جامعة بيرزيت

BIRZEIT UNIVERSITY

INVESTIGATING CONFORMATIONAL AND DYNAMIC DIFFERENCES AMONG RAS ONCOGENIC MUTANTS THROUGH MD SIMULATIONS AND MARKOV MODELS

دراسة الفروق الديناميكية والتكوينية بين بعض الطفرات المسرطنة لبروتين (Ras) من خلال المحاكاة الديناميكية للجزيئات ونماذج ماركوف

Mohammed Khaled

June 2019

Thesis committee:

Dr. Abdallah Sayyed-Ahmad (Principal advisor)

Dr. Wael Karain (Member)

Dr. Hazem Abu Sara (Member)

This Thesis was submitted in partial fulfillment of the requirements for the degree of Master of Physics from the Faculty of Graduate Studies at Birzeit University, Palestine.

INVESTIGATING CONFORMATIONAL AND DYNAMIC DIFFERENCES AMONG RAS ONCOGENIC MUTANTS THROUGH MD SIMULATIONS AND MARKOV MODELS

دراسة الفروق الديناميكية والتكوينية بين بعض الطفرات المسرطنة لبروتين (Ras) من خلال المحاكاة الديناميكية للجزيئات ونماذج ماركوف

By

Mohammed Khaled

Accepted by the Faculty of Graduate Studies, Birzeit University, in partial fulfillment of the requirements of the degree of Master of Physics.

Thesis committee:

_____

Abdallah Sayyed-Ahmad Ph.D. (Principal advisor)

_____

Wael Karain Ph.D. (Member)

_____

Hazem Abu Sara Ph.D. (Member)

June, 2019

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to my thesis supervisor Dr. Abdallah Sayyed-Ahmad for supporting me to work on the exciting field of molecular dynamic simulations and Markov state models. He assisted me to improve my research and computational skills throughout this work. I would also like to thank of my thesis committee members, Dr. Hazem Abusara and Dr. Wael Karain, for reviewing my thesis.

DEDICATION

To my beloved parents, who have been a source of encouragement and inspiration to me throughout my life. To my brother and sisters; particularly my twin sister Nebras. To my teachers and friends who shared their advice and encouragement to finish this study.

ABSTRACT

Mutations in protein amino acids are usually associated with several challenging diseases such as cancer, sickle cell anemia, and Alzheimer. Ras proteins are small GTPase proteins that regulate the signaling pathway for cell growth, proliferation and differentiation. Mutations of Ras proteins are observed in many human cancers. Therefore, numerous studies of Ras proteins are carried out to investigate mutations on their conformational differences, dynamics, allosteric communications and signaling. In this work, we used molecular dynamics simulations and Markov state models to study the effects of TYR32 phosphorylation on G12D K-Ras. Markov state models provided us with a coarse-grained picture consisting of few metastable state conformations. They also helped us identify the probability of each one of these states along with the kinetic transition rates among them. We show that the phosphorylation significantly alters switch region conformations and dynamics. In addition, we show that G12D K-Ras and its phosphorylated mutant exhibit different conformational states.

ملخص

تعتبر الطفرات في تركيبة سلسلة الحموض الأمينية للبروتينات أحد العوامل التي تتسبب في العديد من الأمراض المستعصية مثل السرطان وفقر الدم المنجلي والزهايمر. تعد بروتينات Ras جزءً من عائلة إنزيمات تسمى GTPase الصغيرة . هذه الإنزيمات مسؤولة عن عدة وظائف داخل الخلية منها نقل الإشارات داخل الخلية، حيث تقوم بدورها في عملية نمو وانقسام وتكاثر الخلايا. تسبب بعض طفرات بروتين Ras العديد من الاورام السرطانية. لذلك تمت دراسة الفروق التكوينية والتأثيرات الديناميكية والاشارات والتفاعلات الخلوية لهذه البروتينات من خلال عدد كبير من الدراسات. في هذا البحث استعملنا المحاكاة الديناميكية للجزيئات ونماذج ماركوف لدراسة تأثير فسفرة الحمض الأميني تيروسين 32. نماذج ماركوف تصور لنا التغيرات التكوينية والديناميكية للبروتين بشكل أبسط من خلال تعريف عدد قليل من الحالات الشبه مستقرة وتحديد احتمالاتها حدوثها ومعدلات الانتقال فيما بينها. في الدراسة الحالية أيضا أثبتنا أن الفسفرة تؤثر بشكل خاص على ديناميكية وتكوينية البروتين. بشكل خاص على مناطق معينة في البروتين تسمى المفاتيح وتتميز كل طفرة بحالات تكوينية مختلفة عن الأخرى.

TABLE OF CONTENTS

Acknowledgement ii
Dedication iii
Abstract iv
ملخص v
List of Figures viii
List of Tables x
Chapter 1: Introduction 1
1.1. K-Ras protein 1
1.2. Molecular dynamic simulations 5
Chapter 2: Markov State Models 7
2.1. Theory 8
2.1.1. Exact dynamics in full configuration space 8
2.1.2. Approximation of slowest timescales and the related eigenfunctions: 10
2.1.3. Best approximation of the eigenfunctions 11
2.2. Dimension reduction and discretization 13
2.2.1. Principal component analysis (PCA) 13
2.2.2. Time-lagged independent component analysis (TICA) 14
2.2.3. Discretization of state space 15
2.3. Transition matrix 16
2.4. Estimation of Markov state models 18
2.4.1. Count matrix 18
2.4. 2. Likelihood and Bayesian estimators 19
2.5. Kinetic model validation: Chapman-Kolmogorov test 20
2.6. Transition path theory 21
2.7. Summary 22
Chapter 3: Material and Methods 24
3.1. Initial structures and system preparation 24
3.2. Molecular dynamics simulation 25
3.3. Building Markov state models and their validation 26

Chapter 4: Results and Discussion 31
4.1. Effects of Tyr32 phosphorylation on structure and dynamics of G12D K-Ras. 31
4.2. Analyze side chain torsion for Tyr32 in both cases 35
4.3. Principal component analysis 36
4.4. GTP binding site configuration 37
4.5. Sodium ion interaction 38
4.6. Markov state models analysis 39
4.7. Conclusions 44
References 46
Appendices 51
Appendix A: NAMD Molecular dynamics scripts 51
A.1: TCL script to add harmonic restrains on CA atoms 51
A.2: Equilibration script 51
A.3: Production script 54
Appendix B: Markov state models software scripts 56

LIST OF FIGURES

Figure 1.1: G12D K-Ras sequence and structure. (A) The sequence of the amino acids of the catalytic domain (amino acids 1−166) and HVR (amino acids 167−189) of G12D K-Ras proteins. The SI and SII regions are highlighted by bold font, while the residues 12 and 32 highlighted in red. (B) G12D K-Ras structure shown in cartoon with the location of the mutations studied in this work. Residues 12 and 32 are highlighted by purple and green spheres respectively. Switches SI and SII (residues 60-75) are in red and blue respectively. 2

Figure 2.1: Flowchart for building MSMs. 23

Figure 3.1: A snapshot from one of the simulations. The catalytic domain is in cartoon colored in dark grey and switches SI (residues 25-40) and SII (residues 60-75) are in red. Magnesium, sodium and chloride ions as shown in blue, yellow and green spheres, respectively. The bound GTP is in purple sticks. 25

Figure 3.2: Projection of the trajectories onto the two largest independent components. Projection of the trajectory onto the two largest independent components subset estimated by TICA. Gray lines represent data sampled every 100 ps while thick black lines represent 10 ns running averages. 27

Figure 3.3: Clustering the trajectory into microstate. The trajectory is assigned to the 100 cluster centers using k-means clustering. (A) for pTyr32-G12D and (B) G12D K-Ras. 27

Figure 3.4: Validation of Markov state model. Implied relaxation timescales for the first six eigenvalues calculated from the transition matrix at different lag times. All relaxation timescales become approximately constant beyond the used lag-time 2 ns to construct MSMs for (A) G12D and (B) pTyr32-G12D. 28

Figure 3.5: The Spectral analysis of timescale separation. The Spectral analysis revealed the largest timescale separation is between the first and the second relaxation timescales for G12D, and third and fourth relaxation timescales for pTyr32-G12D, respectively. 29

Figure 3.6: The Chapman-Kolmogorov test. The Chapman-Kolmogorov test for MSMs for (A) pTyr32-G12D and (B) G12D with the five states. The data for MSM (black line) and the MDs trajectory (blue dotted line, with estimated error). 30

Figure 4.1: $C\alpha$ root mean square fluctuation (RMSF). RMSF calculated after alignment excluding SI and SII regions for both G12D K-Ras (black) and its phosphorylated variant (red). SI and SII are highlighted by cyan and purple shadows, respectively. 32

Figure 4.2: Time evolution of root-mean square deviation (RMSD). RMSD values of backbone (BB), Switch 1 (SI) and Switch 2 (SII) structures sampled from the initial X-Ray structure for both G12D K-Ras and its phosphorylated variant. RMSDs were calculated after structural alignment excluding the flexible switch regions. Gray lines represent data sampled every 100 ps while thick black lines represent 10 ns running averages. 33

Figure 4.3: The time evolution of distance between the C_α atoms of the residues Asp38 an Asp57 of both mutants G12D K-Ras (black) and its phosphorylated variant (red). 35

Figure 4.4: The probability of the dihedral angle $\chi 1$(N-Cα-Cβ-Cγ) of Tyr32 for G12D K-Ras (black) and its phosphorylated variant (red). 36

Figure 4.5: Global conformational dynamics of mutants G12D K-Ras and its phosphorylated variant. Projection of simulated trajectories into the first and the second principal components of mutants G12D K-Ras (black) and its phosphorylated variant (red). 37

Figure 4.6: Long-residence sodium-ion binding sites. A snap shot of pTyr-G12Dshowing a sodium ion interacting with GTP and SI. 39

Figure 4.7: The five metastable states grouped from microstates. The trajectory cluster into microstate by assigning it to the 100 cluster centers using k-means clustering. The microstates grouped by (PCCA++) method into five metastable states. (A) pTyr32-G12D. The color code of metastable states 1 (blue), 2 (gray), 3 (black), 4 (green), 5 (purple). (B) G12D K-Ras. The color code of metastable states I (blue), II (gray), III (black), IV (green), V (purple). 40

Figure 4.8: A network diagram of the five metastable states identified by the Markov state model. The metastable states are represented by circles, the arrows indicate the transition probabilities between the states. The structures describe the metastable states found in the MSM analysis, each circle illustrating ten representative protein conformations (generated using MSM), which identify also the SI (red) and SII (blue) regions. (A) pTyr32-G12D. (B) G12D K-Ras. The circle colors are the same as in figure 4.7. 42

Figure 4.9: RMSD matrices of SI and selected C_α atoms computed from MSMs metastable states trajectories. The metastable states of pTyr32-G12D indicate by numbers (1-5) and the metastable states of G12D K-Ras indicate by Roman numbers (I-V). 43

LIST OF TABLES

Table 4.1: Average Distance of T35 and G60 from GTP. 38

Table 4.2: The stationary probability and the free energy of metastable states of G12D K-Ras and its phosphorylated variant. 40

Table 4.3: The maximum four fluxes path of G12D K-Ras and its phosphorylated variant. 43

CHAPTER 1: INTRODUCTION

Understanding the dynamics of a protein and its three-dimensional structure from their amino acid sequences is important to decipher its function and malfunction. For example, many diseases are associated with changes in protein structure such as sickle cell anemia, Alzheimer and cancer. In this thesis, we are interested in studying Ras mutations which have been shown to play a critical role in many human cancers. In particular, phosphorylation of Tyr32 in K-Ras has evidently been shown to influence its catalytic activity and function by disrupting its GTPase cycle, and hence leading to different types of cancer. In this work, we investigated the conformational and dynamical effects of Tyr32 phosphorylation in G12D K-Ras by molecular dynamics (MD) simulation and Markov state models. The thesis is organized as follows: First, we give a brief introduction about K-Ras protein and MD simulations. In the second chapter, we briefly introduce Markov state models and the different related techniques required to build it. In chapter 3, we describe the setup of our MD simulations systems and the procedure used to carry out these simulations. We also describe the method used to build Markov state models from MD simulation trajectories. Finally, the last chapter discusses the results obtained from MD simulations and Markov state models constructed for G12D K-Ras and its phosphorylated variant.

1.1. K-Ras protein

Kirsten Rat Sarcoma (K-Ras) protein is one of the best characterized and ubiquitously expressed GTPases in most human cells [1]. It selectively attaches to the inner leaflet surface of the plasma via a farnesylated C-terminus and polybasic domain [2, 3]. K-Ras structure features two main domain components: the catalytic domain (Figure 1.1B, amino acids 1−166) and the membrane

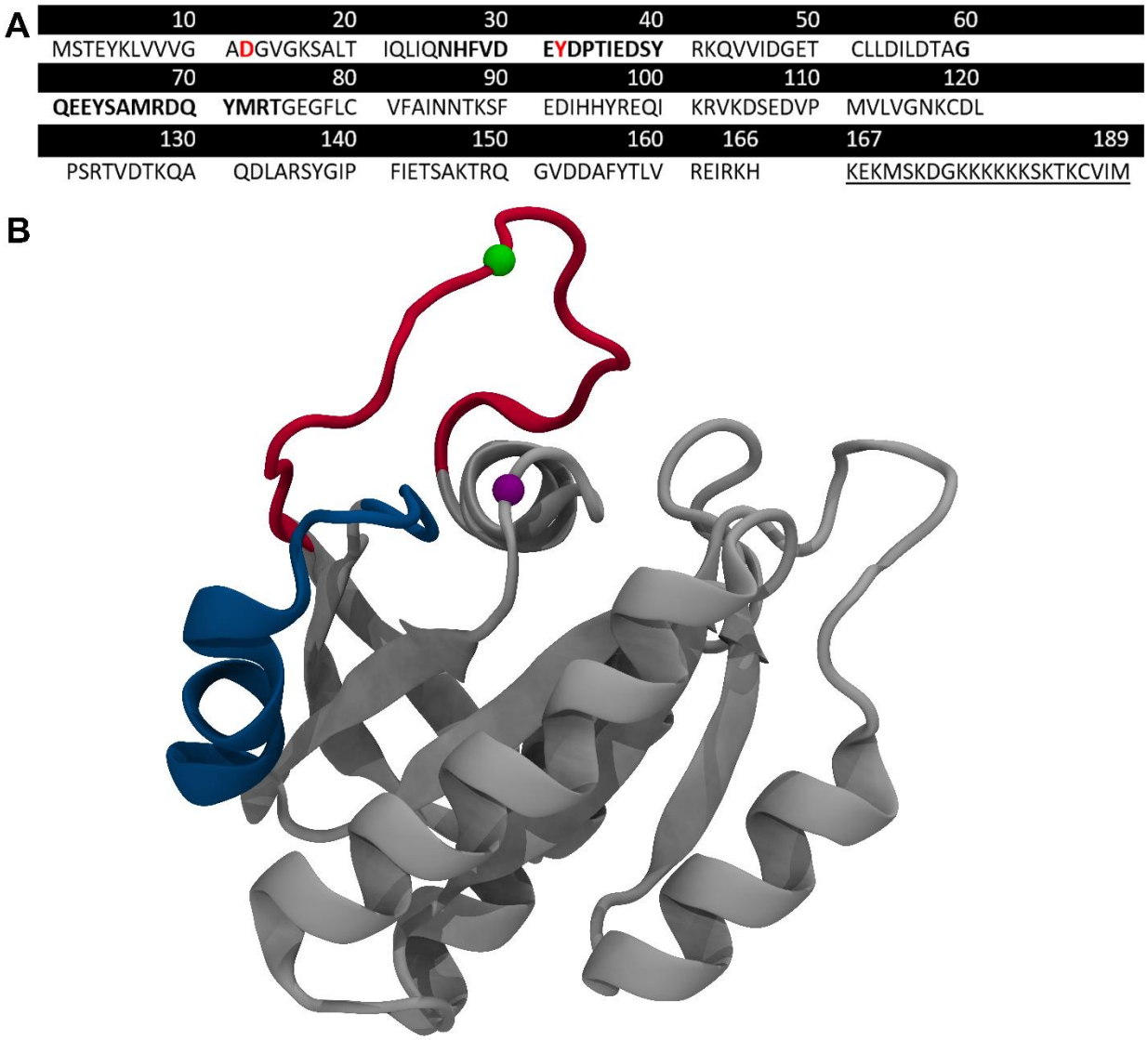

Figure 1.1: G12D K-Ras sequence and structure. (A) The sequence of the amino acids of the catalytic domain (amino acids 1−166) and HVR (amino acids 167−189) of G12D K-Ras proteins. The SI and SII regions are highlighted by bold font, while the residues 12 and 32 highlighted in red. (B) G12D K-Ras structure shown in cartoon with the location of the mutations studied in this work. Residues 12 and 32 are highlighted by purple and green spheres respectively. Switches SI and SII (residues 60-75) are in red and blue respectively.

targeting (HVR region, amino acids 167−189 + farnesyl anchor). The latter membrane anchoring region is not conserved and has notable sequence differences among Ras isoforms (Figure 1.1B). Conversely, the highly conserved catalytic domain of K-Ras interacts with effectors and

exchange factors by modulating the conformations of two flexible canonical switches: switch 1 (SI: residues 25−40) and 2 (SII: residues 60−75). These conformation changes are primarily modulated by the hydrolysis of GTP molecule to GDP which represents the molecular switch state changing from the active to the inactive state [4, 5]. In GTP bound state, the two switches and the P-loop (residues 10−17) form the closed conformation of GTP binding site. Ras is activated by guanine nucleotide-binding site and the Ras-GTP binding activate downstream effector effectors, including Raf kinase, PI3K, and Ral guanine nucleotide dissociation stimulator (RalGDS) [6-9]. Similar to other Ras isoforms, it regulates the slinging pathways controlling growth, proliferations and differentiation of cells [10-12].

It has been shown that in oncogenic Ras, mutations affect GTP hydrolysis [13, 14], and that mutations at 32 position particularly exhibit reduced catalytic activity [15]. The activity of GTPases is also decreased by the mutation of K-Ras G12X that lead to increased K-Ras signaling and more active GTP-bound present [16, 17]. Notably, G12D K-Ras mutation is the most frequently mutated oncogenic found in human cancer. Most G12X mutations show insensitivity to GTPase-activating protein (GAPs) that accelerate GTP hydrolysis [19]. Furthermore, oncogenic Ras mutants activate the downstream effectors that promote cell proliferation, consequently leading to tumor development. [18, 20]. The role of Tyr32 in determining configuration of the active site has been long established. It has been shown that Tyr32 has a critical role in inducing the conformational change in Ras that modulates its GTPase activity and the effector binding [21-23]. Gorfe and Coworkers suggested that Tyr32 orientation along with the relative arrangement of SI and SII can be used to uniquely determine the active and inactive conformations in many experimental and simulated structures [24].

The phosphorylation of protein has a significant effect on its function effects and conformational states. It usually alters the local chemical environment by creating a chemical shift in the modified residues and their adjacent residues [25, 26]. For example, the phosphorylation and dephosphorylation process of Ras modulate its activity and they are mainly mediated by Src and SHP2 protein tyrosine phosphate (PTP) [27]. The phosphorylation process induced by Src regulates the GTPase cycle by impairing Raf binding [28]. In contrast, the dephosphorylation process induced by SHP2 negatively impact the GTPase cycle by enhancing Raf binding [29]. Therefore, any disruption to the balance between these processes may lead to adverse functional effects and various cancers.

Recent biophysical experiments have suggested that phosphorylation of Tyr32 of K-Ras attenuates its sensitivity to GAP and GEF activities which induced intrinsic nucleotide exchange and impair intrinsic GTP hydrolysis. It reduces the binding affinity of K-Ras to its effector Raf [30]. Tyr32 phosphorylation is thought to alter SI and SII conformations as a result of the additional electrostatics repulsion against the negatively charged Asp38 and Asp57 residues within the nucleotide-binding site [31]. The conformation changes in the orientation of SI which significantly affects the affinity of Ras for its effector proteins Raf, lead to reducing downstream signaling mitogen-activated extracellular signal-regulated kinase (MEK)-to-extracellular signal-regulated kinase (ERK) and phosphoinositide-3 kinase-to-AKT signaling [31].

Although these experimental studies have revealed the importance of Tyr32 phosphorylation, its effects on K-Ras structure and dynamics are still not fully understood. Several MD simulation studies investigated the structure, dynamics and function of Ras oligomerization, isoforms or mutants [32-40]. To that end, in this work we investigate the underlying structural and dynamical changes that lead to effects observed due to phosphorylation

of Tyr32 of K-Ras. We therefore carried out two 500 ns MD simulations of G12D K-Ras and pTyr32-G12D. Also, we identified metastable conformational states and the kinetic network of transitions between them using Markov state model (MSM).

1.2. Molecular dynamic simulations

MD simulations are commonly utilized to study the physical and chemical properties of a system of molecules or atoms. In fact, they are currently one of the important methods to study the structure and function of proteins. They also enable us to estimate protein physical properties that might be difficult to access through experiments. In the last decades, protein simulations had a big evolution due to the use of supercomputers and development of new more efficient techniques [36, 41-49].

In MD simulations, the physical motions of atoms in a protein are resolved by the integration of Newton equations of motion for every atom in the system in which the initial coordinates are taken from X-ray crystal or NMR structures [50]:

$$m_i \frac{d^2\vec{r}_i}{dt^2} = -\vec{\nabla}_i V(\vec{r}_1, \vec{r}_2, \ldots\ldots, \vec{r}_N) \tag{1.1}$$

where m_i and $\vec{r}_i$ are the mass and coordinates of each atom in the system respectively. The potential energy, V, is estimated using a force field for a system of N interacting atoms. It is given by

$$V = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{\substack{improper \\ dihedrals}} \frac{1}{2} k_\xi (\xi - \xi_0)^2 + \sum_{dihedrals} k_\phi [1 + \cos(n\phi - \delta)] + \sum_{\substack{atom \\ pairs}} \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}} + \epsilon_{ij} \left(\frac{R_{min}^{12}}{r_{ij}^{12}} - \frac{R_{min}^{6}}{r_{ij}^{6}} \right) \quad (1.2)$$

where r_0 and θ_0 are the equilibrium bond length and angle, respectively. ϕ is the dihedral angle and n is the multiplicity of the function. k_b, k_θ, k_ξ and k_ϕ are the bond, angle, improper dihedral and dihedral constants, respectively. ϵ and R_{min} are the well depth and zeros of the Lennard-Jones potential. ϵ_0 is the free space permittivity and q is the localized charge on each atom. Finally, r_{ij}'s are the inter-particle distances. The Force field function contains bonded and non-bonded terms. Bonded terms include harmonic oscillator energy of bond lengths, bond angles, and sometimes improper dihedrals as well as torsional dihedral angles. The non-bonded terms include Van der Waals interactions and electrostatic interactions. Examples of force fields available to study proteins are CHARMM [51], AMBER [52] and GROMOS [53].

CHAPTER 2: MARKOV STATE MODELS

Markov state models [54] deal with a statistical process called Markov chains in which only the current state of a system can affect its transition to the next state. The state is considered memoryless if the future state depends only on the current state and not on previous states in the system. The system is called Markovian if all states in the system are memoryless. In the late 1990s, Schütte and his colleagues used Markov state models to understand MD trajectories [55]. Markov state models and related techniques have also been developed with the availability of significant computing power for a few research groups since the mid 2000s [56-58]. Nowadays, Markov state models (MSMs) have become a popular technique in computational biophysics for the identification of stationary and kinetic states from MDs trajectories. User-friendly software is available for building Markov state models such as MSMbuilder [59] and PyEMMA [60]. MSMs have been used to analyze many complex molecular processes such as protein folding [61], protein-ligand binding [62, 63], peptide dynamic [64] and peptide aggregations [65].

MSMs are also used for determining molecular kinetics [66]. They can readily describe the slow relaxation processes by kinetic characterization. They can identify the structural changes for these processes and approximate the rates and time scales at which they occur. The model built to approximate a MD trajectory by a Markov chain requires partitioning the conformation space into discrete states. It estimates the kinetic behavior of the system with transition probability matrix that helps find the system in any discrete state after a fixed time τ. The discrete description of molecular kinetics approximates the exact eigenvector and eigenvalue of the propagator of continuous dynamics.

$$\mathcal{P}\phi_i = \lambda_i \phi_i \tag{2.1}$$

where $\mathcal{P}$ is the transfer operator that propagates probability densities of molecular configurations, ϕ_i are its eigenfunctions, and λ_i are the associated eigenvalues. Eq.(2.1) defines all stationary and kinetic quantities when solved for eigenvalues and associated eigenfunctions.

Let $r_1, \ldots, r_d \in \mathbb{R}$ be a possible large set of d order parameters of a molecular system that are *a priori* specified. We aim to find a linear combination of these order parameters that optimally approximate the subspace spanned by the dominant eigenfunctions. Here, the variational principle of conformation dynamics is used to get the best solution for the problem. Furthermore, Time-lagged independent component analysis (TICA) that combine information from the covariance matrix and a time-lagged covariance matrix of the data is used for constructing Markov models. Principal component analysis (PCA) is another method used for dimension reduction of an order parameter space by projecting it on its linear subspace of the largest amplitude motion. But, in this case slow modes are not necessarily associated with large amplitudes.

2.1. Theory

2.1.1. Exact dynamics in full configuration space

Let x_t be the full molecular configuration at time t in a state or phase space Ω. We assume that the MD are statistically reversible Markovian in Ω and the stationary density $\mu(x)$ is given by a Boltzmann distribution density:

$$\mu(\mathrm{x}) = Z^{-1} e^{-\beta H(\mathrm{x})} \tag{2.2}$$

where H is the Hamiltonian of the system, Z is the partition function, and $\beta = (k_B T)^{-1}$ is the inverse temperature.

If the propagator $\mathcal{P}(\tau)$ acts on the probability density of molecular configuration ρ_t, it will describe the probability that a trajectory at configuration x_t at time t will be found at a configuration $\mathrm{x}_{t+\tau}$ after τ time interval:

$$\rho_{t+\tau} = \mathcal{P}(\tau)\rho_t \tag{2.3}$$

The propagator can be written by expanding it in terms of its eigenvalues as

$$\lambda_i(\tau) = e^{-\tau/t_i} \tag{2.4}$$

where t_i are the corresponding timescales; it relates to eigenvalues and experimentally measurable relaxation rate κ_i of the system as

$$t_i^{-1} = \kappa_i = -\frac{ln\lambda_i}{\tau} \tag{2.5}$$

And its eigenfunctions ϕ_i can be written as:

$$\rho_{t+\tau}(y) = \mathcal{P}(\tau)\rho_t(\mathrm{x}) = \sum_{i=1}^{\infty} e^{-\tau/t_i}\langle\psi_i, \rho_t\rangle\phi_i \tag{2.6}$$

The first eigenvalue is $\lambda = 1$ with first relaxation timescales $t_1 = \infty$ and correspond to the stationary distributions, while the remaining eigenvalues have a norm strictly smaller than 1 with finite relaxation timescale t_i. $\psi_i(\mathrm{x})$ are the weighted eigenfunctions by a stationary density where $\psi_i(\mathrm{x}) = \mu^{-1}(\mathrm{x})\phi_i(\mathrm{x})$ .The scalar product $\langle\psi_i, \rho_t\rangle$ represents the overlap of the starting density ρ_t with i^{th} eigenfunction. It determines the amplitude by which the eigenfunction

contributes to the dynamics. The contributions of all basis function ϕ_i to the probability density $\rho_{t+\tau}$ decrease with time. After infinite time $\tau \to \infty$ only the first term with $t_1 = \infty$ is left and the stationary density is reached: $\lim_{\tau\to\infty} \mathcal{P}(\tau)\rho_t = \phi_1 = \mu$.

At large times, the dynamics will be governed by m largest timescales. So, we are interested just in slowest timescales with $\tau \gg t_{m+1}$. At these timescales all the kinetic properties and stationary distributions can be accurately approximated when only the dominant m eigenvalues and eigenvectors are used [66]:

$$\rho_{t+\tau} = \mathcal{P}(\tau)\rho_t \approx \sum_{i=1}^{m} e^{-\tau/t_i}\langle\psi_i, \rho_t\rangle\phi_i \tag{2.7}$$

2.1.2. Approximation of slowest timescales and the related eigenfunctions:

From Eq.(2.6) the time autocorrelation function of some function of molecular configuration $f(\mathrm{x})$ as a function of τ is given by:

$$\langle f(\mathrm{x}_t) f(\mathrm{x}_{t+\tau})\rangle_i = \sum_{i=1}^{\infty} e^{-\tau/t_i}\langle\phi_i, f\rangle^2 \tag{2.8}$$

Since the two eigenfunctions $\psi_i(\mathrm{x})$ and $\phi_i(\mathrm{x})$ are interchangeable, we can use the time autocorrelation function of the eigenfunction $\psi_i(\mathrm{x})$ to yield the exact i^{th} eigenvalue [67], and thus permit us to get the exact i^{th} timescale:

$$\hat{\lambda}_i(\tau) = \langle\psi_i(\mathrm{x}_t)\psi_i(\mathrm{x}_{t+\tau})\rangle_i = e^{-\tau/t_i} \tag{2.9}$$

$$\hat{t}_i = -\frac{\tau}{ln\left|\hat{\lambda}_i(\tau)\right|} = t_i \tag{2.10}$$

In practice, we cannot know the exact eigenfunctions $\psi_i(\mathrm{x})$.Hence, the variational principle of conformation dynamics [67] can be used to construct a model function for $\psi_i(\mathrm{x})$ such that the normalized time-autocorrelation function ($\hat{\psi}_i(\mathrm{x})$) approximates the true eigenvalues and timescales

$$\langle \hat{\psi}_i(\mathrm{x}_t)\hat{\psi}_i(\mathrm{x}_{t+\tau})\rangle \leq e^{-\tau/t_i} \tag{2.11}$$

and

$$\hat{t}_i \leq t_i \tag{2.12}$$

Therefore, we must look for a function $\hat{\psi}_i$ that has the maximum timescale t_i for finding the best approximation of the i^{th} timescale and its associated eigenfunction. All of the first m timescales will however be underestimated when the Markov model is used to approximate the slowest processes. It was shown [68, 69] that the estimation error becomes smaller when τ is increased. As a result, when plotting the estimated timescales $\hat{t}_i(\tau)$ as a function of τ one obtains the well-known implied timescale, where the estimated timescales $\hat{t}_i(\tau)$ slowly converge to the true timescale as τ is increased.

2.1.3. Best approximation of the eigenfunctions

To approximate the eigenfunctions ψ_i by a functions $\hat{\psi}_i$that is a linear combination of basis functions (χ_k) which must be *a priori* defined by

$$\hat{\psi}_i(\mathrm{x}) = \sum_{k=1}^{n} b_{ik}\chi_k(\mathrm{x}) \tag{2.13}$$

The problem is to find the optimal parameters b_{ik} that will be denoted by a vector $\boldsymbol{b}_i \in \mathbb{R}^n$.The coefficients $\boldsymbol{b}_i$ will give us the optimal approximation of the eigenvalues and its corresponding timescales. We aim to get the optimal set of coefficients for an orthogonal basis set that requires them to be uncorrelated at lag-time 0:

$$c_{ij}^{\chi}(0) = \langle \chi_i, \chi_j \rangle_\mu = \langle \chi_i(\mathrm{x}_t)\chi_j(\mathrm{x}_t) \rangle_t = \delta_{ij} \tag{2.14}$$

If the covariance matrix at lag-time τ between functions is defined by

$$c_{ij}^{\chi} = \langle \chi_i(\mathrm{x}_t)\chi_j(\mathrm{x}_{t+\tau}) \rangle_t \tag{2.15}$$

then the eigenvector $\mathbf{b}_i$ gives us the optimal set of the coefficient

$$C^{\chi}(\tau)b_i = b_i\hat{\lambda}_i(\tau) \tag{2.16}$$

For the more general case of a non-orthonormal basis set, the optimal approximation to the exact eigenvalues and eigenfunctions is obtained by solving the generalized eigenvalue problem:

$$C^{\chi}(\tau)b_i = C^{\chi}(0)b_i\hat{\lambda}_i(\tau) \tag{2.17}$$

The two equations Eq.(2.16) and Eq.(2.17) are known from variational calculus. To get the optimal approximation of exact eigenfunctions we need to solve Eq.(2.17) with correlation matrix for lag- time 0 (principal component analysis) and τ (Time-lagged independent component analysis) that will provide us with the linear combination of order parameters. You need to make clear that PCA and TICA give two different results. They are not complementary, are they?

2.2. Dimension reduction and discretization

2.2.1. Principal component analysis (PCA)

MD usually utilizes principal component analysis to identify a linear subspace in which the large amplitude motions. PCA transforms linearly the coordinates provided that their instantaneous correlation vanishes [70, 71].

.The elements of the covariance matrix $\boldsymbol{C^r}$ of the order parameter $\boldsymbol{r}$ is defined by

$$c_{ij}^{r}(0) = \langle r_i r_j \rangle \tag{2.18}$$

PCA transforms the data into orthogonal basis, where the new coordinates are uncorrelated for $i \neq j$ .The principle eigenvectors w_i can be obtained by solving the eigenvalue problem:

$$C^r w_i = w_i \sigma_i^2 \tag{2.19}$$

or in matrix form

$$C^r W = W \Sigma^2 \tag{2.20}$$

where $W = [w_1, \dots, w_d]$ is the eigenvector matrix and $\Sigma^2 = diag(\sigma_1^2, \dots, \sigma_d^2)$ is the variance matrix. The eigenvalues of the matrix measure the variance of the data along the principal direction, while the eigenvectors are used to transform an original coordinate vector $\boldsymbol{r}$ into principal components:

$$y^T = r^T W \tag{2.21}$$

If the variance σ_i^2 is decay quickly with i, one often selects a threshold and ignores all PCs with smaller σ_i^2. PCA used as dimension reduction tool by using the first m dominant column vectors of **W**. The fraction of variance of dimension reduction is

$$V_d = \frac{\sum_{i=1}^{m} \sigma_i^2}{TV} \tag{2.22}$$

where TV is the total variance: $TV = \sum_{i=1}^{d} \sigma_i^2$.

2.2.2. Time-lagged independent component analysis (TICA)

TICA [72] applies a linear transformation to the order parameter. It is an optimal method to detect the slow reaction coordinates and their relaxation timescales. TICA uses a time-lagged covariance matrix $\boldsymbol{C}^r(\tau)$ to get a new set of order parameter that are uncorrelated and their autocovariances at a fixed lag-time τ are maximum.

$$c_{ij}^r(\tau) = \langle r_i(t) r_j(t+\tau) \rangle \tag{2.23}$$

To get uncorrelated independent component $\boldsymbol{C}^r(0)$ diagonalized by transformation matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_d]$, and maximizes the autocorrelations $c_{ii}^z(\tau) = \boldsymbol{u}_i^T \boldsymbol{C}^r(\tau) \boldsymbol{u}_i$ for every column $\boldsymbol{u}_i$ of **U**. One then solves the generalized eigenvalue problem:

$$\boldsymbol{C}^r(\tau)\boldsymbol{u}_i = \boldsymbol{C}^r(0)\boldsymbol{u}_i \hat{\lambda}_i(\tau) \tag{2.24}$$

The independent components $\boldsymbol{y}(t)$ are then obtained from the original coordinate vector $\boldsymbol{r}(t)$ as follows

$$\boldsymbol{y}^T = r^T U \tag{2.25}$$

The new order parameter or independent component with largest autocovariances $\hat{\lambda}_i(\tau)$ will be called dominant. If we have m dominant IC's which describe most of slow processes, the fraction of the kinetic variance retained by the dimension reduction is

$$c_m = \frac{\sum_{i=2}^{m} \lambda_i^2(\tau)}{TKV} \tag{2.26}$$

where the total kinetic variance is

$$TKV = \sum_{i=2}^{d} \lambda_i^2(\tau) \tag{2.27}$$

The squared eigenvalues λ_i^2 are in the range [0, 1], where values near 1 are for slow processes and value near are 0 for fast processes. This means that TKV measures the number of slow processes found in the data.

2.2.3. Discretization of state space

In Markov models a step-function basis is chosen to build it. This gives an optimal step-function approximation to the eigenfunctions and maximal eigenvalues amongst all choices of functions that can be supported by the clustering. Markov models assigning every configuration (x) uniquely to one of the geometric clusters that will be used to construct the model. It can be shown [73] that this operation is equivalent to using basis functions. The discretization of state space Ω into n sets $S = \{S_1, \dots, S_n\}$ which entirely partition the state space and have no overlap. The probability of a point x to belong to set i is found by a membership function $\chi_i(\mathrm{x})$ with the property $\sum_{i=1}^{n} \chi_i(\mathrm{x}) = 1$. The state space is partitioned using crisp partition with step functions:

$$\chi_i(\mathrm{x}) = \begin{cases} 1 & \mathrm{x} \in S_i \\ 0 & \mathrm{x} \notin S_i \end{cases} \tag{2.28}$$

Each basis function χ_i is a step function which has a constant value for all configurations belonging to the i^{th} cluster and is zero elsewhere. The crisp partition considers n centers $\bar{\mathrm{x}}_i$, $i = 1, \dots, n$, and set S_i defines as all the point $\mathrm{x} \in S_i$ that are closer to $\bar{\mathrm{x}}_i$ more than any other center [74]. This basis is an orthonormal basis set where

$$\langle \chi_i, \chi_j \rangle_\mu = \frac{1}{\pi_i} \int_{\mathrm{x} \in S_i}^{\cdot} \mu(\mathrm{x}) d\mathrm{x} = \delta_{ij} \tag{2.29}$$

The stationary probability π_i to be in set i according to the full stationary density given as:

$$\pi_i = \int_{\mathrm{x} \in S_i}^{\cdot} \mu(\mathrm{x}) d\mathrm{x} \tag{2.30}$$

and the local stationary density $\mu_{i(\mathrm{x})}$ restricted to set i is

$$\mu_i(\mathrm{x}) = \begin{cases} \dfrac{\mu(x)}{\pi_i} & \mathrm{x} \in S_i \\ 0 & \mathrm{x} \notin S_i \end{cases} \tag{2.31}$$

where the local partition of state space does not require information about the full state space.

2.3. Transition matrix

Markov models require defining transition matrix elements $T_{ij}(\tau)$ that give the probability of finding the system in state j at time $t + \tau$ starting in state i at time t [69, 75].

$$T_{ij}(\tau) = \mathbb{P}\left[\mathrm{x}(t+\tau) \in S_j | \mathrm{x}(t) \in S_i\right] \tag{2.32}$$

$$= \frac{\mathbb{P}\left[\mathrm{x}(t+\tau) \in S_j \cap \mathrm{x}(t) \in S_i\right]}{\mathbb{P}[\mathrm{x}(t) \in S_i]}$$

$$= \frac{\int_{\mathrm{x}\in S_i} \mu_i(\mathrm{x}) p\left(\mathrm{x}, S_j; \tau\right) d\mathrm{x}}{\int_{\mathrm{x}\in S_i} \mu_i(\mathrm{x}) d\mathrm{x}}$$

We have to integrate over local equilibrium sets $\mu_i(\mathrm{x})$ as a weight to obtain the transition matrix, which facilitates the estimation of the transition probabilities. This approach does not require any information about the global equilibrium of the system and just gives the dynamic information over time period τ. The transition matrix is also related to the correlation function by

$$T_{ij}(\tau) = \frac{\mathbb{E}_t\left[\chi_i(\mathrm{x}_t)\chi_j(\mathrm{x}_{t+\tau})\right]}{\mathbb{E}_t[\chi_i(\mathrm{x}_t)]} = \frac{c_{ij}^{corr}}{\pi_i} \tag{2.33}$$

The probability density of the system of state *j* at time $t + \tau$ will be given by *the P should be a rho to be consistent*

$$p_j(t+\tau) = \sum_{i=1}^{n} p_i(t) T_{ij}(\tau) \tag{2.34}$$

or in matrix notation

$$\boldsymbol{P}^T(t+\tau) = \boldsymbol{P}^T(t)\boldsymbol{T}(\tau) \tag{2.35}$$

The stationary probability of the transition matrix **T** for any time τ corresponds to the highest eigenvalue of norm 1. it is given by

$$\pi^T = \pi^T \boldsymbol{T}(\tau) \tag{2.36}$$

Markov models use $\boldsymbol{T}(\tau)$ to predicate the probability distribution of long-time dynamics of discretized space for later times $t + k\tau$ as

$$\boldsymbol{P}^T(t + k\tau) \approx \boldsymbol{P}^T(t)\boldsymbol{T}^k(\tau) \tag{2.37}$$

2.4. Estimation of Markov state models

2.4.1. Count matrix

Based on that the simulation data is saved at a fixed time interval, a count method is used to sample the trajectory at lag-time τ and estimate the transition matrix. If the trajectory is saved the data of the configuration every fixed time interval $\Delta\mathrm{t}$ for N times [69]:

$$\mathrm{X} = [\mathrm{x}_1 = \mathrm{x}(\mathrm{t} = 0), \mathrm{x}_2 = \mathrm{x}(\mathrm{t} = \Delta\mathrm{t}), \dots, \mathrm{x}_\mathrm{N} = \mathrm{x}(\mathrm{t} = (\mathrm{N} - 1)\Delta\mathrm{t})] \tag{2.38}$$

where every structure is assigned to one discrete state. The discrete count matrix $\boldsymbol{C}^{obs}(\tau)$ can be defined at lag-time τ that must be an integer multiple of the time interval Δt, where $c_{ij}^{obs}(\tau)$ is the total number overall times t of times the trajectory was observed in state i at time t and in state j at time $t + \tau$:

$$c_{ij}^{obs}(\tau) = c_{ij}^{obs}(l\Delta t) = \sum_{k=1}^{N-l} \chi_i(\mathrm{x}_k)\, \chi_j(\mathrm{x}_{k+l}) \tag{2.39}$$

The count matrix considers as the estimator of the correlation function in eq.(2.32) by

$$\hat{c}_{ij}^{corr}(\tau) = \frac{c_{ij}^{obs}(\tau)}{N - l} \tag{2.40}$$

The total number of times the trajectory was in state i can be defined as a row sum of the count matrix as:

$$c_i^{obs} = c_i^{obs}(\tau) \coloneqq \sum_{k=1}^{n} c_{ik}^{obs} \tag{2.41}$$

2.4. 2. Likelihood and Bayesian estimators

The transition matrix in the limit of infinitely long trajectory can be expressed in term of the count matrix as the fraction of times the transition from state i to state j led out of state i [69].

$$\hat{T}_{ij}(\tau) = \frac{c_{ij}^{obs}(\tau)}{c_i^{obs}(\tau)} \tag{2.42}$$

However, the transition matrix $\boldsymbol{T}(\tau)$ for limited length trajectory is not uniquely determined. All the observed data must be statistically independent or uncorrelated counts when jump process is Markovian at lag-time τ. Assuming that count matrix elements are statistically independent. The likelihood probability that a particular $\boldsymbol{T}(\tau)$ would generate a sequence $s_1, \dots, s_n$ the observed trajectory is given by the product of the individual jump probabilities asthis sentence is too long and unclear:

$$p(C^{obs}|T) = \prod_{i.j=1}^{n} T_{ij}^{c_{ij}^{obs}} \tag{2.43}$$

In a Bayesian approach, the posterior probability of the transition matrix being $\boldsymbol{T}(\tau)$ is

$$p(T|C^{obs}) \propto p(T)p(C^{obs}|T) = p(T)\prod_{i.j=1}^{n} T_{ij}^{c_{ij}^{obs}} \tag{2.44}$$

where $p(\boldsymbol{T})$ is the prior probability of transition matrices before observing any data. The two approaches can be used for MD simulations. But, Bayesian estimation is usually used for estimating reversible Markov models [76] where $\boldsymbol{T}(\tau)$ for MD in equilibrium should obey detailed balanced

$$\pi_i T_{ij} = \pi_j T_{ji} \tag{2.45}$$

While for nonreversible estimation transition matrix with $\sum_j T_{ij} = 1$ and $T_{ij} \geq 0$ maximum likelihood estimator is analytically available.

2.5. Kinetic model validation: Chapman-Kolmogorov test

Obtaining a good kinetic model that describes the true dynamics of the system depends on the choice of appropriate lag-time τ, and discretization that minimizes the discretization errors of the Markov state model. Chapman-Kolmogorov(CK) test [69] can test if the transition matrix $\widehat{\boldsymbol{T}}(\tau)$ is approximately Markovian. The CK test with statistical errors for Markovian matrix is

$$\left[\widehat{\boldsymbol{T}}(\tau)\right]^k \approx \widehat{\boldsymbol{T}}(k\tau) \tag{2.46}$$

The test compares the MSM transition probability estimated at lag-time $k\tau$, where k is an integer larger than one with the estimated MSM transition probability matrix to the power k^{th}. The test can be done by comparing the probability for few observables when starting from stationary distributions restricted to set of states A at later time $k\tau$. The stationary probability for a set of states A is

$$w_i^A = \begin{cases} \dfrac{\pi_i}{\sum_{j\in A}\pi_j} & i \in A \\ 0 & i \notin A \end{cases} \tag{2.47}$$

The probability to be at set A after $k\tau$ with starting distribution $\boldsymbol{w}^A$ according to Markov model is given by

$$p_{MSM}(A, A; k\tau) = \sum_{i\in A} [(\boldsymbol{w}^A)^T \boldsymbol{T}^k(\tau)]_i \tag{2.48}$$

where $p(A, A; k\tau)$ is the probability to be at set A at later time $k\tau$ started from the set A at time t. Validation of the Markov model requires us to compare the MSM probability with probability obtained from the trajectory data within statistical errors:

$$p_{MD}(A, A; k\tau) = p_{MSM}(A, A; k\tau) \tag{2.49}$$

where $p_{MD}(A, A; k\tau)$ is the estimated transition probability to be at set A after $k\tau$ with starting distribution $\boldsymbol{w}^A$ according to trajectory data is given by

$$p_{MD}(A, A; k\tau) = \sum_{i \in A} w_i^A p_{MD}(i, A; k\tau) \tag{2.50}$$

The probability $p_{MD}(i, A; k\tau)$ is given by

$$p_{MD}(i, A; k\tau) = \frac{\sum_{j \in A} c_{ij}^{obs}(k\tau)}{\sum_{j=1}^{n} c_{ij}^{obs}(k\tau)} \tag{2.51}$$

and the statistical errors of $p_{MD}(A, A; k\tau)$ is given by

$$\epsilon_{MD}(A, A; k\tau) = \sqrt{k \frac{p_{MD}(A, A; k\tau) - [p_{MD}(A, A; k\tau)]^2}{\sum_{i \in A} \sum_{j=1}^{n} c_{ij}^{obs}(k\tau)}} \tag{2.52}$$

2.6. Transition path theory

Transition path theory (TPT) [77, 78] is used to estimate the probability of the transition pathway (fluxes). We define a start state as A, and a final state as B and other intermediate states as I. The forward committer q_i^+ is defined as the probability, when being at state I, that the system will reach the state B before A, can be estimated by solving the equations:

$$q_i^+ - \sum_{j \in I} T_{ij} q_i^+ = \sum_{j \in B} T_{ij} \tag{2.53}$$

Also, the probability of backward transition is computed as:

$$q_i^- = 1 - q_i^+ \tag{2.54}$$

The average number of transitions of the different pathways from i to j as part of the transition from A to B is defined as

$$f_{ij} = \pi_i q_i^- T_{ij} q_i^+ \tag{2.55}$$

The net fluxes f_{ij}^+ can be computed using eq.(2.56):

$$f_{ij}^+ = max\left(0, f_{ij} - f_{ji}\right) \tag{2.56}$$

2.7. Summary

We briefly introduced MSM building from MD data (Figure 2.1). The first step is to define the input coordinates (feature) which can be used to characterize the MD trajectory. For example, one can choose cartesian coordinates, dihedral angles or contact of distant pairs. The next step is the dimension reduction using TICA to transform the feature into a set of slowest coordinates that can identify the slowest process in the MD data. The TICA output is then clustered into a set of microstates using a clustering algorithm such as K-means clustering which assigned each frame of the trajectory to one microstate [79]. At this stage, we can estimate MSM from the discretized trajectories of the microstate by approximating the transition probability matrix at specified lag-time. Next, one can validate that the system as memory-less after a specified lag-time by calculating implied timescales that are independent of lag-time. To get a simple picture of the system dynamic with few states that contain important information such as structural and

kinetic information. To do this the Perron-cluster cluster analysis (PCCA++) method [80] can use to assign the microstates into macrostates (metastable states). In addition, we can apply Chapman-Kolmogorov test to check that our system is Markovian. If the system is valid we can apply TPT to compute the transitions probability among metastable states.

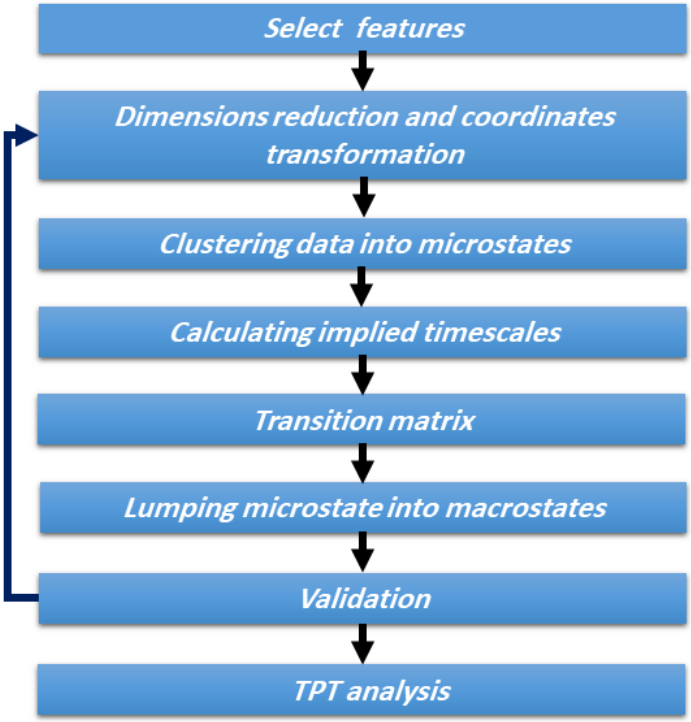

Figure 2.1: Flowchart for building MSMs.

CHAPTER 3: MATERIAL AND METHODS

We simulated the GTP-bound catalytic domain of G12D K-Ras and its phosphorylated mutant (pTyr32) to investigate the effects of Tyr32 phosphorylation on the conformational states of K-Ras. We also identified metastable conformational states and the kinetic network of transitions between them using Markov state model (MSM).

3.1. Initial structures and system preparation

The starting structure was downloaded from the RCSB protein data bank (PDB ID: 4DSO). Since no high-resolution crystal structure was available for the phosphorylated Tyr32, we used 4DSO structure to generate the initial configuration by mutating Tye32 to pTyr32.

In both cases, we replaced GSP with GTP molecule and removed all molecules in the PDB file except for water molecules and Mg^{+2} ions. The C-terminus and anionic residues were deprotonated while the N-terminus and cationic residues were protonated assuming neutral pH. The resulting structure was placed in a cubic box containing TIP3P water molecules, and Na^{+} and Cl^{-} ions were added to neutralize the system and achieve an ionic strength of 150 mM. A minimum of 10 Å buffer between the edges of the box and protein atoms was used to ensure that the protein does not interact with its periodic images (Figure 3.1).

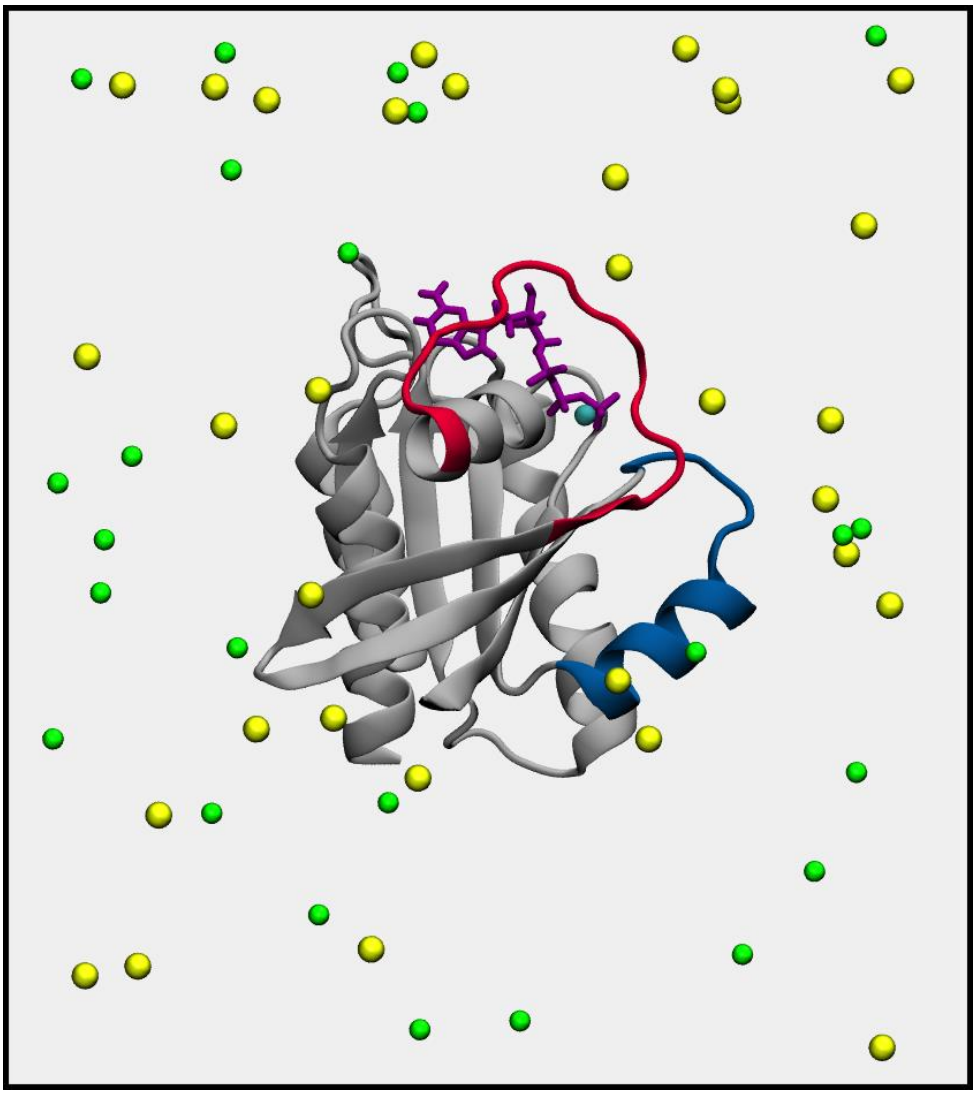

Figure 3.1: A snapshot from one of the simulations. The catalytic domain is in cartoon colored in dark grey and switches SI (residues 25-40) and SII (residues 60-75) are in red. Magnesium, sodium and chloride ions as shown in blue, yellow and green spheres, respectively. The bound GTP is in purple sticks.

3.2. Molecular dynamics simulation setup

Short-range van der Waals interactions were smoothly switched off between 10 Å and 12 Å, with a 14 Å cutoff used for non-bonded pair list updates. Long-range electrostatic interactions were computed using the Particle Mesh Ewald (PME) method [81] with a grid density of about one grid point per Å. The solvated systems were energy minimized (5000 steps of conjugate

gradient), gradually heated keeping the Cα and GTP heavy atoms restrained by a harmonic restraint of force constant $k = 4\ kcal\ mol^{-1}Å^{-2}$, and equilibrated with k progressively reduced to zero by decrements of 1 $kcal\ mol^{-1}Å^{-2}$ every 100 ps. An integration time step of 2 fs was used with the SHAKE [82] algorithm applied to covalent bonds involving hydrogen atoms. The isothermal-isobaric (NPT) ensemble and periodic boundary conditions were used. The temperature was maintained at the physiologic value of 310K using Langevin dynamics with a damping coefficient of 10 ps^{-1}. The Nose-Hoover Langevin piston method was used with a piston period of 200 fs and decay time interval of 100 fs to maintain constant pressure at 1.0 atm. Each system was simulated for 500 ns with NAMD2.11 [83] using the CHARMM27 empirical force field and CMAP dihedral angle correction [51].

3.3. Building Markov state models and their validation

Markov state model(MSMs) [66, 69] were built from MD simulations using PyEMMA software package version 2 [60]. Multiple definitions of microstates were tested. We found that the distances among Cα coordinates of residues 12, 32, 34, 36, 48, 56, 59, 63, 66 ,67 , 74, 105, 108, 122, 126, 138, 148 and 153 are sufficient to resolve protein conformational changes observed in the two trajectories. This is because the aforementioned residues have a relatively higher dynamical nature as quantified from root mean square displacement calculations. Time-lagged independent component (TICA) [72] with 1 ns as a lag-time was also used to find the slow linear subspace of the input coordinates (Figure 3.2) and subsequent dimension reduction by projecting on the two slowest TICA components.

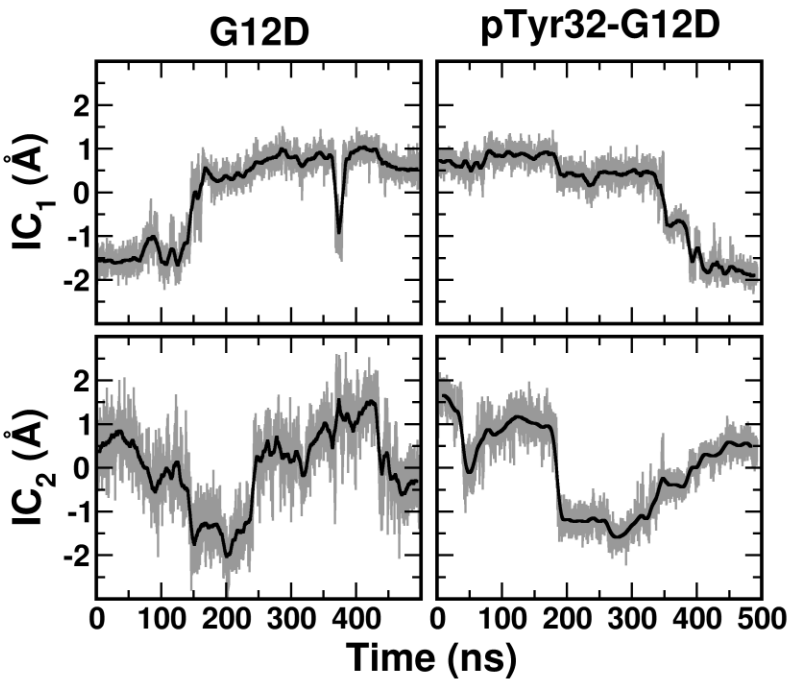

Figure 3.2: Projection of the trajectories onto the two largest independent components. Projection of the trajectory onto the two largest independent components subset estimated by TICA. Gray lines represent data sampled every 100 ps while thick black lines represent 10 ns running averages.

K-means clustering [79] was utilized to get a set of 100 microstates represented by cluster centers (Figure 3.3).

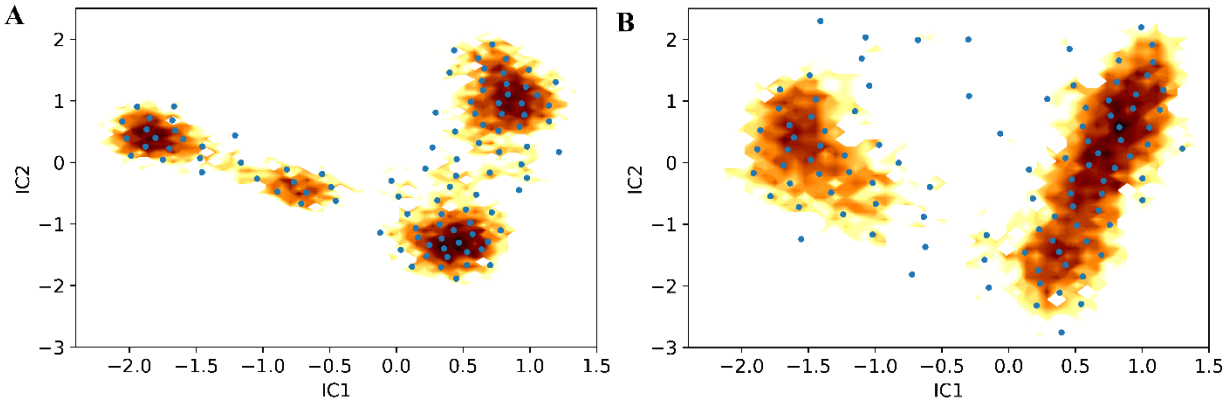

Figure 3.3: Clustering the trajectory into microstate. The trajectory is assigned to the 100 cluster centers using k-means clustering. (A) for pTyr32-G12D and (B) G12D K-Ras.

The resulting discretized trajectories were used to construct the Bayesian MSM using 2 ns lag-time for which the system is considered Markovian (i.e. timescales become and independent of the lag time itself, see Figure 3.4).

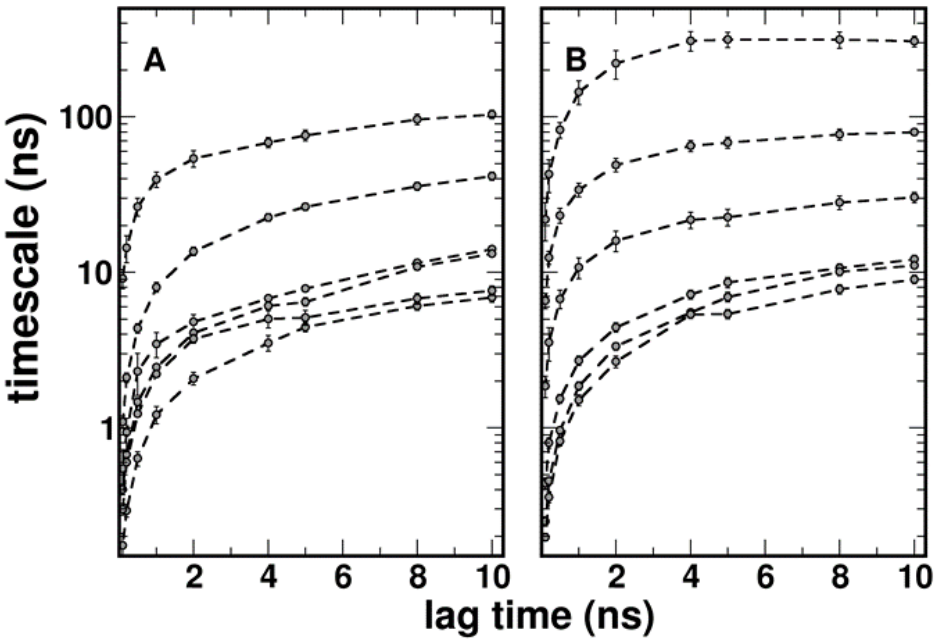

Figure 3.4: Validation of Markov state model. Implied relaxation timescales for the first six eigenvalues calculated from the transition matrix at different lag times. All relaxation timescales become approximately constant beyond the used lag-time 2 ns to construct MSMs for (A) G12D and (B) pTyr32-G12D.

Spectral analysis of timescale separations shows that the largest timescale separation is between the first and the second relaxation timescales for G12D, and third and fourth relaxation timescales for pTyr32-G12D, respectively (Figure 3.5).

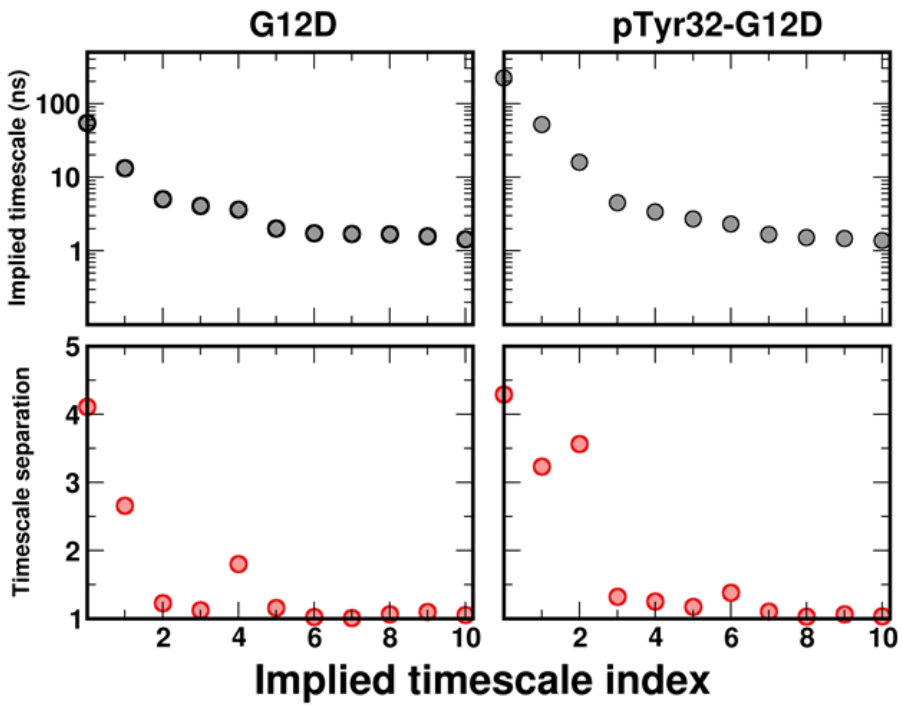

Figure 3.5: The Spectral analysis of timescale separation. The Spectral analysis revealed the largest timescale separation is between the first and the second relaxation timescales for G12D, and third and fourth relaxation timescales for pTyr32-G12D, respectively.

This suggests that retaining four relaxation times or five metastable states is sufficient to coarse-graining the dynamics of both systems. Thus, the microstates were then grouped into five metastable states using the Perron-cluster cluster analysis (PCCA++) method [80]. The free energy of each metastable state was computed by comparing the probabilities of its constituent microstates. The Chapman-Kolmogorov test [69] was employed to further validate the reliability of both five metastable state Markov state models. As shown in Figure 3.6, the probability predicted from MSMs for a given metastable state has small deviations from the probability counts from MD simulations. Finally, Transition path theory [78, 84] was utilized to compute the transition path fluxes among metastable states using the forward committer probabilities because MSM here is reversible.

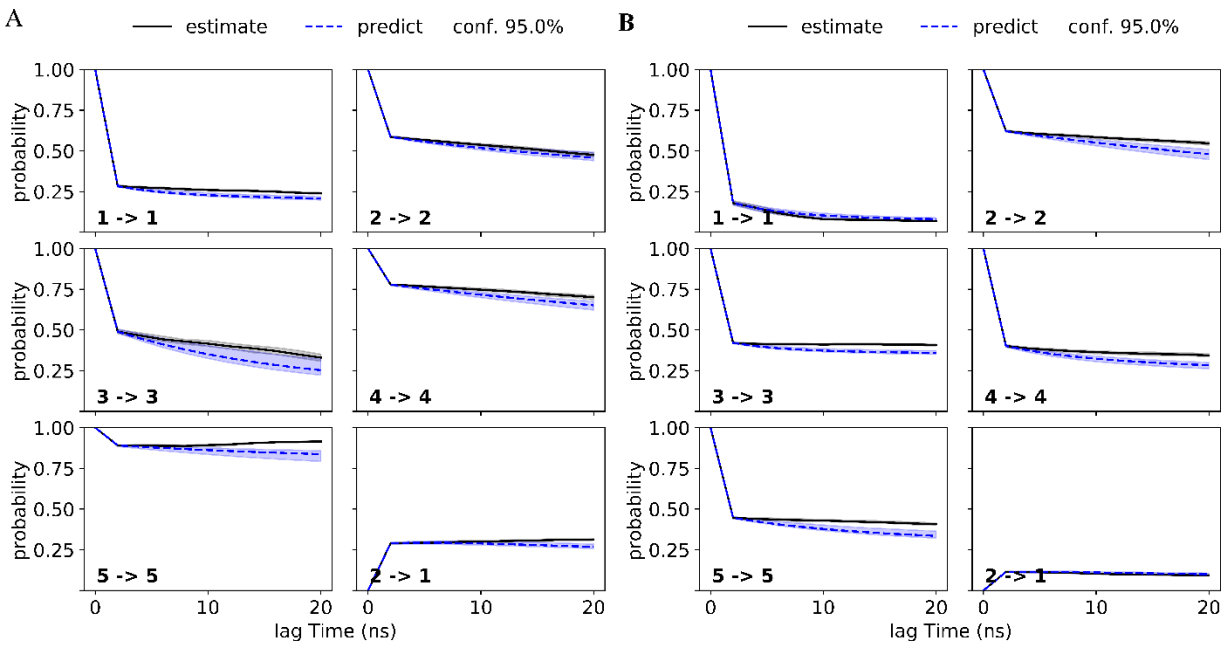

Figure 3.6: The Chapman-Kolmogorov test. The Chapman-Kolmogorov test for MSMs for (A) pTyr32-G12D and (B) G12D with the five states. The data for MSM (black line) and the MDs trajectory (blue dotted line, with estimated error).

CHAPTER 4: RESULTS AND DISCUSSION

4.1. Effects of Tyr32 phosphorylation on structure and dynamics of G12D K-Ras.

To investigate the effects of Tyr32 phosphorylation, we calculated the root mean square fluctuations (RMSF). RMSF of a dynamical particle is a measure of the deviation between its position and some reference position:

$$RMSF = \sqrt{\frac{1}{T}\sum_{t_j=1}^{N}\left|\vec{r}(t_j) - \vec{r}_{ref}\right|^2} \tag{4.1}$$

where T is the simulation time and $\vec{r}(t)$ is the position of the particle at time t and $\vec{r}_{ref}$ is the reference position of the particle. The reference position is usually taken to be the average position of the particle.

RMSFs are typically linked to the stability and mobility of protein structures. Therefore, we carried out RMSFs calculations on G12D K-Ras protein and it phosphorylated variant trajectories as shown in Figure 4.1. The RMSF of each residue reveals the mobility of various parts of the protein and the effect of phosphorylation on the mobility of the G12D K-Ras. The effects of Tyr32 phosphorylation appear clearly in the flexibility of SI and SII. SI shows more flexibility in pTyr32-G12D than G12D K-Ras. This is because the Tyr32 phosphorylated among SI, but SII shows reduction in the flexibility in pTyr32-G12D than G12D K-Ras. Furthermore, the flexibility of helix $\alpha 3$ region (residues 87−104) is increased in the case of Tyr32 phosphorylation.

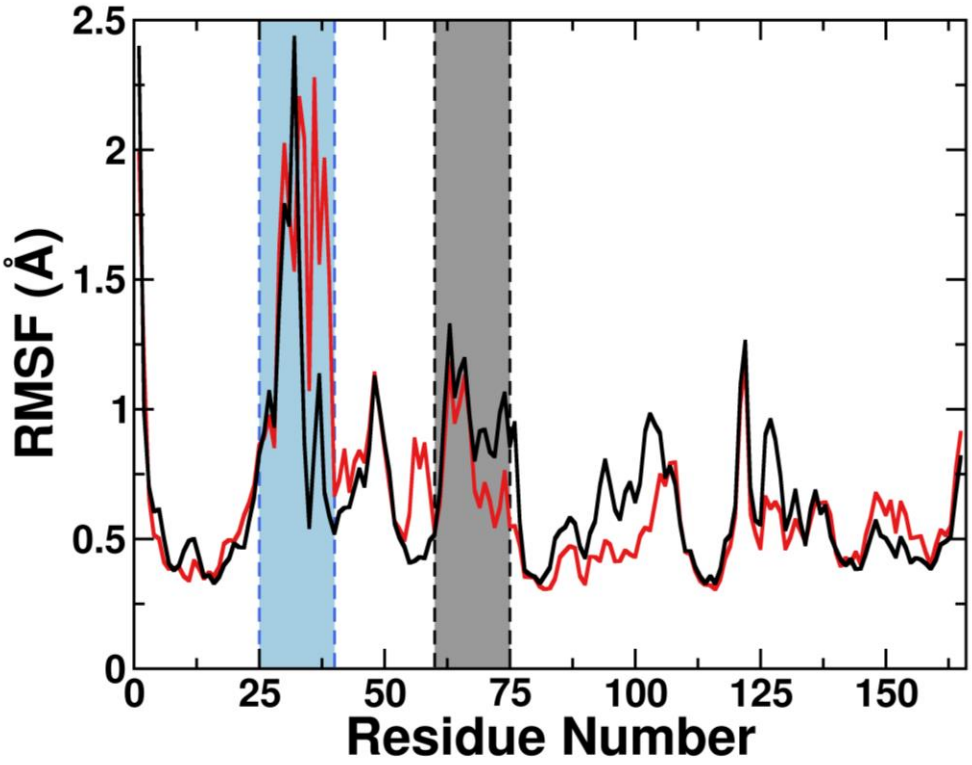

Figure 4.1: C_α root mean square fluctuation (RMSF). RMSF calculated after alignment excluding SI and SII regions for both G12D K-Ras (black) and its phosphorylated variant (red). SI and SII are highlighted by cyan and purple shadows, respectively.

RMSD is a measure of how similar two conformations of a protein are. It is calculated using

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\vec{r}_i(t) - \vec{r}_i^{ref}\right|^2} \tag{4.2}$$

where N is the number of atoms whose positions are being compared. $r_i(t)$ is the position of the i-th atom at time t, while r_i^{ref} is the corresponding reference position of the same atom.

To further study the conformational dynamics of both mutants, we also calculated the root mean square deviation (RMSD) of their backbone atoms. The results of these RMSD

calculations are shown in Figure 4.2. The structures are very stable for both mutants. The RMSDs of the backbone from the initial X-ray structure is between 0.8 and 1.5 Å for both G12D K-Ras and its phosphorylated variant.

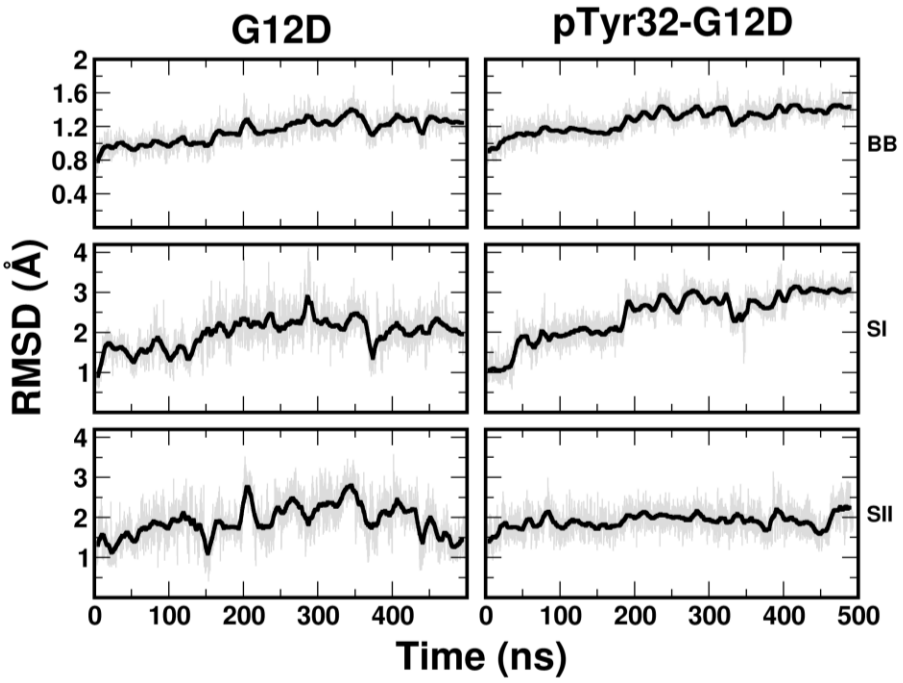

Figure 4.2: Time evolution of root-mean square deviation (RMSD). RMSD values of backbone (BB), Switch 1 (SI) and Switch 2 (SII) structures sampled from the initial X-Ray structure for both G12D K-Ras and its phosphorylated variant. RMSDs were calculated after structural alignment excluding the flexible switch regions. Gray lines represent data sampled every 100 ps while thick black lines represent 10 ns running averages.

Nonetheless, the two switches show different dynamical behavior. In particular, SI has the average RMSD of G12D K-Ras between 0.9 and 2.9 Å, while its phosphorylated variant has it between 1.1 and 3.2 Å. Similarly, the RMSD of SII is 1.1-2.8 Å and 1.4- 2.2 Å for G12D and pTyr-G12D, respectively. Overall, Tyr32 phosphorylation of G12D K-Ras do not induce severe structural changes but it increased the flexibility of SI and the decreased flexibility of SII and helix α3.

The changing of the flexibility of helix α3 and SII encouraged us to investigate the relation between these two neighboring regions. For G12D K-Ras, helix α3 is shifted toward SII due to sidechain interactions between helix α3 and SII. Namely, presence of many transient hydrogen bonds with the polar and charged residues between (ARG68:NH1-TYR96:OH), (ARG73:N-VAL103:HG1) and (GLY75:O-LYS104:HZ2) was observed. This explains the differential flexibility of SII and α3 regions between the two proteins and communications between their two lobes (lobe 1: residues 1–86 and lobe 2: residues 87–171). It is also consistent with having functional significance as suggested previously when correlated motions observed between helix α3 and helix α2 regions in GTP-bound Ras simulations [85].

The GTP binding pocket of Ras is strongly positively charged which stabilized by the negatively charged phosphates of GTP. The important role of Tyr32 on SI in intrinsic hydrolysis is to coordinate a water molecule adjacent to γ-phosphates of GTP which bridged to a hydroxyl group of Tyr32 [86, 87]. Insertion of polar and negatively charged phosphate group at Tyr32 may alert the charge distribution or cause some reordering of the solvent. The conformational differences of SI are directly altered by Tyr32 phosphorylation that induces the electrostatic repulsion between the negatively charged carboxyl groups between proximal residues Asp38 and Asp57 (Figure 4.3) consistent with early results from biophysical experiments [31].

Tyr32 is found in the middle of Ras effector binding site. Its motion is coupled with the movement of four adjacent negatively charged residues in the SI region (Glu31, Asp33, Glu37, and Asp38) which play a critical role in Raf binding. This alteration of SI confirmation illustrates the reduction of the affinity of Ras to its effector proteins [88, 89]. Moreover, Tyr32 phosphorylation may be affecting Ras-GAP interactions. A previous study revealed that mutation of negatively charged residues in SI region affect the electrostatic interaction between the Ras

and GAP and that lead to reduce their affinity to each other [13]. SII displays complicated dynamic interactions with its effectors. Hence, the decreased local flexibility observed in the SII of pTyr-G12D could significantly affect the role of SII in the binding and activation of its effectors through disordering contacts with the effectors in regions outside the Ras effectors binding domain [21, 90].

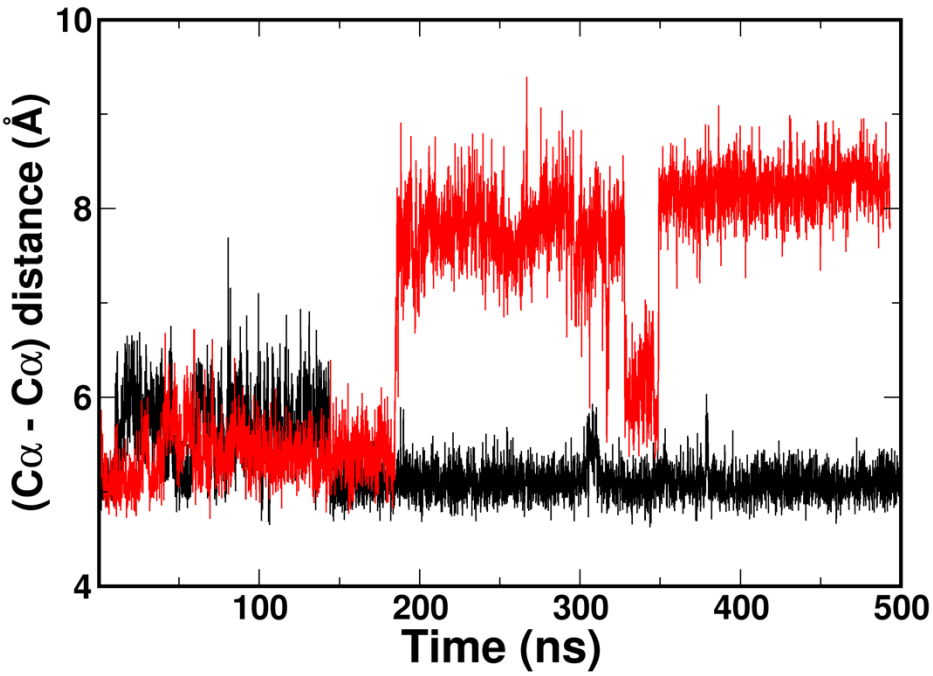

Figure 4.3: The time evolution of distance between the C_α atoms of the residues Asp38 an Asp57 of both mutants G12D K-Ras (black) and its phosphorylated variant (red).

4.2. Analysis of side chain torsions

The side chain of Tyr32 is located across the nucleotide binding site in the GTP bound form. It has been also shown that the specific orientation of Tyr32 affects the GTP hydrolysis. The phosphorylation of Tyr32 affects the side chain reorientation of Tyr32 as it shows different behavior for both mutants. A large change in the orientation of the side-chain of Tyr32 was

observed as characterized by the dihedral angle χ_1 (N-Cα-Cβ-Cγ). Each mutation exhibits distinguishable states (Figure 4.4). These variations in Tyr32 sidechain orientation are likely to contribute to changes in SI conformations and dynamics. Therefore, that may have a functional significance since Tyr32 undergoes a reorientation through nucleotide exchange [91, 92].

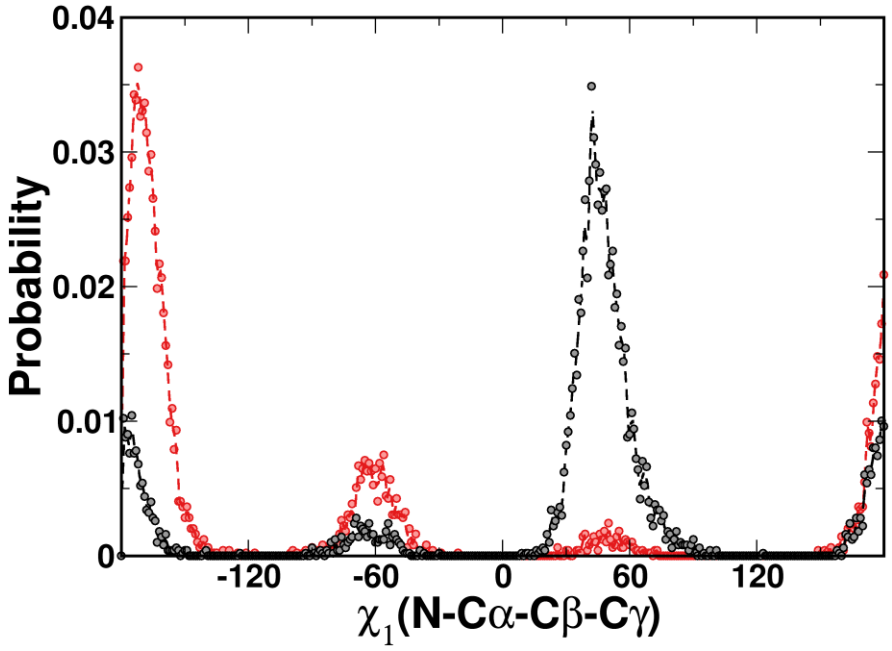

Figure 4.4: The probability of the dihedral angle $\boldsymbol{\chi_1}$(N-Cα-Cβ-Cγ) of Tyr32 for G12D K-Ras (black) and its phosphorylated variant (red).

4.3. Principal component analysis

To gain better insight into the phosphorylation effects on the global dynamics and find the most significant large scale motions, we used principal component analysis [71]. We defined the two largest principal components of MD trajectories obtained from analyzing them for the selected atoms. Figure 4.5 highlights the differences between the two mutants as both seem to occupy distinct and mutual conformational states. It also shows that the profiles of both mutants large-scale motions are distinct. The phosphorylated variant sampled a larger conformational space

than the unphosphorylated one, indicating that the phosphorylated variant is more dynamic. This can be related to the observed fluctuations along PC_1 due to large fluctuations of SI for the phosphorylated variant.

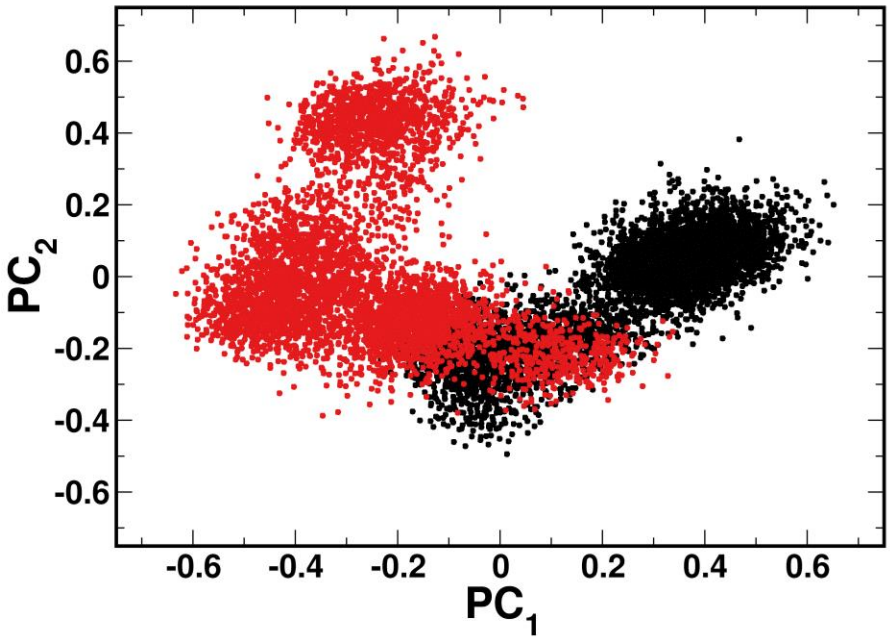

Figure 4.5: Global conformational dynamics of mutants G12D K-Ras and its phosphorylated variant. Projection of simulated trajectories into the first and the second principal components of mutants G12D K-Ras (black) and its phosphorylated variant (red).

4.4. GTP binding site configuration

Previous [93] P-NMR spectroscopic studies of wild-type H-Ras catalytic domain bound to guanosine 5`-(β, γ -imido) triphosphate (GppNHp) found a slow chemical shift changes for the values of the nucleotide phosphorus atoms of the α-, β-, and γ-phosphates showed that H-Ras−GppNHp exists in two distinct conformational states, inactive state 1 and active state 2 [94, 95]. The two states determined the stability of SI and SII by the interaction of the two residues T35 and G60 in switch regions with γ-phosphate in GTP. The two-states are different among Ras mutations, they are associated with different biochemical interactions. Inactive state 1 contains

three substates described by the loss of interaction of T35 or G60 with γ-phosphate this state favors nucleotide exchange while inhibiting the interaction with effector protein. However, active state 2 contain described by the interaction of T35 or G60 with γ-phosphate this state is associated with effector binding and GTP hydrolysis [96, 97]. A previous study showed that the G12D mutation shifts conformations to the active state 2 which have a larger connection with nucleotide binding site [32]. Understanding the effects of TYR32 phosphorylation at protein–GTP interactions is important because these interactions can affect rates of nucleotide exchange and GTPase activity. The State 2 which delineate by the interactions of T35 and G60 with GTP it is very similar for both mutations. The average G60:N-GTP:Oγ2 and T35:Oη-GTP:Oγ3 distances do not have any significant differences (Table 4.1). In our simulation timescale, we did not observe conformational transitions between the active and inactive states. Both mutations mostly remained close to the active state 2 configurations.

	G60:N-GTP:Oγ2 (Å)	T35:Oη-GTP:Oγ3 (Å)
G12D	3.1 ± 0.3	2.8± 0.1
pTyr32-G12D	3.1 ± 0.3	2.9 ± 0.1

Table 4.1: Average Distance of T35 and G60 from GTP.

4.5. Sodium ion interaction

We noted by visual inspection of the trajectory of pTyr32-G12D mutant that the active site of the mutant is temporarily occupied by sodium ion in the last 120 ns of simulation (Figure 4.6). The Na^+ in the active site is bound by the neighboring oxygen atoms of GTP:O2' and the polar amino acids and negatively charged residues Asp30, Glu31 and Asp33 in SI region. We suggest that

mediating of a metal ion at this site stabilize the motion of pTyr32 sidechain through the displacement of pTyr32 sidechain to point to Mg^{+2} ion and γ-phosphate of GTP.

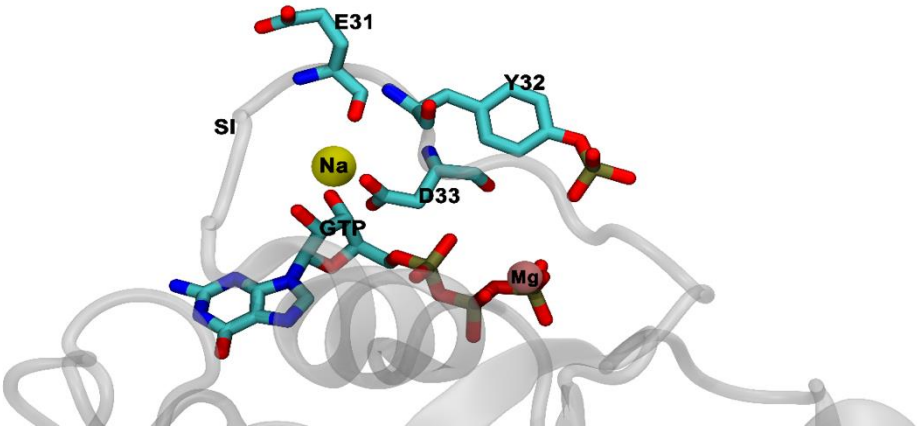

Figure 4.6: Long-residence sodium-ion binding sites. A snap shot of pTyr-G12Dshowing a sodium ion interacting with GTP and SI.

4.6. Markov state model analysis

To get a better insight into the conformational changes for both mutations, we analyzed the simulation trajectories using Markov state models (MSMs) to explore the long-lived conformational dynamics for each system. The MSM identified five metastable states by using Perron-cluster cluster analysis (PCCA++) method, where each of these clusters represent highly identical conformational states (figure 4.7). The free energies for metastable states can be computed from their stationary probabilities by the relation

$$G_{s_i} = -k_B T ln \sum_{j \in S_i} \pi_j \tag{4.3}$$

where π_j is MSM stationary weight of the j^{th} microstate (Table 4.2).

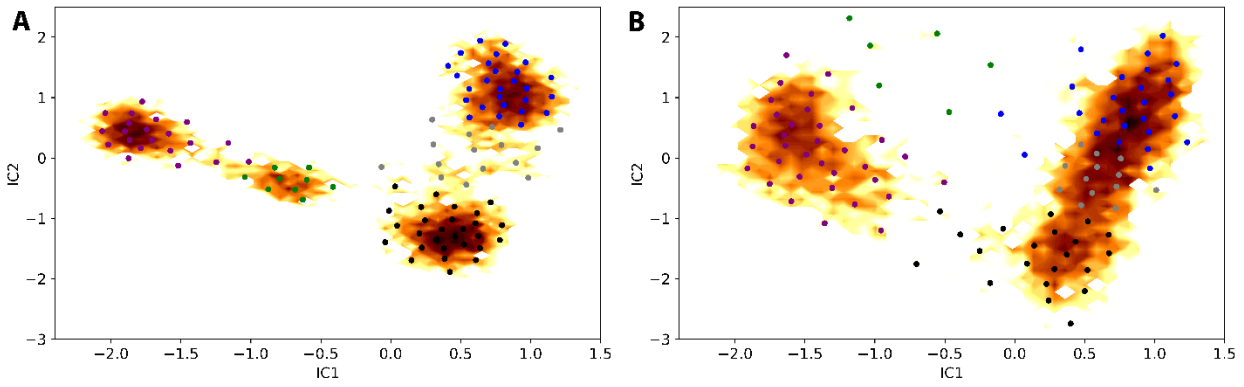

Figure 4.7: The five metastable states grouped from microstates. The trajectory cluster into microstate by assigning it to the 100 cluster centers using k-means clustering. The microstates grouped by (PCCA++) method into five metastable states. (A) pTyr32-G12D. The color code of metastable states 1 (blue), 2 (gray), 3 (black), 4 (green), 5 (purple). (B) G12D K-Ras. The color code of metastable states I (blue), II (gray), III (black), IV (green), V (purple).

G12D K-Ras

Metastable state S_i	**π_{Si}**	**free energy** *(k_BT)*
I	21%	1.556
II	39%	0.939
III	20%	1.610
IV	1%	4.282
V	18%	1.692

pTyr32-G12D

Metastable state S_i	**π_{Si}**	**free energy** *(k_BT)*
1	10%	2.304
2	3%	3.635
3	14%	1.969
4	11%	2.230
5	63%	0.467

Table 4.2: The stationary probability and the free energy of metastable states of G12D K-Ras and its phosphorylated variant.

The conformational changes between the different conformations are a slow process. The slowest MSM implied timescale of pTyr32-G12D is about 220 ns, but it is faster for G12D K-Ras with 55 ns (see Figure 3.5). This suggests that simulation time is not enough to detect all the transitions among metastable states for pTyr32-G12D. Therefore, observing enough transitions will require simulations on the order of many μs long. In crystal structures, it is difficult to observe these dynamic metastable states because the structures of the switch regions are quickly disordered.

Both mutations display different metastable state population distributions. pTyr32-G12D has one highly populated metastable state with probability 63% and low populated states with probability 3-14%. G12D K-Ras most populated metastable state has a probability of 39% and low populated states with probability 21-19% and its lowest populated state with probability 1%. The density of population indicates that several highly populated intermediate conformations were sampled. We also used the transition path theory to obtain the transition pathways and gain insights into how they are affected by mutations. Figure 4.8 shows metastable states and the transitions network among them. Table 4.3 shows the maximum transition pathway. The MSM metastable states confirmed the effect of Tyr32 phosphorylation on the switch regions conformations and dynamics

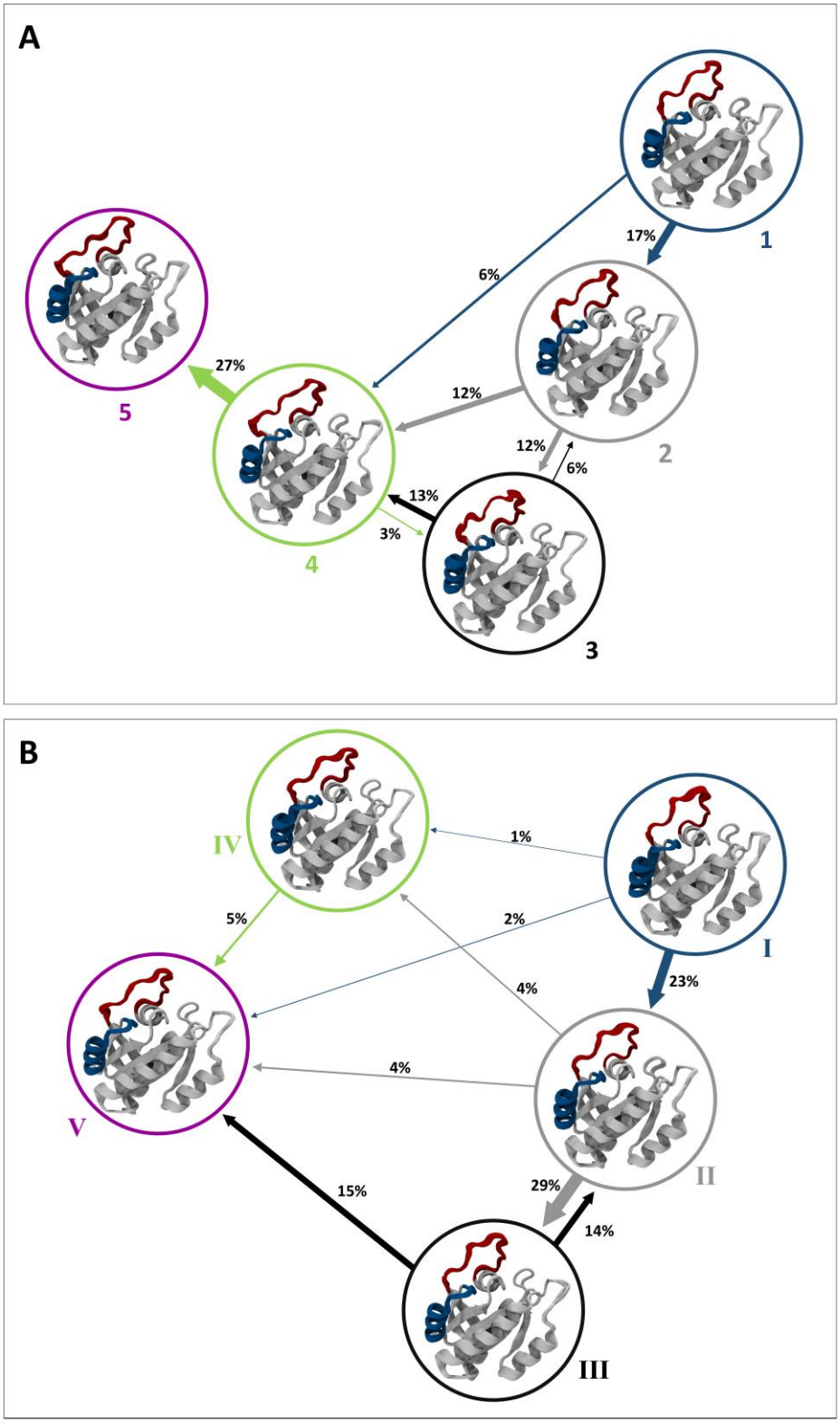

Figure 4.8: A network diagram of the five metastable states identified by the Markov state model. The metastable states are represented by circles, the arrows indicate the transition probabilities between the states. The structures describe the metastable states found in the MSM analysis, each circle illustrating ten representative protein conformations (generated using MSM), which identify also the SI (red) and SII (blue) regions. (A) pTyr32-G12D. (B) G12D K-Ras. The circle colors are the same as in figure 4.7.

G12D K-Ras

Path	Percentage
I → II → III → V	55%
I → II → V	16%
I → II → IV → V	14%
I → V	8%

pTyr32-G12D

Path	Percentage
1 → 2 → 4 → 5	42%
1 → 4 → 5	24%
1 → 2 → 3 → 4 → 5	23%
1 → 3 → 4 → 5	12%

Table 4.3: The maximum four fluxes path of G12D K-Ras and its phosphorylated variant.

To get further insight of mutants and metastable states differences we construct the RMSD matrix (Figure 4.9).

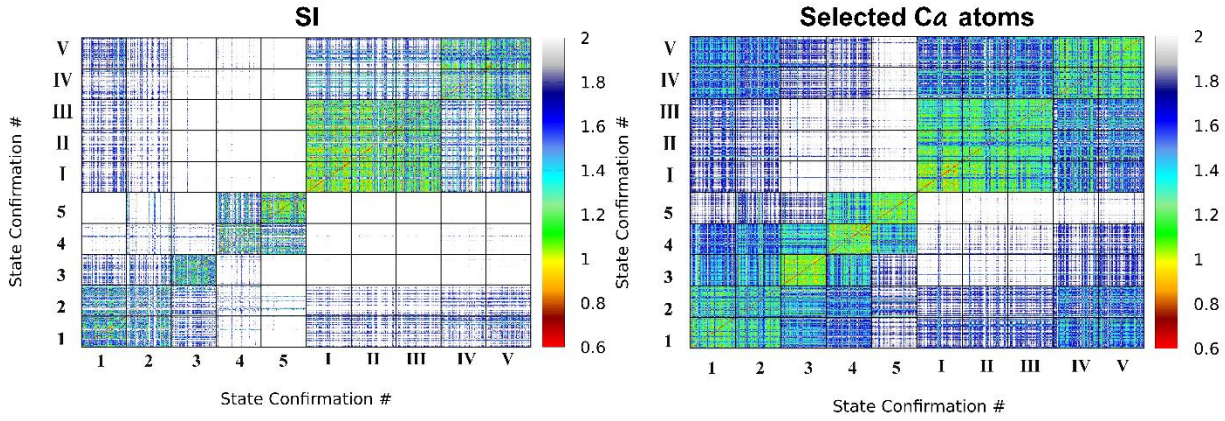

Figure 4.9: RMSD matrices of SI and selected Cα atoms computed from MSMs metastable states trajectories. The metastable states of pTyr32-G12D indicate by numbers (1-5) and the metastable states of G12D K-Ras indicate by Roman numbers (I-V).

pTyr32-G12D switch regions show a different behavior. Unlike SII, SI shows remarkable dynamics differences across all metastable states. State 1 and state 2 have very similar conformation and Tyr32 side chain orientation does not show significant difference. Comparing state 4 and state 5 shows that conformations have similar Tyr32 side chain orientation. This suggests that state 2 and 4 represent intermediate states. However, states 1, 3 and 5 show critical

conformational changes in SI region due to the change in sidechain orientation of Tyr32. Notably, the last 120 ns of pTyr32-G12D trajectory mentioned in the previous subsection is represented by state 5. The sidechain orientation of Tyr32 seems to have a critical role in determining the metastable state of pTyr32-G12D. In G12D K-Ras the metastable states I, II and III have similar conformations with relatively stable Tyr32 orientation. Nonetheless, it differs from the metastable states V and IV especially in SI region where states V and VI have unstable Tyr32 sidechain orientation.

The differences in metastable states of both mutants are more significant in SI region. These differences in mutant dynamics and conformation revealed by different metastable states may affect protein activity and can be important for modulating a specific protein binding and pathway activation [21].

4.7. Conclusions

The aim of this work is to understand the effects of specific mutations on the structure and dynamics of a protein. These effects might lead to differences in protein functions and hence give rise to different signal outputs. In particular, we identified the structural and dynamical differences between G12D K-Ras and its phosphorylated variant pTyr32-G12d. This was achieved by comparing conformational and dynamics differences between two 500ns unbiased MD simulation trajectories of both mutants. We also built Markov state models for both trajectories. In addition, we analyzed conformational fluctuations and GTP binding site configuration using other global measures such as RMSF, RMSD, PCA and torsion angle for Tyr32 to identify any other subtle changes in protein dynamics. Differential dynamics are particularly common in the vicinity of the GTP, but there are also variations at a number of loops distal from the active site. The switch regions of both proteins are significantly more flexible

than other parts of proteins. However, SI shows more flexibility in p Tyr32-G12D while SII shows more flexibility in G12D K-Ras. The analysis of the GTP binding site shows that both mutations remain in active state 1 which can interact with its effectors and leading to oncogenic signal output. Our results show that differential dynamics and conformations observed have implications for functional specialization including in GTPase activity and effector interaction. The simulations also revealed an interaction of sodium ions with the GTP and neighboring residues. These interactions were more prominent in the phosphorylated mutant. In this mutant, the entrance of a sodium ion is directly coupled with residues 30, 31 and 33. Ion binding also appears to be contributing to the orientation and displacement of Tyr32 toward γ-phosphate of GTP position. MSMs confirmed the effects of phosphorylation on conformations and dynamics. Namely, the switch regions of K-Ras. Finally, MSMs reveal the role of Tyr32 in determining the metastable states of both mutations. This suggests a direct potential of Tyr32 movement and orientation on GTP hydrolysis and effectors binding.

References

1. Vetter, I.R. and A. Wittinghofer, *The Guanine Nucleotide-Binding Switch in Three Dimensions.* Science, 2001. **294**(5545): p. 1299-1304.
2. Hancock, J.F., et al., *All ras proteins are polyisoprenylated but only some are palmitoylated.* Cell, 1989. **57**(7): p. 1167-1177.
3. Cox, A.D. and C.J. Der, *Ras history: The saga continues.* Small GTPases, 2010. **1**(1): p. 2-27.
4. Hall, B.E., D. Bar-Sagi, and N. Nassar, *The structural basis for the transition from Ras-GTP to Ras-GDP.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(19): p. 12138-12142.
5. Bourne, H.R., D.A. Sanders, and F. McCormick, *The GTPase superfamily: a conserved switch for diverse cell functions.* Nature, 1990. **348**: p. 125.
6. Campbell, S.L., et al., *Increasing complexity of Ras signaling.* Oncogene, 1998. **17**(11): p. 1395.
7. Sprang, S.R., *G proteins, effectors and GAPs: structure and mechanism.* Current opinion in structural biology, 1997. **7**(6): p. 849-856.
8. Cherfils, J. and M. Zeghouf, *Regulation of small gtpases by gefs, gaps, and gdis.* Physiological reviews, 2013. **93**(1): p. 269-309.
9. Corbett, K.D. and T. Alber, *The many faces of Ras: recognition of small GTP-binding proteins.* Trends in biochemical sciences, 2001. **26**(12): p. 710-716.
10. Drosten, M., et al., *Genetic analysis of Ras signalling pathways in cell proliferation, migration and survival.* The EMBO journal, 2010. **29**(6): p. 1091-1104.
11. Crespo, P. and J. Leon, *Ras proteins in the control of the cell cycle and cell differentiation.* Cellular and Molecular Life Sciences CMLS, 2000. **57**(11): p. 1613-1636.
12. Downward, J., *Targeting RAS signalling pathways in cancer therapy.* Nature Reviews Cancer, 2003. **3**(1): p. 11.
13. Gao, C. and L.A. Eriksson, *Impact of mutations on K-Ras-p120GAP interaction.* Computational Molecular Bioscience, 2013. **3**(02): p. 9.
14. Nassar, N., et al., *The 2.2 A crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue.* Nature, 1995. **375**(6532): p. 554-60.
15. Klink, B.U., R.S. Goody, and A.J. Scheidig, *A newly designed microspectrofluorometer for kinetic studies on protein crystals in combination with x-ray diffraction.* Biophysical journal, 2006. **91**(3): p. 981-992.
16. Lito, P., et al., *Allele-specific inhibitors inactivate mutant KRAS G12C by a trapping mechanism.* Science, 2016. **351**(6273): p. 604-608.
17. Gibbs, J.B., et al., *Intrinsic GTPase activity distinguishes normal and oncogenic ras p21 molecules.* Proceedings of the National Academy of Sciences, 1984. **81**(18): p. 5704-5708.
18. Prior, I.A., P.D. Lewis, and C. Mattos, *A comprehensive survey of Ras mutations in cancer.* Cancer research, 2012. **72**(10): p. 2457-2467.
19. Hunter, J.C., et al., *Biochemical and structural analysis of common cancer-associated KRAS mutations.* Molecular cancer research, 2015: p. molcanres. 0203.2015.
20. Pantsar, T., et al., *Assessment of mutation probabilities of KRAS G12 missense mutants and their long-timescale dynamics by atomistic molecular simulations and Markov state modeling.* PLOS Computational Biology, 2018. **14**(9): p. e1006458.
21. Baussand, J. and J. Kleinjung, *Specific conformational states of Ras GTPase upon effector binding.* Journal of chemical theory and computation, 2012. **9**(1): p. 738-749.

22. Geyer, M., et al., *Conformational transitions in p21 ras and in its complexes with the effector protein Raf-RBD and the GTPase activating protein GAP.* Biochemistry, 1996. **35**(32): p. 10308-10320.
23. Milburn, M.V., et al., *Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins.* Science, 1990. **247**(4945): p. 939-945.
24. Scarabelli, G. and B.J. Grant, *Mapping the Structural and Dynamical Features of Kinesin Motor Domains.* PLOS Computational Biology, 2013. **9**(11): p. e1003329.
25. Markevich, N.I., J.B. Hoek, and B.N. Kholodenko, *Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades.* The Journal of Cell Biology, 2004. **164**(3): p. 353-359.
26. Theillet, F.-X., et al., *Cell signaling, post-translational protein modifications and NMR spectroscopy.* Journal of biomolecular NMR, 2012. **54**(3): p. 217-236.
27. Kano, Y., et al. *New structural and functional insight into the regulation of Ras*. in *Seminars in cell & developmental biology*. 2016. Elsevier.
28. Wee, P. and Z. Wang, *Epidermal growth factor receptor cell proliferation signaling pathways.* Cancers, 2017. **9**(5): p. 52.
29. Bunda, S., et al., *Inhibition of SHP2-mediated dephosphorylation of Ras suppresses oncogenesis.* Nature Communications, 2015. **6**: p. 8859.
30. Kano, Y., et al., *Tyrosyl phosphorylation of KRAS stalls GTPase cycle via alteration of switch I and II conformation.* Nature Communications, 2019. **10**(1): p. 224.
31. Bunda, S., et al., *Src promotes GTPase activity of Ras via tyrosine 32 phosphorylation.* Proc Natl Acad Sci U S A, 2014. **111**(36): p. E3785-94.
32. Lu, S., et al., *The Structural Basis of Oncogenic Mutations G12, G13 and Q61 in Small GTPase K-Ras4B.* Sci Rep, 2016. **6**: p. 21949.
33. Kapoor, A. and A. Travesset, *Differential dynamics of RAS isoforms in GDP- and GTP-bound states.* Proteins, 2015. **83**(6): p. 1091-106.
34. Sayyed-Ahmad, A., P. Prakash, and A. Gorfe, *Distinct dynamics and interaction patterns in H- and K-Ras oncogenic P-loop mutants.* 2017. **85**: p. 1618–1632.
35. Prakash, P., et al., *Computational and biochemical characterization of two partially overlapping interfaces and multiple weak-affinity K-Ras dimers.* Scientific Reports, 2017. **7**: p. 40109.
36. Prakash, P., A. Sayyed-Ahmad, and A.A. Gorfe, *The role of conserved waters in conformational transitions of Q61H K-ras.* PLoS computational biology, 2012. **8**(2): p. e1002394.
37. Prakash, P., A. Sayyed-Ahmad, and A.A. Gorfe, *pMD-Membrane: a method for ligand binding site Identification in membrane-bound proteins.* PLoS computational biology, 2015. **11**(10): p. e1004469.
38. Sarkar-Banerjee, S., et al., *Spatiotemporal analysis of K-Ras plasma membrane interactions reveals multiple high order homo-oligomeric complexes.* Journal of the American Chemical Society, 2017. **139**(38): p. 13466-13475.
39. Sayyed-Ahmad, A., et al., *Computational Equilibrium Thermodynamic and Kinetic Analysis of K-Ras Dimerization through an Effector Binding Surface Suggests Limited Functional Role.* J Phys Chem B, 2016. **120**(33): p. 8547-56.
40. Sayyed-Ahmad, A. and A.A. Gorfe, *Mixed-Probe Simulation and Probe-Derived Surface Topography Map Analysis for Ligand Binding Site Identification.* 2017. **13**(4): p. 1851-1861.
41. van Gunsteren, W.F. and H.J.C. Berendsen, *Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry.* Angewandte Chemie International Edition in English, 1990. **29**(9): p. 992-1023.
42. Bolintineanu, D.S., et al., *Poisson-Nernst-Planck models of nonequilibrium ion electrodiffusion through a protegrin transmembrane pore.* PLoS computational biology, 2009. **5**(1): p. e1000277.

43. Bolintineanu, D.S., et al., *Investigation of changes in tetracycline repressor binding upon mutations in the tetracycline operator.* Journal of Chemical & Engineering Data, 2014. **59**(10): p. 3167-3176.
44. Langham, A.A., A. Sayyed-Ahmad, and Y.N. Kaznessis, *On the nature of antimicrobial activity: a model for Protegrin-1.* 2008.
45. Naddaf, L. and A. Sayyed-Ahmad, *Intracellular crowding effects on the self-association of the bacterial cell division protein FtsZ.* Archives of biochemistry and biophysics, 2014. **564**: p. 12-19.
46. Prakash, P., et al., *Aggregation behavior of ibuprofen, cholic acid and dodecylphosphocholine micelles.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2012. **1818**(12): p. 3040-3047.
47. Sayyed-Ahmad, A. and Y. Kaznessis, *Determining the orientation of the Beta-hairpin antimicrobial peptide Protegrin-1 in a DLPC lipid bilayer using an implicit solvent-membrane model.* 2009.
48. Sayyed-Ahmad, A., H. Khandelia, and Y.N. Kaznessis, *Relative free energy of binding between antimicrobial peptides and SDS or DPC micelles.* Molecular simulation, 2009. **35**(10-11): p. 986-997.
49. Sayyed-Ahmad, A., L.M. Lichtenberger, and A.A. Gorfe, *Structure and dynamics of cholic acid and dodecylphosphocholine− cholic acid aggregates.* Langmuir, 2010. **26**(16): p. 13407-13414.
50. Patodia, S., A. Bagaria, and D. Chopra, *Molecular Dynamics Simulation of Proteins: A Brief Overview.* Journal of Physical Chemistry & Biophysics, 2014. **4**(6): p. 1.
51. Buck, M., et al., *Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme.* Biophysical journal, 2006. **90**(4): p. L36-L38.
52. Kollman, P.A., *Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules.* Accounts of Chemical Research, 1996. **29**(10): p. 461-469.
53. Schuler, L.D., X. Daura, and W.F. van Gunsteren, *An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase.* Journal of Computational Chemistry, 2001. **22**(11): p. 1205-1218.
54. Basharin, G.P., A.N. Langville, and V.A. Naumov, *The life and work of AA Markov.* Linear algebra and its applications, 2004. **386**: p. 3-26.
55. FISCHER, C.S.A. and W.H.P. DEUFLHARD, *A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo.*
56. Noé, F., et al., *Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states.* The Journal of chemical physics, 2007. **126**(15): p. 04B617.
57. Chodera, J.D. and F. Noé, *Probability distributions of molecular observables computed from Markov models. II. Uncertainties in observables and their time-evolution.* The Journal of chemical physics, 2010. **133**(10): p. 09B606.
58. Singhal, N. and V.S. Pande, *Error analysis and efficient sampling in Markovian state models for molecular dynamics.* The Journal of chemical physics, 2005. **123**(20): p. 204909.
59. Beauchamp, K.A., et al., *MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale.* Journal of chemical theory and computation, 2011. **7**(10): p. 3412-3419.
60. Scherer, M.K., et al., *PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models.* Journal of Chemical Theory and Computation, 2015. **11**(11): p. 5525-5542.
61. Bowman, G.R. and V.S. Pande, *Protein folded states are kinetic hubs.* Proceedings of the National Academy of Sciences, 2010. **107**(24): p. 10890-10895.
62. Plattner, N. and F. Noé, *Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models.* Nature Communications, 2015. **6**: p. 7653.

63. Held, M., et al., *Mechanisms of protein-ligand association and its modulation by protein mutations.* Biophysical journal, 2011. **100**(3): p. 701-710.
64. Singhal, N., C.D. Snow, and V.S. Pande, *Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin.* The Journal of chemical physics, 2004. **121**(1): p. 415-425.
65. Schor, M., et al., *Shedding light on the dock–lock mechanism in amyloid fibril growth using Markov state models.* The journal of physical chemistry letters, 2015. **6**(6): p. 1076-1081.
66. Pérez-Hernández, G., et al., *Identification of slow molecular order parameters for Markov model construction.* The Journal of chemical physics, 2013. **139**(1): p. 07B604_1.
67. Noé, F. and F. Nuske, *A variational approach to modeling slow processes in stochastic dynamical systems.* Multiscale Modeling & Simulation, 2013. **11**(2): p. 635-655.
68. Djurdjevac, N., M. Sarich, and C. Schütte, *Estimating the eigenvalue error of Markov state models.* Multiscale Modeling & Simulation, 2012. **10**(1): p. 61-81.
69. Prinz, J.-H., et al., *Markov models of molecular kinetics: Generation and validation.* The Journal of chemical physics, 2011. **134**(17): p. 174105.
70. Amadei, A., A.B. Linssen, and H.J. Berendsen, *Essential dynamics of proteins.* Proteins: Structure, Function, and Bioinformatics, 1993. **17**(4): p. 412-425.
71. Hotelling, H., *Analysis of a complex of statistical variables into principal components.* Journal of educational psychology, 1933. **24**(6): p. 417.
72. Molgedey, L. and H.G. Schuster, *Separation of a mixture of independent signals using time delayed correlations.* Physical Review Letters, 1994. **72**(23): p. 3634-3637.
73. Sarich, M., F. Noé, and C. Schütte, *On the approximation quality of Markov state models.* Multiscale Modeling & Simulation, 2010. **8**(4): p. 1154-1177.
74. Medvedev, N., *The algorithm for three-dimensional Voronoi polyhedra.* Journal of computational physics, 1986. **67**(1): p. 223-229.
75. Swope, W.C., J.W. Pitera, and F. Suits, *Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory.* The Journal of Physical Chemistry B, 2004. **108**(21): p. 6571-6581.
76. Trendelkamp-Schroer, B., et al., *Estimation and uncertainty of reversible Markov models.* The Journal of Chemical Physics, 2015. **143**(17): p. 174101.
77. Noé, F., et al., *Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations.* Proceedings of the National Academy of Sciences, 2009. **106**(45): p. 19011-19016.
78. Metzner, P., C. Schütte, and E. Vanden-Eijnden, *Transition path theory for Markov jump processes.* Multiscale Modeling & Simulation, 2009. **7**(3): p. 1192-1219.
79. Arthur, D. and S. Vassilvitskii. *k-means++: The advantages of careful seeding*. in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007. Society for Industrial and Applied Mathematics.
80. Röblitz, S. and M. Weber, *Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification.* Advances in Data Analysis and Classification, 2013. **7**(2): p. 147-179.
81. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems.* The Journal of chemical physics, 1993. **98**(12): p. 10089-10092.
82. Ryckaert, J.-P., G. Ciccotti, and H.J. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of computational physics, 1977. **23**(3): p. 327-341.
83. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of computational chemistry, 2005. **26**(16): p. 1781-1802.

84. E, W. and E. Vanden-Eijnden, *Towards a Theory of Transition Paths.* Journal of Statistical Physics, 2006. **123**(3): p. 503-523.
85. Grant, B.J., A.A. Gorfe, and J.A. McCammon, *Ras conformational switching: simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics.* PLoS computational biology, 2009. **5**(3): p. e1000325.
86. Kearney, B.M., et al., *DRoP: a water analysis program identifies Ras-GTP-specific pathway of communication between membrane-interacting regions and the active site.* Journal of molecular biology, 2014. **426**(3): p. 611-629.
87. Buhrman, G., et al., *Allosteric modulation of Ras positions Q61 for a direct role in catalysis.* Proceedings of the National Academy of Sciences, 2010. **107**(11): p. 4931.
88. Thapar, R., J.G. Williams, and S.L. Campbell, *NMR characterization of full-length farnesylated and non-farnesylated H-Ras and its implications for Raf activation.* Journal of molecular biology, 2004. **343**(5): p. 1391-1408.
89. Fetics, S.K., et al., *Allosteric effects of the oncogenic RasQ61L mutant on Raf-RBD.* Structure, 2015. **23**(3): p. 505-516.
90. Pacold, M.E., et al., *Crystal structure and functional analysis of Ras binding to its effector phosphoinositide 3-kinase γ.* Cell, 2000. **103**(6): p. 931-944.
91. Ma, J. and M. Karplus, *Molecular switch in signal transduction: reaction paths of the conformational changes in ras p21.* Proceedings of the National Academy of Sciences, 1997. **94**(22): p. 11905-11910.
92. Hall, B.E., et al., *Structure-based mutagenesis reveals distinct functions for Ras switch 1 and switch 2 in Sos-catalyzed guanine nucleotide exchange.* Journal of Biological Chemistry, 2001. **276**(29): p. 27629-27637.
93. Adjei, A.A., *Blocking oncogenic Ras signaling for cancer therapy.* Journal of the National Cancer Institute, 2001. **93**(14): p. 1062-1074.
94. Spoerner, M., et al., *Conformational states of Ras complexed with the GTP analogue GppNHp or GppCH2p: implications for the interaction with effector proteins.* Biochemistry, 2005. **44**(6): p. 2225-2236.
95. Spoerner, M., et al., *Dynamic properties of the Ras switch I region and its importance for binding to effectors.* Proceedings of the National Academy of Sciences, 2001. **98**(9): p. 4944-4949.
96. Spoerner, M., et al., *Conformational states of human rat sarcoma (Ras) protein complexed with its natural ligand GTP and their role for effector interaction and GTP hydrolysis.* Journal of biological chemistry, 2010. **285**(51): p. 39768-39778.
97. Liao, J., et al., *Two conformational states of Ras GTPase exhibit differential GTP-binding kinetics.* Biochemical and biophysical research communications, 2008. **369**(2): p. 327-332.
98. Bowman, G.R., X. Huang, and V.S. Pande, *Using generalized ensemble simulations and Markov state models to identify conformational states.* Methods, 2009. **49**(2): p. 197-201.

APPENDICES

Appendix A: NAMD Molecular dynamics scripts

A.1: TCL script to add harmonic restrains on CA atoms

File name: Restrain.tcl

```
mol load pdb "kras-wt.pdb"

set all [atomselect top "all"]
set fixatom0 [atomselect top "name CA"]

$all set beta 0.0
$fixatom0 set beta 0.0
$all writepdb  kras-wt-restrain.pdb

exit
```

A.2: Equilibration script

File name: equ.namd

```
set MOL kras-wt
set temp 310.15
set boxw 64
set boxz 64

#Can get this from ?xtla parameter of last frame in the
#charmm trajectory.

structure $[98].psf
coordinates $[98].pdb
temperature 0

paraTypeCharmm on
parameters          toppar/par_all36m_prot.prm
parameters          toppar/par_all36_na.prm
parameters          toppar/par_all36_carb.prm
parameters          toppar/par_all36_lipid.prm
parameters          toppar/par_all36_cgenff.prm
parameters          toppar/toppar_water_ions.str
parameters          toppar/toppar_all36_na_nad_ppi.str

outputEnergies  5000
outputTiming    5000
xstFreq         5000
dcdFreq         5000
wrapAll         on
```

```
wrapNearest     on

rigidbonds      all
timestep        2
nonBondedFreq   1
fullElectFrequency 2
stepsPerCycle   10
pairlistsPerCycle    2

switching       on
vdwForceSwitching   yes
switchDist      8
cutoff          10
pairlistdist    12

outputname $[98].0
binaryoutput on

restartname $[98].0
restartfreq 5000

#########################################
cellBasisVector1        $boxw      00.00   00.00
cellBasisVector2        00.00   $boxw   00.00
cellBasisVector3        00.00   00.00   $boxz
cellOrigin          0.    0.    0.
#########################################
Pme             on
PmeGridsizeX    $boxw
PmeGridsizeY    $boxw
PmeGridsizeZ    $boxz

exclude         scaled1-4
1-4scaling      1.0


########################################
# PRESSURE AND TEMPERATURE CONTROL
########################################

langevin                on
langevinDamping         1
langevinTemp            $temp
langevinHydrogen        no

langevinPiston          on
langevinPistonTarget    1.01325
langevinPistonPeriod    200
langevinPistonDecay     100
langevinPistonTemp      $temp
```

```
useGroupPressure      yes     #THIS WILL ALLOW THE SYSTEM SMALLER FLUCTUATIONS
useFlexibleCell no
useConstantRatio no

#
# Restrained atoms for initial heating-up steps
#
constraints on
consRef $[98]-restrain.pdb
consKFile $[98]-restrain.pdb
consKCol B
constraintScaling       10.0



########################################
#Minimize
########################################
minimize 10000
output $[98].min
########################################
# Heat
########################################

set tem 99.15;
langevinPiston on
while { $tem < $temp } {
 langevinTemp $tem
 run 50
 output ${MOL}.heat
 set tem [expr $tem + 20.0]
}

run 25000
constraintScaling       5.0
run 25000
constraintScaling       2.5
run 25000
constraintScaling       1.25
run 25000
constraintScaling       0.0

########################################
# Run
########################################
run 500000
output ${MOL}.0
```

A.3: Production script

File name: prod.namd

```
set MOL kras-wt
set temp 310.15
set boxw 64
set boxz 64


#Can get this from ?xtla parameter of last frame in the
#charmm trajectory.
set freq 5000
set i  0
set j [expr $i + 1]
set previousfile ${MOL}.$i
set nextfile  ${MOL}.$j
firsttimestep 5060000

structure ${MOL}.psf
coordinates ${MOL}.pdb
bincoordinates     ./${previousfile}.coor
binvelocities      ./${previousfile}.vel
extendedSystem     ./${previousfile}.xsc


paraTypeCharmm on
parameters         toppar/par_all36m_prot.prm
parameters         toppar/par_all36_na.prm
parameters         toppar/par_all36_carb.prm
parameters         toppar/par_all36_lipid.prm
parameters         toppar/par_all36_cgenff.prm
parameters         toppar/toppar_water_ions.str
parameters         toppar/toppar_all36_na_nad_ppi.str


outputEnergies  $freq
outputTiming    $freq
xstFreq         $freq
dcdFreq         $freq
wrapAll         on
wrapNearest     on

rigidbonds      all
timestep        2
nonBondedFreq   1
fullElectFrequency 2
stepsPerCycle   10

switching       on
```

```
vdwForceSwitching   yes
switchDist      8
cutoff          10
pairlistdist    12

outputname ${MOL}.$j
binaryoutput off

restartname ${MOL}.$j
restartfreq $freq

Pme             on
PmeGridsizeX    $boxw
PmeGridsizeY    $boxw
PmeGridsizeZ    $boxw

exclude         scaled1-4
1-4scaling      1.0


########################################
# PRESSURE AND TEMPERATURE CONTROL
########################################
langevin                on
langevinDamping         10
langevinTemp            $temp
langevinHydrogen        no

langevinPiston          on
langevinPistonTarget    1.01325
langevinPistonPeriod    200
langevinPistonDecay     100
langevinPistonTemp      $temp

useGroupPressure        yes     #THIS WILL ALLOW THE SYSTEM SMALLER FLUCTUATIONS
useFlexibleCell no
useConstantRatio no



output ${nextfile}
########################################
# Run
########################################
run 500000000
```

Appendix B: Markov state models software scripts

To build Markov state models we used the Python package PyEMMA [60] with the following scripts containing the inputs with some comments:

```
 import pyemma
pyemma.__version__

import numpy as np
%pylab inline

import os
%pylab inline
matplotlib.rcParams.update([93])

import pyemma.coordinates as coor

# some helper funcs
def average_by_state(dtraj, x, nstates):
    assert(len(dtraj) == len(x))
    N = len(dtraj)
    res = np.zeros((nstates))
    for i in range(nstates):
        I = np.argwhere(dtraj == i)[:,0]
        res[i] = np.mean(x[I])
    return res

def avg_by_set(x, sets):
    # compute mean positions of sets. This is important because of some technical points the set order
    # in the coarse-grained TPT object can be different from the input order.
    avg = np.zeros(len(sets))
    for i in range(len(sets)):
        I = list(sets[i])
        avg[i] = np.mean(x[I])
    return avg


# input dcd and pdb files
trajfile = 'input.dcd'
topfile = 'input.pdb'

feat = coor.featurizer(topfile)

inp = coor.source(trajfile, feat)
print('trajectory length = ',inp.trajectory_length(0))
print('number of dimension = ',inp.dimension())

# TICA with 1 ns (10 steps ) lag time
lag=10
tica_obj = coor.tica(inp, lag=lag,dim=2, kinetic_map=False)
```

```
# here we get the data that has been projected onto the first 2 IC's.
Y = tica_obj.get_output()[0]

# plot IC1 and IC2
subplot2grid((2,1),(0,0))
plot(Y[:,0])
ylabel('ind. comp. 1')
subplot2grid((2,1),(1,0))
plot(Y[:,1])
ylabel('ind. comp. 2')
xlabel('time (100 ps)')

print('Mean values: ', np.mean(Y, axis=0))
print('Variances:   ', np.var(Y, axis=0))

# relaxation timescales
print(-lag/np.log(tica_obj.eigenvalues[:5]))
#The eigenvalues of the TICA transform are the values of these autocorrelations at the chosen lag time. We can even interpret them in terms of relaxation timescales


# histogram data
z,x,y = np.histogram2d(Y[:,0],Y[:,1], bins=50)
# compute free energies
F = -np.log(z)
# contour plot
extent = [x[0], x[-1], y[0], y[-1]]
#contourf(F.T, 50, cmap=plt.cm.seismic, extent=extent,interpolation='nearest')
contourf(F.T, 50, cmap=plt.cm.hot, extent=extent)
#save_figure('kras_ph_histogarm.png')
plt.axis([-2.5 , 2.5, -2.5 ,2.5])

# kmeans clustering with 100 cluster centers and 20 iterations
cl = coor.cluster_kmeans(data=Y, k=100, stride=1,max_iter=20)
dtrajs = cl.dtrajs
cc_x = cl.clustercenters[:,0]
cc_y = cl.clustercenters[:,1]

# plot histogram with kmeans cluster crnters
contourf(F.T,50, cmap=plt.cm.hot, extent=extent)
plot(cc_x,cc_y, linewidth=0, marker='.')

# MSM estimation
import pyemma.msm as msm
import pyemma.plots as mplt

# MSM estimation
# To estimate a Markov model at each of the given lag times τ (that are multiples of our saving step, multipl es of 100 ps), compute the eigenvalues of each transition matrix, λi(τ), and then compute the relaxation time scales
lags = [1,2,5,10,20,50,100]
its = msm.its(dtrajs, lags=lags, nits=7, reversible=True, connected=True, weights='empirical', errors=None , nsamples=50, n_jobs=None, show_progress=True, mincount_connectivity='1/n')
```

```
#compute errors using Bayesian MSMs and plot
its = msm.its(dtrajs, lags=lags, nits=7,errors='bayes')

mplt.plot_implied_timescales(its)
ylim(0,10000)
xlim(0,100)

# Bayesian Markov model estimation, discrete trajectories obtained from the clustering and the lag time 2 n
s(20 steps):
M = msm.bayesian_markov_model(dtrajs,20,reversible=True)
print('fraction of states used = ', M.active_state_fraction)
print('fraction of counts used = ', M.active_count_fraction)

# Spectral analysis
# timescale computed from MSM eigenvalues
plot(M.timescales()/10,linewidth=0,marker='o')
xlabel('index'); ylabel('timescale (ns)'); xlim(-0.5,10.5)
# timescale separation
plot(M.timescales()[:-1]/M.timescales()[1:], linewidth=0,marker='o')
xlabel('index'); ylabel('timescale separation'); xlim(-0.5,10.5)

# PCCA++ do with 5 states now
M.pcca(5)
pcca_dist = M.metastable_distributions

# Representative Structures with 10 frame indexes
pcca_samples = M.sample_by_distributions(pcca_dist, 10)

coor.save_traj(inp, pcca_samples[0], 'output1.dcd')
coor.save_traj(inp, pcca_samples[1], 'output2.dcd')
coor.save_traj(inp, pcca_samples[2], 'output3.dcd')
coor.save_traj(inp, pcca_samples[3], 'output4.dcd')
coor.save_traj(inp, pcca_samples[4], 'output5.dcd')

# Transition pathways and Committors
# do PCCA++ with 5 states now
M.pcca(5)
pcca_sets_5 = M.metastable_sets

# Plot PCCA++ macrostates
figure(figsize=(8,5))
pcca_sets = M.metastable_sets
contourf(F.T, 50, cmap=plt.cm.hot, extent=extent)
size = 50
cols = ['gray' ,'blue','green','black','purple']
for i in range(5):
    scatter(cc_x[pcca_sets_5[i]], cc_y[pcca_sets_5[i]], color=cols[i], s=size)

#we select as two end-states the leftmost and the rightmost of these 5 sets
# The average positions along the first TICA coordinate:
xavg = avg_by_set(cc_x, pcca_sets_5)
A = pcca_sets_5[xavg.argmax()]
```

```
B = pcca_sets_5[xavg.argmin()]


# compute mean positions of sets. This is important because of some technical points the set order in the coa
rse-grained TPT object can be different from the input order.
avgpos = np.zeros((5,2))
avgpos[:,0] = avg_by_set(cc_x, cg)
avgpos[:,1] = avg_by_set(cc_y, cg)

# TPT with reversible process
# Plot the metastable states and the transitions
fig, _=mplt.plot_flux(cgflux, avgpos, attribute_to_plot='gross_flux')
cf=contourf(F.T, 50, cmap=plt.cm.hot, extent=extent, fig=fig, zorder=0)

# decompose the flux into individual pathways, along with their fluxes by:
paths, path_fluxes = cgflux.pathways(fraction=0.99)
print('percentage       \tpath')
print('-------------------------------------')
for i in range(len(paths)):
    print((path_fluxes[i] / np.sum(path_fluxes)),' \t', paths[i])

# Validate MSM with 2 ns (20 steps ) lag time and 5 metastable states by  CK test.
ck = M.cktest(5, mlags=11, err_est=True)
mplt.plot_cktest(ck, diag=True, figsize=(7,7),layout=(3,2),  padding_top=0.1, y01=True, padding_between
=0.1, dt=0.1, units=' ns')
```