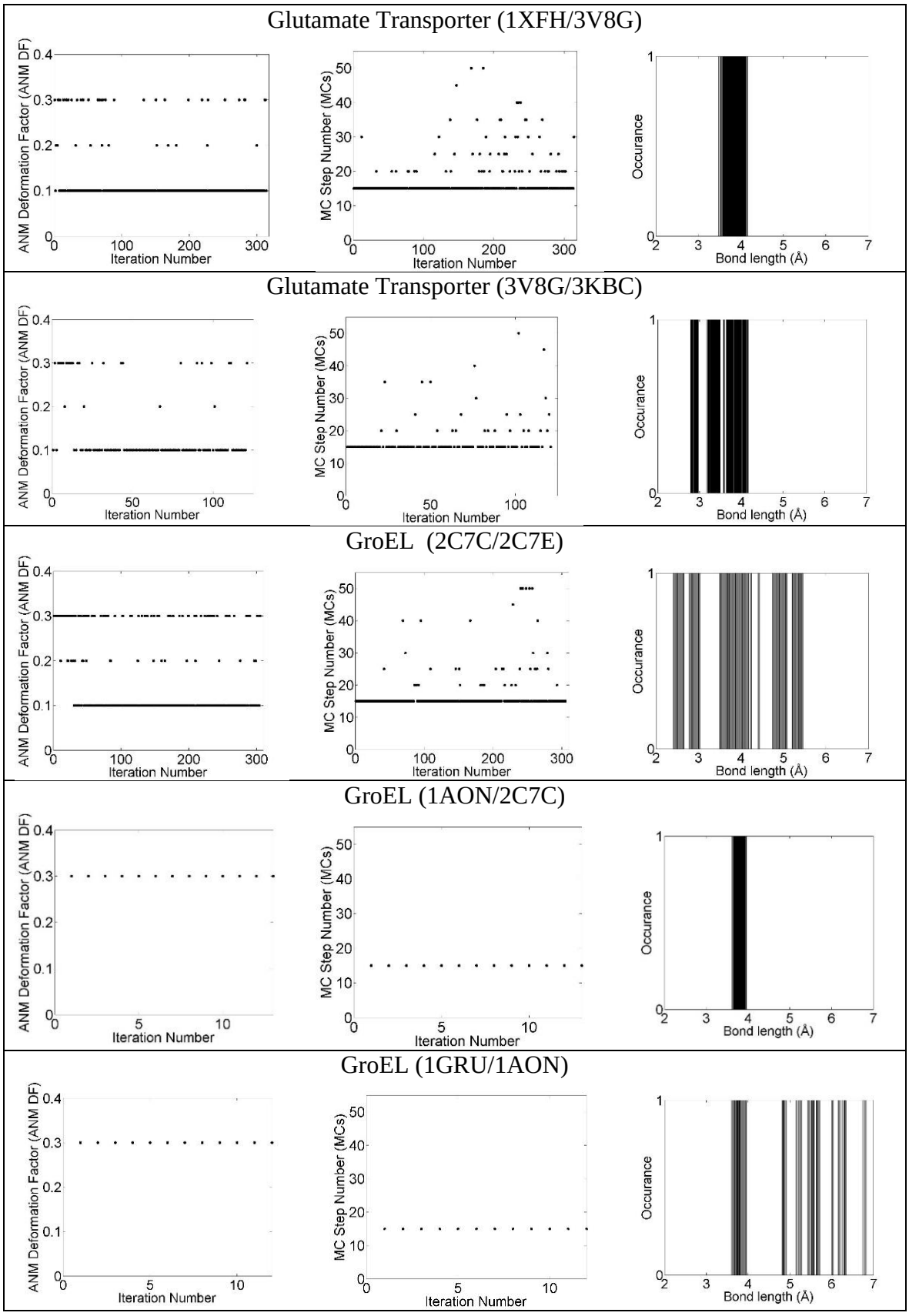


Figure B.2. Profiles of ANM-MC parameters (cont.)

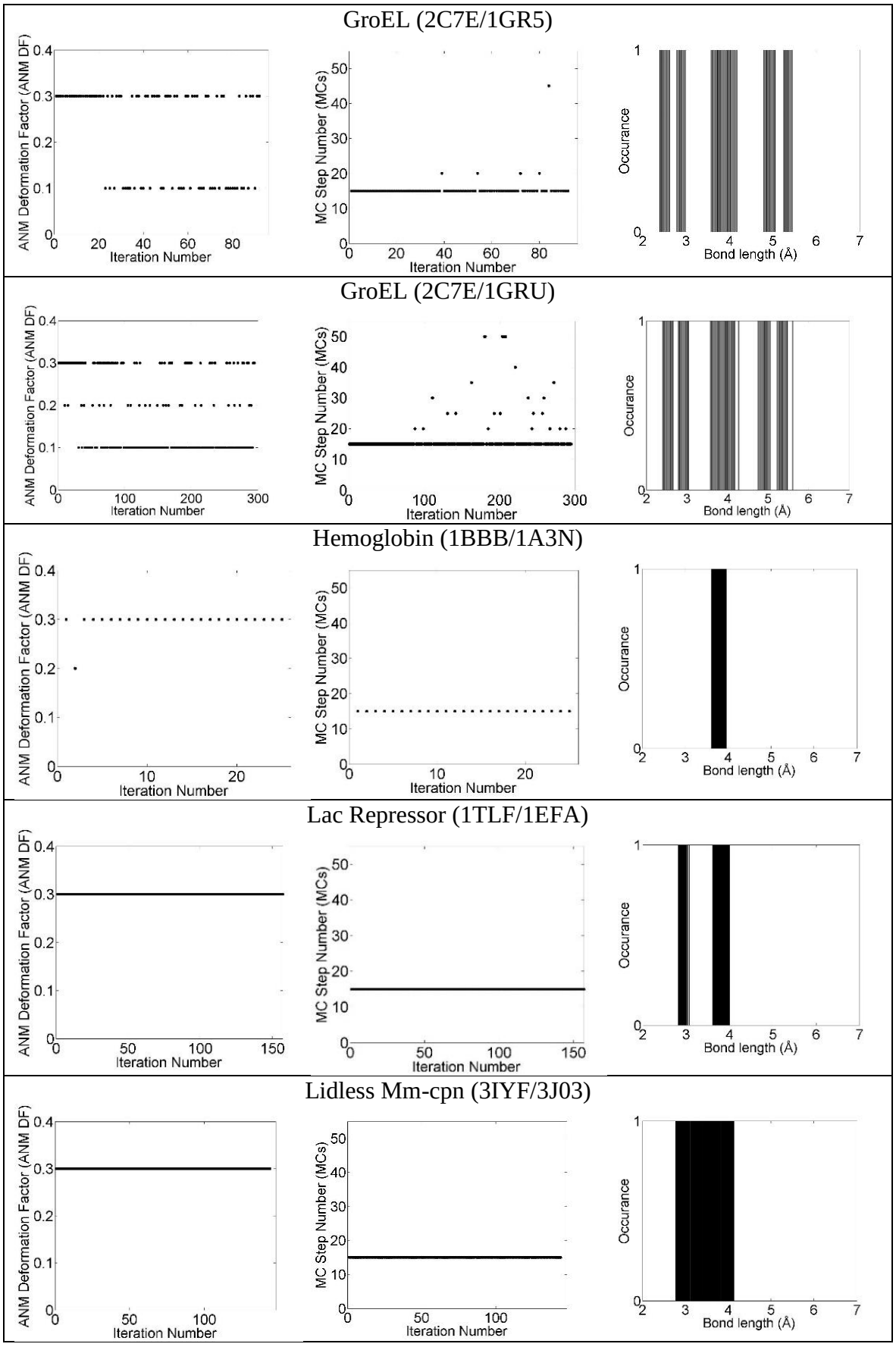


Figure B.2. Profiles of ANM-MC parameters (cont.)

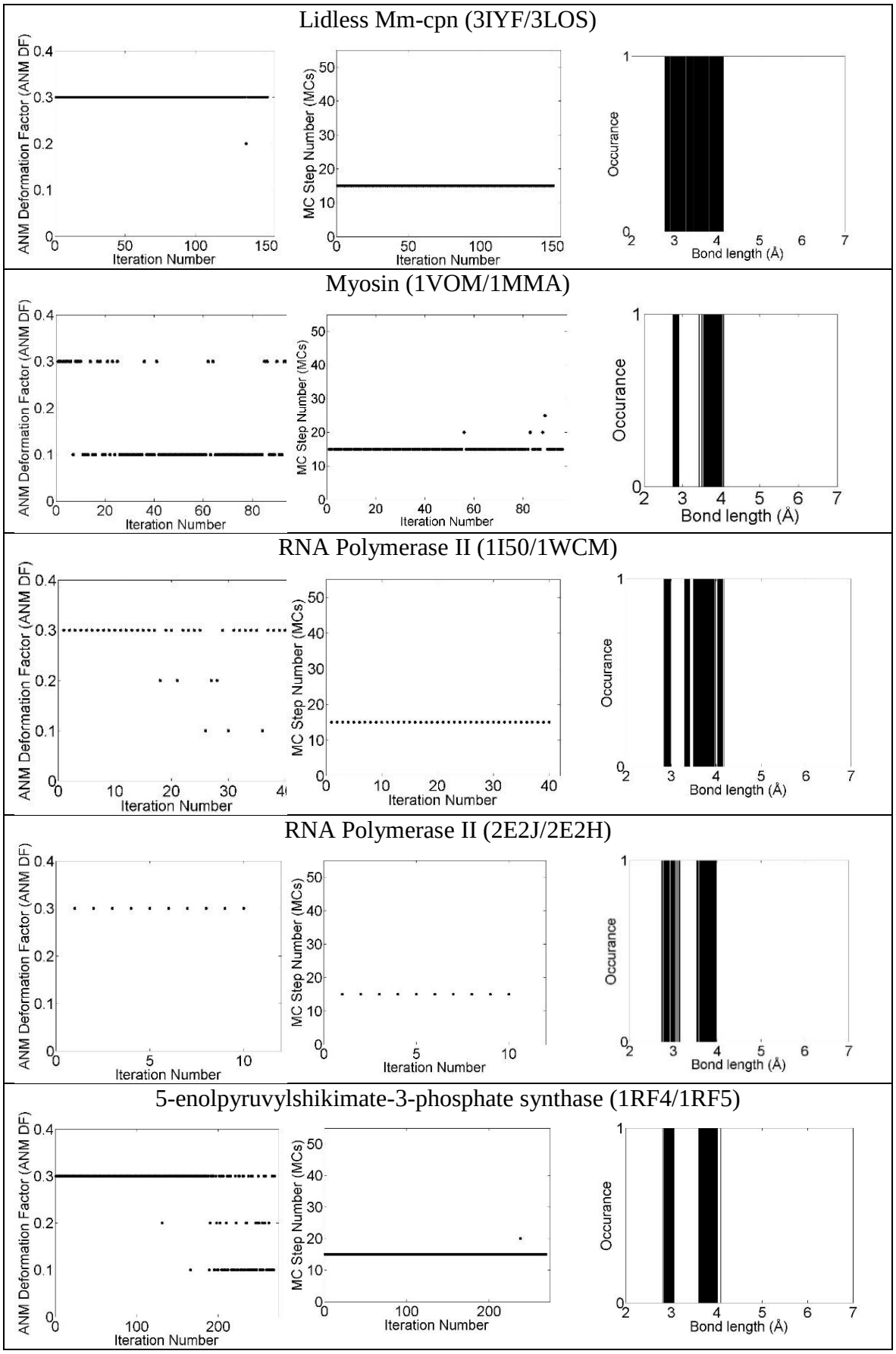


Figure B.2. Profiles of ANM-MC parameters (cont.)

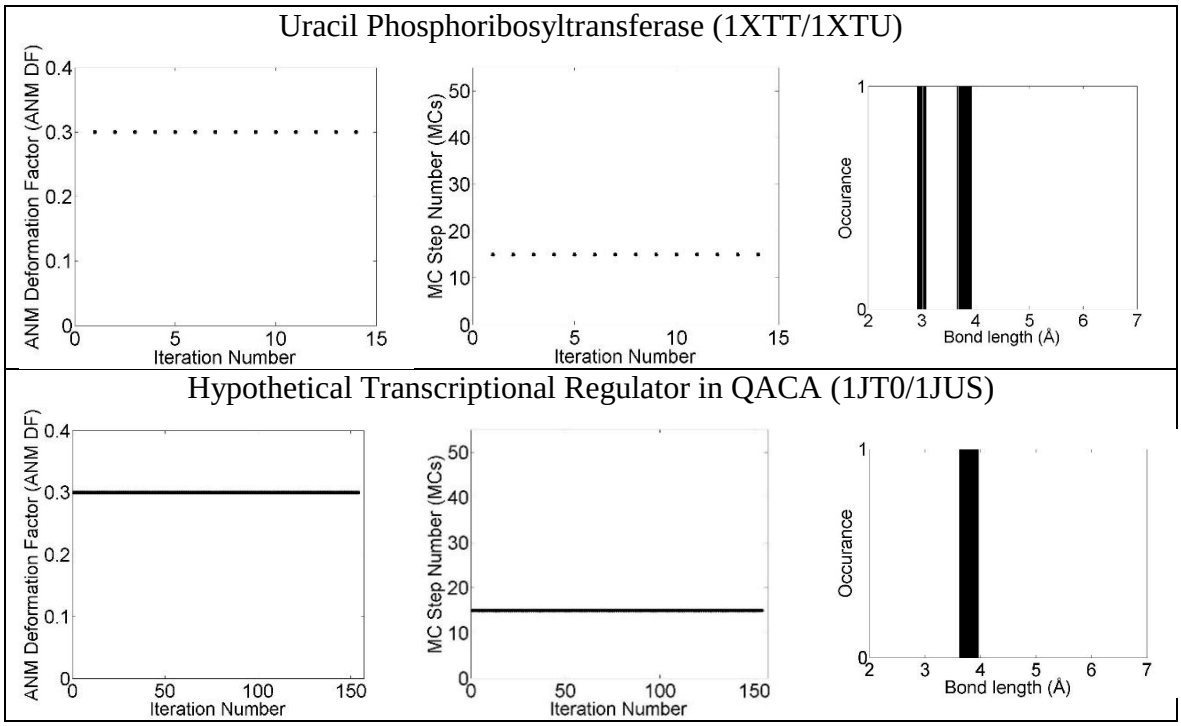


Figure B.2. Profiles of ANM-MC parameters (cont.)

APPENDIX C: WALL-CLOCK TIMES OF ANM-MC SIMULATIONS

Table C.1. Wall-clock times of ANM-MC runs.

Proteins	**Time (ANM) (s)**	**Time (MC – 1 step) (s)**	**Number of iterations**	**Wall-clock Time (dd:hh:mm)**
APS	20	14	30	00:01:52
ATCase I	6	7	30	00:00:58
ATCase II	6	7	33	00:01:04
Cpn I	398	63	144	02:05:43
Cpn II	377	63	151	02:07:33
CS	5	7	18	00:00:32
EPSPS	19	6	269	00:17:39
GluT I	26	10	391	01:05:39
GluT II	21	10	314	00:23:43
GluT III	10	10	129	00:09:54
GroEL I	320	60	305	06:05:30
GroEL II	388	67	13	00.05:01
GroEL III	426	67	43	00:12:07
GroEL IV	368	60	295	06:05:11
GroEL V	290	60	93	01:12:53
Hb	2	5	32	00:00:37
LacI	2	8	158	00:05:01
MYO	4	6	76	00:03:43
QacR	4	6	154	00:03:57
RnaP I	78	51	31	00:07:17
RnaP II	67	51	10	00:02:19
UPRTase	5	6	15	00:00.26