Combined Computational and Solution Scattering Methods to Study the Dynamical Properties of Ras

by Hugo Guterres

B.S. Biochemistry, University of Richmond

A dissertation submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

March 26, 2018

Dissertation directed by

Carla Mattos
Professor of Chemistry and Chemical Biology
Dedication

This dissertation is dedicated to my parents, Honório Sarmento Guterres and Florinda Freitas da Silva.
Acknowledgement

My journey of pursuing this PhD degree has been filled with many collaborations and countless supports from fellow scientists, friends and family. I would like to acknowledge all of you for your kind supports.

Mom and dad, thank you for your love and support throughout this journey. Both of you have raised me with great real-life examples on how to follow my dream. To my siblings, Dr. Helio Guterres and Dr. Bernardette Guterres, the two of you have inspired my journey to pursue higher degree in science. I was always lucky to have the two of you leading the way from the very beginning.

Thank you to the department of chemistry and chemical biology at Northeastern University. Your faculty, staff and financial assistance have been instrumental in my graduate career. It has been a privilege to pursue my doctoral degree in this department and for that, I am thankful.

To the Mattos lab members, both past and present, thank you for the supportive and collaborative environment. Dr. Ryan Knihtila and Dr. Christian Johnson, thank you for training me in the lab with protein expression, purification and crystallization as well as hydrolysis experiments. Dr. Tim Colussi, thank you for training me with protein crystallography, your skills and expertise were very helpful during my time in the lab. Dr. Jillian Parker, thank you for all the insightful Ras discussions and your help in protein biochemistry lab techniques. Your knowledge and expertise in the lab were always helpful during the difficult times I had in trying to get positive results. Dr. Kendra Marcus, thank you for insightful discussions on Ras computational simulations and for being a helpful computational expert in the lab. Your expertise in Ras superfamily and in computational methods have always been very helpful during my graduate
career, given that I do a lot of computational work. Derion Reid and Alicia Volmar, thank you for being helpful lab mates during my time in the lab. The three of us started around the same time in the wet lab, and I have always worked very well alongside the two of you. Thank you to the youngest lab members, Morgan Packer, Monica Ojeda, Dr. Sorabh Agarwal, Thamires Froes and Trinity Cookis. Thank you to the undergraduates that I have worked with, Justin Clark and Kelsey Charlie, for your help in the purification of RasY71A mutant.

Throughout my graduate career I have had the honor of collaborating with many computational chemists, including Dr. Buyong Ma, Dr. Ruth Nussinov, Dr. Jiali Gao, Dr. Alessandro Cembran, and Dr. Rebecca Swett. Special thanks to Dr. Buyong Ma and Dr. Ruth Nussinov for the very helpful initial training I had in the summer of 2013 to run, set up and analyze molecular dynamics simulations. Thank you for your valuable collaborations throughout the first few years of my graduate career and for answering many of my questions throughout the years. Thank you to Dr. Jiali Gao and Dr. Alessandro Cembran for the training with CHARMM MD simulations in the summer of 2014. Thank you to Dr. Rebecca Swett for introducing me to the computational methods in pharmaceutical industry drug discovery, during the summer of 2017. I gained much of the computational expertise from all of you who have mentored me throughout my graduate career, and I am eternally grateful for that. Special thanks to Dr. Nilay Roy, former director of Northeastern Discovery cluster, high-performance computing (hpc) system. Thank you for all your help throughout the years in working with the hpc system. Thank you also to Northeastern discovery cluster for the state-of-the-art computing system that has made it possible to conduct many of my computational simulation work.

Thank you to my long-standing collaborators, Dr. Lee Makoswki and Hao Zhou for the amazing project of combining computational simulations with X-ray solution scattering. I have
learned a lot of X-ray solution scattering methodology, data collection, processing, and analysis from the two of you. I am very excited for the results from this collaboration.

Thank you to my thesis committee members, Dr. Mary Jo Ondrechen, Dr. Lee Makowski and Dr. Paul C. Whitford. The consistent inputs, comments and ideas from all of you throughout my graduate career have been tremendously helpful in guiding my research direction. I am honored to have the three of you serving in thesis committee.

To my mentor and thesis advisor, Dr. Carla Mattos, thank you for everything. Thank you for your training in protein crystallography, biochemistry, structural biology of Ras GTPase, and computational chemistry. You have inspired me throughout these years with your great talent and enthusiasm in structural biochemistry. Your optimism and perseverance to tackle very difficult questions within the structural biology of Ras have been inspirational. Thank you for all the opportunities for summer internships, at the National Cancer Institute (NCI), the University of Minnesota, and Vertex pharmaceuticals. Thank you for the opportunities to present at numerous science conferences throughout the years. Most of all, thank you for your mentorship, you have shaped the way I think, write and talk about science.

This work was funded by NSF grants, NSF MCB-1244203 and MCB-1517295.
Abstract of Dissertation

Ras is a small GTPase that acts as a molecular switch in several signal transduction networks to control cell proliferation, survival and apoptosis. It is the first oncogene product that was discovered more than four decades ago and is found in about 20% of human cancers. It cycles between inactive GDP-bound and active GTP-bound states. Ras in its GTP-bound form is highly dynamic. Among its many conformational states, are the open conformation (state 1) or closed conformation (state 2) of switch I, while switch II can be found in a disordered state (T state) or an ordered conformation (R state). There are three main isoforms of Ras, H-, K-, and N-Ras which share about 95% sequence identity in their soluble domain. Recent biochemical data show that the three isoforms have different hydrolysis rate constants. The work presented in this thesis characterize the balance of conformational states in H-, K-, and N-Ras using accelerated MD paired with X-ray solution scattering. We show that there are differences in the balance of conformational states with different access to the catalytic conformation in each of the isoforms. The isoform specific residues and their influences on the conformational states of Ras are discussed. Furthermore, we analyze the allosteric effects in Ras that controls intrinsic GTP hydrolysis in the presence of Raf. Simulations of Ras/Raf and oncogenic mutant RasQ61L/Raf are presented, where we show substantial long-range effects of mutant RasQ61L on both Ras and Raf-RBD proteins. Recent studies have shown the importance of Ras dimerization in the activation of Ras/Raf/MEK/ERK signaling in the cell. We conduct Ras dimer simulations in the presence of Raf-RBD to provide insights into the dimer effects on Ras dynamics. Additionally, we present a crystal structure of H-RasY71A showing its effect on the conformational balance and intrinsic hydrolysis of H-Ras.
Table of Contents

Copyright 2
Dedication 3
Acknowledgements 4
Abstract of Dissertation 7
Table of Contents 8
List of Figures 12
List of Tables 15

Chapter 1: Introduction 16
  Ras in cancer 17
  Ras: Structure and dynamics 21
  Molecular dynamics of Ras 27
  Summary and thesis outline 32
  Reference 36

Chapter 2: Allosteric effects of the oncogenic mutant RasQ61L on Raf-RBD 42
  Introduction 43
  Materials and Methods 48
  Molecular dynamics simulations 48
  Dynamical network analysis 49
  Results 51
  Wild type Ras bound to Raf-RBD 51
  RasQ61L-GppNHp bound to Raf-RBD 56
  MD Simulations show altered flexibility due to Q61L mutation 62
  Dynamic networks of communication between the Ras allosteric site and Raf-RBD L4 69
  Discussion 74
  Reference 77

Chapter 3: Structure, Dynamics and Conformational States of K-, H-, and N-Ras 81
Assessed by a Combination of MD Simulations and Solution X-ray Scattering 82
  Introduction 82
Chapter 4: RasY71A Mutant Increases the Flexibility of Switch II and Intrinsic Hydrolysis Ras Constant

Introduction

Materials and Methods

Protein expression and purification

Intrinsic hydrolysis experiments
Chapter 5: Ras/Raf Dimerization Effects on Allosteric Networks

Introduction

Materials and Methods

Molecular dynamics

Dynamical network analysis

Results

Dimer formation affects the flexibility of residues at switch I, switch II and the allosteric site

Dynamical network analysis shows the effect of dimer formation that is similar to an oncogenic RasQ61L effect on the Ras/RafRBD allosteric network

Discussion

Reference

Appendices

Appendix A: Assessing MD Force Fields for RNA Simulations

Introduction
Methods

Results

Simulation of 4loop structure is stable under AMBER12 and OPLS3, but not CHARMM27 or CHARMM36 force fields

Simulation of LCS1co structure is stable in OPLS3, but not in AMBER12, CHARMM27, and CHARMM36 force fields

Simulation of siG12D duplex structure is most stable under OPLS3 and AMBER12, and unstable under CHARMM36

Intermolecular interaction in ZIKV sfRNA tetramer is only well maintained in AMBER12, but not in OPLS3 and CHARMM36 force fields

Simulations with magnesium ions introduce internal base pair fraying and rigid RNA structure for all three force fields

Simulations with TIP4P water increases the stability of RNA using OPLS3 force field

Simulations with CHARMM36 using two different initial ions positioning dramatically affects the stability of RNA

Summary

Reference

Appendix B: Laboratory Protocols for Computational Simulations

Molecular dynamics

Data Analysis

RMSD

RMSF

Distance over time between atoms

Conformational clustering

Average structure

Dynamical network analysis

Principal component analysis

Accelerated MD

WAXS
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Simplified schematic representation of Ras GTPase cycle</td>
<td>18</td>
</tr>
<tr>
<td>1.2</td>
<td>Ras isoforms in cancer</td>
<td>21</td>
</tr>
<tr>
<td>1.3</td>
<td>G-domain of Ras in its active state, residues 1-166</td>
<td>23</td>
</tr>
<tr>
<td>1.4</td>
<td>Conformational states in active Ras-GTP form</td>
<td>26</td>
</tr>
<tr>
<td>1.5</td>
<td>A simple potential energy function</td>
<td>29</td>
</tr>
<tr>
<td>1.6</td>
<td>Ras isoforms sequence variation around the allosteric lobe</td>
<td>35</td>
</tr>
<tr>
<td>2.1</td>
<td>The Ras and Raf proteins</td>
<td>45</td>
</tr>
<tr>
<td>2.2</td>
<td>The wild type Ras structure in complex with Raf-RBD</td>
<td>54</td>
</tr>
<tr>
<td>2.3</td>
<td>The wild type Ras/Raf-RBD interface</td>
<td>56</td>
</tr>
<tr>
<td>2.4</td>
<td>The RasQ61L structure in complex with Raf-RBD</td>
<td>58</td>
</tr>
<tr>
<td>2.5</td>
<td>Changes in Raf-RBD due to the Q61L mutation in Ras</td>
<td>61</td>
</tr>
<tr>
<td>2.6</td>
<td>Average Cα RMSFs for Ras proteins</td>
<td>63</td>
</tr>
<tr>
<td>2.7</td>
<td>RMSFs for uncomplexed wtRas and uncomplexed RasQ61L</td>
<td>63</td>
</tr>
<tr>
<td>2.8</td>
<td>RMSFs of uncomplexed RasQ61L bound to Raf-RBD</td>
<td>65</td>
</tr>
<tr>
<td>2.9</td>
<td>RMSFs of Raf-RBDs</td>
<td>65</td>
</tr>
<tr>
<td>2.10</td>
<td>Average structures for wtRas/Raf-RBD and RasQ61L/Raf-RBD</td>
<td>68</td>
</tr>
<tr>
<td>2.11</td>
<td>Community networks formed in the Ras/Raf-RBD complexes based on MD simulations</td>
<td>70</td>
</tr>
<tr>
<td>2.12</td>
<td>Optimal and suboptimal paths connecting R97 in the Ras allosteric site to K109 at the end of Raf-RBD loop 4</td>
<td>73</td>
</tr>
<tr>
<td>3.1</td>
<td>Accelerated MD equations</td>
<td>87</td>
</tr>
<tr>
<td>3.2</td>
<td>Average Cα RMSFs for Ras comparing cMD to aMD</td>
<td>92</td>
</tr>
</tbody>
</table>
Figure 3.3: Schematic of calculations used to combine MD simulation data with WAXS

Figure 3.4: aMD simulations produced calculated scattering that fit well to WAXS data

Figure 3.5: Sigma-r plots of Ras

Figure 3.6: Average Cα RMSFs for H-, K-, and N-Ras

Figure 3.7: Salt bridge interactions between switch II and helix 3

Figure 3.8: Highly flexible SII in K-Ras yields two different stable conformations

Figure 3.9: Switch II and helix 3 in tractions in K-Ras and N-Ras

Figure 3.10: High flexibility in switch I of K-Ras enables the transitioning to state 1

Figure 3.11: The stability of salt bridge R123-E143

Figure 3.12: N-Ras shows a unique interaction between residues Y166-K104

Figure 3.13: Conformational sampling of H-Ras aMD

Figure 3.14: Conformational sampling of K-Ras aMD

Figure 3.15: Conformational sampling of N-Ras aMD

Figure 3.16: Community network analysis of K-, N-, and H-Ras

Figure 3.17: H-Ras and K-Ras show differences in binding to calcium acetate

Figure 4.1: Snapshots from H-Ras-GTP showing highly dynamic residue Y71

Figure 4.2: Crystal structure comparison of wild type and Y71A

Figure 4.3: Kinetic assays for intrinsic hydrolysis of Y71A and wild type H-Ras

Figure 4.4: Changes in RMSFs for Ras residues showing differences in Y71A mutant

Figure 4.5: Final structures of Y71A in MD simulations

Figure 4.6: Plots of solution scattering for Y71A mutant

Figure 4.7: WAXS intensities for wild type and Y71A

Figure 5.1: Ras/Raf dimer structure obtained from PDB 4G0N
Figure 5.2: Average RMSFs of Ras residues comparing dimer to monomer

Figure 5.3: Comparing initial and final structures of Ras residues at the dimer interface

Figure 5.4: Dimer interface H-bonds as observed during MD simulations

Figure 5.5: Community networks formed in Ras/Raf-RBD complexes

Figure 5.6: Optimal and suboptimal paths in dimer and monomer

Figure 6.1: Tetraloop simulation

Figure 6.2: LCS1co simulation

Figure 6.3: Duplex siG12D simulation

Figure 6.4: ZIKV simulation

Figure 6.5: LCS1co simulation with sodium and magnesium ions

Figure 6.6: LCS1co simulations with magnesium ions

Figure 6.7: TIP4P simulations using OPLS3 force field

Figure 6.8: LCS1co simulations with different initial ions positions
List of Tables

Table 3.1: Scale factor values for H-, K-, and N-Ras comparing aMD to cMD  
Table 4.1: H-RasY71A crystal structure refinement statistics  
Table 4.2: Conformational clustering for H-Ras and H-RasY71A mutant  
Table 5.1: Relative dimer and monomer population for Ras wild type and mutants  
Table 6.1: RNA structures and systems used for simulations
Chapter 1: Introduction
Ras in cancer

Ras is a small GTPase and the first oncogene product that was discovered more than four decades ago [1]. Ras protein was translated from Ras oncogene as a 21 kDa protein that binds to guanine-containing nucleotides [2]. Fascinating discoveries came after the identification of Ras as an oncoprotein, including groundbreaking works showing that a single point mutation was sufficient to confer transforming oncogenic properties in human cancer cells [3]. In the early 1980’s, many biochemical studies helped to characterize Ras and its mechanism of action as an oncoprotein. Ras was shown to localize to the inner plasma membrane as a required step to activate its functions [4, 5]. Oncogenic mutations were shown to greatly reduce Ras’ enzymatic ability to hydrolyze GTP [6]. By mid 1980’s, Ras was linked to mitogen factors as a downstream signaling protein that was essential for EGF (epidermal growth factor) oncogenes malignant activity [7-9]. Subsequently the discovery of Ras regulator proteins in the late 1980’s, GAP (GTPase activating proteins) and GEF (Guanine nucleotide exchange factors) helped explain the tremendous differences in the enzymatic activities between oncogenic and wild type Ras [10, 11]. From then on, Ras was viewed as a binary switch, activated by binding to a GEF protein that helped exchange GDP to GTP, and deactivated with a GAP protein that helped speed up the hydrolysis of GTP [12]. These findings further revealed the mechanism of action of Ras oncoprotein, where oncogenic mutants were shown to become insensitive to GAP proteins, thus showing 300-fold lower enzymatic activity than wild type in vivo [13]. Subsequent studies in the 1990’s helped identify downstream effectors of Ras signaling. In 1993, Raf kinase was found to interact with the activated, Ras-GTP, and transduced signals to downstream molecules through the Ras/Raf/MEK/ERK pathway [14]. This signaling pathway was shown to be hyper activated in Ras oncogenic mutants. Other effector proteins that also bind to Ras were identified
independently around mid 1990’s, including PI3K (phosphatidylinositol 3-kinase) and Ral-GDS [15, 16]. These works helped Ras researchers to place Ras at the hub of signaling pathways that functions to regulate cell proliferation, differentiation, apoptosis and survival (Figure 1.1) [17].

**Figure 1.1.** Simplified schematic representation of Ras GTPase cycle and signaling pathways. In red is Ras-GDP signaling off form, which is activated by a GEF (guanine nucleotide exchange factor) which promotes the release of GDP, allowing GTP to bind. In green is Ras-GTP in its signaling active conformation, where it can bind several effector proteins to send signal to downstream molecules that are involved in the control of cell proliferation, endocytosis and survival. To help turn off the signal, GTP is hydrolyzed to form GDP, either intrinsically or with the help of GAP (GTPase activating protein).
Oncogenic mutations in Ras are found in about 20% of human cancer [18, 19]. Thus, it has been at the center of cancer research for decades and in 2013 the NCI (National Cancer Institute) started the Ras initiative to specifically target this so called “undruggable” oncogene [20]. There are three primary oncogenic point mutations and they are found at residues G12, G13 and Q61 (COSMIC database). As mentioned earlier, these point mutations abrogate the signaling events through Ras proteins that result in uncontrolled cell growth. Oncogenic mechanism can arise from slower intrinsic hydrolysis rate, insensitivity to GAP catalyzed hydrolysis reaction, and an increase in nucleotide exchange, all of which will extend the activated form of Ras-GTP. To make matters more complicated, there are also three Ras genes that encode for four Ras isoform proteins, including K-Ras4A and K-Ras4B (Kirsten rat sarcoma), N-Ras (neuroblastoma Ras) and H-Ras (Harvey rat sarcoma). K-Ras4A and K-Ras4B are isoforms that arise from alternative RNA splicing, however K-Ras4B is the majority splice variant in cancer tissues [18]. K-Ras4B will be referred to as K-Ras from here on. The frequency of oncogenic mutations from these isoforms in human cancers are strikingly different, with K-Ras having an 85% share, N-Ras with 11% and H-Ras with 4% [21]. The three isoforms share 95% sequence identity in their G-domain (residues 1-166): 100% sequence identity in their effector lobe (residues 1-86) and 90% sequence identity in their allosteric lobe (87-166) [19]. Most of the sequence convergence is found beyond the G-domain at the C-terminal HVR (hypervariable region), with only 15% sequence identity. The hypervariable regions of the Ras isoforms are post translationally modified with lipid attachments that take Ras to the inner leaflet of the membrane, where it encounters its signaling partners [22, 23]. All of the Ras isoforms are farnesylated and methylated at the C-terminus. The striking difference between their HVRs is responsible for the difference in Ras isoforms membrane localization. K-Ras has polybasic sequence of long lysine
residues, which allows K-Ras to be uniquely tethered to acidic lipid enriched membrane regions. Active H-Ras-GTP and inactive N-Ras-GDP localize to non-ordered lipid domain. Inactive H-Ras-GDP and active N-Ras-GTP bind to lipid rafts [24]. In addition to the farnesyl group, N-Ras is modified by a single palmitic acid, whereas H-Ras is palmitoylated at two different sites, and K-Ras has a phosphorylation site at S181 [18].

Three oncogenic mutation hot-spots and three Ras isoform proteins give rise to specific cancer-type preference for Ras in human cancer. The mutation frequency of G12, G13, and Q61 are strikingly different. The predominant mutant found in K-Ras driven cancer is G12 with 83%, follow by G13 with 14% and Q61 with 2%. In contrast, the predominant mutant in N-Ras driven cancer is Q61 with 63%, follow by G12 with 23% and G13 with 12%. H-Ras harbors similar frequency for each mutant, where Q61 has 37%, G12 has 33%, and G13 has 27% [18]. Additionally, specific Ras isoform is preferentially mutated in specific cancers. K-Ras mutants are found 100% of the time in pancreatic ductal adenocarcinoma (PDAC), 86% in colorectal adenocarcinoma (CRC), and 96% in lung adenocarcinoma (LAC) (Figure1.2). N-Ras is frequently mutated in melanoma with 94%, whereas H-Ras shows preference towards head and neck squamous cell carcinoma (HNSCC) with 86% incidence [18]. These difference in cancer mutant location within Ras as well as the different distribution of Ras isoforms in human cancer suggest that oncogenic mechanism of actions is different across the different mutants and different isoforms. This opens a venue for Ras researchers, like us, to investigate the subtle functional differences amongst the isoforms and mutant specific Ras to aid in rational and specific structure-based drug design. There has been over four decades of intense efforts and there is still no FDA approved drug in the market to target Ras oncoprotein. However, there has been plenty of lessons learned and exciting discoveries of the specific roles of Ras in cancer as
well as inhibitors designed over recent years to target Ras, all of which have culminated a new hope to dissociate the term “undruggable” from the Ras oncoproteins.

**Figure 1.2.** Ras isoforms in cancer. Each isoform protein is preferentially coupled to a specific cancer type, K-Ras (blue), N-Ras (orange), and H-Ras (grey). Pancreatic ductal adenocarcinoma (PDAC), colorectal adenocarcinoma (CRC), lung adenocarcinoma (LAC), multiple myeloma (MM), acute myeloid leukemia (AML), head and neck squamous cell carcinoma (HNSCC) [18].

*Ras: Structure and Dynamics*

Full length sequence of Ras consists of 189 residues, where it can be divided into two domains, the G-domain soluble region (residues 1-166) and the C-terminal hypervariable region
(HVR) (residues 167-189). The HVR is a highly dynamic and disordered region of the protein, and it has been shown to impair crystallization efforts for structure determination. Structural biology studies of Ras have thus been largely focused on the G-domain. This domain can be divided into two lobes, the effector lobe and the allosteric lobe. The effector lobe consists of residues 1-86 and the allosteric lobe contains residues 87-166 (Figure 1.3). The effector lobe, as the name suggests, consists of binding sites to effector and regulator proteins as well as the active site that binds to guanine containing nucleotide, GTP and GDP, with picoMolar affinity [25]. There are two switch regions within the effector lobe, switch I (residues 30-40) and switch II (residues 60-76) that change conformations readily during binding and catalysis (Figure 1.3). The switch regions adopt two different conformations depending on which nucleotide is bound at the active site, with an open conformation in Ras-GDP and a closed conformation in Ras-GTP [17]. Additionally, there is a phosphate-binding loop, called P-loop (residues 10-18) at the active site that functions to stabilize the negatively charged phosphate group from the nucleotides. Two of the three prominent oncogenic mutants mentioned earlier reside in this P-loop (residues G12 and G13) and the third mutant resides on switch II (residue Q61). The three Ras isoforms share 100% sequence identity in this effector lobe (residues 1-86) that contains the oncogenic mutation sites, the active site, as well as the effector and regulator protein binding sites. This sequence identity has led many Ras researchers to believe that the three isoforms are equivalent and can be treated as such in studying their structural and biochemical properties. The allosteric lobe consists of residues 87 to 166, and here is where the three isoforms start to diverge in their sequence identity. The isoforms share 90% sequence identity in this lobe. This lobe contains two of the highly-conserved motifs in small GTPases, including NKxD motif (residues 116-119) and ExSAK motif (residues 143-147) that stabilize nucleotide binding and set binding specificity for
guanine containing nucleotides [26]. Studies of Ras membrane orientation has shown that some residues at the allosteric lobe interact directly with the membrane, specifically two arginine residues on helix 4 [27]. Additionally, an allosteric site

Figure 1.3. G-domain of Ras in its active state, residues 1-166. Ras is bound to GppNHp (non-hydrolysable analog of GTP). The effector lobe is colored green with the switch regions in blue, switch I and switch II. The allosteric lobe is shown in grey, where helix 3 and helix 4 are labelled.

was proposed by our lab, that binds calcium acetate at the intersection of helix 3, loop 7 and helix 4, directly affecting the conformation of switch II region on the effector lobe [28, 29]. The acetate molecule was proposed to be a mimic of a negatively charged membrane head group.
Water network analysis previously done in our lab, discovered a novel helix 5 network that links the active site all the way to the back to the nucleotide sensing residues R161 and R164 on helix 5 [29]. This network was specific to the active form of Ras-GTP and was shown to be sensitive to the conformational changes of residue Y32 at the active site. More recent works have been centered on Ras dimerization, where almost all the proposed dimer interfaces reside at the allosteric lobe, mostly on helix 3, helix 4 and helix 5 [30-33]. Taken together, the allosteric lobe harbors important regions of the protein that can directly affect the conformational balance and enzymatic activity at the effector lobe, specifically the active site through allosteric modulation.

One of the reasons Ras is dubbed “undruggable” is that it is a small globular protein with smooth surfaces and lacks hydrophobic pockets for tight drug binding [34]. Fortunately, advances in structural and dynamical studies over the past few years have shed lights into novel conformations of Ras that can potentially be utilized for small molecule or peptide binding. NMR studies have revealed that active GTP-bound Ras can exist in two conformational states, state 1 and state 2 [35]. State 1 is characterized with an open switch I, a direct binding site to effector protein and as a result, Ras loses binding affinity to effector proteins and is unable to propagate signaling events [36, 37]. Whereas, in state 2, switch I is closed over to the nucleotide, positioning Y32 near the gamma phosphate in a conformation that makes Ras poised for effector binding and activation of signaling events (Figure 1.4a) [38]. In state 2, Ras can adopt either an R-state (reactive) or a T-state (tardy) through allosteric modulation. Crystallography studies from our lab discovered a ligand binding allosteric switch that shifts helix 3/loop 7 and results in the ordering of switch II. In the absence of a ligand bound at the allosteric site, switch II of Ras is disordered, whereas in the presence of calcium acetate at the allosteric site, switch II is ordered and a critical catalytic residue, Q61 is placed near the active site, interacting directly with a water
molecule that bridges between switch I, residue Y32, and the gamma-phosphate of GNP, in a conformation that is proposed to be poised for catalysis (Figure 1.4b) [28, 29]. Since it has been shown that the allosteric lobe interacts with membrane components, the acetate molecule in this structure was proposed to mimic a negatively charged membrane head group. This allosteric switch mechanism is proposed for the Ras/Raf/MEK/ERK pathway, since Raf is a unique effector protein that only binds to the switch I region of Ras, leaving switch II free for allosteric modulation [39, 40]. This hypothesis is strengthened by the fact that the GAP-catalyzed hydrolysis reaction for this pathway is unlikely due to its inability to outcompete Raf in binding to Ras. Raf has a binding affinity of 3.5 nM, which is a lot tighter than GAP binding, with 5 µM [10, 41]. Therefore, another mechanism is needed to regulate Ras/Raf/MEK/ERK pathway, which is critical in controlling the regulation of cell proliferation, survival, apoptosis, and has big implications in cancer [42]. The allosteric switch mechanism is a more likely candidate to attenuate Ras signaling through Ras/Raf/MEK/ERK pathway. When Raf is bound to Ras, switch I is in state 2 conformation, while the highly dynamic switch II is free of contact, leaving a catalytic residue, Q61 in a disordered state. Calcium acetate binding at the allosteric site shifts helix 3/loop 7 toward helix 4, orders switch II and places Q61 at the catalytic site (Figure 1.4b).
**Figure 1.4.** Conformational states in active Ras-GTP form. (a) State 1 (blue-4EFL[43]) with an open switch I and state 2 (green-3K8Y[28]) with a closed switch I. State 2 can bind effector proteins and transduce signals, but not state 1. (b) In state 2, switch II can be modulated by a ligand bound at the allosteric site to adopt either an R-state (green-3K8Y) or a T-state (yellow-2RGE[44]). R-state is reactive and conducive for intrinsic hydrolysis, whereas T-state is tardy and non-conducive for hydrolysis.
Molecular Dynamics of Ras

In addition to NMR and X-ray crystallography studies, there have been many computational simulations using molecular mechanics force fields conducted to study the dynamics and thermodynamics behavior of Ras in solution at atomistic level. Molecular dynamics (MD) simulations has been at the forefront of macromolecular simulations that can provide structural information, such as conformational changes and folding at atomistic-level details, that are often unobservable in conventional biophysics methods. In 2011, twelve structurally diverse proteins were successfully folded to their native structures, each to under 1 Å rmsd compared to the experimental structure, using a force field with no biased sampling parameters [45]. It manifested the powerful nature of molecular mechanic’s force field in reproducing experimentally observed molecular geometry and other biophysical properties through a simple physics-based energy function. In 2013, the Nobel prize in Chemistry was awarded to Karplus, Warshel, and Levitt for the development of molecular mechanic’s force fields to study chemical systems, which recognized the invaluable contribution of computational simulations to the advancements in the field of structural biology and biophysics [46].

Additionally, the accuracy of MD simulations to unveil and characterize large conformational changes in proteins has earned its place in drug discovery research practices in many pharmaceuticals pipeline methods [47]. Our understanding of protein being a dynamic machine that can readily change conformations to accommodate binding and to carry out catalytic functions are keys in drug discovery methodologies to tackle challenging problems like drugging Ras, a smooth globular protein [34]. Moreover, free energy perturbation (FEP) computational calculations have been shown to be robust in optimizing small molecule binding conformation to a dynamic protein target. In 2011, using FEP computations Jorgensen and colleagues...
successfully optimized a 5 µM initial hit to a highly potent 55 pM small molecule inhibitor of HIV reverse transcriptase [48].

Molecular Dynamics methods work by describing the motions of a set of interacting atoms over time by solving Newton’s equation of motions [49]. Calculation of the force on each atom in the system yields new atom position throughout a specified time length. The force on atoms can be computed as the derivative of the potential energy with respect to changes in atoms position. The energy term originates from potential energy function that is calculated and approximated semi-empirically and packaged as molecular mechanic’s force field (Figure 1.5) [50]. The force field is an integral part of MD simulations, where it determines the numerical constants and mathematical functions used to describe the potential energy of an MD system. A simple potential energy function shown on equation 1.1. consists of formulas to describe energy changes in bond stretching, angle bending, bond rotation, as well as nonbonding terms including Van der Waals and electrostatics interactions. Terms for bond lengths and bond angles are modeled using harmonic potentials, where deviations from a determined equilibrium value result in energy penalties. The third term, dihedral potential is modeled using a periodic function that changes in energy as the bond rotates. The last two terms are non-bonded, where hydrophobic interactions are calculated using Lennard-Jones potential and electrostatic interactions are calculated using Coulomb’s law [50]. Numerical constants within the potential energy functions are collected from semi-empirical calculations and experimental values, including X-ray crystallography, IR spectra, and high level quantum mechanics calculations [49]. Often, force constant parameters are obtained from small model compounds and transferability to bigger biomolecules is assumed, if simulation environments are similar. A limitation in working with molecular mechanics force field is that bond parameters are defined and therefore cannot be
broken (or new ones be formed), excluding the study of chemical reactions. However, as discussed earlier MD force fields give great performances in describing large structural changes and energies.

Figure 1.5. A simple potential energy function that contains intra- and intermolecular forces in a molecular dynamics system [50]. The first two terms describe bond stretching and angle bending using a harmonic potential. The third term is described using a cosine function for torsion angle that changes in energy as the bond rotates. The last two terms are used for non-bonded interactions, including Van der Waals (described using Lennard Jones potential) and electrostatics (described using Coulomb’s law).

When the first structure of wild type Ras was solved in 1990, differences in active GTP-form and inactive GDP-form were immediately observed on switch I and switch II [26, 51]. However, subsequent structures of oncogenic mutants showed no major structural differences when compared to the wild type structure [26, 52]. It made it difficult to determine how the catalytic mechanism became dysfunctional at atomistic level. Years after that, with the advances in computational simulation, molecular simulation studies have shed lights into our
understanding of the differences in wild type versus oncogenic Ras functions at atomistic level [53]. Early short MD simulations showed slight differences in the flexibility of residues around the P-loop, switch I and switch II when comparing active Ras-GTP against inactive Ras-GDP [54]. More current simulations explained the complete conformational changes that occur during the transition of active Ras-GTP to inactive Ras-GDP, detailing the changes in flexible residues on switch I and switch II [55]. Furthermore, it showed that conformations adopted by mutants G59A and Y32C were accessible by wild type Ras as intermediates conformations during the nucleotide dependent conformational transitioning. This helped explain the experimental kinetic studies of G59A that significantly reduced the rate of nucleotide exchange, which could be trapping RasG59A at an intermediate state and limiting its structural flexibility to be able to carry out normal exchange function [56, 57].

As mentioned earlier, structural studies of Ras have been limited to the G-domain (residues 1-166), however with the help of molecular modeling, studies including full length Ras (residues 1-189) including membrane interactions have been made possible. Gorfe and colleagues have conducted many full-length studies of Ras and helped explain its differential interactions with the membrane through MD simulations [53, 58-61]. In 2007, their computational study characterized the differences in nucleotide dependent full length Ras-membrane orientation, where Ras-GTP makes larger contact with the membrane (including G-domain residues R128 and R135) than Ras-GDP that only includes HVR residues [58]. Their recent simulations showed correlation between specific plasma membrane composition (their interactions with Ras anchor) and signaling outputs [61]. They showed atomistic-level details on the binding specificity between Ras residue side chains and prenyl groups that explained lipid sorting during Ras-membrane binding. These simulations of full-length Ras and its membrane
interactions have unveiled new windows for Ras drug targeting that are now being pursued by many Ras researchers.

Computational simulations of Ras have also extended towards comparing Ras isoforms and their distinct dynamics [57, 62, 63]. In 2008, using H-Ras as a homology model template to construct K- and N-Ras, Gorfe and colleagues ran MD simulations to show that the three isoforms have different level of flexibility, with K-Ras being the most flexible isoform out of the three proteins [57]. Recently, another group ran longer MD simulations at two different temperatures to unveil further dynamical behaviors between the three isoforms in either GDP- or GTP-bound forms. K-Ras was still shown to be the most flexible isoform, specifically in its GTP-bound form with signaling on conformation [62]. This group also showed that active Ras-GTP had unique enhanced correlation between switch II and helix 3 that is not seen in inactive Ras-GDP. These simulation studies contributed partially to our current understanding of dynamical difference seen in the soluble domain of Ras isoforms, that have been long put aside due to the misconception that the HVR anchor domain had been thought to be the only region that gave rise to isoform differences.

The importance of correlation and communication networks that connect allosteric lobe to the effector lobe, and the active site has been critical in Ras drug targeting because the picoMolar affinity of GTP and GDP have made it impossible to design an inhibitor that could compete for the nucleotide binding pocket. MD simulations have contributed a lot of important insights into the allosteric communication between the two lobes [53, 57, 58, 62]. Cross correlation analysis for the simulation of Ras structures have often pointed out dynamical correlations between helix 3/loop 7 regions and the two switches, switch I and switch II. Until recently, a novel data analysis tool, called conditional time-delayed correlation (CTC), revealed
the directionality of the correlations showing that motions in helix 3/loop 7 drive motions in switch II and that motions in switch II drive switch I motions [64]. As discussed in the previous section, crystallography studies from our lab have shown an allosteric switch that originates at helix 3/loop 7/helix 4 that drives the ordering of switch II residues for catalysis [28]. Through MD simulations this allosteric communication can be viewed and analyzed in motions to reveal further communication pathways that might not be observed in static crystal structures.

**Summary and Thesis Outline**

Much has been accomplished in studying Ras from the field of structure biology and biophysics, however there are still many challenging problems that need to be addressed to aid in the structure-based drug design for Ras related cancers. For instance, what are the oncogenic effects on the global structure of Ras? How does oncogenic mutant affect the structure and dynamics of effector proteins? What are the underlying conformational differences between Ras isoforms? How can we target each isoform with better specificity? What are the effects of dimer formation to the structure and dynamics of Ras? Which amino acid residues are important to stabilize different conformational states seen in Ras proteins? These questions are addressed and answered in great details throughout this thesis. Many powerful biochemistry and biophysics tools are used to dissect the questions regarding Ras proteins, including: molecular dynamics simulations, X-ray crystallography, X-ray solution scattering, and kinetics hydrolysis assays.

Computational methods such as MD simulations are powerful and capable of contributing a lot to our understanding of large conformational changes of proteins and other biomolecules. In chapter 2, I conduct many MD simulation runs to study the dynamical behavior of RasQ61L and differentiate it from the wild type Ras. Ras is presented in complex with Raf to
test our existing hypothesis that allosteric modulation is the mechanism through which this signaling pathway can be turned off. My results show long range oncogenic mutation effects that are manifested in RasQ61L, both on the structure of Ras and Raf. It is, to our knowledge, the first results showing global effects of an oncogenic point mutation on the protein structures and allosteric networks.

The three Ras isoforms are preferentially mutated in specific cancers and our lab recently showed that they exhibit different biochemical properties [65]. Sequence variations between these isoforms are quite small, where there are only 14 different residues that cluster at the allosteric lobe (Figure 5.1). These residue differences are located near the important motifs for guanine nucleotide binding, including NKxD and ExSAK motifs [66]. Additionally, these residues neighbor the allosteric site that binds calcium acetate and modulates the conformation of switch II residues near the active site [28]. We hypothesize that the sequence variations on the allosteric lobe are responsible for the differences seen in their biochemical properties. To observe how these residues differentially interact across the three isoforms at atomistic level, I conduct MD simulations in chapter 3. I present results from accelerated MD simulations of K-, H-, and N-Ras differentiating the balance of conformational states and dynamics of these three isoforms. The simulation results are validated using wide-angle solution X-ray scattering (WAXS) experimental results to strengthen analysis that are made from computational simulations. In collaboration with Dr. Lee Makowski and his graduate student Hao Zhou, we convert MD simulation trajectories into X-ray solution scattering intensity and compare the results to experimentally observed intensity of the protein seen in WAXS data. This is a very exciting and novel method that can directly validate computational results using experimental work.
Chapter 4 contains results that I have obtained from my time in the wet lab, where I successfully crystallized an allosteric mutant RasY71A, conduct hydrolysis kinetic assays and run X-ray solution scattering experiments. I show that Ras residue Y71 increases the flexibility of both switch regions, switch I and switch II, as well as increases the intrinsic hydrolysis rate constant. Additionally, I show that there are significant differences that arise from MD simulations that start with a crystal structure as opposed to a homology model, given the specific mutation is of importance to conformational states adopted by the protein.

In chapter 5, I present analysis of Ras dimer simulation results, focusing on the dimer of Ras/Raf complex, with a system that contains four proteins. Much of recent efforts to fully understand the activation mechanism of Ras has been centered on Ras dimer formation [31-33]. Dr. Jillian Parker, from our lab, has amassed a lot of data from the study of Ras dimer, including the structure of the dimer. To complement her work, I conducted MD simulations of the dimer to investigate the effects of dimer formation on the residues at the dimer interface and their relations to allosteric modulation.

The results from these experiments unveil many important characteristic of Ras oncogenic mutants and differences in Ras isoforms which are critical information for structure-based drug design. I believe that these results have significant contributions to aid in better rational drug design to target this “undruggable” oncogenic protein, Ras.

In the appendices section, I present additional work that I conducted during my internship at Vertex pharmaceuticals and general lab protocols from all the projects in the lab. At vertex, I ran simulations on RNA systems to assess MD force fields and inform modelers on best practices that can be applied quickly and easily in a fast-paced drug discovery pipeline in pharmaceutical industry settings.
Figure 1.6. Ras isoforms sequence variation around the allosteric lobe. (a) Tertiary structure of Ras showing residues 1-166 with sequence differences shown in blue, while NKxD and ExSAK motifs are colored orange. The important motifs are surrounded by residue variations amongst the isoforms. (b) Allosteric lobe sequence of Ras isoforms, where residue differences are highlighted in blue.
Reference


Chapter 2: Allosteric effect of the Oncogenic mutant RasQ61L on Raf-RBD
Introduction

Ras proteins are small monomeric GTPases that function as molecular switches in several signal transduction networks that control cell proliferation, differentiation and survival [1]. Its mutants are found in about 20% of human cancers [2]. Signal propagation through Ras is dependent on the bound nucleotide. In the GDP-bound form, the largely disordered active site, composed of the P-loop (residues 10-17), switch I (residues 30-40) and switch II (residues 60-76), samples a range of conformations that excludes effector binding. When bound to GTP, Ras recruits effector proteins, which in turn are activated for specific protein-protein interactions that ultimately lead to a change in cell behavior [1]. This nucleotide cycle is regulated by Guanine nucleotide Exchange Factors (GEF), which promote the loading of GTP, and GTPase Activating Proteins (GAP), which increase the hydrolysis rate of GTP to GDP on Ras by 5 orders of magnitude [3]. The three Ras isoforms, H-Ras, K-Ras and N-Ras, have identical effector binding regions and nucleotide-binding sites located in the N-terminal half of the catalytic domain (effector lobe). They differ primarily at the C-terminus hypervariable region (HVR) and to a much lesser extent in the second half of the catalytic domain (allosteric lobe) distal from the active site [4]. H-Ras has therefore been used as representative for the structure of all three isoforms at the interface with effector and regulator proteins.

The Ras/Raf/MEK/ERK mitogen activated signaling cascade, one of the most well-studied pathways, is involved in the control of cell proliferation and is associated with a variety of aggressive human cancers [5]. C-Raf (also known as Raf-1) is a protein of 648 amino acid residues with the N-terminal half containing two highly conserved regions 1 and 2 (CR1 and CR2), and the C-terminal half containing the conserved kinase domain (CR3) [6]. CR1, CR2 and CR3 are also present in the two other human isoforms A-Raf and B-Raf, although outside of
these regions the three isoforms differ substantially [6]. The CR1 contains two Ras binding
domains: the originally discovered Ras Binding Domain, Raf-RBD (residues 51-131), which
binds Ras with an affinity of 18nM; and the Cysteine Rich Domain, Raf-CRD (residues 139-
184), which binds with a much lower affinity of 20uM [7] (Figure 2.1). The binding of C-Raf-
CRD to Ras is enhanced by farnesylation of Ras and requires the coordination of two zinc ions
by CRD residues [7, 8]. The overall affinity of the Ras/Raf interaction, including the RBD and
CRD, is 3.5nM [9]. Both domains are required for Raf activation by Ras and for Raf membrane
localization in vivo [10], but the mechanism through which the interaction with Ras leads to the
activation of the kinase domain is currently not known.

Despite the importance of the Ras/Raf complex, the best model to date is that of a
complex between Raf-RBD and a Ras homologous protein, Rap1A, in which switch I residues
30 and 31 have been mutated to those found in Ras (this construct was thus dubbed Raps) (PDB
code 1GUA), resulting in identical switch I sequences between Ras and the mutant Raps [11].
Although the Raps/Raf-RBD complex has served as a good overall model of the interaction,
recent work from our laboratory [12, 13] and others [14] has shown that Raf has an important
influence on the structure of Ras and conformational states associated with hydrolysis of GTP to
GDP. These conformational states involve switch I, which is well modeled in Raps, but they also
involve switch II, which has seven residues that are different between Ras and Rap, and therefore
is most likely affected differently by Raf-RBD. More recently, structures of Ras and Rap bound
to a mutated Raf-RBD have also been solved [15]. Raf-RBD residue A85 has been mutated to
Lysine to increase the affinity to the GDP-bound form of the GTPases. However, the A85K
mutation is at the interface with the active site of Ras, again perturbing the view of an intact
complex in this region.
Figure 2.1. The Ras and Raf proteins. (a) Ribbon diagram of the Ras/Raf-RBD complex. The Ras effector lobe is in light gray and the allosteric lobe is in white. Regions of Ras known to interact with Raf-CRD are in black and the Raf-RBD domain is in dark gray. (b) Schematic of the Ras and Raf sequences. Location of the two lobes of Ras and the various domains of Raf are shown. Dashed lines identify regions known to interact in the complex.
We have shown that in crystals with symmetry of the R32 space group, switch II is free of crystal contacts and is allosterically coupled to a site on the allosteric lobe at the interface between helix 3/loop 7/helix 4 of Ras. Upon binding calcium and acetate at that site there is a clear disorder to order transition that places catalytic residue Q61 in the active site, interacting directly with a water molecule that bridges between switch I residue Y32 and the $\gamma$-phosphate of GTP (the bridging water molecule) in a conformation that we propose is poised for intrinsic hydrolysis [12]. Interestingly, this crystal form has switch I stabilized precisely in the conformation observed in the Raps/Raf-RBD complex, including the bridging water molecule. Based on this observation and the fact that it has been shown by NMR that the binding of Raf promotes a conformation competent for intrinsic hydrolysis [14], we hypothesized that Raf plays an important role in the allosteric activation of intrinsic hydrolysis by stabilizing switch I, such that switch II can complete the active site modulated by the allosteric switch [12]. This hypothesis was supported by the oncogenic mutant RasQ61L which crystallized with symmetry of the same space group R32. Unlike the wild type which has a disordered switch II in the absence of bound calcium and acetate, switch II in this mutant is completely ordered to a conformation that interferes with hydrolysis of GTP and that we now call the anticatalytic conformation [12, 16-18]. In this conformation, the bridging water molecule is absent and Y32 makes a direct H-bond with the $\gamma$-phosphate of GTP, while switch II closes over the active site, isolating it from bulk solvent [16]. We showed that the RasQ61L mutant could hydrolyze GTP in the absence of Raf, albeit slower than wild type, but that in the presence of Raf there was no observed hydrolysis [13, 16]. Furthermore, signaling through the Ras/Raf/MEK/ERK pathway, but not the Ras/PI3K/Akt pathway, leads to nearly saturating levels of MEK and ERK phosphorylation in the presence of the RasQ61L mutant, suggesting that a long-lived
RasQ61L/Raf complex plays a role in increased Ras dependent pERK signaling and is likely to contribute significantly to oncogenesis in RasQ61L mutants [13]. Interestingly we could stabilize the anticatalytic conformation in wild type Ras in the presence of the reducing agent DTE (or DTT), which binds to a site between switch II and helix 3, promoting the conformation we originally observed for RasQ61L [17]. Our working model, based on structures of Ras in a crystal form where switch I mimics that in the Raps/Raf complex, is that Raf helps to promote the catalytic conformation for intrinsic hydrolysis of GTP on wild type Ras, but that it stabilizes the anticatalytic conformation of the RasQ61L mutant specifically enhancing its oncogenic potential associated with the Ras/Raf/MEK/ERK pathway [18]. Because switch II is very different in Raps we have had to base our hypothesis on its behavior in the R32 crystals of Ras alone, assuming that these crystals provide a good approximation of the effect of Raf-RBD on Ras and its mutants, given that in this crystal form switch I is stabilized by crystal contacts in the conformation seen in Raps/Raf-RBD. Here we present molecular dynamics (MD) simulations of the X-ray crystal structures of the wild type Ras-GppNHp/Raf-RBD and RasQ61L-GppNHp/Raf-RBD complexes. The structures confirm that the R32 crystal form indeed provides a good first approximation of both complexes at their molecular interfaces. In addition, these structures show that the Q61L mutation affects the allosteric calcium-binding site on Ras and that it has substantial global impacts on the structure of the complex with Raf-RBD. We see clear differences in a Raf-RBD loop (L4) distant from the interface due to the mutation and show, with a series of molecular dynamics simulations, that these differences reflect a long-range influence on the conformational range accessed by L4 in the mutant relative to L4 in the wild type complex. The Q61L mutation increases the conformational dynamics of the switch II region in
Ras in the absence of Raf, quenches the switch II motion in the Ras/Raf-RBD complex and affects the dynamics of the distant L4 loop in Raf-RBD.

Methods

Molecular dynamics simulations
Molecular dynamics simulations (90 ns production time) were performed for wild type Ras and RasQ61L in the absence and presence of Raf-RBD, as well as for Raf-RBD alone, at the high-performance Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov) and at the Northeastern Discovery Cluster (http://www.northeastern.edu/rc). The starting coordinates for these simulations, with the GTP analogue, GppNHp, in each coordinate set changed to GTP by replacing the β-γ-bridging nitrogen atom with oxygen, were as follows: Ras-GTP (PDB code 3K8Y) [12], RasQ61L-GTP (PDB code 3OIU) [13], Ras-GTP/Raf-RBD (PDB code 4G0N), RasQ61L/Raf-RBD (PDB code 4G3X) (the two complexes presented here) and Raf-RBD (PDB code 1RRB, average NMR structure) [19]. All crystallographic water molecules were included in the simulations of the four models solved by X-ray crystallography. The calcium and acetate ions bound in the allosteric site were retained in the coordinates for Ras-GTP, RasQ61L-GTP and Ras-GTP/Raf-RBD. The allosteric site was empty in the RasQ61L-GTP/Raf-RBD complex. A molecule of DTE was deleted from the active site in the Ras/Raf-RBD structure. Residues 64 and 65, which are missing in this structure (PDB code 4G0N), were added to the model in the conformation found in Ras-GppNHp structure (PDB code 3K8Y). All simulations were performed with NAMD software using the CHARMM27 force field [20-22].
The five model systems were subjected to the same MD simulation protocol. Each was solvated in a box of TIP3P water molecules extending 10 Å from the solute atoms [23]. This resulted in simulation boxes for Ras-GTP, RasQ61L-GTP, Ras-GTP/Raf-RBD, and RasQ61L-GTP/Raf-RBD and Raf-RBD with volumes of 219,051 Å³, 212,400 Å³, 315,208 Å³, 314,340 and 138,985 Å³ respectively. Each system was neutralized with sodium and chloride ions to reach a concentration of 0.15M. The simulation systems were minimized for 5000 steps of conjugate gradient minimization and then heated gradually from 50 K to 250 K prior to the production runs. Periodic boundary conditions were applied in the x, y, and z directions during isothermal-isobaric simulations at 300 K and 1 atm. The first 30 ns of the production run were done with a time step of 1 fs, followed by 60 ns with a time step of 2fs, for a total of 90 ns simulation of each of the five systems. The long-range non-bonded interactions were calculated using the switch function between 10 and 11 Å. Particle Mesh Ewald was activated to evaluate the electrostatic interactions [24]. Covalent bonds involving hydrogens were constrained using SHAKE algorithm [25].

Dynamical Network Analysis

Dynamical network analysis is a general method used to obtain an accurate picture of network topology and long-range signaling in protein complexes derived from molecular dynamics simulations [26]. Each amino acid residue in the complex is assigned a node centered on its Cα atom and used as a base to construct significant regions of amino acid interactions and pathways of allosteric modulation that connect them. Edges are placed to connect the nodes between residues that remain within 4.5 Å distance for at least 75% of the simulation time. The edges are weighted using pairwise correlation data calculated by the program Carma [27]. This
information can then be mined to define community networks using the Girvan-Newman algorithm [28] and the diversity of paths that connect sites of functional significance in the complex [26]. Nodes in the same community network can communicate with each other easily through multiple paths, whereas those in distinct community networks either do not communicate well or communicate through one or a small number of nodes essential for allosteric modulation. The community networks in the wild type Ras and RasQ61L complexes with Raf-RBD are shown in Figure 11.2 and discussed in the result section. Optimal and suboptimal pathways of communication between nodes in distinct communities are determined by defining one of the nodes as a source and the other as a sink. These are possible paths through which allosteric modulation can occur [26]. In our Ras/Raf-RBD simulation systems (wild type and mutant) the source was R97 in the allosteric site in Ras and the sink was K109 at the end of the Raf-RBD loop L4. The optimal path is defined as the shortest distance between the source and sink nodes, while suboptimal paths are calculated within a tolerance value equal to the average edge weight in the protein [26]. The number of suboptimal paths between source and sink is a measure of the strength of communication between them, with key residues in the allosteric pathway identified as nodes that are present in multiple paths [26].
Results

Both the structures of wild type and mutant complexes crystallized with symmetry of the space group P321 with one complex per asymmetric unit and were solved to 2.45 Å and 3.3 Å resolution respectively by Dr. Susan Fetics in the Mattos lab. The Raf-CRD binding site on Ras-GppNHp has been mapped by NMR [7] and involves a continuous surface consisting of residues 23-30 (just before switch I), with additional contributions from V45 and neighboring residues, as determined by site directed mutagenesis [29]. However, in both crystal structures of the complexes, we only have Ras/Raf-RBD without the CRD domain.

Wild type Ras bound to Raf-RBD

While the original view of Ras portrayed two Ras states associated with GTP and GDP bound forms, a large body of structural biology work has made it increasingly apparent that Ras-GTP (as seen in structures of Ras-GppNHp) samples distinct conformational substates that have significant implications for signaling: switch I states 1 and 2 observed by NMR [14], and ordering of switch II in the active site associated with R state and T states observed by X-ray crystallography [18, 30]. The R state is characterized by ligand bound in the allosteric site (calcium and acetate in our crystals) and by a highly ordered active site stabilized by water-mediated H-bonding interactions that link the effector and allosteric lobes of Ras through switch I/loop 8/helix 5 and through switch II/helix3/loop7 [30]. This highly ordered R state is exemplified in the structure with PDB code 3K8Y. The switch I has the same conformation as seen in the Raps/Raf-RBD complex and switch II has a conformation stabilized by R68 through direct and water-mediated interactions that propagate to the N-terminal end of the switch where
catalytic residue Q61 is found [12]. As expected, the Ras-GppNHp/Raf-RBD structure shows Raf-RBD interacting at switch I as seen in the Raps/Raf-RBD complex, including the bridging water molecule that we also see in our structure of Ras in the R state (PDB code 3K8Y) [12]. Remarkably, in the crystal form from which we obtained the Ras/Raf-RBD structure the elements of the allosteric switch (switch II, helix 3 and loop 7) are away from crystal contacts as was the case for Ras-GppNHp crystals with symmetry R32 [12, 16]. Furthermore, as with our Ras-GppNHp structure, the Ras-GppNHp/Raf-RBD complex also crystallized under conditions containing 200mM Ca(OAc)$_2$. The resulting structure of the complex has calcium and acetate bound in the allosteric site, promoting the R state conformation of helix3/loop7 and the corresponding structure for switch II that places R68 in position to order the active site residues. However, in this complex there is a molecule of DTE bound between helix 3 and switch II in van der Waal’s contact with Y96, causing a shift in the position of R68 and a reorientation of its side chain to optimize the interaction with the DTE molecule. This perturbs the water-mediated network that stabilizes switch II and residues 64 and 65 are disordered in this structure (Figure 2.2A). Q61 adopts a similar conformation as seen in the R state, but due to the change in the H-bonding network that helps order its side chain, the side chain amide functional group is disordered as determined by lack of electron density for these atoms (Figure 2.2B). Switch II in the R state is normally connected to the main core of the catalytic domain through a series of water molecules that link it to helix 3 residues, stabilizing the entire switch. In our structure of the Ras/Raf-RBD complex the binding of DTE results in the almost complete lack of crystallographic water molecules between switch II and helix 3. Of the network associated with the R state, only two water molecules are present: Wat 312 bridges the backbone carbonyl group of A59 to the backbone amide of G10 in the P-loop and Wat 361 bridges the backbone amide of
A59 to the backbone carbonyl of L36 in switch I. In addition, a new water molecule not normally seen in the R state, Wat 313 in the complex, H-bonds simultaneously to the NE atom of R68 in its new position and the side chain hydroxyl group of Y71. The only other water molecule observed to interact with any of the switch II residues is the nucleophilic water molecule, which as usual H-bonds to the backbone amide of Q61. Interestingly, with the shifted position of R68 and the absence of water molecules between helix 3 and switch II, Y96 relaxes to its position associated with the T state, adding to a T state-like feature of a partially disordered switch II. Thus, we have a hybrid structure, where helix 3/loop 7 allows ordering of switch II to the R state, but where the binding of DTE disrupts the central R68 so that some features of the T state begin to appear toward the N-terminal end of switch II.

The RBD domain consists of the classical ubiquitin fold, with 5 β-strands and two α helices connected by loops (β1L1β2L2α1L3β3L4β4L5α2L6β5) [31]. In its complex with Ras, the backbone H-bonds between β2 of Ras and β2 of Raf-RBD form an extended intermolecular β sheet [32]. This interaction is shared by other Ras effectors such as PI3K and RalGDS [33, 34]. In the Ras/Raf-RBD complex presented here the electron density for Raf-RBD has no main chain breaks for residues 54 – 131 and there is electron density for most side chain residues, with notable exceptions being the side chains of residues in the loop 4 (L4) between β-strands 3 and 4 (E104, H105, K106 and K108). The L4 region is in clear contrast to the rest of the Raf-RBD, which is well ordered in the complex.
Figure 2.2. The wild type Ras structure in the complex with Raf-RBD. (a) Superposition of Ras in the Ras/Raf-RBD complex (green) and our structure of Ras solved in the R32 crystal form with PDB code 3K8Y (pink). Residues involved in the allosteric switch to promote the R state are shown in stick. Calcium and acetate are bound in the allosteric site. In the Ras model with PDB code 3K8Y black dashed lines indicate the hydrogen-bonding network from the allosteric site to the active site, with the water molecules shown as red spheres. (b) Active site of Ras in the Ras/Raf-RBD complex with electron density contoured at the 1.0 σ level.
In addition to the β-sheet interactions that form the interface in the Ras/Raf-RBD complex there are significant interactions between Raf-RBD α1 residues K84, V88 and R89 with switch I residues in Ras as previously determined by NMR, alanine scanning mutagenesis and free energy component analysis [7, 19, 33]. Key interactions based on our structure of the complex are shown in Figure 2.3. K84 in α1 interacts simultaneously with Ras residues E31 and D33, and these two residues are in somewhat different conformations than seen in the structure of Ras alone. Water molecules 206 and 301 connect Raf-RBD K84 to the α-phosphate of GppNHp on Ras. These two water molecules found at the interface in our structure are highly conserved in Ras structures in general, as we determined using our recently published conserved water analysis program DRoP [30]. Raf-RBD residue V88 at the C-terminal end of α1 makes good van der Waals interactions with Ras residues I21 and Y40, which come together from either end of switch I. These three residues form a nice hydrophobic cluster at the interface between the two proteins. The conformation of Ras residue Q25 in the uncomplexed Ras structures overlaps with the position of Raf V88 in the complex and is seen rotated to a difference conformer, where it makes a good H-bonding interaction with the side chain of H27, itself in a new conformation in the complex. Raf residue R89 interacts with three Ras residues: it makes a salt bridge with D38, H-bonds to the backbone carbonyl oxygen atom of S39 and stacks over Y40. The side chain of Ras residue D38 has a different \( \chi_2 \) dihedral angle relative to that seen in the Ras-GppNHp structure, optimizing the salt bridge to Raf-RBD residue R89. These side chain interactions with Ras residues at the beginning of β2 are in addition to the backbone interactions involved in forming the β-sheet with β2 of Raf-RBD.
Figure 2.3. The wild type Ras/Raf-RBD interface. Residues involved in salt bridges, H-bonding and van der Waals interactions across the interface are shown in stick. Water molecules are shown in red spheres and H-bonding interactions are in black dashed lines.

*RasQ61L-GppNHp bound to Raf-RBD*

The structure of RasQ61L in the RasQ61L/Raf-RBD complex has helix3/loop7 in the R state, with a fully ordered switch II in a conformation that superimposes well on the conformation that we obtained for the RasQ61L-GppNHp structure from crystals with symmetry R32 in the presence of calcium acetate (PDB code 3OIU) [13]. In that structure, there is clear electron density for calcium and acetate in the allosteric site as observed for the wild type. R68 is in its position to facilitate the water-mediated H-bonding interactions that leads to the active site.
and almost all the associated water molecules are present between helix 3 and switch II. The active site is perturbed, however, by the fact that in the presence of L61 there is no room for the bridging water molecule and there is a direct H-bonding interaction between Y32 in switch I and the γ-phosphate of GppNHp. This direct H-bond is a feature of the T state and is accompanied by the shift in Y96 also associated with the T state. Perturbation of the water network that results from this shift is seen in this structure as previously described [13]. In the RasQ61L/Raf-RBD complex there is continuous electron density for the entire switch II and the conformation of its N-terminus as well as the rest of the active site superimpose well on the structure of RasQ61L alone (Figure 2.4A). The L61 in the active site precludes binding of the bridging water molecule, resulting in a direct H-bond between Y32 and the γ-phosphate of GppNHp (Figure 2.4B). The structure of RasQ61L/Raf-RBD is solved at 3.3 Å and at this low resolution there are only very few resolved crystallographic water molecules. The nucleophilic water molecule, which is present in virtually all high-resolution structures of Ras is not observed and the cleft between helix 3 and switch II is also devoid of crystallographic water molecules. Consistently, there is no calcium or acetate seen in the allosteric site, although the R state implies its effect on the structure.
Figure 2.4. The RasQ61L structure in the complex with Raf-RBD. (a) Active site in the mutant complex (cyan) superimposed on the structure of the RasQ61L mutant solved from crystals with R32 symmetry with PDB code 3OIU (yellow). The water molecule shown in yellow sphere is from the structure with PDB code 3OIU. (b) Active site of RasQ61L in the RasQ61L/Raf-RBD complex. The electron density is contoured at the 1.0 σ level.
Despite the lower resolution of this structure, continuous electron density is seen for most of the Ras and Raf-RBD molecules, including side chains. There are notable exceptions, however, that given the rest of the structure, are of significance. There is no electron density for the side chain of K84. D33, with which it interacts in the wild type structure, is in the position seen in our structure of the uncomplexed wild type Ras (PDB code 3K8Y). In the wild type complex the position of K84 would clash with the D33 rotamer observed here and in the Ras structure with PDB code 3K8Y (Figure 2.5A). The fact that there is clear electron density for D33 even at low resolution and that its position overlaps with that seen in the uncomplexed wild type Ras suggests that the RasQ61L mutant has a weaker interaction between Ras switch I D33 and Raf-RBD K84.

With a weakened K84 interaction at switch I, there is a small shift in the N-terminal portion of Raf-RBD helix α1 residues 77-84 in the mutant relative to the wild type complex. These residues directly contact the β-sheet that connect the binding interface to Raf-RBD loop L4 containing residues 103-108 in Raf-RBD. Surprisingly, in the RasQ61L structure there is clear side chain electron density for residues E104, H105 and K108 in Raf-RBD L4. These side chains are disordered in the wild type complex. Conversely, K109, which is ordered in the wild type complex where it interacts with Q127 and D129, is disordered in the complex with RasQ61L, as indicated by the absence of electron density for this side chain. Consistently, in the mutant complex Q127 is flipped away from K109, making a good H-bond to the backbone carbonyl group of F61, and the side chain of D129 is disordered (Figure 2.5B). The interactions involving Q127 connect back to residue R89 at the Ras interface. R89, on Raf-RBD α1, interacts differently in the mutant complex with D38 at the beginning of Ras β2 leading from switch I, possibly influencing a second pathway of communication between switch I and
Raf-RBD loop L4 (residues 103-108). This pathway includes R89, R67, F61 and Q127 and was previously determined based on free energy calculations in the modeled complex between Ras and Raf-RBD, looking at pairwise interactions and pathways of energetic coupling [33]. The same group also determined that in the wild type Ras/Raf-RBD complex there is very high flexibility in the Raf-RBD L4 loop [35]. In our structure of the RasQ61L/Raf-RBD, Ras residue S39 and Raf-RBD residues R89, R67 and Q127 have different conformations than seen in our structure of the wild type Ras/Raf-RBD complex (Figure 2.5B). These residues lead to interactions with K109 as described above and with L101, two residues at either end of L4. There are clear differences in the ways in which wild type Ras and RasQ61L interact with Raf-RBD that propagate to L4, affecting the loop’s flexibility in the complex. The Q61L mutation in Ras has an impact on the way in which residues K84 and R89 interact at the molecular interface and these local changes affect the connectivity to Raf-RBD L4 residues at the opposite end of the molecule, resulting in diminished flexibility of the loop.
Figure 2.5. Changes in Raf-RBD due to the Q61L mutation in Ras. In the wild type, complex Ras is in green and Raf-RBD in magenta; in the mutant complex Ras is in cyan and Raf-RBD in orange. (a) K84 at the Ras/Raf-RBD interface. In the mutant complex, there is no electron density for K84 beyond the Cβ atom and the side chain has been truncated accordingly. The water molecules shown in red spheres and the hydrogen bonds indicated by black dashed lines belong the wild type Ras/Raf-RBD complex. (b) Changes that propagate from the interface to Raf-RBD K109 and loop L4 containing E104. There is no electron density beyond the Cβ atoms for residues K109 and D129 in the mutant complex and these side chains have been truncated accordingly. Dashed black lines represent H-bonding interactions.
Molecular dynamics simulations show altered flexibility due to the Q61L mutation

To understand the effects of the Q61L mutation in uncomplexed Ras and in the Raf-RBD-bound state, we carried out Molecular Dynamics (MD) simulations in explicit solvent for each of the molecules alone and in complex (details in the Materials and Methods section). Figures 2.6, 2.7 and 2.8 show pairwise comparisons of the α-carbon root mean square fluctuations (RMSFs) for Ras residues, highlighting the effects of Raf-RBD binding or of the mutation on the Ras catalytic domain. Figure 2.9 shows the effects that the wild type and mutant Ras have on the structure of Raf-RBD. A comparison of the RMSFs for wild type Ras alone and in complex with Raf-RBD reveals that there is a significant decrease in switch I flexibility in the complex (Figure 2.6). This difference in flexibility is particularly large for residues 30 – 34, the direct site of binding Raf-RBD, including Y32 which has been observed to be highly mobile in solution, giving rise to switch I states 1 and state 2 observed by NMR [14]. In contrast, switch II becomes significantly more flexible upon binding to Raf-RBD. Residues 60 – 67 become highly dynamic, with RMSF for residue 63 going from 2.5 Å in the uncomplexed protein to nearly 9 Å in the complex with Raf-RBD. There is a slight increase in flexibility at the C-terminus of helix 3 and in loop 7, consistent with this region being adjacent to the more flexible switch II. Most intriguingly, the allosteric site residues R97 and Y137 in helices 3 and 4 that coordinate Ca$^{2+}$ become somewhat more rigid in the complex. The calcium-binding site has been shown by NMR to be highly mobile and have no specificity for calcium over magnesium in solution [36], whereas magnesium does not bind Ras in our crystals with R32 symmetry [12]. It appears that Ras-RBD binding (mimicked in our crystals) helps order the ion binding pocket and this could increase the calcium specificity in the allosteric site, poising Ras for intrinsic hydrolysis upon
Figure 2.6. Average Cα root-mean-square-fluctuations (RMSFs) for Ras proteins. Protein secondary structures are shown along the x-axis of the plots. Protein structures in the inserts are shown in gray with increases in fluctuations shown in blue and decreases shown in red. Uncomplexed wild type Ras (wtRas) and wtRas bound to Raf-RBD. The Ras structure shows the effect of binding Raf-RBD.

Figure 2.7. RMSFs for Ras proteins comparing uncomplexed wtRas and uncomplexed RasQ61L. Ras structure shows the effect of the Q61L mutation on uncomplexed Ras.
calcium signaling as we previously proposed [12]. The increased flexibility of wild type Ras switch II in the complex with Raf-RBD has been reported previously [35]. However, the effects of the Q61L mutation on the structure of Ras and its complex with Raf-RBD reveal new and important insights. When comparing the RMSFs from the wild type Ras simulations with those of RasQ61L it appears that the mutation has a modest effect on the dynamics of the uncomplexed Ras protein (Figure 2.7). The switch I residues have nearly identical RMSFs in both wild type and RasQ61L simulations. Switch II is more flexible in the RasQ61L mutant than in the wild type Ras, with residue 63 having an RMSF of 2.5 Å in the wild type and 5 Å in the mutant, but the effect of the mutation is less than observed for the wild type upon binding Raf-RBD. There is also a small increase in RMSFs for residues 27 and 38, just before and after switch I, in the mutant relative to the wild type Ras. In contrast to binding of Raf-RBD, the Q61L mutation increases the flexibility of residues R97 and Y137 in the allosteric site. This would make it more difficult for calcium to bind and activate hydrolysis, contributing to a long-lived Ras/Raf complex and constitutive activation of the Ras/Raf/MEK/ERK pathway.

The binding of Raf-RBD results in much larger and global effects due to the mutation (Figure 2.8). As is the case for the wild type complex, switch I becomes more rigid in the mutant complex and to a similar extent as in the wild type protein due to direct binding of Raf-RBD in this region. Switch II, on the other hand becomes significantly less flexible upon complex formation and this is associated with a conformation change involving residues 61-65 at the beginning of the switch, towards the conformation seen in the T state structure with PDB code 2RGD [16]. This region forms an ordered $3_{10}$-helix in the mutant complex, with a hydrogen bond between the main chain carbonyl of E62 and the amide of S65, which is absent in wild
Figure 2.8. RMSFs of uncomplexed RasQ61L and RasQ61L bound to Raf-RBD. The Ras structure shows the effect of binding Raf-RBD to the RasQ61L mutant.

Figure 2.9. RMSFs of uncomplexed Raf-RBD, Raf-RBD bound to wtRas and Raf-RBD bound to RasQ61L. Raf-RBD structure shows the effect of RasQ61L relative or wtRas on the complexed Raf-RBD protein.
type Ras/Raf-RBD. Although not identical to the switch II structure seen in the anticatalytic conformation, this may represent a structure where switch II has moved toward the T state within the context of a helix 3/loop 7 conformation representatives of the R state. In contrast to the rigidifying of switch II, the entire β–sheet core of the RasQ61L mutant, as well as additional areas in helix 3 and 4 become somewhat more flexible upon complex formation relative to RasQ61L alone. Most surprisingly, the P-loop and the N-terminal end of helix 1 become significantly more dynamic, with an RMSF of about 2 Å compared to that of 0.5 Å in the uncomplexed RasQ61L. It is clear, that the effects of the Q61L mutation, far from being local, substantially alter the dynamics of the mutant, particularly in the context of the complex with Raf-RBD.

The simulation containing Raf-RBD alone, starting from the average NMR structure [31], shows a highly flexible molecule with residue RMSFs much higher than in the complex with either wild type Ras or RasQ61L (Figure 2.9). The overall average RMSF for Raf-RBD alone is 2.6 Å, whereas in complex with wild type Ras and RasQ61L they are 1.5 Å and 1.6 Å respectively. Complex formation rigidifies all regions of Raf-RBD, with large reductions in flexibility seen at the β2 strand, loop L2, α1 and L4. The differences seen in the Raf-RBD residues RMSFs when comparing the complexes with wild type and mutant Ras are consistent with the relative ordering of the residues as described above for the crystal structures, based on the electron density maps. The side chain of Raf-RBD residue K84, which is clearly seen in the crystal structure of the wild type complex but not in the mutant, has increased RMSF in the mutant complex, as do residues 77-86 in α1 that we observed to be slightly shifted in the mutant. This is consistent with a more dynamic interaction at switch I, which in turn could explain the higher RMSF for the P-loop in the RasQ61L/Raf-RBD complex. Loop L4 residues 103 – 108
belong to another area of the Raf-RBD structure for which we observed more order based on electron density in the complex with RasQ61L relative to wild type Ras. Consistently the RMSF for this region in the structure is significantly lower in the mutant complex (Figure 2.9). This trend is reversed at residue K109, which is disordered in the mutant complex but well-ordered in the wild type. R89, R67, Q127 and D129, in \( \beta \)-strands 2, 3 and 4 comprising the pathway that links the interface with Ras to loop L4 also have higher RMSFs in the RasQ61L/Raf-RBD complex relative to the complex with wild type Ras. F61 in \( \beta \)1 is slightly more rigid in the mutant complex, perhaps due to its backbone interaction with the side chain of Q127 (Figure 2.5b). It is clear both from our crystal structures and from the MD simulations presented here that the previously identified pathway of communication between Ras and L4 in Raf is affected significantly by the Q61L mutation in Ras. Residues in this pathway are consistently more flexible in Raf-RBD when bound to RasQ61L, where the wild type interactions of Q127 and D129 with K109 are not present and loop L4 can achieve a more stable conformation.

To gain further insight into possible factors that result in a more rigid L4 in the complex with RasQ61L relative to the wild type complex, we analyzed the conformations of this loop throughout the 90 ns simulations for the two complexes. Although in crystals from which both structures were solved L4 residues 104 – 109 are away from crystal contacts, R100, located before the beginning of L4, is in contact with Raf-RBD residues S120 and I122 of a symmetry-related molecule and this is expected to restrict motion associated with L4 in crystals of both the wild type and mutant complexes. This constraint is not present in the simulations, allowing access to a greater range of conformational states. Even so, Raf-RBD L4 in the wild type complex fluctuates around the position seen in the crystals, resulting in an average MD simulation structure with L4 in a position close to the starting structure. In the complex with
Figure 2.10. Average MD simulation structures for wtRas/Raf-RBD and RasQ61L/Raf-RBD. H-bonding/salt bridge network from RasQ61L residue E37 to Raf-RBD E104 in L4. E104 does not interact with this network in the wild type complex resulting in the a more flexible loop 4.

RasQ61L, however, E104 flips to make a salt bridge to R100 (Figure 2.10). This salt bridge is present 69.2% of the 90-ns simulation time and is a key feature of the average MD simulation structure. R100 also forms a salt bridge with E124 and through this interaction it is linked to a series of H-bonds leading to Ras switch I residue E37 at the binding interface. While the connection to L4 is only seen in the mutant complex (it is not observed at all during the simulation of the wild type complex) the pathway connecting the Ras interface to R100 is found in both structures (Figure 2.10). It propagates from E37 on Ras through a salt bridge to R59 on
β1 of Raf-RBD, whose backbone amide H-bonds to the backbone carbonyl group of E124 on β5. This interaction network through the central β-sheet connects two salt bridges at either end of this pathway, which terminates with R100 in an accessible position for docking of the L4 residue E104. In summary, differences in the RasQ61L interface with Raf-RBD lead to a disordered K109 in the mutant and this appears to be correlated with flipping of L4 so that E104 can interact with R100. This salt bridge is a key feature contributing to the rigidity of L4 in the mutant.

Dynamic networks of communication between the Ras allosteric site and Raf-RBD L4

The differences in connectivity between residues in the wild type and mutant complexes can be visualized using dynamic network analysis of the MD simulation trajectories [26]. In this analysis, each residue is assigned a node centered on its Cα atom and edges are used to connect nodes with associated residues that interact at least 75% of the time (within 4.5 Å) throughout the 90ns simulation time. Within the global dynamical network in our complexes there are sub-networks, called community networks, made of nodes that are tightly interconnected within a group and more loosely connected to other groups (Figure 2.11). There are 10 community networks in the wild type complex (Figure 2.11A), with the allosteric and active sites residing at opposite extremes of a single community network that connects them, and an interface that leads to Raf-RBD L4 through two distinct communities, divided at Raf-RBD F61, leading to K109. There are only 9 community networks in the mutant. In contrast to the wild type, the allosteric and active sites are in two separate community networks, while the connection from the complex interface to the Raf-RBD L4 is through a single community network that includes the path from E37 at the Ras complex interface to R100 near L4 (Figure 2.11B).
Figure 2.11. Community networks formed in the Ras/Raf-RBD complexes based on MD simulations. Each community has its own color, superimposed on the respective average MD simulation structure. (A) Wild type Ras/Raf-RBD. The allosteric site and active site residue Q61 are in the network community shown in orange. Two communities, gray and pink, separate the interface from L4. (B) RasQ61L/Raf-RBD. The allosteric site and Q61 are in two separate communities, orange and magenta, respectively. The interface is linked to L4 through the network community shown in gray.
Dynamic network analysis of the simulation trajectories allows visualization of the Ras/Raf-RBD complexes in terms of closely interacting groups of residues (community networks) and the identification of communication pathways between two functionally important parts of the complex that can be affected through allosteric modulation [26]. The community network analyses for Ras/Raf-RBD and RasQ61L/Raf-RBD (Figure 2.11) show that the connection between the Ras active and allosteric sites present in the wild type complex is severed in the mutant. In the wild type complex, residue F61 is at the interface between two community networks in Raf-RBD linking the interface and loop L4, and it is therefore a likely key residue in allosteric communication between regions of the complex (see Materials and Methods). In the mutant complex, there is a single community network between the interface and L4 of Raf-RBD, with a direct link from E37 on Ras to R100 on Raf-RBD, with which L4 residue E104 interacts substantially during the simulations. The shortest path (optimal path) between the R97 in the Ras allosteric site and K109 at the end of Raf-RBD loop L4 was calculated and serves as the main path between the source (R97) and the sink (K109) nodes (Figure 2.12). The alternate paths (suboptimal paths) were also determined for both complexes. There are 39 and 6 suboptimal paths in the wild type and mutant complexes respectively, indicative of a high connectivity between the two sites in the wild type that is not present in the mutant. The optimal path in the wild type protein involves helix 3 and switch II on Ras, as well F61 and interestingly R100 on Raf-RBD, while suboptimal paths also include the Ras β-sheet core. Note that the optimal and several suboptimal paths are directed through F61 in the wild type complex, consistent with the idea that this residue is key in the allosteric communication between the Ras allosteric site and Raf-RBD loop L4 (see Materials and Methods). The presence of several suboptimal paths indicates degeneracy of communication, with good connection between these
two sites in the wild type complex. In contrast, there is only a single path leading from the Ras allostERIC site to the interface with Raf-RBD, which goes from helix 3 through the β-sheet core and does not include switch II. Of note, it also does not include R100, which in the mutant is primarily engaged in a salt bridge with E104 and participates in the link to K109 only through one of the suboptimal paths. The number of suboptimal paths leading from the complex interface to L4 is greatly reduced in the mutant relative to the wild type complex. Overall, communication between the allostERIC site in Ras and L4 in Raf-RBD is weak in the mutant complex, indicative of severed allostERIC network between the two sites.
Figure 2.12. Optimal and suboptimal paths connecting R97 in the Ras allosteric site to K109 at the end of Raf-RBD loop L4. (A) The optimal path (red) and several suboptimal paths (blue) found for the wild type Ras/Raf-RBD complex. There are 39 suboptimal paths in this complex. (B) The optimal path (red) and suboptimal paths (blue) found for the RasQ61L/Raf-RBD complex. Only 6 suboptimal paths are found in the mutant complex. The backbone trace of the Ras and Raf-RBD corresponding to the average MD simulation structures of the complexes are shown in cyan. Residue nodes in the paths are indicated by spheres. The thickness of each edge is proportional to the number of suboptimal paths that cross it during the calculation. Note the thin edges in the mutant complex.
Discussion

Over the past several years we have accumulated structural evidence suggesting that binding of Raf-RBD would be expected to order the switch II of RasQ61L to an anticatalytic conformation that would keep Ras in its GTP-bound state [12, 13, 16, 17]. Our insight is based on structures of uncomplexed Ras and its mutants solved from a crystal form in which switch I is in the same conformation as seen in the Raps/Raf-RBD complex, the only experimental model available for nearly 20 years to represent the Ras/Raf-RBD interaction [32]. The Ras/Raf-RBD complexes presented here not only support our insights based on the structures of Ras crystallized with R32 symmetry, but also provide information on how the Ras mutation changes the local dynamics at the RasQ61L/Raf-RBD interface, which propagate across the Ras and Raf-RBD structures to areas likely to directly impact activity.

Our analysis of the crystal structures of the wild type and mutant complexes, coupled with results of MD simulations, clearly links the Q61L mutation to residues L101 and K109, at either end of Raf-RBD loop L4. The connection is through a pathway of communication that had been previously identified for the wild type Ras/Raf-RBD complex from simulations based on a homology model starting with the Raps/Raf-RBD structure [33, 35]. The MD simulations presented here show that while the L61 residue has the effect of significantly increasing the flexibility of switch II, it has the opposite effect in the complex with Raf-RBD, where switch II becomes more rigid. This is, despite of the fact that binding of Raf-RBD to the wild type protein dramatically increases the flexibility of switch II. Stabilization of switch I in the mutant complex positions Y32 such that it can participate in a hydrophobic cluster involving switch II residues, with L61 at its core [13, 16]. In contrast to switch I and switch II, the P-loop becomes more
flexible relative to the wild type protein (uncomplexed or in complex with Raf-RBD) or to uncomplexed RasQ61L. These changes particularly affect the interactions of Raf-RBD residues R67 and K84 at the interface with Ras, both of which have higher RMSFs in the mutant complex, with repercussions throughout the Raf-RBD structure ultimately resulting in disordered L101 and K109 in RasQ61L/Raf-RBD (Figure 2.5). The fact that K109 in the mutant complex has more flexible interactions with Q127 and D129 than in the wild type, coupled to increased L101 flexibility, may allow a predominant conformation of L4 where residue E104 forms a salt bridge with R100, not seen in the complex with wild type Ras (Figure 2.10). The long-range effect of the Q61L mutation is also clearly seen in our community network analysis, where the mutation results in a disconnection between residues in the allosteric and active sites, and where the connection between the interface and Raf-RBD L4 becomes directed through a single community network, which includes the R100 residue that interacts with E104. The discovery that the Q61L mutation has a global effect is highly significant and novel, as the current assumption is that the effects of oncogenic mutations are local, with the common active site mutants at G12, G13 and Q61 affecting intrinsic hydrolysis rates and sensitivities to GAPs, thus prolonging the duration of the GTP-bound form of Ras.

It has become increasingly apparent that Ras oncogenic mutants occur at very different rates in different cancers and that they affect distinct pathways in the cell [2]. The G12V and Q61L mutants, for instance, behave differently in terms of their interactions with GAPs [37] and we have shown that they have different switch II structures and respond differently to allosteric modulation associated with intrinsic hydrolysis [13, 16]. Of note, they respond differently to Raf both in terms of their hydrolysis rates and in terms of MEK and ERK phosphorylation [13]. RasQ61L cannot hydrolyze GTP in the presence of Raf-RBD and it aggressively activates the
Ras/Raf/MEK/ERK pathway [13, 16]. This correlates well with the prevalence of the Ras Q61 mutations in melanomas, which can also be driven by V600 mutations in Raf [38]. Given the unique relationship between Ras mutated at residue 61 and Raf at the structural, biochemical and cell biology levels, it is imperative that we make progress in understanding the mechanisms through which the Q61 mutants constitutively activate Raf kinase. Yet, to date, there is no definitive information on the molecular mechanism through which activation of the Raf kinase domain is affected by the Ras-GTP/Raf interaction, even in the case of the wild type protein. We focused here on progressing beyond the current understanding of local effects due to the Q61L mutation and showed that it has a global impact on structure, particularly in the complex with Raf-RBD. In addition to stabilizing switch II, L61 promotes greater flexibility in the Ras allosteric calcium-binding site and in the pathway of communication linking the Ras/Raf-RBD interface to the Raf-RBD loop L4, which in turn has been proposed as a key element in activating the kinase domain [33]. Elucidating the ways in which structural pathways of communication in protein complexes within signaling networks are affected by oncogenic mutants is key in promoting novel approaches to target Ras in human cancers [39]. The present work represents a shift in paradigm in the way we regard the effects of oncogenic Ras mutations and opens new exciting venues for future research in this area.
Reference


Chapter 3: Structure, Dynamics and Conformational States of K-, H-, and N-Ras Assessed by a Combination of MD Simulations and Solution X-ray Scattering
**Introduction**

Ras is a small GTPase that acts as a molecular switch [1]. Inactive Ras binds to GDP and can be activated by replacing GDP with GTP, with the help of GEFs (guanine nucleotide exchange factors). Active GTP-bound Ras binds to several effector proteins, including Raf and PI3K, to send signals to downstream biomolecules that are involved in the control of cell proliferation, survival, and apoptosis [2]. GAPs (GTPase activating proteins) help to inactivate Ras by catalyzing the hydrolysis of GTP to form GDP. The G-domain of Ras that consists of residues 1-166, can be divided into two lobes, the effector lobe, residues 1-86, and the allosteric lobe, residues 87-166. The effector lobe binds to effector and regulator proteins. It contains switch I (residues 30-40) and switch II (residues 60-76) that change conformations readily during binding and catalysis [1, 3]. In its active GTP-bound form, Ras can exist in two states, state 1 with an open switch I and state 2 with a closed switch I [4-6]. State 2 is the activated form that can bind to effector proteins. Previous work in the Mattos Lab has shown that in the state 2 conformation Ras can exist in either a T-state (tardy) or an R-state (reactive) [7]. In the T-state, switch II of Ras is disordered, and in the R-state switch II becomes ordered through allosteric modulation by calcium binding at a remote site. In our crystals, when calcium and acetate bind at the allosteric site, helix 3 and loop 7 shift toward helix 4, resulting in the ordering of switch II and the positioning of Q61 near the catalytic center, interacting directly with a water molecule that bridges between switch I residue Y32, and the γ-phosphate of GTP, in a conformation that is conducive for intrinsic hydrolysis [7]. In summary, Ras in its GTP-bound form is highly dynamic. Among its many conformational states are the open conformation (state 1) or closed conformation (state 2) of switch I, while switch II can be found in a disordered state (T state) or
an ordered conformation (R state). In order to hydrolyze GTP and stop the signaling events, Ras needs have its switch I in state 2 and switch II in the R state.

There are four main Ras isoforms, including H-Ras, N-Ras, K-Ras4B, and K-Ras4A. In our study, we focus on the three most prominent isoforms, H-Ras, N-Ras, and K-Ras4B (here on referred to as K-Ras). Mutations from all three isoforms are found in over 20% of all human cancers [8]. The G-domain of Ras isoforms share 100% sequence identity in their effector lobe, and 90% sequence identity in the allosteric lobe [9]. K-, N-, and H-Ras, each localizes to different plasma membrane microdomains [10]. Active H-Ras and inactive N-Ras localize to the non-ordered domain of the plasma membrane. Inactive H-Ras and active N-Ras bind to lipid rafts. K-Ras, on the other hand, uniquely localizes to acidic lipid enriched membrane microdomains [10]. Additionally, each isoform has been shown to be preferentially mutated in specific cancers, where K-Ras is the most predominant isoform in human cancers, mutated in pancreatic, colorectal and lung cancer. Melanoma has a high prevalence of N-Ras mutants. H-Ras mutants are seen mostly in bladder and head and neck squamous cell carcinoma [8]. Our recent biochemical studies have shown that K-, N-, and H-Ras have different intrinsic hydrolysis rate constants in the presence and absence of Raf effector protein [11]. H-Ras has the highest rate constant, about three times those of K- and N-Ras in the absence of Raf. In the presence of Raf, K-Ras, but not N-Ras, experiences a doubling of its GTP hydrolysis rate constant [11]. In this project, we hypothesize that given 100% sequence identity in the effector lobe between the three Ras isoforms, the differences in the hydrolysis rate constants are due to changes in the balance of conformational states, with different access to the catalytic conformation in each of the isoforms. The conformational ensembles of Ras-GTP, including state 1/state 2 and T/R states have been studied extensively only on H-Ras, but not the other two isoforms, K- and N-Ras [4, 7]. In this
chapter, we characterize the balance of conformational states in K-, N-, and H-Ras using accelerated molecular dynamics (aMD) simulations paired with Wide-Angle X-ray solution scattering (WAXS). There have been many molecular dynamics (MD) studies of Ras conducted over the years [12-22]. Gorfe and colleagues reported in 2008 the isoform dependent dynamical features of Ras proteins, where they showed that K-Ras was more dynamic than N- and H-Ras [15]. In 2015, Kapoor et al. presented the differential dynamics of Ras isoforms using MD simulations and further confirmed the higher degree of flexibility in K-Ras as compared to the other two isoforms [13]. Last year, Harrison and colleagues correlated experimental HDX studies with MD simulations and showed correlated motions between the P-loop and switch II in the Ras isoforms [17]. In this study, we went further and conducted accelerated MD to characterize the difference in the balance of conformational states that are seen in experimental studies of Ras isoforms [11]. We show that all three isoforms can transition between R- and T- states in our simulations, and that they prefer the T-state in the absence of ligand binding in the allosteric site. K- and N-Ras also transition from state-2 to state-1 during our simulation. We use wide angle X-ray solution scattering (WAXS) experiments to confirm that the aMD simulations are sampling a range of conformations consistent with that found in solution, which is not the case when classical MD simulations are used. We also show that the simulations more closely reflect the experimental results when starting with the crystal structure of the specific isoform being studied as opposed to mutating residues in the crystal structure of a different isoform.
Materials and Methods

Molecular Dynamics

Molecular Dynamics (MD) simulations were performed for 200 ns production run for each of the isoforms, K-, H-, and N-Ras, at the Northeastern Discovery Cluster (http://www.northeastern.edu/rc). The starting structures were obtained from X-ray crystallography structures in their active conformations bound to GTP analog, GppNHp, with PDB accession codes 3K8Y (H-Ras), 3GFT (K-RasQ61H), and 5UHV (N-Ras) [7, 11]. GppNHp molecules were converted GTP by replacing the $\beta - \gamma$ bridging nitrogen with oxygen. Residue H61 in K-Ras was mutated back to wild type Q61. Crystallographic waters were retained for the simulation. The NAMD and VMD software packages were used to prepare, perform and analyze all the simulations and the CHARMM27 force field was used in all cases [23, 24]. Three systems were made and solvated in TIP3P water molecules extending 10 Å from the protein molecule. Charges in the systems were neutralized by adding sodium and chloride ions. The simulations started with energy minimization for 5,000 cycles, followed by gradual heating from 50 K to 250 K prior to the production runs. Periodic boundary conditions were used at all three x, y, and z directions. Time step of 1 fs was applied during the production run at isothermal-isobaric condition of 300 K and 1 atm for the 30 ns. Time step was increased to 2 fs for the subsequent production run for 170 ns. The switch function was used between 10 and 11 Å to calculate long range non-bonded interactions. Electrostatic interactions were evaluated using Particle Mesh Ewald parameters [25]. Covalent bonds involving hydrogen bonds were constrained using the SHAKE algorithm [26].
Accelerated Molecular Dynamics

Accelerated MD (aMD) is a variation of conventional MD (cMD) where a bias potential is added to increase conformational sampling by enhancing the escape rates from potential wells. The form of the modified potential is shown on equation 1.4, where boost potential is added to the true potential when the overall potential energy is below a user calculated value. Our study followed the method described previously in the McCammon’s lab, where we applied dual boost potentials to the total potential energy and the dihedral energy [27, 28]. The form of the boost potential requires two parameters, $E$ and $\alpha$, to be specified based on equations determined in McCammon’s lab (Figure 3.1b) [27]. The first nanosecond of the simulation was done using cMD, where $E$ and $\alpha$ for the total potential energy and the dihedral energy of the protein, were calculated and added to the conformation file to continue the subsequent production run in aMD mode up to 200 ns simulation time. Boosting the potential for the total potential energy is done to increase the degree of diffusivity for the system, where the solvents are also accelerated. The additional boosting potential for the dihedral energy is done to increase the conformational sampling of the protein, as protein conformational changes mostly happen through torsional rotations.
Figure 3.1. Accelerated molecular dynamics equations as described in McCammon’s papers [27, 28]. (a) Modified potential, \( V^*(r) \), that is used when the potential energy is below a user-specified value threshold of \( E \). (b) Form of the boost potential, \( \Delta V(r) \), that includes two parameters, \( E \) and \( \alpha \), that must be specified by the user. (c) \( E \) and \( \alpha \) for the total potential energy that is aimed to increase the diffusivity of the system as the solvents are also accelerated. (d) \( E \) and \( \alpha \) for the dihedral energy of the protein, as changes in protein conformations are largely due to torsional rotations.

Conformational Clustering

Conformational clustering is embedded within the NAMD/VMD interface, where it performs a clustering analysis by finding clusters of time steps that are similar with respect to a given distance function for the atoms in selection [29]. The method is based on the quality threshold (QT) algorithm [30]. RMSD of the backbone was used as the distance function excluding hydrogen atoms. The number of clusters was determined to be five and the maximum distance value between two frames, or the cutoff distance, was determined to be 1.5 Å.
**Principal Component Analysis**

Principal Component Analysis (PCA) were done using Carma software [31]. Coordinates from the aMD trajectories were stripped of solvents molecules and neutralizing ions and rotations and translations were removed. A new PSF (protein structure file) was built for the stripped-down system. PCA calculations were done using orthogonal transformation that converted the trajectory frames into a set of values of linearly uncorrelated variables. PCA identified principal directions where the data varied. In our analysis, we projected the first PC1 and the second PC2 for all Ras conformations. In addition, Carma also calculates the principal components for which ΔG (energy landscape) varied [31]. PCA-based cluster analysis is embedded in Carma software, where it automatically performed an analysis of the three largest principal components with the aim of identifying prominent molecular conformations corresponding to heavily populated clusters [31].

**Community Network Analysis**

Dynamical network analysis is a general method used to obtain an accurate picture of network topology and long-range signaling in protein complexes derived from molecular dynamics simulations [32]. Each amino acid residue in the complex is assigned a node centered on its Cα atom and used as a base to construct significant regions of amino acid interactions and pathways of allosteric modulation that connect them. Edges are placed to connect the nodes between residues that remain within 4.5 Å distance for at least 75% of the simulation time. The edges are weighted using pairwise correlation data calculated by the program Carma [31]. This information can then be mined to define community networks using the Girvan-Newman algorithm [33] and the diversity of paths that connect sites of functional significance in the
complex [32]. Nodes in the same community network can communicate with each other easily through multiple paths, whereas those in distinct community networks either do not communicate well or communicate through one or a small number of nodes essential for allosteric modulation.

*Wide Angle Solution Scattering*

Protein expression, purification, and nucleotide exchange were done on all three Ras isoforms (residues 1-166) according to a previously published protocol [34]. GDP-bound proteins were concentrated and exchanged into buffer for 5’-guanylyl imidodiphosphate (GppNHp) as described previously [35]. Purified Ras proteins were exchanged into stabilization buffer (20 mM HEPES pH 7.5, 5 mM MgCl2, 50 mM NaCl, 1 mM DTT), concentrated to two different concentrations, 5 mg/mL and 10 mg/mL for WAXS experiments. For each sample, two measurements were conducted, one of protein in buffer solution and another one of the buffer solution by itself. Scattering data from the buffer was subtracted from the protein in buffer solution to get protein scattering data. Scattering data were collected at the G1 beam line at Cornell High Energy Synchrotron Source (CHESS). X-ray source was of 9.846 keV and \( \lambda = 1.259 \) Å with a beam size of 250 \( \mu \)m * 250 \( \mu \)m [36]. Data collection was in the range of 0.009 < \( q \) < 0.072 Å. They were processed using the software RAW [37]. WAXS and MD data were combined to directly compare conformational flexibility of protein in solution seen in experimental setting and *in silico*. Initially, we use the software package XS [38] to calculate the scattering of a rigid protein with the crystal structure that was used to start the MD simulations. The MD trajectory was converted to produce a sigma-r plot, a plot of intermolecular distances as a function of that distance [39]. Vector length convolution was used to obtain the calculated
WAXS intensity, where each interatomic vector in the pair correlation function is replaced by a Gaussian distribution of vector lengths [40]. The two calculated intensities were then fitted on to the experimentally observed intensity for comparison.

Results

 Accelerated MD helps reveal conformational changes of Ras in solution

Two simulations systems, classical MD (cMD) and accelerated MD (aMD) show differences in the range of conformational space that are sampled throughout the simulation time. In cMD, all three isoforms show no major conformational changes, sampling R state and state 2 conformations throughout the simulation time. To circumvent this problem, we conduct aMD simulations for all three isoforms and show that we are able to capture the diverse conformations of Ras-GTP, including R/T-states conformations and state 1/state 2 conformations. R- and T-states conformations were previously shown through X-ray crystallography experiments using H-Ras [7]. In here, we show that all three isoforms, H-, K-, and N-Ras undergo conformational changes from R- to T-state. Additionally, state 2 to state 1 conformational changes are observed in K- and N-Ras isoforms, but not in H-Ras. These conformational changes are unique to aMD simulations and are not observed in cMD simulations. This is due to the additional boost potential added to the total potential energy and the dihedral energy of the protein. It has been shown previously that the bias potential still echoed the original shape of the true unaccelerated potential energy landscape [27]. Furthermore, the true potential can be obtained by reweighting points in the phase space on the bias potential by Boltzmann factor, since the level of applied boost is known [28]. We show here that the locations of flexible residues are similar when
comparing cMD to aMD simulations in H-Ras (Figure 3.2). In aMD the residues have higher flexibility in almost all parts of the protein, and this facilitates the conformational transition from R- to T-state. The overall fold of the protein is conserved and only regions that have been shown previously through experimental, X-ray crystallography and NMR studies, to be of higher flexibility show changes in conformations during our accelerated MD simulations [4, 5, 7, 41]. Comparing cMD to aMD, we show that residues around the allostERIC site and particularly on helix 3/loop 7 experience greater dynamic range in aMD than cMD (Figure 3.2). This increase in the flexibility of these residues allows for a kink to form on helix 3, resulting in a shift of helix 3/ loop 7 toward switch II, characteristics of the T-state structure [41]. This conformational change triggers an increase in the flexibility switch II, consistent with the disordered state of switch II in the crystal structure of Ras in the T-state [7, 41]. RMSF plot comparing aMD to cMD also shows that significant increases of more than 0.5 Å in the flexibility of residues on the five helices are only observed on helix 2 and 3, the active site and the allostERIC site helices (Figure 3.15). The fact that aMD captures conformational states previously observed in crystal structures and associated with biologically relevant states, further confirms the relevance of aMD simulations for sampling biologically relevant protein conformations, as has been reported previously for other proteins through several case studies. An example was shown in the case of neuraminidase that could captured a new wide-open state through aMD simulation, but not with cMD simulation [42].
Figure 3.2. Average Cα root-mean-square-fluctuations (RMSFs) for Ras comparing classical MD to accelerated MD method. Protein secondary structure is shown along the x-axis of the plot, where [tube, α-helix], [arrow, β-strand], and [line, loop]. Protein structure is shown in grey with increases in fluctuations shown in blue and decreases shown in red (>0.25 Å). The two switches are labeled SI and SII. Almost all regions of Ras experience increase in fluctuations when simulated in accelerated MD as compared to classical MD method. The inter-switch region and helix 3/loop 7 experience significant increases in fluctuations.

*Accelerated MD simulations produce calculated scattering patterns that are in excellent agreement with the experimentally observed data*

We use wide angle X-ray solution scattering (WAXS) experiments to directly verify our MD simulations data [43]. X-ray solution scattering captures the behavior of protein in solution where data analysis can quickly determine radius of gyration, molecular envelopes and the pair-distribution function [44]. While typically X-ray solution scattering data are of low resolution,
wide angle X-ray solution scattering extends data in the SAXS regime out to higher resolution that is sensitive enough to detect structural changes in proteins [40]. In collaboration with Lee Makowski and his graduate student Hao Zhou, we conduct WAXS experiments to directly validate the results from our accelerated MD simulations. A schematic for our method of calculating and combining MD simulations data with WAXS experimental data is shown in Figure 3.3. Our calculations start with a crystal structure, from which we conduct MD simulations and calculate the scattering intensity curve for a rigid protein using the program XS [38]. We then Fourier transform the intensity curve of the rigid protein from reciprocal space to real space, showing the pair distribution function of the interatomic distances [39]. Using high performance computation, we calculate the sigma-r plot of the flexible protein from the MD trajectory [43]. Sigma-r plot is obtained by calculating the interatomic distances of the protein throughout the simulation time and plots their standard deviations as a function of that distance [39]. To calculate the intensity curve of the flexible protein by incorporating MD data, we use vector length convolution (VLC), where each interatomic vector in the pair distribution function of the rigid protein is replaced by a Gaussian distribution of vector length from the sigma-r plot [45]. In other words, we merge the sigma-r plot of the flexible protein obtained from the MD trajectory with the pair distribution function of the rigid protein to obtain the calculated scattering intensity curve for the flexible protein. Comparing this predicted scattering pattern obtained from the MD trajectory to the observed scattering pattern from experimental WAXS data, we can directly assess the quality of our MD simulations in predicting protein motion in solution.
Figure 3.3. Schematic of calculations used to combine MD simulation data with WAXS experimental data. A crystal structure is used to conduct MD simulations and to calculate the intensity curve for a rigid structure using the XS program [38]. Using Fourier transformation, calculated intensity is converted into a pair distribution function of the interatomic distances of the protein. MD trajectory is converted into a sigma-r plot using high performance computation (HPC), where it calculates the interatomic distances of the protein throughout the simulation time and plots the standard deviation of the interatomic distances as a function of that distance [39].

Vector length convolution (VLC) is used to merge sigma-r plot with the pair distribution function from the rigid protein to obtain the calculated intensity for a flexible protein [45]. The final step is to fit the calculated intensity from MD trajectory (in blue) onto the observed WAXS intensity (in red) to directly verify MD data.

To validate the dynamics of Ras in silico with X-ray scattering data, we fit the calculated intensity curve to the observed intensity curve by scaling their degree of flexibility using sigma-r plots. A scale factor close to one suggests similar Ras dynamics in silico as compared to the one
observed in solution. A value greater than one means that Ras is more flexible in the experimental X-ray scattering data than in MD simulations. Scale factor less than one suggests that the protein is more flexible in MD simulation than is shown in solution scattering data. We show for all three isoforms that the calculated intensity curves using aMD simulation data fit well to the experimentally observed intensity curves for all three isoforms up to 0.1 \( \text{1/d} \), which is 0.1 Å\(^{-1}\) (10 Å spacing) (Figure 3.4). In particular H- and K-Ras show excellent fitting between the calculated intensity (blue curve) and the observed intensity (red curve) (Figure 3.4a, b). The rigid Ras proteins (black curve), show sharper features than the flexible Ras proteins obtained from MD trajectories and the experimental data. The flattening of the intensity peak at around 0.045 \( \text{1/d} \) for H- and K-Ras (real space = 22 Å) is characteristics of an increase in protein flexibility (Figure 3.4a, b). The flattening of this intensity peak is more prominent in K-Ras than H-Ras, indicating that K-Ras is more flexible in solution as compared to H-Ras. This is further shown in the sigma-r plot where K-Ras has higher values for the standard deviation of interatomic distances as compared to H-Ras (Figure 3.5a). N-Ras shows relatively less agreement between the calculated intensity curve and the observed intensity curve (Figure 3.4c).

Additionally, we show that aMD simulations show better agreement with the WAXS experimental data when compared to cMD simulations. Scale factors that we obtain by scaling up the sigma-r plots of MD simulations data to match the observed intensity are all smaller in aMD as compared to cMD simulations (Table 3.1). This suggests that aMD simulations better represents the overall dynamics of Ras in solution than cMD, as the enhanced flexibility in aMD simulations is verified through WAXS experimental data. Sigma-r plot comparing the scaling of K-Ras from aMD simulation against cMD simulation is shown in Figure 3.5b. K-Ras cMD
Figure 3.4. Accelerated MD simulations produce calculated scattering patterns that fit well to the experimentally observed WAXS data. In black is the reference curve calculated from a rigid crystal structure. In red is the experimentally observed WAXS scattering pattern. In blue is the calculated scattering pattern from aMD simulations. All three isoforms show excellent agreement from MD simulations to the experimentally observed data.
simulation data is scaled to 2.46 to match the target experimental sigma-r value, whereas K-Ras aMD data is only scaled to 1.18.

**Figure 3.5.** Sigma-r plots of Ras. In solid lines are sigma-r values from MD simulations and in dotted-lines are scaled sigma-r values to match the experimental data. (a) Comparing K- and H-Ras, K-Ras in blue is more flexible than H-Ras in red, with higher sigma-r values. (b) Comparing K-Ras from aMD simulation to cMD simulation, where sigma-r values of aMD simulations in blue are closer to the scaled sigma-r values to match the target experimental data than cMD simulations in red.

**Table 3.1.** Scale factor values for H-, K-, and N-Ras comparing accelerated MD simulations to classical MD simulations. Values closer to one in aMD than cMD suggests that aMD shows dynamics that are more similar to the experimentally observed WAXS data than cMD.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Scale factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>aMD</td>
</tr>
<tr>
<td>H-Ras</td>
<td>1.12</td>
</tr>
<tr>
<td>K-Ras</td>
<td>1.18</td>
</tr>
<tr>
<td>N-Ras</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Ras isoforms show differences in residue flexibilities, specifically around the active and the allosteric sites.

The two switch regions of Ras, switch I (residues 30-40) and switch II (residues 60-76) change conformations readily during binding and catalysis [46]. These regions are located at the effector lobe (residues 1-86) that is 100% conserved across the three isoforms, K-, H-, and N-Ras. The allosteric site elements include the C-terminus of helix 3, loop 7, and the C-terminus of helix 4 (residues 97-108, and 137) [3, 7, 11]. They are located at the allosteric lobe (residue 87-166) which is 90% conserved between the isoforms. Our results from accelerated MD compare the α-carbon root-mean-square-fluctuations (RMSFs) for Ras residues and show that at specific important sites, including the active and the allosteric sites, Ras isoforms show differences in flexibility (Figure 3.6). Specifically, K-Ras shows highest overall flexibility at these sites with 3.5 Å average, followed by N-Ras with 2.9 Å and H-Ras with 2.5 Å. At switch I, K-Ras has average fluctuations of 3.1 Å, followed by H-Ras with 2.6 Å and N-Ras with 2.4 Å. There are two main residues on switch I that are responsible for promoting state 1 and state 2 conformational transition, Y32 and T35. These simulations start with state 2 conformation and fluctuates between states 1 and 2 based on the positioning of residue Y32 relative to the γ-phosphate of the nucleotide. RMSF data show that residue Y32 is most flexible in K-Ras (8.1 Å), follow by N-Ras (5.5 Å) and H-Ras (5.0 Å). It suggests that K- and N-Ras sample more state 1 conformations during the simulations than H-Ras. For K-Ras this is consistent with NMR data showing a more open state 1 conformation for K-Ras-GppNHp relative to H-Ras-GppNHp [47].

K-Ras-GppNHp has an even higher average fluctuation at switch II, with 4.3 Å, followed by N-Ras with 3.2 Å, and H-Ras with 2.5 Å. This again is consistent with our NMR experiments
The allosteric site residues have the same fluctuations average for K- and N-Ras with 3.0 Å and lower for H-Ras with 2.3 Å. The general trend for the active site and the allosteric site residues is that K-Ras is the most flexible isoform, followed by N-Ras and H-Ras. H-Ras shows relatively little fluctuations. Helix 2 and helix 3 are located at the interface between the effector and allosteric lobes, where they show dynamics guided by electrostatic interactions between amino acid residues at belonging to the two helices. Three prominent salt bridges connecting helix 2 and helix 3 were identified throughout the simulation, formed between E62-K88, R68-D92, and D69-R102. The stability of these salt bridges varies significantly across Ras isoforms, where they are most stable in H-Ras, followed by N-Ras and K-Ras (Figure 3.7c). The salt bridge E62-K88 is not sampled in K-Ras but appears 26% of the simulation time for H-Ras and 40% for N-Ras. This salt bridge is located at the N-terminus of helix 2 and helix 3, which partially explains the highest flexibility observed at the N-terminal portion (residues 60-65) of helix 2 seen in K-Ras as compared to N- and H-Ras. This salt bridge conformation, E62-K88, that is found in N- and H-Ras was previously seen in Ras-GTP/SOS structure complex (PDB 1NVW), where Ras adopts an intermediate conformation between the T and the R states [48]. This conformation was suggested to prolong the active form of Ras bound to GTP, when Ras is in bound to the allosteric site of SOS [3]. The lack of this salt bridge in K-Ras may be an indication that this isoform could have different kinetics of allosteric feedback when bound to the allosteric site of SOS.

At the N-terminal portion of helix 2, K-Ras has average fluctuations of 6.0 Å, followed by N-Ras with 3.6 Å and H-Ras with 2.8 Å. Amino acid residues at the N-terminal end of helix 2 in K-Ras are solvent exposed and do not form stable interactions with neighboring residues at either helix 3 or switch I. This has an additional implication on the conformation of K-Ras as
compared to the other two isoforms. K-Ras has an open conformation between helix 3 and switch II, where the amino acid residues that are supposed to be forming salt bridges as in seen in H- and N-Ras are far apart (Figure 3.7). In K-Ras, these residues interact with adjacent residues on the same helix, including R68-D69 and K88-D92 (Figure 3.7b).

**Figure 3.6.** Average Cα root-mean-square-fluctuations (RMSF) for H-, K-, and N-Ras. Protein secondary structures are shown on the x-axis. The plot shows that K-Ras is the most flexible isoform, specifically at the switch regions (SI and switch II) and the allosteric site (α3/loop7).
Figure 3.7. Salt bridge interactions between switch II and helix 3. (a) H-Ras has 3 stable salt bridge interactions, in contrast (b) those salt bridges are absent for the most part in K-Ras. (c) Table depicting the lifetime of salt bridge interactions for H-, K- and N-Ras.

Switch II in K- and N-Ras adopts two different stable conformations

High flexibility of switch II in K- and N-Ras enables them to adopt two different stable conformations, one interacting with switch I and the other interacting with helix 3. K-Ras simulation initially starts with a partially empty allosteric site, where E107 is solvent exposed and away from the allosteric site. Meanwhile, switch II is stacked against helix 3, where Y64 interacts with Q99 and E62 forms a hydrogen bond with H95. As loop 7 changes conformation
and E107 moves into the allosteric site, the close interaction between switch II and helix 3 breaks, and switch II opens and flips toward switch I. Y64 has moved 10 Å away from its initial conformation. Following this, helix 3/loop 7 adopts T state conformation, where Y137 forms a hydrogen bond with the backbone of H94 that results in a kink in the middle of helix 3. Meanwhile, R97 pulls away toward helix 4 and forms a hydrogen bond with D108 and the C-terminal end of helix 3/loop 7 shifts toward switch II, pushing the C-terminal of switch II down. This changes the overall conformation of switch II and switch I, where Y64 now forms a stable conformation with I36 and P34 at the end of switch I, and with A59 at the beginning of switch II. Residue Y64 remained in this conformation through the rest of the simulation. It is the most flexible residue in K-Ras and seems to mediate the switch II transition between the two very different conformations of switch II in K-Ras (Figure 3.8).

Figure 3.8. Highly flexible SII in K-Ras yields two different stable conformations of SII guided by residue Y64. In cyan, K-Ras at 15ns with SII interacting with helix 3, residue H95 on helix 3 forms a hydrogen bond with E63 on SII. In green, K-Ras at 45ns with SII interacting with S1, residue Y64 forms a hydrophobic interaction with I36 on S1.
N-Ras simulation initially starts with Y64 forming a hydrophobic pocket near the active site by interacting with I36 and A59. Like K-Ras, this conformation of Y64 is accompanied by D107 and R97 occupying the allosteric site. When the allosteric residues, R97 and D107 move away from the allosteric site, it triggers a big conformational change at switch II that flips Y64 away from switch I toward helix 3. Switch II is stacked against helix 3, and the backbone of Y64 has moved 8 Å away from its initial conformation near switch I. Y64 forms a stable hydrophobic interaction with L95 and Y96 (Figure 3.9). These two different stable conformations of switch II are similar to how they are formed in K-Ras, although the conformation of Y64 interacting with helix 3 is different due to an isoform specific residue. At position 95, N-Ras has a leucine that can form a hydrophobic interaction with Y96, but K-Ras contains a histidine at position 95 that interacts with E63 while Y64 interacts with Q99 (Figure 3.9).

**Figure 3.9.** Switch II and helix 3 interactions in K-Ras (green) and N-Ras (purple) are guided by isoform specific residue 95, H95 in K-Ras and L95 in N-Ras. In green is K-Ras at 15 ns, where H95 interacts with E63 stabilizing a switch II conformation. In pink is N-Ras at 90 ns, where L95 forms a hydrophobic interaction with Y64 to stabilize a switch II conformation.
Switch I in K- and N-Ras, but not in H-Ras, is highly flexible and transitions to state 1 conformation from state 2 during the simulation.

GTP-bound Ras can adopt two different conformations that affects its affinity toward effector proteins, they are state 1, where switch I is open, and state 2, where switch I is closed over to the nucleotide [4, 6]. State 2 is conducive for signaling events, whereas state 1 is not because the open switch I disrupts the binding site for effector proteins. In our accelerated MD simulations, the additional boost potential helps to transition K- and N-Ras from state 2 to state 1, but this does not happen in the H-Ras aMD simulation. Early in the simulation for K-Ras, switch II shows a high degree of flexibility and swings from interacting with helix 3 to interacting with switch I. At 15-ns simulation time, the switch II region moves closer to switch I, forming hydrogen bond interactions between the hydroxyl group of Y71 and the backbone carbonyl group of E37. Meanwhile, the most flexible residue, Y64 forms a hydrophobic interaction with I36 and P34 pushing Y32 away from the nucleotide. At 50-ns simulation time, Y32 swings away from the \( \gamma \)-phosphate of GTP due to the positioning of Y64 near the active site. Y32 is flipped all the way to form a hydrogen bond with Q25 on helix 1. Superimposing this conformation to the initial structure, it shows a major difference at switch I, mimicking a state 1 open conformation (Figure 3.9). This conformation of switch I is held throughout the 200-ns simulation time, suggesting that state 1 is the predominant state for K-Ras in solution. Our lab has indeed shown that K-Ras-GTP predominantly adopts state 1 conformation in solution and was also crystallized in state 1 conformation (Parker J. et al, 2018).

N-Ras eventually also transitions to state 1 conformation from its state 2 starting structure during the simulation. However, while K-Ras transitioned from state 2 to state 1 at around 50 ns simulation time, N-Ras transitioned at 130 ns, much later in the trajectory. Unlike K-Ras
simulation events, N-Ras starts with Y64-I36-A59 forming a hydrophobic pocket near the active site when N-Ras is in the state 2 conformation. The transitioning from state 2 to state 1 happens when Y64 flips over to helix 3 to form a hydrophobic interaction with L95 and Y96. As Y64 leaves the active site, residue Y32 flips away from the γ-phosphate of GTP to form a hydrogen bond with Q25 at the C-terminal of helix 1. This is open conformation is similar to that observed in K-Ras.

**Figure 3.10.** High flexibility in switch I of K-Ras enables it to transition from state 2 conformation (gray) to state 1 conformation (green). In grey, the starting structure of K-Ras where switch I adopts a closed state 2 conformation. In green, K-Ras at 50 ns, where switch I opens up and residue Y32 flips away to form a hydrogen bond with Q25 on helix 3, stabilizing an open state 1 conformation.

*Isoform specific residues on the allosteric lobe influences the stability of hydrogen bonding networks for nucleotide sensing*

The Mattos Lab developed and used the program DRoP, a water analysis method, to identify a water-mediated network of H-bonding interactions that connects the active site to
nucleotide sensing residues R161 and R164 on helix 5 of the allosteric lobe that directly contact the membrane (the Helix 5 network) (53). Residue N85 is critical in this network and it was demonstrated with in vitro hydrolysis experiments that it significantly affects nucleotide binding and catalysis [49]. Thus, our analysis shows that, at least in H-Ras, the network involving N85 stabilizes nucleotide binding, is important in the intrinsic GTP hydrolysis reaction (in the absence of GAP) and senses whether Ras is bound to GDP or GTP [49]. In our simulation data, we find that the allosteric lobe residue differences at position 122, where it is a Serine in K-Ras, Threonine in N-Ras and Alanine in H-Ras, influences the hydrogen bonding network through N85 for nucleotide sensing. The hydroxyl side chains of Serine and Threonine interact with the side chain of N85, interrupting the network, but this is not the case for Alanine in H-Ras.

Additionally, the Helix 5 network and conserved motifs NKxD and ExSAK in Ras proteins are joined by the R123-E143 salt bridge on the allosteric lobe. This salt bridge interaction is 100% maintained throughout the simulation in H-Ras, 78% in N-Ras, and 61% in K-Ras. There are a few isoform-specific residues around this salt bridge that help to stabilize it differently in the three proteins (Figure 3.11). Residue Y141 in H-Ras interacts with E143 and, together with the backbone carbonyl group of V125, the side chain hydroxyl group of S127 and the side chain of Q131, forms a network of H-bonding interactions that stabilizes the R123-E143 salt bridge and thus the active site (E143 is part of one of the conserved ExSAK active site motifs). This interaction is different in both N- and K-Ras where residue 141 is a phenylalanine and residue 127 is a threonine that are in van der Waals contact and no longer part of the network. This isolates residue 131 (Q in K-Ras, H in N-Ras) from the network and in essence removes one of the anchors that tie helix 4 to one of the core β-sheet strands (b6 that leads into a5) and stabilizes the R123-E143 salt bridge (Figure 3.11). The side chain of R123 in H-Ras has a unique
conformation as compared to the other two isoforms, with its three-carbon aliphatic side chain in its preferred rotamer stabilized by the interaction between the V125 backbone and the guanidinium group of R123.

**Figure 3.11.** The stability of salt bridge R123-E143 is highly influenced by the residue differences at the allosteric lobe. (a) Superimposition of the three isoforms centered on R123-E143 salt bridge. (b) H-Ras in cyan. (c) K-Ras in green. (d) N-Ras in purple.
**N-Ras shows a unique interaction between Y166-K104 that destabilizes switch II**

N-Ras has a more flexible switch II and helix 3/loop 7 than H-Ras, and a more flexible C-terminus of switch II and loop 7 than K-Ras (Figure 3.6). Analysis of the protein conformation throughout the simulation reveals that N-Ras has a unique hydrogen bonding interaction between Y166-K104 that destabilizes switch II (Figure 3.12). This hydrogen bond is not found in K- and H-Ras because of sequence differences at residue 166, where K- and H-Ras contain Histidine and N-Ras has a Tyrosine. This hydrogen bond is temporary and occurs at 12-ns for a mere 3-ns simulation time. Residue K104 has been shown previously to be acetylated in K-Ras and was proposed to interfere with the stable conformation of switch II [50]. Residue K104 is proposed to form a salt bridge with the carbonyl backbone of R73 that stabilizes switch II. Acetylation of K104 eliminates this salt bridge. We suggest that N-Ras undergoes a similar conformational change seen in acetylated K-Ras, where Y166 sequesters K104 from interacting with switch II, thus temporarily destabilizing switch II, particularly at its C-terminus.

![Figure 3.12](image.png)

**Figure 3.12.** N-Ras shows a unique interaction between residues Y166-K104 that sequesters K104 from interacting with switch II and destabilizes the catalytic conformation of switch II.
RMSD conformational clustering and PCA analysis show that all three isoforms prefer to be in the T-state

Root-mean-square-deviation (RMSD) clustering finds clusters of similar conformations from the total number of frames in MD trajectory based on a given distance function [23]. In accelerated MD we increase the chance of sampling biologically relevant conformations every once in a while by adding boost potentials to better navigate the potential energy landscape [28]. We set the number of clusters to five to properly separate relevant conformations from the rest. For each of the isoforms, we found that a T-state like conformation is the major conformer. This conformation has a kinked helix 3, where the C-terminal end of helix 3 shifts toward switch II and destabilizes the helical conformation of switch II. T state population is 41% in K-Ras, 55% in H-Ras, and 54% in N-Ras. Our laboratory has shown previously that H-Ras exists in two conformational states in GTP-bound form, a catalytically ordered R state (reactive) and a disordered T state (tardy) [7]. R state has a signature conformation with ordered allosteric site elements at helix 3/loop 7/ helix 4 that culminates in a hydrogen bond network that connects it to the active site and stabilizes the active site components. In our clustering, we find that all three isoforms prefer to reside in the T state conformation where the active site is disordered. Additionally, we find that R state conformation occurs more frequently in H-Ras with 26%, than either K- or N-Ras, which have 11% and 14%, respectively. This is consistent with RMSF data where H-Ras is shown to be the most stable isoform, and with our NMR results showing more flexible and open switch regions in K- compared to H-Ras (Figure 3.6).

PCA projection of the first two principal components show that R state starting structure of Ras converted to T state during the MD simulation (Figure 3.13, Figure 3.14, Figure 3.15). We calculate the average structure in the highest density clusters and superimpose onto the
starting structure, where we show that the highest density structure from all three isoforms resemble T state, with a signature kink in the middle of helix 3 (Figure 3.13c, Figure 3.14c, Figure 3.15c).

To understand the network connections between residues in the allosteric and the effector lobes, we run a community network analysis. Network analysis assigns each residue as a node, and edges are used to connect nodes that interact >75% of the simulation time [32]. Community networks are subnetworks that have stronger connections within the community than to nodes in other communities [33]. Calculations of the community network for the three isoforms show that H-Ras has the largest community that connects helix 3 and switch II (Figure 3.16). The effector lobe also shows differences in the network of communities that connect the two switches, switch I and switch II. K-Ras is the unique isoform that has a separate community for each of the switch regions, whereas H- and N-Ras contain a community that connects the active site residue Y32 at switch I to the residues at the N-terminus of switch II (Figure 3.16). It suggests that K-Ras is more prone to adopting a state 1 conformation with an open switch I than the other two isoforms.

**Figure 3.13.** Conformational sampling of H-Ras aMD. (a) Projection of PC1 and PC2. In black is the cluster with the highest density and in grey is the starting structure. (b) Free energy
landscape showing the highest density structure occupying the lowest energy wells. (c) Highest density structure resembles T state conformation when superimposed on to the starting structure.

**Figure 3.14.** Conformational sampling of K-Ras aMD. (a) PC1 and PC2 for all the conformations in the simulation trajectory. In green is the highest density cluster and in grey is the starting structure. (B) Free energy landscape showing highest density cluster in the lowest energy wells. (C) Highest density structure resembles the T state when superimposed on to the starting structure.

**Figure 3.15.** Conformational sampling of N-Ras aMD. (a) PC1 and PC2 for all configurations. In red is the highest density cluster and in grey is the starting structure. (B) Free energy
landscape from the same PC1 and PC2, showing highest density cluster in the lowest energy wells. (C) Highest density structure resembles T state when superimposed on to the starting structure.

Figure 3.16. Community network analysis, showing the largest community connecting the effector and the allosteric lobe in H-Ras when compared to the K- and N-Ras. H- and N-Ras has a shared community in grey that connects active site residue Y32 on switch I to the residues in switch II, but not in K-Ras.

**H- and N-Ras bind calcium acetate at the allosteric site, but not K-Ras**

Calcium acetate was shown to bind at the allosteric site for H-Ras [7]; using this crystal structure K- and N-Ras were simulated using classical MD simulation. Around the allosteric site, the only residue variation is found at position 107, where it is an aspartic acid in H- and N-Ras, and a glutamic acid in K-Ras. At the allosteric site, R97 is the only residue side chain that interacts directly with the acetate molecule by forming a hydrogen bond. During the simulation, residue 107 forms a stable salt bridge interaction with R97 for the most part. An extra carbon on the side chain on E107 shows a more flexible interaction with R97 as compared to D107. E107 guides R97 away from the allosteric site toward E98 and breaks the hydrogen bond between R97 and the acetate molecule in the K-Ras simulation (Figure 3.17b). D107 in H- and N-Ras did not
do the same thing, but instead they form a stable head on salt bridge with R97 while keeping it at the allosteric site and stabilizing its interaction with the acetate molecule. The final conformation seen in H-Ras is shown in Figure 3.17a, where calcium acetate is retained in the allosteric site, while D107 helps to stabilize R97-acetate hydrogen bond interaction.

![Figure 3.17](image.png)

**Figure 3.17.** H-Ras and K-Ras conformations from cMD simulation show differences in binding calcium acetate. (a) H-Ras at 90 ns shows calcium acetate retained in at the allosteric site, where R97-acetate hydrogen bond is stabilized by D107. (b) In contrast, K-Ras at 32 ns expels calcium acetate from the allosteric site with the help of E107 that breaks the hydrogen bond between R97 and the acetate molecule.

**Discussion**

* Differences in dynamics and balance of conformational states in K-, N-, and H-Ras are guided by variation of residues in the allosteric lobe*
The allosteric lobe of the soluble domain of Ras, residues 87 to 166 contains 17 residues that are different between K-, N-, and H-Ras. Out of these 17 residues, each of the isoforms contains several unique residues, where N-Ras has nine, H-Ras has seven, and K-Ras has five residues that do not appear in the other two isoforms. We have shown through our simulations that these residues affect the overall conformations of switch I and switch II and the hydrogen bonding networks that connect them to the allosteric lobe of Ras. Residue 95 which is different in each isoform, Histidine in K-Ras, Leucine in N-Ras, and Glutamine in H-Ras, affects the interaction between helix 3 and switch II. Highly dynamic switch II in K- and N-Ras has Y64 guiding the interaction between switch II and helix 3 differently than in H-Ras. Residue Y64 in switch II is one of the most flexible in the Ras proteins, particularly in K-Ras where it has an RMSF of 12.1 Å and N-Ras where its RMSF is 5.5 Å, but only 3.4 in H-Ras (Figure 3.6). In the K-Ras simulation, Y64 forms a hydrogen bond with Q99 and away from H95, whereas in N-Ras, Y64 stacks up against L95 on helix 3. These interactions give rise to two different conformations that opens switch II and exposes the P-loop at the active site toward the solvent. Previous HDX experiments have shown that there is a correlation between switch II and p-loop deuterium exchange levels amongst Ras isoforms, when P-loop experiences high deuterium exchange, switch II experiences lower deuterium [17]. We show here through our simulations that the opening of switch II to expose P-loop is different across the three isoforms. This helps explain the difference in the level of deuterium exchange amongst the three isoforms from structural perspective (Figure 3.8, 3.9). K-Ras has higher P-loop exchange than N- and H-Ras, because switch II is packed up higher on to the C-terminal end of helix 3, exposing more of P-loop to the solvent. H-Ras has a more rigid switch II that does not open enough time to expose P-loop for deuterium exchange. Additionally, residue Y64 has been shown through mutagenesis studies to
be important in effector and regulator proteins binding, as well as dimer formation [51-54]. Our work suggests that the highly flexible residue Y64 is of biological relevance in the context of conformational changes that are associated with effectors and regulators binding for signaling events of Ras.

Furthermore, in accelerated MD each isoform shows correlation between residues at helix 3/loop 7 and residues at switch I and switch II. Helix 3/loop 7 residues that affect the conformation of the allosteric site are 94-98, 101 and 107, 108 and 137 at helix 4. The allosteric site cleft is usually occupied by the side chains from residues R97 and D/E107, and when these two residues change conformations and leave the allosteric site, it results in major conformational changes at switch I and switch II which directly affect the active site. This connectivity between residues at helix 3 and residues at switch I and switch II have been reported using conditional time-delayed correlation (CTC) calculation [12]. Vatansever and colleagues show a causality of correlated motions, in that helix 3/loop 7 motions drive switch I and switch II motions in K-Ras. We see the same event happening in our simulation for all three Ras isoforms, where each isoform shows allosteric effects from helix 3/loop 7 to the active site regions, through switch I and switch II, although, there are differences in the way the switch conformations are affected across the isoforms.

Residue K104 has been shown to be acetylated and ubiquitylated in Ras and plays a role in allosteric modulation of switch II [50]. As discussed before, in N-Ras, Y166 disrupts the salt bridge between K104 and the C-terminal end of switch II, which in turn destabilizes switch II residues and its helical structure. The simulation trajectory shows a momentary destabilization of the helical conformation of switch II (Figure 3.12). Over time switch II is re-stabilized and it reforms its helical conformation. As evidenced by the residue RMSFs, we see that on average K-
Ras has the most flexible switch II as compared to N- and H-Ras (Figure 3.6) in spite of the fact that residue 166 is a histidine, as it is in H-Ras. If K104 salt bridge to the C-terminal end of switch II is critical to stabilize switch II, we would have seen a more flexible switch II in N-Ras as well as a major conformational change for switch II. However, as shown in the simulation trajectory, the destabilization of switch II is temporary and does not affect much of the overall flexibility of switch II. This result is consistent with the finding from K-Ras K104Q cell biology and NMR studies that reveal a temporary disturbance of switch II in K104Q relative to WT K-Ras. However, the overall binding affinity to effector proteins are not disturbed. There are, however, reduction in the efficiency of regulator proteins in catalyzing the nucleotide exchange, GEF, and hydrolyzing GTP, GAP. These disturbances offset each other and resulted in no overall effect in signaling and oncogenic transformation ability of K-RasK104Q variant [55].

State 2 to state 1 transition is accessible through aMD simulations for K- and N-Ras. K-Ras goes through conformational change from state 2 to state 1 after 50 ns simulation time, whereas N-Ras takes longer to transition, after 130 ns simulation time. This result is consistent with the kinetics and ITC data collected from our lab for the three isoforms, in which the data suggest that state 1 to state 2 conversion in K-Ras has lower energy barrier than in N-Ras [11]. ITC data of Ras isoforms binding to Raf show that H- and K-Ras have the same binding affinity which is twice as tight when compared to N-Ras. The lower binding affinity of N-Ras to Raf suggests that conformational change from state 2 to state 1 in N-Ras has higher energy barrier than in K-Ras.

Based on a novel method developed in the Makowski Lab [43], we have used WAXS to show that the scattering curves obtained from aMD simulation trajectories for H-, K- and N-Ras more closely reflect the respective experimental WAXS scattering in solution than do the cMD
simulations. We therefore use the aMD trajectories in the analysis presented here. The results show that K-Ras is the most flexible of the three isoforms and that in the absence of other factors it has an effector-binding lobe and active site largely disconnected from the allosteric lobe containing sites of interactions with the membrane [20, 56] and the Ras dimerization interface [56]. The connection between the two lobes is driven by a compact closed conformation of switch I and switch II, not observed in K-Ras. In spite of a more open conformation at the C-terminal end of switch II due to the presence of Y166, N-Ras partially maintains connections between switch II and the allosteric site, which remains integral to ligand binding, while switch I is mostly in an open conformation, driven by changes in loop 8 and areas surrounding the R123-E143 salt bridge as in K-Ras. H-Ras on the other hand maintains more ordered switch regions with switch I stabilized in a state 2 conformation, as confirmed by NMR [47], and switch II fluctuating in a narrow range of conformations that allow frequent accessibility to the reactive R-state. This is consistent with our experiments comparing GTP hydrolysis rate constants between the three Ras isoforms [11]. Although all three proteins can access similar conformational states for switch I and switch II, the population of states is clearly different between them, driven by isoform-specific residues in the allosteric lobe of the G-domain. It is therefore likely that each isoform requires different levels of regulation by other factors in order to attain the relevant conformations associated with their common mechanism of interaction with effector proteins to propagate signaling in the cell, and GTP hydrolysis to turn the signal off.
Reference


Chapter 4: RasY71A Mutant Increases the flexibility of Switch II and Intrinsic Hydrolysis Rate Constant
Introduction

Ras is a small GTPase of 21 kDa that switches between GTP-bound active state and GDP-bound inactive state [1]. Structural studies based on X-ray crystallography and NMR have shown that in its active GTP-bound state, Ras can adopt multiple conformations, with the switch I state 1, and state 2 and the switch II R-(reactive) state and T-(tardy) state, being among the most well studied [2, 3]. Switch I must be in state 2 to interact with effector proteins [3] and given this conformation, switch II must be in the R state with an ordered active site to catalyze hydrolysis of GTP to GDP, modulated by ligand binding at a remote allosteric site [2]. Analyses of accelerated molecular dynamics (aMD) data for wild type H-Ras-GTP presented in Chapter 3 show that the R-state and T-state conformations are sampled throughout the simulations. Looking at the trajectory of aMD over time, we can see that the side chain of the switch II residue Y71 is highly dynamic and interacts with the side chain of residue Y96 on helix 3 and the main chain of residue E37 on switch I (Figure 4.1). Switch II is located between helix 3 and switch I, with Y71 sampling a range of conformational space to interact with either one or the other of these structural elements. Since, R- and T-states are associated with switch II conformations and its interaction with helix 3, we hypothesized that residue Y71 must be important in facilitating the transition between the R- and T-states. Indeed, the dynamic behavior of residue Y71 accompanies the transitioning from R- to T-state conformations. Based on this observation, we mutated Y71 to A to test whether eliminating the interactions with helix 3 and switch I, changes the overall flexibility of switch II, with consequences to the balance of conformational states in H-Ras. In this Chapter we present a crystal structure of H-RasY71A that is solved to 1.8 Å resolution and analyze switch I and switch II dynamics with respect to access
to the R state conformation. The structure shows that the elimination of tyrosine side chain at position 71 changes the overall hydrogen bond network from the allosteric site to the active site. Without Y71, R68 has more room to change conformation and flips to interact with the backbone of Q61. Additionally, Y96 shifts slightly relative to the wild type and forms a hydrogen bond with G60 on switch II. Although the crystal structure of the RasY71A mutant is in a conformation similar to the T state, the R state is accessed more frequently in the mutant than in the wild type simulations, consistent with an increase in the rate constant observed for the hydrolysis of GTP in the mutant relative to the wild type protein.

We propose that residue Y71 is important for conformational changes in Ras, particularly at switch II. To test this hypothesis, we initially ran aMD simulations of the H-RasY71A mutant starting from the wild type crystal structure of H-Ras (PDB 3K8Y) mutated in silico to make H-RasY71A by replacing the tyrosine side chain with an alanine residue. We refer to this model as the in silico model of H-RasY71A. Once we obtained the crystal structure of the mutant we ran an additional aMD simulation starting with the experimental crystal structure and found significant differences in the dynamics of Ras and the conformational states accessed in the two simulations. We thus used the method described in Chapter 3 of combining MD simulation data with wide-angle X-ray solution scattering (WAXS) [4] to compare the fitting and the scale factors from the two simulations with the behavior of the RasY71A mutant in solution based on scattering data. We find that the simulations started from the crystal structure much more accurately reflects the experimental data than that starting from making the mutation in silico. Thus, the change at residue 71 somehow affects the structure of Ras in at global level reflected in the crystal structure in a way not captured by mutating the single residue in the context of the
wild type structure. This work shows the importance of using the most appropriate starting structure for MD simulations to accurately predict protein’s dynamical behavior.

**Figure 4.1.** Snapshots from H-Ras-GTP simulations showing the highly dynamics residue Y71. In grey is a snapshot from 10-ns simulation time showing Y71 forming a hydrogen bond with Y96. In cyan is a snapshot from 14-ns showing Y71 flipped toward switch I and formed a hydrogen bond with E37.

**Methods**

*Protein expression and purification*

Crystallization, hydrolysis and solution scattering experiments were performed using the soluble domain of H-Ras proteins, residues 1-166 (EC 3.6.5.2). Protein expression and purification for H-Ras wild type and H-RasY71A were done based on previously published protocol for H-Ras [5, 6]. Proteins for hydrolysis experiments were stored at -80 °C in their GDP
bound forms [7]. In order to crystallize H-RasY71A in an active form, bound to GppNHp (Guanylyl-5-imidophosphate, a GTP analogue), nucleotide exchange reaction was performed following a previously described protocol [6, 8]. H-RasY71A was concentrated to 10 mg/mL for crystallization, 3.5 mg/mL for hydrolysis, and 5 mg/mL and 10 mg/mL for solution scattering experiments. Bradford assay was used to determine protein concentrations [9].

*Intrinsic hydrolysis experiments*

Kinetics hydrolysis were conducted following the single turnover experimental procedure explained in a previously published protocol [5, 10]. Ras-GDP at 5μM was incubated with radioactive γ³²Pi-GTP (Perkin Elmer) in the presence of a reducing agent, EDTA for 5 minutes. To start the reaction, 4 μL of protein solution was added to 16 μL of hydrolysis buffer that contains 20 mM Tris pH 8.0, 100 mM NaCl, 5 mM MgCl₂, 2 mM DTT, at 37°C. The final concentration of Ras was 1 μM and γ³²Pi-GTP was at 10 nM. Samples from each time step was collected up to 120 minutes, t_{max}, at the following time points = 0, 2, 6, 10, 12, 14, 16, 20, 30, 40, 50, 60, 80, 90, 100, and 120 min. Organic extraction protocol was used to isolate ³²Pi product [11, 12]. Radioactivity was measured using a scintillation counter (300SL, LabLogic). Best-fit values from the global fit gives k_{hyd} from the triplicates [13].

*X-ray crystallography*

Mutant H-RasY71A protein was concentrated at 15 mg/mL for crystallization. The protein was stored in a stabilization buffer containing 20 mM HEPES pH 7.5, 50 mM NaCl, 20 mM MgCl₂, and 1 mM DTT. Crystals were grown at 18 °C using hanging drop crystallization trays with 2 μL protein solution and 2 μL reservoir solution. The reservoir solution contained 500 μL of 200 mM
calcium acetate hydrate, 20% w/v PEG 3350 (PEG ION 28 Hampton Research PEG ION screen). Crystals were grown for one week and cryoprotected in a solution containing 10% glycerol and 90% reservoir solution. Data collection was conducted on MicroMax-007 Rigaku X-ray generator mounted on an R-axis IV++ detector. Diffraction data were scaled using HKL3000 [14]. The structure was solved using molecular replacement with H-Ras-GppNHp (PDB 3K8Y) as the initial phase model on PHENIX suite [15]. The structure was solved to 1.81 Å resolution with refinement statistics shown on table 4.1.

Wide-angle X-ray solution scattering

Protein expression, purification, and nucleotide exchange were done for H-RasY71A (construct of 166 residues) according to a previously published protocol [6]. GDP-bound protein was concentrated and exchanged into buffer for 5′-guanylyl imidodiphosphate (GppNHp) as described previously [16]. Purified H-RasY71A was exchanged into stabilization buffer (20 mM HEPES pH 7.5, 5 mM MgCl2, 50 mM NaCl, 1 mM DTT), concentrated to two different concentrations, 5 mg/mL and 10 mg/mL for WAXS experiments. For each sample, two measurements were conducted, one of protein in buffer solution and another one of the buffer solution by itself. Scattering data from the buffer was subtracted from the protein in buffer solution to get protein scattering data. Scattering data were collected at the G1 beam line at Cornell High Energy Synchrotron Source (CHESS). X-ray source was of 9.846 keV and lambda= 1.259 Å with a beam size of 250 μm * 250 μm [17]. Data collection was in the range of 0.009 < q < 0.072 Å. They were processed using the software RAW [18]. WAXS and MD data were combined to directly compare conformational flexibility of protein in solution seen in experimental setting and in silico. Initially, we use the software package XS [19] to calculate the
scattering of a rigid protein with the crystal structure that was used to start the MD simulations. The MD trajectory was converted to produce a sigma-r plot, a plot of intermolecular distances as a function of that distance [20]. Vector length convolution was used to obtain the calculated WAXS intensity, where each interatomic vector in the pair correlation function is replaced by a Gaussian distribution of vector lengths [21]. The two calculated intensities were then fitted on to the experimentally observed intensity for comparison.

Molecular Dynamics

Molecular Dynamics (MD) simulations were performed for 90 ns production runs for two systems, H-RasY71A-GTP \textit{(in silico} model\textit{)} and H-RasY71A-GTP (crystal structure) at the Northeastern Discovery Cluster (http://www.northeastern.edu/rc). The first starting structure was obtained from the X-ray crystallography structure of wild type H-Ras bound to GTP analogs, GppNHP, with PDB accession codes 3K8Y (H-Ras-GppNHP). The second structure is a novel crystal structure of HRasY71A-GppNHP. For the simulations, the nucleotide analogue molecules were converted to GTP by replacing the β-γ- bridging nitrogen with oxygen and the appropriate parameters for GTP were used. Crystallographic waters were retained for the simulation. NAMD and VMD softwares were used to prepare, perform and analyze all the simulations and CHARMM27 force fields were used [22-25]. Two systems were made and solvated in TIP3P water molecules extending 10 Å from the protein molecule. Charges in the systems were neutralized by adding sodium and chloride ions. The simulations started with energy minimization for 5,000 cycles, followed by gradual heating from 50 K to 300 K prior to the production runs. The periodic boundary condition was used at all three x, y, and z directions. Time step of 1 fs was applied during the production run at isothermal-isobaric condition of 300
K and 1 atm for the first 30 ns. Time step was increased to 2 fs for the subsequent production run for 60 ns. The switch function was used to calculate non-bonded interaction at long range between 10 and 11 Å. Electrostatic interactions were evaluated using Particle Mesh Ewald parameters [26].

Accelerated molecular dynamics (aMD) is a variation of conventional MD (cMD) where a bias potential is added to increase conformational sampling by enhancing the escape rates from potential wells. Our study follows the method described previously in the McCammon’s lab, where we apply dual boost potentials to the total potential energy and the dihedral energy [27, 28]. The form of the boost potential requires two parameters, E and α, to be specified based on equations determined in McCammon’s lab [27]. The first two nanoseconds of the simulation were done using cMD, where E and α were calculated and added to the configuration file to continue the subsequent production runs in aMD mode for up to 90ns simulation time.

**Conformational clustering**

Conformational clustering is embedded within the NAMD/VMD interface, where it performs a clustering analysis by finding clusters of time steps that are similar with respect to a given distance function for the atoms in selection [24]. This method is based on the quality threshold (QT) algorithm [29]. RMSD of the backbone was used as the distance function excluding hydrogen atoms. The number of clusters was determined to be five and the maximum distance value between two frames, or the cutoff distance, was selected to be 1.5 Å.
Results

Crystal structure of H-RasY71A mutant bound to GppNHp is found in the T-state conformation

The crystal structure of H-RasY71A is solved to 1.8 Å resolution, from crystal with symmetry of the space group P3_{2}2_{1}. Refinement statistics are shown in Table 4.1. Superimposition of H-RasY71A crystal structure on H-Ras wild type (PDB 1CTQ) of the same space group, P3_{2}2_{1}, shows that the elimination of Y71 side chain changes the overall hydrogen bond network from the allosteric site to the active site. H-RasY71A shows a shift on helix 3 toward switch II that resembles the T-state conformation. Without the Y71 side chain, residue R68 has more room to change conformations and shifts away from helix 3 to interact with the backbone of Q61 on switch II. Residue Y96 shifts slightly toward switch II, relative to the wild type, and forms a hydrogen bond with G60 on switch II, which is a signature of the T-state conformation (Figure 4.2). On the structure of H-RasY71A, we can trace a water-mediated hydrogen bond from the allosteric residue R97 to D69 on switch II (Figure 4.2). Residue R97 makes a hydrogen bond to a water molecule that interacts with K101, which in turn connects to E98. E98 on wild type H-Ras adopts a different conformation, facing away from the C-terminal portion of helix 3. E98 on the mutant H-RasY71A connects to R102 through a water molecule hydrogen bonding. R102 interacts with the carbonyl group of Q99 that bridges over to D69 on switch II through a hydrogen bond. Wild type H-Ras has a straight helix 3, where the side chain of R102 is 7.2 Å away from E98 and 5.4 Å away from Q99. The water mediated hydrogen bond network from R97 on helix 3 to D69 on switch II helps to shift the C-terminal portion of helix 3 toward switch II, which is representative of the T-state conformation. As a result, the side chains
of residues E62 and E63 are disordered on switch II of the mutant H-RasY71A-GppNHp crystal structure.

Figure 4.1. Crystal structures comparison of wild type H-Ras and H-RasY71A bound to GppNHp, both solved from crystals with symmetry P3\_2\_1. The model for wild type H-Ras (PDB 1CTQ) is shown in cyan and that of mutant H-RasY71A-GppNHp is shown in orange. The R- and T-states conformational differences are indicated with an arrow on helix 3. Residue side chains are shown in stick model to show their conformational changes from wild type to mutant H-RasY71A. A water mediated hydrogen bond can be traced from the allosteric site residue R97 to switch II residue D69 that helps to stabilize T-state in H-RasY71A mutant.
The H-RasY71A mutant has a higher hydrolysis rate constant than wild type H-Ras

Single turnover experiments of H-RasY71A mutant and H-Ras wild type are conducted to obtain rate constants for intrinsic hydrolysis of GTP on Ras. Hydrolysis assays are performed based on previously described protocol that measures the production of $^{32}$Pi over time [5, 10]. The radioactive nucleotide $^{32}$Pi-GTP (Perkin Elmer) is loaded on Ras to start the hydrolysis reaction and the reaction progress is measured by the release of $^{32}$Pi. Samples are removed from

---

**Table 4.1. H-RasY71A crystal structure refinement statistics**

<table>
<thead>
<tr>
<th></th>
<th>H-RasY71A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelength (Å)</td>
<td>1.54</td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>25.9 - 1.81 (1.875 -1.81)</td>
</tr>
<tr>
<td>Space group</td>
<td>P 3$_2$ 2 1</td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>39.223 39.223 160.113</td>
</tr>
<tr>
<td>α, β, γ</td>
<td>90 90 120</td>
</tr>
<tr>
<td>Total reflections</td>
<td>13323 (1341)</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>1341</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>96.29 (98.68)</td>
</tr>
<tr>
<td>I/ σ (I)</td>
<td>48.73 (24.92)</td>
</tr>
<tr>
<td>Wilson B-factor</td>
<td>18.52</td>
</tr>
<tr>
<td>R-work</td>
<td>0.1926 (0.2158)</td>
</tr>
<tr>
<td>R-free</td>
<td>0.2421 (0.3276)</td>
</tr>
<tr>
<td>Number of atoms</td>
<td>1462</td>
</tr>
<tr>
<td>protein</td>
<td>1304</td>
</tr>
<tr>
<td>ligands</td>
<td>34</td>
</tr>
<tr>
<td>water</td>
<td>124</td>
</tr>
<tr>
<td>RMS (bonds)</td>
<td>0.008</td>
</tr>
<tr>
<td>RMS (angles)</td>
<td>1.2</td>
</tr>
<tr>
<td>Ramachandran favored (%)</td>
<td>98</td>
</tr>
<tr>
<td>Ramachandran outliers (%)</td>
<td>0</td>
</tr>
<tr>
<td>Clash score</td>
<td>6.47</td>
</tr>
<tr>
<td>Average B-factor</td>
<td>20.5</td>
</tr>
</tbody>
</table>
reaction mixture throughout specific time points for two hours. The $^{32}$Pi is extracted using a previously described organic extraction method [11, 12]. Radioactivity of the collected samples is measured using a scintillation counter (300SL, LabLogic). The rate constants are obtained by plotting the concentration of $^{32}$Pi against reaction time using regression analysis provided in the software package DynaFit4 [13, 30, 31]. The kinetic experiments were done for the wild type and mutant proteins for comparison. The rate constant of $0.021 \pm 0.003 \text{ min}^{-1}$ obtained for the wild type protein agrees well with those previously published [10]. The H-RasY71A mutant has a higher rate constant of $0.039 \pm 0.006 \text{ min}^{-1}$, indicating that the hydrolysis reaction on the mutant Ras is twice as fast than on the wild type protein (Figure 4.3). This suggests that the mutation stabilizes state 2 and increases sampling of R-state to facilitate the intrinsic hydrolysis reaction.

Figure 4.3. Kinetics assays for intrinsic hydrolysis rate constants of H-RasY71A and wild type H-Ras. (a) Progress curve for single turnover hydrolysis experiments done in triplicates for H-RasY71A mutant. (b) Hydrolysis rate constants for H-RasY71A as compared to wild type H-Ras, where mutant Y71A has about double the rate constant of wild type.
Molecular dynamics show differences in simulation results between using crystal structure and in silico model as starting structures

In addition to studying the biochemical and structural properties of the Y71A mutant, I analyze a pairwise comparison of accelerated MD (aMD) simulations for the mutant Y71A by running a simulation with two different starting structures. One of the starting structures is the crystal structure of H-RasY71A (Y71A-cs) and the other being the wild type H-Ras structure that is modified in silico to make H-RasY71A (Y71A-model). Both simulations are performed under the same conditions and are compared to study the effects, if any, of starting structure on the dynamic properties of the protein from aMD simulations. Many previously published MD simulations have worked with in silico mutations, where the starting structure for the simulation originates from either the wild type crystal structure or a different mutant crystal structure [32, 33]. Specifically for Ras, Gorfe and colleagues in 2008 used a crystal structure of H-RasG12V-GDP bound to perform simulations for wild type H-Ras, N-Ras, and K-Ras in both GDP and GTP bound forms after performing in silico mutations [32], and in 2017 reported a simulation of WT K-Ras starting from the crystal structure of its G12D mutant with D mutated back to G in silico, with different results from those published in 2008. In 2015, Lu et al. performed MD simulations of K-Ras mutants G12C, G12D, G12V, G13D, E37K, Q61H, and R164Q, all using the crystal structure of K-RasQ61H (PDB 3GFT) as a starting structure to be mutated in silico [33]. The method of performing in silico mutations to represent a starting structure of a desirable mutant is the common practice in running MD simulations of mutants for which experimental structures are not available. However, our simulations data show that there are differences in the overall flexibility of regions of the protein depending on which starting structure is used as a starting conformation. When starting with the wild type structure that is modified and minimized
computationally, we see more similarities between this simulation and the one of wild type structure. The similarities are seen throughout important regions of the protein, namely switch I.

**Figure 1.4.** Changes in the RMSF (root-mean-square-fluctuations) for Ras residues to indicate differences between Y71A mutant and wild type, where Y71A-CS is shown in blue and Y71A-model is shown in orange. (b) Comparison of flexibility between Y71A-CS and wild type, where increases in fluctuation are shown in blue and decreases in fluctuations are shown in red. (c) Comparison of flexibility between Y71A-model and wild type.
inter-switch region containing loop 3, helix 3/loop 7 and helix 4 (Figure 4.4). Switch II, an important region for catalytic activity, and the site where residue Y71 resides, is the only region that shows differences between wild type and the mutant Y71A, suggesting that local changes due to the mutation are captured best in this simulation, with the structure sampling conformations in the vicinity of the global wild type structure. When starting the simulations with the crystal structure of H-RasY71A, we see more differences compared to wild type dynamics, focused on regions of previously noted biological importance, including switch I, the inter-switch loop 3, helix 3, loop 7, and helix 4, in addition to switch II, where the mutation resides (Figure 4.4). At switch II, Y71A-cs and Y71A-model have opposite effects when compared to the wild type H-Ras. Y71A-cs triggers an increase in flexibility at switch II, whereas Y71A-model causes a decrease in flexibility of switch II, possibly due to that area of the structure being stuck in deep potential energy well not overcome by the boost potential we apply during the relatively short 90 ns simulation time (Figure 4.4). Additionally, parts of the allosteric site, specifically loop 7 is more disordered in Y71A-cs than Y71A-model simulation. Overall, the mutant becomes globally more flexible relative to the wild type H-Ras protein, although there are areas, such as loop 3 and the C-terminal ends of helices 4 that become more rigid.

Although the Y71A-model simulation gives some hints of these changes, it is clear that within the 90-ns simulation time the Y71A-cs simulation starts from a more representative global minimum for the mutant and is therefore more likely to represent the changes in dynamics throughout the structure that results as a consequence of the mutation.

In order to better understand the differences in dynamics between Y71A-cs and Y71A-model affect access to the states associated with function, such as catalysis of GTP hydrolysis to GDP, we look into the conformational sampling throughout the simulation time. Y71A-cs
simulation starts with switch II in the T-state conformation as described above for the crystal structure of the mutant. On the other hand, Y71A-model starts with the R-state conformation that originates from the wild type H-Ras-GppNHp structure (PDB ID 3K8Y). We cluster the conformations of H-RasY71A into five groups based on rmsd distance function with a cutoff at 1.5 Å as described in the methods section [24]. The clusters for Y71A-cs and Y71A-model are then compared to those for the H-Ras wild type described in Chapter 3, focusing on the two major clusters that correspond to the T- and R-states (Table 4.2). The most populated cluster in all three simulations is the T-state, with 55% in the wild type, 58% in Y71A-cs, and 59% in Y71A-model. R-state population shows more difference amongst the three simulations, where in the wild type there is 26%, in Y71A-model it is 28%, and in Y71A-cs it is 35%. The increase in R-state population for Y71A mutant simulations, both from the in silico model and the crystal structure, suggests that the mutation somewhat decreases the energy barrier for conformational changes between R- and T-states relative to wild type H-Ras. More interestingly, the nearly 10% increase in accessibility to the R state when comparing the wild type simulation to the Y71A-cs simulation is consistent with the increase in hydrolysis rate constant that we observe for the mutant. While the T state is still the most accessed conformation as expected in the absence of ligand bound in the allosteric site [2], we find that conformations other than R and T states are more sampled in the wild type protein than in the mutant, resulting in the significant increase that we observe in accessibility of the R state important for hydrolysis.

The final structures from aMD simulation of Y71A-cs and Y71A-model adopt completely different switch II conformations, and this is also the case for helix 3 and loop 7 (Figure 4.5). The differences in the residue flexibility data, RMSF (Figure 4.4), shows that the
residues at switch II behave differently when comparing Y71A-cs to Y71A-model. In Y71A-model, residues

**Table 4.2.** Conformational clustering based on backbone rmsd distance function for H-Ras, H-RasY71A-model, and H-RasY71A-cs.

<table>
<thead>
<tr>
<th>Protein</th>
<th>T-state (%)</th>
<th>R-state (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-Ras</td>
<td>55</td>
<td>26</td>
</tr>
<tr>
<td>H-RasY71A-model</td>
<td>59</td>
<td>28</td>
</tr>
<tr>
<td>H-RasY71A-cs</td>
<td>58</td>
<td>35</td>
</tr>
</tbody>
</table>

Y64 on switch II has a 1.6 Å lower RMSF relative to wild type, which results in Y64 forming a tight hydrophobic interaction with I36 on switch I. On the other hand, in Y71A-cs, residue Y64 is more flexible with an RMSF 0.5 Å higher than in the wild type, allowing it to more easily flip toward the core of the protein to interact with helix 3 on the opposite site of switch II from switch I (Figure 4.5). This conformational change of residue Y64 is facilitated by the increase in flexibility of neighboring residues, including E63, which has an RMSF 1.3 Å higher than in the wild type Ras simulation. Additionally, residue R68 on switch II in Y71A-cs gains 2.4 Å increase in flexibility of its backbone that results in a different side chain conformation relative to wild type. Residue R68 is completely solvent exposed at the end of the Y71A-cs simulation, whereas it adopts the same conformation as wild type in Y71A-model, interacting with helix 3 residues (Figure 4.5). The flexibility of residue R68 in Y71A-model is the same as wild type H-Ras.

The increase in flexibility of switch II residues in Y71A-cs relative to wild type H-Ras, is also accompanied by an increase in flexibility of residues in loop 7 (Figure 4.4). Residues S106
and D107 experiences an increase in RMSF of 1.7 and 1.9 Å respectively. The communication network between the allosteric site, helix 3/loop 7 to switch II, is apparent in these data, where changes in flexibility occur in both regions and further changes the balance of conformational states between R-and T-state in Y71A mutant relative to wild type H-Ras (Table 2.4).

**Figure 4.5.** Final structures of from Y71A-cs and Y71A-model simulations. (a) Simulation started from a crystal structure of the mutant yields a different conformation for switch II and helix 3, where it opens the active site. (b) Simulation from a in silico model using wild type template, results in a similar conformation to a wild type simulation of H-Ras, with a closed active site.

WAXS data show better agreement with Y71A-crystal structure simulation than with Y71A-in silico model simulation results

The method of combining MD simulation data with solution scattering data has been described in the previous chapter (Figure 3.3). Briefly, MD trajectories from both simulations, Y71A-cs and Y71A-model are compared to X-ray solution scattering using sigma-R plot and vector length convolution [20, 21]. Calculated scattering patterns from MD simulations are
scaled up and fitted on to the observed WAXS data collected at CHESS (Cornell High Energy Synchrotron Source) [19, 34, 35]. The fitting errors and scaled factor values can directly relate *in silico* simulation to experimental data.

This method was used to determine whether the aMD simulation trajectories qualitatively represent the range of conformations sampled in solution and to compare the results from Y71A-cs and Y71A-model simulations with the experimental data. This was done in collaboration with the Makowski’s lab. Working with Hao Zhou, we conducted WAXS (wide angle X-ray solution scattering) experiments using the same protein construct that I used to crystallize H-RasY71A. We fit the calculated intensity data of flexible structures from Y71A-cs and Y71A-model simulations onto the experimentally observed intensity data for the mutant (Figure 4.6). The black curves in panels A and B are the scattering intensity calculated from the rigid crystal structure of the mutant. The blue curves are the calculated intensity of flexible structures from MD simulations, and the red curves are the experimentally observed data. The fitting errors of calculated to the observed data show that Y71A-cs has a better value of 2.0*10^{-2} as compared to Y71A-model with 3.0*10^{-2}, indicating a better fit of the simulations starting with the crystal structure of the mutant to the solution scattering data for the mutant. The scale factors for Y71A-cs and Y71A-model are 2.37 and 2.62, respectively. This is the factor by which the intensities calculated from the simulation trajectories need to be adjusted to maximize agreement with the experimental WAXS data. The lower value for Y71A-cs suggests that it shows a more similar flexibility to the observed data than Y71A-model. Both scale factors are greater than one, meaning that the H-RasY71A is more flexible in solution than is seen in either of the aMD simulations. This may be a consequence of the short simulation time. Comparing the WAXS
intensity data for H-RasY71A-GppNHp to wild type H-Ras-GppNHp shows that the mutant is more flexible than wild type in solution (Figure 4.7).

**Figure 4.6.** Plots showing solution scattering patterns of experimentally observed data (red), MD simulation calculated scattering pattern (blue), and rigid crystal structure calculated scattering pattern (black). (a) Comparison of Y71A-crystal structure simulation with the observed data. (b) Comparison of Y71A-homolgy model simulation with the observed data. The scale factor for Y71A-cs is lower than Y71A-model, which suggests that Y71A-cs simulation better mimics the protein behavior in solution.
Figure 4.7. Wide-angle X-ray solution scattering intensities for wild type H-Ras-GppNHp (red) superimposed on mutant H-RasY71A-GppNHp (blue). Mutant H-RasY71A is more flexible in solution than wild type H-Ras.

Discussion

Based on accelerated MD simulation analysis of wild type H-Ras-GppNHp, residue Y71 on switch II stands out as one of the most flexible side chains, spanning a conformational range sufficiently large to interact with Y96 on helix 3 on one side of switch II, and with E37 on switch I. In addition, this dynamic behavior of Y71 tends to accompany R- and T-state conformational transitions. These observations lead to further study of the role of residue Y71 in H-Ras. Here we present a crystal structure of H-RasY71A solved to 1.81 Å resolution and intrinsic hydrolysis rates for H-RasY71A relative to its wild type counterpart. Furthermore, we study the differences in dynamics of these proteins using accelerated MD simulations. The T-state conformation is stabilized in the crystal structure of H-RasY71A, showing a different H-bond network than the wild type connecting helix 3 to switch II. MD simulations show that switch I and switch II are more flexible in H-RasY71A mutant relative to the wild type H-Ras. Interestingly, intrinsic
hydrolysis assays give a higher rate constant for H-RasY71A than for wild type. We suggest that a higher hydrolysis rate constant is due to a reduction in energy barrier to cross from the T-state to the R-state in the Y71A mutant relative to wild type H-Ras.

Furthermore, we show that running MD simulation on the crystal structure of H-RasY71A is critical in getting results that are more reflective of the mutant behavior in solution. The overall dynamics of the in silico model of H-RasY71A (Y71A-model) is similar to wild type H-Ras, whereas we see more differences in the simulation of the crystal structure of H-RasY71A (Y71A-cs). WAXS results comparing wild type to mutant Y71A confirm that the mutation increases the overall flexibility of Ras. This result is only consistent with the simulation from the crystal structure Y71A-cs, where H-RasY71A shows increases in flexibility of switch II and loop 7 relative to wild type. On the other hand, Y71A-model simulation results suggests that switch II is more rigid than the wild type H-Ras, likely due to the conformation being stuck in a global minimum inherent in the wild type structure from which the simulation started after making the mutation in silico. The flexibility of switch II has been shown to play a critical role in activating and deactivating Ras for its signaling events [2, 36]. The discrepancy in the simulation results suggests that the treatment of starting structure is important to ensure that the simulation extends and converges to the correct conformations, particularly in cases where the simulations are relatively short. We showed in Chapter 3 that aMD simulations is able to at least partially compensate for the lack of sampling in short cMD simulations. Here we show that this is even more so when starting with the crystal structure of the mutant, which is already in a global minimum associated with the target structure for the simulations. A notable difference in the starting conformations between the crystal structure H-RasY71A and in silico model is that the crystal structure starts in the T-state conformation, whereas the in silico model starts in the R-
state conformation. The hydrogen bond networks that connect helix 3 to switch II are different between the R- and T-state, and this affects the flexibility of switch II. Therefore, we see differences in the conformational sampling between Y71A-cs and Y71A-model simulations. In both cases the trajectories sample R- and T-state populations, but these are sampled differently in each case, in terms of sampling frequency and in terms of the range of conformations for switch II. Our overall conclusion from this study is that, whenever possible, MD simulations should be conducted with appropriate experimental structures of the specific mutant to yield more accurate results.
Reference


Chapter 5: Ras/Raf Dimerization effects on allosteric networks
Introduction

Ras is a small GTPase that is involved in several signaling pathways to regulate cell proliferation, differentiation and apoptosis [1]. Ras can function as a molecular switch that cycles between GDP-inactive state and GTP-active state. This cycle is catalyzed by GEF (guanine nucleotide exchange factors) and GAP (GTPase activating proteins) [2]. Ras has been shown through many biophysical and cell biology experiments to form dimers and oligomers when it is in the active GTP-bound form [3-13]. Findings from our lab, with size exclusion chromatography (SEC) dimerization assays developed by Dr. Jillian Parker, show that Ras dimerizes when Raf-RBD and membrane phospholipid headgroups are simultaneously present in solution. In collaboration with Lee Makowski’s we collected small angle X-ray solution scattering (SAXS) data in an experimental setup where peaks elute from a SEC column straight into a cell for SAXS data collection (SEC-SAXS). The low resolution SAXS envelope obtained from the elution peak containing the dimer of the Ras/Raf-RBD complex fits the crystal structure of the dimer generated from the Ras/Raf-RBD complex (PDB ID 4G0N), which contains a monomer in the asymmetric unity (discussed in Chapter 2) [14]. The dimer is obtained by rotation about a two-fold symmetry axis, with a dimer interface composed of helices α4 and α5, also found in several crystal structures of Ras-GppNHp in the absence of Raf-RBD (eg. PDB ID 3K8Y) [15]. Mutations at this interface obstruct dimerization in solution [16], and impair signaling through the Ras/Raf/MEK/ERK pathway in cells and in mice [17]. We hypothesize that dimerization stabilizes the allosteric site near the membrane for Ca^{2+} binding, which in turn is linked to intrinsic hydrolysis of GTP on Ras [18]. Here I present the results of MD simulations of the Ras/Raf-RBD dimer generated from our crystal structure as described above (PDB 4G0N) [14].
The structure shows that Ras the dimer interface through interactions between helix 4, helix 5 and loop 3 (interswitch), are mainly electrostatics in nature (Figure 5.1). MD simulations of the Ras dimer have been conducted before, always in the absence of Raf-RBD. Furthermore, several dimer interface models have been proposed, and their simulations are in different contexts as compared to our simulations. The N-Ras dimer simulation conducted in 2012 by Guldenhaupt and colleagues, had a starting model generated by two-fold rotation from crystal structures in the PDB with helix 4 and helix 5 at the interface, but did not contain calcium and acetate in the allosteric site [7]. In 2015, Ras dimer simulations ran by the Nussinov lab showed different dimer interfaces as compared to ours, where in one model they had helix 3 and helix 4 interface, and in another model, they had β-sheet dimer interface between β2-strands of each Ras molecule [5]. The second model of β-strand interface blocks the interactions of Ras with effector proteins, including Raf and PI3K. The most recent Ras dimer simulation came from Gorfe’s lab, where they proposed two dimer interfaces, one similar to ours between helix 4 and helix 5, and another one between helix 3 and helix 4 [3]. However, their simulations did not contain Raf-RBD and calcium acetate molecules. The simulations presented in this chapter are extensions of those discussed in chapter 2 (Ras/Raf monomer studies), where we aim to see the effects of dimer formation on the dynamics of Ras and Raf molecules, as well as allosteric networks connecting the allosteric site bound to calcium acetate and the active site bound to GTP.

The importance of Ras dimer has been more emphasized recently after works done by Nan and colleagues in cells which show that Ras dimerization is a required step to activate Ras/Raf/MEK/ERK signaling pathway and drive oncogenic effects [4]. It demonstrates the therapeutic possibility of targeting Ras dimer for cancer therapy. Our existing hypothesis ties together the dimer formation with the allosteric modulation through the Ras/Raf/MEK/ERK
signaling pathway. The connection between Ras dimer and residues near the active site has been demonstrated previously through mutagenesis study of residue Y64A. Lin and colleagues show that mutant Y64A abolishes dimer formation [6]. Residue Y64 is at the effector lobe and it is directly involved in binding to effector and regulator proteins, including SOS (son of sevenless) and PI3K [19, 20]. In our dimer model, where the dimer interface is located at the allosteric lobe, helix 4 and helix 5, we suggest that residue Y64 is allosterically coupled to the dimer interface. We hypothesize that when Raf binds Ras at switch I, switch II is free to be modulated through allosteric network that stems from the dimer interface, at helix 4 and helix 5, as well as the allosteric site, helix 3 and loop 7. Dimer formation would modulate switch II to further activate hydrolysis of GTP, resulting in attenuated signaling through allosteric modulation by calcium binding at the allosteric site.

Figure 5.1. Ras/Raf dimer structure obtained from PDB 4G0N. Ras molecule 1 (Ras1) is shown in green and Ras molecule 2 (Ras2- cyan) was obtained by a two-fold rotation of the asymmetric
unity in the crystal structure containing the Ras/Raf-RBD monomer. The dimer interface consists of helix 4, helix 5 and interswitch (loop 3) regions. The allosteric site that consists of calcium acetate is located near the dimer interface and labeled. GppNHp (GNP) is shown in sticks at the active site of each Ras molecule. RafRBD molecules are shown in blue.

Methods

Molecular Dynamics

MD simulations of 90 ns production run were performed for Ras/Raf monomer and Ras/Raf dimer, where both structures were started from coordinates with PDB ID 4G0N [14]. The simulations were conducted at the Northeastern Discovery Cluster (http://www.northeastern.edu/rc). In each of the PDB files, GTP analogue, GppNHp, were modified to form GTP by replacing beta-gamma-bridging nitrogen atom with oxygen. Calcium acetate molecules were left at the allosteric site. All the crystallographic waters were included in the simulation. The simulations were prepared using VMD and performed on NAMD software [21, 22]. CHARMM27 force field was used for the simulation [23].

Dynamical Network Analysis

Dynamical network analysis is a general method used to obtain an accurate picture of network topology and long-range signaling in protein complexes derived from molecular dynamics simulations [24]. Each amino acid residue in the complex is assigned a node centered on its Cα atom and used as a base to construct significant regions of amino acid interactions and pathways of allosteric modulation that connect them. Edges are placed to connect the nodes between residues that remain within 4.5 Å distance for at least 75% of the simulation time. The edges are
weighted using pairwise correlation data calculated by the program Carma [25]. This information can then be mined to define community networks using the Girvan-Newman algorithm [26] and the diversity of paths that connect sites of functional significance in the complex. Nodes in the same community network can communicate with each other easily through multiple paths, whereas those in distinct community networks either do not communicate well or communicate through one or a small number of nodes essential for allosteric modulation.

**Results**

*Dimer formation affects the flexibility of residues at switch I, switch II and the allosteric site*

Classical MD simulations are conducted for 90 ns for each system, Ras/RafRBD monomer and dimer in explicit solvent. Dimer interface for Ras/RafRBD is located across the back of Ras protein, spanning helix 4, helix 5, and loop 3. Residues that are involved in dimer formation are shown in Figure 5.2, where most of them become more rigid upon dimer formation. At the interswitch region, β2/ loop3/ β3, eight residues are rigidified. On helix 5, four residues at the C-terminal end have reduced flexibility upon dimer formation. A different pattern is observed on helix4, where only two residues are rigidified, and five residues become more flexible upon dimer formation. Residue E49 on loop 3 that interacts with Q131 on helix 4 at the start of the simulation, changes conformation and forms a salt-bridge with R128 by the end of the simulation (Figure 5.3). This increases the flexibility of residue R128 relative to R128 in Ras/RafRBD monomer (Figure 5.2).
Calcium acetate at the allosteric site is retained throughout the simulation in both the monomer and dimer Ras/RafRBD simulations. However, the dimer formation increases the flexibility of the allosteric residues, R97 and Y137, and additionally five of their neighboring residues on helix 3. Changes in the flexibility of the allosteric residues affect the dynamics of residues on switch-II, changing the fluctuations of twelve residues. Seven of these residues are rigidified and five residues have higher flexibility upon dimer formation. Residue Y64 and its immediate neighboring residues, E62, E63, and S65 have reduced flexibility in the dimer simulation relative to the monomer (Figure 5.2). Residue Y64 has been shown previously to be a critical residue in the dimer formation, where its mutant Y64A abolished dimer formation [6]. In the monomer simulation, binding of RafRBD to Ras at switch I is shown to increase the flexibility of switch II residues and helped stabilize allosteric residues [14]. Comparing to the dimer simulation, we see that the binding of the second Ras molecule at helix 4, helix 5, and loop 3, increases the flexibility of the allosteric site residues and stabilizes most of switch II residues (Figure 5.2). Additionally, we see increases in the flexibility of switch I residues upon dimer formation. The electrostatic interactions at the interface of Ras and RafRBD are slightly altered during the dimer simulation (Figure 5.3b). Initially, residue E31 and D33 form a salt bridge with K84, and during the simulation they change conformation and form a salt bridge with R73 (Figure 5.3b).

The crystal structure of Ras/Raf-RBD dimer shows that residue D154 forms a hydrogen bond with D154 on the second Ras molecule. However, this interaction is quickly terminated at the start of MD simulation, where D154 on Ras-1 shifts up to form a salt-bridge with R161 on Ras-2. The same interaction is observed for D154 on Ras-2 that interacts with R161 on Ras-1.
Figure 5.2. Average root-mean-square-fluctuations (RMSFs) of Ras residues comparing Ras/RafRBD dimer to monomer. (A) Ras/Raf monomer is shown in solid line and Ras/Raf dimer is shown in dashed line. Protein secondary structures are labeled and shown along the x-axis of the plot. (B) back view of the protein (C) front view of the protein. Folded protein structure is shown in grey, with increases in fluctuations shown in blue and decreases in fluctuations shown in red (>0.25Å). Residues involved in dimer formation are labeled, including D47, E49, Y64, S127, Y131, R135, D154, and Q165.
Figure 5.3. Comparing initial and final conformations of residues at the dimer interface and the interface between Ras and RafRBD. (A) Ras dimer interface. Initial conformations are shown in cyan (Ras1) and pink (Ras2), final conformations are shown in green (Ras1) and red (Ras2). (B) Ras-RafRBD interface, initial Ras structure is in cyan and Raf in orange, and final structure of Ras is shown in green and Raf in purple.

(Figure 5.4A). These interactions are buried in the middle of the interface and have been shown through our mutagenesis studies that when abolished can significantly reduce dimer population by 52% (Table 5.1). Here it is worth pointing out that during the simulations D154 is negatively charged, while in the dimer in solution there may be a $pK_a$ shift resulting in protonation of the D154 side chain due to its proximity across the dimer interface. If this is the case, the results of the simulation need to be taken with caution, because incorrectly modeling the dimer interface may result in errors throughout the structure, as the dimer interface is allosterically connected to the active site and other areas of the protein.
Table 5.1. Relative dimer and monomer population (%) for Ras wild type and mutants. Data reproduced from Dr. Jillian Parker’s Ph.D. thesis [16].

<table>
<thead>
<tr>
<th>KRAS 1-173</th>
<th>ADDITIVE</th>
<th>MONOMER</th>
<th>DIMER</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>No glycerol</td>
<td>91</td>
<td>9</td>
</tr>
<tr>
<td>WT</td>
<td>2% glycerol</td>
<td>25</td>
<td>75</td>
</tr>
<tr>
<td>D154A</td>
<td>2% glycerol</td>
<td>77</td>
<td>23</td>
</tr>
<tr>
<td>R135A</td>
<td>2% glycerol</td>
<td>69</td>
<td>31</td>
</tr>
<tr>
<td>R164A</td>
<td>2% glycerol</td>
<td>64</td>
<td>36</td>
</tr>
<tr>
<td>E168A</td>
<td>2% glycerol</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>Y64A</td>
<td>2% glycerol</td>
<td>57</td>
<td>43</td>
</tr>
<tr>
<td>Q131A</td>
<td>2% glycerol</td>
<td>42</td>
<td>58</td>
</tr>
<tr>
<td>E49A</td>
<td>2% glycerol</td>
<td>35</td>
<td>65</td>
</tr>
<tr>
<td>T127A</td>
<td>2% glycerol</td>
<td>31</td>
<td>69</td>
</tr>
<tr>
<td>Q61L</td>
<td>No glycerol</td>
<td>37</td>
<td>63</td>
</tr>
<tr>
<td>Q61L</td>
<td>2% glycerol</td>
<td>47</td>
<td>53</td>
</tr>
</tbody>
</table>

On helix 4 of Ras-1, residue R135 forms a hydrogen bond with the backbone of R164 on helix 5 of Ras-2 (Figure 5.4B). This interaction was also shown to be critical in dimer formation, where Jillian’s mutagenesis study show that R135A mutant reduces dimer population by 44%. Also, on helix-4, residue Q131 forms H-bonds with E49 on the inter-switch loop and R164 on helix 5. These interactions hold together helix 4 of Ras-1 with the inter-switch loop 3 and helix 5 of Ras-2 (Figure 5.4C). Jillian’s mutagenesis study shows that without the side chain of Q161, there is only a slight decrease in dimer population, a reduction of 17%. It suggests that this interaction is not very critical for dimer formation. Similarly, a mutant of E49A only reduces dimer population by 10%. On the other hand, when residue R164 is mutated to an Alanine, the dimer population decreases by 39%. The more significant effect from residue R164 suggests that...
it is critical in stabilizing the H-bond cluster that is formed between Q131, R164, and E49 (Figure 5.4C). At loop 9 of Ras-1, the carbonyl backbone of residue G138 forms a hydrogen bond with Q165 on helix 5 of Ras-2 (Figure 5.4D). This interaction appears similar to the one formed between D154 and R161 that is located near the center of the interface. Residue G138 is adjacent to an allosteric site residue Y137 on helix 4 that shows an increased in flexibility due to the dimer formation (Figure 5.3). Overall, our MD simulations show that the dimer formation increases the flexibility of the allosteric site residues, which rigidifies switch-II residues and stabilizes Raf-RBD binding at switch-I relative to Ras/Raf-RBD monomer.

*Dynamic network analysis shows the effect of dimer formation that is similar to an oncogenic RasQ61L effect on the Ras/RafRBD allosteric network*

We use dynamical network analysis tool to study the differences in the connectivity between Ras/Raf-RBD monomer and dimer complexes. Dynamical network analysis is a tool that is developed by Sethi and colleagues at the University of Illinois at Urbana-Champaign to visualize and analyze pathways of allosteric communication in protein complexes [24]. In this analysis, amino acid residues are assigned as nodes that are connected by edges based on their interactions throughout the simulation time, with a minimum of 75% interaction time within 4.5 Å distance. Cross correlation data from MD simulation are incorporated into this network of nodes and edges to identify the shortest communication path from residues at the allosteric site of Ras and loop 4 of Raf-RBD [25]. Communities within the networks are calculated from the time averaged connectivity of the nodes using the Girvan-Newman algorithm [26]. A community corresponds to a set of residues that have stronger connections and move in concert with one
Figure 5.4. Dimer interface H-bonds as observed during MD simulations. Ras proteins are shown in cartoon representation, where Ras-1 is shown in green and Ras-2 is shown in cyan. Residues that form H-bonds at the interface are shown in sticks representation. (A) Residue D154 on Ras-1 forms a salt bridge with R161 on Ras-2, (B) Residue R135 forms an H-bond with the carbonyl backbone of R164. (C) Residue Q131 forms H-bonds with E49 and R164, (D) The carbonyl backbone of residue G138 forms an H-bond with Q165.
another. To compare Ras/Raf-RBD monomer complex with the dimer complex, we focus particularly on the connection between the allosteric site residues and the active site residues. Additionally, we also compare the dimer complex to the monomer RasQ61L/Raf-RBD. Comparing wild type dimer to monomer complexes, the community that connects Ras allosteric site residues, including R97, Q99 and D107, to the active site residues, is smaller in the dimer complex than in monomer complex (shown in orange) (Figure 5.5A, B). The population of this community is 43 residues in the monomer wild type, but there are only 24 residues in the wild type dimer complex. The active site residue, Q61 that has been shown to be directly affected by allosteric modulation with calcium acetate is included in a separate community for the dimer Ras/Raf-RBD complex, as opposed to being on the same community as shown for the monomer complex. It suggests that the connectivity between the allosteric site residues to the active site residue, Q61, is disturbed due to the dimer formation. A similar result is seen in the effect of the oncogenic mutant RasQ61L/Raf-RBD monomer complex, where residue L61 is excluded from the community that connects the allosteric site residues to the active site (shown in orange) (Figure 5.5C). Additionally, in the dimer wild type complex there is one community shown in grey that directly connects Ras/Raf-RBD interface to loop 4 (Figure 5.5). This is similar to what is seen in RasQ61L/Raf-RBD monomer, but it is different when compared to the wild type monomer. In the wild type monomer, there are two communities that connect Ras to loop L4 on Raf-RBD, they are shown in gray and pink (Figure 5.5B).

Using the dynamical network matrix, we calculate the suboptimal and optimal communication pathways leading from the allosteric site of Ras, residue R97, to loop 4 residue K109 on Raf-RBD [25]. We use this communication pathway to assess the strength of allosteric communication between the two sites on Ras and Raf-RBD. The wild type monomer Ras/Raf-
**Figure 5.5.** Community Networks formed in Ras/Raf-RBD complexes, in the dimer wild type, monomer wild type and mutant Q61L. (A) Dimer wild type complex. The allosteric site residues and active site residue Q61 are separated in the orange and purple communities. (B) Monomer wild type complex. The allosteric site residues and active site residue Q61 are in the same community shown in orange. (C) Mutant monomer RasQ61L/Raf-RBD complex. L61 is on a separate community shown in purple, and the allosteric site residues are in the orange community.
Figure 5.6. Optimal and suboptimal communication paths leading from the allosteric site of Ras to loop L4 of Raf-RBD in Ras/Raf-RBD dimer wild type, monomer wild type and mutant Q61L. Source node is R97 on Ras and the sink node is K109 on Raf-RBD. The optimal paths are shown in red and suboptimal paths are shown in blue. (A) There are 13 suboptimal paths on dimer Ras/Raf-RBD complex and the residues involved in the optimal path (red) are labeled. (B) There are 39 suboptimal paths in the monomer complex. (C) In the mutant RasQ61L/Raf-RBD monomer complex there are 6 suboptimal paths. The thickness of edges that connect the nodes are proportional to the number of suboptimal paths that cross their pathway of communication.

The dimer complex (A) and the mutant monomer (C) have thinner edges than the wild type monomer complex (B).
RBD complex contains 39 suboptimal paths connecting residue R97 on Ras to K109 on Raf-RBD, whereas dimer Ras/Raf-RBD only has 13 suboptimal paths and RasQ61L/Raf-RBD monomer has 6 suboptimal paths (Figure 5.6). This is indicative of weaker connectivity in the dimer complex than in the monomer complex, which is similar to the effect of oncogenic mutation RasQ61L. The optimal path, which is the shortest path, connecting R97 on Ras to K109 on Raf-RBD shows differences when comparing dimer to monomer. The optimal path in the monomer goes from helix 3 through switch II to switch I on the interface of Ras/Raf-RBD, whereas the dimer’s optimal path excludes switch II and goes through the core β sheet (Figure 5.6A, B). Interestingly, the optimal path in the dimer complex is the same as the one seen in RasQ61L/Raf-RBD monomer complex (Figure 5.6C).

**Discussion**

We show here that the dimer complex of Ras/Raf-RBD is intact throughout the 90-ns simulation time. The dimer interface consists of helix 4, helix 5 and loop 3, where their interactions remain stable throughout the MD simulation. Additionally, the binding of calcium acetate at the allosteric site is retained during the simulation. It has been reported previously that some Ras dimer models dissociate quickly during MD simulations due to weak protein-protein interactions, specifically that helix 3/helix 4 dimer interface is more stable than our model of helix 4/helix 5 dimer interface in the absence of Raf-RBD [3]. We reason that our dimer simulation is stable because of two factors, the incorporation of calcium acetate and the binding of Raf-RBDs. We hypothesize that the acetate molecule acts as a membrane mimic that can help stabilize the binding of two Ras molecules to form a dimer. The concentrating effect of
membrane in assisting dimer formation for Ras has been reported through many experimental data as well as MD simulations [3, 6, 7, 10]. Additionally, Raf-RBD binding on Ras can help stabilize state 2 conformation on Ras and through allosteric pathway affects the residues at the dimer interface. With this we can analyze the effect of dimer formation on the complex of Ras/Raf-RBD. We show that dimer formation rigidifies Ras residues at the interface of Ras/Raf-RBD, which suggests that it stabilizes the interaction between Ras and Raf-RBD. This may have an enhancing effect on signaling through the Ras/Raf/MEK/ERK pathway.

Dimer formation increases the flexibility of allosteric residues on Ras, including Y137 and R97, which in turn reduces the flexibility of switch II residues, including Y64. Through community network analysis, we see that allosteric-active sites communication is reduced, where the dimer simulation shows two different communities for allosteric site residues and the active site residue, Q61. Additionally, the communication pathway from Ras allosteric site residue R97 to residue K109 on loop 4 of Raf-RBD is also reduced in suboptimal paths and altered in the optimal path. Interestingly, this effect on allosteric network and communication pathway from Ras to Raf-RBD is similar to how it is shown in RasQ61L mutant [14]. We suggest that the dimer formation in the context of Ras/Raf-RBD complex might increase the activation of Ras/Raf/MEK/ERK pathway relative to the monomer complex.

We have started an exciting study of Ras/Raf-RBD dimer simulation here that can be followed up with more simulations to further look into the effects of Ras dimer on dynamics and catalysis. Accelerated MD can be used to increase conformational sampling for Ras/Raf-RBD dimer and to quantify changes in the R-and T- conformational balance, if there is any. We also see that there is D154-D154 head on side chain interactions from the crystal structure, which should be optimized with additional parameter for MD simulations. A protonated state of
aspartic acid might be needed to stabilize the dimer interface. It will also be interesting to study the effects of Ras/Raf-RBD dimer on oncogenic mutants of Ras, specifically RasQ61L/Raf-RBD for which we already have a crystal structure.
Reference


Appendix A:

Assessing Molecular Dynamics Force Fields for RNA simulations

Work conducted at

Vertex pharmaceuticals, Inc. 50 Northern Ave, Boston, MA 02210

In collaboration with Dr. Rebecca Swett
Introduction

Molecular dynamics (MD) simulation has become an important and powerful technique in biochemical research by providing invaluable structural and dynamical information at atomistic level that are often unobservable through conventional biophysics methodologies. The Chemistry Nobel Prize in 2013 awarded to Karplus, Warshel, and Levitt was one of the moments that marked the powerful nature of computational models of complex biochemical systems [1]. At the heart of MD simulations is the use of force fields along with their approximations to reproduce molecular geometry and properties of biomolecular structures. Thus, the reliability of MD simulation results is highly dependent on the quality of the force field that was used. Force fields for protein simulations have been very successful in properly reproducing molecular geometry and other biophysical properties that are consistent with experimental data. In 2011, twelve proteins that were structurally diverse, were successfully folded to their native structures by Shaw et al. using one force field and no biased sampling parameters [2]. On the other hand, nucleic acids force fields have been lagging and are underdeveloped [3-5]. Attempts to reversibly fold RNA tetraloop have also been done, but were much less successful [6, 7]. The best result was shown by Chen and coworkers at folding 8-mer RNA tetraloop motifs by extensively calibrating the force field parameters and running on replica exchange MD methods [8]. RNA force fields have been more difficult to parameterize and properly calibrate due to the high flexibility and large conformational dynamics of RNA polymers. In addition, electrostatics, which is one of the most complex terms to model in a force field, has a central role in stabilizing the highly-charged RNA molecules.
RNA-based therapeutics has seen some of the highest growth over the past few years, filling up R&D biotechnology pipeline. In 2015, nucleic acid therapeutics had 12% year-over-year growth, the second highest growth just behind antibody therapy [9]. Since the discovery of RNA interference technology from almost twenty years ago, RNAs have been viewed as promising therapeutic biomolecules to regulate gene and protein expression. As of 2017, there are six FDA-approved oligonucleotide therapies that have given positive clinical results [10]. RNA-based therapeutics utilize a diverse set of cellular mechanisms to treat diseases that are currently difficult to treat using small molecules, including various types of cancer, genetic diseases, diabetes, and Alzheimer’s. RNA-based therapeutic types include antisense oligonucleotide (ASO), small-interfering RNA (siRNA), microRNA (miRNA), aptamers, synthetic mRNAs and most recently CRISPR-Cas9, all show great potential in advancing this line of therapy in biopharmaceuticals [11, 12].

Advances in MD simulation force fields and MD engines have helped to integrate MD as one of the standard biophysical tools to study molecular motions and molecular interactions in many contexts, including drug design. In drug development, MD simulations can assist in structure based virtual screening, free energy perturbation (FEP) calculations for rational design, and allosteric modulation [13]. Using FEP calculations, Jorgensen et al. optimized a docking hit of $5 \mu M$ to a potent $55 \text{ pM}$ inhibitor of HIV reserve transcriptase [14]. Our view of the role of MD simulations in structure based drug design is that it plays an important role in designing and optimizing molecular interactions at atomistic level details. The lack of assessments on RNA force fields is concerning and can hinder the ability of computational chemists to properly conduct MD simulations on RNA biomolecules. Especially in a fast-paced biopharmaceuticals research setting. In this study, we hope to objectively assess and inform computational scientists
on best practices that can be applied quickly and reliably in running MD simulations on RNA biomolecules.

In this study, we select four RNA molecules of different sizes, functions and structures to assess three standard MD force fields, AMBER, CHARMM and OPLS. Additionally, we evaluate the effects of two neutralizing ions and two different water models in stabilizing RNA simulations. Although other studies have been done on comparing RNA force fields, they have been limited to using highly stable, small hairpin tetraloop structures [8, 15-21]. Those assessments, although useful, are limited in their ability to inform molecular modelers in choosing the most suitable force fields for their specific RNA simulations. In this study, we select one of the hairpin tetraloop structures as a baseline and add three other RNA molecules that are of relevance to biopharmaceuticals, including LCS1co miRNA-mRNA complex, siG12D RNA, and Zika Virus (ZIKV) sfRNA. LCS1co miRNA-mRNA complex comprises of let-7 miRNA and lin-41 mRNA molecules, where let-7 miRNA functions to regulate cell differentiation [22]. This miRNA has been proposed to be a potential biomarker and target in lung cancer therapy, since reduced levels of let-7 have been correlated with increased levels of oncogenes in lung cancer [23]. LCS1co is a designed construct that preserves the non-canonical interactions between let-7 miRNA and lin-41 mRNA, including an adenine bulge and an asymmetric internal loop, and the two strands are connected by a GAAA tetraloop. We also select an siRNA, siG12D, which is in phase II/b clinical trial under Silenseed pharmaceutical company to treat locally advanced pancreatic cancer (LAPC) [24-26]. It is a 19-mer duplex with eleven GC pairs. Lastly, we choose a viral RNA, ZIKV sfRNA, with a multi-pseudoknot structure that blocks cellular antiviral responses [27]. The biological assembly of this structure is a tetramer complex of 71-mer subunits that are held together by intersubunit interactions. With
these four structures, we aim to properly evaluate RNA force fields in AMBER ff12, CHARMM36, and OPLS3 on structures of considerably higher complexity than the standard tetraloop motifs. Our results suggest that OPLS3 is the best suited force field for RNA simulations, however it still has room to improve in terms of stabilizing the quaternary structure of ZIKV sfRNA. AMBER12 force field is second best with the best result in stabilizing ZIKV sfRNA quaternary structure, however it fails to stabilize the tertiary structure of LCS1co in our system. CHARMM36 has the most room to improve in terms of internal base pair fraying, opening of G-U wobble pairs, unstable duplex simulation and weak intersubunit electrostatics interactions. In this study, we evaluate the performances of common force fields to establish best practices for RNA simulations.

**Methods**

RNA structures were selected based on their relevance to biopharmaceuticals. The first structure was a 14-mer UUCG tetraloop motif with 5 stem base pairs (PDB 2KOC) that we wanted to use a baseline for our assessments [28]. The second structure was a 34-mer of let-7 miRNA and lin-41 mRNA complex (PDB 2KPV) [22]. Both were NMR structures with multiple conformations and we selected the last conformer from each set as our model structures. The third structure was modeled *de novo* using Chimera structure-building to form a 19-mer RNA duplex A-form, where the sequence was obtained from k-ras G12D siRNA used to treat locally advanced pancreatic cancer (LAPC) [24-26, 29]. The fourth structure was a crystal structure of 71-mer Zika virus sfRNA (PDB 5TPY), where we included all four subunits in the biological assembly with a total of 284 nucleotides [27]. Simulations with AMBER12 force field were prepared using MOE tools and production runs were done using NAMD engine with batch
scripts generated from MOE [30-34]. Structures were neutralized using sodium and magnesium ions according to their respective systems as shown in table 1. Neutralizing ions were placed before water molecules were added using MOE ion placement algorithm, where the ions were placed sequentially along the phosphate backbone of the RNAs [30]. Structures were solvated using TIP3P water models extending at least 10 Å from the solute atoms to form a cubic box [35]. Energy minimizations were done on each structure using conjugate gradient algorithm until a local minimum is reached. The structures were slowly heated up from 0 K to 300 K for 1000 ps with positional restraints. The restraints were released slowly during equilibration for 500 ps prior to starting the production runs. Periodic boundary conditions were applied at NPT ensemble using langevin thermostat and barostat at 300 K and 1 atm. Production runs were done with an integration time step of 1 fs. Non-bonded interactions were cut at 10 Å and particle mesh Ewald was activated to calculate the electrostatic interactions [36]. Light bonds constraints were applied to all the covalent bonds involving hydrogen atoms using LINCS algorithm [37].

Simulations with CHARMM27 and CHARMM36 force fields were prepared using CHARMM-GUI input generator and VMD [38-40]. All simulations using CHARMM36 force field were prepared using CHARMM-GUI, except for ZIKV sfRNA and one of the two siG12D simulations with sodium ions. CHARMM27 force field simulations were all prepared using VMD. Equilibration and production runs were done using NAMD engine [31]. Structures were neutralized using sodium and magnesium ions as described in table 1. Additionally, simulations prepared using VMD were set to have salt concentration of 0.15 M of either NaCl or MgCl2. Simulations of LCS1co and siG12D were conducted using two different initial ion position algorithms, one with CHARMM-GUI imbedded method of 2000 steps Monte Carlo simulations and another one with MOE initial ion configuration as described above [30, 38]. Structures were
then solvated in a cubic box using TIP3P water extending at least 10 Å from the solute [35]. The structures were minimized for 10,000 steps and equilibrated for 1,500 ps by heating up slowly from 50 K to 300 K prior to production runs. The integration time step was 1 fs. Periodic boundary conditions were applied and particle mesh Ewald was used to calculate the electrostatics potentials. Lennard-Jones (LJ) potentials were truncated at 10 Å. Covalent bonds involving hydrogens were constrained using SHAKE algorithm [41].

Simulations with OPLS3 force field were prepared using Maestro and ran on Desmond MD engine [42-45]. Structures were prepared by adding a water box extending at least 10 Å from the solute to form a cubic box and neutralized using their respective ions as shown on table 1. Initial ion placements were random. Salts, of either NaCl or MgCl2, were added to reach a concentration of 0.15 M. In addition to TIP3P, we conducted additional systems with TIP4P water models using OPLS3 force field [35]. Structures were minimized and equilibrated using Desmond NPT equilibration protocol that extends to 1500 ps [43]. Solutes were restrained at 10K and heated up to 300K by slowly removing restraints over time prior to the start of production runs. The integration time step was set to 1 fs using Langevin thermostat and non-bonded interactions were cutoff at 9 Å.

All the simulations were conducted using NVIDIA GPU computing nodes at Vertex Pharmaceuticals, Inc. High Performance Computing Cluster in Boston. Data analysis were done using VMD software [40]. The structures of the RNA molecules were evaluated on their average root mean square deviation (RMSD), root mean square fluctuations (RMSF) of the nucleotide residues, Watson-Crick (WC) base pairing, and conformational clustering. WC base pairings are calculated between base pair atoms N1-N3 distances. Conformational clustering is embedded within the VMD interface, where it performs a clustering analysis by finding clusters of time
steps that are similar with respect to a given distance function for the atoms in selection [40]. The
method is based on the quality threshold (QT) algorithm [46]. RMSD of the structures were used
as the distance function excluding hydrogen atoms. The number of clusters was determined to be
5 and the maximum distance value between two frames, or the cutoff distance, was determined
to be 1.5 Å.
Table 6.1. RNA structures and systems used for this study and their total simulation time (ns).

<table>
<thead>
<tr>
<th>RNA</th>
<th>Force field</th>
<th>Na(^+), TIP3P</th>
<th>Mg(^{2+}), TIP3P</th>
<th>Na(^+), TIP4P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tetraloop</td>
<td>AMBER12</td>
<td>300</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CHARMM27</td>
<td>300</td>
<td>30</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CHARMM36</td>
<td>300</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OPLS3</td>
<td>400</td>
<td>500</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>LCS1co</td>
<td>AMBER12</td>
<td>200</td>
<td>200</td>
<td>-</td>
</tr>
<tr>
<td>CHARMM27</td>
<td>120</td>
<td>30</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CHARMM36</td>
<td>400*</td>
<td>400*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OPLS3</td>
<td>200</td>
<td>200</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>siG12D</td>
<td>AMBER12</td>
<td>200</td>
<td>200</td>
<td>-</td>
</tr>
<tr>
<td>CHARMM36*</td>
<td>370*</td>
<td>200*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OPLS3</td>
<td>500</td>
<td>200</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>ZIKV sfRNA</td>
<td>AMBER12</td>
<td>50</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CHARMM36</td>
<td>85</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OPLS3</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Aggregate time from two simulations with different initial ion placements
Results

Our assessment of the force fields start by comparing the results from the four RNA systems running with sodium ions, then move to compare the effects of magnesium ions and TIP4P water model.

*Simulation of the tetraloop structure is stable under AMBER12 and OPLS3, but not under CHARMM27 or CHARMM36 force fields*

Our first model is a 14-mer hairpin tetraloop (GGCAC-UUCG-GUGCC) with five canonical WC base pairings and one G-U wobble pair at the loop. This structure has been simulated previously and found to be very stable under standard MD force fields [8, 15, 17, 19]. Indeed, we see the same trend with two of the force fields, AMBER12 and OPLS3, where this structure shows high stability throughout the simulation time. CHARMM force fields, CHARMM36 and CHARMM27 show high degree of instability for this structure. In CHARMM27, there are major openings of WC base pairs, especially at the termini and fraying of G-U wobble pair. CHARMM36 improves the stability of WC base pairs, however it fails to maintain G-U wobble pair at the tetraloop as shown in RMSF plot (Figure 6.1). It suggests that improvements made on the 2’-hydroxyl conformation for CHARMM36 from CHARMM27 help to stabilize canonical base pairings but not the non-canonical G-U wobble pair [47]. Long unbiased μs MD simulation using AMBER12 force field for 20 NMR ensembles of this structure has been done by Bergonzo and co-workers, where they found 15 stable structures and 5 unstable structures. The commonality between the 5 unstable simulations has to do with the opening of G-
U wobble pair at the tetraloop which is similar to what we observe for CHARMM36 force field simulation [15].

*Simulation of LCS1co structure is stable in OPLS3, but not in AMBER12, CHARMM27, and CHARMM36 force fields*

MD simulations of LCS1co structure can elucidate the role of two non-canonical interaction sites between let-7 miRNA and lin-41 mRNA on maintaining the stability of the complex. Force fields evaluation can show the extent to which each force field can maintain these non-canonical interactions. The nucleotide sequence of LCS1co (GGAGGUAGUAGUCGAAGACCGUUCUACACUCC) forms a stem-loop structure with an upper and a lower stem, an asymmetric internal loop, an adenine bulge, and a tetraloop. In our simulations, OPLS3 force field gives the most stable results through RMSD, RMSF, as well as WC base pairing assessments of LCS1co structure (Figure 6.2). Slightly higher average RMSD for OPLS3 is caused by the non-canonical interactions formed by residues at the asymmetric loop, 9, 10, 23 to 25 and the tetraloop residues 15-18, as well as the Adenine bulge, residue 30 (Figure 6.2a, b). OPLS3 maintains the A-form twist tertiary structure and has the lowest RMSD when only WC base pairing are included. CHARMM36 force field shows comparable RMSD and RMSF to OPLS3, however its WC base pairing shows a peak at 4.1 Å (Figure 6.2c), which corresponds to the opening two internal base pairs, G5-C29 and U6-A28, during the simulation. The base pairs fraying is caused by the non-canonical residue adenine bulge, A30, which hydrogen bonds to G5 and disrupts the G5-C29 base pairs. This internal base pair fraying eventually reforms during the 200ns simulation. AMBER12 force field shows the least stable RMSD and RMSF, however its WC base pairings are stable throughout the simulation (Figure
The higher RMSD in AMBER12 simulation corresponds to its inability to stabilize the A-form tertiary structure (Figure 6.2c). It becomes unstable due to the interactions with highly flexible residues at the internal loop. The internal loop residues in AMBER12 simulation are highly disordered with residue 9 reaching up to 10 Å RMSF and residues 10, 24 and 25 having an average of 7.5 Å RMSF. These highly flexible nucleobases interact with the neighboring residues which results in the destabilization the tertiary structure from its native twist. The final structure has an RMSD of 5.9 Å to the native structure, mainly because of the change in its tertiary structure. The conformation of the asymmetric internal loop can be divided into two families as described by NMR structures, where in both families U9 and A10 are stacked, but the residues G23, U24, and U25 adopt different conformations [22]. In family 1, G23 orients into the minor groove, U24 points into the helix and U25 flips out to the solvent. In family 2, G23 flips out to the solvent, U24 hydrogen bonds to A10, and U25 orients into the major groove. Our starting structure adopts family 1 conformation and this conformation is disrupted in AMBER12 force field simulation. During the first few nanoseconds of the simulation, the nucleobase of G23 forms a hydrogen bond with the 2’-hydroxyl ribose on G12 follows by U25 that forms hydrogen bond to the nucleobase of G11. It results in a major conformational change at the internal loop that allows for U9-U24 base pairing interactions. Conformational change at the internal loop negatively affects the overall twist of the A-form tertiary structure at the upper and lower stems. As shown in the RMSF plot, the internal loop residues are highly flexible and do not maintain this conformation throughout the 200ns simulation time (Figure 6.2b). The U9-U24 base pair breaks apart and U9 flips out to the solvent resulting in its high RMSF of 10 Å. Although, RMSF of the termini residues in AMBER12 simulation show higher values than OPLS3 and CHARMM36, it is not due to base pair fraying as evidenced by a unimodal distance distribution.
of N1-N3 in WC base pair plot (Figure 6.2d). The high flexibility of the termini residues is caused by the swaying of the helical structure during the conformational change. Additionally, these residues are unable to reform their native conformations during the 200ns simulation time, which disrupts the helical tertiary structure of LCS1co. NMR results for relaxation measurements of nucleobases $^{13}$C atoms in the internal loop suggest that there are large fast motions in the ps-ns time scale for these nucleobases, which is consistent with our RMSF data [22]. However, we expect to maintain the tertiary structure under nanoseconds simulation time, which does not happen in AMBER12 force field simulation.

*Simulation of siG12D duplex structure is most stable under OPLS3 and AMBER12 and unstable under CHARMM36*

The duplex structure of siG12D is stable because it contains eleven G-C base pairs and eight A-U base pairs, where all the base pairings are canonical WC pairs d(GUUGGAGCUGAUUGCGUAG). However as reported before, simulations of long oligonucleotides tend to experience base pairs opening at the end termini [48]. Our simulations also experience the base pair opening at the end termini, with three opening in OPLS3 simulation, four in AMBER12 and six in CHARMM36. In OPLS3 simulation, the broken base pairs are G1-C19, A18-U2, and G19-C1. AMBER12 has all those three base pairs open with an addition of U2-A18. CHARMM36 has all four of those termini base pairs opening with additional base pairs of U17-A3 and G16-C4. CHARMM36 also shows internal base pair opening between G5-C15 and G10-C10. Although these base pair openings are temporary and the final structures from all three force fields superimpose well to the native structure with, it is a
force field problem that should be fixed. Especially the internal base pair openings, which is rarely seen experimentally in the short time scales [49, 50].

OPLS3 force field once again shows the most stable simulation for a duplex RNA. The A-form twist structure of siG12D is maintained in OPLS3 and AMBER12 simulations but not in CHARMM36, which results in higher RMSD for CHARMM36 (Figure 6.3).

Intermolecular interaction in ZIKV sfRNA tetramer is only well maintained in AMBER12, but not in OPLS3 and CHARMM36 force fields

Electrostatic interactions between RNA molecules are important to consider in RNA simulations, because they play a role in holding the quaternary structures together and are important for target binding, recognitions and inter-domain communication pathways [51-55]. In our simulations, we test the extent to which each of the force fields can stabilize the intermolecular interactions between four subunits of RNA molecules in ZIKV sfRNA. Our result shows that only in AMBER12 force field these interactions are strong enough and properly described that the quaternary structure stays put throughout the simulation time (Figure 6.4).

Simulations with Magnesium ions introduce internal base pair fraying and rigid RNA structure for all three force fields

Magnesium ions have been shown to assist in the folding of RNA biomolecules through biochemical experiments [56], which leads us to assess its effect in our RNA simulations using all three force fields. Simulations with magnesium ions slightly increase stability of RNA molecules (Figure 6.5), but at the expense of some base pair fraying (Figure 6.6).
**Figure 6.1.** Tetraloop simulation comparing AMBER12 (purple), OPLS3 (red), and CHARMM36 (blue).

**Figure 6.2.** LCS1co simulation results comparing AMBER12 (green), CHARMM36 (black), OPLS3 (red) force field performances. All systems were neutralized using sodium ions and ran for 200ns simulation time. (A) All heavy atom RMSD difference over time, (B) Average RMSFs for each residue, (C) Final structures from each of the force field simulations superimposed onto the starting structure (grey), (D) Watson-Crick base pairs distance distribution.
Figure 6.3. Duplex siG12D simulation results comparing AMBER12 (green), CHARMM36 (black), OPLS3 (red) force field performances. All systems were neutralized using sodium ions and ran for 200ns simulation time. (A) All heavy atom RMSD difference over time, (B) Average RMSFs for each residue from both chains, chain A (o) and chain B (A), (C) Final structures superimposed onto the initial structure (grey), (D) Watson-Crick base pairs distance distribution.

Simulations with TIP4P water increases the stability of RNA biomolecules using OPLS3 force fields

Simulations with TIP4P water molecules using OPLS3 force field shows better stability when compared to TIP3P simulation using the same force field (Figure 6.7). The backboned rmsd are slightly more stable for tetraloop, LCS1co and siG12D duplex simulations.
Simulations with CHARMM36 using two different initial ions positioning dramatically affects the stability of RNA

CHARMM36 force field looks to be lacking in performance when comparing to AMBER12 and OPLS3 force fields for RNA simulation. In order to circumvent the problem, we run simulation of CHARMM36 using different initial ions positioning algorithms. The first one is the ions positioning from MOE, where ions are placed sequentially around the backbone of RNA at energetically optimal locations [30]. The other initial ion position was calculated using Monte Carlo short simulation that is provided through CHARMM GUI web developer, where the ions are placed based on scaled coulombic and VDW interactions [38]. Our results show that simulations with MOE initial ions placements significantly improve the stability of the RNA using CHARMM36 force fields (Figure 6.8).
Figure 6.5 LCS1co simulation results comparing sodium ions against magnesium ions simulations. AMBER12 with sodium (purple) and magnesium (green), CHARMM 36 with sodium (blue) and magnesium (green), OPLS3 with sodium (red) and magnesium (green). Simulation with Magnesium seem to slightly increase stability, but at the expense of base pair fraying.
Figure 6.6. Snapshots of simulations with magnesium ions, (A) LCS1co showing some base pair fraying due to magnesium ions that are bound tightly to the tertiary structure. Magnesium binding was permanent throughout the simulation time. (B) siG12D duplex structure, where magnesium binding changes the tertiary structure at the beginning of the simulation and holds it rigid throughout the simulation time. Three internal base pairs are broken in this structure.
Figure 6.7. TIP4P simulations using OPLS3 force fields compared with TIP3P. (A) Tetraloop simulation showing a slight stabilization when using TIP4P. (B) LCS1co simulation that also shows lower average RMSD using TIP4P than TIP3P. (C) siG12D duplex simulation showing better stability using TIP4P water model.
Figure 6.8. LCS1co simulations using CHARMM36 force fields with two different initial ions positioning algorithms. In blue is the simulations with MOE initial ions placement and in green is the simulation using CHARMM GUI initial ions placement.

Summary

Through our assessments we show that LCS1co with non-canonical interactions show better simulation results with OPLS3 and CHARMM36 than AMBER12. Duplex simulations of siRNA G12D with canonical base pairs show better simulation results with OPLS3 and AMBER12 than CHARMM36. More complex RNA structure, like ZIKV sfRNA show better simulation results with AMBER12 than OPLS3 or CHARMM36. Overall, OPLS3 shows the
best and most consistent results for all four types of RNA molecules that we run simulations on, from the smallest tetraloop to the biggest ZIKV sfRNA.

Additionally, we show that divalent ions, Mg$^{2+}$ can introduce instability to RNA structures by breaking base pair hydrogen bonds due to the tight binding to the backbone of RNAs. To slightly increase the stability of RNA simulations, one can change the water models to TIP4P. And to circumvent early instability of RNA simulations, one can choose better initial ion position algorithm.
References


Appendix B: Laboratory protocols for computational simulations
**Molecular dynamics simulation protocol** as performed on Unix machine using NAMD and VMD software packages. H-Ras-GppNHp structure is used as an example structure in this simulation preparation.

1. Download PDB (protein data bank) file 3K8Y
2. Download CHARMM topology file, top_all27_prot_lipid_na.inp
3. Launch VMD GUI
4. Load 3k8y.pdb on VMD. Click on File ➔ New Molecule … and browse to find the appropriate PDB file
5. Separate components of the PDB file to ease in the making of PSF (protein structure file).

   3K8Y was separated into six components, including:
   - Ras (166 residues)
   - GNP (32 atoms – convert to GTP using text file by changing the bridging Nitrogen atom N3B into O3B Oxygen atom)
   - Calcium, CA (2 atoms – change the name to CAL, because that is the nomenclature for CHARMM force field)
   - Magnesium, MG (2 atoms)
   - Acetate, ACT (1 molecule)
   - Water, SOLV (184 water)

6. Create PSF file using psfgen.tcl script. On VMD go to Extensions ➔ Tk Console menu item. In the folder where psfgen.tcl script resides, call the script by typing “source psfgen.tcl.” The content of the psfgen.tcl is as following:

   ```tcl
topology top_all27_prot_lipid_na.inp
pdbalias atom ILE CD1 CD
pdblalias residue HIS HSD
pdblalias residue HOH TIP3
```
pdbaliases atom HOH O OH2

segment RAS {pdb Ras.pdb
  first nter
  last cter
} segment GTP {pdb GTP.pdb}
segment MG {pdb MG.pdb}
segment CAL {pdb CAL.pdb}
segment ACT {pdb ACT.pdb}
segment SOLV {pdb SOLV.pdb}

# (6) Read protein coordinates from PDB file
coordpdb Ras.pdb RAS
coordpdb GTP.pdb GTP
coordpdb MG.pdb MG
coordpdb CAL.pdb CAL
coordpdb ACT.pdb ACT
coordpdb SOLV.pdb SOLV
guesscoord
writepsf 1HRas.psf
writepdb 1HRas.pdb

# End of psfgen commands

7. On VMD, load 1HRas.psf and load the new pdb file coordinate, 1HRas.pdb, into the PSF file.

8. Now you should see hydrogen bonds on your protein structure. It is time to solvate the protein. On VMD, go to Extensions → Modeling → add solvation box. Check the box for “rotate to minimize volume” and “use molecule dimension.” For box size, go to box padding and add 10 Angstrom for every direction, X, Y, and Z. Click solvate.

9. Add ions. On VMD, go to Extensions → Modeling → add ions. Autoionize with NaCl at physiological concentration 0.15 mol/L.

10. Check the dimension of your box system. On VMD go to Extensions → Analysis → PME electrostatics. In the box for “pad by = 0” add zero and click enclose. This should
tell you the dimension of your box and PME grid sizes. H-Ras is in a box of 63*61*57 A$^3$.

11. Now you have the complete system, a protein with solvated and neutralized in a water box, in a PSF and a PDB file. Exit VMD and transfer your files to a cluster. This simulation was performed on NIH Biowulf HPC cluster.

12. Configuration file for the system is needed to perform a simulation, as it contains specific as to how the simulation is to be performed. H-Ras configuration file for minimization is as following:

```
# NAMD configuration file for H-Ras minimization

# molecular system
structure cell-final_xplor.psf
coordinates cell-final.pdb

# force field
paratypecharmm on
parameters /gs1/users/mabuyong/src/charmm/toppar/par_all27_prot_na.prm
exclude scaled1-4
1-4scaling 1.0

# approximations
switching on
switchdist 10
cutoff 11
pairlistdist 12
margin 1.0
stepspercycle 20
nonbondedFreq 2

cellBasisVector1 63.0 0.0 0.0
cellBasisVector2 0.0 62.0 0.0
cellBasisVector3 0.0 0.0 57.0
cellOrigin 0.0 0.0 0.0

#protocol
temperature 0 #initial t
seed 78133
rigidBonds all
```
reassignFreq 1000
timestep 1.0

langevin on
langevinDamping 10
langevinTemp 50.0

# Constant Pressure Control (variable volume)
useGroupPressure yes ;# needed for rigidBonds
useFlexibleCell no
useConstantArea no

langevinPiston on
langevinPistonTarget 1.01325 ;# in bar -> 1 atm
langevinPistonPeriod 100.
langevinPistonDecay 50.
langevinPistonTemp 50.0

reassignTemp 0
reassignIncr 5
reassignHold 50

# output
DCDunitcell on
wrapall on
outputname abeta-run.out # output PDB file
restartname abeta-run.rst
DCDfile abeta-run.dcd
restartfreq 1000
outputenergies 5000
outputtiming 5000
dcdfreq 5000
binaryoutput no

minimize 5000
run 20000

13. H-Ras configuration file for heating up and equilibration are as following:

# NAMD configuration file for H-Ras heat 1

# molecular system
structure cell-final_xplor.psf
coordinates abeta-run.out.coor
bincoordinates abeta-run.rst.coor
binvelocities abeta-run.rst.vel
extendedsystem  abeta-run.rst.xsc
firsttimestep  0

# force field
paratypecharmm  on
parameters  /gs1/users/mabuyong/src/charmm/toppar/par_all27_prot_na.prm

exclude  scaled1-4
1-4scaling  1.0

# approximations
switching  on
switchdist  10
cutoff  11
pairlistdist  12
margin  1.0
stepspercycle  20

#PME
PME yes
PMEGridSizeX  63
PMEGridSizeY  61
PMEGridSizeZ  57
FFTWUseWisdom no
nonbondedFreq 2

#protocol
#temperature  0   #initial t
seed  78133
rigidBonds  all
reassignFreq  1000
timestep  1.0

langevin  on
langevinDamping  10
langevinTemp  100.0

# Constant Pressure Control (variable volume)
useGroupPressure  yes ;# needed for rigidBonds
useFlexibleCell  no
useConstantArea  no

langevinPiston  on
langevinPistonTarget  1.01325 ;# in bar => 1 atm
langevinPistonPeriod  100.
langevinPistonDecay  50.
langevinPistonTemp  100.0

reassignTemp  50
reassignIncr  2
reassignHold  100

# output
DCDunitcell on
wrapall on
outputname    abeta-heat.out  # output PDB file
restartname   abeta-heat.rst
DCDfile       abeta-heat.dcd
restartfreq   1000
outputenergies  5000
outputtiming   5000
dcdfreq        5000
binaryoutput   no

numsteps = 50001

# NAMD configuration file for H-Ras heat 2

# molecular system
structure    cell-final_xplor.psf
coordinates  abeta-run.out.coor
bincoordinates  abeta-heat.rst.coor
binvelocities   abeta-heat.rst.vel
extendedsystem  abeta-heat.rst.xsc
firsttimestep  0

# force field
paratypecharm on
parameters /gs1/users/mabuyong/src/charmm/toppar/par_all27_prot_na.prm
exclude scaled1-4
1-4scaling  1.0

# approximations
switching on
switchdist   10
cutoff 11
pairlistdist  12
margin  1.0
stepspercycle  20

#PME
PME yes
PMEGridSizeX 63
PMEGridSizeY 61
PMEGridSizeZ 57
FFTWUseWisdom no
nonbondedFreq 2

#protocol
seed 78133
rigidBonds all
  reassignFreq 1000
timestep 1.0

langevin on
langevinDamping 10
langevinTemp 250.0

# Constant Pressure Control (variable volume)
useGroupPressure yes ;# needed for rigidBonds
useFlexibleCell no
useConstantArea no

langevinPiston on
langevinPistonTarget 1.01325 ;# in bar -> 1 atm
langevinPistonPeriod 100.
langevinPistonDecay 50.
langevinPistonTemp 250.0

reassignTemp 100
reassignIncr 2
reassignHold 250

# output
DCDunitcell on
wrapall on
outputname abeta-heat2.out # output PDB file
restartname abeta-heat2.rst
DCDfile abeta-heat2.dcd
restartfreq 1000
outputenergies 5000
output_timing 5000
dcdfreq 5000
binaryoutput no

numsteps = 150001
14. Production run is started once we reach 300 K temperature. The configuration file is as following:

```
# NAMD configuration file for H-Ras prun 1

# molecular system
structure       cell-final_xplor.psf
coordinates abeta-run.out.coor
bincoordinates abeta-heat2.rst.coor
binvelocities abeta-heat2.rst.vel
extendedsystem abeta-heat2.rst.xsc
firsttimestep 0

# force field
paratype charmm on
parameters /home/guterres.h/Toppar/par_all27_prot_na.prm

exclude scaled1-4
1-4scaling 1.0

# approximations
switching on
switchdist 10
cutoff 11
pairlistdist 12
margin 1.0
stepspercycle 20

# PME
PME yes
PMEGridSpacing 1.0
FFTWUseWisdom no
nonbondedFreq 2

# protocol
#temperature 0 initial t
seed 78133
rigidBonds all
reassignFreq 1000
timestep 1.0

langevin on
langevinDamping 10
langevinTemp 300.0
```
# Constant Pressure Control (variable volume)
useGroupPressure yes ;# needed for rigidBonds
useFlexibleCell no
useConstantArea no

langevinPiston on
langevinPistonTarget 1.01325 ;# in bar -> 1 atm
langevinPistonPeriod 100.
langevinPistonDecay 50.
langevinPistonTemp 300.0

reassignTemp 250
reassignIncr 1
reassignHold 300

# output
DCDunitcell on
wrapall on
outputname abeta-prun1.out # output PDB file
restartname abeta-prun1.rst
DCDfile abeta-prun1.dcd
restartfreq 5000
outputenergies 5000
outputtiming 5000
dcdfreq 5000
binaryoutput no

# 0.5 ns a chunk
numsteps = 500001

15. The production run at 1 femtoseconds timestep was ran for 30 ns with 60 configuration files. The next 60 production runs were conducted at 2 femtoseconds timestep.
Data processing and analysis. Many of data analysis were done using Tc/Tk console on VMD.

The scripts are as following:

1. **Root mean squared deviation (rmsd)** script (rmsd.tcl)

   ```
   package require psfgen
   package require molefacture
   package require alchemify
   package require readcharmmtop
   package require pbctools

   resetpsf
   pdbalias residue HIS HSD
   pdbalias atom ILE CD1 CD

   topology /gs1/users/mabuyong/src/charmm/toppar/top_all27_prot_lipid_na.inp

   mol new ../cell-final_xplor.psf
   mol addfile ../cell-final.pdb
   mol addfile ../abeta-prun1.dcd type dcd waitfor all
         set cell [pbc get -now]
   mol addfile ../abeta-prun2.dcd type dcd waitfor all
   mol addfile ../abeta-prun3.dcd type dcd waitfor all
   mol addfile ../abeta-prun4.dcd type dcd waitfor all
   mol addfile ../abeta-prun5.dcd type dcd waitfor all
   mol addfile ../abeta-prun6.dcd type dcd waitfor all
   mol addfile ../abeta-prun7.dcd type dcd waitfor all
   mol addfile ../abeta-prun8.dcd type dcd waitfor all
   mol addfile ../abeta-prun9.dcd type dcd waitfor all
   mol addfile ../abeta-prun10.dcd type dcd waitfor all
   mol addfile ../abeta-prun11.dcd type dcd waitfor all
   mol addfile ../abeta-prun12.dcd type dcd waitfor all
   mol addfile ../abeta-prun13.dcd type dcd waitfor all
   mol addfile ../abeta-prun14.dcd type dcd waitfor all
   mol addfile ../abeta-prun15.dcd type dcd waitfor all
   mol addfile ../abeta-prun16.dcd type dcd waitfor all
   mol addfile ../abeta-prun17.dcd type dcd waitfor all
   mol addfile ../abeta-prun18.dcd type dcd waitfor all
   mol addfile ../abeta-prun19.dcd type dcd waitfor all
   mol addfile ../abeta-prun20.dcd type dcd waitfor all
   mol addfile ../abeta-prun21.dcd type dcd waitfor all
   mol addfile ../abeta-prun22.dcd type dcd waitfor all
   mol addfile ../abeta-prun23.dcd type dcd waitfor all
   mol addfile ../abeta-prun24.dcd type dcd waitfor all
   mol addfile ../abeta-prun25.dcd type dcd waitfor all
   ```
mol addfile ../abeta-prun26.dcd type dcd  waitfor all
mol addfile ../abeta-prun27.dcd type dcd  waitfor all
mol addfile ../abeta-prun28.dcd type dcd  waitfor all
mol addfile ../abeta-prun29.dcd type dcd  waitfor all
mol addfile ../abeta-prun30.dcd type dcd  waitfor all
mol addfile ../abeta-prun31.dcd type dcd  waitfor all
mol addfile ../abeta-prun32.dcd type dcd  waitfor all
mol addfile ../abeta-prun33.dcd type dcd  waitfor all
mol addfile ../abeta-prun34.dcd type dcd  waitfor all
mol addfile ../abeta-prun35.dcd type dcd  waitfor all
mol addfile ../abeta-prun36.dcd type dcd  waitfor all
mol addfile ../abeta-prun37.dcd type dcd  waitfor all
mol addfile ../abeta-prun38.dcd type dcd  waitfor all
mol addfile ../abeta-prun39.dcd type dcd  waitfor all
mol addfile ../abeta-prun40.dcd type dcd  waitfor all
mol addfile ../abeta-prun41.dcd type dcd  waitfor all
mol addfile ../abeta-prun42.dcd type dcd  waitfor all
mol addfile ../abeta-prun43.dcd type dcd  waitfor all
mol addfile ../abeta-prun44.dcd type dcd  waitfor all
mol addfile ../abeta-prun45.dcd type dcd  waitfor all
mol addfile ../abeta-prun46.dcd type dcd  waitfor all
mol addfile ../abeta-prun47.dcd type dcd  waitfor all
mol addfile ../abeta-prun48.dcd type dcd  waitfor all
mol addfile ../abeta-prun49.dcd type dcd  waitfor all
mol addfile ../abeta-prun50.dcd type dcd  waitfor all
mol addfile ../abeta-prun51.dcd type dcd  waitfor all
mol addfile ../abeta-prun52.dcd type dcd  waitfor all
mol addfile ../abeta-prun53.dcd type dcd  waitfor all
mol addfile ../abeta-prun54.dcd type dcd  waitfor all
mol addfile ../abeta-prun55.dcd type dcd  waitfor all
mol addfile ../abeta-prun56.dcd type dcd  waitfor all
mol addfile ../abeta-prun57.dcd type dcd  waitfor all
mol addfile ../abeta-prun58.dcd type dcd  waitfor all
mol addfile ../abeta-prun59.dcd type dcd  waitfor all
mol addfile ../abeta-prun60.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun1.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun2.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun3.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun4.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun5.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun6.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun7.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun8.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun9.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun10.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun11.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun12.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun13.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun14.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun15.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun16.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun17.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun18.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun19.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun20.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun21.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun22.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun23.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun24.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun25.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun26.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun27.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun28.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun29.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun30.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun31.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun32.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun33.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun34.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun35.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun36.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun37.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun38.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun39.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun40.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun41.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun42.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun43.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun44.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun45.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun46.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun47.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun48.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun49.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun50.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun51.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun52.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun53.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun54.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun55.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun56.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun57.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun58.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun59.dcd type dcd  waitfor all
mol addfile ../run2/abeta-prun60.dcd type dcd  waitfor all

 pbc unwrap -sel "all not water"

# Prints the RMSD of the protein atoms between each timestep
# and the first timestep for the given molecule id (default: top)
proc print_rmsd_all {{mol top}} {
    set outDataFile [open RMSDout w]
    # use frame 0 for the reference
    set reference [atomselect $mol "protein" frame 0]
    # the frame being compared
    set compare [atomselect $mol "protein"]
    set num_steps [molinfo $mol get numframes]
    for {set frame 0} {$frame < $num_steps} {incr frame} {
        # get the correct frame
        $compare frame $frame
        # compute the transformation
        set trans_mat [measure fit $compare $reference]
        # do the alignment
        $compare move $trans_mat
        # compute the RMSD
        set rmsd [measure rmsd $compare $reference]
        # print the RMSD
        puts "RMSD of $frame is $rmsd"
        puts $outDataFile "$frame $rmsd"
    }
    close $outDataFile
    return
}

print_rmsd_all
exit

2. **Root mean squared fluctuations** (Average RMSD of the amino acid residues – rmsf.tcl)

script.

package require psfgen
package require molefacture
package require alchemify
package require readcharmmtop
package require pbctools

resetpsf
pdbalias residue HIS HSD
pdbalias atom ILE CD1 CD

topology /gs1/users/mabuyong/src/charmm/toppar/top_all27_prot_lipid_na.inp

mol new ../cell-final_xplor.psf
mol addfile ../cell-final.pdb
mol addfile ../abeta-prun1.dcd type dcd waitfor all
   set cell [pbc get -now]
mol addfile ../abeta-prun2.dcd type dcd waitfor all
.
.
oml addfile ../run2/abeta-prun60.dcd type dcd waitfor all

   pbc unwrap -sel "all not water"

set sel_resid {[atomselect top "segid RAS and alpha"] get resid]}

# Prints the RMSF of the protein atoms between each timestep
# and the first timestep for the given molecule id (default: top)
proc rmsd_residue_over_time {{mol top} res} {

   # use frame 0 for the reference
   set reference [atomselect $mol "segid RAS" frame 0]
   # the frame being compared
   set compare [atomselect $mol "segid RAS"]
   # get the number of frames
   set num_steps [molinfo $mol get numframes]
   # open file for writing
   set fil [open RMSFout w]

   foreach r $res {
       set rmsd($r) 0
   }

   # loop over all frames in trajectory
   for {set frame 0} {$frame < $num_steps} {incr frame} {
       puts "Calculating rmsd for frame $frame"
       # get the correct frame
       $compare frame $frame

       set rmsd($r) 0
       foreach a [get $r] {
           set dist [dist $a $compare]
           set rmsd($r) [expr $rmsd($r) + $dist**2]
       }

       # average over all residues
       set rmsd($r) [expr $rmsd($r)/$num_res]

       # write to file
       puts $fil "Frame $frame Residue $r $rmsd($r)"
   }

   close $fil
}
# compute the transformation
set trans_mat [measure fit $compare $reference]
# do the alignment
$compare move $trans_mat

# compute the RMSD
# loop through all residues
foreach r $res {
    set ref [atomselect $mol "segid RAS and resid $r and noh"
    frame 0]
    set comp [atomselect $mol "segid RAS and resid $r and noh" frame $frame]
    set rmsd($r) [expr $rmsd($r) + [measure rmsd $comp $ref]]
    $comp delete
    $ref delete
}

set ave 0
foreach r $res {
    set rmsd($r) [expr $rmsd($r)/$num_steps]
    set ave [expr $ave + $rmsd($r)]
}

set ave [expr $ave/[llength $res]]
puts "Average rmsd per residue: $ave"
close $fil
}

return

rmsd_residue_over_time top $sel_resid

exit

3. **Distance over time between two atoms.** This is useful to quantify hydrogen bonds, salt bridge, etc. This script was for K-Ras salt bridge between helix 3 and switch II residues,
as presented in chapter 3. The output files are a distance list and a distance distribution

histogram, which are very useful to properly quantify hydrogen bond interactions.

package require psfgen
package require molefacture
package require alchemify
package require readcharmmtop
package require pbctools

resetpsf
pdbalias residue HIS HSD
pdbalias atom ILE CD1 CD
topology /home/guterres.h/Toppar/top_all27_prot_lipid_na.inp

mol new ../cell-final_xplor.psf
mol addfile ../cell-final.pdb
mol addfile ../abeta-prun1.dcd type dcd waitfor all
    set cell [pbc get -now]
.
.
mol addfile ../run2/amd-prun60.dcd type dcd waitfor all

    pbc unwrap -sel "all not water"

proc distance {seltext1 seltext2 N_d f_r_out f_d_out} {

    set sel1 [atomselect top "$seltext1"]
    set sel2 [atomselect top "$seltext2"]

    set nf [molinfo top get numframes]
    # Loop over all frames.
    set outfile [open $f_r_out w]
    for {set i 0} {$i < $nf} {incr i} {

        puts "frame $i of $nf"
        $sel1 frame $i
        $sel2 frame $i

        set com1 [measure center $sel1 weight mass]

        set com2 [measure center $sel2 weight mass]
set com2 [measure center $sel2 weight mass]

set simdata($i.r) [veclength [vecsub $com1 $com2]]
puts $outfile "$i $simdata($i.r)"
}
close $outfile

# Obtain the distribution.
set r_min $simdata(0.r)
set r_max $simdata(0.r)
for {set i 0} {$i < $nf} {incr i} {
set r_tmp $simdata($i.r)
if {$r_tmp < $r_min} {set r_min $r_tmp}
if {$r_tmp > $r_max} {set r_max $r_tmp}
}
set dr [expr ($r_max - $r_min) /($N_d - 1)]
for {set k 0} {$k < $N_d} {incr k} {
set distribution($k) 0
}
for {set i 0} {$i < $nf} {incr i} {
set k [expr int(($simdata($i.r) - $r_min) / $dr)]
incr distribution($k)
}
set outfile [open $f_d_out w]
for {set k 0} {$k < $N_d} {incr k} {
puts $outfile "[expr $r_min + $k * $dr] $distribution($k)"
}
close $outfile

#distance for helix 3 residues and switch II residues
distance "segid KRAS and resid 62 and name OE1" "segid KRAS and resid 88 and name NZ" 100 Dis-list-E62-K88a.txt Dis-dist-E62-K88a.txt
distance "segid KRAS and resid 62 and name OE2" "segid KRAS and resid 88 and name NZ" 100 Dis-list-E62-K88b.txt Dis-dist-E62-K88b.txt
distance "segid KRAS and resid 68 and name NH1" "segid KRAS and resid 92 and name OD1" 100 Dis-list-R68-D92a.txt Dis-dist-R68-D92a.txt
distance "segid KRAS and resid 68 and name NH2" "segid KRAS and resid 92 and name OD1" 100 Dis-list-R68-D92b.txt Dis-dist-R68-D92b.txt
distance "segid KRAS and resid 68 and name NH1" "segid KRAS and resid 92 and name OD2" 100 Dis-list-R68-D92c.txt Dis-dist-R68-D92c.txt
distance "segid KRAS and resid 68 and name NH2" "segid KRAS and resid 92 and name OD2" 100 Dis-list-R68-D92d.txt Dis-dist-R68-D92d.txt
distance "segid KRAS and resid 69 and name OD1" "segid KRAS and resid 102 and name NH1" 100 Dis-list-D69-R102a.txt Dis-dist-D69-R102a.txt
distance "segid KRAS and resid 69 and name OD2" "segid KRAS and resid 102 and name NH1" 100 Dis-list-D69-R102b.txt Dis-dist-D69-R102b.txt
distance "segid KRAS and resid 69 and name OD1" "segid KRAS and resid 102 and name NH2" 100 Dis-list-D69-R102c.txt Dis-dist-D69-R102c.txt
distance "segid KRAS and resid 69 and name OD2" "segid KRAS and resid 102 and name NH2" 100 Dis-list-D69-R102d.txt Dis-dist-D69-R102d.txt
distance "segid KRAS and resid 60 and name O" "segid KRAS and resid 96 and name OH" 100 Dis-list-G60-Y96.txt Dis-dist-G60-Y96.txt

4. **RMSD conformational clustering, with a cutoff 1.5 Å between frames.** This analysis was performed in chapter 3 and 4.

source fep.tcl
package require psfgen
package require molefacture
package require alchemify
package require readcharmmtop
resetpsf
pdbalias residue HIS HSD
pdbalias atom ILE CD1 CD
topology top_all27_prot_lipid_na.inp

mol new ../cell-final_xplor.psf
mol addfile ../cell-final.pdb
mol addfile ../abeta-prun1.dcd type dcd waitfor all
    set cell [pbc get -now]
.
.
.

mol addfile ../run2/amd-prun60.dcd type dcd waitfor all

    pbc unwrap -sel "all not water"
animate delete beg 0 end 0

set reference [atomselect top "protein and backbone and noh" frame 0]
set num_beg [molinfo top get numframes]
set compare [atomselect top "protein and backbone and noh"]
set all [atomselect top "all"]

for {set frame 0} {$frame < $num_beg} {incr frame} {
    # get the correct frame
    $compare frame $frame
    # compute the transformation
    set trans_mat [measure fit $compare $reference]
    # do the alignment
    animate goto $frame
    # $compare move $trans_mat
    $all move $trans_mat
    # compute the RMSD
    set rmsd [measure rmsd $compare $reference]
    puts "RMSD of $frame is $rmsd"
}

set allcluster [measure cluster [atomselect top "protein and backbone and noh"] num 5 distfunc rmsd cutoff 1.5]
#set allcluster [measure cluster $compare num 100 distfunc rmsd cutoff 1.5]

# puts "cluster is $allcluster"

#puts "Item at index 0 of the list is: [lindex $allcluster 0]"

set sizecl0 [llength [lindex $allcluster 0]]
set firstcon [lindex [lindex $allcluster 0] 0]
puts "first cluster has $sizecl0 conformers"
puts "first conformer is $firstcon"
#puts "Item at index 1 of the list is: [lindex $allcluster 1]"
#puts "Item at index 2 of the list is: [lindex $allcluster 2]"

puts "number of frame now is $num_beg"

for { set ncluster 0 } {$ncluster < 10} { incr ncluster 1 } {
set sizecl [llength [lindex $allcluster $ncluster]]
}
puts "The cluster $ncluster has $sizecl conformers\n"
set compare [atomselect top "protein and backbone and noh" frame [lindex [lindex $allcluster $ncluster] 0 ]]
#   # compute the transformation
#   set trans_mat [measure fit $compare $reference]
#   # do the alignment
#   $compare move $trans_mat
#   # compute the RMSD
#   set rmsd [measure rmsd $compare $reference]
puts "The cluster $ncluster has $rmsd from crystal conformers\n"

for { set i 0 } {$i < $sizecl } { incr i 1 } {
    animate dup frame [lindex [lindex $allcluster $ncluster] $i ] 0
    #puts "the confom
    }}

set num_end [molinfo top get numframes]
set num_end [expr {$num_end -1}]
puts "number of frame now is $num_end \n"
animate write dcd $ncluster.dcd beg $num_beg end $num_end waitfor all
animate delete beg $num_beg end $num_end

set num_beg [molinfo top get numframes]
puts "number of frame now is $num_beg \n"

} exit

5. Obtaining average structure from a simulation or from a cluster of frames.

package require psfgen
package require molefacture
package require alchemify
package require readcharmmtop
package require pbctools

resetpsf
pdbalias residue HIS HSD
pdbalias atom ILE CD1 CD

=(topology /home/guterres.h/Toppar/top_all27_prot_lipid_na.inp

217
mol new ../cell-final.xplor.psf
mol addfile ../cell-final.pdb
mol addfile 0.dcd type dcd waitfor all
# set cell [pbc get -now]

    pbc unwrap -sel "all not water"

# Prints the RMSD of the protein atoms between each timestep
# and the first timestep for the given molecule id (default: top)
proc print_rmsd_all          {{mol top}} {
    set outDataFile [open RMSDout w]
    # use frame 0 for the reference
    # set reference [atomselect $mol "protein and resid 10 to 200" frame 0]
    set reference [atomselect $mol "protein" frame 0]
    # the frame being compared
    # set compare [atomselect $mol "protein and resid 10 to 200"]
    set compare [atomselect $mol "protein"]

    set num_steps [molinfo $mol get numframes]
    for {set frame 0} {$frame < $num_steps} {incr frame} {
        # get the correct frame
        $compare frame $frame
        # compute the transformation
        set trans_mat [measure fit $compare $reference]
        # do the alignment
        $compare move $trans_mat
        # compute the RMSD
        set rmsd [measure rmsd $compare $reference]
        # print the RMSD
        puts "RMSD of $frame is $rmsd"
        puts $outDataFile "$frame $rmsd"
    }
}

close $outDataFile
return

print_rmsd_all

set outfile [open rmsf.dat w]
set sel [atomselect top "name CA"]
for {set i 0} {$i < [$sel num]} {incr i} {
set alpharmsf [measure rmsf $sel first 0 last 4929 step 1]
# puts $outfile "[expr {$i+1}] [lindex $rmsf $i]"
puts $outfile "[expr {$i+1}] [lindex $alpharmsf $i]"
# puts $outfile "[measure rmsf $sel first 1 last 6000 step 1]"
}
close $outfile

set average [atomselect top "protein"]
$average set {x y z} [measure avpos $average first 0 last 4929 step 1]
$average writepdb averagestructure0.pdb

exit

6. **Dynamical Network Analysis/ Community Network Analysis.** This analysis assigns each residue as a node and edges are used to connect the nodes that interact at least 75% of the time during the simulation. This method was conducted following procedures described in the tutorial file by John Eargle, Li Li, and Zan Luthey Schulten ([http://www.scs.illinois.edu/schulten/tutorials/network/](http://www.scs.illinois.edu/schulten/tutorials/network/)). Download and compile: carma 0.8 or later ([http://utopia.duth.gr/~glykos/Carma.html](http://utopia.duth.gr/~glykos/Carma.html)), CatDCD 4.0 precompiled binary ([http://www.ks.uiuc.edu/Development/MDTools/catdcd/](http://www.ks.uiuc.edu/Development/MDTools/catdcd/)), gncommunities and subopt from John Eargle’s tutorial file. Follow these steps on a cluster terminal that you’re working on:

a. cp gncommunities subopt network.config into /communities

b. Concatenate all trajectories to make one trajectory file. Open on VMD and save the last frame as a new pdb file. Use this pdb file to make a new psf file, use psfgen.tcl.

c. Edit the script for network.config. This is an example for Ras/Raf-RBD dimer

```
>Psf
dimer.psf

>Dcds
dimer.dcd

>SystemSelection
```
(segnames: RAS1, RAS2, RAF1, RAF2) and (not hydrogen)

>NodeSelection
(name CA A)

>Restrictions
notSameResidue
notNeighboringCAAlpha
notNeighboringPhosphate

d. Run network configuration by typing “networkSetup network.config” on your cluster terminal

e. You can load your network on VMD, go to Extensions → Analysis → NetworkView. Choose File menu and select Load Network, the output file should be called contact.dat

f. The second part is the calculation of the network communities. Communities are sets of residues that have stronger connections and move in concert with each other. On your terminal type: ./gncommunities contact.dat communities.out

g. View this community on VMD, in a similar manner as how you’d view the dynamical network.

h. The third part is calculating optimal and suboptimal paths between two nodes. For this, you would have to find out the identity number for each node that you want to measure the optimal and suboptimal paths for. Go to Tk console on VMD for that, in the folder where you have all these files type:
::NetworkView::getNodesFromSelection “Chain A and resid 137 and name CA”

This will return an ID for your node, record it. Do it again for another node that you want to connect to.

Then to run the suboptimal path calculation, run this command on your working terminal:
The first path and the shortest path is the optimal path. You can view this the same way as how you’d view the dynamical network through VMD, Extensions → Analysis → NetworkView.

7. **Principal Component Analysis (PCA)** was performed on carma software packages. Carma was downloaded in dynamical network analysis tutorial, in section 6. This analysis was performed following the tutorial posted on 


a. You should already have one trajectory file, one pdb and one psf file. On a working terminal, in the folder for your PCA analysis type:

    carma -v -w -fit -atmid ALLID -segid A 1KRas.dcd 1KRas.psf

    This is done to remove rotations and translations from your trajectory file.

b. To calculate PCA, run:

    carma -v -w -col -cov -eigen -proj 5 3 300 1KRas.dcd 1KRas.psf

    Number 5 is the number for principal components that will be written by carma. Number 3 is the number for principal components for which DeltaG (energy landscapes) diagrams will be produced by carma. And 300 is the temperature of the simulation in Kelvin.

c. To calculate Cartesian-PCA-based cluster analysis, run:
awk '{if ($2==1) prin $1, $3, $4, $5}' carma.clusters.dat > c_01.dat

Use Microsoft XL to make a scatter plot of the second and third columns of the file carma.PCA.fluctuations.dat file, and superimpose on this another scatter plot of the second and third columns from c_01.dat file.

d. To see the motion of the selected principal components, do:

awk '(print $2)' carma.PCA.fluctuations.dat | sort -n | tail -1
awk '(print $2)' carma.PCA.fluctuations.dat | sort -n | head -1

Use the two numbers that are returned from the commands above, and run:
carma -verb -write -col -cov -eigen -play 1 4 5 cluster_01.dcd CAs.psf

Your new dcd file that is produced from the command line above can be played on VMD to see the specific motions that correspond to your principal components.
Accelerated Molecular Dynamics protocol

Accelerated MD (aMD) is a variation of MD that alters the potential energy function by adding boost potentials to increase the sampling of protein conformational states during the simulation.

For aMD, we follow the previously published protocol by Jeff Wereszczynski and J. Andrew McCammon in Computational drug discovery and design, methods in molecular biology, vol. 189, chapter 30 (https://link.springer.com/protocol/10.1007/978-1-61779-465-0_30). The parameters for aMD that would be added to a regular configuration file are as following:

```bash
# Parameters for Accelerated MD, see excel spread sheet for parameter calculation####
accelMD on
accelMDdihe on
accelMDE XXX
accelMDalpalpha XXX
accelMDDual on
accelMDTE XXX
accelMDlalpalpha XXX
accelMDfirstStep XXX
accelMDLastStep XXX
accelMDOutFreq XXXX
# END of AMD parameter
```

As seeing above, we use two boost potential, one for the total potential energy and the other one for the dihedral energy. The full procedure is as following:

1. Perform a classical MD simulation preparation as mentioned above on appendix B.
2. Extract values for total potential energy and dihedral energy as using the following tcl script namdstat.tcl

```bash
puts "Usage: data_avg <logfile> \
[<first timestep> <last timestep>]"
pduts " <first timestep> and <last timestep> may be entered as numbers or"
pduts " <first timestep> = 'first' will start at the beginning of the simulation"
pduts " <last timestep> = 'last' will go to the end of the simulation"
pduts "Usage: data_time <data stream> <logfile> \
[<first timestep> <last timestep>]"
pduts " <data stream> = BOND, ANGLE, DIHED, IMPRP, ELECT, VDW,
BOUNDARY, MISC, KINETIC, TOTAL, TEMP, TOTAL2, TOTAL3, TEMPAVG"
```

```bash
proc data_avg {logfile {first 0} {last -1}} {
```
set file [open $logfile r] while {[gets $file line] != -1} {
    if [regexp "ETITLE:" $line] {set etitles $line}
} close $file
puts "Calculating averages..."

set file [open $logfile r]
set ener " " while {[gets $file line] != -1} {
    if [regexp "ENERGY: " $line] {set ener "$ener $line"}
} close $file

set l [llength $etitles]
set lc [expr [llength $ener]/$l]

for {set k 0} {$k < $l} {incr k} {
    set tstest [lindex $etitles $k]
    if {$tstest == "TS"} {set tsnum $k}
}

for {set k 0} {$k < $lc} {incr k} {
    set ts([expr $k+1]) [lindex $ener [expr $l*$k+$tsnum]]
} set lastts $ts($k)
set lastline $lc
set firstts $ts(1)
set firstline 1

if {$first != "first"} {
    set k 1
    while {$first > $ts($k)} {
        set firstts $ts([expr $k+1])
        set firstline [expr $k+1]
        incr k
    }
}

if {$last != "last"} {
    set k 1
    while {$last >= $ts($k)} {
        set lastts $ts($k)
        set lastline $k
        incr k
    }
}
puts "CALCULATING DATA FROM TIMESTEP $firstts TO $lastts:"

set k 0
for {set i 1} {$i <= $lc} {incr i} {
    for {set j 1} {$j <= [llength $etitles]} {incr j} {
        set e($i,$j) [lindex $ener $k]
        incr k
    }
}

for {set j 3} {$j <= [llength $etitles]} {incr j} {
    set total 0
    for {set i $firstline} {$i <= $lastline} {incr i} {
        set total [expr $total + $e($i,$j)]
    }
    set avg [expr $total/(1+$lastline-$firstline)]
    puts "[lindex $etitles [expr $j-1] ]: $avg"
}

unset ener
unset ts
}

proc data_time {data logfile {first 0} {last -1}} {
    set file [open $logfile r]
    while { [gets $file line] != -1 } {
        if [regexp "ETITLE:" $line] {set etitles $line}
    }
    close $file
    puts "Getting $data data..."

    set file [open $logfile r]
    set ener ""
    while { [gets $file line] != -1 } {
        if [regexp "ENERGY: " $line] {set ener "$ener $line"}
    }
    close $file

    set lc [expr [llength $ener]/[llength $etitles]]

    set k 0
    for {set i 1} {$i <= $lc} {incr i} {

for {set j 1} {j <= [llength $etitles]} {incr j} {
    set e($i,$j) [lindex $ener $k]
    incr k
}
}

set lastts $e($lc,2)
set lastline $lc
set firstts $e(1,2)
set firstline 1

if {$first != "first"} {
    set k 1
    while {$first > $e($k,2)} {
        set firstts $e([expr $k+1],2)
        set firstline [expr $k+1]
        incr k
    }
}

if {$last != "last"} {
    set k 1
    while {$last >= $e($k,2)} {
        set lastts $e($k,2)
        set lastline $k
        incr k
    }
}

set j 0

switch $data {
    BOND {set j 3}
    ANGLE {set j 4}
    DIHED {set j 5}
    IMPRP {set j 6}
    ELECT {set j 7}
    VDW {set j 8}
    BOUNDARY {set j 9}
    MISC {set j 10}
    KINETIC {set j 11}
    TOTAL {set j 12}
    TEMP {set j 13}
    POTENTIAL {set j 14}
    TOTAL3 {set j 15}
    TEMPAVG {set j 16}
}
3. On VMD go to Extensions → Tk console, and type “source namdstats.tcl” to run the above script. In a folder with all your output files for 10 ns (prun1.out, prun2.out etc),
type:

```
data_time DIHED prun1.out prun2.out ...prun20.out
data_time POTENTIAL prun1.out prun2.out ...prun20.out
```

4. The output files are text files for potential energy and dihedral energy during the simulation time. You can use Microsoft excel to calculate the average and calculate the boost potential. An example of the spreadsheet calculation is shown here.

```
<table>
<thead>
<tr>
<th>Time</th>
<th>Potential</th>
<th>Dihedral</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>87672.591</td>
<td>87672.591</td>
<td>0</td>
</tr>
<tr>
<td>10000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>20000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>30000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>40000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>50000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>60000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>70000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>80000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>90000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
<tr>
<td>100000</td>
<td>87227.845</td>
<td>87227.845</td>
<td>587.732</td>
</tr>
</tbody>
</table>
```

5. The formulas for boost potentials are as following:
a. Total potential energy

\[ V = E(\text{total}) + \frac{N_{\text{atom}}}{5} \]

accelMDTE
\[ \text{alpha} = \frac{N_{\text{atom}}}{5} \]

b. Dihedral energy

\[ V = E(\text{dihe}) + 4\times \text{number-of-residue} \]

accelMDE:
\[ \text{alpha} = \frac{4}{5}\times \text{number-of-res} \]

6. Modify all the subsequent configuration files to include the parameters for aMD. An example configuration file with aMD parameter is as following (with aMD parameters in bold):

```
# NAMD configuration file for H-RasY71A aMD

# molecular system
structure   cell-final_xplor.psf
coordinates abeta-run.out.coor
bincoordinates  abeta-prun20.rst.coor
binvelocities   abeta-prun20.rst.vel
extendedsystem  abeta-prun20.rst.xsc
firsttimestep  10000000

# force field
paratypecharmm on
parameters /home/guterres.h/Toppar/par_all27_prot_na.prm
exclude scaled1-4
1-4scaling 1.0

# approximations
switching on
switchdist 10
cutoff 11
pairlistdist 12
margin 1.0
stepspercyle 20
```
#PME
PME yes
PMEGridSpacing 1.0
FFTWUseWisdom no
nonbondedFreq 2

#protocol
#temperature 0  #initial t
seed 78133
rigidBonds all
#reassignFreq 1000
timestep 1.0
langevin on
langevinDamping 10
langevinTemp 300.0

# Constant Pressure Control (variable volume)
useGroupPressure yes ;# needed for rigidBonds
useFlexibleCell no
useConstantArea no

langevinPiston on
langevinPistonTarget 1.01325 ;# in bar -> 1 atm
langevinPistonPeriod 100.
langevinPistonDecay 50.
langevinPistonTemp 300.0

#reassignTemp 25
#reassignIncr 25
#reassignHold 300
#tCouple on
#tCoupleTemp 300

# Parameters for Accelerated MD, see excel spread sheet for parameter calculation####
accelMD on
accelMDdihe on
accelMDE 1324.3
accelMDalpha 132.8
accelMDdual on
accelMDTE -77910.9
accelMDTalpha 4827
accelMDFirstStep 0
accelMDLastStep 0
accelMDOutfreq 5000
# END of AMD parameter

# output
DCDunitcell on
wrapall on
outputname amd-prun21.out  # output PDB file
restartname amd-prun21.rst
DCDfile  amd-prun21.dcd
restartfreq 5000
outputenergies 5000
outputtiming 5000
dcddfreq 5000
binaryoutput no

# 0.5 ns a chunk
numsteps = 10500001
Wide-angle X-ray solution scattering data processing. Data collection for WAXS consists of both wide-angle and small-angle in which the two-data set need to be merging to obtain a full data set for a WAXS intensity plot. For each sample, we collected two data sets and averaged them. The average is the subtracted from the buffer data. This initial data processing is done using ATSAS (https://www.embl-hamburg.de/biosaxs/software.html). To merge the SAXS and WAXS data set, we use the following script (originally written by Hao Zhou):

```python
from Tkinter import Tk
from tkFileDialog import askopenfilename
import numpy as np
import sys
import matplotlib.pyplot as plt
import os

def average_fir(d, pts, ft = 1, op = 0):
    # op = 0, subtract the minimum and maximum value in the window
    # op = 1, don't subtract
    # ft: the number of running time
    if np.mod(pts,2) == 0 | pts < 3:
        raise Warning('The size of the window must be an odd number!')
        sys.exit()
    else:
        num_d = len(d)
        half_pt = int(np.floor(pts/2))
        for j in xrange(ft):
            new_d = np.zeros(num_d)
            if op == 0:
                for i in xrange(half_pt,num_d-half_pt):
                    new_d[i] = np.true_divide(np.sum(d[i-half_pt:i+half_pt+1]) - np.min(d[i-half_pt:i+half_pt+1]) - np.max(d[i-half_pt:i+half_pt+1]),(pts-2))
                    new_d[0:half_pt] = d[0:half_pt]
            elif op == 1:
                for i in xrange(half_pt,num_d-half_pt):
                    new_d[i] = np.true_divide(np.sum(d[i-half_pt:i+half_pt+1]),pts)
                    new_d[0:half_pt] = d[0:half_pt]
        d = new_d

    return new_d, half_pt
```
def get_data_arr(f):
    lines = f.readlines()
    num_lines = int(lines[3].strip())
    norm_num = float(lines[num_lines+12].strip().split('/')[1].strip('"')) # data from cornell is normalized by a factor that recorded in the data file
    x_data = []
    for line in lines[4:4+num_lines]:
        x_data.append(line.strip().split())
    x_data = np.array(x_data,'float')
    x_data[:,1] = x_data[:,1]*norm_num
    return x_data

Tk().withdraw()

saxs_filename = askopenfilename(filetypes = (('Data files', '*.dat' ),('All files', '*.*')),title = 'Open the SAXS file') # show an "Open" dialog box and return the path to the selected file
print(saxs_filename)
saxs_f = open(saxs_filename)

Tk().withdraw()
waxs_filename = askopenfilename(filetypes = (('Data files', '*.dat' ),('All files', '*.*')),title = 'Open the WAXS file') # show an "Open" dialog box and return the path to the selected file
print(waxs_filename)
waxs_f = open(waxs_filename)

saxs_data = get_data_arr(saxs_f)
[smooth_saxs, halfs] = average_fir(saxs_data[:,1],7,3,0)
plt.figure(1)
plt.plot(saxs_data[:,0],smooth_saxs,'r')
plt.plot(saxs_data[:,0],saxs_data[:,1],'b')
plt.show()

waxs_data = get_data_arr(waxs_f)
[smooth_waxs, halfw] = average_fir(waxs_data[:,1],7,8,0)
plt.figure(2)
plt.plot(waxs_data[:,0],smooth_waxs,'r')
plt.plot(waxs_data[:,0],waxs_data[:,1],'b')
plt.show()

inter_saxs = saxs_data[1,0] - saxs_data[0,0]
new_waxs_start = saxs_data[-1-halfs,0]
um_new_waxs = int(round((waxs_data[-1,0]-new_waxs_start)/inter_saxs))
new_waxs_x =
np.arange(new_waxs_start,new_waxs_start+inter_saxs*num_new_waxs,inter_saxs)
new_waxs_i = np.interp(new_waxs_x, waxs_data[halfw::,0],smooth_waxs[halfw::])

over_saxs = smooth_saxs[-1-halfls::]
over_waxs = new_waxs_i[0:halfls+1]

des = np.sum(over_waxs)/np.sum(over_saxs)
error = 9999999
f_des = des
scale = np.arange(des-0.25,des+0.25,0.01)
for i in xrange(len(scale)):
    err = np.sum(np.absolute(scale[i]*over_saxs-over_waxs))
    if err < error:
        error = err
        f_des = scale[i]

over_data = (f_des*over_saxs+over_waxs)/2
smooth_saxs = smooth_saxs*f_des
smooth_saxs[-1-halfls::] = over_data

merge_i = np.concatenate((smooth_saxs, new_waxs_i[halfls+1::]),axis = 1)
merge_x = np.concatenate((saxs_data[:,0],new_waxs_x[halfls+1::]),axis = 1)
[smooth_all,halfall] = average_fir(merge_i,11,7,0)

final_x = merge_x.reshape((len(merge_x),1))
final_i = smooth_all.reshape(len(smooth_all),1)
merge_data = np.concatenate((final_x,final_i),axis = 1)
plt.figure(3)
plt.plot(merge_x,smooth_all,'r',linewidth = 2)
plt.show()

plotsave = raw_input('Save the result?(y/n):')
if plotsave == 'y':
    file_dir = os.path.dirname(saxs_filename)
    file_name = 'M ' + os.path.basename(saxs_filename).split('.')[0] + '.txt'
    np.savetxt(file_dir+'/'+file_name, merge_data,fmt = '%5.5f')
else:
    sys.exit()
Sigma-r plot calculation to convert MD trajectory into a sigma-r plot (author Hao Zhou). This is a python code that can be run to convert xyz coordinates from MD trajectory frames into a one dimensional plot showing the standard deviation of interatomic distances in the protein as a function of the distance.

```python
import numpy as np
import math
from numbapro import cuda
from numba import float64
import numba
from timeit import default_timer as timer
import progressbar
from time import sleep
import os
import shutil
import sys
import matplotlib.pyplot as plt

cuda.select_device(1)
@numba.cuda.jit('void(float64[:], float64[:], float64[:], float64[:], float64[:], float64[:], float64[:], int32)', target = 'gpu')
def mean_std(atom1_x,atom1_y,atom1_z,atom2_x,atom2_y,atom2_z,dist,select_frame):
    i = cuda.threadIdx.x + cuda.blockIdx.x * cuda.blockDim.x
    if i <= select_frame:
        dist[i] = math.sqrt((atom1_x[i] - atom2_x[i])*(atom1_x[i] - atom2_x[i]) + (atom1_y[i] - atom2_y[i])*(atom1_y[i] - atom2_y[i]) + (atom1_z[i] - atom2_z[i])*(atom1_z[i] - atom2_z[i]))
    cuda.syncthreads()

def get_coor():
    global coor
    traj_file_dir = raw_input('Input the directory of the RAS trajectory file: ')
    traj_file = open(traj_file_dir)
    lines = traj_file.readlines()
    atom_num = int(lines[0].strip())
```

---

234
print 'The number of atom is %d' %atom_num
frame_num = len(lines)/(atom_num+2)
print 'The number of frame is %d' %frame_num
start_frame = input('Start Frame is:')
end_frame = input('End Frame is:')
print 'The frame range you choose is: %d ~ %d' %(start_frame,end_frame)
select_frame = end_frame - start_frame
lines = lines[(start_frame-1)*(atom_num+2):(end_frame-1)*(atom_num+2)]
coor = []
for line in lines:
    if line.strip() != str(atom_num) and line.strip() != 'generated by VMD':
        coor.append([line[10:23].strip(),line[26:39].strip(),line[42:53].strip()])
coor = np.array(coor,'float')
coor = np.reshape(coor,(select_frame, atom_num*3))
return coor, atom_num,select_frame

def have_mean_std():
a = raw_input('Where do you want to start?(0:From the beginning; 1:Already have the mean and std matrix; 
  2:Already have the sigma-r matrix.):')
return a
def input_step():
a = input('Input the step of r value of sigma-r plot:')
return a
def get_sigma_r(mean_all, std_all):
    step = input_step()
    num_interval = int(np.ceil(np.max(mean_all)/step))
    sigma_r = []
    r = []
    for i in xrange(num_interval):
        r_value = i*step
        sigma_ave = np.mean(std_all[np.where(np.logical_and(mean_all>i*step, mean_all <= (i+1)*step))])
        r.append(r_value)
        sigma_r.append(sigma_ave)
return r, sigma_r

def get_mean_all():
    a = raw_input('Input the directory of the mean_all matrix:')
    return a

def get_std_all():
    a = raw_input('Input the directory of the std_all matrix:')
    return a

def get_sigma_r_mat():
    a = raw_input('Input the directory of the sigma_r matrix:')
    return a

def plot_sigma_r(sigma_r_mat):

    plt.figure(1)
    plt.plot(sigma_r_mat[:,0], sigma_r_mat[:,1],'b', label='Sigma_r')
    plt.xlabel('Interatomic Distance')
    plt.ylabel('Average Standard Deviation')
    plt.legend(loc=2)
    plt.savefig(current_path+'/sigma_r_all_atom', dpi=300, facecolor='w', edgecolor='w',
                orientation='portrait', papertype=None, format=None,
                transparent=False, bbox_inches=None, pad_inches=0.1,
                frameon=None)
    plt.show()

def heatmap(sigma_matrix, save_name):
    num_of_atoms = np.shape(sigma_matrix)[0]
    for i in xrange(0, num_of_atoms-1):
        for j in xrange(i+1, num_of_atoms):
            sigma_matrix[j,i] = sigma_matrix[i,j]
    plt.figure(2)
    plt.imshow(sigma_matrix,interpolation = 'none')
    plt.jet()
    plt.colorbar()
    plt.xlabel('Atom Number')
    plt.ylabel('Atom Number')
    plt.title('Heatmap for the standard deviation of each interatomic vector')
    plt.savefig(current_path+'/'+save_name, dpi=300, facecolor='w', edgecolor='w',
                orientation='portrait', papertype=None, format=None,
                transparent=False, bbox_inches=None, pad_inches=0.1,
                frameon=None)
plt.show()

def main():

    control_num = have_mean_std()

    if control_num == '0':
        global mean_all
        global std_all
        global current_path
        dir_this_file = os.path.dirname(sys.argv[0])
        folder_name = raw_input('Create a new folder for saving the results. Folder name:')
        result_path = os.path.join(dir_this_file, folder_name)
        os.mkdir(result_path)
        print result_path

[coor,atom_num,select_frame] = get_coor()
mean_all = np.zeros((atom_num, atom_num))
std_all = np.zeros((atom_num,atom_num))
run_times = int(atom_num*(atom_num-1)/2)
bar = progressbar.ProgressBar(maxval = run_times,)
widgets = [progressbar.Bar(), ', ', progressbar.Percentage()]).start()
step = 0
start = timer()
for i in xrange(atom_num-1):
    atom1_x = coor[:, i*3]
    atom1_y = coor[:, i*3+1]
    atom1_z = coor[:, i*3+2]
    dA = cuda.to_device(np.ascontiguousarray(atom1_x, dtype = np.float64))
    dB = cuda.to_device(np.ascontiguousarray(atom1_y, dtype = np.float64))
    dC = cuda.to_device(np.ascontiguousarray(atom1_z, dtype = np.float64))
    for j in xrange((i+1), atom_num):
        atom2_x = coor[:, j*3]
        atom2_y = coor[:, j*3+1]
        atom2_z = coor[:, j*3+2]
blockdim = 500, 1
griddim = int(math.ceil(float(select_frame)/blockdim[0])), 1

dD = cuda.to_device(np.ascontiguousarray(atom2_x, dtype = np.float64))
dE = cuda.to_device(np.ascontiguousarray(atom2_y, dtype = np.float64))
dF = cuda.to_device(np.ascontiguousarray(atom2_z, dtype = np.float64))
dG = cuda.device_array(np.shape(atom1_x),dtype = np.float64)
mean_std[griddim, blockdim](dA, dB, dC, dD, dE, dF, dG, select_frame)
cuda.synchronize()
dist = dG.copy_to_host()
cuda_cost = timer() - start
#print ('Cuda took %f seconds' % cuda_cost)

mean_all[i,j] = np.mean(dist)
std_all[i,j] = np.std(dist)

bar.update(step)
step = step + 1
bar.finish()

cuda_cost = timer() - start
time_per_cuda = float(cuda_cost/step)
print ('Cuda took average %f seconds for each atom pair calculation' % time_per_cuda)
print step
np.savetxt("mean_ras.txt",mean_all,fmt="%5.3f")
np.savetxt("std_ras.txt",std_all,fmt = "%5.3f")
shutil.move(dir_this_file+'/mean_ras.txt', result_path)
shutil.move(dir_this_file+'/std_ras.txt', result_path)

[r, sigma_r] = get_sigma_r(mean_all, std_all)
std_d_r = std_all/mean_all
sigma_r_mat = np.transpose(np.array([r,sigma_r]))
np.savetxt(result_path+'/sigma_r_all_atom.txt', sigma_r_mat, fmt = "%5.3f")
current_path = result_path
plot_sigma_r(sigma_r_mat)

heatmap(std_all[0:atom_num-1,0:atom_num-1],'std_heatmap')
heatmap(std_d_r[0:atom_num-1,0:atom_num-1],'std_divide_d')
if control_num == '1':
    global current_path
    mean_all_loc = get_mean_all()
    mean_all = np.loadtxt(mean_all_loc)
    std_all = np.loadtxt(get_std_all())
    std_d_r = std_all/mean_all
    [r, sigma_r] = get_sigma_r(mean_all, std_all)
    sigma_r_mat = np.transpose(np.array([r,sigma_r]))
    print 'sigma_r.txt will be save into the same location of mean and std files'
    current_path = os.path.dirname(mean_all_loc)
    np.savetxt(current_path+'/sigma_r_all_atom.txt', sigma_r_mat, fmt = "%5.3f")
    plot_sigma_r(sigma_r_mat)
    heatmap(std_all[0:atom_num-1,0:atom_num-1],'std_heatmap')
    heatmap(std_d_r[0:atom_num-1,0:atom_num-1],'std_divide_d')

if control_num == '2':
    global current_path
    sigma_r_mat = np.loadtxt(get_sigma_r_mat())
    plot_sigma_r(sigma_r_mat)

if __name__ == '__main__':
    main()