PARAMETER OPTIMIZATION ON THE CONVERGENCE SURFACE OF PATH SIMULATIONS

SRINIVAS NIRANJ CHANDRASEKARAN

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry and Biophysics in the School of Medicine

Chapel Hill
2016

Approved by:

Sharon L. Campbell

Charles W. Carter Jr.

Nikolay V. Dokholyan

Jan Hermans

Qi Zhang

ABSTRACT

Srinivas Niranj Chandrasekaran: Parameter optimization on the convergence surface of PATH simulations
(Under the supervision of Charles W. Carter Jr.)

Computational treatments of protein conformational changes tend to focus on the trajectories themselves, despite the fact that it is the transition state structures that contain information about the barriers that impose multi-state behavior. PATH is an algorithm that computes a transition pathway between two protein crystal structures, along with the transition state structure, by minimizing the Onsager-Machlup action functional. It is rapid but depends on several unknown input parameters whose range of different values can potentially generate different transition-state structures. Transition-state structures arising from different input parameters cannot be uniquely compared with those generated by other methods. I outline modifications that I have made to the PATH algorithm that estimates these input parameters in a manner that circumvents these difficulties, and describe two complementary tests that validate the transition-state structures found by the PATH algorithm. First, I show that although the PATH algorithm and two other approaches to computing transition pathways produce different low-energy structures connecting the initial and final ground-states with the transition state, all three methods agree closely on the configurations of their transition states. Second, I show that the PATH transition states are close to the saddle points of free-energy surfaces connecting initial and final states generated by replica-exchange Discrete Molecular Dynamics simulations. I show that aromatic side-chain rearrangements create similar potential energy barriers in the transition-state structures identified by PATH for a signaling protein, a contractile protein, and an enzyme. Finally, I observed, but cannot account for, the fact that trajectories obtained for all-

atom and $C\alpha$-only simulations identify transition state structures in which the $C\alpha$ atoms are in essentially the same positions. The consistency between transition-state structures derived by different algorithms for unrelated protein systems argues that although functionally important protein conformational change trajectories are to a degree stochastic, they nonetheless pass through a well-defined transition state whose detailed structural properties can rapidly be identified using PATH. In the end, I outline the strategies that could enhance the efficiency and applicability of PATH.

To my parents; my wife

ACKNOWLEGDEMENTS

Firstly, I would like to thank my mentor, Dr. Charles Carter, for encouraging me to pursue a challenging project and helping me at each step to overcome many hurdles. I would like to thank him for not losing patience with me, the tens of times I came up with a new solution to the parameter optimization problems, raising everyone's hopes only to came back the very next morning to tell him that method worked only for special cases. His perseverance and his enthusiasm for the project motivated me during times when the project was stuck and progress was hard to come by. As a mentor, he also provided me with many anecdotal advices, which have been invaluable in shaping my career.

I would also like to thank Dr. Jan Hermans for helping me think about PATH in a new way which was crucial in finding a solution to the optimization problem and in the development of the current version of PATH. He graciously spent several afternoons with me listening to and critiquing my ideas which were important for my development as a scientist.

I would like to thank my committee members Drs. Nikolay V. Dokholyan, Sharon Campbell and Qi Zhang. I collaborated with Dr. Dokholyan for validating the PATH results, during which I had the opportunity of interact with him several times which were instrumental in figuring out the correct approach to test PATH. Dr. Campbell guided me when I was not able to find a suitable lab during my lab rotation and helped me join Dr. Carter's lab. With Dr. Zhang I have had several fruitful discussions during my committee meetings for which I am thankful.

I am also thankful to Dr. Tishan Williams for her wonderful ability to simplify complex subjects and providing me the viewpoint of an experimentalist during group meetings, which has helped me several times to rethink my approach to solving the PATH equations.

Finally, I would like to thank my family. My parents have always been supportive and have stood by my side in every academic decision that I have made. They have always encouraged me to pursue by goals and choose the career that I was interested in. Last of all, I would to thank my wife, Priya, who has made the pursuit of science an enjoyable experience. It is great to be able to discuss my work with her, as she is a fellow scientist, and she has elevated science from being just a career to a way of life.

PREFACE

Part of the work described here was published in the journal *Structural Dynamics*:

Chandrasekaran, S. N., Das, J, Dokholyan, N. V., & Carter C. W. (2016). **A modified PATH algorithm rapidly generates transition states comparable to those found by other well established algorithms**. Structural Dynamics, 3(1), p.012101.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANM	Anisotropic Network Model
CHARMM	Chemistry at HARvard Molecular Mechanics
DMD	Discrete Molecular Dynamics
MD	Molecular Dynamics
NMA	Normal Mode Analysis
NMR	Nuclear Magnetic Resonance Spectroscopy
RMSD	Root Mean Squared Deviation
TrpRS	Tryptophanyl-tRNA Synthetase

CHAPTER 1: INTRODUCTION

Enzyme catalyzed reactions form an integral part of biology and are fundamentally important for replication of genetic material and for the survival of life. Hence enzyme catalyzed reactions have been studied in great detail, starting from simple catalysis of oligosaccharides by lysozyme (Chipman 1971) to complex reactions like mRNA translation by Ribosomes (Fluitt et al. 2007). In general, enzyme catalyzed reactions often take place in two steps, the fast chemical reaction step and the slower, rate-limiting protein conformational change step (Watt et al. 2007). The former is well understood for many enzyme catalyzed reactions because the transition states of the reaction are well characterized. One of the common ways to study these transition states is to arrest the reaction at the transition state using a transition state analog (Secemski et al. 1972). This is possible because the structure of the chemical species at the transition state is well known. Also, the chemical reaction is a localized phenomenon, that is, the parts of the protein that are involved in the chemical reaction are in and around the reactants in the binding pocket

Characterization of the other step of enzyme catalyzed reactions, the conformational transition of the protein, is more difficult because it is not a localized phenomenon. Many large scale transition events in the protein are brought about by allosteric effects where a change in one part of the protein affects another part, mostly the active site (Weinreb et al. 2012). Hence the approach of designing transition state analogs to characterize conformational transition states is not straightforward.

1.1 Conformational transition states impose multistate behavior in proteins

Conformational transition states are the most energetic structures along the conformational change pathways of proteins. Understanding the nature of these transition states is important as they hold information regarding what causes proteins to exist in multiple states (Kapustina et al. 2007). The multistate behavior of proteins is fundamental to life as it provides a way to generate states with a free energy differential and the transition between such states happens as a response to stimulus. Such conformational transitions can act as molecular timers to help regulate amplitude and duration of cellular processes (Nicholson & Lu 2007), significantly enhancing function by creating the capacity for a protein to transmit time and ligand-dependent information and/or mechanical motion necessary for signaling and other free-energy transduction processes. Structures of conformational transition states should therefore reveal valuable information about the energy barriers that separate one equilibrium structure from another.

1.2 Identification of conformational transition states by computational methods

Traditional experimental methods that are used for determination of macromolecular structure, like NMR and X-ray crystallography, cannot be used to identify the structures of conformational transition states, due to their fleeting existence. Hence computational methods have to be used to identify and characterize conformational transition states.

Molecular Dynamics (MD) simulations are the most commonly used computational methods to study the time evolution of macromolecular structures. There are several well-established force fields and algorithms, such as, GROMACS (Lindahl et al. 2001; Hess et al. 2008), CHARMM (Brooks et al. 1983; Brooks et al. 2009), AMBER (Cornell et al. 1995) and NAMD (Phillips et al. 2005), that are the most widely used tools for performing molecular dynamics simulations. MD simulations have been successful in studying domain motions of large proteins (Gumbart et al. 2009) and have also been an invaluable tool in

studying folding pathways of smaller proteins (Shaw et al. 2010). But to identify conformational transition states MD simulations are inefficient tools, because, the protein conformational changes occur on the timescale of milliseconds and it is difficult to simulate large proteins for that time. Also as conformational changes are very rare compared to the rest of the time the protein spends at the equilibrium state identifying these conformational transition states in a statistically significant ensemble would require several transitions between the equilibrium states, which is really difficult to simulate.

In spite of these problems, MD simulations can still provide useful information about protein conformational changes and transition states when coupled with sampling algorithms. One such popular algorithm is Steered Molecular Dynamics (Baker et al. 2013) where a force is applied to a part of a protein so that the transition from one state to another is induced. This speeds up the simulation and the different conformations can be sampled more rapidly. Another such algorithm is the Umbrella Sampling method (Torrie & Valleau 1977) where the energy barrier separating the two conformations is flattened such that the two states are sampled. Another recent, but popular, method is the replica exchange algorithm (Sugita & Okamoto 1999) which provides comprehensive mapping of the conformational free energy landscape. The replica exchange algorithm efficiently searches the configuration space of proteins by overcoming the sampling problem that affects single temperature simulations, which is that, at low temperatures the structures do not have enough energy to overcome conformational barriers and at high temperatures, the structures are unfolded and are far from the equilibrium states. In replica exchange simulations, multiple replicas of the starting structure are simulated at different temperatures and at defined time intervals, structures at different temperatures are exchanged. By doing this, replica exchange simulations allow systems to explore structures at different temperatures, thus sampling the conformational landscape, quickly and efficiently.

There are other molecular dynamics simulation algorithms that increase the speed of simulations by different kinds of simplification like Discrete Molecular Dynamics (DMD) (Ding et al. 2008; Dokholyan et al. 1998; Shirvanyants et al. 2012). DMD uses a step-wise potential energy function instead of the smooth functions that other force fields use, which decreases the frequency at which the force has to be computed. This also means that the forces are computed not at regular intervals of time but is event based, that is, the force is computed when an event occurs. Hence DMD speeds up molecular dynamics simulations and with the help of a sampling algorithm (Williams II et al. 2015), can rapidly generate conformational landscapes.

Even though there are good sampling algorithms and quick MD simulations techniques, they are still time and resource intensive. Some dedicated algorithms (Fujisaki et al. 2010) sample transition paths in the neighborhood of the most probable pathways between two equilibrium states, thereby not requiring massive computational resources. Once such algorithm is the String method (E et al. 2002b; Ovchinnikov et al. 2011) which furnishes an analytical algorithm for mapping the most probable path through conformational free energy landscapes using intervals between nodes defined in terms of collective variables along the path. It describes the transition pathway as the curve that connects successive metastable states so as to maintain a tangential projection of the curvature of the collective variables with respect to Cartesian space onto the free energy surface defined by the collective variables. Using collective variables reduces the number of degrees of freedom over which MD simulations are required. The progress between successive states is monitored in the String method with the help of a reaction coordinate called the committor function, which is the fraction of molecules that complete the trajectory from each node. The transition state along a trajectory between the two equilibrium states is achieved when the committor function reaches a value of 0.5. The all-atom CHARMM potential and the analytical formulation of the gradient mean that the string method can be

considered to be the gold standard in the field. In spite of the success of the string method, it is nevertheless resource intensive.

Many functional conformational changes are distinct from protein folding reactions in that they entail primarily large amplitude motions that are independent of individual covalent bond vibration. Often, these conformational changes are rigid-body motions that can be replicated by the superposition of a few large amplitude normal modes. Numerous algorithms have been introduced to exploit Elastic Network Models (ENM) (Bahar et al. 1997) in the computation of conformational change trajectories like Plastic Network Model (Maragakis & Karplus 2005) and adaptive Anisotropic Network Model (aANM) (Yang et al. 2009). In PNM, the conformational change trajectory is computed by minimizing the path integral of a free energy functional corresponding to the action or “resistance” along the path. This computed trajectory is a minimum energy path and is a differentiable curve through the centers of a smooth tube in pathspace containing the most probable paths (Durr & Bach 1978; Pinski & Stuart 2010). In aANM, the trajectory is calculated by an iterative method in which the intermediate state between the two equilibrium states are identified along the distance vector connecting the initial and the final states. Another related algorithm is ANMPathway (Das et al. 2014) which uses an Anisotropic Network Model (ANM) (Atilgan et al. 2001) to describe the potential energy wells of the two equilibrium states. In this method, the two structures are linearly extrapolated such that the energy of the two intermediates states, relative to the end states are equal. These two intermediate states now define a cusp hyperspace on which the lowest energy structure is identified by energy minimization, which is the transition state. Then by steepest descent energy minimization the trajectory from the transition state, back to the equilibrium states is computed.

Curiously, despite the relative importance of conformational transition states, few, if any of the computational studies on conformational changes to date have focused on the

transition state structures. It can be argued that in many ways transition-state structures, not the exact path, may be what are most important about conformational transitions. Hence my project was on the investigation of the possibility that the simplified potentials may furnish a sufficient basis set to identify valid transition state structures for large domain motions. Thus, whereas most treatments focus on the trajectories; I focused on the transition states themselves because they contain information about the barriers that impose multi-state behavior on proteins.

1.3 PATH rapidly computes the most likely path and transition state

PATH (formerly MinActionPath (Franklin et al. 2007)) is an algorithm that rapidly computes conformational transition states and the associated trajectories by minimizing the Onsager-Machlup (OM) functional (Onsager & Machlup 1953). The probability of finding a stochastic system at a given position and time is given by the Fokker-Planck equation. The OM functional is derived from the solution to the Fokker-Planck equation (Onsager & Machlup 1953) such that its minimization by a variational computation, implemented using the Euler-Lagrange equations, furnishes equations of motion describing the most probable path.

PATH defines the structures of equilibrium states using a linearized ANM potential. This approximation of the complex potential energy landscape works because most protein conformational changes are small displacements from the equilibrium states. PATH uses either all atom or more limited ANM models to identify the transition state. Then, it computes paths to and from that transition state using the OM equations of motion. PATH is an efficient algorithm for identifying the transition state and also the conformational change pathway that passes through it.

A comparison all the computational methods discussed above is summarized in Appendix 1

1.4 Minimum action pathways depend on several input parameters

PATH models the two equilibrium structures, between which the path has to be computed, as harmonic potential wells and the point of intersection of the two wells as the transition state. The shapes of the harmonic wells are defined by force constants k_l and k_r for the left and the right potential wells respectively (Fig. 1)

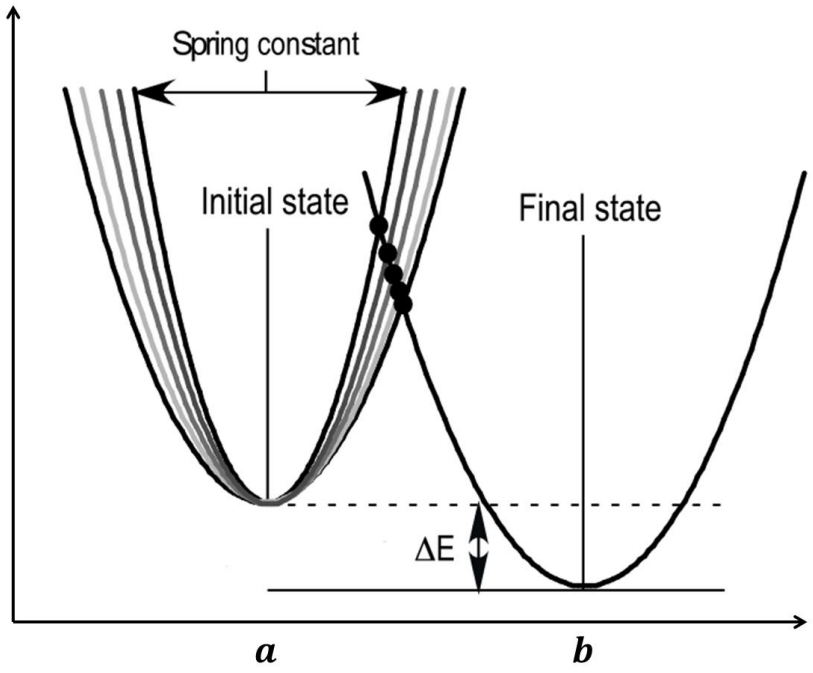

Figure 1. The double well: The two states of the protein between which the transitions are studied, can be approximated by the double well system. a and b are the two equilibrium states which are separated by an energy barrier. ΔE is the energy difference between the two minima. The width of the well is given by the value of the force (spring) constants, the narrower the well, the greater the value of the force constants.

The two structures are input crystal structures, a and b, and the force constants are calculated from the second derivative of the potential, called the Hessian Matrix. At the point of intersection of the two wells, which is the transition state $\bar{x}$, the two wells have the same energy $U^{\ddagger}$. If the total time taken to make the full transition is considered to be t_f, then the time taken to

reach the transition state from the initial state, $\bar{t}$, is a fraction of the total time and it uniquely identifies each minimum action path at that t_f.

From Fig. 1, it can be seen that if either force constant, k_l or k_r, the relative energy difference between the two wells (ΔE), or t_f are changed, then the minimum action path that the system will take would be different. This means that for different values of ΔE and t_f and as noted previously (Pinski & Stuart 2010) there are different minimum action paths between the given equilibrium states, each defined by a different $\bar{t}$. As previously mentioned, since $\bar{t}$ uniquely identifies each path, when plotted against different values of ΔE and t_f it gives rise to the surface that I call the convergence surface (Fig. 2).

This surface represents all the possible minimum action trajectories between a given pair of structures and it is different for different pairs of structures. This surface also shows the bi-sigmoidal behavior of $\bar{t}$ with respect to both ΔE and t_f. This surface also means that multiple, locally minimum action paths are possible for the same pair of structures. Appropriate values of both ΔE and t_f must therefore be chosen to identify a single minimum action path and transition state that is closest to what is observed in nature.

As mentioned earlier, the force constants are calculated from the Hessian matrix, which is built using a scale constant that is obtained by fitting crystallographic B values to the mean-square fluctuations of atoms in the structure (Bahar et al. 1997). Hence their accuracy depends strongly on the resolution of the X-ray data. This restriction appears to limit the application of PATH to high resolution crystal structures. Alternately, force constants can, in principle be determined iteratively by perturbative methods. Parameter estimation can thus require tens of simulations, compromising on the relative speed of PATH simulations. An alternate method to calculate the force constants must be used to improve the applicability of PATH to that of a general method for studying protein conformational changes.

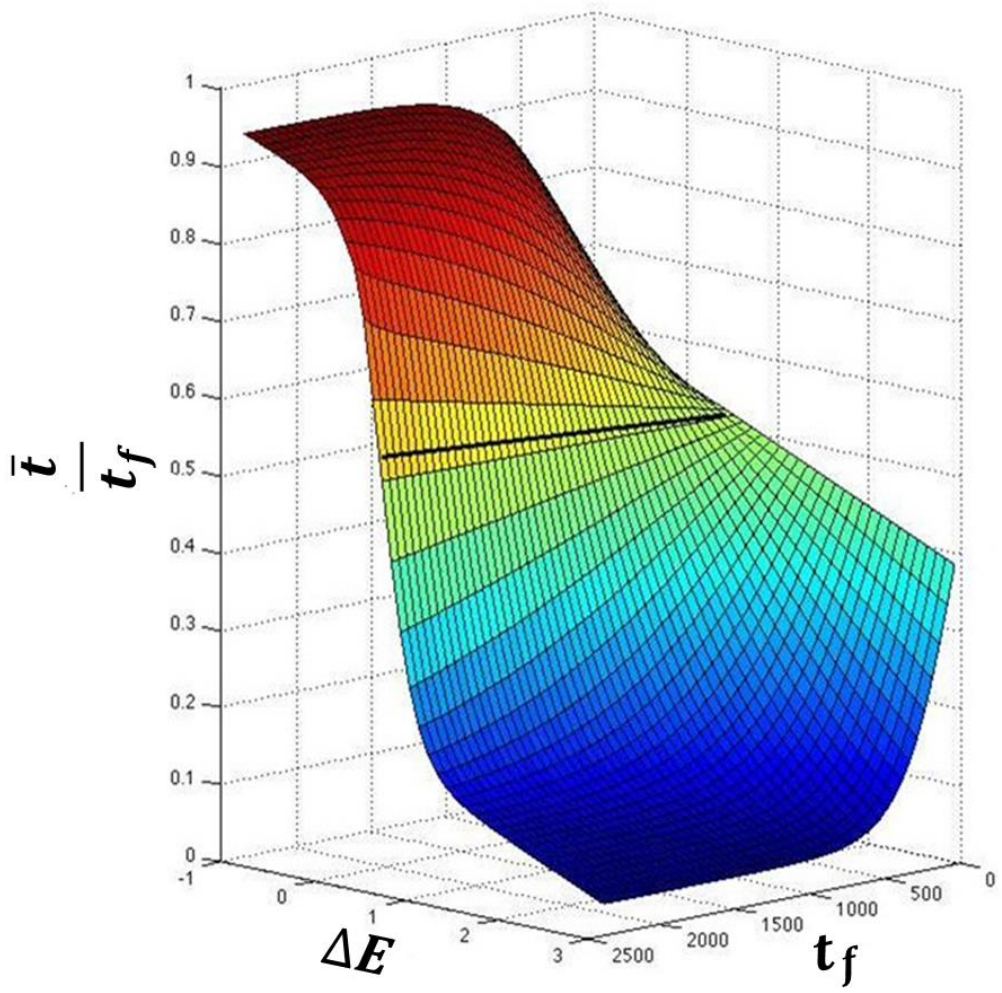

Figure 2. The convergence surface: From Fig. 1, it can be seen that the path must depend on both ΔE and t_f. Since $\bar{t}$, at each value of t_f, uniquely identifies a path as a function of ΔE, it gives rise to the convergence surface shown in this figure. The surface was obtained from simulations of the catalytic step of Tryptophanyl-tRNA synthetase and fitted to an empirical equation ($R^2 = 0.99$). The surface shows a sigmoidal dependence of $\bar{t}$ on both ΔE and t_f. Since only positive values of t_f are used in the simulations, only the lower half of the sigmoid is seen along the t_f axis and it can be fitted approximately to a rectangular hyperbola.

CHAPTER 2: THEORY OF PATH

2.1 Protein conformational change as a stochastic process

Protein conformational changes are diffusive processes, which can be modeled using an overdamped Langevin equation. In the overdamped regime there is no acceleration which means that the energy gained by interaction with the random force, is also lost quickly due to friction.

In one dimension, the Langevin equation can be written as

$$m\gamma\dot{x} = -\frac{\mathrm{d}V}{dx} + \sqrt{2m\gamma k_B T}\xi \tag{1}$$

where, γ is the diffusion coefficient, $-\frac{dV}{dx}$ is the force that causes the drift and ξ is a delta-correlated Gaussian random force, with zero mean. That is,

$$\langle \xi(t) \rangle = 0 \tag{2}$$

$$\langle \xi(t)\xi(t') \rangle = \delta(t - t') \tag{3}$$

In the case of proteins, the drift force arises from the interatomic interactions which are modeled in PATH as linearized Anisotropic Network Model (ANM). The potential in ANM is written as

$$V = \frac{k}{2}(x - a)^2 \tag{4}$$

where, k is the force constant and a is the equilibrium state. The potential is quadratic in nature and displacement away from the equilibrium position results in a force that restores structure to the equilibrium state. Using the ANM potential the Langevin equation becomes

$$m\gamma\dot{x} = -k(x-a) + \sqrt{2m\gamma k_B T}\xi \tag{5}$$

Because of the stochastic nature of the Langevin equation, only the probability of the states that the protein could be in, can be computed, which is in contrast to the ballistic equations that gives deterministic paths. The probability of these states can be calculated using an alternate form of the Langevin equation called the Fokker-Planck equation.

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(-k(x-a)p(x,t)\right) - \frac{1}{2}\frac{\partial^2}{\partial x}\left(\frac{2k_B\,T}{m\gamma}\,p(x,t)\right)$$

Then the probability of the protein to reach state x_2 at time t_2 given that that the system was at state x_1 at time t_1 can be written as

$$p(x_2,t_2|x_1,t_1) = \frac{e^{-\frac{k}{4k_BT}\left((x_1-a)-(x_2-a)e^{\frac{k\Delta t}{m\gamma}}\right)^2\left(-1+\coth\left(\frac{k\Delta t}{m\gamma}\right)\right)}}{\left(\frac{1}{4\pi k_B T}\left(k\left(1+\coth\left(\frac{k\Delta t}{m\gamma}\right)\right)\right)\right)^{-\frac{1}{2}}} \tag{6}$$

where, k_B is the Boltzmann constant, and T is the temperature. This is a solution to the Fokker-Planck equation.

If the total path is a succession of such states, then the joint probability can be calculated as the product of the probabilities of the individual segments. By this method, the probability of each path that goes from one state of the protein to another can be calculated. Since in the case of PATH, only the most probable pathway is of interest, it can be calculated by minimizing the exponent of (6). To do this Onsager and Machlup came up an ingenious method (Onsager & Machlup 1953), where they were able to derive the equation of motion for the most probable path by writing (6) as

$$p \propto e^{-\frac{S_{OM}}{2mk_BT}} \tag{7}$$

where, S_{OM} is an integral called the Onsager-Machlup action functional and it is of the form

$$S_{OM} = \frac{1}{2\gamma}\int_0^t \left(m\gamma\dot{x} + k(x-a)\right)^2 dt \tag{8}$$

2.2 Equation of motion of the most probable path

To understand the derivation of the resulting equation of motion for the most probable path from Onsager-Machlup action, it is a useful exercise to compare and calculate the ballistic equations of motion using classical action.

2.2.1 Classical Action

Figure 3. Toy model: The two atoms of a diatomic system interact with each other and this interaction is modeled as a one-dimensional spring that follows Hook's law, $F = -kx$, where k is the force constant and x is the displacement from the mean position.

Consider a 1D diatomic system (Fig. 3) following Newtonian dynamics in a single potential well. Let the interaction between the two atoms be modeled by a 1D Hookean spring. On the basis of the principle of least action, the equation of motion can be derived by identifying the path that minimizes action. Classical action of a path is defined as the sum over the Lagrangian at every time instant. It has the mathematical form

$$S_{cl} = \int_0^t L.\,dt \tag{9}$$

The Lagrangian is the difference between the potential energy and the kinetic energy of the system. Hence the action can be rewritten as

$$S_{cl} = \int_0^t (T - V)dt \tag{10}$$

Hence to compute the action of any path and to identify the most probable path, it is required to know the kinetic and potential energies of the system.

Since the potential energy is described by a Hookean spring, it is written as equation (4)

$$V = \frac{k}{2}(x - a)^2$$

As the kinetic energy is written as $T = \frac{1}{2}m\dot{x}^2$, the Lagrangian can be written as

$$L = T - V = \frac{1}{2}(m\dot{x}^2 - k(x - a)^2) \tag{11}$$

Using (11), (10) can be written as

$$S_{cl} = \int_0^t \frac{1}{2}(m\dot{x}^2 - k(x - a)^2)dt \tag{12}$$

Equation (12) computes the action of any given path but the path of minimum action can be identified by calculating the extremum of the action functional. In Lagrangian mechanics (using the Lagrangian to derive Newton's equation of motion), this boils down to finding the solution to the Euler-Lagrange equation

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) = 0 \tag{13}$$

The solution to the Euler-Lagrange equation is the path of least action.

On applying the boundary conditions, B_1 and B_2, which are, at time t_1, $x(t_1) = x_1$ and at t_2, $x(t_2) = x_2$, the solution to the Euler-Lagrange equation gives the following equation of motion

$$x(t) = a + \frac{1}{\sin(\omega(t_2 - t_1)}\Big((x_1 - a)\sin\big(\omega(t_2 - t)\big) - (x_2 - a)\sin\big(\omega(t_1 - t)\big)\Big) \tag{14}$$

where, $\omega = \sqrt{\frac{k}{m}}$ is the angular frequency. This is the equation of motion of a spring following Newtonian dynamics which also minimizes classical action with boundary conditions B_1 and B_2.

From (14) the velocity of the system can be calculated as

$$\dot{x}(t) = \frac{1}{\sin(\omega(t_2 - t_1)}\big(-\omega(x_1 - a)\cos\big(\omega(t_2 - t)\big) + \omega(x_2 - a)\cos\big(\omega(t_1 - t)\big)\big) \quad (15)$$

Using (14) and (15), the classical action for a 1D diatomic system can be written as

$$S_{cl} = \frac{m\omega}{2\sin\big(\omega(t_2 - t_1)\big)}\Big(((x_1 - a)^2 + (x_2 - a)^2)\cos\big(\omega(t_2 - t_1)\big) - 2(x_1 - a)(x_2 - a)\Big) \quad (16)$$

2.2.2 Onsager-Machlup equations of motion

Using the same approach outlined for deriving the equation of motion for the classical system, the Onsager-Machlup equations of motion for the most probable path can be derived from equation (8)

$$S_{OM} = \frac{1}{2\gamma}\int_0^t \big(m\gamma\dot{x} + k(x - a)\big)^2 dt$$

On solving the Euler-Lagrange equation and applying the same boundary conditions, B_1 and B_2, the trajectory equation can be written as

$$x(t) = a + \frac{1}{\sinh\big(\Gamma(t_2 - t_1)\big)}\Big((x_1 - a)\sinh\big(\Gamma(t_2 - t)\big) - (x_2 - a)\sinh\big(\Gamma(t_1 - t)\big)\Big) \quad (17)$$

where, $\Gamma = \frac{k}{m\gamma}$. This is the equation of motion of the minimum action path undergoing stochastic dynamics, modeled by Langevin equation.

From (17) the velocity equation can be calculated as

$$\dot{x}(t) = \frac{1}{\sinh\big(\Gamma(t_2 - t_1)\big)}\big(-\Gamma(x_1 - a)\cosh\big(\Gamma(t_2 - t)\big) + \Gamma(x_2 - a)\cosh\big(\Gamma(t_1 - t)\big)\big) \quad (18)$$

Using (17) and (18) the Onsager-Machlup action for the diatomic system be written as

$$S_{OM} = \frac{mk}{2\sinh\big(\Gamma(t_2 - t_1)\big)}\Big(\big((x_2 - a)^2 e^{\Gamma(t_2 - t_1)} + (x_1 - a)^2 e^{-\Gamma(t_2 - t_1)}\big) - 2(x_2 - a)(x_1 - a)\Big) \quad (19)$$

2.3 Using PATH for a two state system

In the previous sections the Onsager-Machlup equations of motion were derived for a 1D system in one state, that is, the potential is defined by a single potential energy well. To study protein conformational changes, the trajectory of transition from one state to another has to be computed. In the case of PATH, this is done by considering that each state is defined by a different potential energy well and the transition from one well to another occurs at the intersection of the two wells. Since the equations are easier to understand for a 1D diatomic system, I will describe the PATH algorithm for a 1D diatomic system and then extend the equations to a 3D N atom system, where N is the number of atoms in the protein of interest.

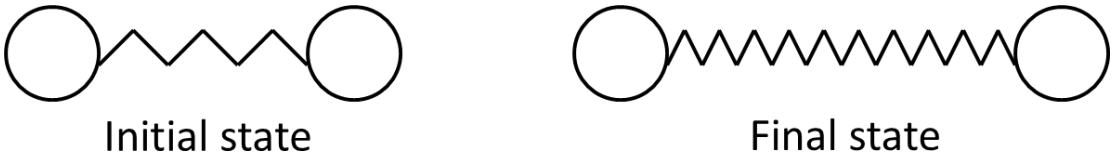

Figure 4. Two state of the 1D system: In a double well system, each state is represented by a different potential well. The width of the well is determined by the force constants and the equilibrium states are the minima of these wells. The force constants also determine the strength of interaction between the two atoms, along with the interatomic distance of separation.

2.3.1 1D diatomic system

Consider a double well system, where each well represents a different state of the diatomic systems. Let the two equilibrium states be a and b. Then the Onsager-Machlup equation of motion within each well is written as

$$x_l(t) = a + (\bar{x} - a)\left(\frac{\sinh(k_l t)}{\sinh(k_l \bar{t})}\right) \; when \; t \; \leq \bar{t} \tag{20}$$

$$x_r(t) = b + (b - \bar{x})\left(\frac{\sinh\left(k_r(t - t_f)\right)}{\sinh\left(k_r(t_f - \bar{t})\right.}\right) \; when \; t \; > \bar{t} \tag{21}$$

where, $\bar{t}$ is the time taken to reach the transition state from a, t_f is the total time for transition from a to b, $\bar{x}$ is the transition state structure, k_l and k_r are the force constants for the initial and the final states, respectively.

For a smooth transition from one well to the other, the paths have to satisfy boundary conditions based on position, velocity and energy at the transition state. These conditions can be expressed mathematically in the following way:

$$\begin{aligned} x_l(t \to \bar{t}) &= x_r\left(\bar{t} - t_f\right) \\ \dot{x}_l(t \to \bar{t}) &= \dot{x}_r\left(\bar{t} - t_f\right) \\ \frac{1}{2}(\bar{x} - a)^2 + \Delta E &= \frac{1}{2}(\bar{x} - b)^2 \end{aligned} \tag{22}$$

Also, $x_l(0) = a$, $x_r\left(t_f\right) = b$ and $x(\bar{t}) = \bar{x}$, where x_l and x_r are the trajectories in the left and right well, respectively and ΔE is the potential energy offset between the minima of the two energy wells.

Using (20) and (21) the velocity continuity equation in (22) can be written as

$$k_r(\bar{x} - b)\coth\left(k_r\left(\bar{t} - t_f\right)\right) = k_l(\bar{x} - a)\coth(k_l \bar{t}) \tag{23}$$

The above equation relates $\bar{t}$, t_f and ΔE. But to make the relationship more explicit, it is necessary to compute the structure of the transition state.

If the initial structure is $a = (a_1, a_2)$ and the final structure is $b = (b_1, b_2)$, the structure of the transition state can be calculated by writing the energy continuity equation in (22) as

$$\frac{1}{2}(\bar{x} - a)\begin{pmatrix} k_l & -k_l \\ -k_l & k_l \end{pmatrix}(\bar{x} - a)^T + \Delta E = \frac{1}{2}(\bar{x} - b)\begin{pmatrix} k_r & -k_r \\ -k_r & k_r \end{pmatrix}(\bar{x} - b)^T \tag{24}$$

The two matrices in the (23) are the Hessian matrices on the initial and final states, which are described in next section

(23) can be rewritten as

$$\frac{k_l}{2}(\bar{x} - a)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}(\bar{x} - a)^T + \Delta E = \frac{k_r}{2}(\bar{x} - b)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}(\bar{x} - b)^T \tag{25}$$

On simplification, the above equation becomes

$$\frac{k_l}{2}\big((x_1 - x_2) - (a_1 - a_2)\big)^2 + \Delta E = \frac{k_r}{2}\big((x_1 - x_2) - (b_1 - b_2)\big)^2 \tag{26}$$

Substituting $\bar{X} = x_1 - x_2$, $A = a_1 - a_2$ and $B = b_1 - b_2$, (23) becomes

$$\frac{k_l}{2}(\bar{X} - A)^2 + \Delta E = \frac{k_r}{2}(\bar{X} - B)^2 \tag{27}$$

Solving for $\bar{X}$,

$$\bar{X} = \frac{(k_l A - k_r B) - (A - B)\sqrt{k_r k_l + \frac{2\Delta E(k_r - k_l)}{(B - A)^2}}}{(k_l - k_r)} \tag{28}$$

For a diatomic system centered on the origin, $x_1 + x_2 = 0$, giving, together with $\bar{X}$, the transition state $\bar{x}$.

On substituting $\bar{x}$ and rearranging, the velocity continuity equation (23) can be rewritten as

$$\frac{\sinh\left(\lambda_r t_f - (\lambda_r + \lambda_l)\bar{t}\right)}{\sinh\left(\lambda_r t_f - (\lambda_r - \lambda_l)\bar{t}\right)} = \left(\frac{\lambda_r + \lambda_l}{\lambda_r - \lambda_l}\right) - \left(\frac{2\lambda_r\lambda_l}{Z_{\Delta E}(\lambda_r - \lambda_l)}\right) \tag{29}$$

where, $Z_{\Delta E} = \sqrt{\lambda_r\lambda_l + \frac{2\Delta E(\lambda_r - \lambda_l)}{(B-A)^2}}$, and λ_r and λ_l correspond to eigenvalues of the respective Hessian matrices.

In any spring system, in one dimension, the overall motion is comprised of N independent modes, each with its own force constant. In the case of the diatomic system, there is one translational mode, whose force constant is zero and one vibrational mode. Since each mode behaves independently from the other, the spring constant associated with each mode is calculated from the eigenvalues of the respective Hessian matrices. Since t_f is known, $\bar{t}$ of the 1D diatomic system can calculated, numerically. Thus, the entire landscape of path trajectories shown in Fig. 2 can be computed from equation (29). This equation also describes the bi-sigmoidal behavior of the convergence surface. For the values of t_f, $\bar{t}$ has a sigmoidal relationship to ΔE. Similarly, at constant ΔE, $\bar{t}$ has a sigmoidal relationship to t_f, though the shape of the curve in Fig. 2 is that of a rectangular hyperbola. This behavior rises from the use of positive values of t_f, as negative values of t_f are meaningless.

Though equation (29) cannot be solved analytically, by calculating the structure of the transition state using (28), the $\bar{t}$ values at different t_f and ΔE can be computed numerically. Once the $\bar{t}$ is known, using (20) and (21) the most probable path connecting the two minima, passing through the transition state can be computed.

As PATH is based on stochastic dynamics, it could be argued that the velocity continuity equation, which forms the basis of PATH, is meaningless because velocity continuity imposes conservation of momentum at the transition state. Momentum is not conserved in stochastic systems due to the random force and friction. But this does not affect the velocity continuity equation in PATH because, the equation of motion in PATH is that of a single

trajectory, the most probable PATH, which is a continuous function. Since the equation of motion of this most probable is similar to that of a ballistic equation, except for the sinh term, using velocity continuity to establish continuity in the trajectory as it transitions from one potential well into another, is meaningful.

2.3.2 3D N atom system

The 1D toy model described in the previous section is effective in deriving the equations that used in PATH to generate the most probable pathway between the minima of two harmonic potentials. But the equations and the approach aren't useful for any real world applications, especially for studying proteins because the proteins have more than two atoms and there are also in three dimensions. Though the equations of motion for a 3D multiatom system are similar to equations (20) and (21), it is not possible to derive a convergence surface equation and solve for $\bar{t}$ numerically. Hence a different approach has to be taken, as outlined below.

For a multiatom 3D system, the interactions between the atoms are more complex than in the case of the 1D diatomic system. PATH uses a linearized ANM potential to represent interatomic interactions where each atom pair is connected to each other in some manner via springs with a single force constant k. According to ANM (Atilgan et al. 2001), there is a pair potential between any two atoms, which is given by

$$U(r_i, r_j) = \frac{1}{2} k(r_{ij} - \bar{r}_{ij})^2 \tag{30}$$

where, r_i is the position of the i^{th} atom, r_j is the position of the j^{th} atom and $\bar{r}$ is the equilibrium distance between the two atoms.

A Hessian is a $3N \times 3N$ matrix of second derivatives that basically gives the curvature of a surface defined by a function, where N is the number of atoms. For example, the Hessian for a function F, relative to the variables x, y and z is written as

$$Hessian = \begin{pmatrix} \frac{\partial^2 F}{\partial x \partial x} & \frac{\partial^2 F}{\partial x \partial y} & \frac{\partial^2 F}{\partial x \partial z} \\ \frac{\partial^2 F}{\partial y \partial x} & \frac{\partial^2 F}{\partial y \partial y} & \frac{\partial^2 F}{\partial y \partial z} \\ \frac{\partial^2 F}{\partial z \partial x} & \frac{\partial^2 F}{\partial z \partial y} & \frac{\partial^2 F}{\partial z \partial z} \end{pmatrix} \quad (31)$$

In the case of the Hookean spring the interatomic interaction energy can be expressed using the Hessian as

$$U(r_i, r_j) = \frac{1}{2} \Delta X H \Delta X^T \quad (32)$$

where, ΔX is the difference vector $\big((x_1 - \bar{x}_1), (y_1 - \bar{y}_1), (z_1 - \bar{z}_1) \ldots (x_N - \bar{x}_N), (y_N - \bar{y}_N), (z_N - \bar{z}_N)\big)$

For small displacements about the equilibrium positon, the second derivative of the potential relative to the Cartesian coordinates can be calculated in the following way:

From (30), the potential is written as

$$U = \frac{k}{2}\left(r_{ij} - \bar{r}_{ij}\right)^2 \quad (33)$$

If $r_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2}$,

Then,

$$\frac{\partial V}{\partial x_i} = \frac{k}{2}.2.\left(r_{ij} - \bar{r}_{ij}\right).\frac{1}{2}.2\frac{(x_i - x_j)}{r_{ij}} \quad (34)$$

which can be rewritten as

$$\frac{\partial V}{\partial x_i} = k\left(r_{ij} - \bar{r}_{ij}\right)\frac{(x_i - x_j)}{r_{ij}} \quad (35)$$

Or

$$\frac{\partial V}{\partial x_i} = k\left(1 - \frac{\bar{r}_{ij}}{r_{ij}}\right)(x_i - x_j) \tag{36}$$

Then,

$$\frac{\partial}{\partial x_j}\left(\frac{\partial V}{\partial x_i}\right) = k\left(\left(1 - \frac{\bar{r}_{ij}}{r_{ij}}\right).1 + (x_i - x_j).1.\frac{\bar{r}_{ij}}{r_{ij}}.\frac{(x_i - x_j)}{r_{ij}}\right) \tag{37}$$

On simplification,

$$\frac{\partial^2 V}{\partial x_j \partial x_i} = k\left(1 - \frac{\bar{r}_{ij}}{r_{ij}}\left(1 - \frac{(x_i - x_j)^2}{r_{ij}^2}\right)\right) \tag{38}$$

Since $\bar{r}_{ij} = r_{ij}$ at equilibrium,

$$\frac{\partial^2 V}{\partial x_j \partial x_i} = k\left(\frac{(x_i - x_j)^2}{r_{ij}^2}\right) \tag{38}$$

Similarly, the other terms of the Hessian can be calculated and a single 3×3 block of the Hessian for a two atom interaction can be written as

$$h_{ij} = \frac{s}{\bar{r}_{ij}^2}\begin{pmatrix} (x_i - x_j)(x_i - x_j) & (x_i - x_j)(y_i - y_j) & (x_i - x_j)(z_i - z_j) \\ (y_i - y_j)(x_i - x_j) & (y_i - y_j)(y_i - y_j) & (y_i - y_j)(z_i - z_j) \\ (z_i - z_j)(x_i - x_j) & (z_i - z_j)(y_i - y_j) & (z_i - z_j)(z_i - z_j) \end{pmatrix} \tag{39}$$

This is the Hessian appropriate to the linearization of the spring connecting atom i with atom j, but there are many such connections in general. Here, s is a scale constant that is generally derived from fitting the mean square fluctuation of the atoms to the crystallographic B values (Bahar et al. 1997). For non-high resolution crystal structures and for computational mutants, the B values cannot be used to estimate the scale constants. Alternate methods have to be developed for evaluate the scale constants for such systems.

The three-by-three block in (39) can be used to build the full Hessian H. By referring to the 3×3 coordinates of the i^{th} atom as x_i in X (so that X has N such entries), and H_{ij} gives the

3×3 block of H at row i, column j, then the Hessian is constructed by adding h_{ij} from (39) to H_{ii} and H_{jj} and subtracting h_{ij} from H_{ij} and H_{ji}.

Using this Hessian, the Onsager-Machlup equation of motion can be written as

$$x_l(t) = V\left[\begin{pmatrix} \frac{t}{\bar{t}} & 0 \\ 0 & \frac{\sinh(\lambda_l^i t)}{\sinh(\lambda_l^i \bar{t})} \end{pmatrix}\bar{\psi}\right] + a \tag{40}$$

$$x_r(t) = W\left[\begin{pmatrix} \frac{t_f - t}{t_f - \bar{t}} & 0 \\ 0 & -\frac{\sinh\left(\lambda_r^i(t - t_f)\right)}{\sinh\left(\lambda_r^i(t_f - \bar{t})\right)} \end{pmatrix}\bar{\phi}\right] + b \tag{41}$$

where, $\bar{\psi} = V^T(\bar{x} - a)$, $\bar{\phi} = W^T(\bar{x} - b)$. V and W are the eigenvectors of the Hessian matrices of the initial and the final wells, and λ_l^i and λ_r^i are their eigenvalues. The eigenvalues replace the force constants in the trajectory equations because by diagonalizing the Hessian matrix, $3N$ normal modes are generated whose individual motion depends on the rate at which the structure changes, which is given by the eigenvalues. The final trajectory is generated by a linear combination of the normal modes.

Unlike the 1D system where the transition state structure is identified by solving the energy continuity equation, the transition state is computed in the 3D case from the velocity continuity equation as the latter is easier to solve for a 3D system.

From (40) and (41) the velocity continuity equation can be written as

$$V\left[\begin{pmatrix}\frac{1}{\bar{t}} & 0\\ 0 & \frac{\lambda_l^i\cosh(\lambda_l^i t)}{\sinh(\lambda_l^i \bar{t})}\end{pmatrix}\bar{\psi}\right] = W\left[\begin{pmatrix}\frac{1}{\bar{t}-t_f} & 0\\ 0 & \frac{\lambda_r^i\cosh\left(\lambda_r^i(\bar{t}-t_f)\right)}{\sinh\left(\lambda_r^i(t_f-\bar{t})\right)}\end{pmatrix}\bar{\phi}\right] \quad (42)$$

Considering the matrix on the left hand side of equation (42) to be L and the matrix on the right hand side to be R, equation (42) becomes

$$VLV^T(\bar{x}-a) = WRW^T(\bar{x}-b) \quad (43)$$

Equation (43) can be solved for $\bar{x}$ to get

$$\bar{x} = \frac{VLV^T a - WRW^T b}{VLV^T - WRW^T} \quad (43)$$

Equation (42) can be used to calculate the structure of the transition state at particular values of $\bar{t}$ and t_f. The following is the PATH algorithm to identify the transition state and then calculate the trajectory.

- For two equilibrium structures, the Hessian for the initial and final state can be computed using (39) if the scaling constant s is known from crystallographic B values (Bahar et al. 1997).
- The two Hessians are diagonalized to compute the eigenvalues and the eigenvectors for both the Hessians.
- For a given value of t_f, using equation (43) the structure of the transition state is identified for an assumed value of $\bar{t}$.
- Using this structure, the energy of the transition state is computed relative to both the equilibrium states to check for energy continuity. In the case of the 3D system, the energy continuity equation is written as

$$\frac{1}{2}(\bar{x}-a)H_l(\bar{x}-a)^T + \Delta E = \frac{1}{2}(\bar{x}-b)H_r(\bar{x}-b)^2 \quad (44)$$

- If $\bar{x}$ from (43) satisfies (44), then the transition state has been identified. If not, a different value of $\bar{t}$ is assumed and the process is repeated until the $\bar{x}$ from (43) satisfies (44)
- Once the transition state is identified, the most probable path connecting the two equilibrium states is computed from (40) and (41).

As this algorithm shows, to compute trajectories and identify conformational transition states with PATH, apart from the equilibrium states, it is required to know the scale constant to build the Hessian matrices, the values of t_f and ΔE. All these affect the most probable paths and are not easily determinable for all protein systems. The scale constants can be obtained only for high resolution crystal structures. But there are so many systems for which this is not possible. Also, there are computationally designed mutants or modified proteins which also lack information about the thermal fluctuation of the protein. These values can be determined by running molecular dynamics simulations but that adds another layer of complexity to determining transition states. It is even more difficult to determine t_f and ΔE, as the former is a time like parameter whose value depends on the scale constant while the latter is the difference in the relative potential energies of the two equilibrium state which is not readily available. Thus the lack of knowledge of the input parameters severely limits the scope of applicability of PATH.

CHAPTER 3: MODIFICATION OF PATH

PATH is a rapid algorithm for computing the most probable transition pathway between two equilibrium states. But for reasons outlined in last section, including the parameter optimization on the convergence surface, the applicability of PATH is limited. It can be used to obtain meaningful results only if the two equilibrium states are high resolution crystal structures and the time for transition and the potential energy difference between the two equilibrium states are known in advance, before the simulations can be performed. Therefore, the values of the four input parameters, k_l, k_r, t_f and ΔE have to be optimized every time a new system has to be simulated. From the convergence surface in Fig 2, it is clear that there is a relationship between ΔE and t_f.

Due to the relationship between ΔE and t_f as seen in the convergence surface and the relationship between force constants and t_f as seen from the equations of motion in PATH, in which, the product of the force constants and the time of transition determines the most probable pathway, I approached the problem of optimizing t_f first, as its value might be useful to evaluate the optimum values of the other two parameters.

3.1 Optimization of t_f

All the optimization studies were performed using the 1D diatomic system, unless it is explicitly mentioned. As the value of t_f can be any positive number, I split it into three regimes, small, intermediate and large.

3.1.1 Small values of t_f

Small values of t_f are those for which the ratio of the sinh terms in the equations of motion of PATH are such that

$$\frac{\sinh(kt)}{\sinh(k\bar{t})} = \frac{t}{\bar{t}} \tag{45}$$

because,

$$\lim_{x\to 0} \sinh(x) = x \tag{46}$$

In this regime, the convergence surface equation (29) can be solved for a special case where $\bar{t} = \frac{t_f}{2}$ to compute ΔE. This special case is important because, the ΔE value computed at this $\bar{t}$ is amount of energy that has to be given to the initial state of the system such that system spends equal amount of time in both the wells. This value of ΔE is henceforth referred to as $\Delta E_{0.5}$. This particular definition of $\Delta E_{0.5}$ is important because, the free energy difference between two conformations of a system can be written as

$$\Delta G_{conf} = -RT \ln K_{eq} \tag{47}$$

Where, K_{eq} is the equilibrium constant. When the equilibrium constant is 1, the system spends the same amount of time in both the conformations. In other words, there is 50% chance of finding the system in either conformation. Therefore, free energy can be defined the amount of energy that is added to one of the two conformations such that it spends equal time in both the wells. Since the equilibrium constant is related to the rate of conformational change and $\bar{t}$ is a time-like parameter in PATH, which is a reciprocal of the rate, $\Delta E_{0.5}$ can be considered to be indirectly related to ΔG_{conf}.

Solving the convergence surface equation (29) for the special case, the equation for $\Delta E_{0.5}$ is derived to be

$$\Delta E_{0.5} = \frac{(k_r - k_l)(b-a)^2}{8} \tag{48}$$

or

$$\Delta E_{0.5} = \frac{E_{right} - E_{left}}{4} \tag{48}$$

Where, E_{right} is the energy of the initial state relative to the potential energy function of the final states and E_{left} is the energy of the final state relative to the potential energy function of the initial state.

Though in the small t_f regime $\Delta E_{0.5}$, can be calculated, the trajectories that are generated are unrealistic. This is because the equation of motion in this regime is linear with respect to time. For example, the equation of motion in the left well becomes,

$$x_l(t) = a + (\bar{x} - a)\frac{t}{\bar{t}} \; when \; t \; \leq \bar{t} \tag{49}$$

This means that the trajectory is only a linear interpolation between the initial and final states. Also, there is no dependence on ΔE. This is demonstrated in Fig. 5, where the similarity of a state on the trajectory with respect to the initial state, Q_1, is plotted as a function of time. It can be seen that the line is almost a straight line connecting the initial and the final states. Also, trajectories at different values of ΔE are exactly the same which means that the trajectory is independent of the potential energy wells.

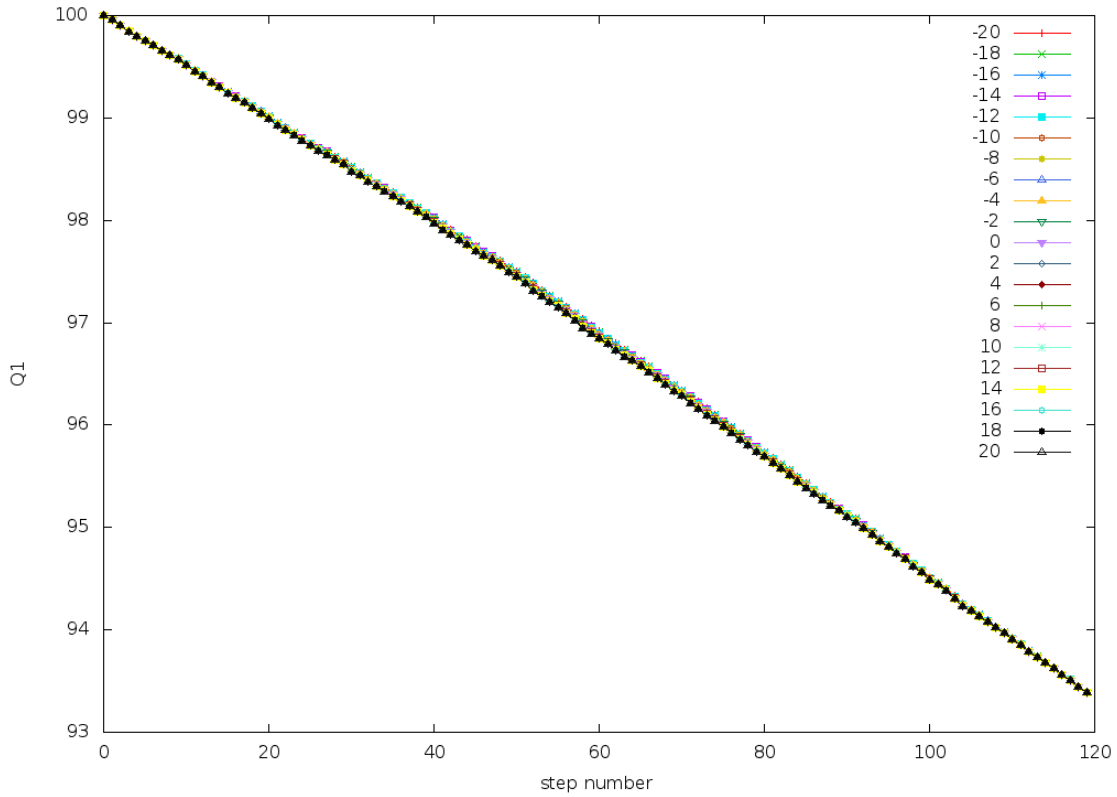

Figure 5. Q_1 vs. t at small t_f: Q_1 of a frame of the trajectory is the similarity of that frame to the initial state of the trajectory. These trajectories were generated by simulating the PreTS to Pdt transition of the TrpRS system (described later) at t_f=0.0003 at a range of ΔE values. The similarity metric increases almost linearly with respect to time. This figure shows that the trajectory is a linear extrapolation of the initial state. Also there is no dependence on ΔE.

This behavior at low values of t_f arises from the fact the system doesn't have enough time to undergo motion based on the potential and the equations of motion. Since the system is forced to reach the final state at t_f and yet not enough time to given, the system changes its conformation by a linear interpolation method. This linear interpolation is called *morphing*.

3.1.2 Intermediate values of t_f

The problem with small values of t_f doesn't affect the system in the intermediate regime as seen from the Q_1 plot (Fig. 6). But the problem with the intermediate regime is that it is hard

to define the boundaries of this regime. There is no special value of t_f which works better than others. But what is clear from the plot is that the dependence of the trajectory on the equations of motion and on ΔE reappears in this regime. The value of kt at which the ratio is the sinh terms is greater than the ratio of the time values is about 0.3. Since there is no special value of t_f, even though the trajectories are more reasonable than those in the small t_f regime, this regime is not particularly useful in identifying the optimum values of the PATH parameters.

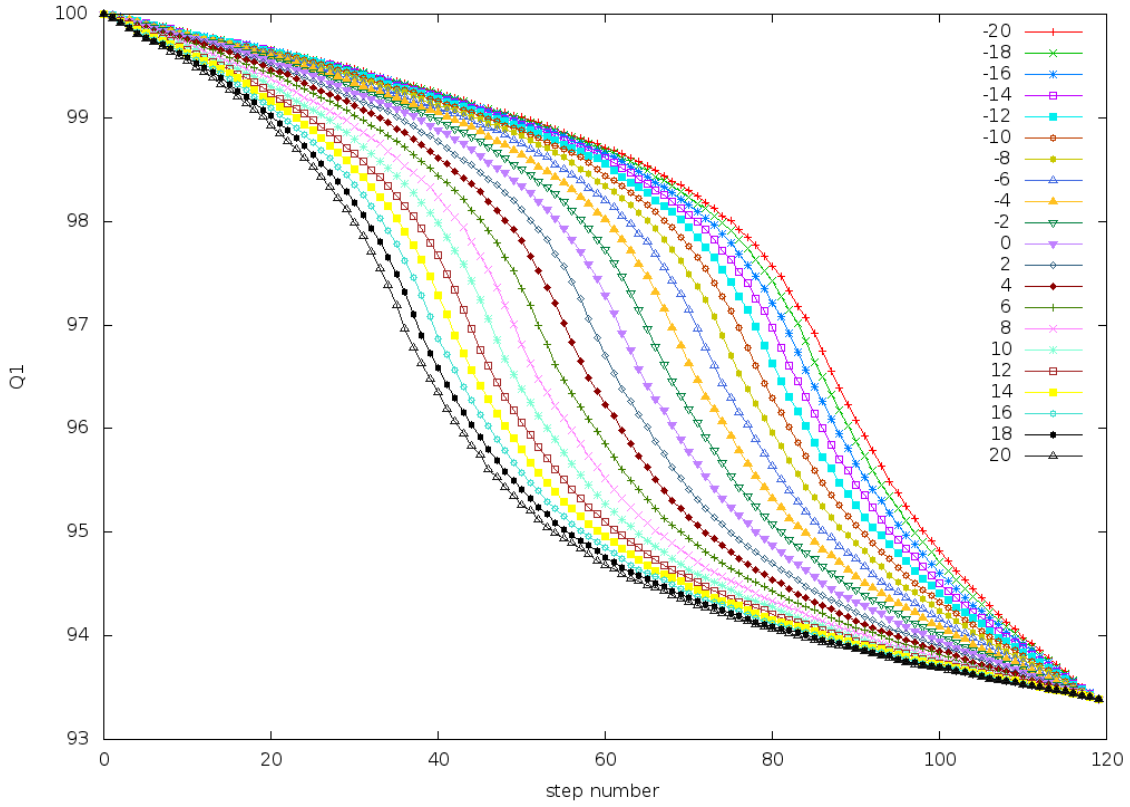

Figure 6. Q_1 vs. t at intermediate values of t_f: Q_1 of a frame of the trajectory is the similarity of that frame to the initial state of the trajectory. These trajectories were generated by simulating the PreTS to Pdt transition of the TrpRS system (described later) at t_f=30 at a range of ΔE values. The trajectories at intermediate regime show dependence on ΔE and the Q_1 values have a sigmoidal dependence on time. Tis behavior arises from the sinh terms in the trajectory equations.

3.1.3 Large values of t_f

As shown by Figs. 5 and 6, with increase in t_f the dependence on ΔE and on the sinh terms of the equations of motion reappear.

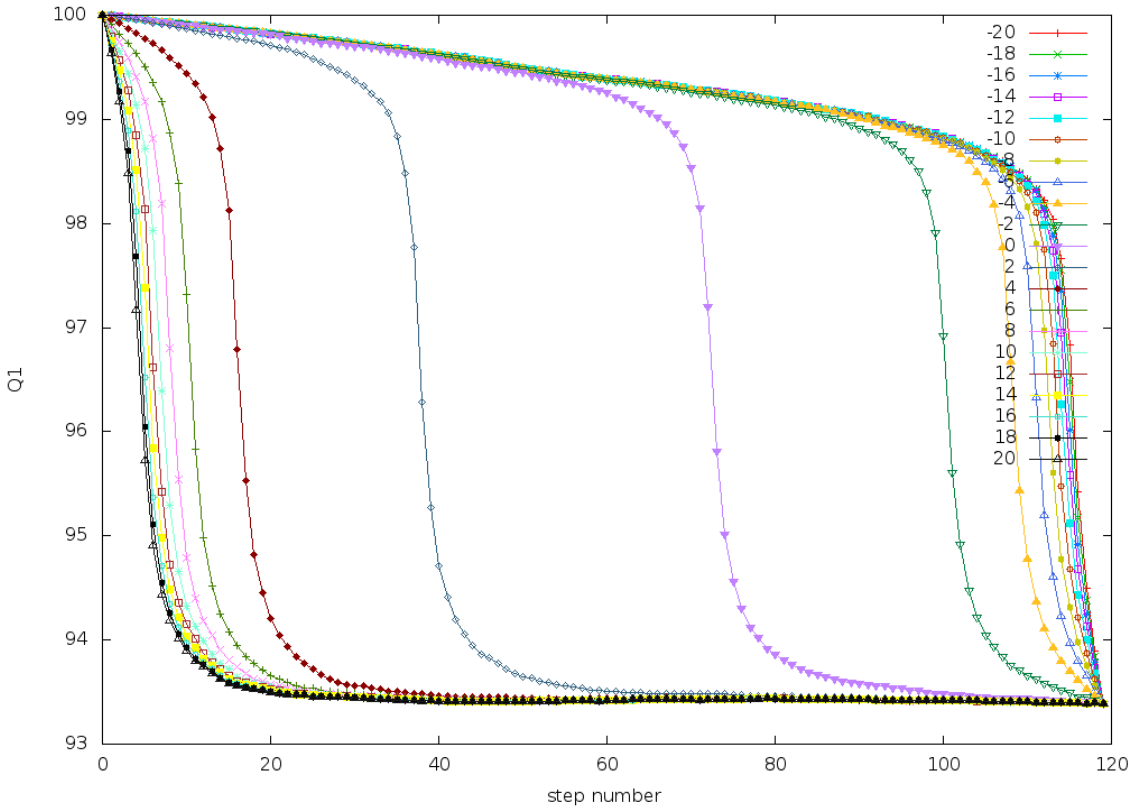

Figure 7. Q_1 vs. t at large values of t_f: Q_1 of a frame of the trajectory is the similarity of that frame to the initial state of the trajectory. These trajectories were generated by simulating the PreTS to Pdt transition of the TrpRS system (described later) at t_f=300 at a range of ΔE values. The sigmoidal curves that were observed in the intermediate regime become steeper in the large t_f regime. Also at ΔE values that are farther away from 0, the sigmoidal curves start to resemble a step function. This happens because in the large regime, the step size used in plotting the curve might be too large and the entire transition could take place within a single step.

In this regime, the system has enough time to undergo transition under the influence of the equations of motion and the potential. Also, the convergence surface equation can be solved at $\bar{t} = \frac{t_f}{2}$ to get $\Delta E_{0.5}$ as

$$\Delta E_{0.5} = \left(E_{right} - E_{left}\right)\left(\frac{-krkl}{(k_r + k_l)^2}\right) \tag{50}$$

It is interesting to note that when $k_r = k_l$,

$$\lim_{t_f \to \infty} \Delta E_{0.5} = -\lim_{t_f \to 0} \Delta E_{0.5} \tag{51}$$

Another observation that lends support to the hypothesis that using a large value of t_f to compute the trajectories is the correct approach, comes from the values of action. Since PATH is based on the minimization of the Onsager-Machlup action to derive the equations of motion, it is not surprising that the values of action might provide useful information for the optimization of PATH input parameters on the convergence surface.

For a double well system, the action of the most probable path in both the wells can be written as

$$S = \frac{1}{2\gamma}\left(\int_0^{\bar{t}} \left(m\gamma\dot{x} + k_l(x-a)\right)^2 dt + \int_{\bar{t}}^{t_f} \left(m\gamma\dot{x} + k_r(x-b)\right)^2 dt\right) \tag{52}$$

Integrating the above equation for the left well is straight forward. On assuming that at time $t = 0$, $x = a$ and at $t = \bar{t}$, $x = \bar{x}$ then the equation for the action in the left well is

$$S_l = \frac{\Gamma_l}{2}(\bar{x} - a)^2 \frac{e^{\Gamma_l \bar{t}}}{\sinh(\Gamma_l \bar{t})} \tag{53}$$

The above equation was derived assuming that the system starts at the equilibrium state and ends up at the transition state in the left well.

In the right well, when the reaction is moving forward in time, the assumption that is made is that at time $t = \bar{t}$, $x = \bar{x}$ and at $t = t_f$, $x = b$. This assumption leads to the following equation

$$S_r = \frac{\Gamma_r}{2}(\bar{x} - b)^2 \frac{e^{-\Gamma_r(t_f - \bar{t})}}{\sinh\left(\Gamma_r(t_f - \bar{t})\right)} \tag{54}$$

At large t_f the two action equations give rise to Fig. 8. It can be seen from the figure that the right action has a tendency to go to 0 even when the values of ΔE are such that the transition state is close to the initial state and most of the trajectory is in the right well. This would mean that, in a case where the transition state falls on the initial state, the total action will be 0, which gives rise to a deterministic path with probability 1. This result is counter-intuitive and would mean that the value of ΔE has no effect on the trajectory in the right well.

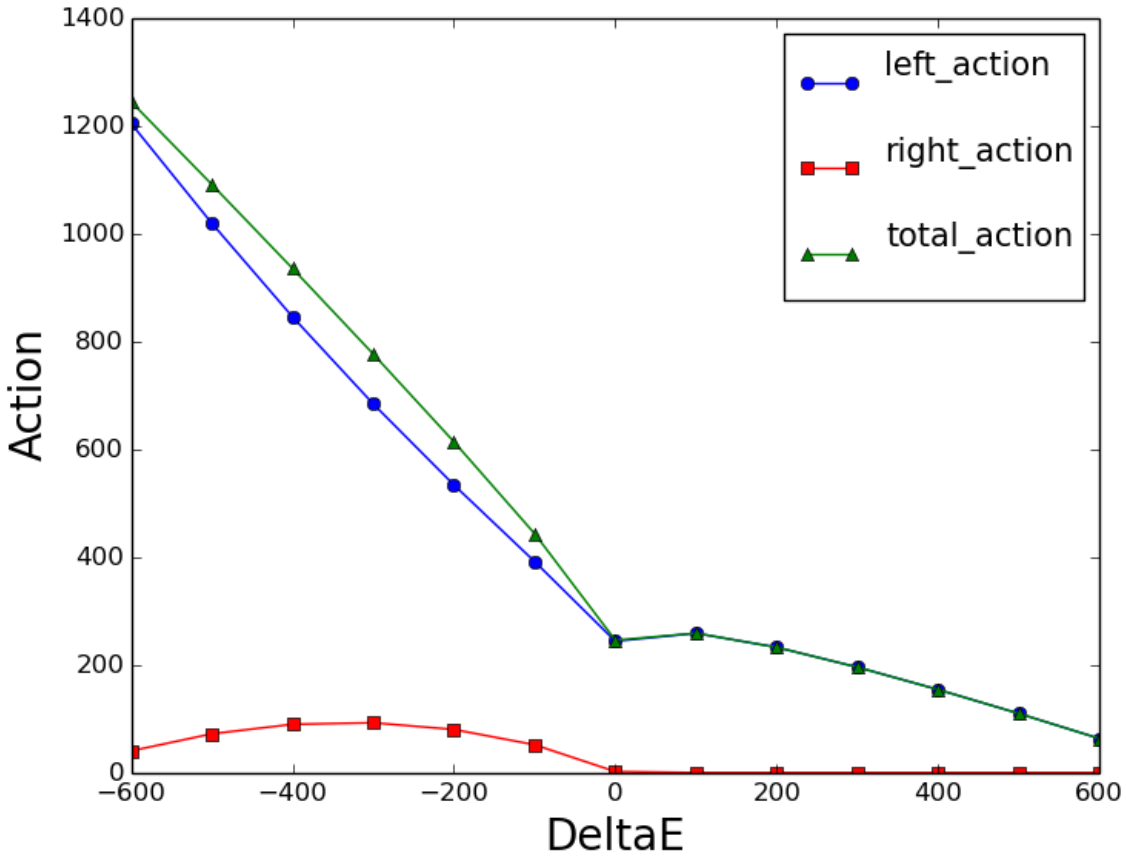

Figure 8. Action vs. ΔE with the old equation for right action: Simulations of transition from rigor state to the prepowerstroke state of myosin VI converter domain (described later) were performed at $t_f = 300000$ at a range of ΔE values. Left, right and total action plotted as a function of ΔE at this constant large value of t_f. The left action goes to zero at large positive ΔE values because the initial state has almost the same energy as the transition state and most of the trajectory is in the right well. But right action also goes to 0 which is not the expected behavior.

The reason behind this behavior of the right action arises from the nature of the Onsager-Machlup equations of motion. Unlike the classical equations of the motion, which are periodic in nature, that is, the time dependence of the equation arises from the sine term in the equation, OM equations of motion are dependent instead on sinh terms, which are exponential functions. Hence, the trajectory is always a displacement away from the equilibrium when the

direction of time is forward. The system can move from the transition state towards the equilibrium state only when time is reversed.

This consideration of the direction of time and displacement from the equilibrium state means that the trajectory and action in the right well have to be computed from state b to the transition state and then the time is reversed so that the overall reaction goes from state a to b. This modification in the assumptions gives rise to following equation for the right action

$$S_r = \frac{\Gamma_{\mathrm{r}}}{2}(\bar{x} - b)^2 \frac{e^{\Gamma_{\mathrm{r}}(t_f - \bar{t})}}{\sinh\left(\Gamma_r(t_f - \bar{t})\right)} \tag{55}$$

This modification gives rise to Fig. 9 which shows that a particular value of ΔE, the total action goes through a minimum and this ΔE value is the same as $\Delta E_{0.5}$.

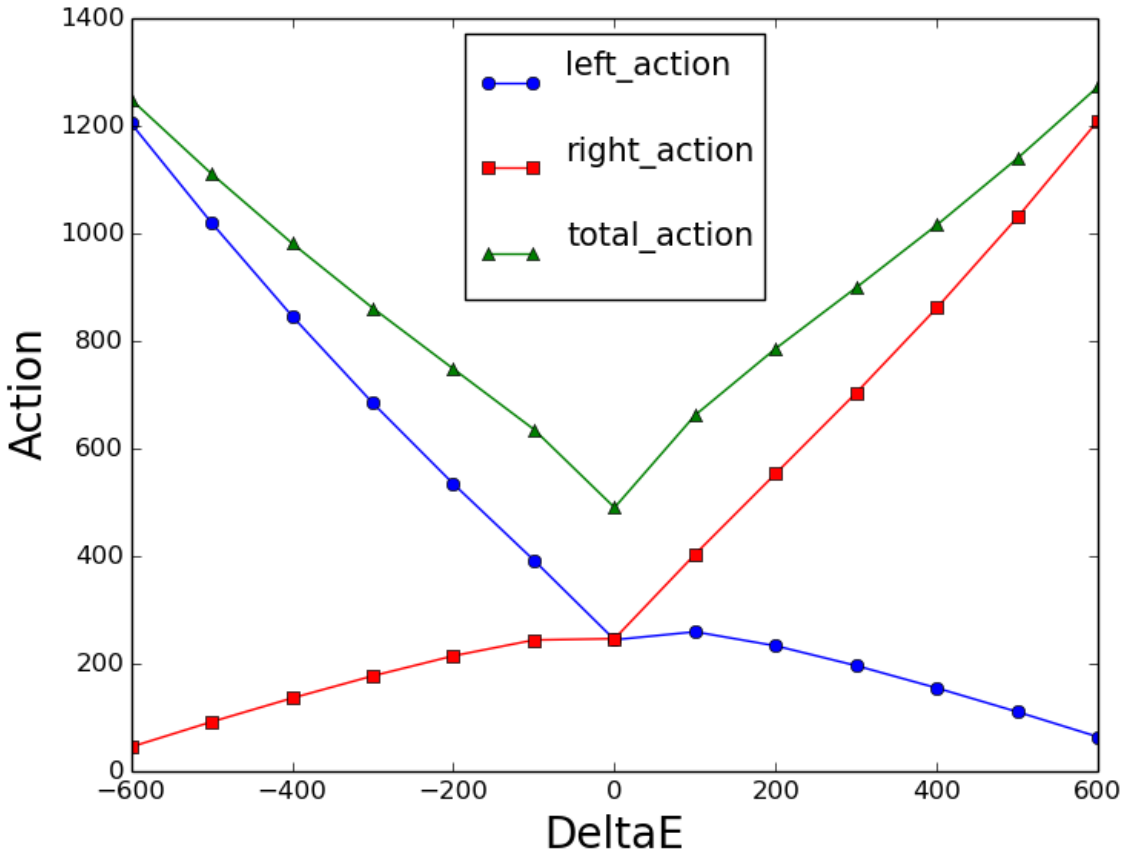

Figure 9. Action vs. ΔE with the new equation for right action: Simulations of transition from rigor state to the prepowerstroke state of myosin VI converter domain (described later) were performed at $t_f = 300000$ at a range of ΔE values. Left, right and total action plotted as a function of ΔE at this constant large value of t_f. The modified right action gives rise to a minimum in total action at $\Delta E_{0.5}$.

The minimum of action, when plotted as a function of t_f exponentially decays and reaches an asymptotic value (Fig. 10). This is the global minimum of action that occurs at large values of t_f and at $\Delta E_{0.5}$. As the value of action is related to the probability of a path, the global minimum of action would mean that it is the most probable path.

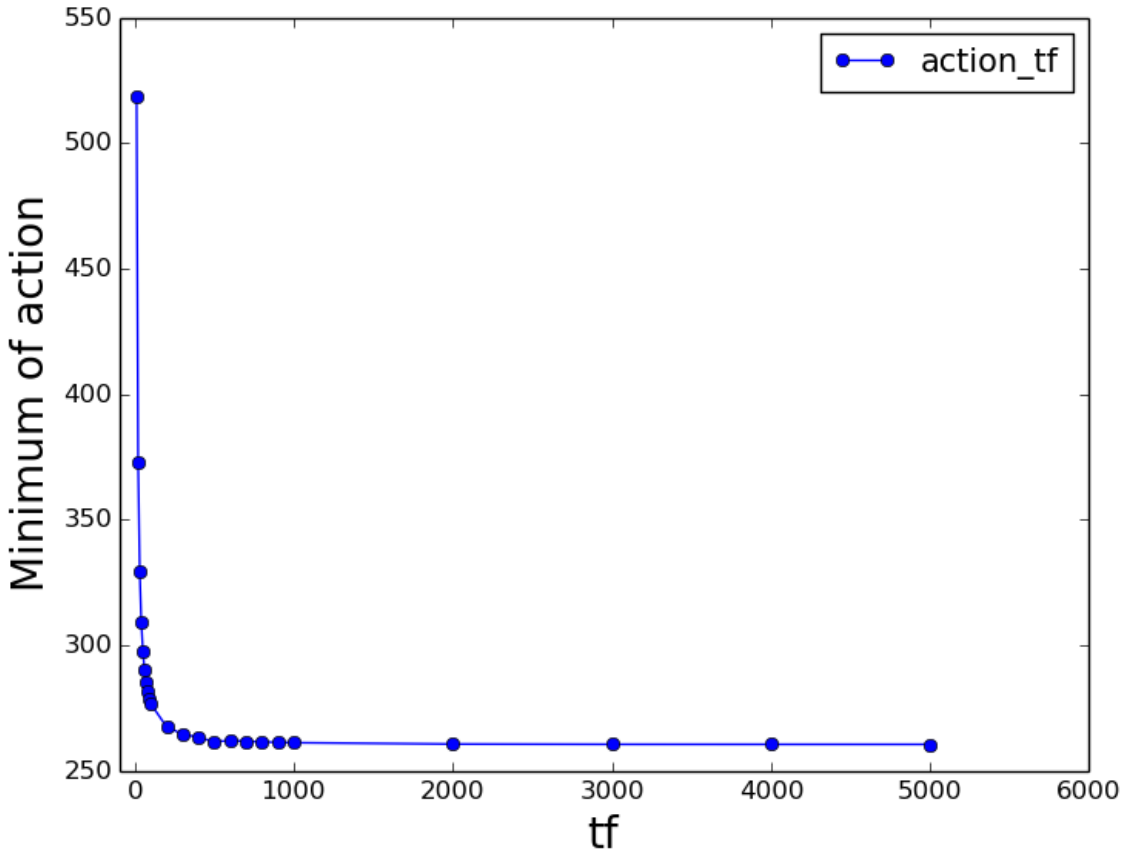

Figure 10. Minimum of action vs. t_f: Simulations of transition from rigor state to the prepowerstroke state of myosin VI converter domain (described later) were performed at different values of t_f and ΔE. The minimum of action at each value of t_f behaves asymptotically with respect to t_f. This curve also fits to an exponential equation with a R^2 of 0.99

Based on Fig. 10, it is clear that running simulations at large values of t_f is beneficial and it also gives the value of ΔE at which the simulation has to be performed. But there is still no informational generated by the PATH method about the values of the force constants. It could be argued that since the product of kt is what determines whether the value of t_f used is sufficiently large, it is only the product that matters and the value of the force constant independent of t_f is not. Also, simulations of several systems using PATH indicate that the

structure of the transition state becomes invariant at large values of t_f. This means that when the action is at its global minimum, the structure of the transition state converges.

Even though large values of t_f generate the global minimum of action and invariant transition state structures, it gives rise to a new problem. At extremely large values of t_f, the path spends most of its time near the equilibrium structures and uses only a fraction of the total time to change the conformation of the protein. Also, the system spends more time in the narrower (more energetic) well than in the wider well. This behavior contradicts statistical mechanics. But, at the same time, once the conformational change starts, the system takes less time to climb up the potential well in the narrower well than in the wider well, which is consistent with statistical mechanics.

The origin of these behaviors can be understood in the following way. As described earlier, converting the equations of motion from those defined by classical action to those defined by OM action changes the fractional increment in position, $x(t)$, from an oscillatory motion to the hyperbolic sine function. As a consequence, the system invariably spends most of its time at the origin (i.e., at $x(t) = a$) and commences its climb to the transition state after an inordinately long time. This problem of the system spending most of the time in the initial state has previously been observed (Faccioli et al. 2006; Ghosh et al. 2002; Pinski & Stuart 2010). A solution to this problem can be obtained by transforming the Lagrangian from the time-dependent Newtonian description to the dual, energy-dependent Hamilton-Jacobi description (Faccioli 2008). This transformation changes the reaction coordinate into an energy based one where even a small change in the energy has to be accompanied by changes in structure. That elegant coordinate transformation affords a more complete solution to the problem. It is possible that for complex dynamic processes like *ab initio* protein folding, where important structural changes may occur at the level of bond vibration, neglecting part of the trajectory may entail the loss of relevant information.

For protein conformational changes, like domain motions that depend on large frequency rigid-body motions, I describe multiple lines of evidence that no essential information is lost by truncating the initial, invariant portion of the trajectory during which the structure does not change. To solve this problem, the system must be given just enough time for the transition state to converge and no more. Thus, the PATH trajectory must be truncated by beginning only when the system has moved away from a by at least 10% of the total distance between the equilibrium state and the transition state. This is an arbitrary choice; using 1% of the distance from the equilibrium state would change the resulting transition state almost imperceptibly.

An appropriate value of t_f can be calculated for the 1D diatomic system in the following way. Using the equation of motion in 1D in the left well, a general OM trajectory can be written as

$$x(t) = a + (\bar{x} - a)\frac{\sinh(kt)}{\sinh(k\bar{t})} \tag{56}$$

when $t_f \to \infty$, (55) becomes

$$x(t) = a + (\bar{x} - a)e^{-k(\bar{t}-t)} \tag{57}$$

Since the time of interest is the one at which the system has changed by at least 10%,

$$e^{-k(\bar{t}-t)} = 0.1 \tag{58}$$

which gives,

$$\bar{t}^{opt} = \bar{t} - t = \frac{2.302}{k} \tag{59}$$

There is a $\bar{t}^{opt}$ for each well and t_f^{opt} is the sum of the two. Thus from (59) the optimum value of t_f can be solved. By using this value of t_f and $\bar{t}$ in the velocity continuity equation the invariant $\bar{x}$ can be identified and the $\Delta E_{0.5}$ can also be calculated. Since (59) relates the force constant to t_f, it is not necessary to compute the exact value of the force constant for each well

as errors in the values can be compensated by the use of t_f that is appropriate for the chosen force constant.

This equation directly computes $\bar{t}$ for a 1D diatomic system but for multiatom systems in 3D, there are multiple interatomic interactions, and hence multiple force constants associated with the diagonalized Hessian matrix. Hence, the average force constant for a structure can calculated which is the average of the trace of the Hessian

$$\bar{k} = \frac{tr(H)}{3N} \tag{60}$$

where, N is the number of atoms. And (59) becomes

$$\bar{t}^{opt} = \bar{t} - t = \frac{2.302}{\bar{k}} \tag{61}$$

CHAPTER 4: VALIDATION AND RESULT FROM PATH SIMULATIONS

To validate the new PATH algorithm, I compared the results from PATH with those from other simulation algorithms, namely the String Method (E et al. 2002a; Ovchinnikov et al. 2011) and ANMPathway (Das et al. 2014). I also performed Discrete Molecular Dynamics (DMD) simulations (Dokholyan et al. 1998; Ding et al. 2008; Shirvanyants et al. 2012) using the replica exchange algorithm (Sugita & Okamoto 1999) to see if the PATH transition states are at the saddle points of the conformational free energy landscapes. Finally, I compared the transition states from a signaling protein, a contractile protein, and an enzyme and found similar structural elements contributing to the energy barrier at the transition state.

4.1 PATH, ANMPathway and the String trajectories agree most closely with each other at their transition states

I compared trajectories from the simulations of the converter domain from myosin VI performed using the string method (E et al. 2002a; Ovchinnikov et al. 2011) ANMPathway (Das et al. 2014), and PATH. Since the reaction coordinates of the three trajectories are different, it would be difficult to compare them at every instant. In Fig. 11, I compare the structural similarity and energetic properties of the string transition state as evaluated according to the linearized ANM force field used by PATH. For both comparisons, the subset of structures in the string trajectory that was structurally most similar to the PATH transition state (Fig 11(a)) was the same subset for which the absolute potential energy difference, calculated using the PATH energy function, between those calculated with respect to the initial and final states, was closest

to zero (Fig. 11(b)). In the context of PATH, the structure whose corresponding potential energy difference is zero is, by definition, the transition state.

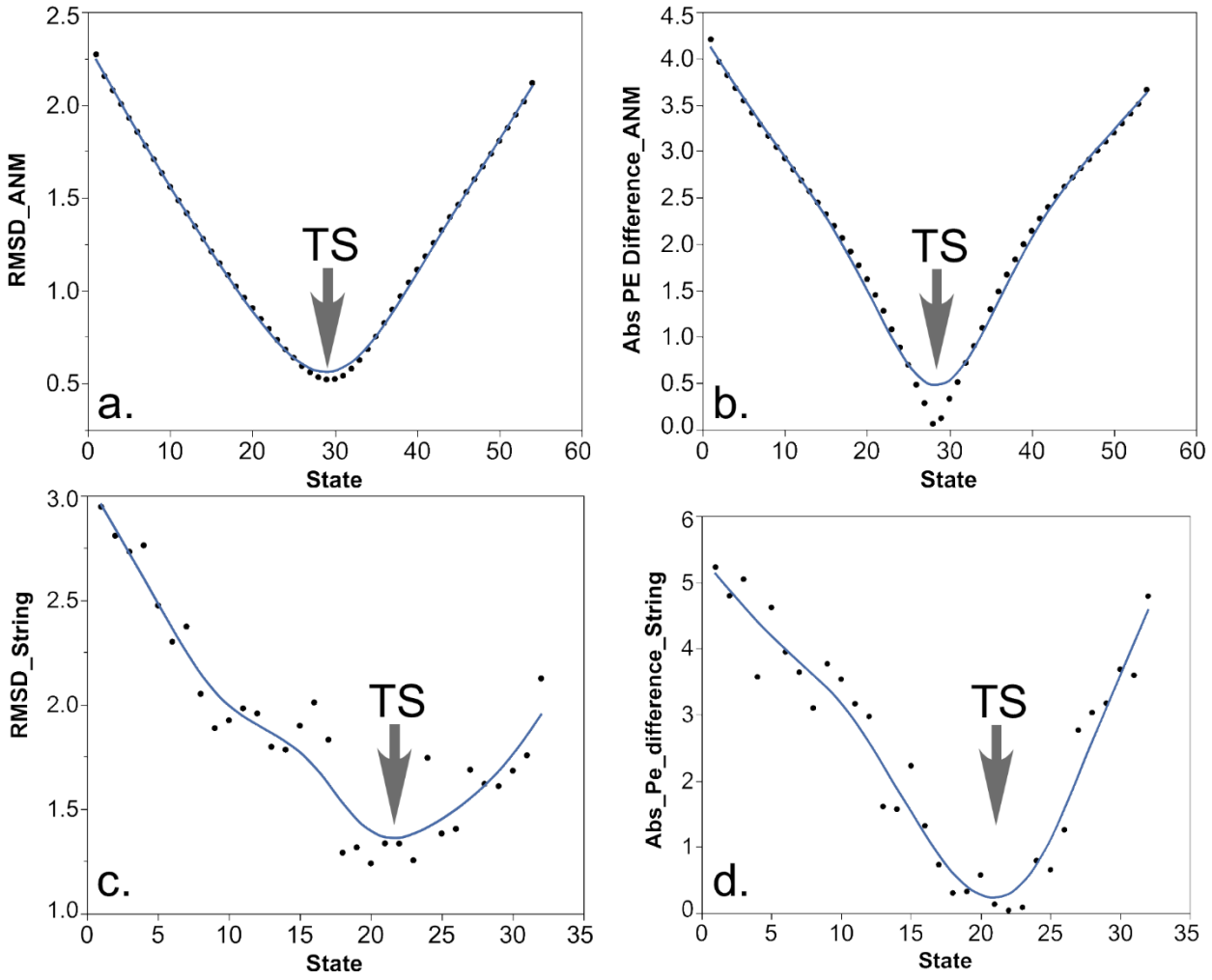

Figure 11. PATH vs. ANMPathway and String: The ANMPathway trajectory and the string trajectories were compared with the PATH trajectory. In (a), I calculated the RMSD between the transition state from the PATH trajectory and all the states along the ANMPathway trajectory. States 28–31 are structurally similar to the PATH transition state. In (b), I calculated the potential energy (PE) of each state in the ANMPathway trajectory with respect to the potential energy well of the initial and the final state and their absolute difference was plotted. States 27–30 have the lowest potential energy difference, which coincides with the states in (a). I performed a similar comparison between the string trajectory (G3c) and the PATH transition states in (c) and (d). States 18–23 are structurally similar to PATH transition state,

and the same states also have the lowest potential energy difference, implying their proximity to the same transition state.

I performed a similar analysis with the myosin conformational change trajectory from the ANMPathway method (Das et al. 2014). I found that when the PATH transition state is compared with the ANMPathway trajectory the structures are the closest [root mean squared deviation (RMSD) 0.52 Å] near the transition state of the ANMPathway trajectory (Fig 11(c)). Similarly, the same group of structures have the absolute potential energy difference closest to zero (Fig 11(d)).

4.2 Discrete molecular dynamics replica exchange simulations verify that transition states identified by path are close to saddle points in the free energy surface connecting initial and final states

Previous work in the lab (Kapustina et al. 2007) has established that *Bacillus Stearothermophilus* Tryptophanyl-tRNA synthetase (TrpRS) passes through three distinct structural states:

- an Open state that can be stabilized either by stoichiometric amounts of tryptophan or by substoichiometric amounts of Mg•ATP (adenosine triphosphate)
- a closed, Pre-TS, stabilized by stoichiometric amounts of Mg•ATP and a tryptophan analog
- a closed, Product state (Pdt), stabilized either by the bound intermediate adenylate product, tryptophanyl-5'AMP, or by stable analogs thereof

As the ligands bind to the Open state, the protein undergoes an induced fit conformational change and goes to the Pre-TS state. At the Pre-TS state, a subsequent catalytic step takes the Pre-TS state to the Pdt state. Both induced-fit and catalysis are slow, relative to the chemical transformation of the substrates; each is therefore associated with a

different conformational transition state. Preliminary analysis with the PATH program had given us descriptive accounts of the two transitions.

- Induced-fit proceeds by and early and higher energy barrier that matches the behavior seen by MD simulations of the TrpRS monomer (Laowanapiban et al. 2009)
- Catalysis proceeds by a later, lower barrier transition state in which the volume of the tryptophan binding pocket assumes a minimum value immediately after the conformational transition state identified by PATH (Weinreb et al. 2014)

The earlier MD calculations relating to the Induced-fit transition were short, 10 ns simulations, and represented what appeared to be a slower conformational change. As MD simulations led to a confirmation of the results PATH had given for the Induced-Fit transition (Laowanapiban et al. 2009), I decided to see whether similar, but more detailed simulations might allow a more stringent test of results the PATH algorithm had given for the catalytic transition. As the catalytic transition represents what is likely a more rapid conformational change with a lower barrier, I carried out replica exchange calculations using DMD (Dokholyan et al. 1998; Shirvanyants et al. 2012; Ding et al. 2008) with sufficiently long equilibrations to appropriately sample the free energy surface connecting the Pre-TS and Pdt states.

DMD simulations were set up with the same configuration of ligands that I had used for PATH: AMP (adenosine monophosphate), Tryptophan, and Pyrophosphate. These ligands were configured as before (Weinreb et al. 2014) to allow an approximation to the actual chemical reaction displacing pyrophosphate from ATP with tryptophan. From the resulting snapshots, I computed the internal coordinates used previously to describe the Induced-Fit transition (Twist and Hinge (Kapustina & Carter 2006)). Sufficiently many steps were computed to visualize the relative populations centered on the two states. For each case, I identified representative structures for the two different distributions. Free energy surfaces were then computed by fitting an empirical equation, built using two 2D Gaussian curves for the equilibrium states and a bi-

variate quadratic for the transition state, to the points between the two equilibrium states. These representative structures reflect the stable, equilibrium structures of the two states in the DMD force field (Ding & Dokholyan 2006; Shirvanyants et al. 2012) as obtained from the DMD simulations. They were then input as initial and final states to PATH.

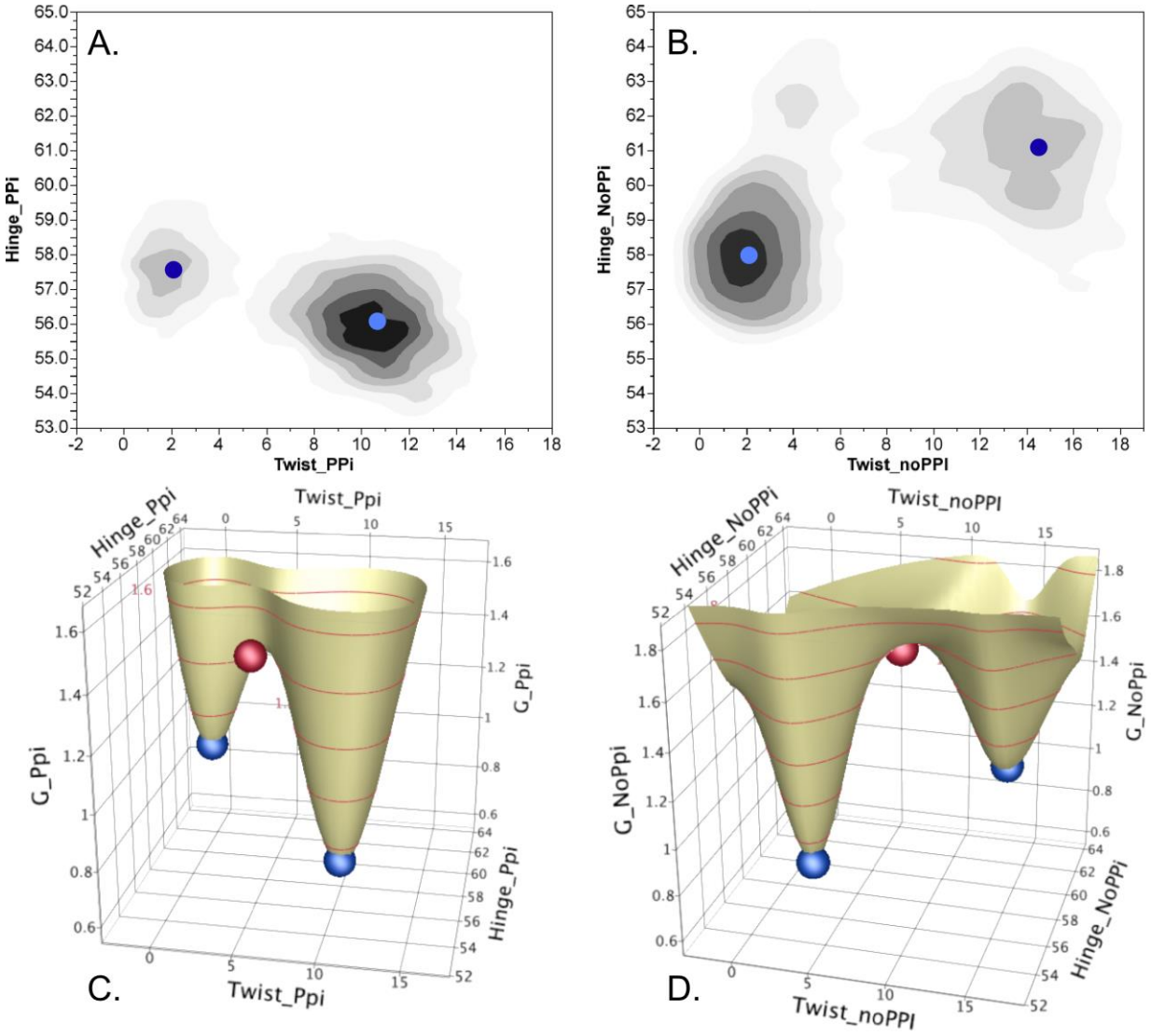

Figure 12. Free energy surface from Replica exchange DMD: Free energy surfaces for the fully liganded TrpRS monomer derived from DMD replica exchange computations and plotted as a function of the two conformational angles, Twist and Hinge, which represent collective variables for the catalytic conformational change derived by Kapustina (Kapustina & Carter 2006). The structures (2000 snapshots) generated at the lowest DMD temperature (~175 K) were used in the analysis. (A)

Distributions of the TrpRS Pre-transition state and Product derived from simulations initiated from the Product state in the (harmonically restrained) presence of AMP, tryptophan, and pyrophosphate. (B) Distributions of these two states in similar simulations without pyrophosphate and without restraining potentials. In (A) and (B), the dark blue circles represent the free energy minima of the less populated state fitted to a bivariate quadratic response surface. Light blue circles represent free energy minima computed using the same approach for the more highly populated states. (C) Free energy surface derived from (A). (D) A similar plot derived from (B). Blue spheres represent the initial and final states input to PATH computations; red spheres represent the coordinates of the transition states produced by PATH.

These calculations produced two notable results:

- The apparent free energy difference between the Pre-TS and Product states depends strongly on the presence of the bound product, pyrophosphate (PPi). If the PPi was retained in the binding pocket by a harmonic potential, the equilibrium was far on the side of the Pre-TS state (Fig 12 (a)). On the other hand, if this potential (or constraints) mimicking PPi binding was relaxed or omitted, rapid PPi release is observed and the distribution of states exhibits higher probability towards product state (Fig. 12(B)). This behavior is especially interesting in view of the possibility that early release of orthophosphate following actin binding triggers the myosin V powerstroke (Ovchinnikov et al. 2010).
- Transition states for the transitions with and without the harmonic potential restraining the PPi output by PATH fall close to the coordinates of the saddle points of the respective energy surfaces (Figs. 12(C) and 12(D)).

4.3 Transition states identified by PATH display comparable rate-limiting structures in three different systems

I began these studies to access structural information about the transient conformational transition state(s) that appear to be rate-limiting for TrpRS catalysis (Kapustina et al. 2007). In

the course of the work, I found it useful also to investigate PATH behaviors of other model systems, including the 1D system described above. Two well-defined protein conformational transitions – Ca^{2+} release by the Ca^{2+}-binding domain of calmodulin and the converter domain of myosin VI. These studies reveal a remarkable similarity in all three transition-states (Fig. 13). In each case, the rate-limiting conformational change involves re-packing of multiple aromatic side chains (Burley & Petsko 1985; Burley & Petsko 1986; Lanzarotti et al. 2011) associated with subtle rearrangements of the surrounding backbone chains. Such rearrangements are known to occur on a far slower timescale (μs to ms (Skalicky et al. 2001)) than rotamer exchanges of aliphatic side chains in hydrophobic core regions.

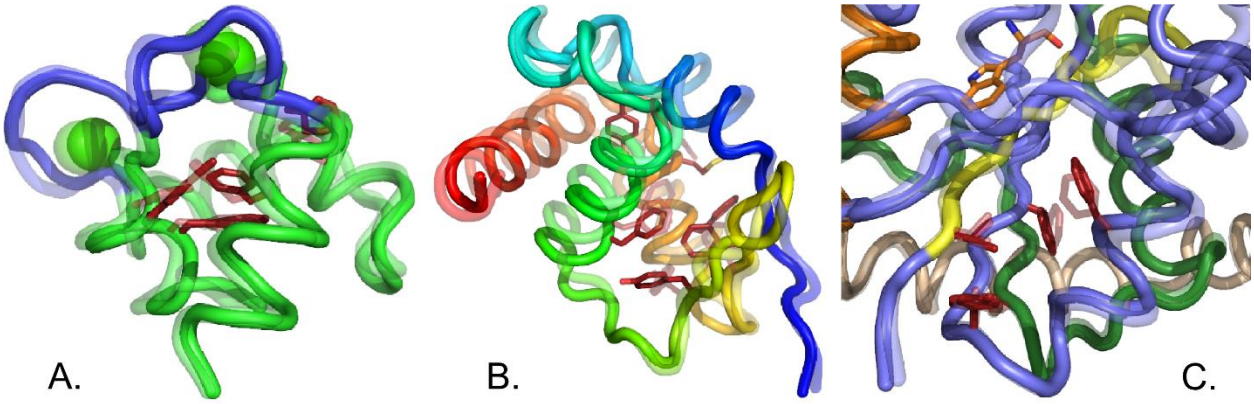

Figure 13. Transition states of three different systems: Conformational transition state structures for Calmodulin Ca^{2+}-binding domain (A), Myosin VI converter domain rigor to Prepowerstroke (B), and the TrpRS induced-fit (C) transition states. Aromatic residues that flip at the transition state are highlighted in red. The initial state is 50% transparent, to distinguish the states before and after the rate-limiting step.

CHAPTER 5: MATERIALS AND METHODS

5.1 Structures

Three TrpRS structures were used in these studies. These structures were derived from the crystal structures of the three conformations of TrpRS, namely, Open (1MAW,1MB2), Pre-TS (1MAU) and Pdt (1I6L). The terminal aminoacid (R328) was excised from the structures as it is not observed in most of the crystal structures. I believe that its absence would not affect the conformational change of the rest of the protein. The ligands in the binding pockets are different for the different states of TrpRS. To make the ligands consistent in all the three structures, I used Tryptophan, AMP, PPi as separate molecules in the binding pocket; the distance between these molecules changes, depending on the state and the chemical species that they represent. The arrangement has been used previously (Laowanapiban et al. 2009; Weinreb et al. 2014) and this allows approximating the chemical reaction without requiring the use of quantum calculations. The myosin VI structures for the rigor state and the prepowerstroke state were derived from 2BKH and 2V26, respectively. As described in (Ovchinnikov et al. 2011), only residues 703–788, which form the converter domain, were used in the simulations. The equilibrium structures for calmodulin were derived from 1CMF and 1FW4, which are carboxy-terminal domains of calmodulin.

5.2 PATH simulations

To run PATH simulations, the number of atoms in the two equilibrium states and their relative order in the two PDB files must be the same. Only the heavy atoms are used. The modified algorithm requires no input parameters other than the two equilibrium states, because

the force constants are assumed to be 0.01 for both states, and errors in this assumption are compensated by the evaluation of $\bar{t}$ for the forward and reverse reactions from equation (61).

5.3 ANMPathway simulations

The ANMPathway calculations were set up on the ANMPathway server. Default input force constants = 0.1 were used for both the energy wells. All the other parameters were set to their default values – Cutoff – 15 Å, Energy offset - 0, Step size (on cusp) - 0.8, Step size (slide down) - 0.04 and Target RMSD - 0.1 Å.

5.4 DMD simulations

Replica Exchange Discrete Molecular Dynamics (REX/DMD) simulations were set up with the Pdt state structure described previously. A harmonic potential was applied between the atoms of the ligands and all the surrounding atoms within 3.5 Å to retain the ligands within the binding pocket. In general, replica exchange simulations are used for efficient sampling of the conformational landscape of a given system. However, I was only interested in monitoring the transition between the Pdt and Pre-TS state. To facilitate the exploration of this particular transition event as well as to expedite the sampling, I introduced weak harmonic constraints to guide the system progressing from Pdt to Pre-TS state. By comparing the native contacts within the two systems (as obtained from their crystal structures), I extracted the unique contacts that were present in the Pre-TS and not the Pdt state. Those contacts were used as experimental constraints. The DMD force field is currently equipped to work only with Cu^{2+} or Zn^{2+}. Since ATP is complexed with Mg^{2+} in the Pre-TS state, it was replaced with Zn^{2+}. I believe that this replacement would not affect the conformational change of the protein in a significant way. I simulated parallel replicas at 24 temperatures ranging from ~175 K to ~405 K for a total duration of 2.5 million steps (~125 ns) as described in (Williams II et al. 2015). As the system requires 500 000 steps to equilibrate, all our analyses were performed with the remaining 2

million steps. Snapshots were generated every 1000 steps, hence all our analyses (Fig. 12) include 2000 snapshots

5.5 Fitting the Free energy surfaces

Each of the 2000 snapshots from the lowest temperature replica exchange DMD simulation was segregated in 225 bins of equal size, based on their Hinge and Twist angles. Based on the distribution of structures within these bins, the free energy surface is computed using the formula

$$\Delta G = -k_B T \ln\left(100 * \left[\frac{n_i}{N}\right]\right) \tag{62}$$

where, n_i is the number of structures in the i^{th} bin and N is the total number of structures.

Then, these free energy values are fitted to the following equation to generate the free energy surface in Fig 12,

$$\begin{aligned} \Delta G = C + Ae^{-\left(\frac{(X-Tw1)^2}{2SigTw1}+\frac{(Y-H1)^2}{2SigH1}+\frac{J(X-Tw1)(Y-H1)}{2\sqrt{SigTw1^2+SigH1^2}}\right)} \\ + Be^{-\left(\frac{(X-Tw2)^2}{2SigTw2}+\frac{(Y-H2)^2}{2SigH2}+\frac{L(X-Tw2)(Y-H2)}{2\sqrt{SigTw2^2+SigH2^2}}\right)} + D(X-Twt) + F(X-Twt)^2 \\ + G(Y-Ht) + H(Y-Ht)^2 \end{aligned} \tag{63}$$

where, X and Y are the Twist and Hinge angles and the constants Tw1, H1, Tw2, and H2 are the twist and hinge, respectively, of the Pdt and Pre-TS structures and Twt and Ht are coordinates of the saddle point.

5.6 Design of computational mutants

Computational mutants of TrpRS were designed using Rosetta Backrub server (Lauck et al. 2010). The backrub algorithm perturbs the structure minimally in the neighborhood of the mutation site, hence the overall structure of the protein doesn't change significantly. In order to

maintain the consistency of the structure I generated the wild-type structure by mutant one of the single point mutants back to the wild-type residue.

CHAPTER 6: CONCLUSION AND FUTURE DIRECTIONS

The modified PATH algorithm generates transition states comparable to those generated by other well established methods like ANMPathway and the String method. Also, in the case of TrpRS, myosin and Calmodulin, it generates transition states that have similar structural features that contribute to the energy barrier at the transition state, which is the conformation of aromatic side chains. The rearrangement of these aromatic sidechains causes the switch from one conformation to another.

Because of these predications made by PATH, I wanted to explore other areas where the PATH algorithm could be useful to understanding the nature of protein conformational changes. One such area is the study of kinetics of mutants of TrpRS. TrpRS contains a region of four residues (I4, F26, Y33 and F37) which were previously identified (Kapustina et al. 2007) as the fulcrum of motion that caused TrpRS to undergo conformational change. These are the same residues which are far away from the active site of the protein but undergo a significant rearrangement during the catalytic cycle of TrpRS. Previous experimental kinetics studies in the lab (Weinreb et al. 2012) have shown that mutation at each of the sites affect the protein in a different manner and alter the rate of catalysis by TrpRS. The action of the quadruple mutant along with Mg^{2+} contributes $\Delta\Delta G \sim -6\ kcal/mol$ of transition state stabilization. Since there is data available on every interaction possible between the mutational sites, this system provides an opportunity for simulation algorithms like PATH to predict the effect of the mutants. I used this system to test if the parameters output by the modified PATH program could predict these experimentally determined kinetic values.

6.1 Output parameters from PATH can be used to model experimentally determined kinetic $\Delta\Delta G$ values for TrpRS mutants

From the kinetics experiments performed using the mutants of TrpRS, the rate of catalysis k_{cat} were calculated (Weinreb et al. 2012). These kinetic rate constants were converted into a free energy term ΔG_{kcat} using the equation,

$$\Delta G_{kcat} = -k_B T \ln(k_{cat}) \tag{64}$$

These ΔG_{kcat} can then be compared with the PATH parameters using regression models.

Of the several parameters that can be calculated from PATH, the one that closely represents the rate of the reaction is $\bar{t}$, as it is the time taken to reach the transition state from the first equilibrium state, which in a way can be considered to be the inverse of the reaction rate. From this rate, a theoretical ΔG_{kcat} can be calculate using equation (64). Similarly, a free energy term related to the reverse reaction can also be calculated from the time to the transition state in the right well, $\bar{t}_l$.

Apart from $\bar{t}$, there are two energy parameters that can also determine the rate of the reaction, namely, the energy of the transition state relative to both the equilibrium states ($U_l^{\ddagger}$ and $U_r^{\ddagger}$) and the difference in energy between the two equilibrium states $\Delta E_{0.5}$. It should be noted that even though the value of $\Delta E_{0.5}$ computed by the new algorithm is the same as the one computed by the old algorithm, the definition of $\Delta E_{0.5}$ has since changed. In the new algorithm, since evaluation of $\bar{t}$ does not require an iterative procedure, unlike the old algorithm where the velocity and energy continuity equations were used in tandem to identify $\bar{t}$, the transition state structure is calculated from the velocity continuity equation (43) directly. The energy continuity equation (44) is then rearranged to evaluate the value ΔE associated with this $\bar{t}$ and t_f. That is.

$$\Delta E = \frac{1}{2}(\bar{x} - b)H_r(\bar{x} - b)^2 - \frac{1}{2}(\bar{x} - a)H_l(\bar{x} - a)^T \quad (65)$$

As mentioned earlier, this ΔE is still equivalent to the $\Delta E_{0.5}$ computed using the old algorithm, the value of ΔE at which the action is at global minimum, but $\bar{t}_l$ and $\bar{t}_r$ are not equal. Hence the value of ΔE computed from the new algorithm (equation (65)) will henceforth be mentioned only as ΔE.

Using these PATH parameters, the experimental kinetic free energy was fitted and the fitting had a correlation coefficient, R^2 of 0.59

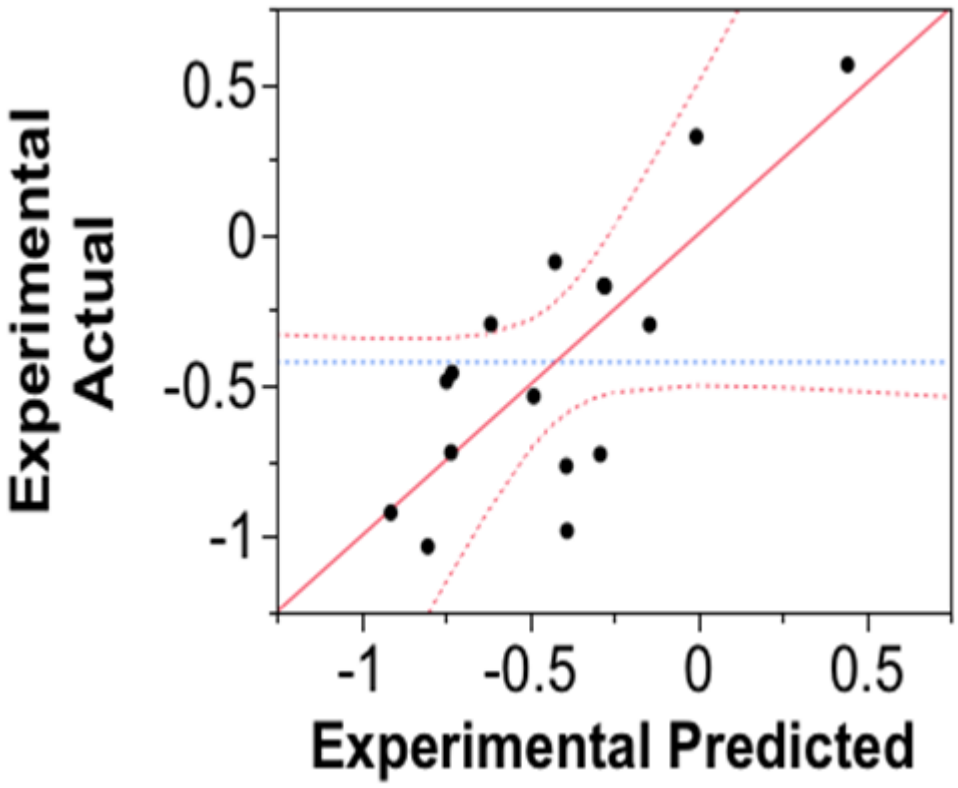

Figure 14. Prediction of experimentally determined kinetic free energy using PATH: The experimental free energy values of 16 TrpRS mutants are predicted using different parameters from PATH. The variables used in this fit were $U_r^{\ddagger}$, ΔE and $\ln\left(\frac{1}{\bar{t}_r}\right)$ and their higher order terms. The correlation coefficient for the fit was found to be $R^2 = 0.59$.

Term	Estimate	Standard Error	Prob.>|t|
$\left(U_r^{\ddagger} - 64.7969\right) * \left(\ln\left(\frac{1}{\bar{t}_r}\right) + 2.12779\right)$	162.6334	46.46563	0.0057
$(\Delta E - 1.35413) * \left(U_r^{\ddagger} - 64.7969\right)$	1.3417523	0.399572	0.0073
ΔE	-1.019092	0.593475	0.1167
$U_r^{\ddagger}$	0.1484907	0.08906	0.1264
$\ln\left(\frac{1}{\bar{t}_r}\right)$	-31.50726	48.11258	0.5273

Table 1: The estimates, standard error in the estimation and the significance of The PATH parameters used to fit the experimental kinetic values are tabulated here. The PATH simulations were performed on 16 TrpRS mutants which alter the kinetics of the proteins in different ways.

Though the correlation coefficient, R^2 is 0.59, the fit is not particularly good and also, as it can be seen from the table, apart from the two second order interaction terms, the individual parameters are not significant. Hence either a different set of parameters or a different way to estimate the parameters have to be used for this correlation.

6.2 Modifying the PATH Hessian

Use of appropriate spring constants is essential for generating PATH parameters than could be used for correlation studies with experimental parameters. The importance of the force constant cannot be understated because all the three parameters used in the above correlation, are derived from the force constants. I have argued previously that it is reasonable to use a constant value of the force constants for both the wells for all the systems and any error in the

force constants will be compensated by the estimation of t_f as it is the product of these two quantities which is important and not the actual values. But if it is possible to evaluate more accurate force constants, it could improve the trajectories and also generate parameters that agree better with experiments.

One general approach to deriving the values of force constants comes from the group of Konrad Hinsen (Hinsen et al. 2000). They evaluate the force constant for an ANM potential by using an empirical equation that was derived by fitting the motion of $C\alpha$ atoms in the AMBER force field (Cornell et al. 1995). There is one equation for $C\alpha$ atoms that are within 0.4 nm and another equation for atoms that are farther away. The following is the equation

$$k = 8.6 \times 10^5\ kJ\ mol^{-1}nm^{-3}.r - 2.39 \times\ 10^5\ kJ\ mol^{-1}nm^{-2}\ for\ r < 0.4\ nm$$
$$k = 128\ kJ\ nm^4mol^{-1}.r^{-6}\ for\ r \geq 0.4\ nm \quad (66)$$

In both the regimes, the force constant is dependent on the distance of separation between the atoms.

Another parameter that can be evaluated differently is $\bar{t}$. Currently, in the new PATH algorithm, it is calculated from equation (61) where the denominator is the average force constant calculated from the mean of the trace of the Hessian, or in other words the mean of the Hessian. The reason behind this approach is make sure that the system has converged, that is the product of the force constant and the time parameter is large enough that any increase in this product does not change the estimated structure of the transition state. If each mode of a vibrating molecule is considered separately, the vibrational mode which has the most significant effect on the PATH is the smallest vibrational mode. This means that if the smallest mode converges then all the other modes will also converge. Hence equation (61) can rewritten as

$$\bar{t}^{opt} = \bar{t} - t = \frac{2.302}{\lambda_{smallest}} \quad (67)$$

Using equation (66) to calculate the force constants, and (67) to calculate the time to the transition state on both the wells, the PATH parameters that are generated for the TrpRS mutants are different and the correlation with experimental parameters increases significantly, with a R^2 value of 0.76

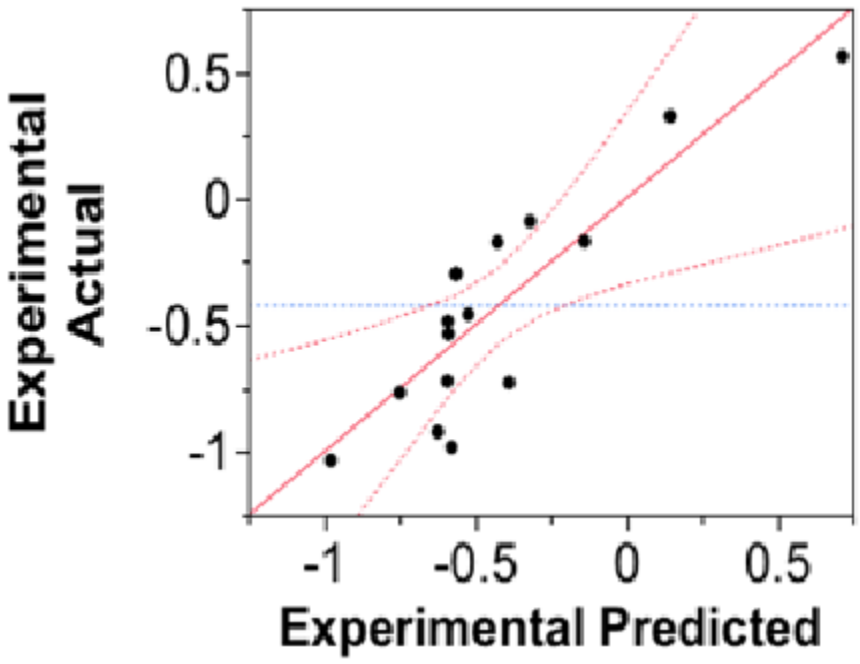

Figure 15. Prediction of experimentally determined kinetic free energy using a modified Hessian: The experimental free energy values of 16 TrpRS mutants are predicted using different parameters from PATH. The variables used in this fit were $U_r^{\ddagger}$, ΔE and $\ln\left(\frac{1}{\bar{t}_l}\right)$ and their higher order terms. The correlation coefficient for the fit was found to be $R^2 = 0.76$.

Term	Estimate	Standard Error	Prob.>|t|
$U_r^{\ddagger}$	1.8057062	0.525875	0.0064
$(\Delta E + 0.53131) * \left(U_r^{\ddagger} - 5.14956\right)$	-74.67219	25.94635	0.0164
$\ln\left(\frac{1}{\bar{t}_l}\right)$	274.54033	106.1404	0.0271
ΔE	-13.28052	6.094605	0.0543
$(\Delta E + 0.53131) * \left(\ln\left(\frac{1}{\bar{t}_l}\right) + 7.54065\right)$	15967.629	7987.513	0.0735

Table 2. The estimates, standard error in the estimation and the significance of The PATH parameters used to fit the experimental kinetic values are tabulated here. The PATH simulations were performed on 16 TrpRS mutants which alter the kinetics of the proteins in different ways. There are more significant parameters in this fit than in the one above.

Comparing Table 1 and Table 2 shows that the parameters are more significant when the Hessian is modified. Also there are more parameters that are significant in Table 2 than in Table 1. Added to the fact the R^2 is higher, it can be argued that the second model is much better than the first one and this improvement is because of the modified Hessian. Thus a better estimation of the force constants results in significant improvement of the PATH parameters. An interesting observation in this context is that, even though the modified Hessian is only based on $C\alpha$ atoms, it still behaves differently for different mutants. Though the backbone trace of the transition states generated by both Hessians are very similar, yet the results are quite different. It has been shown previously (England 2011) that simple backbone based potentials along with information about the hydropathy of each aminoacid can be used for *ab initio* folding of proteins.

Since the dynamics in the case of domain conformational changes is much simpler, maybe a backbone based potential is still sufficient to capture the dynamics.

6.3 Including potential to constrain the torsional angles

One of the problems with the modified Hessian described in the previous section is that it cannot provide information about the side chains of aminoacids and their conformations at the transition state. Hence analysis like what is shown in Fig. 13 cannot be reproduced using this approach. One solution to this problem is to use an all atom potential to build the Hessian. There have been such approaches taken previously (Hinsen & Kneller 1999) but an empirical like equation (66) hasn't be derived.

An interesting approach taken by Hyuntae Na and Guang Song (Na & Song 2014) to connect ANM to more complete techniques for studying protein dynamics like full potential Normal Mode Analysis (NMA) provides a partial solution to this problem. In their paper, they start with a simple ANM potential, just like the one used in PATH, and compare the trajectory to that of a full potential NMA. They find that the correlation coefficient is 0.41. Then they add a term for the torsional angle to the potential and the correlation increases to 0.79 immediately. Adding more terms for bond angle, van der Waals interaction, Urey-Bradley and improper increases the correlation coefficient to 0.88. Thus the correlation between ANM and NMA almost doubles just with an additional torsional term. Hence the PATH Hessian could be modified to include the torsional potential, which could increase the correlation with experimental parameters and at the same time, also provide information about the aminoacid sidechains.

The new Hessian can be built by adding the Hessian for ANM, which was described earlier to a new Hessian for the torsional potential. The process of building the torsional Hessian is similar to that of the ANM Hessian. The following is the derivation of the single 12×12 block of the Hessian.

The torsional angle is calculated between 4 atoms are connected by three consecutive covalent bonds. Let's consider these 4 atoms be $a = (x_1, y_1, z_1)$, $b = (x_2, y_2, z_2)$, $c = (x_3, y_3, z_3)$ and $d = (x_4, y_4, z_4)$. The difference vectors connecting the 4 atoms then are $l = (x_2 - x_1, y_2 - y_1, z_2 - z_1)$, $m = (x_3 - x_2, y_3 - y_2, z_3 - z_2)$ and $n = (x_4 - x_3, y_4 - y_3, z_4 - z_3)$. The torsional angle between the 4 atoms is then written as

$$\phi = \cos^{-1}\left(\frac{u.v}{|u||v|}\right) \tag{68}$$

where, u and v are normal to the places of the vectors l, m and m, n respectively.

The potential arising from the torsional angle constraint is written as

$$V(\phi) = K_\phi\big(1 - \cos\big(n(\phi - \phi_0)\big)\big) \tag{69}$$

where, ϕ_0 is the torsional angle of the four atoms in the equilibrium state and K_ϕ is the force constant associated with the torsional potential and n is a multiplicity factor. If K_ϕ and n are assumed to be 1 (Na & Song 2014), then (69) becomes,

$$V(\phi) = \big(1 - \cos\big((\phi - \phi_0)\big)\big) \tag{70}$$

The Hessian, as described above is a matrix of second derivatives of the potential. Hence the second derivative of (70) has to be computed.

By Taylor expansion, (70) can be rewritten as

$$V(\phi) = \frac{(\phi - \phi_0)^2}{2} \tag{71}$$

The first derivative of (71) relative to the first coordinate of the 4 atoms, $x_!$, is

$$\frac{\partial V(\phi)}{\partial x_1} = (\phi - \phi_0).\frac{\partial \phi}{\partial x_1} \tag{72}$$

And the second derivative is

$$\frac{\partial^2 V(\phi)}{\partial x_1^2} = (\phi - \phi_0).\frac{\partial^2 \phi}{\partial x_1^2} + \left(\frac{\partial \phi}{\partial x_1}\right)^2 \quad (72)$$

At equilibrium, since $\phi = \phi_0$,

$$\frac{\partial^2 V(\phi)}{\partial x_1^2} = \left(\frac{\partial \phi}{\partial x_1}\right)^2 \quad (72)$$

For a general coordinate X, the above equation can be written as

$$\frac{\partial^2 V(\phi)}{\partial X_i \partial X_j} = \frac{\partial \phi}{\partial X_i}\frac{\partial \phi}{\partial X_j} \quad (73)$$

There are 78 unique combinations of $\frac{\partial \phi}{\partial X}$ which make up the 12×12 Hessian matrix and each of these terms can be calculated as the derivative of the angle (68) relative to the coordinates.

Considering,

$$R = \frac{u.v}{|u||v|} \quad (74)$$

$\frac{\partial \phi}{\partial X}$ can be calculated as

$$\frac{\partial \phi}{\partial X} = -\frac{1}{\sqrt{1 - R^2}}.\frac{\partial R}{\partial X} \quad (75)$$

A potential defined in this manner, if added to the ANM potential could improve the trajectories and at the same time generate PATH parameters that correlate better with experimental methods. PATH can then be used to study the effects of mutations on protein conformational changes and protein stability.

6.4 Conclusion

PATH was originally developed as a rapid algorithm (Franklin et al. 2007) for computing conformational changes between equilibrium states of proteins if several input parameters were

known before the simulations can be set up. With the modifications to the algorithm that I have made, PATH now requires only the crystal structures for perform the calculations. Also, since the new algorithm does not require to check for energy continuity by an iterative method, it is considerably faster than the original algorithm.

Previously, the results generated from PATH had not been compared with those generated by other methods. Simulations of Adenylate Kinase were performed and most of the validation of PATH was based on qualitative tests of this trajectory (Franklin et al. 2007). But now, there are two quantitative tests showing that the transition states generated by PATH agree with those from the String method and ANMPathway and also with a more general purpose algorithm like DMD.

PATH also generates several parameters whose significance haven't been fully understood yet. Though the use of parameters from PATH to predict experimentally determined kinetic parameters is a good first step towards understanding these parameters, using PATH on several other well characterizing enzymes will be essential to understand these parameters fully.

The current version of PATH, even though it is efficient and correctly predicts conformational transition states, still leaves scope for improvement. As outlined above, the potential functions can be modified to generate better results without compromising on the speed of PATH or its efficiency.

APPENDIX 1: COMPARISON OF COMPUTATIONAL METHODS

Method	Pros	Cons
Molecular Dynamics simulation	• Provides information about the dynamics of the protein at an atomistic level	• Time and resource intensive
Replica exchange Discrete Molecular Dynamics simulation and other MD simulation methods coupled with a sampling algorithm	• Faster than traditional MD simulations • Provides information about the dynamics of the protein at an atomistic level	• Time and resource intensive
String Method	• Much faster than MD simulations • Provides information about the dynamics of the protein at an atomistic level • Gold standard	• Still dependent on MD simulations. • Resource intensive • Not readily accessible
ANMPathway	• Fast • Accessible	• Not dependent on time hence the method does not provide any kinetic information about the conformation change process • Backbone only

PATH	• Fast • All atom simulations • Time dependent; kinetic information available • Accessible	• Dynamics of the sidechain is not available

BIBLIOGRAPHY

Atilgan, A.R. et al., 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophysical journal, 80(1), pp.505–15.

Bahar, I., Atilgan, A.R. & Erman, B., 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and design, 2(3), pp.173–81.

Baker, J.L., Biais, N. & Tama, F., 2013. Steered Molecular Dynamics Simulations of a Type IV Pilus Probe Initial Stages of a Force-Induced Conformational Transition M. Nilges, ed. PLoS Computational Biology, 9(4), p.e1003032.

Brooks, B.R. et al., 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry, 4(2), pp.187–217.

Brooks, B.R. et al., 2009. CHARMM: The biomolecular simulation program. Journal of Computational Chemistry, 30(10), pp.1545–1614.

Burley, S.K. & Petsko, G.A., 1986. Amino-aromatic interactions in proteins. FEBS Letters, 203(2), pp.139–143.

Burley, S.K. & Petsko, G.A., 1985. Aromatic-Aromatic Interaction: A Mechanism of Protein Structure Stabilization. Science, 229, pp.23–28.

Chipman, D.M., 1971. A kinetic analysis of the reaction of lysozyme with oligosaccharides from bacterial cell walls. Biochemistry, 10(9), pp.1714–22.

Cornell, W.D. et al., 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. Journal of the American Chemical Society, 117(19), pp.5179–5197.

Das, A. et al., 2014. Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model. PLoS computational biology, 10(4), p.e1003521.

Ding, F. et al., 2008. Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure (London, England : 1993), 16(7), pp.1010–8.

Ding, F. & Dokholyan, N. V., 2006. Emergence of protein fold families through rational design. PLoS Computational Biology, 2(7), pp.0725–0733.

Dokholyan, N. V et al., 1998. Discrete molecular dynamics studies of the folding of a protein-like model. Folding and design, 3(6), pp.577–87.

Durr, D. & Bach, A., 1978. The Onsager-Machlup Function as Lagrangian for the Most Probable Path of a Diffusion Process. Communications in Mathematical Physics, 170, pp.153–170.

E, W., Ren, W. & Vanden-Eijnden, E., 2002a. Energy landscapes and rare events. Proceedings of the Intenattional Congress of Mathematicians, I, pp.621–630.

E, W., Ren, W. & Vanden-Eijnden, E., 2002b. String method for the study of rare events. Physical Review B, 66(5), p.052301.

England, J.L., 2011. Allostery in Protein Domains Reflects a Balance of Steric and Hydrophobic Effects. Structure, 19(7), pp.967–975.

Faccioli, P., 2008. Characterization of protein folding by dominant reaction pathways. The journal of physical chemistry. B, 112(44), pp.13756–64.

Faccioli, P. et al., 2006. Dominant Pathways in Protein Folding. Physical Review Letters, 97(10), p.108101.

Fluitt, A., Pienaar, E. & Viljoen, H., 2007. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. Computational Biology and Chemistry, 31(5-6), pp.335–346.

Franklin, J. et al., 2007. MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. Nucleic acids research, 35(Web Server issue), pp.W477–82.

Fujisaki, H., Shiga, M. & Kidera, A., 2010. Onsager-Machlup action-based path sampling and its combination with replica exchange for diffusive and multiple pathways. The Journal of chemical physics, p.20.

Ghosh, A., Elber, R. & Scheraga, H.A., 2002. An atomically detailed study of the folding pathways of protein A with the stochastic difference equation. Proceedings of the National Academy of Sciences, 99(16), pp.10394–10398.

Gumbart, J. et al., 2009. Regulation of the Protein-Conducting Channel by a Bound Ribosome. Structure, 17(11), pp.1453–1464.

Hess, B. et al., 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation, 4(3), pp.435–447.

Hinsen, K. et al., 2000. Harmonicity in slow protein dynamics. Chemical Physics, 261(1-2), pp.25–37.

Hinsen, K. & Kneller, G.R., 1999. A simplified force field for describing vibrational protein dynamics over the whole frequency range. The Journal of Chemical Physics, 111(24), p.10766.

Kapustina, M. et al., 2007. A conformational transition state accompanies tryptophan activation by B. stearothermophilus tryptophanyl-tRNA synthetase. Structure (London, England : 1993), 15(10), pp.1272–84.

Kapustina, M. & Carter, C.W., 2006. Computational studies of tryptophanyl-tRNA synthetase: activation of ATP by induced-fit. Journal of molecular biology, 362(5), pp.1159–80.

Lanzarotti, E. et al., 2011. Aromatic Aromatic Interactions in Proteins: Beyond the Dimer. J. Chem. Inf. Model., 51, pp.1623–1633.

Laowanapiban, P. et al., 2009. Independent saturation of three TrpRS subsites generates a partially assembled state similar to those observed in molecular simulations. Proceedings of the National Academy of Sciences of the United States of America, 106(6), pp.1790–5.

Lauck, F. et al., 2010. RosettaBackrub--a web server for flexible backbone protein structure modeling and design. Nucleic acids research, 38(Web Server issue), pp.W569–75.

Lindahl, E., Hess, B. & Spoel, D. Van Der, 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. Molecular modeling annual, pp.306–317.

Maragakis, P. & Karplus, M., 2005. Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. Journal of Molecular Biology, 352(4), pp.807–822.

Na, H. & Song, G., 2014. Bridging between normal mode analysis and elastic network models. Proteins.

Nicholson, L.K. & Lu, K.P., 2007. Prolyl cis-trans Isomerization as a Molecular Timer in Crk Signaling. Molecular Cell, 25(4), pp.483–485.

Onsager, L. & Machlup, S., 1953. Fluctuations and irreversible processes. Physical Review, 91(6), pp.1505–1512.

Ovchinnikov, V., Karplus, M. & Vanden-Eijnden, E., 2011. Free energy of conformational transition paths in biomolecules: the string method and its application to myosin VI. The Journal of chemical physics, 134(8), p.085103.

Ovchinnikov, V., Trout, B.L. & Karplus, M., 2010. Mechanical coupling in myosin V: a simulation study. Journal of molecular biology, 395(4), pp.815–33.

Phillips, J.C. et al., 2005. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry, 26(16), pp.1781–1802.

Pinski, F.J. & Stuart, A.M., 2010. Transition paths in molecules at finite temperature. The Journal of Chemical Physics, 132(18), p.184104.

Secemski, I.I., Lehrer, S.S. & Lienhard, G.E., 1972. A transition state analog for lysozyme. Journal of Biological Chemistry, 247(15), pp.4740–4748.

Shaw, D.E. et al., 2010. Atomic-Level Characterization of the Structural Dynamics of Proteins. Science, 330(6002), pp.341–346.

Shirvanyants, D. et al., 2012. Discrete molecular dynamics: An efficient and versatile simulation method for fine protein characterization. Journal of Physical Chemistry B, 116(29), pp.8375–8382.

Skalicky, J.J. et al., 2001. Aromatic Ring-Flipping in Supercooled Water: Implications for NMR-Based Structural Biology of Proteins. J. Am. Chem. Soc., 123, pp.388–397.

Sugita, Y. & Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters, 314(1-2), pp.141–151.

Torrie, G.M. & Valleau, J.P., 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. Journal of Computational Physics, 23(2), pp.187–199.

Watt, E.D. et al., 2007. The mechanism of rate-limiting motions in enzyme function. Proceedings of the National Academy of Sciences of the United States of America, 104(29), pp.11981–6.

Weinreb, V. et al., 2014. Enhanced amino acid selection in fully evolved tryptophanyl-tRNA synthetase, relative to its urzyme, requires domain motion sensed by the D1 switch, a remote dynamic packing motif. The Journal of biological chemistry, 289(7), pp.4367–76.

Weinreb, V., Li, L. & Carter, C.W., 2012. A master switch couples Mg2+-assisted catalysis to domain motion in B. stearothermophilus tryptophanyl-tRNA Synthetase. Structure (London, England : 1993), 20(1), pp.128–38.

Williams II, B. et al., 2015. ApoE4-specific Misfolded Intermediate Identified by Molecular Dynamics Simulations N. Ben-Tal, ed. PLOS Computational Biology, 11(10), p.e1004359.

Yang, Z., Májek, P. & Bahar, I., 2009. Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. PLoS computational biology, 5(4), p.e1000360.