MOLECULAR DYNAMICS OF ADENYLATE KINASE IN DIFFERENT CONFORMATIONAL TRANSITION STATES

A Dissertation Presented
By
Yujing Wang
to
Department of Bioengineering
Northeastern University

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
In the field of
Bioengineering

Northeastern University
Boston, Massachusetts

December 2016
# Contents

List of Tables 4  
List of Figures 5  
List of Abbreviations 7  
Abstract 8  

## Chapter 1. Introduction

1.1 Enzyme structure and dynamics provide support for protein function 10  
1.2 Adenylate kinase (ADK) as a model system for application of MD 16  
1.3 SAXS/WAXS as an experimental tool for the study of structure and dynamics 18  

## Chapter 2. Methods - Computational

2.1 Technical details 20  
2.2 Implementation of molecular dynamics (MD) program for ADK 21  
  2.2.1 GROMACS 21  
  2.2.2 NAMD 22  
  2.2.3 VMD 23  
  2.2.4 SMOG 24  
  2.2.5 Biopython Project 26  
2.3 Custom Built Software 27  
  2.3.1 K-means clustering 27  
  2.3.2 Radius of gyration (ROG) 31  
  2.3.3 Cα atom pseudodiheiral angle fluctuations 33  
  2.3.4 Root mean square fluctuation (RMSF) between each Cα atom pair 34  
2.4 Comparison to SAXS/WAXS experimental results 36  
  2.4.1 Implementation of XS 36  
  2.4.2 SAXS for determination of conformational equilibria 40
Chapter 3. Result and Discussion

3.1. GROMACS MD simulation details ............................................................... 43
3.2. Calculation of RMSD for each trajectory ..................................................... 44
3.3. The comparison between 45 crystal structures and the ensemble generated by an MD trajectory ................................................................. 46
3.4. Clustering of structures from the MD trajectory ............................................ 48
3.5. Properties of clusters ................................................................................ 51
3.6. Circulation .................................................................................................. 52
3.7. Comparison among the average structures of the three partial open clusters..... 57
3.8. Changes of radius of gyration of domains of ADK in different clusters .......... 60
3.9. Standard deviation of interatomic distance variations for each state............. 64
3.10. Pseudodiheral angle changes along the MD trajectory ............................... 73
3.11. The comparison of ADK secondary structure RMSD change ..................... 81
3.12. The comparison of RMSF for mutant (MT) G85V and Wild type (WT) ....... 92

Chapter 4. Conclusions .................................................................................. 95

Chapter 5. Future directions .......................................................................... 97
List of Tables

Table 1. Transfer rate frequency (TRF) for 16 MD simulations that completed the transition to closed state in under 500 ns (limiting time for each simulation). .......................... 54

Table 2. The percentage of transition from one cluster to the other three clusters........ 55
List of Figures

Figure 1. The location of ligands (ATP, AMP) in the closed conformation of ADK..................12

Figure 2. The P-loop within the crystal structure of E. coli ADK 1AKE in a closed conformation.................................................14

Figure 3. The giant anion hole within the crystal structure of E. coli ADK 1AKE in a closed conformation.................................................16

Figure 4. The LID and NMP domain undergo large conformational transitions in this phosphotransferase reaction.................................................17

Figure 5. Layout of beamline X9 at Brookhaven National Laboratory at NSLS..............18

Figure 6. The RMSD of four MD trajectories, which are plotted along the time sequence .................................................................46

Figure 7. The 45 crystal structures and the ensembles of MD trajectories plotted in two-dimensions.................................................................48

Figure 8. The elbow plot for K selection.................................................................49

Figure 9. The ensembles of previous MD trajectory are clustered into 5 groups by K-means and projected onto the domain mass center distance map ..............................................52

Figure 10. Connectivity map for conformational transitions during the simulation of ADK rearrangement...............................................................53

Figure 11. Testing different algorithms for reconstructing all atom structures from Cα structures ...........................................................................58
Figure 12. The comparison of the backbone of the three average structures of the partial open clusters

Figure 13. The distribution of the ROG of ADK’DK three domains in different clusters

Figure 15. The RMS deviation of distance matrix between Ca atom pairs in open, partially-open and closed states

Figure 16. The comparison of the RMSF of each residue in different state

Figure 17. The blue residues (residue 56 and alpha helix 159-176) shows intense relative motion in open and partial open state

Figure 18. Pseudodihedral angle change along the MD trajectory

Figure 19. The fluctuation change of pseudodihedral angle in different cluster of the MD trajectory

Figure 20. The top 20 most unstable pseudodihedral angles in the trajectory

Figure 21. The comparison of Ca dihedral angle fluctuation standard deviation in different states

Figure 22. The RMSD change of LID (122-160) and CORE(60-122) domain in trajectory 1.27

Figure 23. The RMSD change of N-terminus (1-42) and C-terminus (178-214) region of ADK in trajectory 1.27

Figure 24. The RMSD change of NMP domain (30-60) of ADK in trajectory 1.27

Figure 25. The RMSD change of the long long alpha-helix (161-190) links LID and CORE domains of ADK in trajectory 1.27
Figure 26. The RMSD change around the LID and CORE domains and CORE domains of ADK in trajectory 1.271.27pen clustersap a……………………………………………………………………………………………… 87

Figure 27. The RMSD change of the secondary structures in CORE domains (60-122) in trajectory 1.27……………………………………………………………………………………………… 89

Figure 28. The comparison of RMSF map of MT G85V and WT in partial open state… 94
List of Abbreviations

SAXS = small-angle regime scattering
WAXS = wide-angle regime scattering
MD = Molecular Dynamics
ADK = Adenylate kinase
GROMACS = GROningen MAchine for Chemical Simulation
ROG = radius of gyration
RMSF = root mean square fluctuation
RMSD = root mean square deviation
STD = standard deviation
PME = Particle Mesh Ewald
DIMS = dynamic importance sampling
MT = mutant
WT = Wild type
CAPF = Cα Atom Pair Factor
PDB = Protein Data Bank
NSLS = National Synchrotron Light Source
BNL = Brookhaven National Lab
AP5A = diadenosine pentaphosphate
Abstract

Adenylate kinase (ADK) catalyzes the reversible Mg$^{2+}$-dependent phosphoryl transfer reaction $\text{Mg}^{2+} + 2\text{ADP} \leftrightarrow \text{Mg}^{2+} + \text{ATP} + \text{AMP}$ in essential cellular systems. This reaction is a major player in cellular energy homeostasis and the isoform network of ADK plays an important role in AMP metabolic signaling circuits. As a well-studied protein, the structures of ADK from various organisms have been solved by x-ray crystallography. ADK has three domains, the LID, NMP and CORE domains. Comparison of crystal structures reveals that the three domains undergo large conformational rearrangement during ADK’s catalytic cycle. The LID and NMP domains are directly involved in the dynamics and close over the enzyme’s ATP and AMP binding sites on ligand binding. The conformational switching and protein dynamics are recognized to be important for the enzyme functions. In this thesis I explore the molecular motions which connect the simple static crystal structures.

The conformational transitions of ADK are characterized while it undergoes conformational changes required for catalytic cycling in order to understand the contributions of intermediate sub-states to catalysis. As a first step, the intermediate, partially open conformation is divided into sub-states using K-means clustering based on RMS differences. We then characterize the dynamic behavior of each structural domain in the different sub-states using a number of metrics including radius of gyration, dihedral angle fluctuation, interatomic pair fluctuation and others. This allows us to determine whether or not there are correlations between the transitions among sub-states and the conformational fluctuations in the different domains of ADK. In this research, we use the SMOG model (a structure based model for biomolecular dynamics) to perform long time scale molecular dynamics simulations. In these simulations, the MD parameters are chosen such that ADK transits from an initial open conformation to a closed state. Interestingly, for some choices of force parameters a circulation among the intermediate partially open states is observed. The structural differences among the average structures of the intermediate sub-states and local structural changes in different sub-states are also characterized to help understand the molecular mechanism of catalysis. Besides the wild type ADK, we also use the mutant form MT G85V as a reference structure in MD
simulations. The MT ADK exhibits an unstable local structure in the CORE domain that influences the structural dynamics of the sub-domains.
Chapter 1

Introduction

1.1 Enzyme structure and dynamics provide support for protein function

Adenylate kinases (ADK) are ubiquitous, monomeric phosphotransferase enzymes. They maintain energy homeostasis in essential cellular systems by catalyzing the reversible Mg$^{2+}$-dependent phosphoryl transfer reaction Mg$^{2+}$+2ADP ↔Mg$^{2+}$+ATP + AMP. ATP, as the energy currency of the cell, releases high amount of energy. This reaction involves the breakage of the high energy phosphate bond of ATP between the gamma and beta phosphates resulting in transfer of a phosphoryl group from the ATP donor molecule to a lower energy substrate AMP. Without ADK, this phosphoryl transfer between the nucleotides takes around 8000 years to complete [13]. ADK helps to accelerate the phosphorylation in a timely manner by reducing the energy barrier of the reaction. This shortens the time-scale of the reaction to milliseconds. The huge acceleration on the ubiquitous phosphoryl transfer reaction is important in biological processes and prevalent in many species. The reaction regulates energy homeostasis in cells and mediates the balance of the concentrations of AMP, ADP and ATP. Large changes in cellular equilibrium of AMP/ATP will influence those metabolic sensors with high fidelity and sensitivity to stress signals [3]. Depending on the function of the isoforms of ADK, the reaction can activate or deactivate the downstream enzymes by releasing a large amount of chemical energy.
Adenylate kinases are small globular proteins. They are around 20 to 26 kDa. Structurally, ADK consists of three domains: the NMP-binding domain, the LID domain, and the CORE domain as seen in Fig. 1. The NMP and LID domains undergo dynamic motions during function. The NMP-binding site is on the small NMP domain of ADK. As a phosphate receptor site, the NMP domain is not as selective as the phosphate donor site. In this study, the NMP domain of *E. coli* ADK is defined as including residues 30 to 59. The motion of the NMP domain during function is executed by two hinges linking the CORE and NMP domains. These hinges are formed by residues 59-60 and residues 29–30. [19]
The LID domain is a lid-shaped ATP binding domain. During catalysis, the LID domain closes over ATP at the active center and increases catalytic efficiency (Muller et al., 1996). The flexible LID domain of ADK isoforms can be either of a short type or long type. In our research, E. coli ADK is used as the reference structure which is a long type. Its LID domain extends from residue 122-159.

The large central CORE domain is important for overall stability, and also contains the catalytic residues critical to phosphoryl-transfer. The core domain consists of three regions, 1-29, 60-121, 160-214. It has five parallel beta sheets, which are surrounded by four alpha-helices. One of the most important components of CORE domain is the P-loop.

A P-loop is a universal characteristic of adenylate kinases, and is also called as Walker A sequence as originally identified in the beta-subunit of F1 -ATPase (Walker et al., 1982; Saraste et al., 1990). This short, highly conserved polypeptide is shown to exist in a variety of nucleotide binding and hydrolysis enzymes, and is shared by 13 super families. The P-loop is also frequently used as a diagnostic tool for identifying new nucleotide dependent proteins. In enzymes, this loop connects a beta-sheet to an alpha helix, with some exceptions. The highly conserved G-X-P-G-X-G-K-G-T sequence, forms hydrogen bonds between its backbone atoms and a phosphate group. [14] The P-loop fragment starts from a glycine at its N-terminus. In E. coli ADK, the sequence is G-A-P-G-A-G-K-G-T, and spans from residue 7 to 15. Its position in E. coli ADK is shown in the rendition of the crystal structure of the closed form, 1AKE, in Fig. 2. Gly7 forms the hydrogen bonds with Lys13. This hydrogen bond not only stabilizes the structure of the P-loop, but also confirms the structural importance of Lys13. Ala8 also forms hydrogen bonds with Lys13. This hydrogen bond acts as a bridge between the ATP phosphate chain and Ala8. The Gly7, Ala8 and Lys13 define the two ends of the P-loop. The two hydrogen bonds help to maintain the integrity of the P-loop. Pro9 and Gly10 both have a dramatic influence on the K_m values of the reaction.[17] Gly10 forms the only interaction between the gamma phosphate of ATP and the P-loop in the crystal structure (1AKE). Gly12 interacts with the alpha phosphate of ATP through a hydrogen bond. Lys 13 is a very important residue in P-loop. It forms a bridge between the AMP and ATP phosphates, and stabilizes the active site.[18] In most MD simulations, a permanent hydrogen bond is formed between Ala8 and
Lys13 [20]. So Lys13 also acts as a bridge between the ATP phosphate chain and Ala8. Gly14 and Thr15 both form stable hydrogen bonds with phosphates of ATP in MD simulation.[20] Thr15 plays a role in catalysis through direct interaction with ATP. Both Lys13 and Thr15 are very important to the binding affinity of ATP to the P-loop and ADK. [16][17]

Figure 2. The P-loop (orange) within the crystal structure of *E. coli* ADK 1AKE in a closed conformation. The CORE domain (green), lid domain (yellow) and NMP domain (blue) are indicated by colors. The red molecule is the AP5A inhibitor. The orange loop below the inhibitor is the P-loop. Lys13 is on the P-loop.

The giant anion hole is another important structural feature at the active site of ADK. It is rendered in the crystal structure in Fig. 3. Parts of the CORE and LID domains come together to form the giant anion hole. It acts as a catalytic cleft with the P-loop buried deeply within it. In comparison with the ribose and adenine parts of ATP, the triphosphate
moiety has extensive contacts with active site residues of ADK. Several arginines and the P
loop (glycine-rich) sit on the LID and CORE domains to form the highly charged pocket
which makes contacts with the phosphates of ATP. This is called the giant anion hole.
Arg36, Arg88, Arg123, Arg156, Arg167, and Lys13 sit in the hole, all interacting with the
P-loop backbone in a position to mediate the binding, orientation, and transfer of phosphate.
Arg36 and Arg88 both interact with the alpha phosphate of AMP. [21] Arg88 has a large
influence on the Kₘ values for both ATP and AMP. It apparently stabilizes the transferable
gamma phosphate group during its transition from ATP to AMP. [21] Arg156 interacts
weakly with alpha phosphate of AMP. It forms hydrogen bonds with gamma phosphate of
ATP and stabilizes the transferred phosphoryl group. Arg156 is very important, since it is
one of the two arginine residues which work as a “trigger” rearranging on ATP binding
(Muller-Dieckmann & Schulz, 1995). Arg123 is on the LID domain and Arg167 is on the
CORE domain. These two residues also help stabilize the transferred phosphoryl group.
The Arg123 interacts with beta phosphate and forms a further contact with the alpha
phosphate.[21]
Figure 3. The giant anion hole within the crystal structure of *E. coli* ADK 1AKE in a closed conformation. The red molecule is the AP5A inhibitor. The CORE domain (green), LID domain (yellow) and NMP domain (blue) are indicated by color. Arg36, Arg88, Arg123, Arg156, Arg167, and Lys13 are positioned around the giant anion hole.

### 1.2 AdK as a model system for application of MD

ADK is responsible for functions in immune response, signal transduction and enzymatic activity. Typically, these kinds of functions are carried out by the interactions between proteins and other proteins or between proteins and ligands. These interactions often involve allosteric regulation, which usually requires large-scale conformational alterations. To understand the mechanism of these large scale conformational transitions is an important problem. Currently, there are two models for catalytic function. One is the
In the absence of substrates, ADK usually adopts an open conformation in solution. After ATP and AMP are aligned to the proper positions in the active site of ADK, large conformational transitions are induced. The LID and NMP domains close over the CORE domain, trapping the two reactants and forming a closed conformation as seen in Fig. 3. During the whole reaction, the CORE domain remains nearly rigid. The large scale structural rearrangements are confined to change in the positions of the NMP and LID domains as seen in the comparison of open and closed conformations in Fig. 4. A double set of hinges connects the LID domain to the CORE domain. Due to the flexible nature of the LID domain, little elastic energy is required to move the LID domain from an open position to a closed position. In the reaction, the closing of the LID region always precedes the closing of the NMP domain. [19]

Figure 4. The LID and NMP domain undergo large conformational transitions in this phosphotransferase reaction as can be seen by comparing the crystal structure of the open form (4AKE - left) and closed form (1AKE - right).
1.3 SAXS/WAXS as an experimental tool for the study of structure and dynamics

The internal motion and conformational changes of a protein play an important role in protein function. X-ray crystallography provides essential information on protein structure. However, the crystal contacts restrict protein motions and the conformation observed represents but a single snapshot of the ensemble of conformations present when the protein is in solution. In order to reveal the natural dynamics of proteins in solution, solution scattering in the small-angle regime (SAXS) and wide-angle regime (WAXS) provides a source of information about protein dynamics and conformational changes difficult to obtain by other methods. The advantages of SAXS/WAXS include ease of sample preparation (crystallization is not needed), study of proteins in a dilute aqueous solution where intramolecular motions are not constrained by crystal contacts, and the fact that studies can be carried out on proteins during function (e.g. catalytic cycling) or under conditions where multiple conformations may be present.

Figure 5. Layout of beamline X9 at Brookhaven National Laboratory at NSLS. NSLS has been shut down and is no longer in operation [40].

Most of our experiments were done at the x-ray scattering synchrotron beamline-X9 at the National Synchrotron Light Source (NSLS), Brookhaven National Lab (BNL) [34] with other data collected at BioCAT at Argonne National Lab (ANL). The A collimated, monochromatic X-ray beam incident on the sample generates scattered X-rays, which are imaged by a detector. The solution scattering data captured on the 2D detector are processed for scattering geometry, absorption, and polarization and then circularly averaged. Subsequently, background is measured by scattering from buffer (without
protein present) and then used to correct the scattering from the protein solution by background subtraction.

X-ray solution scattering in both the small-angle (SAXS) and wide-angle (WAXS) regimes is helping us to understand the relationship between protein function, rearrangement and dynamics. SAXS data is limited to spacings of ~ 15–20 Å resolution. SAXS is more sensitive to large conformational changes. It can provide information on transient protein-protein interactions and is very sensitive for detecting aggregation. A pair distribution function can be calculated from SAXS data to determine oligomeric form and characterize structural rearrangements during protein function or ligand binding or changes in environment [42].

WAXS extends data present in the SAXS regime out to scattering angles comparable to those used in crystallography. It is sensitive to small conformational changes in proteins; useful for calculation of the pair-distribution function at relatively high resolution; provides a means to characterize the breadth of the structural ensemble in solution; and can be used to identify proteins with similar folds. [7]

Compared to the X-ray scattering of proteins in a crystal, SAXS/WAXS patterns, which can be produced from protein solution sample, contain the structural rearrangement information of enzyme in the real reaction environment. SAXS is sensitive to changes in the shape, size and structure of the protein. Without the constraint of crystal contacts, the structural dynamics of protein in solution can be recorded by X-ray scattering pattern. With the help of signal processing, “fingerprints” of the protein structure can be extracted from the scattering pattern. It offers the possibility of determining the putative models of conformational states and intermediate transitions of the protein reactions.[9] X-ray solution scattering can provide a great deal of structural information on biomolecules in solution and is compatible with a wide range of experimental conditions. [6][7]
Chapter 2

Methods – Computational

2.1 Technical details

The known crystallographic or other structures of nucleic acids, proteins and lipids can be obtained from Protein Data Bank. The PDB files include the full atomic coordinates of those biological molecules and explicit water and ions and are used in our molecular dynamics simulation. The force between atoms in MD is described by empirical energy functions defined within the MD software packages used. Electrostatic terms, approximations of covalent interactions and long-range Lennard-Jones are all considered in these functions. The motion of atoms in MD is directed by Newtonian equations using for this work, a time-step of 2 femtoseconds (fs). Pressure and temperature can be controlled during the simulation by editing the appropriate equations.

In study of the enzymatic activity of ADK, the goal is to capture the large-length scale and functionally important slow domain motions. To do this, we use coarse-grained and/or simplified Hamiltonians in MD simulations. Simplified models and structure-based Hamiltonians are required to access the very slow motions with an accessible amount of computer time. They are developed from the energy landscape theory of protein folding. Those models are often used to study Biomolecular dynamics [25].

Since the domain motions of ADK take place on the scale of microseconds, conventional, all-atom MD would require a scale of computer resources not readily available. As an alternative, we chose to use coarse-grained simulation models supported by the programs GROMACS and SMOG.
2.2 Implementation of molecular dynamics (MD) program for ADK

2.2.1 GROMACS

GROMACS is short for GROningen MAchine for Chemical Simulation [22]. It is a versatile and efficient package that has broad capabilities for multiple modes of molecular dynamics. Motion of systems with hundreds to millions of particles can be simulated by Newtonian equations in GROMACS. GROMACS was originally designed for application to biological macromolecules, like proteins and lipids. GROMACS is extremely fast at calculating bonded and non-bonded interactions. Its application has been expanded and many researchers now use the program on non-biological systems, such as polymers. It can perform bio(macro) molecule simulations in both aqueous and membrane environments. This program can be implemented in both single processor and parallel message-passing computer systems. Beside microcanonical Hamiltonian mechanics, GROMACS also can carry out stochastic dynamics including Langevin and Brownian dynamics, and energy minimization. GROMACS does not have a force field of its own. It supports GROMOS, OPLS, ENCAD, and AMBER force fields. Flexible constraints and polarizable shell models can also be handled by GROMACS. We utilized the AMBER force field for the calculations described here. Users can add custom force routines to this program, and specify tabulated functions, which makes analyses of simulations easily customized. Free energy determinations and nonequilibrium dynamics are also incorporated in the program. There are various coupling methods in GROMACS which can be applied for use of temperature and pressure baths. Most important for the work described here, steered Molecular Dynamics can be run in GROMACS, which allows biological process is on time scales accessible to MD simulation. In steered MD, external forces are applied to induce non-equilibrium dynamics. Specific atoms can be selected and organized into groups, then, detailed analysis of energies and selective participation in the dynamics can be run on those
groups. GROMACS can process massless interactions by constructing virtual sites at atom positions. A large panel of analysis tools are included in GROMACS, such as normal mode and principal component analysis of structural fluctuations, and extensive graphical trajectory analyses.[22] Utilization of these tools are specified below and in the 'Results' section.[22]

2.2.2 NAMD

NAMD is an object oriented, parallel MD simulation program designed for high performance simulation of large biomolecular systems [21]. It has prioritized message-driven execution capabilities that enable parallel scaling on either workstation clusters or massively parallel supercomputers. NAMD is designed to work with VMD (see below) for interactive problem solving in structural biology. Connecting the simulation and visualization programs makes it possible to apply forces to a set of atoms to alter or rearrange parts of the molecular structure.

NAMD uses the same force field as the programs CHARMM and X-PLOR. Simulations produced by any of these three MD programs can be universally transferred one to the other. NAMD takes advantage of the Particle Mesh Ewald (PME) algorithm. This algorithm is efficient at taking the full electrostatic interaction into account. In NAMD's force field, the bonded interactions between 2, 3, and 4 atoms are included in the local interaction terms. Van der Waals forces and electrostatics are both included in the pairwise interactions. The electrostatic interactions beyond some specified threshold distance are defined as longer range interactions. Local interactions are calculated at each time step. A longer time interval is utilized for the longer range interactions. This is feasible because the short-range portion of the electrostatic interaction changes fast and the long range portions of the electrostatic interaction vary relatively slowly. A 'smooth splitting function' is used to separate the different length ranges of the electrostatic interaction. The multiple time step scheme substantially reduces the cost of the calculation of long-range electrostatic forces. NAMD also gains efficiency by use of the velocity Verlet integration method to help to calculate the positions and velocities of the atoms.
NAMD uses the same input and output file formats as used by X-PLOR and CHARMM, making interchangeability much easier. In particular, the structure files are in X-PLOR PSF format, coordinate files in PDB format, and energy parameter files in either X-PLOR or CHARMM formats. After MD simulation, binary DCD trajectory files and PDB coordinate files are produced as output files. The molecular dynamics trajectories produced from CHARMM, X-PLOR, or NAMD can be analyzed by the program package supplied by any of the three.

A lot of simulation options are offered by NAMD. For instance, the user can choose to carry out MD at constant energy, constant temperature, or constant pressure. Periodic boundary conditions can be defined in NAMD. In NAMD, user can set up energy minimization, fixed atoms, rigid waters, rigid bonds to hydrogen, harmonic restraints, spherical or cylindrical boundary restraints. Those options help the user to manipulate MD simulation to achieve different research goals.

NAMD uses a modular design and is written in the object-oriented style with C++. Since molecular dynamics technology is developing fast, this design allows new algorithms to be integrated and tested in NAMD easily.

2.2.3 VMD

VMD is the visualization counterpart to NAMD. It is also written as an object oriented program in C++ and its source code, extensive documentation and program code is available. Multiple molecular structures can be displayed on VMD simultaneously in various coloring methods and rendering styles. Molecules can be displayed in several different “representations”. Each representation is displayed in certain coloring scheme and rendering method for a selected group of atoms. The specific representation can be assigned to atoms by an extensive atom selection syntax. VMD also includes a number of photo quality image-rendering applications to produce raster images of displayed molecules, a graphical user interface and a text interface to control the program. The MD simulation trajectories generated in NAMD can be animated by the tools offered by VMD.
2.2.4 SMOG

SMOG uses the information in a structure file, along with additional parameters, to produce a topological file which is subsequently used by GROMACS to produce an MD simulation trajectory.[22] Variations can be designed and implemented into it for specific needs of research. In the structure based models, a known conformation is defined as a potential energy minimum. Based on this requirement, there are a lot of ways to construct a structure based model. RNA-specific and protein-specific variants can be implemented. Various parameters defining resolution and multiple minima can also be set. The variations used to construct the structure based models can be edited to answer specific questions about biomolecular processes.

In the simplest form of structure-based model, a single conformation is defined as the global potential energy minimum. All inter- and intra- molecular interactions intrinsic to this conformation are assigned as minima. However, in the energy landscapes of biomolecules, there is often some degree of energetic roughness. Both native and nonnative interactions must be included in the structure-based models to make it possible to cross over energy barriers. In the SMOG v2 software package, two energetically unfrustrated models can be generated and additional interactions can be added into them.

In this project, we created structure-based models for the open and closed conformations of ADK (1AKE, 4AKE) using the online SMOG server. After submitting the PDB file of ADK, a well-characterized structure-based model was created which is ready to be implemented on GROMACS.

Cellular enzyme catalysis often occurs at the micro to millisecond time scale. We want to study the conformational transitions of ADK in the catalysis processes. ADK has 214 residues. The traditional all-atom molecular modeling technique is difficult to investigate its large scale conformational motion. So we use the coarse graining (CG) model of SMOG to study the dynamics of long time scales and large systems. CG model can reduce the computational complexity and the number of degrees of freedom in MD simulation to a level of computer resources currently available.
The open (4AKE) conformation is used in our research as a local energy minimum. In order to construct trajectories that navigate the conformational change from open to closed in a time frame feasible for MD we chose MD parameters that defines the open (4AKE) and closed (1AKE) conformation as the energetic basins. The coarse-grained Cα structure-based model of ADK created on SMOG server reduces the energetic frustration and energetic heterogeneity among residue-residue native interactions. In the Cα model of SMOG, each residue of ADK is coarse-grained as single bead of unit mass located at the position of the residue’s Cα carbon and the charge of each atom is set to zero. The harmonic interactions between the Cα beads is calculated by the bond lengths, bond angles, dihedral angles of the neighboring Cα atom which are calculated from the coordinate values found in 1AKE and 4AKE pdb files. Non- neighboring Cα atom pairs that are in contact in the provided structure between residues i and j, are given an attractive 10-12 potential.

The force field of SMOG Cα model is given by the following equation. The parameter \( \mathbf{x}_0 \) represents the coordinates of the starting point. In this study we use 4AKE as the starting point. Any subscript 0 in the algorithm is a value taken from the starting coordinates. Epsilon is chosen according to the desired depth of the attractive well. In the equation, we set \( \epsilon_r = 100\epsilon \), \( \epsilon_\theta = 20\epsilon \), \( \epsilon_D = \epsilon_C = \epsilon_{NC} = \epsilon \).

\[
V_{C\alpha}(\mathbf{x}, \mathbf{x}_0) = \sum_{\text{bonds}} \epsilon_r (r - r_0)^2 + \sum_{\text{angles}} \epsilon_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} \epsilon_D F_D(\phi - \phi_0) + \sum_{\text{contacts}} \epsilon_C \left[ 5 \left( \frac{\sigma_{ij}}{r} \right)^{12} - 6 \left( \frac{\sigma_{ij}}{r} \right)^{10} \right] + \sum_{\text{non-contacts}} \epsilon_{NC} \left( \frac{\sigma_{NC}}{r_{ij}} \right)^{12} \quad (A.3)
\]

The dihedral potential \( F_D \) is defined by:

\[
F_D(\phi) = [1 - \cos(\phi)] + \frac{1}{2} [1 - \cos(3\phi)].
\]

The parameters \( \mathbf{x} \) describe the Cα coordinates of the trajectory ensemble. The bond lengths of the nearest neighbors are represented by \( r \), three body angles are represented by \( \Theta \), four body dihedrals are represented by \( \phi \), the distance between Cα atoms \( i \) and \( j \) is represented
by $r_{ij}$. Every Cα atom has excluded volume which has an apparent radius of $\sigma_{NC} = 4 \text{ Å}$. The radius is maintained by a hard wall interaction. The native state $x_0$ is defined by the set of atomic coordinates and by a set of parameters which provides the native bias to the simulation.[8][25] The MD simulation starts from the structure-based potential of the open conformation (4AKE), then, based on this SMOG force field, the gradual perturbations which are based on the closed conformation (1AKE) are added to the simulation. These perturbations drive the transition of the structural ensemble through a minimally frustrated energy landscape on the basis of the energy wells defined by the start and end points of the transition (open and closed conformation). As we will see, the trajectory resides for a variable length of time in an intermediate, partially closed ensemble of structures that contains significant local organization. This residence time is a function of the energetic roughness of the energy landscape that is typical of biomacromolecules. Generating these trajectories, both native and nonnative interactions are required to be included in the structure-based models in order to drive a cross-over of the energy barrier. (Whitford et al., 2007). The analysis of clusters of conformations in the partially closed ensemble of structures provides some insights into the degree of roughness of the landscape.

### 2.2.5 Biopython Project

The biopython project is a python toolbox which is used for computational biology and bioinformatics. In our custom built program, a lot of PDB file and protein structure need to be read and analyzed. This library package allowed us to rapidly localize the multilevel hierarchy of protein. When we use the biopython package, we add the name of biopython library package that we want to use on the beginning of our python program. Then, we can use the class and function of this biopython module directly later.

As a biopython module, Bio.PDB is used to access crystal structures of biological macromolecules. This module of biopython includes PDB and mmCIF parsers, a module to keep a local copy of the PDB up-to-date, a Structure class, selective IO of PDB files, and so on. In these programs, we often need to extract the coarse grain Cα structure from the trajectory iteratively. When we access protein structure, a structure object need to be
created to save the data from the PDB file. The get_structure function of PDBParser object returns the protein structure from the PDB file. The protein coordinates are stored in the Bio.PDB.Model object, which represents a model in a structure. There are three important function tools in this object: get_chains(self), get_residues(self), get_atoms(self). The first function is used to access the peptide chain which forms the protein. The second function is used to access all the residues in the protein. The third function is used to access all the atoms in the protein. The lower layer of the structure is the Bio.PDB.Chain object. As its name suggests, this structure object stores one peptide chain. The next layer of the structure is the Bio.PDB.Residue object. The get_resname() function of the Residue object returns the residue name, which is very useful for searching certain type of amino acid. The next layer of the structure is the Bio.PDB.Atom object. The get_coord() function returns the atomic coordinate of this atom. The get_name() function returns the atom’s name.

The Bio.PDB object also can be used to navigate through a Structure object. Superimposer object of Bio.PDB is used to superimpose two structures. The set_atoms function of Superimposer inputs one set of reference atoms, and one set of moving atoms. The set of moving atoms can be translated and rotated to match the set of fixed atoms by minimizing the RMSD. The minimized RMSD between the moving and fixed atoms is also saved at variable rms of superimposer object.

2.3 Custom Built Software

2.3.1 K-means clustering (program 1)

To analyze the huge amount of data produced by these MD simulations, we used K-means clustering to classify the structures from MD trajectories. Each data point is a coordinate of ADK Ca structure in PDB format. The Ca structure can be viewed as a 214*3 matrix. The high dimensional data will be classified by similarity between structures into 5 clusters.
In the K means program, a data point \( x \) belongs to cluster \( k \) if it is closer to centroid \( k \) than any other centroids. The centroid \( k \) is called as the dominating centroid of \( x \). It is a 214*3 matrix. The average of the coordinate sets of all the structures in one cluster is the centroid of this cluster. The goal is to discover any patterns within the data points from the trajectory. The k-means algorithm helps to minimize distortion of the data set. To reduce distortion, every round, every data point is re-classified to the closer dominating centroid. The choices of centroids are also refined in each step of the k-means algorithm to reduce distortion. The number of data point shifting to a new dominating centroid in one round is used as a stopping criterion. When the number of times that data points shift their dominating centroid dips below some chosen threshold, the k-means algorithm cannot make sufficient progress anymore and terminates. The maximum number of iterations 1000 is defined in the program, when the program loops over this number, it also terminates.

In this program, some external libraries are imported to facilitate and simplify the programing. Bio.PDB, scipy, fnmatch, os, os.path, numpy, glob, string, random, gc, are libraries used in the program. Scipy and numpy facilitate the mathematical computations. Bio.PDB belongs to the Biopython library. We use it to deal with the PDB file and protein structure. The glob module of python uses the rules of Unix shell to find all the pathnames which match the specific pattern. The fnmatch module compares filenames with the glob-style pattern. The os module offers a way to run the operating system dependent functionality. The os.path module offers some useful functions to manipulate pathnames. The string module contains useful string operations. The random module generates pseudo-random numbers on various distributions. The gc module is an interface of python’s optional garbage collector.

**Cluster:** Two classes are built for this program. One is class Cluster, the Cluster object has a list of points and a calculated centroid point, three functions. The def \_init\_(self, points): function initialize a Cluster object. In the function, a list of data points is stored at Cluster.points, the centroid of this cluster is calculated by function Cluster._calcCenter(), and then the centroid is stored at Cluster.cenpoint. The function def _calcCenter(self): calculates the centroid of this cluster. After adding the new data point, the coordinates of the list of data points are summed and averaged. The data point which is most close to the
center of the data points is the centroid of the cluster. The function `def update(self, points):` saves the new data point passed in by parameter and re-calculate the centroid of this cluster with the function `Cluster._calcCenter()`.

**Stru:** The other is class Stru, Stru object has coordinates (x,y,z,...) and an optional label. In the Stru object, four functions are built. First is `def __init__(self, pdb_structure, tstep, label=-1)`, this function is used to initialize a data point and variables in the k means algorithm. Three variables are included in this object, a Bio.PDB.structure object, int label, int tstep. This Bio.PDB.structure object stores the coordinates of a protein structure. The int label stores the dominating centroid number which the data point belongs to. The int tstep stores the time sequence number of this data point in the MD trajectory. The function `def dist(self, other):` of Stru object helps user to access the RMSD between itself and another data point indirectly. The function `def str_dist(self, a, b):` offers a tool to calculate the RMSD between two data points by their time sequence numbers. The function `def __str__(self):` helps user to access the coordinates and dominating centroid number of this data point.

The `def DomainCentermass(pdbStr,staAtom,endAtom):` this program is used to calculate the center of mass (COM) of a set of coordinates. The `pdbStr` is a list which stores the coordinate sets of protein structure. In our research, we use this function to calculate the COM of a continuous peptide segment in the protein. The `staAtom` and `endAtom` store two integers. The `staAtom` is the sequence number of the starting residue of the peptide segment. The `endAtom` is the sequence number of the ending residue of the peptide segment. These three variables are the input parameters of this function. The structure of a protein is defined on five levels, ‘structure’, ‘model’, ‘chain’, ‘residue’ and ‘atom’. From ‘structure’ to ‘model’, ‘model’ to ‘chain’, ‘chain’ to ‘residue’, ‘residue’ to ‘atom’, the information of every structure level of a protein is parsed by tools of Bio.PDB object. The sequence number of a residue in ‘chain’ can be extracted from the ‘residue’ object. Each sequence numbers of the residues in ADK is matched with the starting and ending residue number of this domain. The set of ‘atom’ objects which belong to the domain are cut out from the whole protein structure. The coordinates of the set of ‘atom’ objects are returned by the function Bio.PDB.Atom.get_coord() and stored into a separate list ‘coord’. The
residue names of each Cα atoms are returned by the function Bio.PDB.Residue.get_resname(). The abbreviation of the names of 30 amino acids and their corresponding weights are saved in a global dictionary variable MW_DICT in main function. With the residue names of each Cα atoms, the weights of each residue in the domain are extracted from the dictionary and saved in the list ‘masses’. The weights of each residue in the subdomain are stored in the list ‘masses’ and summed together by the function sum(masses). The result is saved in the variable ‘total_mass’. Then, the list ‘masses’ is divided by the ‘total_mass’ and saved in a list ‘weights’. The list ‘weights’ is summed together on each axis of Cartesian coordinates and saved in variable ‘center’. The variable ‘center’ is the COM of the set of coordinates and returned at the end of the function.

The def kmeans(clusterPoints, points, k, rmsMat,struMax, max_iter=1000, min_shift_frac=0.001): is one of the most important functions in program. The 5 starting points and the list of the structure of the ensembles are the input parameters for this function. 5 clusters are initialized with the 5 starting point by Cluster object. The 5 starting points are set as the centroids of the 5 starting clusters. According to the RMSD between the 5 centroids and the structures, each structure is classified into the closest centroid. A new centroid of each cluster will be recalculated at the end of each loop. The number of data points that shift their dominating centroid in each round is recorded, and divided by the numbers of data points. If the ratio is smaller than 0.001 or the shifting number is smaller than 3, the loop will break and return the 5 clusters. Alternatively, when the number of iterations reaches 1000, the loop also stops and returns the 5 clusters.

In the main function, the 4AKE from PDB is defined as the reference structure at the beginning of the program. For this calculation, the start and end point of CORE domain is defined between 60-121 since only a single contiguous peptide chain can be considered. The start and end point of LID domain is defined between 122-159. The start and end point of NMP domain is defined between 30-59. Then, the structures of the trajectory are traversed. The Cα structures are accessed by Bio.PDB.PDBParser.get_structure function and these coordinates are aligned to the reference structure. All those data points are initialized by Stru object and saved in the list trajStru_set with their time sequence number.
For analysis of the structures, we plot the result in a two dimensional coordinate system in which the distance between the COM of CORE and LID domains is defined as the X-axis, and the distance between the COM of CORE and NMP domains as the Y-axis. With the structure of each data point, the COM of CORE, LID and NMP domains are calculated by DomainCentermass(pdbStr,staAtom,endAtom) function and the coordinates saved in list CoreCenter, LipCenter, AmpCenter. The norm of the vector between COM and every atom of the domain is calculated by the function numpy.linalg.norm(). Before performing k-means clustering, initial starting points are necessary as input parameters. Sometimes, different results may be raised from different starting points. We know from the crystal structures of ADK that it interconverts between open and closed conformations. The conformational transitions we follow from the MD simulation are important for probing the conformational landscape. To divide the trajectory into 5 clusters, we need 5 starting points. One starting point is from open state and one is from closed state, and three from intermediate states. The list of 5 starting points and the list of the structure of each ensemble are passed in the kmeans function as parameters. 5 new clusters are created from the kmeans function, each of them contains a list of data points and its centroid. The time sequence number and cluster number of each data point are extracted from the result of kmeans function and output into the external file.

2.3.2 Radius of gyration (ROG) (Program 2)

We define five clusters with the five sample structures as the five centroids. According to the distance between structures of the trajectory, each structure is assigned to the cluster to which the closest centroid belongs. According to the previously defined start and end Cα sites of each domain, the ADK structure is separated into three parts, LID, NMP and CORE domains. In each structure of trajectory, the ROG can be calculated for each subdomain of ADK. The ROG is defined as the root mean square distance from each atom of the protein to their centroid. The ROG is the average squared distance of Cα atoms of the domain from center of mass. The ROG distribution plot of data points in each state is visualized.
In the main function, the structures of the trajectory are traversed by the time sequence. There are 61600 data points of the trajectory used in the calculation. The string variable ‘XSWorkFileN’ stores the prefixes of the PDB file of the frames in the trajectory. According to the number of the current frame, the file name of the frame is composed and restored in the variable ‘PDBFile’. Bio.PDB.PDBParser(PERMISSIVE=1) is a function in Biopython. It parses the PDB file. Bio.PDB in the toolbox can store the protein structure into the customized object. The pdbPar.get_structure(structure_id, PDBFile) can extract the PDB structure by its file name. The structure is saved in the object ‘pdb_structure’.

The cluster file which contain time sequence number and cluster number of structures in a trajectory is parsed by the readClusterFile(PDBFile) function. The paired timeline sequence numbers and cluster numbers of each frame are returned and stored in the list ‘SteClusSet’. One set of ADK Ca coordinates is saved in each frame of the trajectory. The Ca structures are accessed by Bio.PDB.PDBParser.get_structure function. ADK Ca structure is cut into three parts by the residue sequence of the Ca atoms. The start and end point of CORE domain is defined between 60-121. The start and end points of NMP domain are defined between 30-59. The start and end points of LID domain are defined between 122-159. The start and end point of each domain and Ca structure are input parameters for the PRadiusofGyration(pdbStr,staAtom,endAtom) function. The ROG of each domain is returned and classified by the paired timeline sequence numbers and cluster numbers which are stored in the list ‘SteClusSet’. The classified ROG information of each domain is stored in the list ‘clusterRG’. and output by function numpy.savetxt ()

def readClusterFile(PDBlist): this function is used to read the cluster file. The time sequence numbers of each data point and their corresponding cluster numbers are stored in this file. The name of the cluster file is the input parameter of this function. The data in the file is extracted out by the open (PDBlist, 'r') function and saved in the file pointer fp. The file data is traversed line by line with the function fp.readlines(). In each line, there is a pair of the time sequence numbers of the data point and the corresponding cluster numbers. The sets of information is saved into a list and returned.

The def DomainCentermass(pdbStr,staAtom,endAtom): is the same function which is used in program 1.
def PRadiusofGyration(pdbStr, staAtom, endAtom): this function is used to calculate the radius of gyration (ROG) of a set of coordinates. The ‘pdbStr’ is a list which stores the coordinate sets of protein structure. In our research, we use this function to calculate the ROG of ADK’s subdomains. The ‘staAtom’ and ‘endAtom’ store two integers. The ‘staAtom’ is the sequence number of the defined first residue of ADK’s one domain. The ‘endAtom’ is is the sequence number of the defined last residue of the domain. These three variables are the input parameters of this function. The Cα structures stored in ‘pdbStr’ are accessed by Bio.PDB.PDBParser.get_structure function which belongs to biopython toolbox. The structure of protein is defined as five levels, ‘structure’, ‘model’, ‘chain’, ‘residue’ and ‘atom’. From ‘structure’ to ‘model’, ‘model’ to ‘chain’, ‘chain’ to ‘residue’, ‘residue’ to ‘atom’, the information of every structure level of a protein is parsed by tools of Bio.PDB object. The sequence number of a residue in ‘chain’ can be extracted from the ‘residue’ object. Each sequence numbers of the residues in ADK is matched with the starting and ending residue number of this domain. The set of ‘atom’ objects which belong to the domain are cut out from the whole protein structure. The coordinates of the set of ‘atom’ objects are returned by the function Bio.PDB.Atom.get_coord() and stored into a separate list ‘coord’. The residue names of each Cα atoms are returned by the function Bio.PDB.Residue.get_resname(). The abbreviation of the names of 30 amino acids and their corresponding weights are saved in a global dictionary variable MW_DICT in main function. With the residue names of each Cα atoms, the weights of each residue in the domain are extracted from the dictionary and saved in the list ‘masses’. The weights of each residue in the subdomain are stored in the list ‘masses’ and summed together by the function sum(masses). The result is saved in the variable ‘total_mass’. The coordinate of COM of the set of residues are calculated by the function DomainCentermass(pdbStr, staAtom, endAtom) and saved in the list variable ‘CoM’. The square of distance between every Cα atom and the ‘CoM’ is timed with the list weights and summed together. The result is saved in float variable ‘r_sq’. The variable ‘r_sq’ is divided by ‘total_mass’ and the result saved in ‘r_sq’. The square root of ‘r_sq’ equals to the ROG of the subdomain and the result is returned.

2.3.3 Cα atom pseudodihedral angle fluctuations (program 3)
In the main function, the cluster file which contain time sequence number and cluster number of structures in trajectory is parsed by the readClusterFile(PDBFile) function. The time sequence numbers of each snapshot and its dominating centroid numbers are returned and stored in the list ‘clusters’. Along the MD simulation time sequence, the PDB files of each snapshot is parsed and stored by the Bio.PDB.PDBParser structure. The Cα structures in PDB files are accessed the Bio.PDB.PDBParser.get_structure function. The corresponding dominating centroid information of each snapshot is accessed from the list ‘clusters’ and paired with the Cα structure. In our research, 4 consecutive Cα atoms are defined as a pseudodihedral angle. The Cα structure of ADK is formed by 214 Cα atoms. In each ADK Cα structure, 211 pseudodihedral angles are defined along the peptide chain of ADK. The pseudodihedral angles are calculated by the Bio.PDB.calc_dihedral function and stored in the list ‘clusterDiTrSet’. The Bio.PDB.calc_dihedral function is a Biopython tool to calculate the dihedral angle between 4 vectors representing 4 connected points. All the pseudodihedral angles of 61601 Cα structures of the trajectory 1.27 are stored in the list ‘clusterDiTrSet’ by time sequence. The standard deviation of the angular change of the same pseudodihedral angle in different state is computed along the MD simulation time axis. The calculation is done by the numpy.std function. This numpy function returns the standard deviation along the specified axis of an array. The variation of the pseudodihedral angle reflects the fluctuation of the local structure.

2.3.4 Root mean square fluctuation (RMSF) between each Cα atom pair (Program 4)

The RMSF is defined as the root mean square fluctuation between every pair of Cα atom in the trajectory. The RMSF is the root mean squared deviation of all the Cα atom pair distances in the trajectory. The RMSF distribution plot of data points in each state is visualized.

The RMSF of one residue pair in the trajectory can be calculated by the following equation. The parameter \( v_i \) is the coordinates of the residue \( v \) at time \( i \). The parameter \( w_i \) is the coordinate set of the residue \( w \) at time \( i \). In the equation, the distance between residue
pair \( v \)-\( w \) at each point in the trajectory is calculated. Any subscript 0 in the algorithm is a value taken from the starting coordinates. The epsilon is based on depth of the attractive well, it is equal to energy scale \( k_B \) (the reduced units \( k_B = 0.00831451 \)). \( \varepsilon_r=100\varepsilon, \varepsilon_0=20\varepsilon, \varepsilon_D=\varepsilon_C=\varepsilon_{NC}=\varepsilon, \)

\[
\text{RMSD}(v, w) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} ||v_i - w_i||^2}
\]

In the main function, the structures of the trajectory are traversed by the time sequence. There are 61600 data points of the trajectory used in the calculation. The string variable ‘XSWorkFileN’ stores the prefix of the PDB file of the frames in the trajectory. The cluster file which contains time sequence number and cluster number of structures in trajectory is parsed by the readClusterFile(PDBFile) function. The paired timeline sequence numbers and cluster numbers of each frame are returned and stored in the list ‘SteClusSet’. The stores the prefix of the PDB file of the frames in the trajectory and the cluster numbers of each frame are the input parameters for the TsqCreatCAdevMax(XSWorkFileN, SteClusSet) function. A 214*214 matrix is returned from this function, which is the RMSF of this trajectory. The result is output by function numpy.savetxt().

readClusterFile(PDBFile) is same as in program 1.

TsqCreatCAdevMax(XSWorkFileN, SteClusSet) This function takes the cluster information of each frame file and the coordinates of the structures from the trajectory as input. The PDB file name of each frame is composed by their number of the current frame with the prefix. The file name is stored in the string variable ‘PDBFile’.

Bio.PDB.PDBParser(PERMISSIVE=1) is a function of Biopython toolbox. It parses the PDB file. Bio.PDB in the toolbox can store the protein structure into the customized object. The pdbPar.get_structure(structure_id, PDBFile) can extract the PDB structure by its file name. The structure is saved in the object ‘pdb_structure’. From the list object ‘pdb_structure’, the set of coordinates of ADK can be extracted as list by list(pdb_structure.get_atoms()). The set of coordinates is saved in ‘StruCoor’. We built a
214*214 matrix, ‘CAdisMax’. As we know, the ADK has 214 Cα atoms. So each time, we need to go through the 214 atoms and record their coordinates. Under this loop, we built another loop. When we go through each Cα atom, we calculate the distances between first Cα atom i and all the other 214 atoms. In each step, the coordinate of Cα atom i is saved in list ‘coord[0]’. The other Cα atom was saved in list ‘coord[1]’. The distance is calculated by the function ‘distance(coord[0],coord[1])’. In the distance() function, the Cartesian coordinate distance of a pair coordinates are calculated. Then we travel to the next Cα atom, (i+1). Now we need to calculate the distance between the new Cα atom i and the other (214-i) atoms. We do not want to calculate the repeating Cα atom pair. Finally, the matrix ‘CAdisMax’ will look like a triangle. Each new matrix ‘CAdisMax’ is saved in the list ‘CAdevMax’. After we go through the PDB files of the trajectory, the length of list ‘CAdevMax’ will be 61600 frames.

2.4 Comparison to SAXS/WAXS experimental results

2.4.1 Implementation of XS

The internal motion and conformational changes of protein play important roles in their function. The X-ray crystallography provides the essential information of the protein structure. However, the crystal contacts restrict the protein motions. To reveal the natural dynamics of protein in solution, solution scattering helps to study protein dynamics and conformational changes. Compared with X-ray crystallography, crystallization is not needed in preparation for solution scattering. In the X-ray solution scattering, protein is free to tumble when X-rays are passing through the sample, therefore the solution scattering pattern can only reveal the spherically averaged scattering intensity which is not enough for reconstructing the 3D structure of the protein beyond about 20 Å resolution. However, the solution scattering intensity still contains a lot of information for studying of protein motion and conformation. The theoretical modeling is an important assistance for extracting more information from the experimental data. The theoretical modeling provides a basis for verifying our inference of protein motion. The program XS offers a tool to construct the solution scattering pattern of a protein from its atomic coordinates [25].
CRYSOL is a popular program for calculating solution scattering pattern, continuum models of water is used in it for constructing the solvent in sample [27]. This continuum water model brings significant discrepancies at the 10 Å resolution and beyond, so the calculated protein solution scattering pattern is undependable when it is extended to the resolution of WAXS. The WAXS of solution scattering pattern contains a lot information of protein folds. The XS program uses relatively short MD simulation to build up an atomistic description for water. The solution scattering patterns calculated by XS has high accuracy in both the WAXS and SAXS regimes. The following equation describes how the solution scattering intensity is calculated from the protein coordinates. [25]

\[
\Delta I_{\text{scaled}}(q) = I_A(q) - (1 - f)I_B(q),
\]

\[
I_A(q) = \langle |\tilde{A}(q)|^2 \rangle,
\]

\[
\tilde{A}_1(q) = \sum_l e^{-iqr_l} f_l(q),
\]

The \( \Delta I_{\text{scaled}} \) is the scattering intensity from the structure of protein in the solvent. The f is the estimated uncertainty that exists in the experiment. The \( I_A \) is the scattering intensity of system A. System A is the protein and with water shell wrapping around it. System B is pure water. Since the protein is tumbling in the solvent, \( \langle \ldots \rangle \) averages the degrees of freedom of protein and solvent in rotation, translation and internal motion. In each MD simulation, 50 snapshots are taken for the protein with water shell, 50 for pure water box. The instantaneous electron density of each snapshot can be calculated by the above equations. The \( r_l \) represents the coordinates of lth atom in the system. The thickness of the water shell is defined by \( \xi \). In our search, we set \( \xi \) as 7 Å which is adequate for most proteins.[26] The \( f_l(q) \) is the atomic form factor of the lth atom. The 'q' is the scattering wave vector.

The program XS is written in python and separated into two files, ‘runxs’ and ‘xs’. The ‘runxs’ contains the main function. The ‘xs’ contains the function which are used in main function. There are total eleven functions in the ‘runxs’ library: defgetConfig
(configFile); def spiral (N): def readpdb (pdbFile); def formFactor (q, atype); def fourier (q, spherePoints, pdbFile); def cutProteinWaterShell (thickness, psfFile, dcdFile); def cutWaterBlob (thickness, psfFile, dcdFile, proteinPDBfile, blobDir); def build (pdbFile, psfFile); def minimize (pdbFile, psfFile, prmFile, boxSize, minSteps, n_proc); def equilibrate (pdbFile, psfFile, prmFile, xscFile, runSteps, trajSteps, n_proc); def buildWater (boxSize).

The function getConfig reads user’s parameter file ‘xs.cfg’ and store the information into the variable config. In the ‘xs.cfg’ file, user can define minimum and maximum of the scattering wave vector q. We define them as 0.01 and 2.5 Å⁻¹ in this work. The minimal increment of q is defined as 0.01. The ‘n_sphere_points’ define the number of points used for spherical quadrature on the unit sphere. In this research, we define it as 1000. The ‘prm_file’, ‘pdb_file’ and ‘psf_file’ are the parameters for the path information of MD parameter file, PDB file and PSF file. The ‘work_directory’ stores the work directory. The ‘temperature’ is the environment temperature in the MD simulation. We use 277 K which is appropriate for most proteins. The thickness of water shell is define by ‘shell_thickness’. The following several parameters helps user to direct the time length of different stages in MD simulation. The ‘n_min_steps’ defines the number of steps in the minimization stage. The ‘n_pre-equilibration_steps’ defines the number of steps in the pre-equilibration stage. The ‘n_production_run_steps’ defines the number of steps in the production stage. The ‘n_steps_bet_snapshots’ defines the time interval between the snapshots been taken in the production stage. The ‘n_processors’ sets the number of computer processors for simulation.

The rotational average is evaluated in the function spiral (). In our search, this function places 1000 points on the unit sphere along a spiral. Spherical quadrature will be used in the Fourier equation.

The function readpdb () takes the name of PDB file as parameter. The coordinates of protein structure and atom type are read from PDB file and stored.
In the function formFactor(), the atomic form factor of each independent atom is calculated along the scattering wave vector q. The electron charge density distribution of the atom in reciprocal space is obtained by Fourier transform of the function.

In the function fourier(), the coordinates of the structures from the MD simulation and the form factor of each independent atom are used. The scattering intensity from the structure is obtained by summation of the Fourier transform of the form factor of each independent atom.

The function cutProteinWaterShell() contains a VMD script. It directs the VMD program to cut out a thickness 7Å water shell around protein in each protein and solvent snapshot. In our research, 50 water shells are created for each MD trajectory. The function cutWaterBlob() contains a VMD script. It directs the VMD program to cut out a protein shaped water blob from each pure water box snapshot. In our research, 50 water blobs are created from each MD trajectory. The function build() contains a VMD script. It constructs a PSF file from the PDB file of the protein. The solvent and protein are both put into the simulation box. The information of the simulation box is saved in the boxinfo.txt. The function minimize() contains a NAMD script. It directs the NAMD to run the minimization stage of the simulation. The function equilibrate() contains a NAMD script. It directs the NAMD to run the equilibration and production stage of the simulation. The function buildWater() contains a VMD script. The pure water box is prepared and created in the function.

In the main function, the MD simulation is run in the first step. With the PDB file of the protein, the simulation box which contains protein and solvent is created. The coordinate and topological information of the simulation box are saved into the new PDB and PSF files. After preparation, the simulation is separated into 2 parts: first, the MD simulation for the protein and water shell: simulation for the water which wraps around the fixed protein; second, the simulation for the pure water box. Each step follows the same process: minimization, pre-equilibration, production. Each step in the MD simulation is 2 fs. The energy minimization stage is 100 steps. In this stage, the potential energy of the system is minimized to as low as possible. The coordinates are written to trajectory file every 1000 steps and the energies are written every 100 steps. The pre-equilibration stage
is 5000 steps. In this stage, the protein is fixed and the solvent in the box undergoes a relaxation. The coordinates are written to trajectory file every 1000 steps and the energies are written every 500 steps. The production stage is 50000 steps. In this stage, the protein stays fixed. The trajectory produced in this stage will be used in scattering intensity calculation. The coordinates are written to trajectory file every 1000 steps and the energies are written every 500 steps. The box size used in the simulation is depended on the size of protein. After the simulation, 7 angstrom thick water layer is cut out around the protein.

In the next step, the 100 snapshots (50 for protein and solvent, 50 for pure water box) are produced from trajectory. The function cutProteinWaterShell() and cutWaterBlob() trim the structure of protein and water shell, and the structure of water blob. Depending on the atomic coordinates of the solvent and protein, the scattering intensity is calculated by the function fourier(). The solution scattering intensity distribution from the experiment is isotropic. The only magnitude of the distribution is the scattering wave vector. So in the final processing section of the program, the intensity is averaged over the solvent and protein degrees of freedom. After the scattering intensity of protein and water blob are obtained separately, the intensity of protein and water shell is subtracted by the intensity of water blob. This is the background subtraction scheme we used.

2.4.2 SAXS for determination of conformational equilibria

In parallel with the work described here, Dr. Emre Onuk carried out a series of studies to develop a new 'maximum a posteriori' estimation approach to estimating relative abundances of protein conformations from SAXS data (Onuk et al., 2014) and applied it to ADK data (Onuk, Badger, Wang, et al., 2016). Much of the computational work described in this thesis contributed to that work, particularly the re-engineering and use of the program XS, and the utilization of K-means clustering for analyses of conformational ensembles. Therefore, Dr. Onuk's work is described in brief here.

In this phosphotransferase reaction, AdK has two stable end states, the open conformation (4AKE) and the closed conformation (1AKE). In previous research, the open to closed transition of this reaction was studied by a dynamic importance sampling (DIMS) molecular dynamics method. [39] An ensemble of 330 DIMS trajectories was used to
simulate the progression of the transition. The 45 *E. coli* ADK protein structures were derived from crystallography (either from *E. coli* ADK submissions to the protein data bank or homology models derived from ADK structures from other species) and used to characterize the progression of the conformational change that occurs on ligand binding. The progression of the transition was measured by the ‘ΔRMSD’ measure which is the difference in RMSD of the current conformation from the reference open structure (4AKE) and closed structures (1AKE).

In solution, ADK and its ligand undergoes the open ↔ closed transition. A lot of intermediate conformations of ADK that may lie on or near the transition pathway occur in solution at the same time. To characterize this ensemble of conformational states of ADK in solution, x-ray solution scattering data were collected from ADK in apo form and titrated with AMP, ADP, ATP, a non-functional ADP analog and the tightly binding ligand diadenosine pentaphosphate (AP5A). Apo ADK scattering data are quite well predicted by XS using the crystal structure of *E. coli* ADK in the open conformation whereas, scattering data from ADK samples saturated with AP5A are in poor agreement with calculated scattering from crystal structures of ADK in the closed conformation.

In this research, we used the XS program to calculate the scattering intensity of the 45 ADK crystal structures reported by Beckstein [39]. These scattering patterns were clustered using a subset selection method based on k-means clustering and the Cramer-Rao bound, resulting in a sparse basis set of conformations. According to the result of the analysis, five clusters are defined and five crystal structure are chosen as the representatives of each cluster. The five crystal structures of AdK include an apo form (PDB ID: 4AKE_A), a structure with the LID domain in a partially closed conformation (PDB ID: 2RH5_C), a structure with the LID domain closed and containing ATF (phosphodifluoromethylphosphonic acid-adenylate ester - a non-hydrolyzable ATP analog) under the LID domain (PDB ID: 1DVR_B), a homology model based on the structure of maize ADK in an almost closed conformation and containing the inhibitor, Ap5A, that mimics the transition state of AMP with ATP (PDB ID: 1ZAK_B) and a structure in the closed conformation also contains Ap5A but closes tighter than the maize structure (PDB ID: 1AKE). For manipulation of these structures, and replacing crystal ligands with ligands appropriate for comparison with
solution scattering data, we used YASARA which is a computer program for molecular visualization, modelling, and dynamics to reconstruct the structures containing all possible ligands[41]. The Ap5A was removed from the 4AKE structure and replaced with ATP. Similarly, the apo 1DVR_B was combined with an ATP and the apo 1DVR_B combined with an ATP and an AMP. Those five structures were then used as a basis set of representative ADK structures. XS program was used to calculate the scattering intensity for each of them. The resultant basis sets of scattering intensity distributions were then used to determine populations of AdK conformations in open, partially open, partially closed and closed conformations that best fit all data sets.

In the experiment, X-ray solution scattering data were obtained from ADK in the presence of different concentrations of ligands, including ATP, ADP and AMP. The proportion of proteins in the liganded ensemble and the unliganded ensemble are estimated by expressing each scattering pattern as a linear combination of scattering in the absence of ligand and scattering at saturation of each ligand. The transition from unliganded to liganded ensembles exhibits the sigmoidal behavior expected for ligand binding. The experiment intensity data, $I_m$ can be fitted by the linear combination of known basis conformation intensities, $I_c$, and with noise multiplicative log-normally distributed (Onuk et al., 2014; 2016). From these fits, the relative abundances of different conformations under all experimental conditions used were determined (Onuk et al., 2016).
Chapter 3

Results and Discussion

3.1 GROMACS MD simulation details

The overall goal of this work is to characterize the conformational transitions of ADK at the initiation of the catalytic reaction cycle. In particular, we analyzed the process whereby ADK transitions from open to closed conformations, slowly moving down what proved to be a rough binding funnel. Molecular Dynamics (MD) simulations were used to simulate the conformational changes of ADK. In the coarse grain simulations, each residue is represented by a single bead centered at the C\textalpha position. We are using the SMOG model (a structure based model for biomolecular dynamics), in which the VDW radius and molecular weight of each bead is defined by different residue type. Two sets of MD trajectory starting points were created one from the open conformation and the other based on the closed conformation, 25 MD simulations were used for each conformation. The 4AKE, chain A and 1AKE, chain A from PDB were used as the representative open and closed (starting and ending) structures for simulation, at room temperature. Focus was placed on the open-to-closed transition.

Initially, we carried out two 5 microsecond, all atom simulations on 1AKE and 4AKE with regular simulation of GROMACS and no-water environment. These simulations never resulted in a conformational change that would constitute passing over the energy barrier between open and closed states, indicating that longer simulation times, as enabled by coarse-graining would be essential for this work. In particular, the MD simulation should be coherent and long enough that we can view the conformational transitions as the trajectory climbs over energetic barriers and travels between the two reference structures. To achieve that, we used the multi-state energy function (described in Methods) to combine
the two potentials created by simulations of 1AKE and 4AKE, and preserve the shape of energy surface near the energy minimum of each reference structure while transitions are allowed between them.

GROMACS and SMOG were used to produce the MD simulation trajectories of ADK from the open (4AKE) to closed (1AKE) state. The MD simulations start from the open conformation (4AKE) of ADK and end in the closed conformation (1AKE). 2fs time steps are used in the simulation. Of the simulations attempted, 16 simulations reached the closed state prior to the end of the 500 nsec time course. For these simulations, we chose CAPF (Cα Atom Pair Factor) to be in the range 1.25-1.35, and the rVDW (cutoff of van der waals radius in MD simulation) to be 2.5-3.0 nm. From start to around 400 ns, most of the time RMSD was near 0.45 Å, with ADK in a partially open state (LID domain mostly closed, NMP domain largely open).

3.2 calculation of RMSD for each trajectory

100000 data point (snap shots) are generated by each trajectory. Every single data point includes a structure which is formed by 214 atoms. The root mean square deviation (RMSD) is a very helpful tool to view the fluctuation of the protein during the simulation. The RMSD of each structure from the trajectory to the reference open (4AKE) structure is computed. The deviation of trajectory to the reference structure along the timeline show the path of the structural change of ADK in simulation.

GROMACS contain the command ‘gmx_rms’, which compares two structures by computing the root mean square deviation (RMSD). An open conformation of ADK is used as a reference state in the RMSD calculation. Each structure from the trajectory is compared to the reference structure. We used the starting frame of the trajectory as the reference. The RMSD of each structure of the trajectory to the reference structure was saved in an output file and plotted along the timeline in the Fig. 6.

The RMSD of the backbone atoms between the reference conformation and the closed structure is about 7 Å. The progression of the transition was measured by the ‘ΔRMSD’, which is the difference in RMSD of each conformation in trajectory frames from the reference open structures (4AKE). In the conformational transition, the NMP and LID
domains of ADK undergo large conformational rearrangements relative to the more stable CORE domain. Based on the distance between the COM of LID and NMP domains to the CORE domain, the procession of ADK transition can be roughly viewed as three states. The open state, the distance of COM of LID and CORE domains is more than 27 Å while the distance of COM of NMP and CORE domains is more than 16.8 Å. In the partial-open state, the distance of COM of LID and CORE domains is less than 27 Å, while the distance of COM of NMP and CORE domains remains more than 16.8 Å. In the closed state, the distance of COM of LID and CORE domains decreases to less than 27 Å while the distance of COM of NMP and CORE domains is less than 16.8 Å. ADK in these simulations transfers from one state to another very quickly. This can be seen from the observation of the ‘ΔRMSD’ change along the trajectory.

In all cases, immediately after the start of the simulation the LID domain of ADK closes over the active site. Then, ADK stays in the partially closed conformation for a variable length of time. During this period, the LID domain occasionally opens and then re-closes over the active site. The conformational transition of ADK between open state and partially open and closed states is very fast. In contrast, before the conformation of ADK reaches a more favorable closed state, it prefers to reside in the partially open state for a variable length of time.

In Fig. 6, the RMSD of several trajectories are plotted along the time sequence:
3.3 The comparison between 45 crystal structures and the ensemble generated by an MD trajectory

In Fig. 7, the structures along the trajectory are plotted on the coordinate system in which the Y axis is the distance between the COM of NMP domain and CORE domain and the X axis is the distance between the COM of LID and CORE domains. The data points from the trajectory naturally fall into three clusters in the plot. The longest cluster on the upper-right of the plot corresponds to the open state. In the early stages of the simulation, the LID domain undergoes the largest rearrangement. The COM distance change between the CORE and LID domain decreases dramatically. The large cluster on the upper-left of
the plot corresponds to a partial-open state. This intermediate state cluster has the biggest population. In the early stage of the simulation, ADK sporadically transfers between the open and partially open states. In this cluster, the LID domain is in an almost closed form and the NMP domain is open. Both the LID and NMP domains exhibit fluctuations in this state. The small round cluster on the lower-left of the plot corresponds to the closed state. In this state, the LID and NMP domains are both closed. The fluctuation of the two domains are highly constrained in this configuration.

In a previous study, an ensemble of 330 DIMS trajectories was compared to 45 ADK protein structures.[39] These 45 ADK crystal structures were used to define a hypothetical open ↔ closed transition pathway of ADK reaction. Here, we wanted to study the structural differences between the crystal structures and structures generated from an MD trajectory. We calculated the distance between LID, NMP and CORE domains for these 45 crystal structures and plotted them with the ensemble of structures from one trajectory. The arrows in the plot start from the positions of each crystal structures on the plot and point from one crystal structures to the next in the order defined by the ‘ΔRMSD’ change along the trajectory as defined by Beckstein[39]. In Onuk et al. (2016) the 45 crystal structures were classified into 5 clusters, and the conformations in Fig. 7 are colored to define those clusters - one color for each cluster. The average COM distance between the CORE and LID domain of the crystal structures which belong to the open and partially open states are smaller than those from the simulated ensembles in the same state. The LID domain in the crystal structure may be compressed by the crystal contacts leading to this discrepancy. The apo ADK in real experiment environment may actually be more open and flexible than observed in the crystal. In the simulation, the COM distance between the CORE and LID domain can reach to much smaller than in crystal. This may be due to the obstruction of ligand in crystal structures. In the simulation of SMOG model, the ligand is not required and the LID domain can close farther to the active site.
Figure 7. The 45 crystal structures defined by Beckstein et al. (2009) and the ensembles of MD trajectories plotted in two-dimensions: core-lid distance (x-axis) and core-NMP distance (y-axis).

3.4 clustering of structures from the MD trajectory

Use of K-means to cluster trajectory into 5 clusters

Thousands of structures are produced from each MD trajectory. Each structure corresponds to a Cα coarse grain ADK structure PDB file, which can be viewed as a data point in a large dimensional space. The distance between each pair of data points is the root mean square distance (RMSD) of the pair of conformations. To study the fine structure of the conformational transitions and define the sub-classes of conformations within the transitional, partially closed ensemble, we want to cluster all conformations generated by
the MD simulations and find representative conformations for each cluster. We use the unsupervised clustering algorithm, the K-means clustering, which is used to study the structural features of the transition intermediates. Clustering is the process of partitioning a large number of data points into a small number of clusters.

The most important part of the algorithm is defining k centroids for k different clusters and determining the maximum number of clusters that can be distinguished using the coordinate data of the structures in each cluster. To do this, we used a score function based on the average ROG of the K clusters. As seen in Fig. 8, when we cluster the data points of the trajectory into 2 clusters, the average ROG value will be very high which is around 1.7. If we increase the number of clusters, the average ROG value drops rapidly. We define f(K) as the following equation. In the equation, the sum of the squared distance between each member of the cluster and its centroid is divided by selected K. In the K means program, a data point x belongs to cluster k if it is closer to centroid k than any other centroids. It is a 214*3 matrix. The average of the coordinate sets of all the structures in one cluster is the centroid of this cluster. The x is each member of cluster i. The ci is the centroid for cluster i. Both of them are 214*3 matrix.

\[ f(K) = \sum_{i=1}^{K} \sum_{x \in c_i} \text{dist}(x, c_i)^2 / K \]

f(K) provides a measure of the statistical significance of adding additional clusters. At a point between 4 and 5 clusters it levels out as can be seen in Figure 8. After we classify more clusters, the decrease of the average ROG value slows down. Additional information cannot be obtained by adding new clusters once the average ROG plateaus. Therefore, we chose K = 5 as the number of clusters we use in the subsequent analysis of the intermediate states.
Figure 8. The elbow plot for K selection. The slope of the f(K) function drops rapidly until the number of clusters reaches 3. At that point it levels off and becomes much flatter after 5.

After the number of clusters, K is defined, we begin to cluster the data points from the trajectory around these K centroids. In the loop of the python program, each data point is traversed through and assigned to the nearest cluster. K-means is a clustering method that aims to find the positions $\mu_i, i=1...k$ of the clusters that minimize the square of the distance from the data points to the cluster.

$$\arg\min_{c} \sum_{i=1}^{k} \sum_{x \in c_i} d(x, \mu_i)^2 = \arg\min_{c} \sum_{i=1}^{k} \sum_{x \in c_i} \|x - \mu_i\|_2^2$$

The K-means clustering algorithm will use the factor A as the distance between any pair of conformations to partition the conformational transitions into K clusters.
We use the RMSD of structures within the ensembles as the distance metric in the K-means algorithm. The 45 structures were also clustered by K-means using the RMSD of X-ray scattering intensity as distance metric (as opposed to RMSD of atomic coordinates).

The factor A will be used to define the distance between any two conformations from trajectories. The A factor equals to the RMSD of all the Cα that belong to the non-contact residue pairs of the two conformations.

**3.5 Properties of clusters**

By plotting the center of mass distance from Core to NMP and LID domains of each ensembles of the MD trajectory, the consecutive track of structural movement of ADK from open to closed state can be viewed in Fig. 9. The CORE-LID and CORE-NMP distance of 45 conformations of *E. coli* AdK that were modeled by homology from PDB structures are compared with the structures from the trajectory[1]. After aligning each structure in the trajectory to the reference structure 4AKE, the unsupervised clustering algorithm, K-means clustering, is used to study the structural feature of the transition intermediates. We use the RMSD between structures as distance metric in K-means algorithm. The 45 structures are also clustered by K-means, using the RMSD of X-ray scattering intensity as distance metric. The average CORE-NMP distance of MD structures in open and partial open states is greater than the crystal structure. The CORE-LID distance of MD structures can open wider in the open state than observed in the crystal structure 4AKE. These observations suggest that the actual structure of ADK in solution may be more flexible than the crystal structure in the sense of being 'more open' than observed in the crystal. Compared with the ADK crystal structure in open state (4AKE), the NMP domain in the real solution environment may stay in a wider open position when ADK is in the open and partially open states. The movement of the LID domain is very large and quick in solution. The wider open LID and NMP domain may help to expose active site and better assist the ligands getting into the right position.
Figure 9. The ensembles of previous MD trajectory are clustered into 5 groups by K-means and projected onto the domain mass center distance map (above). The X axis is the LID to CORE domain mass center distance, Y axis is the NMP to CORE domain mass center distance. The black points represents the open state of ADK, the blue, green and red points represent the three partial open state, the light blue points down and to the left represent the closed state.

3.6 Circulation

The structures in the partially open state are classified into three clusters by K-means clustering (the red, dark blue and green clusters in Fig. 9 and Fig. 10. Among the three clusters, the red cluster is largest and is much closer to the open state that the other two intermediate structure clusters. The structures in the green cluster have shorter average distance of COM of LID and CORE domain than structures in the red cluster and have longer average distance of COM of NMP and CORE domain than structures in the dark blue cluster. The dark blue is much closer to the closed state. We calculated the average structure of each cluster. In order to analyzed this 'average structure' in detail, the Cα coarse grain structures were reconstructed to all atom structures by first reconstructing the
backbone structure using PD2 [31] and then adding residues to the backbone structures using SCWRL4. [32]

Figure 10. Connectivity map for conformational transitions during the simulation of ADK rearrangement. The arrows connecting each cluster to one another and the number represent the number of times transition from one cluster to another occurred during the simulation. Numbers within the colored ellipses indicate the number of simulation steps that resulted in the conformation remaining within a single cluster. The trajectory present in this picture is 1.3-2.

In each trajectory, before the NMP domain closes over the active site and ADK reaches the closed state, the ADK generally resides in a partially open state for some time, during which it transfers between the three partially open clusters multiple times. To create a framework for understanding the possible allosteric pathway underlying the catalytic function of ADK, we chose to study the transfer rate between the clusters. We defined a measure of the circulation in the clockwise direction according to:

\[
100\% \left[ \frac{(R_{rg} - R_{gr})}{(R_{gr} + R_{rg})} + \frac{(R_{rb} - R_{br})}{(R_{br} + R_{rb})} + \frac{(R_{gb} - R_{bg})}{(R_{bg} + R_{gb})} \right] = TRF
\]

Where the transfer rate is averaged in one direction, the Rgr, Rrb, Rbg, Rrg, Rbr,Rgb represent the transfer number between the green, red, blue clusters in forward and reverse directions. ADK moves between the partially-opened state clusters many times during a
Table 1. **Transfer rate frequency (TRF)** for 16 MD simulations that completed the transition to closed state in under 500 ns (limiting time for each simulation). Lower CAPF appeared to enhance the clock-wise circulation among the 3 intermediate structure clusters. CAPF is Cα Atom Pair Factor and the rVDW is cutoff of van der waals radius in the MD simulations. The red number in the table is for the trajectory 1.3-2 which is shown in Fig. 10.

<table>
<thead>
<tr>
<th>rVDW</th>
<th>3.0</th>
<th>2.8</th>
<th>2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.35</td>
<td>-1.97</td>
<td>76</td>
<td>-4.59</td>
</tr>
<tr>
<td>1.32</td>
<td>0.867 0.818</td>
<td>2.58</td>
<td>-0.266</td>
</tr>
<tr>
<td>1.3</td>
<td>0.345 0.932</td>
<td>-0.233</td>
<td>-2.47</td>
</tr>
<tr>
<td>1.25</td>
<td>0.728 0.252 0.441</td>
<td>8.3</td>
<td>1.77</td>
</tr>
</tbody>
</table>

Before the NMP domain closes on the CORE domain and the phosphotransferase reaction occurs, ADK stays at the partial open state for a long time. During this period, large-scale conformational fluctuations occur. We apply a general computational strategy to characterize such a conformational change in ADK.

In the transition between the open cluster and the all three partially open clusters, around 90% of the transitions involve the red cluster which appears to act as a hub connecting the other two clusters and the open cluster. ADK does not transfer directly from an open to closed state.

For some choices of force parameters a circulation between the different clusters is observed. For most choices of force parameters, the anti-clockwise direction is more favored. This apparent circulation within the conformational ensemble acts like a molecular ratchet. The origin of this behavior may be truncation of the van der Waals force, other parameters chosen in the coarse graining or may reflect actual behavior of the protein in the partially open state.
<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster 4</th>
<th>Cluster 3</th>
<th>Cluster 0</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 4</td>
<td>56.28%</td>
<td>17.26%</td>
<td>26.1%</td>
<td>0.35%</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>60.2%</td>
<td>16.02%</td>
<td>23.65%</td>
<td>0.14%</td>
</tr>
<tr>
<td>Cluster 0</td>
<td>60.62%</td>
<td>15.32%</td>
<td>24%</td>
<td>0.18%</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>3.7%</td>
<td>0.34%</td>
<td>0.18%</td>
<td>95.82%</td>
</tr>
</tbody>
</table>

| Population | 38735 | 11032 | 16743 | 3872 |
| Percentage(%) | 55 | 16 | 24 | 5 |

Table 2. The percentage of transition from one cluster to the other three clusters.

The transfer connectivity map (Fig. 10) of ADK shows the actual rate of ADK transfer among the open cluster and three partially open clusters. In each cluster, there are a lot of possible conformations of ADK. In a complete MD simulation, a free energy map can be built to show the spatial positions of interacting molecules in a system. The most populated cluster of possible conformations of protein often occur around regions of the lowest free energy local minimums. In our MD simulation of ADK, since the transfer of ADK from partial open state to the closed state only happens once in most the trajectories, then the ADK stays at the closed state and never transfers back, that makes construction of an free energy map of the whole trajectory un-trustworthy. Here, we are only interested in the energy difference between the four open and partial open clusters. In Table 2, we show the percentage of transitions from one cluster to the other three clusters. The transfer rate difference of one cluster to another is a measure of the energetic basin difference between the clusters. The population ratio between the four clusters is, red: green: blue: black 55: 16: 24: 5. The red cluster has the largest population, thereby, among the four clusters, it most likely represents the deepest energy basin. The black cluster (open state)
has the smallest population, which shows that the open state, as defined by the parameters of these simulations, is less stable than the partial open state. Among the three partial open sub-clusters, the green cluster has the smallest population, which is, therefore, least stable.

When ADK takes on a conformation in the red cluster, the percentage of chance that ADK stays in red cluster during any one step in the trajectory is around 56.3%. When ADK sits in the blue cluster, the percentage of transition from blue to red cluster is around 60.6%. When ADK is in the unstable green cluster, ADK likes to transfer to the red cluster, the percentage of transition from green to red cluster is around 60.2%. The ratio of transition from blue to red cluster is almost the same as that from green to red cluster. When ADK sits in the blue cluster, the percentage of chance that ADK stays in blue cluster is around 24%. When ADK sits in the green cluster, the percentage of transition from green to blue cluster is around 23.7%. When ADK in the red cluster, the percentage of transition from red to blue cluster is around 26.1%. The ratio of transition from red to blue cluster is a little bit higher than that from green to blue cluster. When ADK sits in the blue cluster, the percentage of transition from blue to green cluster is around 15.3%. When ADK in the red cluster, the percentage of transition from red to green cluster is around 17.3%. The ratio of transition from red to green cluster is a little bit higher than that from blue to green cluster. The transition from red to blue cluster are both in clockwise direction. The transition from red to green cluster is in anti-clockwise direction. Those transition rate difference may determine the clockwise or the anti-clockwise direction is more favored in circulation.

3.7 Comparison among the average structures of the three partial open clusters

To study the correlation between the transfer rate and the structural features of partial open clusters, we aligned the backbone of the three representative structures of the partial open clusters.

First, we calculated the average structure of the three partial open clusters. In each cluster, the Ca structures are aligned with their centroid structures. Then, the structure coordinates in one cluster are averaged by numpy.mean() function. The results are the representative structures of the sub-states. In our research, we used three different programs
to reconstruct the backbone from the $C\alpha$ structures. The three programs are PD2, maxsprout, and sabbacBB2. The backbone structures are reconstructed to all atom structures by SCWRL4. In Fig. 11, the three reconstructed structures are built from the $C\alpha$ structures of 4AKE. The scattering intensity of the three reconstructed structures are calculated by XS program. The three scattering intensities are compared with the scattering intensity of all atom 4AKE structure which is also calculated by XS program. The Fig. 11 shows the four scattering intensities are very similar and the four programs, PD2, maxsprout, sabbacBB2, and SCWRL4 can construct an all atom structure from the $C\alpha$ structures very well. Finally, we use the $C\alpha$ coarse grain representative structures which are reconstructed to backbone structures by PD2. Then, we aligned the backbone of the three representative structures of the three partial open clusters.
Figure 11. Testing different algorithms for reconstructing all atom structures from Cα structures. The X-ray scattering intensities of 4AKE crystal structure and the open ADK structures which are reconstructed from the Cα structure of 4AKE. The backbone structures are reconstructed by PD2, maxsprout[38], sabbacBB2[37], then the all atom structures are reconstructed from the backbone.
structures by SCWRL4. The intensity of 4AKE crystal structure matches well with the reconstructed structures. Around the 18-21 A, the crystal structure has higher intensity.

![Images of protein structures with annotations](image1.png)

Figure 12. Comparison of the backbone of the three average structures of the partial open clusters show the structural features of each cluster. The different colors of each structure correspond to the color of clusters. Different angles (A, B, and C) allow visualization of the three places where the distinctions between clusters are most obvious.

In the orange circle of Fig. 12(A), the red structure, compared with the other two structures, exhibits structural differences in the lower part of an important long alpha helix, which connects the CORE and LID domains. In the open structure of ADK (4AKE), this long alpha helix is the alpha helix 9. In the closed structure of ADK (1AKE), this long alpha helix breaks into two, designated alpha helix 10 and 11. The structural differences in the orange circle show the transformation of the long alpha helix into two helixes may happens at the early stage of the of the sub-states.

In the yellow circle of Fig. 12(B), the green and blue structures are more similar with each other, compared to the red structure. The red structure forms a better beta sheet
structure than the other two. In the orange circle of Fig. 12(B), this part of the LID domain of the red structure shows tendency to be wide open. In the orange circle of Fig. 12(C), the linker between the NMP and CORE domain of the blue structure shows a structural change, compared with the other two. This kink structure in the orange circle of Fig. 12(C) is on one of the linkers of NMP and CORE domains and facing one side of the linker of LID and CORE domains.

Summarizing, the red cluster is the most open of the sub-states, as seen in Fig. 10, and acts as a hub connecting the other two partial open clusters with the open state. The wide open configuration of the LID domain, combined with the structural changes observed in the lower part of the hinge helix appear to facilitate the movement of LID domain to and from the open state. The distance between the center of mass of the CORE and LID domains of the red cluster is larger than in the other two clusters. In the yellow and orange circles of Fig. 12(B), the green and blue structures have the similar structure and the red structure is different from them. This suggests that the LID domain in the green and blue substates are closer to the CORE domain and the regions in the yellow and orange circles are more influenced by the active site. The blue cluster is the most nearly closed of the substates, as seen in Fig. 10. And in most of our MD simulation trajectories, the last frame before ADK transfer into closed state falls at the blue cluster. This structural difference in the orange circle of Fig. 12(C) shows tendency of closure of NMP domain.

3.8 Changes of radius of gyration of domains of ADK in different clusters

To further explore the structural change of the subdomains in the five clusters, we calculate distribution of the radius of gyration of the structures in different clusters. The distribution of the atoms of a protein around an axis is the radius of gyration of the protein. The ROG is used to describe the dimensions of a polymer chain. The ROG indicates the compactness of protein structure. The ADK is important in many cellular signaling pathways which are helped by its large conformational changes. The flexibility and plasticity of ADK help it to bind ligands, form oligomers, and perform mechanical work. [29] In the reaction, the NMP and LID domain of ADK undergoes large conformational
rearrangement relative to the stable CORE domain. The rearrangement can be observed from quantitating the significant RMSD changes of ADK in the trajectory. The high strain energy which is released by the conformational changes may induce protein unfolding at localized regions. This folding/unfolding are also related to the catalytic efficiency enhancement. [29] We want to study the function of different domain in different step of reaction which requires overall research on the fluctuation of different domain in different state. What kind of change will be brought to the conformation of different domain by cracking. Radius of gyration is a tool to study the overall spread of the molecule, so we use it to find the relationship between the global movement of different domains and their own volume change.

\[ R_g^2 \overset{\text{def}}{=} \frac{1}{N} \sum_{k=1}^{N} (\mathbf{r}_k - \mathbf{r}_{\text{mean}})^2 \]

Where \(N\) is the number of Cα atoms in the domain. The \(\mathbf{r}_{\text{mean}}\) is the coordinate of center of mass (COM) of the domain. The \(\mathbf{r}_k\) is the coordinate of the Cα atom \(k\). The data points from trajectory are classify by k means algorithm into 5 clusters. The 5 clusters are divided into open, partial open and closed states by the average distance between CORE and LID, NMP domains in one cluster. The dimensions change of ADK’s each domain in different state is monitored by ROG which can be viewed in Fig. 13.

Core domain is generally considered as the most stable domain in the rearrangement, however the means of radius of gyration of core domain in five different clusters have the biggest change. The dimensions of core domain expand in the open state and shrink in the close state. This may be due to movement of residues immediately proximal to the hinges connecting core to lid and to NMP domains.
The average of the ROG of the NMP domain does not change much in open and partial open state, and drops down in the closed state. The averages of the ROG of the LID domain stay the same in all five clusters.

The clusters of the LID and NMP have similar averages of the ROG in open and partial state, but the standard deviation of NMP is much higher than LID clusters. The fluctuation of NMP domain is higher than LID domain.

According to the result, the dimension of core domain is undergoing significant changes in different states. These changes may be caused by the stretching of LID and NMP domain in conformational movement and movement of residues proximal to these two domains. The stretch of core domain in open state may help to expose the active site to the ligands. The shrinkage of core domain in partial open and closed state may assist the residues in the active site to adjust their position and interact with ligands. In the NMP kinase family, the length of LID domain ranges from 11-residues to 38-residues. Comparison of crystal structures and amino acid sequences of AK from many species, the varied LID domain is not considered as an essential part for a functional enzyme. [30] The variation leads to drastic change in the conformation of the LID domain. On the other hand, the location of active site on CORE and NMP domain do not change significantly. That may explain why the average of ROG of LID domain does not change so much. The interaction between ligands and the CORE and NMP domains may need more structural changes in these two domains to assist the phosphotransferase reaction. The LID domain plays a more important role in ligands recruiting and stabilization.
cluster0_STD: 0.1230
cluster1_STD: 0.1239
cluster2_STD: 0.1308

cluster0_STD: 0.1180
cluster1_STD: 0.1059
cluster2_STD: 0.1284

cluster0_STD: 0.0675
cluster1_STD: 0.0696
cluster2_STD: 0.0678

cluster0_STD: 0.0679
cluster1_STD: 0.0696
cluster2_STD: 0.0678
Figure 13. The distribution of the ROG of ADK's three domains in different clusters. The distribution of the ROG of ADK's three domains in different clusters. (A) ROG of the CORE domain is smaller in the closed state than the intermediate states, and somewhat larger in the open state. The ROG of the blue cluster is slightly smaller than for the green and red cluster, and corresponds to its place as a gate-keeper to the closed state. (B) The ROG of the NMP domain changes sparingly, being significantly smaller in the closed state, but only slightly different between the intermediate and open states. (C) In these simulations the ROF of the LID domain is apparently insensitive to its position relative to the CORE domain, exhibiting essentially the same distribution in the open, closed and intermediate states.

3.9 Standard deviation of interatomic distance variations for each state

\[
\text{RMSD}(v, w) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|v_i - w_i\|^2}
\]

Fluctuation between atoms is an important tool to observe the relative motion in protein. The ADK has three domains and large conformational rearrangements are involved in protein function. When the chemical reaction occurs, the LID domain and NMP domain close onto the active site, the ligands ATP and AMP are enclosed in the active site. The surface of the LID domain and the CORE domain both contact with the ATP. The surface of the NMP domain and the CORE domain both contact with the AMP. Each ligand contacts different regions in the closed state. The relative motions in those contact areas are important reflections of the interaction forces in those regions and the mechanism of reaction. The root mean square fluctuation between each Cα atom pair shown in Fig. 14 reveals the internal relative fluctuation of ADK in the trajectory 1.27.
Figure 14. The root mean square fluctuation between each Cα atom pair in ADK. There is total 214 Cα atoms in *E.coli* ADK. In this plot red indicates a high fluctuation in a Cα - Cα distance and blue a low level of fluctuation. The dark blue diagonal means the atoms close to one another along the peptide chain move little relative to one another. The yellow and red region indicates interatomic distances undergoing relatively large movements of NMP relative to LID. CAPF = 1.27 for this simulation.

We present the RMSD of distance matrix between CA atoms (lower triangular area), for trajectory 1.27 in which CAPF = 1.27. The section that corresponds to the motion of the LID domain is enclosed by the red square. The motion of LID towards the rest of protein result in a much lighter region that in the left of the square Fig. 14. The section that corresponding to the motion of the NMP domain is enclosed by the pink square in Fig. 14. There are three regions which are resulted by the motions between the NMP and the rest of
protein. The subsequent closure of the NMPbind domain creates the two large red and yellow regions which are around the junction between the LID and NMP domain squares. Most interesting is that the greatest fluctuations are not between the LID and NMP domain, it is between the NMP domain and the joints of CORE and LID domain. Another interesting thing is the other two regions which NMP domain have large motion with are on the P-loop region and N-terminus. The two joints of CORE and LID domain, the P-loop and N-terminus of the protein are in a line. It is speculated that the motion of NMP domain closure may along this same line.

To explore the relative motions among the three domains in different states, the matrix of root mean square fluctuation between each Ca atom pair of ADK for different states are produced. The results are shown in figure 15 and interesting since the domain motion show the characteristic in different state which reflects the relation between the change of Ca atom pair fluctuation and domain rearrangement.
open state
In the open state, (Fig. 15 a, b, c) The motion of LID towards the rest of protein results in a much higher standard deviation of inter-atomic distances in the upper right of the plot representing the open conformation (top)(Fig. 15 c). In this state, the LID domain undergoes large motion. The another region which has relative high fluctuation in this state is around the Gly56. The motion of Gly56 is mainly relative to the LID domain region. The fluctuation of the Gly56 stays at a similar level in the open and partial open state, and the
fluctuation suddenly drops in the closed state. There is only one other region in ADK
conforms to this pattern which is the alpha helix 160-175.

In the partial open state, (Fig. 15 d, e, f) the strongest relative movement happens
among the residues marked in pink in the structural rendering (Fig 15 d). The fluctuation in
the LID domain drops suddenly in this state. The relative motion between the Gly56 and
the LID domain region is more obvious in (Fig. 15 d, e, f). The relative motion between the
Gly56 and P-loop, Ala203 regions are also obvious in (Fig. 15 d, e, f). In this state, the LID
domain is closed on the CORE domain and the NMP domain stays open. The largest
relative motion between the Gly56 and the LID domain is more on the side of the LID
domain which is facing the Gly56. The LID domain may plays an important role in the
closure of NMP domain.

In the closed state, (Fig. 15 g, h, i) the fluctuation of most residues of ADK drops
quickly. There are still several region highlight in (Fig. 15 i). In this state, both LID and
NMP domains are closed over the CORE domain. The three residues Ala66, Lys69, Ala73
are on one helix. The Ser41, Ala99 and the three residues on one helix have strong relative
motion. The three regions are close to each other in space. These residues are facing
outside of protein and remote from the active side. Another relative large motion is
between the loop region of LID domain and Ala203.
Figure 16. The comparison of the RMSF of each residue in different state.
Figure 17. The blue residues (residue 56 and alpha helix 159-176) shows intense relative motion in open and partial open state.

Fluctuation between region around residue 56 and alpha helix 159-176 may occur due to the closure of NMP domain. Those two regions keep similar high fluctuation in open and partial open state. Then, the fluctuation drops sharply after the NMP domain closes over the active site. After the LID and NMP domains both close, Gly56 is very close to the helix 159-176. This observation may prove the fluctuation of these two regions plays an important role in dragging NMP domain.
The long helix 160-188 in 4AKE is separated into two parts in 1AKE. The helix breakage also reveals itself in the fluctuation of the region at a boundary apparently centered on Ala176. In the partial open state, the alpha helix 160-175 exhibits similar fluctuations as in the open state. On the other hand, comparing with the open state, the fluctuation of the second part of the helix 177-188 drops dramatically in the partially open state and then becomes much more rigid after NMP domain closure.

The regions 64-75 and 94-101 have much higher fluctuation in the closed state than the partial open state. These two regions are spatially close to each other. The dihedral angle 81-84 which has the high fluctuations in the closed state is on a beta-sheet which is between these two regions. The high fluctuation of dihedral angle 81-84 may be influenced by them. The fluctuation of 64-75 and 94-101 are more correlative to the region 41-44, which is the loop linking the two alpha helix of NMP domain together. This observation is corresponding to high fluctuation of dihedral angle 41-44 in the closed state.

The N-terminal 1-41 and C-terminal 176-214 regions are very interesting. Since the fluctuation level of these two sections stays the same in the partial open and closed states, the relative motion of the two regions looks like more relative with the closure of the LID domain instead of the closure of the NMP domain.

3.10. Pseudodihedral angle changes along the MD trajectory

The dihedral angle is the angle between two intersecting planes and this angle is on a third plane which is perpendicular to the line of intersection. In our work, the coarse grain SMOG model is used in MD simulation, and the ensembles of trajectories are Cα structures. Since the 4 atoms CA, N, C', and Cα form a plane, the Cα atom traces provides a simplified backbone representation and are adequate for representing the global folds of proteins. The pseudodihedral angle in our research is defined as the angle between the two successive planes which are formed by four consecutive alpha carbon atoms. The purely geometric information of protein is contained in pseudodihedral angle. The analysis of variation of the Cα pseudodihedral angle in the open to closed pathway provides the information of local conformational change of ADK in different domains and different
states. Comparing with the RMSF (°) between Cα atom pairs of ADK, the dihedral angle fluctuation is directly related relative to the previous study of radius of gyration. Comparison of conformational change shows characteristic features of these secondary-structure elements in different states.

Figure 18. Pseudodihedral angle changes along the MD trajectory(1.27). Green represents very small change; yellow and blue increasing levels of change. The large regions of uniform green reflect the fact that most of the structural changes ongoing during these MD simulations are rigid-body motions (at the level of the main chain), with large scale changes limited to a relatively small number of pseudo-dihedral angles.
Figure 19. The fluctuation change of pseudodihedral angles in different cluster of the MD trajectory (1-2.7).
Figure 20. The top 20 most unstable pseudodihedral angles in the trajectory.

The above Fig. 20 shows the most fluctuate pseudodihedral angles. In Fig. 20 we only mark the second residue of each pseudodihedral angle as pink. The high fluctuation dihedral angle residues are gathered on the parallel pack beta sheets in CORE domain and the beta sheets in LID domain. The fluctuation of dihedral angle is more likely to affect the neighbouring beta sheets, which shows the force between beta sheets holding the structure
of core domain. Arg2 which is close to the Ile101 may influence each other. The two high fluctuation pseudodihedral angles which are around Arg88 and ASN102 site between the NMP and LID domains. It may relative with the closure of NMP.

Figure 21. The comparison of Cα dihedral angle fluctuation standard deviation in different states. (1) the STD of Cα dihedral angle fluctuation around the red location has higher value in open state than the other two states (0.45 vs 0.23) (2) the STD of Cα dihedral angle fluctuation around the green location has higher value in closed state than the other two states (3) the STD of Cα dihedral angle fluctuation around the blue location has much lower value in closed state than the other two states.
The Fig. 19 shows the STD of the pseudodihedral angle variation of ADK in 5 clusters. The top unstable pseudodihedrals are almost the same in all 5 clusters. The result is interesting since the 5 pseudodihedral angles behave very different in different states. The 5 pseudodihedral angles can be classified into 3 types, which can be viewed in Fig. 19. The fluctuation of dihedral angle change around the blue residues in Fig 21(3) reduce after closure of LID and NMP domain. The two dihedral angles in the blue region include the residues Ile4, Leu5, Leu6, Gly7, Ser129, Gly130, Arg131, Val132. The Leu5 and leu6 site on the core domain and adjacent to the P-loop of ADK. Since the adjacent residues belong to hydrophobic group, they do not have direct interaction with ligands. They may influence the reaction indirectly. The fluctuation in this region will influence the fluctuation of P-loop, but do not influence the structure stabilization of P-loop. In the open state, the LID domain undergoes large conformational transitions, it transforms between open and approximately closed state. P-loop is a very important component of giant anion hole, which attracts and stabilizes phosphoryl group of ATP. R131 surrounding AP5A’s phosphate groups in 1AKE, presumably to help stabilize the negatively charged phosphates. Y133 forms hydrogen bond with O3A of ATP, contribute significantly to the large movement of domain INSERT. The large fluctuation decrease of the two dihedral angles may be caused by the closure of NMP domain. Those blue regions may play an important role in catching and reorienting ligands.

Gly 12 and Lys13 sit on the P-loop. Lys13 forms a bridge between the AMP and ATP phosphates, and stabilizes the active site. Comparing with the blue regions, the fluctuation of red region drops immediately after LID domain close over to the active state. The dihedral angle in the red region include the residues Ala11, Gly12, Lys13, Gly14. In the partial open and closed state, the red region do not change very much. The stable conformation of P-loop region is important for stabilization of the binding between ATP and ADK. Reorientation of ATP and interaction of AMP and ATP are more helped by the blue region.

The fluctuation of green dihedral angles (41-44, 81-84, 195-197) increase after the NMP closes over the active site. The three dihedral angles in the green region include the residues Ser41, Gly42, Ser43, Glu44, Phe81, Leu82, Leu83, Asp84, Lys195, Val196. The residues
41-44 is a loop on the NMP domain, adjacent to helix alpha3 which is 44-54. In 1AKE complexes I and II, the residues 46–48 have two different secondary structures [16]. So, this region may easily be influenced by conformational changes. The increase fluctuation of dihedral angle 41-44 in closed state may be relative to the transfer of gamma phosphate from ATP to AMP or the release of ADPs. The dihedral angle 81-84 is on the beta sheet 3. One of the most important residue Asp84 in binding the Mg2+ and water complex to ADK is in the dihedral angle 81-84. Mg2+ is a very important component for the chemical and structural role in the reaction. Mg2+ and the 4 Mg2+-coordinated waters help to keep the distance and geometry of the ATP gamma- and beta-phosphates and AMP alpha-phosphate. The associative reaction mechanism for phosphoryl transfer are supported by the complex. Asp84 also forms H bond with Thr31 and Ser30. Thr31 have stable link with adenine of AMP. Ser30 has stable link with Lys13 and bridges to Mg2+. [24] The fluctuation of dihedral angle 81-84 may play an important part in the structural role of Mg2+. It may influence the orientation of AMP and release of ligands. The dihedral angle 195-197 is on the beta sheet 9. Some residues closed to it, Gly198, Tyr199 and K200, form interaction with the bond ligand ATP. Those interactions are more or less in synchronization with the movement of the adenine ring of ATP. So the structural change of the dihedral angle 195-197 may influence the stability of the ATP binding. The beta sheet 9 is parallel to beta sheet 3. There are two more parallel beta sheets, beta sheet 1 and 4, sit between the beta sheet 3 and 9 in spatially. These four beta sheets sit in the middle of the CORE domain. They play a very important role for the stability of the CORE domain. The increase of fluctuation of the dihedral angles 81-84 and 195-197 may influence the fluctuation of the four parallel beta sheets and the stable of CORE domain and active site at closed state.
3.11. the comparison of ADK secondary structure RMSD change

Figure 22. RMSD change of LID ((A), 122-160) and CORE ((B)60-122) domain in trajectory 1.27.
To track the fluctuation of the secondary structure of ADK in MD simulation, we calculate the RMSD change of different fragments of ADK in the trajectory 1.27. The open conformation 4AKE is used as reference structure. The target fragment of ADK is picked and cut from every snapshot of the trajectory. The fragment pieces are aligned with the corresponding section of 4AKE. Then the distance between the fragment of snapshot and 4AKE is calculated. The structural change in those fragment is closely related to the domain rearrangement.

The residues 60-122 of ADK is the NMP domain. The peaks in the orange circle are brought by the rearrangement of LID domain. The peak in the green circle are brought by the closure of NMP domain. These can be viewed as “fingerprints” of ADK rearrangement. The residues 122-160 of ADK is the middle part of the NMP domain. However, the RMSD change of the LID domain is not influenced by the LID or NMP’s movement.
Figure 23. The RMSD change of N-terminus ((A), 1-42) and C-terminus ((B), 178-214) region of ADK in trajectory 1.27.

The N-ter and C-ter do not have structural change when LID domain closes. The N-ter 1-42 contains the linker between the CORE and NMP domain, so the NMP domain closure induces significant structural change in N-ter of ADK. These RMSD of the two regions behave very different with the RMSF between these two and rest of ADK. So when LID
domain close, the N-ter and C-ter do not change their structure, but relatively move as a whole with the LID domain.
Figure 24. The RMSD change of NMP domain (30-60) of ADK in trajectory 1.27.

The NMP domain is formed by two alpha helixes and a short loop. The conformational rearrangement of this domain is only influenced by the motion of NMP domain closure. The structural change happens mainly on the outside alpha helix and loop region. The conformational rearrangement does not happen on the one which is closer to the LID domain.
Figure 25. The RMSD change of the long long alpha-helix (161-190) links LID and CORE domains of ADK in trajectory 1.27.
161-190 is a long alpha-helix in CORE domain of 4AKE. In 1AKE, it is divided in to two alpha-helixes 161-176, 177-190. The influence of LID domain closure only shows on the upper alpha-helix 161-176. When the domain rearrangement happen, most part of the domain stay still. The lower alpha-helix 177-190 is not influenced by the domain rearrangement.
Figure 26. The RMSD change around the LID and CORE domains’ linker region (156-190 and 113-122) of ADK in trajectory 1.27.
The 113-122 and 156-179 regions include the linkers that links the CORE and LID. Both of them are strongly influenced by the rearrangement of the LID domain and barely influenced by the closure of NMP domain. In the closed state, the fluctuation of the joint 156-179 regions is reduced. The joint 156-176 is on the side of NMP domain. The closure of NMP domain may repress the fluctuation of the joint 156-179.
FIGURE 27 (Part 1):

The RMSD change of ADX fragments in the trajectory 1.27

RMSD (Å)

Time (~2 fs)

0 2000 4000 6000 8000 10000

0 0.5 1 1.5 2 2.5 3

"joint-61-72.out"

The RMSD change of ADX fragments in the trajectory 1.27

RMSD (Å)

Time (~2 fs)

0 2000 4000 6000 8000 10000

0 0.5 1 1.5 2 2.5 3

"S sheet:80-86.out"
FIGURE 27 (Part 2)
Figure 27. The RMSD change of the secondary structures in CORE domains (60-122) in trajectory 1.27. Regions chosen for display were judged to be of particular interest, including residues 61-72 (joint - loop connecting two secondary structures), 80-86 (beta sheet), 89-100 (alpha-helix), 104-112 (beta-sheet), 113-122 (loop connecting two secondary structures).

The middle part of CORE domain is formed by two beta-sheets and three alpha-helixes. The beta-sheet 104-112 is interesting. Since it is close to the LID domain, however its conformational change is influenced by the NMP closure. The alpha-helix 113-122 is the linker between CORE and LID domains, it is influenced by LID closure. The alpha-helix 61-72 is the linker between CORE and NMP domains, it is influenced by NMP closure.

The high fluctuation dihedral angle residue 27-30, and the high fluctuation Gly56 may reveal the mechanism of closure of NMP domain. These two regions are on the joint between NMP and CORE domains. The relative fluctuation between 27-30 and the rest of protein stays at the same level in partial open and closed stage. The fluctuation of dihedral angle residue 27-30 is more relative to the structure change in this region. The fluctuation Gly56 is influenced by the relative motion between itself and the alpha helix 160-175. The dihedral angle fluctuation around Gly56 is relatively low. The conformation change of
alpha-helix 45-56 is not influenced by NMP closure. The translation of Gly56 may be relative with the rotation of the dihedral angle residue 27-30. The rotation may push the motion of translation.

The mobile domain is linked to the rest of the protein by two helices packed together in an antiparallel fashion.

3.12 The comparison of RMSF for mutant (MT) G85V and Wild type(WT)

The Gly85 is part of the beta-turn, which connect the beta-sheet 3 and alpha-helix 7. If you check the Fig. 20, you will see the Gly85 is very close to the highest fluctuation pseudodihedral angle 86-89. On the other side, Gly85 adjacent the pseudodihedral angle 81-84. The fluctuation of this dihedral angle increases after the NMP close over to the active site. These makes Gly85 a very interesting residue to study. In others research, Gly85 was mutated to Val.[33] The MT G85V does not alter tertiary and secondary structure of ADK very much. Gly85 and Phe86 play a major role in stabilizing the three dimensional structure of ADK. G85V brought thermodynamic destabilization to ADK. G85V ADK was proteolyzed four times faster than the wild-type protein. After Gly85 mutant, the affinity of ADK for ATP was decreased by a factor of 5.6 as compared to the wild type protein.

The Gly85 is a residue very closely related to the stability of ADK. MT G85V will definitely alter the stability of local structure. And we want to find out how does this MT influences the rearrangement of sub-domains and fluctuation between the Cα atom pairs. Gly85 is on the CORE domain. Many previous research show CORE domain is stable in the reaction. However, in our study, the CORE domain undergoes large ROG change in different states. In the study of the pseudodihedral angle fluctuation, several dihedral angles which are close to center of CORE domain show interesting fluctuation change at different states. On the other hand, in the study of RMSF between each Cα atom pair, most of the Cα atom pairs which show high RMSF in all three states are far away from the center of CORE domain. Those led us to explore the relationship between the local structural change around the center of CORE domain and the fluctuation of the distal part of ADK.
In the ‘ΔRMSD’ measure along the trajectory of MT G85V, ADK stays at the open state for less than 10 frames. The LID domain closes over to the active site quickly and never opens again in this trajectory. After the ADK step into partial open state, it transfers among three partial open sub-states and stays at this stage for less than 2000 frames. Then, ADK transfers into the closed state. The transfer point arrives much earlier in the simulation trajectory of MT G85V than the most of the cases in wild type. The NMP domain closes fast in the MT G85V trajectory.

We study the RMSF between each Cα atom pair of MT G85V ADK in the three different partial open sub-states and compare the maps with the WT. In the red cluster map, the MT G85V is very different from the WT. The RMSF is very strong between the atom pair 56 and 157. In the previous study, 56 shows strong relative motion with the LID domain in the open and partial open states. In the MT G85V trajectory, the residues around 157 have much stronger relative motion with 56 than the other part of LID domain. Since the distance between COM of LID and CORE domains is longer in the red cluster than the other two partial open sub-states, the attraction between residue 57 and the most part LID domain is weaker in the MT G85V. The movement of LID domain of MT G85V looks less like a whole body. In the green cluster, the residues around 157 of MT G85V have strong relative motion with the most part of the NMP domain. In the blue cluster, the yellow region around the joint place between the NMP and LID domain is bigger in the MT G85V than in the WT. In the WT, the relative motion centre around the 56 with the LID domain. In the MT G85V, the most part of NMP domain has significant relative motion with the LID domain. This may be caused by the unstable local structure in the CORE domain. This unstable structure may induce stronger relative motion between the NMP and LID domains.
Figure 28. The comparison of RMSF map of MT G85V and WT in partial open state.
Chapter 4

Conclusions

The clustering of conformations within an MD trajectory provided us with a powerful tool to study the roughness of the folding landscape during the closing of ADK. We produced coarse-grained simulation data of ADK’s phosphoryl transfer reaction using a GROMCS and SMOG model. Each MD trajectory was classified into sub-states by k-means clustering. The transitions among sub-states and the conformational fluctuations in the different domains of ADK were analyzed in our research. In our MD simulation trajectories, the LID domain always closes first before the NMP domain close and then, the ADK reaches the closed state.

In our study, we define the residues 113-122 (close to LID-CORE hinge) and the residues 29-31 (close to NMP-CORE hinge) as the distal hinges, the residues 164-169 (close to LID-CORE hinge) and the residues 55-58 (close to NMP-CORE hinge) as the proximal hinges.

In the study of RMSF between each Cα atom pair of ADK in different states, we found the fluctuation between the residues 55-58 (the NMP-CORE proximal hinge) and the other parts of ADK is very strong. This relative motion starts from the open and partially open states. This motion maybe critical for closure of NMP domain. In the pseudodiheiral angle fluctuation analysis, the residues 29-31 (the NMP-CORE proximal hinge) are very flexible during the whole trajectory. On the other hand, the NMP-CORE proximal hinge lacks local structural change, but has strong motion relative to the other parts of ADK. Closing of the NMP may start at the NMP-CORE proximal hinge which may be helped by the structural change of the distal NMP hinge region. The proximal hinges have translational motion relative to each other. The distal hinges shows local structural changes in simulation. The structural change of the distal hinges may contribute to the closing of the proximal hinges.

The fluctuation of dihedral angles (41-44, 81-84, 195-197) increase after ADK enters the closed state. Residues 195-197 are on the beta-sheet which is closest to the C-terminus. The prominent relative motion between LID domain and C-terminus in the closed state are
also close to the same beta-sheet. The relative motion between LID domain and C-terminus in the closed state may help manipulate the ATP into position for chemical reaction and subsequent release of the ADP product.

The prominent relative motion between NMP and CORE domains may also be influenced by the flexible dihedral angles, 81-84, 194-197 and 41-44, in the closed state. 81-84 is on the beta-sheet which is between the two alpha helixes 91-99 and 61-73. 194-197 is at the neighborhood of Ala203 and LID domain. The residues between 194-197 and Ala203 interact with the AP5A adenine atoms directly in the closed form of ADK (1AKE) (10). And the 41-44 is adjacent to the alpha helix 31-41. The NMP residues that have strong motion relative to other parts of ADK match with the NMP residues that exhibit strong dihedral angle fluctuations. The stability of residues 44-48, which is adjacent to 41-44, may influence the conformational changes of ADK (9). The Ala203, Phe137 are at the ATP site, while the 81-84, Ala37 and Ser41 are close to the AMP side. And Val148 sits between the two ends as a bridge. This arrangement suggests that the relative motion of the NMP and CORE domains in the closed state may aid the docking of the AMP and eventual release of the product ADP.

The closed conformation - although generally most rigid – includes some parts that fluctuate more than in the partially open state. Those exceptions appear to move in unison – correlated with one another – and may be critical for the catalytic activity of this enzyme.
Chapter 5

Future directions

In the current study, our research has analyzed WT and MT G85V ADK. The RMSF between each Cα atom pair analysis of MT G85V ADK demonstrated that structural instability in the CORE domain may induce structural fluctuation changes. There is significant evidence that structural fluctuations may actually support enzyme activity in other systems.[43][44][45] Our results indicate that the approach developed here may have broad utility for analysis of the impact of single amino acid substitutions on enzyme activity. In future, we’ll design the SMOG model MD simulation on other MT ADK. MT on different sites of ADK will decreased thermodynamic stability on different region of ADK. The comparison of structural fluctuation variation of different MT ADK will help us understand more about the relation between the ADK’s structural fluctuations and catalytic activity. A lot of new unsupervised clustering algorithms are developed to deal with the huge MD simulation data nowadays. This research also helps us to explore the reliability of these clusters and help to build the basis sets of ADK structures were then used to determine populations of AdK conformations in real solution experiments.
Reference


conformational ensembles in biomolecular recognition. Nature Chemical Biology, 5(11), 789-796. doi:10.1038/nchembio.232


