UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Molecular Switches Coordinate Dynamically Coupled Allosteric Networks in Protein Complexes

Permalink

https://escholarship.org/uc/item/5zs5t333

Author

Gaieb, Zied

Publication Date

2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Molecular Switches Coordinate Dynamically Coupled Allosteric Networks in Protein Complexes

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Bioengineering

by

Zied Gaieb

August 2016

Dissertation Committee:
Dr. Dimitrios Morikis, Chairperson
Dr. Chia-en A. Chang
Dr. David D. Lo

Copyright by
Zied Gaieb
2016

The Dissertation of Zied Gaieb is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

My PhD has been a life-changing experience and it would not have been possible without the support and guidance that I received from many individuals. I, first, would like to thank my advisor, Professor Dimitrios Morikis, for his support and encouragement throughout the years. His passion and enthusiasm for research was contagious and motivational for me. Joining the lab, I had the opportunity to pursue my own research interests, which taught me to become independent and gave me the chance to design my own project. I also would like to thank my committee members, Professor David Lo, and Professor Chia-en Chang, for their contributions of time and ideas during my PhD. Also, I would like to acknowledge the UCR Graduate Research Mentorship Program for providing funding, and the National Center for Multiscale Modeling of Biological Systems (MMBioS) through Grant P41GM103712-S1 from the National Institutes of Health and the Pittsburgh Supercomputing Center (PSC) for providing Anton computer time. We also thank Dr. Annick Dejaegere (Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France) for providing the CHARMM forcefield parameters for methylated lysine that were essential parameters to run our simulations.

I am also grateful for all the collaborative endeavors I have had during my PhD. I had the opportunity of working on multiple projects with Professor WenYong Chen at the City of Hope, Professor Jiayu Liao and his student Dr. Harbani Kaur Malik Chaudhry at the University of California Riverside, and Professor Neal L. Schiller and his student Dr. Danh Cong Do at the University of California Riverside.

I was lucky to be a member of a remarkable research group. Members of the BioMoDeL group have contributed immensely to my personal and professional time at UCR. I am grateful for the group of original BioMoDeL group members, Drs. Chris Kieslich, Aliana López De Victoria, Ronald Gorham, and Phanos Tamamis, for their tremendous help and advice in many of my projects. I am also thankful for the newest BioMoDeLers, Reed Harrison, Rohith Mohan, and Nehemiah Zewde for helping me get through my final years.

Finally, I would like to thank my family, especially my parents, for all their hard work and support to get me to where I am today. Their support has been unconditional all these years, and they have given up many things for me to be here.

The text of this dissertation, in part, is a reprint of the material as it appears in:

1. Gaieb, Z., D. Lo, D. Morikis. 2016. Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7. Journal of Chemical Information and Modeling. Just Accepted Manuscript.

The co-author Dimitrios Morikis directed and supervised the research which forms the basis for this dissertation. Other co-authors listed provided experimental expertise.

DEDICATION

I dedicate this dissertation to
my parents:
Ridha Gaieb,
and Monia Adhoum Gaieb,
whose sacrificial care and support
made it possible for me to complete this work.

ABSTRACT OF THE DISSERTATION

Molecular Switches Coordinate Dynamically Coupled Allosteric Networks in Protein Complexes

by

Zied Gaieb

Doctor of Philosophy, Graduate Program in Bioengineering
University of California, Riverside, August 2016
Dr. Dimitrios Morikis, Chairperson

Structure and dynamics are essential elements of protein function. Protein structure is constantly fluctuating and undergoing conformational transitions, which are typically captured by molecular dynamics (MD) simulations. Conformational state transitions in a protein involve shifts in its equilibrium conformations that occur either independently or as a response to external perturbations. In this work, we describe the effect of ligand binding and post-translational modifications (PTMs) to proteins as an external perturbation responsible for conformational changes in chemokine receptor 7 (CCR7) and the KU70-KU80 protein complex, respectively. In both systems, we isolate specific side chain rearrangements that act as molecular switches, and mediate the allosteric communication between distant functional sites in a protein, as a mechanism to regulate conformational state transitions and sampling. Specifically, in CCR7, we focus on the role of allostery in regulating the information transduced from the ligand-binding site to the intracellular region of the receptor to allow discrimination in binding intracellular effectors. This phenomenon is known as biased activation and is critical to G protein-coupled receptor function. In our work, we detect a series of molecular switches in CCR7 that are coupled to various ligand-induced allosteric events. Although these molecular switches mediate the transitioning between different states, the receptor remains inactive (absence of the canonical TM6 outward movement), illustrating loose coupling between the extracellular ligand-binding site and the intracellular effector-binding site. This finding might justify the existence of a novel hybrid model in CCR7, consisting of a “rhodopsin-like” sequential network of allosteric events (mediated by molecular switches) and a “β_2-adrenergic-like” loose coupling between the extracellular and intracellular regions of the receptor. Furthermore, MD simulations of the ligand-free receptor highlight the importance of the ligand in coordinating the receptor’s side-chain fluctuations. We also focus on developing new methods to systematically detect coupled molecular switches and large domain motions in membrane proteins. Finally, we used MD simulations and electrostatic calculations to identify the role of PTMs, such as acetylation and methylation, on KU70-KU80’s dynamics. Such PTMs are shown to regulate conformational changes within several of KU70’s functional domains through acetylation-dependent alteration of the electrostatic profile of the DNA-binding and linker-SAP domains, and methylation-dependent molecular switching that is responsible for regulating a “pendulum-like” motion in linker-SAP domain.

TABLE OF CONTENTS

Chapter 1: Introduction **1**
- 1.1 Protein Dynamics 1
- 1.2 G Protein-Coupled Receptors 2
- 1.3 Activation Mechanism of β2-Adrenergic and Rhodopsin Receptors 3
- 1.4 Biased Ligand Activated GPCRs 4
- 1.5 Effect of Post-Translation Modifications on Protein Dynamics 5
- 1.6 Overview 5
- 1.7 References 8

Chapter 2: Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7 11
- 2.1 Introduction 11
- 2.2 Materials and Methods 13
- 2.3 Results 19
- 2.4 Discussion 32
- 2.5 Acknowledgements 38
- 2.6 References 39

Chapter 3: Comparison of Free and Ligand-Bound Conformational Transitions of CC Chemokine Receptor 7, using Molecular Dynamics Simulations 46
- 3.1 Introduction 46
- 3.2 Materials and Methods 48
- 3.3 Results 50
- 3.4 Discussion 55
- 3.5 Acknowledgements 56
- 3.6 References 57

Chapter 4: Detection of Molecular Switches Mediating Domain Motions in Molecular Dynamics Simulations of Membrane Proteins 59
- 4.1 Introduction 59
- 4.2 Materials and Methods 61
- 4.3 Application to Molecular Dynamics Simulation Data 67
- 4.4 Concluding Remarks 70
- 4.5 References 71

Chapter 5: Lysine Acetylation and Methylation Regulate KU70 Functions 73
- 5.1 Introduction 73
- 5.2 Materials and Methods 74
- 5.3 Results 79
- 5.4 Discussion 88
- 5.5 Acknowledgements 91
- 5.6 References 92

Appendices 96

Appendix A: Supporting data for Chapter 2 96
- A.1 Discussion of Previous Mutagenesis Data 96
- A.2 Supplementary Figures 97
- A.3 References 108

Appendix B: Supporting data for Chapter 5 109
- B.1 Supplementary Figures 109

TABLE OF CONTENTS (continued)

Appendix C: Algx Binds Alginate Through Fine-Tuned Intra- and Inter-Molecular Interactions in the C-Terminal Carbohydrate-Binding Module 111
- C.1 Introduction 111
- C.2 Materials and Methods 111
- C.3 Results 113
- C.4 Discussion 116
- C.5 References 119

Appendix D: Roles of APPBP1 and UBA3 in NEDD8 Activation 122
- D.1 Introduction 122
- D.2 Materials and Methods 122
- D.3 Results 124
- D.4 Discussion 126
- D.5 References 128

Appendix E: The Role of Zinc and Calcium Ions in MMP-14 and TIMP-2 Association 129
- E.1 Introduction 129
- E.2 Materials and Methods 131
- E.3 Results and Discussion 133
- E.4 References 138

Appendix F: Input files for simulations and scripts for data analysis 140
- F.1 Side Chain Distance Time Series Calculation 136
- F.2 Backbone Distance Time Series Calculation 144
- F.3 Detection of Side Chain Contact Rearrangements from MD Simulations 148
- F.4 Detection of Large Domain Motions Through DCCM 155
- F.5 Network of the Protein’s Dynamical Components 160

Appendix G: Ligand Design Slides 169

Curriculum Vitae 203

LIST OF FIGURES

Chapter 2: Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7

2.1 Molecular switches in the CCR7 transmembrane domain adopt ligand-specific conformations 23
2.2 During the cMD simulation, CCL21 induces a series of allosteric events that prompt equilibrium shifts in the discrete conformational states of CCR7's molecular switches 25
2.3 During the cMD simulation, CCL19 induces a series of allosteric events that prompt equilibrium shifts in the discrete conformational states of CCR7's tri-tyrosine switch 27
2.4 CCL21 induces global helical motions in CCR7 cMD simulation 30
2.5 Multiple sequence alignment of residues in CC chemokine receptor family displays conserved motifs within the characterized switches 36

Chapter 3: Comparison of Free and Ligand-Bound Conformational Transitions of CC Chemokine Receptor 7, using Molecular Dynamics Simulations

3.1 Molecular graphics of CCL19-bound, CCL21-bound, and apo forms of CCR7 49
3.2 Dynamic cross-correlation map for CCL19-bound, CCL21-bound, and apo forms of CCR7 50
3.3 Root mean-square fluctuations of CCR7 in its CCL19-bound, CCL21-bound, and apo forms 52
3.4 Comparison of Y112$^{3.32}$ interactions between CCL19-bound, CCL21-bound, and apo forms of CCR7 54

Chapter 4: Detection of Molecular Switches Mediating Domain Motions in Molecular Dynamics Simulations of Membrane Proteins

4.1 Schematic of our computational framework to detect molecular switches mediating large domain motions in proteins 62
4.2 Examples of side chain distance probability densities fitted using Gaussian mixture models 64
4.3 Dynamic cross-correlation heat map of pairwise C_α-C_α distance time series are clustered using hierarchical clustering 65
4.4 Dynamic cross-correlation-based network illustration of the protein's dynamical components 66
4.5 DCC-based network of the full CCL21-bound CCR7 MD simulation dataset 69

Chapter 5: Lysine Acetylation and Methylation Regulate KU70 Functions

5.1 Modeled structure of the KU70 (magenta)-KU80 (cyan)-DNA (yellow) complex 79
5.2 Electrostatic contribution of Ku70's acetylation sites 81
5.3 Effect of K74 lysine methylation on its microenvironment 83
5.4 Effect of KU70 lysine methylation on the intra-molecular "pendulum" 85
5.5 Principal component analysis of the helical hinge motion (K510-E537) 87

Appendix A: Supporting data for Chapter 2

A.1 Root mean-square deviation of CCR7 N-terminal domain and CCR7 C-terminal domain calculated in both CCL19-bound and CCL21-bound cMD and aMD simulations 97
A.2 Contact maps between CCR7 N-terminal domain and each of its ligands are in close agreement with published NMR data 98
A.3 CCL19 and CCL21 adopt different conformations in CCR7's binding pocket 99
A.4 Molecular switches in the CCR7 transmembrane domain adopt ligand-specific conformations within aMD simulations 100
A.5 Root mean-square deviation calculated in both CCL19-bound and CCL21-bound cMD simulations to assess the ligand-NTD stability and the receptor fluctuations 100
A.6 CCL21 prompts a hydrogen bond network within TM6 and TM7 in cMD simulation. 101
A.7 π-stacking interactions in the tri-tyrosine switch 102
A.8 Molecular switches characterized in CCL21-bound cMD simulation remain stable in the CCL19-bound cMD simulation 103

A.9 During the aMD simulation, CCL21 induces global motions in CCR7 that show a decrease and increase in helical distances synchronized with the formation of hydrogen bond between Y112$^{3.32}$ and Y255$^{6.51}$ 104
A.10 During the cMD simulation, CCL19 binding promotes global helical motions in CCR7's extracellular TM domains 105
A.11 CCL19 and CCL21 binding results in different arrangements in a polar interaction between TM6 and TM7 in cMD simulations 106
A.12 Multiple sequence alignment of residues in CC chemokine receptor family and CXCR4 displays conserved motifs within the characterized switches 107

Appendix B: Supporting data for Chapter 5

B.1 Backbone representation of KU70 MD trajectories 109
B.2 Hydrogen bond and charge-charge (Coulombic) interaction occupancy changes in the vicinity of K74 between methylated and nonmethylated states of the complex in the MD simulations in its acetylated state 110
B.3 Electrostatic potential of linker-SAP domain of KU70 illustrating the SAP domain positively-charged side (A) and negatively-charged side (B). 110

Appendix C: Algx Binds Alginate Through Fine-Tuned Intra- and Inter-Molecular Interactions in the C-Terminal Carbohydrate-Binding Module

C.1 In silico docking of alginate ligands to AlgX carbohydrate binding module 114
C.2 Intra-molecular interaction network of the CBM 115

Appendix D: Roles of APPBP1 and UBA3 in NEDD8 Activation

D.1 Thermodynamic cycle for calculation of electrostatic free energy of association and solvation 123
D.2 Non-covalent interactions between APPBP1 and UBA3 are critical for NEDD8 activation and conjugation 125

Appendix E: The Role of Zinc and Calcium Ions in MMP-14 and TIMP-2 Association

E.1 Activation mechanism of Pro-MMP2 129
E.2 Thermodynamic cycle for calculating electrostatic free energy of association and solvation 132
E.3 Structure of MMP-14 TIMP-2 complex illustrating Zinc and Calcium ions 133
E.4 Electrostatic Clustering for MMP14 proteins and free energy results of MMP14-TIMP2 complexes. 136

LIST OF TABLES

Appendix E: The Role of Zinc and Calcium Ions in MMP-14 and TIMP-2 Association

E.1 16 MMP14 complexes showing the different combinations of ions used 131
E.2 Ion distances to ionizable amino acids in vicinity with a cutoff distance of 9 Å 134
E.3 pK_a values of the residues coordinating the catalytic ion (Zn289) in the MMP14-TIMP2 complex with ions and without the catalytic zinc 135

CHAPTER 1: INTRODUCTION

1.1 Protein Dynamics

The mechanics of large biological machines arise from the properties of their constituent parts, whether it is the mechanics of biological tissue emerging from those of the cell or the mechanics of protein structures emerging from the dynamics and spatial arrangements of their amino acid residues. Modeling and predicting such phenomenon represents a fundamental challenge in structural biology limited by our understanding of protein dynamics.

Proteins are a central component of cellular function, where they participate in all processes through a highly diverse set of tasks. Protein diversity mainly stems from their diverse three-dimensional structures and amino acid compositions, where protein sizes range from hundreds to thousands of residues, forming up to 100,000 or so different proteins in the human body (1). Despite their diverse roles and structural arrangements, proteins are mainly encoded by up to 20 different amino acid residues, where amino acid residue chain composition determines the protein's emergent structure and dynamics. Consequently, protein function arises from its dynamics as a large ensemble of conformations that can be grouped into different conformational states depending on biological function, free energy, and three-dimensional arrangements (2, 3).

Conformational state transitions in a protein involve shifts in its equilibrium conformation that occur, either independently or as a response to external perturbations, such as ligand binding. Additionally state transitions involve several protein motions that can be classified based on their timescales of occurrence, which range from nanoseconds to milliseconds, depending on the energy barriers separating both conformational states (3). Each energy barrier is composed of enthalpic and entropic contributions that are manifested as large domain motions and atomic fluctuations. Large domain motions orchestrate the protein's transitioning between distinct states that are sampled on hundreds of nanosecond to microsecond or millisecond timescales and are separated by energy barriers of several RT (usually exceeding thermal energy, RT = 0.6 kcal/mol, where R is the gas constant). Within each state, the protein is not static, but instead involves thermal fluctuations of its side chains and backbone atoms, occurring on picosecond and nanosecond timescales. The timescale and amplitude of protein motion orchestrate transitions between states and are a result of the energy barriers making up the energy landscape of the protein (3).

Conformational state transitions emerge from coupled large domain motions and side chain contacts reorganization between remote regions of a protein. This phenomenon is termed allostery and it plays an important role in transmitting information between distant functional sites of the protein as a mechanism to regulate its conformational state transitions and sampling (2–4). Allosteric regulation is mediated by side chain rearrangements that act as molecular switches, which contribute to the entropic and enthalpic components of the energy barrier. Information transfer between distant sites in a protein is facilitated by a network of strongly or loosely coupled molecular switch rearrangements (4, 5).

Molecular dynamics (MD) simulation is one of the many techniques used to study protein allostery at atomic level (3). Several recent advances in enhanced sampling methods and simulation speed and accuracy have allowed us to reach biologically relevant timescale that capture the transitioning of a protein between different states and, consequently, allow the study of allostery (3). Time scales up to several microseconds to millisecond are now readily accessible by MD simulations (6–8). Several studies have explored a number of fast folding proteins (9) and captured the transitioning of membrane proteins between different states (10, 11). With the current accessible time scales, entropic changes have been challenging to estimate due to the large conformational space sampled by proteins (12). Consequently, systems, where allostery is dominated by enthalpic changes, are better suited for study by MD simulations. In a tightly packed environment of the transmembrane (TM) domain of membrane receptors, the entropic changes are limited, and molecular-switch rearrangements are manifested mainly as enthalpic changes.

1.2 G Protein-Coupled Receptors

G protein-coupled receptor family comprises of more than 800 different TM receptor proteins and represent one of most popular drug target, accounting for 30-40% of all FDA-approved drugs (13). Knowledge of the structural and dynamic features is essential for gaining a deeper understanding of how these receptors operate and provides a framework to aid in the rational design of therapeutics that modulate the cell's signaling pathways. The conformational ensemble of the receptor comprises of distinguishable conformational states characterized by its seven-transmembrane helical arrangements, and is a fundamental component of the receptor's selective properties in binding to intracellular (INC) effectors (14–16). To enable the complex function of a GPCR, bound ligands are capable of stabilizing the receptor in many distinct conformational states ranging from inactive to fully active, spanning G protein, arrestin, or GPCR kinase (GRK) biased conformations (14, 16–19). Additionally, receptor conformations are capable of discriminating between the different effector subtypes and arrestin binding modes (14, 20). These attributes emphasize the importance of the receptor's conformational diversity and the role of the ligands in shifting the receptor's equilibrium in sampling different conformational states.

1.3 Activation Mechanism of β2-Adrenergic and Rhodopsin Receptors

Ligand binding to a GPCR drives the receptor to its corresponding conformational states through different molecular switches and helical rearrangements (21). Crystallographic and ^{19}F-NMR spectroscopy studies have identified specific TM helices involved in the INC rearrangements of the receptor. Specifically, upon ligand binding, TM5, TM6, and TM7 undergo large helical displacements to accommodate effector binding (14, 16, 22, 23). Rhodopsin and β2-adrenergic receptor are two of the most studied GPCRs, both experimentally and computationally (11, 24–26). β2-adrenergic receptor was shown to feature loose coupling between the agonist-induced motions in the binding site and the canonical outward motion in the intracellular interface as a result of the absence of molecular switches mediating such information (11). β2-adrenergic receptor was shown to favor an inactive conformation where the canonical outward movement of TM6 is absent unless bound to an intracellular effector. On the other hand, rhodopsin's activation mechanism involves a sequential model of coupled allosteric events within its molecular switches, where initial small conformational changes caused by light-induced isomerization of 11-*cis*-retinal into all-*trans*-retinal are converted to larger changes in the INC region of the receptor, and induce the canonical TM6 outward motion without requiring the presence of an intracellular effector (27–29).

Phylogenetic tree representation of the human GPCR superfamily is based on sequence similarity of the TM region of the receptor, and shows rhodopsin and β2-adrenergic receptor to belong to the same subfamily, indicating high sequence similarity. Despite their close proximity in the phylogenetic tree, both receptors show different activation mechanisms (Figure 1 in (13)). This is due to their inherent biological function, where rhodopsin ensures the rapid and efficient detection of a photon (29), while the β2-adrenergic receptor evolved for a more complex signaling behavior (25). In that aspect, both receptors has evolved to function through different ligands, intended for different biological functions (30).

1.4 Biased Ligand Activated GPCRs

Biased receptors are activated by several endogenous ligands capable of inducing ligand-specific signaling pathways through the same receptor. Therefore, they represent a fundamentally different function from rhodopsin and β2-adrenergic receptors where both receptors are either activated by a photon-induced isomerization or unbiased and endogenous ligand-binding, respectively. Both receptors are unbiased and were shown to indiscriminately activate their respective intracellular pathways in the cell (30). However, unlike β2-adrenergic and rhodopsin receptors, biased receptors have evolved to have a tightly regulated function that ensures discrimination between several intracellular effectors. Given the different activation mechanisms of β2-adrenergic and rhodopsin receptors, we now ask: where do receptors that function through biased endogenous ligands fit within these two models? And how is the information transduced from the ligand-binding site to the intracellular region of the receptor to allow discrimination between intracellular effectors?

As a biased receptor, we examine a key regulator of the adaptive immune response in the CC chemokine receptor family, CC chemokine receptor 7 (CCR7) (31). Its ligands, CCL19 and CCL21, have distinct roles in the homing and functional compartmentalization of T cells and antigen-presenting dendritic cells to and within the secondary lymph nodes, as a result of their differential chemotactic behaviors (31–33). To initiate CCR7's cellular function, CCL19 and CCL21 have been shown to selectively induce distinct signaling pathways in the cell (17). While both ligands mediate their signaling through binding of intracellular G_i protein and GRK6 to CCR7, only CCL19 induces CCR7 internalization and desensitization through receptor phosphorylation by GRK3 and recruitment of β-arrestin 2 (17, 34). This differential binding of CCR7 to β-arrestin 2, GRK3, and GRK6 suggests the presence of selective conformational states in CCR7 induced by its biased ligands, CCL19 and CCL21. Using MD simulations, we detect a series of molecular switches hypothesized to facilitate the receptor's conformational transitions to ligand-specific conformational states. With that, we hypothesize that CCR7 involves a novel activation mechanism that combines the "rhodopsin-like" molecular switches that induce a ligand-initiated relay of interactions and the "β2-adrenergic-like" loose coupling between the ligand-binding site and the intracellular region.

1.5 Effect of Post-Translation Modifications on Protein Dynamics

Post-translational modifications (PTM) are important mechanisms in cellular signaling, which can lead to significant changes in the dynamics of a protein. In chapter 5, we study the effect of PTM on the dynamics of KU70-KU80, which is the first protein recruited to initiate non-homologous end-joining (NHEJ) of Double-strand DNA breaks. Dysregulation of KU70-KU80 through acetylation and methylation of lysine residues has been shown in various cell lines to introduce mutations and error-prone NHEJ (35–37). However, no molecular mechanism describes how such PTMs might affect the interactions of KU70-KU80 with DNA, and its deacetylase and demethylase enzymes, SIRT1, and LSD1. In this regard, our objective is to bridge the knowledge-gap between PTMs on KU70 and fallible DNA repair in NHEJ. The project aim is to assess the structural mechanisms of acetylation and methylation on regulating KU70-KU80 heterodimer's dynamics.

1.6 Overview

The majority of the work described in this thesis aims to investigate the dynamics of biased GPCRs. We study the role of allostery in inducing large domain motions in the receptor coordinated by its constituent side chain motions. This information is subsequently used to develop a computational framework designed to characterize the dynamical behavior of membrane proteins by systematically extracting biologically relevant motions. In chapter 2, we simulate microsecond dynamics of CC chemokine receptor 7 (CCR7) bound to its native biased ligands, CCL19 and CCL21, and detect a series of molecular switches that are mediated by various ligand-induced allosteric events. These molecular switches involve three tyrosine residues (Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$), three phenylalanine residues (F116$^{3.36}$, F208$^{5.47}$, and F248$^{6.44}$), and a polar interaction between Q252$^{6.48}$ and R294$^{7.45}$ in the transmembrane domain of CCR7. Conformational changes within these switches, particularly hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$, lead to global helical movements in the receptor's transmembrane helices and contribute to the transitioning of the receptor to distinct states. Ligand-induced helical movements in the receptor highlight the ability of biased ligands to stabilize the receptor in different states through a dynamic network of allosteric events. In chapter 3, we highlight the importance of the ligand in coordinating the receptor's side-chain fluctuations to drive the conformational changes responsible for state transitions in the receptor. Despite the similarly high side chain fluctuations in all receptor forms, only the ligand-bound receptors show substantial correlated conformational changes in the receptor. We conclude that the lack of correlation in the apo receptor is owed to the absence of a bound ligand capable of inducing conformational changes in the receptor's molecular switches. In chapter 4, we focus on the methodology for detecting side chain interactions that act as molecular switches mediating large domain motions and vice versa. Molecular switches are extracted from persistent side chain interactions that undergo well-defined abrupt changes in distance time series using Gaussian mixture models (GMM), whereas large domain motions are detected using dynamic cross-correlation (DCC). This method allows for the study of allosteric regulation in proteins by relating different molecular switches to the larger domain motions that are essential to their function. This computational framework is suitable for the study of tightly packed proteins, such as membrane proteins, and we use the ligand bound CC chemokine receptor 7 (CCR7) as an example. Overall, the computational framework can be tailored to study different protein environments and dynamics.

In chapter 5, we identify the role of acetylation and methylation on KU70's dynamics, which consequently affects its binding function to different effectors, such as DNA, SIRT1, and LSD1. KU70's acetylation is shown to alter the DNA binding affinity by neutralizing the charge of four lysine residues that reside within the DNA-binding domain. Additionally, the electrostatic profile of the linker domain was altered through charge-removing acetylation of five lysine residues on the linker. The latter contributes to a more negatively charged linker and disrupts alternating positively and negatively charged patches. This linker domain connects KU70's core and SAP domains; and both, lysine acetylation and alternating-charge pattern, are capable of altering the linker dynamics, which consequently can alter the SAP domain's function of binding the KU70-KU80 complex and/or DNA (38).

1.7 References

1. Wilhelm, M., J. Schlegl, H. Hahne, A.M. Gholami, M. Lieberenz, M.M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster. 2014. Mass-spectrometry-based draft of the human proteome. Nature. 509: 582–587.

2. Motlagh, H.N., J.O. Wrabl, J. Li, and V.J. Hilser. 2014. The ensemble nature of allostery. Nature. 508: 331–339.

3. Henzler-Wildman, K., and D. Kern. 2007. Dynamic personalities of proteins. Nature. 450: 964–972.

4. Popovych, N., S. Sun, R.H. Ebright, and C.G. Kalodimos. 2006. Dynamically driven protein allostery. Nat. Struct. Mol. Biol. 13: 831–838.

5. Vanatta, D.K., D. Shukla, M. Lawrenz, and V.S. Pande. 2015. A network of molecular switches controls the activation of the two-component response regulator NtrC. Nat. Commun. 6: 7283.

6. Salomon-Ferrer, R., A.W. Götz, D. Poole, S. Le Grand, and R.C. Walker. 2013. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. 9: 3878–3888.

7. Shaw, D.E., R.O. Dror, J.K. Salmon, J.P. Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, K.J. Bowers, E. Chow, M.P. Eastwood, D.J. Ierardi, J.L. Klepeis, J.S. Kuskin, R.H. Larson, K. Lindorff-Larsen, P. Maragakis, M.A. Moraes, S. Piana, Y. Shan, and B. Towles. 2009. Millisecond-scale molecular dynamics simulations on Anton. In: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. . pp. 1–11.

8. Miao, Y., F. Feixas, C. Eun, and J.A. McCammon. 2015. Accelerated molecular dynamics simulations of protein folding. J. Comput. Chem. 36: 1536–1549.

9. Lindorff-Larsen, K., S. Piana, R.O. Dror, and D.E. Shaw. 2011. How Fast-Folding Proteins Fold. Science. 334: 517–520.

10. Miao, Y., S.E. Nichols, P.M. Gasper, V.T. Metzger, and J.A. McCammon. 2013. Activation and dynamic network of the M2 muscarinic receptor. Proc. Natl. Acad. Sci. 110: 10982–10987.

11. Dror, R.O., D.H. Arlow, P. Maragakis, T.J. Mildorf, A.C. Pan, H. Xu, D.W. Borhani, and D.E. Shaw. 2011. Activation mechanism of the β2-adrenergic receptor. Proc. Natl. Acad. Sci. 108: 18684–18689.

12. Wereszczynski, J., and J.A. McCammon. 2012. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. Q. Rev. Biophys. 45: 1–25.

13. Stevens, R.C., V. Cherezov, V. Katritch, R. Abagyan, P. Kuhn, H. Rosen, and K. Wüthrich. 2013. The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. Nat. Rev. Drug Discov. 12: 25–34.

14. Liu, J.J., R. Horst, V. Katritch, R.C. Stevens, and K. Wüthrich. 2012. Biased Signaling Pathways in β2-Adrenergic Receptor Characterized by 19F-NMR. Science. 335: 1106–1110.

15. Reiter, E., S. Ahn, A.K. Shukla, and R.J. Lefkowitz. 2012. Molecular Mechanism of β-Arrestin-Biased Agonism at Seven-Transmembrane Receptors. Annu. Rev. Pharmacol. Toxicol. 52: 179–197.

16. Rahmeh, R., M. Damian, M. Cottet, H. Orcel, C. Mendre, T. Durroux, K.S. Sharma, G. Durand, B. Pucci, E. Trinquet, J.M. Zwier, X. Deupi, P. Bron, J.-L. Baneres, B. Mouillac, and S. Granier. 2012. Structural insights into biased G protein-coupled receptor signaling revealed by fluorescence spectroscopy. Proc. Natl. Acad. Sci. 109: 6733–6738.

17. Zidar, D.A., J.D. Violin, E.J. Whalen, and R.J. Lefkowitz. 2009. Selective engagement of G protein coupled receptor kinases (GRKs) encodes distinct functions of biased ligands. Proc. Natl. Acad. Sci. 106: 9649–9654.

18. Boguth, C.A., P. Singh, C. Huang, and J.J.G. Tesmer. 2010. Molecular basis for activation of G protein-coupled receptor kinases. EMBO J. 29: 3249–3259.

19. Katritch, V., V. Cherezov, and R.C. Stevens. 2013. Structure-Function of the G-protein-Coupled Receptor Superfamily. Annu. Rev. Pharmacol. Toxicol. 53: 531–556.

20. Shukla, A.K., G.H. Westfield, K. Xiao, R.I. Reis, L.-Y. Huang, P. Tripathi-Shukla, J. Qian, S. Li, A. Blanc, A.N. Oleskie, A.M. Dosey, M. Su, C.-R. Liang, L.-L. Gu, J.-M. Shan, X. Chen, R. Hanna, M. Choi, X.J. Yao, B.U. Klink, A.W. Kahsai, S.S. Sidhu, S. Koide, P.A. Penczek, A.A. Kossiakoff, V.L. Woods Jr, B.K. Kobilka, G. Skiniotis, and R.J. Lefkowitz. 2014. Visualization of arrestin recruitment by a G-protein-coupled receptor. Nature. 512: 218–222.

21. Kahsai, A.W., K. Xiao, S. Rajagopal, S. Ahn, A.K. Shukla, J. Sun, T.G. Oas, and R.J. Lefkowitz. 2011. Multiple ligand-specific conformations of the β2-adrenergic receptor. Nat. Chem. Biol. 7: 692–700.

22. Kang, Y., X.E. Zhou, X. Gao, Y. He, W. Liu, A. Ishchenko, A. Barty, T.A. White, O. Yefanov, G.W. Han, Q. Xu, P.W. de Waal, J. Ke, M.H.E. Tan, C. Zhang, A. Moeller, G.M. West, B.D. Pascal, N. Van Eps, L.N. Caro, S.A. Vishnivetskiy, R.J. Lee, K.M. Suino-Powell, X. Gu, K. Pal, J. Ma, X. Zhi, S. Boutet, G.J. Williams, M. Messerschmidt, C. Gati, N.A. Zatsepin, D. Wang, D. James, S. Basu, S. Roy-Chowdhury, C.E. Conrad, J. Coe, H. Liu, S. Lisova, C. Kupitz, I. Grotjohann, R. Fromme, Y. Jiang, M. Tan, H. Yang, J. Li, M. Wang, Z. Zheng, D. Li, N. Howe, Y. Zhao, J. Standfuss, K. Diederichs, Y. Dong, C.S. Potter, B. Carragher, M. Caffrey, H. Jiang, H.N. Chapman, J.C.H. Spence, P. Fromme, U. Weierstall, O.P. Ernst, V. Katritch, V.V. Gurevich, P.R. Griffin, W.L. Hubbell, R.C. Stevens, V. Cherezov, K. Melcher, and H.E. Xu. 2015. Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. Nature. 523: 561–567.

23. Rasmussen, S.G.F., B.T. DeVree, Y. Zou, A.C. Kruse, K.Y. Chung, T.S. Kobilka, F.S. Thian, P.S. Chae, E. Pardon, D. Calinski, J.M. Mathiesen, S.T.A. Shah, J.A. Lyons, M. Caffrey, S.H. Gellman, J. Steyaert, G. Skiniotis, W.I. Weis, R.K. Sunahara, and B.K. Kobilka. 2011. Crystal structure of the β2 adrenergic receptor–Gs protein complex. Nature. 477: 549–555.

24. Nygaard, R., Y. Zou, R.O. Dror, T.J. Mildorf, D.H. Arlow, A. Manglik, A.C. Pan, C.W. Liu, J.J. Fung, M.P. Bokoch, F.S. Thian, T.S. Kobilka, D.E. Shaw, L. Mueller, R.S. Prosser, and B.K. Kobilka. 2013. The Dynamic Process of β2-Adrenergic Receptor Activation. Cell. 152: 532–542.

25. Manglik, A., T.H. Kim, M. Masureel, C. Altenbach, Z. Yang, D. Hilger, M.T. Lerch, T.S. Kobilka, F.S. Thian, W.L. Hubbell, R.S. Prosser, and B.K. Kobilka. 2015. Structural Insights into the Dynamic Process of β2-Adrenergic Receptor Signaling. Cell. 161: 1101–1111.

26. Tikhonova, I.G., B. Selvam, A. Ivetac, J. Wereszczynski, and J.A. McCammon. 2013. Simulations of Biased Agonists in the β2 Adrenergic Receptor with Accelerated Molecular Dynamics. Biochemistry (Mosc.). 52: 5593–5603.

27. Altenbach, C., A.K. Kusnetzow, O.P. Ernst, K.P. Hofmann, and W.L. Hubbell. 2008. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. Proc. Natl. Acad. Sci. 105: 7439–7444.

28. Park, P.S.-H., D.T. Lodowski, and K. Palczewski. 2008. Activation of G Protein–Coupled Receptors: Beyond Two-State Models and Tertiary Conformational Changes. Annu. Rev. Pharmacol. Toxicol. 48: 107–141.

29. Smith, S.O. 2010. Structure and Activation of the Visual Pigment Rhodopsin. Annu. Rev. Biophys. 39: 309–328.

30. Rajagopal, S., S. Ahn, D.H. Rominger, W. Gowen-MacDonald, C.M. Lam, S.M. DeWire, J.D. Violin, and R.J. Lefkowitz. 2011. Quantifying Ligand Bias at Seven-Transmembrane Receptors. Mol. Pharmacol. 80: 367–377.

31. Förster, R., A.C. Davalos-Misslitz, and A. Rot. 2008. CCR7 and its ligands: balancing immunity and tolerance. Nat. Rev. Immunol. 8: 362–371.

32. Haessler, U., M. Pisano, M. Wu, and M.A. Swartz. 2011. Dendritic cell chemotaxis in 3D under defined chemokine gradients reveals differential response to ligands CCL21 and CCL19. Proc. Natl. Acad. Sci. 108: 5614–5619.

33. Schumann, K., T. Lämmermann, M. Bruckner, D.F. Legler, J. Polleux, J.P. Spatz, G. Schuler, R. Förster, M.B. Lutz, L. Sorokin, and M. Sixt. 2010. Immobilized Chemokine Fields and Soluble Chemokine Gradients Cooperatively Shape Migration Patterns of Dendritic Cells. Immunity. 32: 703–713.

34. Kohout, T.A., S.L. Nicholas, S.J. Perry, G. Reinhart, S. Junger, and R.S. Struthers. 2004. Differential Desensitization, Receptor Phosphorylation, -Arrestin Recruitment, and ERK1/2 Activation by the Two Endogenous Ligands for the CC Chemokine Receptor 7. J. Biol. Chem. 279: 23214–23222.

35. Bennetzen, M., D. Larsen, C. Dinant, S. Watanabe, J. Bartek, J. Lukas, and J.S. Andersen. 2013. Acetylation dynamics of human nuclear proteins during the ionizing radiation-induced DNA damage response. Cell Cycle. 12: 1688–1695.

36. Subramanian, C., M. Hada, A.W. Opipari, V.P. Castle, and R.P.S. Kwok. 2013. CREB-Binding Protein Regulates Ku70 Acetylation in Response to Ionization Radiation in Neuroblastoma. Am. Assoc. Cancer Res. 11: 173–181.

37. Hu, S., and F.A. Cucinotta. 2011. Computational studies on full-length Ku70 with DNA duplexes: base interactions and a helical path. J. Mol. Model. 18: 1935–1949.

38. Walker, J.R., R.A. Corpina, and J. Goldberg. 2001. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature. 412: 607–614.

CHAPTER 2: MOLECULAR MECHANISM OF BIASED LIGAND CONFORMATIONAL CHANGES IN CC CHEMOKINE RECEPTOR 7

2.1 Introduction

Conformational diversity in G protein-coupled receptors (GPCRs) is a fundamental component of the receptor's selective properties in binding intracellular effectors, such as G protein and arrestin (1). The conformational ensemble of the receptor comprises of distinguishable conformational states characterized by its seven-transmembrane (TM) helical arrangements, critical to coordinating the receptor's selective binding to intracellular (INC) effectors (1–3). To enable the complex function of a GPCR, bound ligands are capable of stabilizing the receptor in many distinct conformational states ranging from inactive to fully active, spanning G protein, arrestin, or GPCR kinase (GRK) biased conformations (1, 3–6). Additionally, receptor conformations are capable of discriminating between the different effector subtypes and arrestin binding modes (1, 7). These attributes emphasize the importance of the receptor's conformational diversity and the role of the ligands in shifting the receptor's equilibrium in sampling its different conformational states.

Here, we examine a key regulator of the adaptive immune response in the CC chemokine receptor family, CC chemokine receptor 7 (CCR7) (8). Its ligands, CCL19 and CCL21, have distinct roles in the homing and functional compartmentalization of T cells and antigen-presenting dendritic cells to and within the secondary lymph nodes as a result of their differential chemotactic behaviors (8–10). To initiate CCR7's cellular function, CCL19 and CCL21 have been shown to selectively induce distinct signaling pathways in the cell (8). While both ligands mediate their signaling through binding of intracellular G_i protein and GRK6 to CCR7, only CCL19 induces CCR7 internalization and desensitization through receptor phosphorylation by GRK3 and recruitment of β-arrestin 2 (4, 11). This differential binding of CCR7 to β-arrestin 2, GRK3, and GRK6 suggests the presence of selective conformational states in CCR7 induced by its biased ligands, CCL19 and CCL21.

Ligand binding to a GPCR drives the receptor to its corresponding conformational states through different molecular switches and helical rearrangements (12). Crystallographic and ^{19}F-NMR spectroscopy studies have identified specific TM helices involved in the INC rearrangements of the receptor. Specifically, upon ligand binding, TM5, TM6, and TM7 undergo large helical displacements to accommodate effector binding (1, 3, 13, 14). A recent study by Liu et al. shows that β2-adrenergic receptor explores two independent equilibrium conformations in TM6 and TM7 (1); where, ligands are capable of fine-tuning the relative population of equilibrium conformational states adopted by TM6 and TM7 to accommodate biased or unbiased binding of G protein and β-arrestin (1).

Despite these fundamental advances in characterizing the role of TM helices in selectively binding INC effectors, the role of the ligands and their associated molecular mechanisms and pathways in stabilizing such conformations, in particular the relative positions of TM5, TM6, and TM7, remains unclear. Previous studies using molecular dynamics (MD) simulations have focused on the activation mechanism in response to unbiased agonist (15–18), to the exception of a few studies that discuss "known microscopic characteristics" of the biased conformational states in β2-adrenergic receptor (19). However, to the best of our knowledge, detailed ligand-specific structural allosteric pathways have not been fully characterized in any GPCR, let alone chemokine receptors. Such structural detail is essential for not only gaining a deeper understanding of how ligands operate in a selective manner, but also aids in the rational design of drugs targeting desired signaling pathways.

In the present study, we apply conventional MD (cMD) using Anton (20) and accelerated MD (aMD) (21, 22) to delineate the conformational changes in CCR7 in response to its biased ligands, CCL19 and CCL21. From all simulations, we identify various molecular switches that undergo various ligand-specific conformational changes. Additionally, during the cMD simulation, we capture ligand-associated allosteric pathways, starting at the ligand-binding site, and propagating to these molecular switches. Highly correlated helical movements are coupled to the conformational changes occurring within the molecular switches and illustrate transitioning of the receptor between two distinct states.

2.2 Materials and Methods

Nomenclature. Residues are represented as a one-letter amino acid code and a number corresponding to their order in the sequence. Additional sequence information is represented in the superscripts. Superscripts of ligand residues denote the ligand it belongs to; and receptor residues are numbered according Ballesteros–Weinstein and convey the helix number and position of each residue relative to the most conserved residue in the helix (23).

System setup for molecular dynamics simulations. The structure of CCR7 and CCL19/CCL21-bound CCR7 were modeled after the newly determined crystal structure of the chemokine receptor CXCR4 bound to a viral chemokine antagonist vMIP-II (24) (PDB code 4RWS). The structures of the CCL19 and CCL21 ligands and CCR7 receptor were generated separately and were subsequently assembled into complexes using the vMIP-II-CXCR4 crystal structure as a template. First, the mouse sequences of CCR7, CCL19, and CCL21 were extracted from UniProt (25) (http://www.uniprot.org/). Alignments of receptor and ligand sequences with CXCR4 (PDB code 4RWS) and vMIP-II (PDB code 4RWS) respectively were then performed using the ClustalW2 (26, 27) server (http://www.ebi.ac.uk/Tools/msa/clustalw2/). Sequence alignments revealed a percent identity of 27% between vMIP-II and CCL19 and 23% between vMIP-II and CCL21. Despite the low sequence identity, vMIP-II was used as a template since all chemokine ligands have a common structural motif with a core domain (residues 10-83) that is structurally maintained by two disulfide bridges and a flexible N-terminal domain (28) (ligand-NTD) (residues 1-7). The core domain has a canonical fold: a N-loop, followed by three antiparallel β-strands, an α-helix, and a flexible C-terminal domain (CTD). The three β-strands are connected by two loops: the 30s-loop, and 40s-loop. Unlike CCL19, the full sequence of CCL21 comprises of a long and positively charged CTD that allows binding to glycosaminoglycans. Given that CTD does not interact with CCR7 and not involved in receptor activation (8), a truncated version of CCL21 (residues 1-83) was modeled along with the full sequence of CCL19 (residues 1-83) using Modeller9.11 (29). Both models assume a robust core forming the canonical chemokine motif and show low rmsd to their respective experimental structures: rmsd between CCL19 and CCL21 models and their respective NMR structures is 1.143 Å and 1.184 Å, respectively (28, 30).

Modeling of the ligands using vMIP-II was done to obtain three-dimensional structures of each ligand in their bound conformation to CCR7.

Sequence alignment of CCR7 with CXCR4 (sequence extracted from PDB file 4RWS) showed a 36% identity and 47% similarity (with 39% identity and 63% similarity for the transmembrane domain (TMD)). (Sequence similarities were calculated using SIAS [http://imed.med.ucm.es/Tools/sias.html], by taking into consideration the following physicochemical properties of aligned amino acids: aromatic (F, Y, W), hydrophobic (V, I, L, M, A, F, W), aliphatic (V, I, L), positively charged (R, K, H), negatively charged (D, E), polar (Not charged) (N, Q), or small (A, T, S)). The receptor is composed of three domains: the transmembrane (TMD, residues 26-305), N-terminal (CCR7-NTD, residues 1-25), and C-terminal (CTD, residues 306-354) domains that have different structural motifs and were therefore modeled separately. The CCR7-NTD is an intrinsically disordered segment of the receptor, that could not be detected in several crystallographic and NMR studies of many GPCRs (24, 31–34). Therefore, the CCR7-NTD was modeled using Modeller9.11 in a random coil conformation and was positioned away from the extracellular loops to avoid clashes between the CCR7-NTD and the ligand when positioning CCL19 and CCL21. In contrast, the CTD contains an α-helical sequence motif F(RK)xx(FL)xxx(LF) and an intrinsically disordered segment (32). Given the success of i-TASSER in predicting protein structure, the full CTD sequence was modeled using the i-TASSER server (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) (35, 36). The top generated model had a relatively low confidence score of -2 (C-score range [-5, 2]), which is due to the intrinsically disordered segment of the CTD. However, more importantly, the produced i-TASSER model still generated the canonical helical domain, known as helix 8 (residues 309-319), present in many other GPCRs such as CCR5 (32) (PDB code 4MBS), CXCR1 (34) (PDB code 2LNL) and β2-adrenergic receptors (14, 33, 37, 38) (PDB codes 3P0G, 3SN6, 3NY8, 2RH1). Lastly, The TMD of CCR7 is composed of 7 TM helices (TM1: residues 27-61; TM2: residues 68-96; TM3: residues 102-134; TM4: residues 144-172; TM5: residues 194-228; TM6: residues 239-267; TM7: residues 274-303) connected by three extracellular loops (ECL) and three intracellular loops (ICL). Templates exhibiting more than 35% sequence homology can be used to generate highly reliable GPCR models (39). Therefore, with a 39% sequence identity, CXCR4 was used as a template to generate a three-dimensional structure of CCR7's TMD (residues 19-303) using Modeller9.11 (29).

The modeled ligands and receptor were then used to construct the complex structures following the published binding mode of the CXCR4-vMIP-II crystal structure (24) (PDB code 4RWS). The ligands bind the receptor via a two-site mechanism where the CCR7-NTD binds the core domain of the ligand, site I, followed by ligand-NTD binding inside the receptor's extracellular pocket, site II (24). For site I, various NMR studies of chemokine ligands with their receptor NTDs show similar interactions where the receptor-NTD interacts with the N-loop, 40s loop, and third β-strand of the ligand (28, 30, 40, 41). Therefore, chemokine-binding to its receptor involves two experimentally determined structural constrains: the receptor-NTD and extracellular binding pocket in sites I and II respectively. Published crystallographic structures of chemokine-GPCR complexes (PDB 4RWS and 4XT1/3) show very similar binding modes (root-mean-square deviation (rmsd) of 1.3 Å) and comply with the aforementioned structural constraints (24, 42). In light of the new chemokine-GPCR structures, CCL19 and CCL21 were positioned using CXCR4-vMIP-II crystal structure (PDB code 4RWS) (24) as a template using Chimera (43). The binding mode complies with previously published NMR chemical shift perturbations between CCR7-NTD and its ligands CCL19 and CCL21 (28, 30). These chemical shift perturbation studies, carried by Love et al, indicate the presence of a binding interface in each ligand (N-loop, 40s loop, and third β-strand) that interacts with the CCR7-NTD.

To assess the stability of the receptor terminal domains, we calculate rmsd time series of both domains (CCR7-NTD and CCR7-CTD) in all of our simulations (cMD and aMD for each of the two complexes). All rmsd plots show stabilization of both domains in all simulations (Figure A.1). Furthermore, contact maps between the CCR7-NTD and the ligand calculated from both aMD and Anton simulations reveal very close agreement with NMR data of CCL19 and CCL21 with CCR7-NTD (Figure A.2) (28, 30).

In site II, both ligand-NTDs adopt different poses in the receptor-binding pocket during our simulations (Figure A.3). Figure A.3 displays contacts between CCR7 and its ligand-NTDs at less than 5 Å and a residence time of 50% or more of the equilibrated cMD recorded frames (equilibrated time domain are specified in Figure 2.1 caption). CCL21 N-terminus forms a salt bridge with E169$^{4.60}$, which facilitates the ligand's interaction with TM4, TM5, and TM6. In addition, D2$^{\mathrm{CCL21}}$ forms a salt bridge with the second extracellular loop (ECL2), R185$^{\mathrm{ECL2}}$; and D6$^{\mathrm{CCL21}}$ interacts with K26$^{1.24}$ and R30$^{1.28}$. Unlike CCL21, CCL19 forms no interactions with TM4, TM5, and TM6 and its charged residues are stabilized with the charge-complementary residues in TM1: D4$^{\mathrm{CCL19}}$ interacts with K26$^{1.24}$, R30$^{1.28}$, K33$^{1.31}$, and E94$^{1.64}$; E6$^{\mathrm{CCL19}}$ interacts with R185$^{\mathrm{ECL2}}$ and E181$^{\mathrm{ECL2}}$; D7$^{\mathrm{CCL19}}$ interacts with E181$^{\mathrm{ECL2}}$, and K26$^{1.24}$; CCL19's N-terminus interacts with E94$^{1.64}$, D285$^{7.36}$, R30$^{1.28}$, and K33$^{1.31}$.

Our generated complex models were equilibrated in a model of a lipidic membrane patch through our multi-step molecular dynamics simulations protocol described below and subsequently used to initiate long timescale MD simulations.

Conventional molecular dynamics simulations. Equilibration MD simulations of CCR7-CCL19 and CCR7-CCl21, were performed using NAMD, version 2.9 (44). Initial protein structure files were prepared using the PSFGEN utility in VMD (45) and the CHARMM36 forcefield (46–49). All disulfide bonds were maintained during the simulations: two disulfide bonds in CCR7 (C24-C274 and C105-C186), two disulfide bonds in CCL19 (C8-C34 and C9-C50), and two in CCL21 (C8-C34 and C9-C52). The receptor region embedded in the membrane was determined using the Positioning of Proteins in Membrane (PPM) server (http://opm.phar.umich.edu/server.php) and used to position the lipid bilayer around the receptor (50). The palmitoyl-oleoyl-phosphatidyl-choline (POPC) bilayer was generated using the Membrane plugin in VMD (45) and all overlapping lipid molecules within 1 Å were removed. The lipid-protein system was embedded into a water box using the VMD utility SOLVATE and the TIP3P model for the water molecules. The water box dimensions of the ligand bound receptor was 105 Å × 105 Å × 120 Å respectively. The system was neutralized using sodium and chloride counterions at an ionic strength of 150 mM. The final ligand bound CCR7 systems contained ~270 lipid molecules, ~200 Na^+, ~214 Cl^-, and ~35,000 water molecules, for a total of ~149,000 atoms.

NAMD (44) was used to equilibrate the system in a series of minimization, melting, and equilibration steps at 1 atm pressure and 310 K. The system was freely minimized for 1000 steps of conjugate gradient energy minimization. The protein, solvent, ions, and lipid heads were then fixed and the system was minimized for 1000 steps and simulated for 0.5 ns to allow the lipids tails to equilibrate at 1 fs/step. The following simulation series were run at 2 fs/step. The system was simulated with protein restrained at 5 kcal/mol to allow the environment to relax in 1000 steps minimization and 0.5 ns simulation. The system was then subjected to a minimization step for 1000 steps and five equilibration stages (1 ns each) with all protein atoms harmonically constrained (using force constants of 5, 4, 3, 2, and 1 kcal/mol/Å2, respectively) to their post-minimization positions, and a final unconstrained equilibration stage of 15 ns.

All simulations were performed using periodic boundary conditions and particle-mesh Ewald electrostatics for long-range electrostatic interactions with a grid point density of 1/Å. Nonbonded van der Waals interactions and short-range electrostatic interactions were calculated with an interaction cutoff of 12 Å and switching distance of 10 Å. The SHAKE algorithm was employed to fix the length of all hydrogen-containing bonds, enabling the use of 2 fs integration time steps. Coordinates were sampled every 2 ps.

Final output velocities, dimensions, and coordinates from equilibration NAMD simulations were used as input to simulate our systems on Anton (20), a special purpose supercomputer for biomolecular simulation designed and constructed by D. E. Shaw Research (DESRES). All Anton simulations were performed under the NPT ensemble using a multigrator (51) (310K using a Nosé-Hoover thermostat and an isotropic pressure of 1 atm using the Martyna-Tobias-Klein barostat). Multigrator with thermostat interval of 24 ps and barostat interval of 240 ps were used. All bond lengths to hydrogen atoms were constrained using the M-SHAKE. A RESPA integrator was used with a time step of 2 fs for bonded, VDW, and short-range electrostatic interactions, and 6 fs for long-range electrostatic interactions. Long-range electrostatic interactions were handled with the k-space Gaussian Split Ewald (GSE) method and a 64 × 64 × 64 grid. Interaction parameters such as GSE parameters, and nonbonded cutoffs were determined systematically using Anton scripts, designed to optimize accuracy and performance. For CCL19-CCR7, $\sigma = 3.19$ Å, $\sigma s =$

1.92 Å, Rcut = 13.76 Å, Rspread = 8.30 Å; and for CCL21-CCR7, σ = 3.17 Å, σs = 1.84 Å, Rcut = 13.67 Å, Rspread = 7.97 Å. Initial configuration of our three systems produced root mean-squared force errors of no more than 0.0023 kcal/mol/Å.

Accelerated molecular dynamics simulations. Accelerated molecular dynamics simulations were performed using NAMD2.9 at the dual-boost acceleration level. The dual-boost applies a boost potential to all dihedral angles and all atoms in the system using the input parameters: $E_{dihed} = V_{dihed\_avg} + 0.3*V_{dihed\_avg}$, $\alpha_{dihed} = 0.3*V_{dihed\_avg}/5$, $E_{total} = V_{total\_avg} + 0.2*Natoms$, and $\alpha_{total} = 0.2*Natoms$. Natoms is the number of atoms in each system; and V_{dihed_avg} and V_{total_avg} are the average dihedral and total energies respectively, extracted from 45 ns equilibration cMD simulations performed as described above. Accelerated MD simulations were performed for 150 ns each by restarting from the 45 ns equilibration conventional MD simulations (21, 22).

Analysis Protocols. System snapshots were extracted at a rate of 180 ps and 2 ps during Anton and aMD production simulations, respectively. All frames have been analyzed to extract different measures and create time series of data as shown in figures. These measures include hydrogen bond, torsion angle, and atomic distance calculations. Analysis of the MD trajectories was performed with in-house scripts using R programming language (52), Python (53), Chimera (43), the Bio3D library (54, 55), and TimeScapes (56).

Hydrogen bonds were calculated with Chimera, using hydrogen bond criteria as described in ref. (43, 57). Backbone and side-chain torsion angles were calculated using Bio3D. Atomic distances were calculated between non-hydrogen side-chain atoms using TimeScapes (56) with a cutoff of 5 Å. Interacting residue distance time series are determined by calculating the minimum distances of the set of atomic distance time series of two interacting residues. Calculating side-chain contacts using the minimum-distance interacting atoms allows for an accurate separation distance between both residues, as compared to using the centroid or a representative atom of the side-chain. Distance time series between Cα atoms were also calculated using TimeScapes (56). Distance probability distributions of time series were calculated by computing Gaussian kernel density estimates using R. Percent occupancy of an interaction was calculated as the percentage of frames in which an interaction is within 5 Å in the MD trajectory. Pairwise cross-correlation was calculated between all Cα distance time series using Bio3D (54, 55). Cross-correlation coefficient cutoff of 0.95 was used to cluster correlated Cα distance time series; and correlated time series that show a coupled behavior to the different measures in Figure 2.2 and Figure 2.5 are extracted as ligand-induced global helical motions. Rmsds are calculated using Bio3D based on residue side-chains or Cα carbons distance time series rather than atomic coordinates.

Π-stacking interactions. Aromatic residue trimers within proteins in the protein data bank (PDB) have been studied extensively and found to form two distinct geometrical clusters: symmetrical and ladder conformations (58). To monitor the π-stacking arrangement within the tyrosine triad, we measured the angle formed between the C_ζ carbons in each of the tyrosines, and was found to adopt an obtuse (between 90° and 120°) and acute (around 60°) θ-angle for ladder and symmetric conformations, respectively (58). Molecular mechanics studies and PDB surveys indicate a π-stacking interaction upper-limit distance of 7.5 Å (distance between benzene ring centroids), where the energy drops below the Boltzmann temperature factor (58, 59).

Sequence alignment. The mouse sequences of all CC chemokine receptors were extracted from UniProt (25) (http://www.uniprot.org/). Alignment of all receptors with CCR7 was then performed using the Clustal Omega server (60) (http://www.ebi.ac.uk/Tools/msa/clustalo/). Sequence logos are generated using the WebLogo3 (61, 62) (http://weblogo.berkeley.edu/logo.cgi). Mouse proteins were selected to enable the use of our study to generate testable hypotheses for experimental animal models.

2.3 Results

With the recent available chemokine-bound crystal structure of CXCR4 (24) (vMIP-II-bound CXCR4, PDB code 4RWS), we set out to study the structural mechanism of CCR7 dynamics in response to its native biased chemokine ligands, CCL21 and CCL19. Using the vMIP-II-CXCR4 complex, we modeled the CCL21 and CCL19-bound structures of CCR7 (see Methods), and using Anton, we then performed cMD simulations on the CCL19-CCR7 and CCL21-CCR7 structures for 7 µs each, and we simulated both systems using dual-boost aMD for 150 ns each. Accelerated MD was designed as an enhanced sampling MD method and shown to be able to capture microsecond timescale events in various molecular systems (21, 63–65). Using both MD methods, we aim to identify microsecond timescale conformational changes within the TMD of the receptor in response to each of its native biased ligands.

Chemokine ligands carry their function by binding to the extracellular region of the receptor following a two-site model (24). The core domain of the ligand interacts with the extracellular loops (ECL) of the receptor, ensuring specificity of the ligand to its receptor, and the N-terminal domain of the ligand (ligand-NTD: residues 1-7 preceding the first two conserved cysteines) interacts inside the receptor's extracellular binding pocket, and carries receptor activation (28, 66, 67). This is emphasized further in our system by experimental truncation of the CCL19-NTD which has converted the ligand to an antagonist; confirming that the ligand core domain is not involved in receptor activation (67). Therefore, we delineate the different binding poses adopted by both ligand-NTDs in the receptor's binding pocket (Figure A.3) and their effect on receptor dynamics.

Given the physicochemical properties of the receptor binding pocket and ligand-NTD, both CCL21-NTD and CCL19-NTD stabilize into different binding poses (Figure A.3). The electrostatic profile of both ligand-NTDs displays different positions of its charged residues. CCL21 ($S^1DGGGQD^7CC$) has three charges distributed on either sides of the CCL21-NTD heptapeptide, with one negatively charged residue at position 7 (D) and two charged residues, one positive at the backbone N-terminus, and one negative at position 2 (D). However, CCL19 ($G^1ANDAED^7CC$) has four charged residues distributed uniformly along the CCL19-NTD heptapeptide, with one N-terminal positive charge, and three negatively charged residues at positions 4 (D), 6 (E), and 7 (D). Correspondingly on the receptor side, CCR7 displays two distinct electrostatic regions in its binding pocket, one formed by a salt bridge between TM3 and TM4 ($K113^{3.33}$ and $E169^{4.60}$), and another formed within TM1, TM2, and TM7 ($K26^{1.24}$, $K27^{1.25}$, $R30^{1.28}$, $E94^{2.64}$, $D285^{7.36}$) (Figure A.3C). Consequently, both ligand-NTDs stabilize in different binding pocket of the receptor by forming complementary electrostatic interactions (Figure A.3). CCL21's missing middle negative charge, together with the flexibility of the three consecutive glycines, allows its CCL21-NTD to stretch across and interact with both electrostatic patches in the receptor, while the presence of the middle charge, D4, anchors the CCL19-NTD to interact only with TM1, TM2, TM3, and TM7. The binding of each ligand-NTD in different pockets in CCR7 agrees with experimental mutagenesis data obtained by Ott et al. showing the differential effect of K113$^{3.33}$ mutation to alanine in CCR7, where the mutation affects binding and activation by 3.5- and 22-fold, respectively, in response to CCL21 but shows no effect to CCL19 (66).

CCL21 and CCL19 induce different conformational changes within ligand-specific molecular switches in CCR7. From both conventional and accelerated molecular dynamics of CCL21-bound and CCL19-bound CCR7 complexes, we depict multiple regions in the transmembrane domain (TMD) of CCR7 that show distinct conformational behavior (Figure 2.1). These functionally relevant regions act as molecular switches in the receptor and include: the tri-tyrosine switch (Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$), the tri-phenylalanine switch (F116$^{3.36}$, F208$^{5.47}$, and F248$^{6.44}$), and the polar bridge (Q252$^{6.48}$ and R294$^{7.45}$). Each can take on multiple conformations, and each can adopt ligand-specific conformations, which demonstrates the ligand's ability to regulate the arrangements within its molecular switches.

The multimodal distributions in Figure 2.1 illustrate the effect that each ligand has on the various conformational states sampled by each of the molecular switches. Within the tri-tyrosine switch, all simulations show the formation of π-stacking interactions (Figure 2.1C) following the loss of a hydrogen bond between Y112$^{3.32}$ and Q252$^{6.48}$. (π-stacking interactions are monitored by measuring the angle θ formed between the C_ζ carbons in each of the tyrosines.) The bimodal distributions of the population density plots of angle θ indicate the presence of conformational changes that result in the formation and loss of π-stacking interactions (Figure 2.1C). These conformational changes result in equally distributed sampling of both states in the CCL21-bound simulations, while the equilibrium is shifted towards the loss of π-stacking interactions in the CCL19-bound simulations. Similarly, only the CCL21-bound simulations display hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$ (Figure 2.1B). Probability density plots in Figure 2.1B show a 3-Å peak associated with hydrogen bond formation in the CCL21 and not CCL19 simulations. Moreover, only the CCL21-bound simulations involve conformational changes within the tri-phenylalanine switch, where the interaction between F116$^{3.36}$ and F248$^{6.44}$ is lost, moving from 4 Å to 7 Å in the CCL21-bound but not the CCL19-bound simulations (Figure 2.1D). Lastly, Figure 2.1E displays the polar interaction between Q252$^{6.48}$ and R294$^{7.45}$ as another characteristic molecular switch. This polar bridge resides deeper in the TMD of the receptor and its multi-peak distributions show higher variability than the previous switches and a clear difference in arrangements between the CCL21-bound and CCL19-bound simulations (Figure 2.1E). Both cMD and aMD simulations illustrate the effect of the ligand on each of the molecular switches. In particular, the aMD simulations provide independent reproducibility of conformational changes is each of the ligand-bound systems. Here, probability density plots of the aMD simulations in Figure A.4 show similar ligand-dependent sampling of the conformational states of each of the molecular switches including: hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$ in CCL21 and not CCL19 (Figure A.4A), equally distributed sampling of both gain and loss of π-stacking interactions in the CCL21, while the equilibrium is shifted towards the latter in the CCL19 (Figure A.4B), and clear difference in arrangements within the polar bridge (Figure A.4D).

The different positions of the ligand-NTDs play a critical role in initiating different allosteric mechanisms in the receptor by disrupting and establishing specific contacts within the binding pocket. In that aspect, the Anton simulations capture various allosteric events that are clearly coupled to the transitions between the different conformational states in the molecular switches. In contrast, even though the aMD simulations were able to reproduce the different conformational changes within the molecular switches (Figure A.4), they display uncoupled and stochastic variability within these switches that lacks the correlation documented in the Anton simulations (Figures 2.2 and 2.3). This is owed to the fact that the aMD consists of an enhanced sampling method that applies a dual-boost to the all atoms and dihedral angles in the system (21, 65). The energy boost might disrupt the fine-tuned transitions within the receptor and blur the energy gaps governing the conformational transitions observed in the Anton cMD.

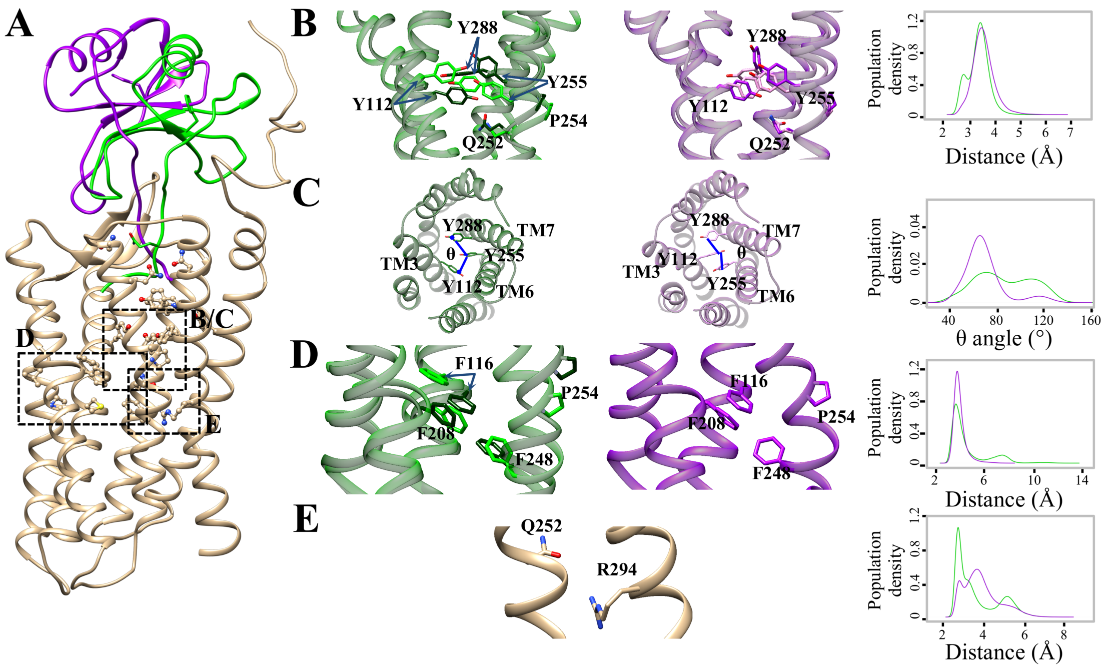


Figure 2.1 Molecular switches in the CCR7 transmembrane domain adopt ligand-specific conformations. Molecular switches with multiple states in our simulations are represented in dark and light colors in their corresponding molecular graphics. The color code in the molecular graphics and population density plots is green for CCL21-bound and purple for the CCL19-bound simulations. (A) Molecular graphics of CCR7 bound to CCL21 and CCL19 with the three molecular switches outlined in boxes. Residues involved in the switches are shown as stick models. Each highlighted region is labeled with a letter corresponding to panels (B)-(E). (B) Side view of representative structures of the tri-tyrosine switch. Side-chain distances between $Y112^{3.32}$ and $Y255^{6.51}$ are plotted as population densities. (C) Top view of representative structures of the tri-tyrosine switch in the CCL21- and CCL19-bound CCR7 simulations illustrating π-stacking interactions. θ is the angle formed by the C_ζ atoms of the three tyrosines. Population densities of the θ angle are plotted. (D) Representative structures of the tri-phenylalanine switch in the CCL21-bound and CCL19-bound CCR7 simulations. Side-chain distances between $F116^{3.36}$ and $F248^{6.44}$ are plotted as population densities. (E) Molecular graphics of the polar bridge. Side-chain distances between $Q252^{6.48}$ and $R294^{7.45}$ are plotted as population densities for the CCL21- and CCL19-bound CCR7 simulations. Criteria for inter-residue distances shown as population densities in all figures are outlined in Methods.

Both ligand-bound structures are each simulated for 7 μs using cMD on the Anton supercomputer (20). Rmsd calculations of CCR7 display an initial increase related to the stabilization of the initial receptor conformations used in both simulations (Figure A.5B). Subsequently, CCL19-bound CCR7 shows a stable structure in the remainder of the simulation (Figure A.5B, right panel), while CCL21-bound CCR7 exhibits a sharp increase at 4 μs related to a large conformational transition in the receptor (Figure A.5B, left panel).

Conformational changes within the characterized molecular switches are coupled to different allosteric events initiated by CCL21-NTD. During the 7-μs CCL21-bound cMD simulation, we capture multiple conformational changes in suboptimal hydrogen bonds and van der Waals tertiary interactions. These types of interactions represent two of the main physical forces stabilizing membrane protein structures, and changes within these interactions constitute allosteric events that are critical for signal propagation in the receptor (31). The time series of all detected allosteric events are listed in Figure 2.2 and clearly depict different conformational states separated at 1.9 μs, 4 μs, and 5 μs. These states represent the initial (0-1.9 μs), intermediate (1.9-5 μs), and final (5-7 μs) states of all depicted events. The intermediate state is a transitional phase where different arrangements occur at different times and is further divided to two intermediate substates I (1.9-4 μs) and II (4-5 μs).

At the ligand-binding pocket, CCL21 induces the formation of a hydrogen-bonding network within $Q6^{CCL21}$, $N266^{6.62}$, $Q262^{6.58}$, and $N281^{7.32}$ (Figure 2.2B, and Figure A.6). The formation of the critical hydrogen bond between $Q262^{6.58}$ and $N281^{7.32}$ is synchronized to various other events in the receptor at 4 μs (Figure 2.2).

The stabilization of the hydrogen bond is associated with a decrease in distance between TM6 and TM7 at the ligand-binding pocket, which in turn prompts an equilibrium shift in the backbone ϕ-torsion angle of $P254^{6.50}$. CCR7 has multiple prolines in its TM helices forming helical kinks. The highly conserved $P254^{6.50}$ in TM6 is part of the WxPF/Y motif and is positioned in the middle of the helix and considered an important dynamic component in the rearrangements of helices in receptor activation (6, 68). In Figure 2.2C, a gradual transition of $P254^{6.50}$ ϕ-torsion angle from a bimodal to a unimodal distribution is illustrated in the population density of torsion angles at different states. The initial state has a clear bimodal distribution of ϕ-torsion angles around -70° and -45°. Then, at intermediate state I, there is an equilibrium shift towards an angle distribution around a torsion angle of -70°, which in turn fully stabilizes to a clear unimodal distribution around -75° at 4 μs.

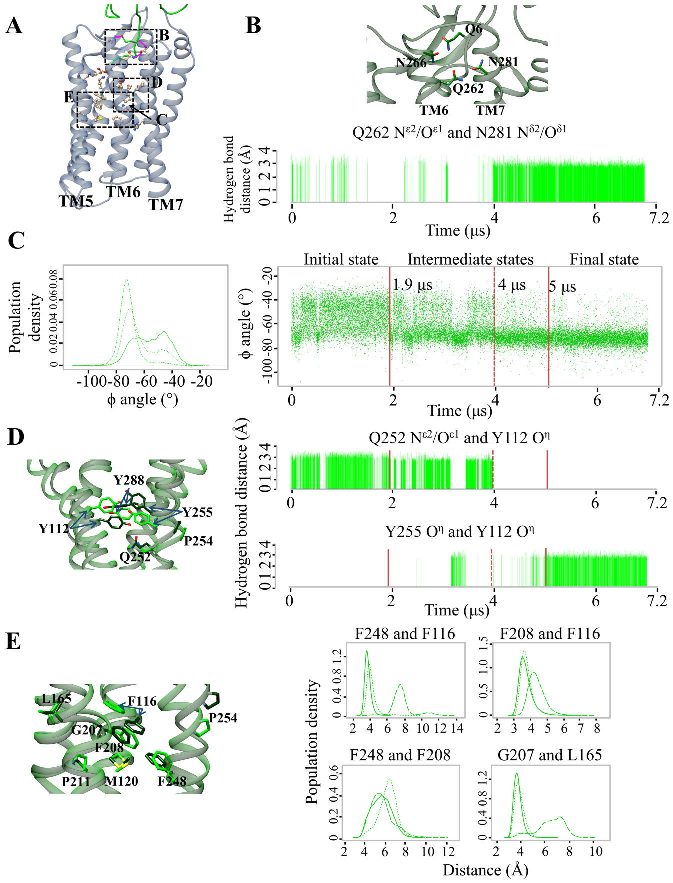


Figure 2.2 During the cMD simulation, CCL21 induces a series of allosteric events that prompt equilibrium shifts in the discrete conformational states of CCR7's molecular switches. (A) Molecular graphics of CCL21-bound CCR7 structure. Residues involved in the various allosteric events are shown as stick models and are outlined. Each highlighted region is labeled with a letter corresponding to panels (B)-(E). (B) Q262$^{6.58}$, N266$^{6.62}$, N281$^{7.32}$, and Q6$^{\mathrm{CCL21}}$, shown as stick models, are involved in a hydrogen-bonding network in the ligand-binding site. The molecular graphics is a representative structure of the dominant conformation showing a hydrogen bond between Q262$^{6.58}$ and N281$^{7.32}$ (4-7 μs). The bar plot displays the hydrogen bond distance time series between Q262$^{6.58}$ and N281$^{7.32}$ side-chains. The bar plots of remaining hydrogen bonds are displayed in Figure A.6. (C) ϕ-torsion angle of P254$^{6.50}$ population density (left) and time series (right) plots. The time series is broken down to four time segments labeled accordingly as initial, intermediate (I and II) and final states. Distributions of the ϕ-torsion angle are plotted for the following time segments: solid line (0-1.9 μs), dashed line (1.9-4 μs), and a dash-dot line (4-7 μs). (D) Molecular graphics showing representative structures of the dominant conformation in intermediate state I (1.9-4 μs) in dark green and the final state (5-7 μs) in light green. Residues involved in the tri-tyrosine switches and neighboring allosteric events are shown as stick models and labeled accordingly. Bar plots display the hydrogen bond distance time series of Y112$^{3.32}$ with Q252$^{6.48}$ and Y255$^{6.51}$. Time series ranges are divided as shown in panel C and indicated by the vertical red lines. (E) Molecular graphics showing the tri-phenylalanine switch region using the same conformations as in panel D. Side-chain distance population densities are plotted for different time segments of the simulation as shown in panel C. The atoms used to calculate the hydrogen bond distance are marked in panels B and D. Data for CCL19-bound cMD simulation, corresponding to panels B, C, D, and E are displayed in Figure A.8.

On the same TM6 helix, P254$^{6.50}$ is flanked by two critical residues: Q252$^{6.48}$ and Y255$^{6.51}$. Both residues are capable of forming hydrogen bonds with Y112$^{3.32}$ on TM3. Throughout the simulation, the Y112$^{3.32}$ side-chain hydrogen bond transitions from Q252$^{6.48}$ to Y255$^{6.51}$ (Figure 2.2D). This transition occurs twice during the simulation (at 3.1 μs and 4 μs) and is coupled with the bimodal to unimodal transition observed in P254$^{6.50}$ ϕ-torsion angle (Figure 2.2C, and 2.2D). The disruption of a hydrogen bond between Y112$^{3.32}$ and Q252$^{6.48}$ is associated with a ϕ-torsion angle of -75° in P254$^{6.50}$. At 4 μs, we observe a disruption of the hydrogen bond between Y112$^{3.32}$ and Q252$^{6.48}$ occurring simultaneously with the proline ϕ-torsion angle adjustment followed by the second intermediate state experiencing a microsecond delay before the formation of a hydrogen bond between Y112$^{3.32}$ and Y255$^{6.51}$.

During that delay, we detect the formation of π-stacking interactions within the tri-tyrosine switch, Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$, between 4 and 5 μs (Figure A.7A). This interaction stabilizes the transitional state resulting from both, the loss of the hydrogen bond between Y112$^{3.32}$ and Q252$^{6.48}$, and the adjustment of P254$^{6.50}$ ϕ angle. In the CCL21-bound cMD simulation, a wide distribution of angles in intermediate state I indicates a disordered arrangement between the three tyrosines and a lack of π-stacking interactions (Figure A.7A). Then, these tyrosines are stabilized in intermediate state II by forming π-stacking interactions, where Y255$^{6.51}$ interacts with Y288$^{7.39}$ and Y112$^{3.32}$ at distances of 5.5 Å and 6 Å, respectively (less than the upper-limit of 7.5 Å (58, 59)). These π-stacking interactions adopt a ladder conformation where θ has an angle distribution around 110°. At 5 μs, π-π interactions between Y255$^{6.51}$ and Y288$^{7.39}$ are maintained. However, centroid distance between Y112$^{3.32}$ and Y255$^{6.51}$ increase to 7.5 Å, indicating a loss in π-stacking (Figure 2.2D and Figure A.7A). In the final state, the tri-tyrosine π-stacking interactions are lost, resulting in hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$.

Another functionally relevant feature in CCR7 is the tri-phenylalanine switch, which resides within the vicinity of the tri-tyrosine switch and is composed of three phenylalanines (F116$^{3.36}$, F208$^{5.47}$, and F248$^{6.44}$) in TM3, TM5, and TM6 (Figure 2.2E). At 4 μs, F116$^{3.36}$ reorients its side-chain away from F248$^{6.44}$ (increasing the distance from 4 Å to 7.5 Å) as well as F208$^{5.47}$ (increasing the distance from 3.5 Å to 4.5 Å). Meanwhile, we also observe changes within TM4 and TM5 where L165$^{4.56}$ and G207$^{5.45}$ show a clear change in both, distance distribution and variance after 4 µs.

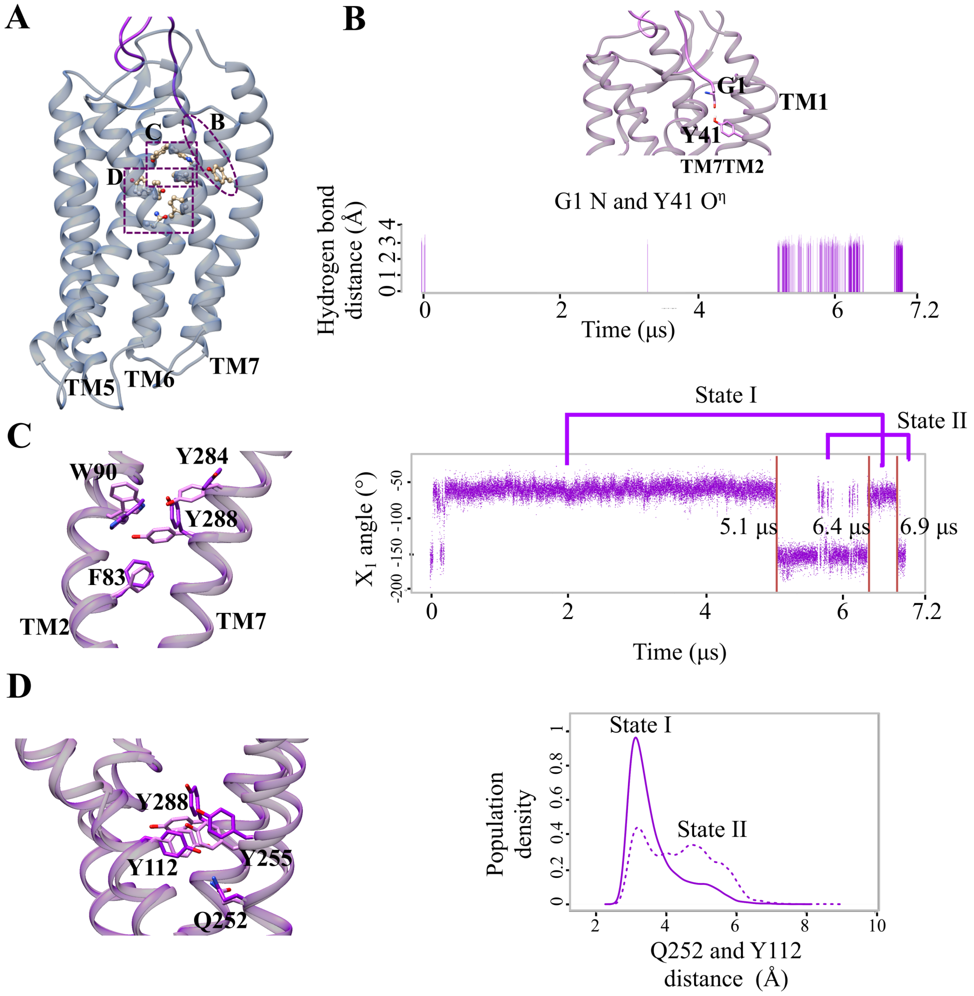


Figure 2.3 During the cMD simulation, CCL19 induces a series of allosteric events that prompt equilibrium shifts in the discrete conformational states of CCR7's tri-tyrosine switch. (A) Molecular graphics of CCL19-bound CCR7 structure. Residues involved in the various allosteric events are shown as stick models and outlined. Each highlighted region is labeled with a letter corresponding to panels (B)-(D). (B) G1^{CCL19} and Y41$^{1.39}$ are shown as stick models as a representative structure of the dominant conformation displaying the hydrogen bond between G1^{CCL19} and Y41$^{1.39}$ (5.1-6.4 µs). The bar plot displays the hydrogen bond distance time series. The atoms used to calculate the hydrogen bond distance are marked top of the panel. (C) Molecular graphics showing representative structures of the dominant CCL19-induced conformation in state I (dark purple) and state II (light purple). The χ_1-torsion angle of Y288$^{7.39}$ time series is plotted. The time series is broken down to two time segments labeled accordingly. (D) Molecular graphics shows a side view of the tri-tyrosine switch region using the same conformations as panel C. Y112$^{3.32}$, Y255$^{6.51}$, Y288$^{7.39}$, and Q252$^{6.48}$ are shown as stick models and labeled accordingly. Q252$^{6.48}$ and Y112$^{3.32}$ distance population densities are plotted for both CCL19 states.

In contrast to the allosteric events observed in the CCL21-bound simulation, the CCL19-bound cMD simulation do not exhibit any of the events characterized in CCL21 and all torsion angles and interactions appear stable throughout the entirety of the simulation with the exception of the hydrogen bond between Y112$^{3.32}$ and Q252$^{6.48}$ explored below (Figure A.8). This difference in receptor behavior is a result of the differential binding and biased nature of both ligands.

Conformational changes within the characterized molecular switches are coupled to different allosteric events initiated by CCL19-NTD. Despite the stability of CCR7's rmsd in the CCL19-bound cMD simulation (Figure A.5B), at 5.1 μs, we observe a brief series of side-chain rearrangements (Figure 2.3). Starting at the ligand-binding pocket, we observe the formation of a hydrogen bond between CCL19's N-terminus and Y41$^{1.39}$ side-chain at 5.1 μs and again 6.9 μs. The hydrogen bond formation induces a series of synchronized allosteric events in the TMD of CCR7. The time series listed in Figure 2.3 clearly depict different conformational states of CCR7's tri-tyrosine switches (states I and II).

The hydrogen bond formation between G1^{CCL19} and Y41$^{1.39}$ is synchronized with a change in the χ_1-torsion angle of Y288$^{7.39}$ from -70° to -170° (Figures 2.3B and 2.3C). The side-chain orientation of Y288$^{7.39}$ allows for a reorganization of the tri-tyrosine switch (Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$) to form π-stacking interactions, where Y112$^{3.32}$ interacts with Y288$^{7.39}$ and Y255$^{6.51}$ at distances of 6 Å or less (less than the upper-limit distance of 7.5 Å (58, 59)). We further measure the θ-angle formed by the C_ζ carbons of each of the three tyrosines to identify the symmetrical/ladder arrangement of π-stacking interactions (Lanzarotti et al. 2011) (Figure A.7B). The difference in angle between states I and II illustrates the transition of π-stacking interactions from symmetrical in state I (angles sampled around 60°) to ladder in state II (angles sampled around 140°). This transition prompts a distance increase between Y288$^{7.39}$ and Y255$^{6.51}$ to 9 Å (58, 59). The π-stacking interactions reached in response to CCL19 show similar arrangement to the tri-tyrosine switch in CCL21-bound cMD simulation (Figure A.7A).

In CCL21-bound cMD simulation, intermediate state II is associated with the loss of a hydrogen bond between Y112$^{3.32}$ and Q252$^{6.48}$ and result in the formation of the tri-tyrosine ladder π-stacking interactions. Similarly, in our CCL19-bound cMD simulation state II, even though the hydrogen bond between Y112$^{3.32}$ and Q252$^{6.48}$ persists (Figure A.8C), we see an equilibrium shift in the side-chain distance between both residues. Population density of the side-chain distance between Y112$^{3.32}$ and Q252$^{6.48}$ shows an increase in the distance sampled (population increase for a distance of more than 4 Å in Figure 2.3D), illustrating a weakening and loss of hydrogen bonding in a large portion of state II. The characterized transitions in the allosteric sites occur twice during in the simulation (at 5.1 μs and 6.9 μs), highlighting the coupled nature of these allosteric events (Figure 2.3). However, hydrogen bond loss between Y112$^{3.32}$ and Q252$^{6.48}$ is loosely coupled to the remaining allosteric events and only shows a slight equilibrium shift. This results in an even weaker coupling to the hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$ characterized in the CCL21-bound simulation. Indeed, the hydrogen bond between Y112$^{3.32}$ and Y255$^{6.51}$ is not formed in the CCL19-bound cMD and is briefly present during the aMD.

Both ligands induce large conformational changes in CCR7's helical domains. In CCL21-bound cMD simulation, the involvement of TM2, TM3, TM4, TM5, TM6, and TM7 in global helical movements is synchronized with the aforementioned localized events (Figure 2.4). This coupling of local side-chain fluctuations and global helical motion is direct evidence of the ligand-induced mechanism involved in the transformation of the receptor global structure.

Large helical motions are characterized as correlated Cα-Cα distance time series. We identify multiple sets of distance time series that are not only correlated at a coefficient of 0.95 (using Pearson's inner-product correlation calculation), but also show coupled changes with the characterized molecular switches at ~3.1, 4, and 5 μs (Figure 2.4). In the extracellular (EXC) region, we detect a set of correlated distance time series involving TM3 interacting with TM2, TM6, and TM7 (Figure 2.4A). For this set of distance time series, an increase in distance between TM3 and its neighboring helices (TM2, TM6, and TM7) is coupled to the formation and loss of hydrogen bond between Y112$^{3.32}$ and Y255$^{6.51}$ (Figures 2.2D and 2.4).

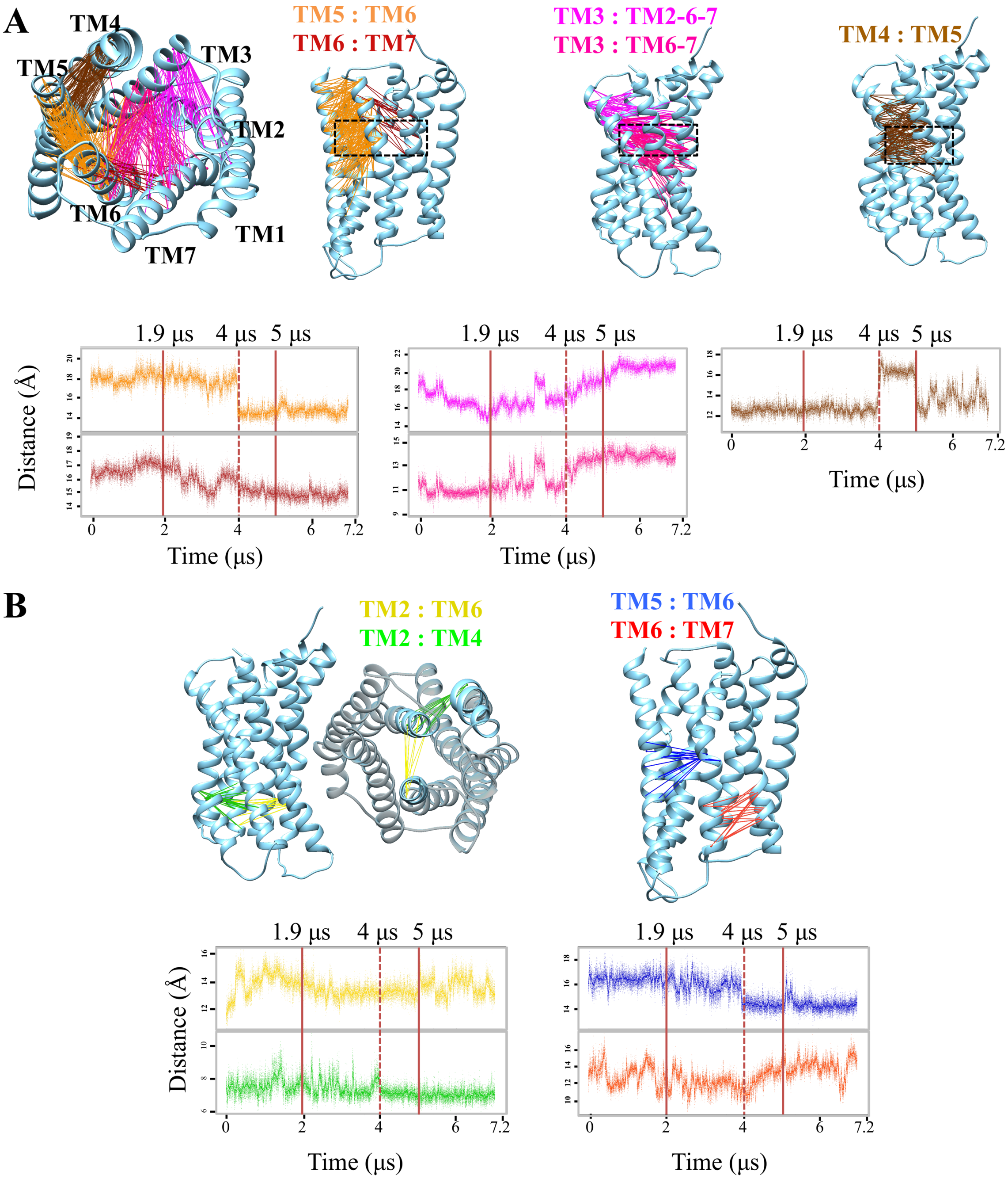


Figure 2.4 CCL21 induces global helical motions in CCR7 cMD simulation. (A) Global motions in the EXC region (upper half) of CCR7. Sets of C_α-C_α distance time series correlated at a coefficient of 0.95 are illustrated in different colors on the molecular structure. Each set is represented by a group of edges and each edge connects two C_α atoms whose distance time series is within a correlated set. A top view and side views of the receptor illustrate the different sets of correlated distance time series and display the TM helices containing the residues involved in the illustrated sets. A representative distance time series from each of the sets is plotted with color code matching that in the molecular graphics. Time series are divided to the same states defined in Figure 2.2. Regions of characterized molecular switches (tri-tyrosine and tri-phenylalanine switches) are highlighted with dashed box. (B) Global motions in the INC region (lower half) of CCR7. Global motions are represented similarly to panel A.

We also observe changes within TM4, TM5, and TM6 where distances of TM5 interacting with TM4 and TM6 increase and decrease, respectively (Figure 2.4A). Both changes are coupled with the rearrangement observed in the tri-phenylalanine switch (Figure 2.2E). TM4 and TM5 move apart due to the detachment between L165$^{4.56}$ and G207$^{5.45}$, changing in distance distribution and variance at 4 μs (Figure 2.2E). In contrast, TM5 and TM6 move closer together, and the distance between TM6 and TM7 decreases as a result of the CCL21-induced formation of the hydrogen bond between Q262$^{6.58}$ and N281$^{7.32}$ (Figures. 2.2 and 2.4A).

Concurrently, we observe larger helical motions within the INC region of TM2, TM4, TM5, TM6, and TM7. We show an interchange of inter-helical behavior between TM2, TM4, and TM6, where, at 4 μs, the distance between TM2 and TM4 is stabilized, while the distance between TM2 and TM6 undergoes periodic fluctuations (Figure 2.4B). More importantly, critical helices, TM5, TM6, and TM7, are involved in helical movements in the INC region of the receptor, where TM6 experiences a shift in its equilibrium position towards TM5 and away from TM7 (Figure 2.4B). As a result of the characterized rearrangements, the INC end of TM6 shows an increase in its distance with other neighboring helices (TM2 and TM7). We observe periodic fluctuations between TM6 and TM2 and an increase in the distance between TM6 and TM7 (Figure 2.4B).

TM4 is involved in movements with different TM helices in the EXC and INC region of the receptor. As a result, we highlight the complementary role of TM4 in propagating the helical movements from the EXC and INC regions of CCR7. The TM4 topology in GPCRs allows it to interact with TM5 and TM2 in the EXC and INC regions respectively. In our CCL21-bound simulation, the separation of TM4 and TM5 in the EXC region (Figure 2.4A) allows for tighter interactions between TM2 and TM4 in the INC region, where the interaction is stabilized (Figure 2.4B). This stabilization complements the increased fluctuations between TM2 and TM6 associated with TM6 conformational changes.

Overall, a series of CCL21-induced molecular switches (Figures. 2.2B, 2.2C, and 2.2D) prompt subtle rearrangements in the movement of TM6 towards TM5 and away from TM2 and TM7, when compared to the larger helical motions in the EXC half of the receptor (Figure 2.4). Similarly, in the aMD simulation, hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$ induces large helical motions in the TMDs as characterized in the CCL21-bound cMD simulation (Figure A.9).

On the other hand, in the CCL19-bound simulation, the lack of hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$ result in minor global helical motions in CCR7 in both the cMD and aMD simulations. CCL19-induced molecular rearrangements show inconsequential movements of the receptor's TM5, TM6, and TM7 EXC regions in cMD simulation, where these movements are synchronized with the localized events in state II (Figure A.10). As CCL19 transitions from states I to II, TM5 exhibits an outward movement away from TM6 and TM7.

2.4 Discussion

Molecular dynamic simulations have made substantial advances in the past years in describing the structural mechanism of GPCR activation (15–19). However, the majority of these studies focus on the unbiased activation of GPCRs; and the molecular basis of ligand-biased signaling remains to be elucidated. In this study, using Anton microsecond dynamics and accelerated MD, we aim to delineate the structural elements that are responsible for mediating the ligand-specific function of chemokine receptors. This entails the characterization of ligand-specific structural events in a biased chemokine receptor system: CCR7 and its endogenous ligands CCL19 and CCL21. During the MD simulations of both ligand-bound systems, we depict functionally relevant regions (molecular switches) in CCR7 that adopt ligand-specific arrangements, and, using the cMD simulations, we detail ligand-induced allosteric pathways capable of mediating the ligand-specific conformational changes within these molecular switches in CCR7.

Both conventional and accelerated MD simulations capture conformational changes within the molecular switches in the TMD of CCR7 involving: the tri-tyrosine switch (Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$), the tri-phenylalanine switch (F116$^{3.36}$, F208$^{5.47}$, and F248$^{6.44}$), and the polar bridge (Q252$^{6.48}$ and R294$^{7.45}$). Additionally, cMD simulations capture a clear connection between ligand-binding and conformational changes within the molecular switches in CCR7 through a series of allosteric events, following a relay of interactions. In CCL21-bound CCR7, changes in the tri-tyrosine and tri-phenylalanine switches are induced through a hydrogen bond between Q262$^{6.58}$ and N281$^{7.32}$ and an adjustment of the ϕ-torsion angle of P254$^{6.50}$. On the other hand, CCL19 is capable of only disrupting the tri-tyrosine switch, and instead it acts through hydrogen bond formation between its N-terminus and Y41$^{1.39}$ and the reorientation of the χ_1-torsion angle in Y288$^{7.39}$. Allosteric events induced by the ligands form allosteric paths between the ligand-NTD and the molecular switches with varying degrees of coupling. Indeed, the tri-tyrosine switch appears to be strongly coupled to the CCL21-induced allosteric events and result in the hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$, while it is loosely coupled to the CCL19-induced allosteric events and do not result in the hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$.

Within the CCL21 allosteric path, the adjustment of the P254$^{6.50}$ ϕ-torsion angle and the loss of the hydrogen bond between Y112$^{3.32}$ and Q252$^{6.48}$ belong to the same group of strongly coupled events, and show synchronized conformational changes at ~3 and 4 μs (Figure 2.2). Conformational changes within this group are a result of an equilibrium shift at 4 μs induced by hydrogen bond formation between Q262$^{6.58}$ and N281$^{7.32}$. As Y112$^{3.32}$ is involved in a relay of hydrogen bonds between Q252$^{6.48}$ and Y255$^{6.51}$, the hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$ occurs stochastically after a 1 μs delay. This hydrogen bond is favored through the formation of π-stacking interactions within the tri-tyrosine switch and hydrogen bond loss between Y112$^{3.32}$ and Q252$^{6.48}$. In contrast, the same hydrogen bond (Y112$^{3.32}$ and Y255$^{6.51}$) does not occur in the CCL19-bound simulations, which is due to the loose coupling of the crucial event of hydrogen bond loss between Y112$^{3.32}$ and Q252$^{6.48}$ to its CCL19-induced allosteric events. Within the CCL19 allosteric pathway, the detected allosteric events (hydrogen bond formation between CCL19's N-terminus and Y41$^{1.39}$ and the reorientation of the χ_1-torsion angle in Y288$^{7.39}$) belong to the same group as highly coupled events occurring at 5.1 μs and 6.9 μs. These events are loosely coupled to the hydrogen bond loss between Y112$^{3.32}$ and Q252$^{6.48}$, whose distance only shows a slight shift in equilibrium upon side-chain rotation in Y288$^{7.39}$.

Another characteristic molecular switch in CCR7 is the tri-phenylalanine switch. Rearrangements in F116$^{3.36}$, F208$^{5.47}$, and F248$^{6.44}$ occur in the CCL21-bound and not CCL19-bound simulations.

The third featured molecular switch, the polar bridge, involves a polar interaction between Q252$^{6.48}$ and R294$^{7.45}$ and belongs to a cluster of conserved polar residues that reside within the TMD of the receptor (69). This polar-residue network accommodates a nearly continuous water-filled passage connecting the EXC and INC side of the receptor, where key rearrangements affect the flow of the water and are critical for receptor activation (69–73). Our simulations (both cMD and aMD) show a highly variable polar bridge involving Q252$^{6.48}$ and R294$^{7.45}$ (Figure 2.1E). These residues connect TM6 and TM7 and contribute to INC rearrangements of the TM helices and to the connectivity of the hydrogen-bonding network regulating the presence of water molecules within the TMD of the receptor. Our cMD simulations exhibit different equilibrium distances between Q252$^{6.48}$ and R294$^{7.45}$ side-chains in the CCL19- and CCL21-bound CCR7 (Figure A.11). The CCL21-bound simulation shows an increase in distance of 3 Å between the initial and final state of the receptor through a highly variable intermediate state, whereas the CCL19-bound simulation lacks any conformational changes within the polar bridge and shows a stable distance of 4 Å (Figure A.11). These conformational changes indicate a differing degree of coupling between the polar bridge and the characterized molecular switches and allosteric events when comparing the ligand-bound cMD simulations.

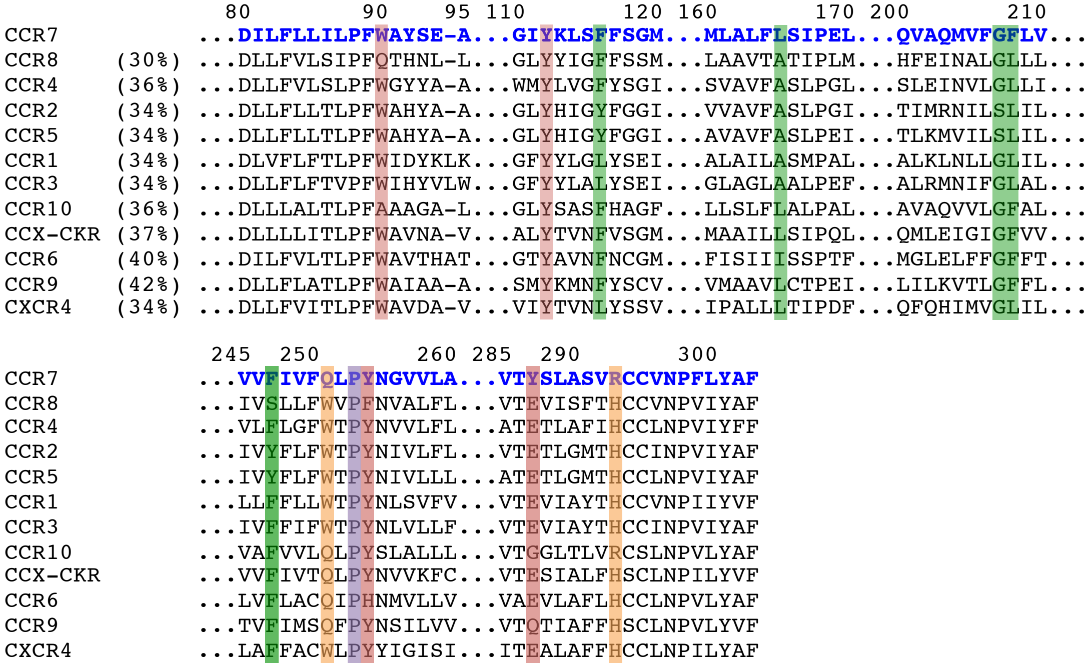
Overall, different parts of the receptor appear to operate through loosely coupled clusters of interlocked allosteric events. In other words, each event (or group of events) is capable of shifting the equilibrium of the various functional regions of the receptor to induce various conformational changes. It is important to note that the degree of coupling between the different events changes in a ligand-specific manner; and it is this concerted coupling that allows the receptor to induce a variety of ligand-specific conformational states that may contribute to its biased signaling. Our simulations may not be sufficiently long to account for equilibrium sampling or to capture all the different allosteric pathways within the receptor. However, we highlight the dependence of each allosteric pathway to its ligand-NTD's interactions inside the receptor-binding pocket, which provides a structural dependence to the differently coupled conformational changes within the switches. It may be that further sampling might result in conformational changes within the tri-tyrosine and tri-phenylalanine switches in the CCL19-bound simulation similar to those in the CCL21-bound simulation. However, these conformational changes may still show different degrees of coupling due to the ligand-NTD's conformation and interactions within the binding pocket of the receptor.

Molecular switches characterized in CCR7 belong to critical TM helices involved in the relative helical conformations responsible for selective-binding of intracellular effectors, namely TM5, TM6, and TM7. Using Cα-Cα distance time series cross-correlation analysis, we detect correlated global helical motions in the ligand and effector binding sites that are synchronized with the aforementioned series of molecular switch rearrangements at the side-chain level (Figure 2.4). In both conventional and accelerated MD simulations, one notable allosteric event is the hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$ that is highly coupled to large helical motions in the EXC region of the TMD (Figures 2.2 and 2.4, and Figure A.9). This interaction occurs in the CCL21 simulations only, and explains the lack of correlated large helical motions in CCL19-bound simulations. Both local side-chain rearrangements and EXC global helical motions result in minor conformational changes in the INC region. These changes involve a slight adjustment of the position of TM6 INC region relative to the rest of its neighboring helices, where TM6 is adjusted for preferential interaction with TM5, and separation from TM2 and TM7. These changes are loosely coupled to the events in the EXC region and do not show any outward movement of TM6 that is indicative of receptor activation (14, 74).

The identified molecular switches are consistent with experimentally reported key conformational changes during the activation of several receptors. Residues at positions 5.50, 3.40, and 6.44 form the "P-I-F" motif and experience discrete conformational states, indicative of active and inactive conformations (17, 75–77). However, in the M2 receptor, activation involves the breaking of the interaction between F195$^{5.47}$ and V199$^{5.51}$ with an unchanged P-I-F motif (15, 77). These studies illustrate that residues involved in the various side-chain rearrangements are not consistent across receptors. However, these alterations occur within the same vicinity. Similarly, in CCR7, even though the P-I-F motif remains unchanged, phenylalanine residues, F248$^{6.44}$ and F208$^{5.47}$, are part of CCR7's tri-phenylalanine switch, where F116$^{3.36}$ flips its side-chain away from F248$^{6.44}$ and F208$^{5.47}$ (Figure 2.2E). Another key molecular change involves

W6.48 (part of WxPF/Y motif) in several GPCRs (15, 78). This residue is mutated to Q252$^{6.48}$ in CCR7 and was shown to participate in key hydrogen bonding and polar interactions with Y112$^{3.32}$ and R294$^{7.45}$ in TM3 and TM7, respectively. Additionally, P254$^{6.50}$ and Y255$^{6.51}$, which are also part of the WxPF/Y motif, participate in signal propagation in CCR7 (68). Our simulations illustrate the presence of several allosteric events that are in line with previously determined molecular switches in the literature.

Additionally, sequence alignment of this family of receptors (Figure 2.5) shows high conservation of the residues involved in the observed molecular switches, which emphasizes their importance in the ligand mediated signal propagation in the receptor. In the tri-tyrosine switch, both Y255$^{6.51}$ and Y112$^{3.32}$ (involved in large helical motions) are conserved in the CC chemokine family. However, the third tyrosine, Y288$^{7.39}$, is mutated to E7.39 in all other receptors. The presence of a negatively charged residue at that



```
                80         90    95 110       120 160       170 200       210
CCR7          ...DILFLLILPFWAYSE-A...GIYKLSFFSGM...MLALFLSIPEL...QVAQMVFGFLV...
CCR8    (30%) ...DLLFVLSIPFQTHNL-L...GLYYIGFFSSM...LAAVTATIPLM...HFEINALGLLL...
CCR4    (36%) ...DLLFVLSLPFWGYYA-A...WMYLVGFYSGI...SVAVFASLPGL...SLEINVLGLLI...
CCR2    (34%) ...DLLFLLTLPFWAHYA-A...GLYHIGYFGGI...VVAVFASLPGI...TIMRNILSLIL...
CCR5    (34%) ...DLLFLLTLPFWAHYA-A...GLYHIGYFGGI...AVAVFASLPEI...TLKMVILSLIL...
CCR1    (34%) ...DLVFLFTLPFWIDYKLK...GFYYLGLYSEI...ALAILASMPAL...ALKLNLLGLIL...
CCR3    (34%) ...DLLFLFTVPFWIHYVLW...GFYYLALYSEI...GLAGLAALPEF...ALRMNIFGLAL...
CCR10   (36%) ...DLLLALTLPFAAAGA-L...GLYSASFHAGF...LLSLFLALPAL...AVAQVVLGFAL...
CCX-CKR (37%) ...DLLLLITLPFWAVNA-V...ALYTVNFVSGM...MAAILLSIPQL...QMLEIGIGFVV...
CCR6    (40%) ...DILFVLTLPFWAVTHAT...GTYAVNFNCGM...FISIIISSPTF...MGLELFFGFFT...
CCR9    (42%) ...DLLFLATLPFWAIAA-A...SMYKMNFYSCV...VMAAVLCTPEI...LILKVTLGFFL...
CXCR4   (34%) ...DLLFVITLPFWAVDA-V...VIYTVNLYSSV...IPALLLTIPDF...QFQHIMVGLIL...

               245  250       260 285  290       300
CCR7          ...VVFIVFQLPYNGVVLA...VTYSLASVRCCVNPFLYAF
CCR8          ...IVSLLFWVPFNVALFL...VTEVISFTHCCVNPVIYAF
CCR4          ...VLFLGFWTPYNVVLFL...ATETLAFIHCCLNPVIYFF
CCR2          ...IVYFLFWTPYNIVLFL...VTETLGMTHCCINPVIYAF
CCR5          ...IVYFLFWTPYNIVLLL...ATETLGMTHCCLNPVIYAF
CCR1          ...LLFFLLWTPYNLSVFV...VTEVIAYTHCCVNPIIYVF
CCR3          ...IVFFIFWTPYNLVLLF...VTEVIAYTHCCINPVIYAF
CCR10         ...VAFVVLQLPYSLALLL...VTGGLTLVRCSLNPVLYAF
CCX-CKR       ...VVFIVTQLPYNVVKFC...VTESIALFHSCLNPILYVF
CCR6          ...LVFLACQIPHNMVLLV...VAEVLAFLHCCLNPVLYAF
CCR9          ...TVFIMSQFPYNSILVV...VTQTIAFFHSCLNPVLYVF
CXCR4         ...LAFFACWLPYYIGISI...ITEALAFFHCCLNPILYAF
```

Figure 2.5 Multiple sequence alignment of residues in CC chemokine receptor family and CXCR4 displays conserved motifs within the characterized switches. Percent identities of each chemokine receptor to CCR7 are reported between parentheses. The transmembrane helical domains are colored in blue in CCR7's sequence and CCR7's residue numbers are displayed above the alignment. Residues involved in the characterized switches are highlighted in different colors: red (W90$^{2.60}$ and tri-tyrosine switch: Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$), orange (polar bridge: Q252$^{6.48}$ and R294$^{7.45}$), purple (P254$^{6.50}$), and green (L165$^{4.56}$, G207$^{5.45}$ and tri-phenylalanine switch: F116$^{3.36}$, F208$^{5.47}$, and F248$^{6.44}$). The sequences are truncated for clarity and the full sequence alignment is displayed in Figure A.12.

position would disrupt the differential binding poses between CCL19 and CCL21. Therefore, Y288$^{7.39}$ is a critical residue unique to CCR7 in inducing its ligands' differential binding. Markedly, this residue is involved in CCL19's allosteric events and not CCL21's. Polar residues at positions Q252$^{6.48}$ and R294$^{7.45}$ in CCR7 are mutated to tryptophan and histidine, respectively, in other CC chemokine receptors. Despite these mutations, important hydrogen bonding capability (tryptophan indole group) and the presence of a positive charge (histidine), that are critical to the characterized allosteric events, are conserved. Lastly, the tri-phenylalanine switch is conserved throughout all CC chemokine receptors with few conserved mutations of leucine to alanine or phenylalanine to leucine (with the exception of CCR2 and CCR5). Additional across species alignment (mouse, human, and bovine) illustrates the conservation of all molecular switch residues in CCR7 and ligand-NTD domains of CCL19 and CCL21 with the exception of less conserved mutations from G6 to A6 in CCL21 and A2 to T2 in CCL19 (data not shown).

The absence of the outward motion of the INC region of TM6 indicates that the receptor remains inactive during our simulations despite the manifestation of conformational changes within the molecular switches and large helical motions in the EXC region of the receptor. We postulate that the characterized networks of allosteric events fine-tune CCR7's molecular switches to drive the receptor to different bound states that show ligand-specific behavior. Similar networks of coupled allosteric events have been characterized previously as a sequential model of activation in rhodopsin, where initial small conformational changes induced by the ligands are converted to larger changes in the EXC regions of the receptor and induce the canonical TM6 outward motion (74, 79, 80). However, despite the presence of a sequential model of conformational changes in CCR7, our simulations show loose allosteric coupling between the EXC and the INC regions. This loose coupling between the agonist-binding site and the intracellular interface has been shown to be responsible for the complex signaling behavior observed for β_2-adrenergic receptor (18). These differences in activation mechanism between rhodopsin and β_2-adrenergic receptors are a result of their inherent biological function, where rhodopsin ensures the rapid and efficient detection of a photon (80), while the β_2-adrenergic receptor evolved for a more complex signaling function (18). In that aspect, unlike CCR7, β_2-adrenergic receptor evolved to function through unbiased small molecules hormones (81), whereas CCR7 functions through biased peptidic chemokine ligands. This finding might justify the existence of a hybrid model in CCR7, consisting of a “rhodopsin-like” sequential network of allosteric events and the “β_2-adrenergic-like” loose coupling between the EXC and INC regions of the receptor. We speculate that the ability of the biased chemokine ligands to regulate the molecular switches ensures efficient transitioning of the receptor to ligand-associated states to carry a ligand-specific function, while the loose coupling between the ligand-binding and INC sites maintains the receptor’s ability to carry out its complex signaling. This presents a hypothesis on how biased ligands may modulate the conformation of the receptor, and further experimental validation is required. Ideally, high-resolution structures of the ligand-bound receptors would depict ligand-stabilized conformations of the receptor’s molecular switches.

Knowledge of the structural and dynamic features of CCR7 provides a framework to aid in the rational design of therapeutics to modulate cell migration or receptor silencing. Here, we identify ligand-dependent conformations in CCR7 that provide the structural detail necessary to rationally design functionally selective drugs. Targeting CCR7 may provide the needed regulation of acute and chronic inflammatory responses involved in many autoimmune diseases and assist in the development of cancer therapies, as CCR7 has been shown to be involved in solid tumors metastasis to the lymph nodes (8, 82–85).

2.5 Acknowledgements

Anton computer time was provided by the National Center for Multiscale Modeling of Biological Systems (MMBioS) through Grant P41GM103712-S1 from the National Institutes of Health and the Pittsburgh Supercomputing Center (PSC). The Anton machine at PSC was generously made available by D.E. Shaw Research.

2.6 References

1. Liu, J.J., R. Horst, V. Katritch, R.C. Stevens, and K. Wüthrich. 2012. Biased Signaling Pathways in β2-Adrenergic Receptor Characterized by 19F-NMR. Science. 335: 1106–1110.

2. Reiter, E., S. Ahn, A.K. Shukla, and R.J. Lefkowitz. 2012. Molecular Mechanism of β-Arrestin-Biased Agonism at Seven-Transmembrane Receptors. Annu. Rev. Pharmacol. Toxicol. 52: 179–197.

3. Rahmeh, R., M. Damian, M. Cottet, H. Orcel, C. Mendre, T. Durroux, K.S. Sharma, G. Durand, B. Pucci, E. Trinquet, J.M. Zwier, X. Deupi, P. Bron, J.-L. Baneres, B. Mouillac, and S. Granier. 2012. Structural insights into biased G protein-coupled receptor signaling revealed by fluorescence spectroscopy. Proc. Natl. Acad. Sci. 109: 6733–6738.

4. Zidar, D.A., J.D. Violin, E.J. Whalen, and R.J. Lefkowitz. 2009. Selective engagement of G protein coupled receptor kinases (GRKs) encodes distinct functions of biased ligands. Proc. Natl. Acad. Sci. 106: 9649–9654.

5. Boguth, C.A., P. Singh, C. Huang, and J.J.G. Tesmer. 2010. Molecular basis for activation of G protein-coupled receptor kinases. EMBO J. 29: 3249–3259.

6. Katritch, V., V. Cherezov, and R.C. Stevens. 2013. Structure-Function of the G-protein-Coupled Receptor Superfamily. Annu. Rev. Pharmacol. Toxicol. 53: 531–556.

7. Shukla, A.K., G.H. Westfield, K. Xiao, R.I. Reis, L.-Y. Huang, P. Tripathi-Shukla, J. Qian, S. Li, A. Blanc, A.N. Oleskie, A.M. Dosey, M. Su, C.-R. Liang, L.-L. Gu, J.-M. Shan, X. Chen, R. Hanna, M. Choi, X.J. Yao, B.U. Klink, A.W. Kahsai, S.S. Sidhu, S. Koide, P.A. Penczek, A.A. Kossiakoff, V.L. Woods Jr, B.K. Kobilka, G. Skiniotis, and R.J. Lefkowitz. 2014. Visualization of arrestin recruitment by a G-protein-coupled receptor. Nature. 512: 218–222.

8. Förster, R., A.C. Davalos-Misslitz, and A. Rot. 2008. CCR7 and its ligands: balancing immunity and tolerance. Nat. Rev. Immunol. 8: 362–371.

9. Haessler, U., M. Pisano, M. Wu, and M.A. Swartz. 2011. Dendritic cell chemotaxis in 3D under defined chemokine gradients reveals differential response to ligands CCL21 and CCL19. Proc. Natl. Acad. Sci. 108: 5614–5619.

10. Schumann, K., T. Lämmermann, M. Bruckner, D.F. Legler, J. Polleux, J.P. Spatz, G. Schuler, R. Förster, M.B. Lutz, L. Sorokin, and M. Sixt. 2010. Immobilized Chemokine Fields and Soluble Chemokine Gradients Cooperatively Shape Migration Patterns of Dendritic Cells. Immunity. 32: 703–713.

11. Kohout, T.A., S.L. Nicholas, S.J. Perry, G. Reinhart, S. Junger, and R.S. Struthers. 2004. Differential Desensitization, Receptor Phosphorylation, -Arrestin Recruitment, and ERK1/2 Activation by the Two Endogenous Ligands for the CC Chemokine Receptor 7. J. Biol. Chem. 279: 23214–23222.

12. Kahsai, A.W., K. Xiao, S. Rajagopal, S. Ahn, A.K. Shukla, J. Sun, T.G. Oas, and R.J. Lefkowitz. 2011. Multiple ligand-specific conformations of the β2-adrenergic receptor. Nat. Chem. Biol. 7: 692–700.

13. Kang, Y., X.E. Zhou, X. Gao, Y. He, W. Liu, A. Ishchenko, A. Barty, T.A. White, O. Yefanov, G.W. Han, Q. Xu, P.W. de Waal, J. Ke, M.H.E. Tan, C. Zhang, A. Moeller, G.M. West, B.D. Pascal, N.

Van Eps, L.N. Caro, S.A. Vishnivetskiy, R.J. Lee, K.M. Suino-Powell, X. Gu, K. Pal, J. Ma, X. Zhi, S. Boutet, G.J. Williams, M. Messerschmidt, C. Gati, N.A. Zatsepin, D. Wang, D. James, S. Basu, S. Roy-Chowdhury, C.E. Conrad, J. Coe, H. Liu, S. Lisova, C. Kupitz, I. Grotjohann, R. Fromme, Y. Jiang, M. Tan, H. Yang, J. Li, M. Wang, Z. Zheng, D. Li, N. Howe, Y. Zhao, J. Standfuss, K. Diederichs, Y. Dong, C.S. Potter, B. Carragher, M. Caffrey, H. Jiang, H.N. Chapman, J.C.H. Spence, P. Fromme, U. Weierstall, O.P. Ernst, V. Katritch, V.V. Gurevich, P.R. Griffin, W.L. Hubbell, R.C. Stevens, V. Cherezov, K. Melcher, and H.E. Xu. 2015. Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. Nature. 523: 561–567.

14. Rasmussen, S.G.F., B.T. DeVree, Y. Zou, A.C. Kruse, K.Y. Chung, T.S. Kobilka, F.S. Thian, P.S. Chae, E. Pardon, D. Calinski, J.M. Mathiesen, S.T.A. Shah, J.A. Lyons, M. Caffrey, S.H. Gellman, J. Steyaert, G. Skiniotis, W.I. Weis, R.K. Sunahara, and B.K. Kobilka. 2011. Crystal structure of the β2 adrenergic receptor–Gs protein complex. Nature. 477: 549–555.

15. Miao, Y., S.E. Nichols, P.M. Gasper, V.T. Metzger, and J.A. McCammon. 2013. Activation and dynamic network of the M2 muscarinic receptor. Proc. Natl. Acad. Sci. 110: 10982–10987.

16. Dror, R.O., D.H. Arlow, P. Maragakis, T.J. Mildorf, A.C. Pan, H. Xu, D.W. Borhani, and D.E. Shaw. 2011. Activation mechanism of the β2-adrenergic receptor. Proc. Natl. Acad. Sci. 108: 18684–18689.

17. Nygaard, R., Y. Zou, R.O. Dror, T.J. Mildorf, D.H. Arlow, A. Manglik, A.C. Pan, C.W. Liu, J.J. Fung, M.P. Bokoch, F.S. Thian, T.S. Kobilka, D.E. Shaw, L. Mueller, R.S. Prosser, and B.K. Kobilka. 2013. The Dynamic Process of β2-Adrenergic Receptor Activation. Cell. 152: 532–542.

18. Manglik, A., T.H. Kim, M. Masureel, C. Altenbach, Z. Yang, D. Hilger, M.T. Lerch, T.S. Kobilka, F.S. Thian, W.L. Hubbell, R.S. Prosser, and B.K. Kobilka. 2015. Structural Insights into the Dynamic Process of β2-Adrenergic Receptor Signaling. Cell. 161: 1101–1111.

19. Tikhonova, I.G., B. Selvam, A. Ivetac, J. Wereszczynski, and J.A. McCammon. 2013. Simulations of Biased Agonists in the β2 Adrenergic Receptor with Accelerated Molecular Dynamics. Biochemistry (Mosc.). 52: 5593–5603.

20. Shaw, D.E., R.O. Dror, J.K. Salmon, J.P. Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, K.J. Bowers, E. Chow, M.P. Eastwood, D.J. Ierardi, J.L. Klepeis, J.S. Kuskin, R.H. Larson, K. Lindorff-Larsen, P. Maragakis, M.A. Moraes, S. Piana, Y. Shan, and B. Towles. 2009. Millisecond-scale molecular dynamics simulations on Anton. In: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. . pp. 1–11.

21. Hamelberg, D., J. Mongan, and J.A. McCammon. 2004. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. J. Chem. Phys. 120: 11919–11929.

22. Wang, Y., C.B. Harrison, K. Schulten, and J.A. McCammon. 2011. Implementation of accelerated molecular dynamics in NAMD. Comput. Sci. Discov. 4: 015002.

23. Ballesteros, J.A., and H. Weinstein. 1995. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In: Sealfon SC, editor. Methods in Neurosciences. Academic Press. pp. 366–428.

24. Qin, L., I. Kufareva, L.G. Holden, C. Wang, Y. Zheng, C. Zhao, G. Fenalti, H. Wu, G.W. Han, V. Cherezov, R. Abagyan, R.C. Stevens, and T.M. Handel. 2015. Crystal structure of the chemokine receptor CXCR4 in complex with a viral chemokine. Science. 347: 1117–1122.

25. Apweiler, R., A. Bairoch, and C.H. Wu. 2004. Protein sequence databases. Curr. Opin. Chem. Biol. 8: 76–80.

26. Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23: 2947–2948.

27. Goujon, M., H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez. 2010. A new bioinformatics analysis tools framework at EMBL–EBI. Nucleic Acids Res. 38: W695–W699.

28. Love, M., J.L. Sandberg, J.J. Ziarek, K.P. Gerarden, R.R. Rode, D.R. Jensen, D.R. McCaslin, F.C. Peterson, and C.T. Veldkamp. 2012. Solution Structure of CCL21 and Identification of a Putative CCR7 Binding Site. Biochemistry (Mosc.). 51: 733–735.

29. Šali, A., and T.L. Blundell. 1993. Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol. 234: 779–815.

30. Veldkamp, C.T., E. Kiermaier, S.J. Gabel-Eissens, M.L. Gillitzer, D.R. Lippner, F.A. DiSilvio, C.J. Mueller, P.L. Wantuch, G.R. Chaffee, M.W. Famiglietti, D.M. Zgoba, A.A. Bailey, Y. Bah, S.J. Engebretson, D.R. Graupner, E.R. Lackner, V.D. LaRosa, T. Medeiros, M.L. Olson, A.J. Phillips, H. Pyles, A.M. Richard, S.J. Schoeller, B. Touzeau, L.G. Williams, M. Sixt, and F.C. Peterson. 2015. Solution Structure of CCL19 and Identification of Overlapping CCR7 and PSGL-1 Binding Sites. Biochemistry (Mosc.). 54: 4163–4166.

31. Venkatakrishnan, A.J., X. Deupi, G. Lebon, C.G. Tate, G.F. Schertler, and M.M. Babu. 2013. Molecular signatures of G-protein-coupled receptors. Nature. 494: 185–194.

32. Tan, Q., Y. Zhu, J. Li, Z. Chen, G.W. Han, I. Kufareva, T. Li, L. Ma, G. Fenalti, J. Li, W. Zhang, X. Xie, H. Yang, H. Jiang, V. Cherezov, H. Liu, R.C. Stevens, Q. Zhao, and B. Wu. 2013. Structure of the CCR5 Chemokine Receptor–HIV Entry Inhibitor Maraviroc Complex. Science. 341: 1387–1390.

33. Cherezov, V., D.M. Rosenbaum, M.A. Hanson, S.G.F. Rasmussen, F.S. Thian, T.S. Kobilka, H.-J. Choi, P. Kuhn, W.I. Weis, B.K. Kobilka, and R.C. Stevens. 2007. High-Resolution Crystal Structure of an Engineered Human β2-Adrenergic G Protein–Coupled Receptor. Science. 318: 1258–1265.

34. Park, S.H., B.B. Das, F. Casagrande, Y. Tian, H.J. Nothnagel, M. Chu, H. Kiefer, K. Maier, A.A. De Angelis, F.M. Marassi, and S.J. Opella. 2012. Structure of the chemokine receptor CXCR1 in phospholipid bilayers. Nature. 491: 779–783.

35. Roy, A., A. Kucukural, and Y. Zhang. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5: 725–738.

36. Zhang, Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 9: 40.

37. Wacker, D., G. Fenalti, M.A. Brown, V. Katritch, R. Abagyan, V. Cherezov, and R.C. Stevens. 2010. Conserved Binding Mode of Human β2 Adrenergic Receptor Inverse Agonists and Antagonist Revealed by X-ray Crystallography. J. Am. Chem. Soc. 132: 11443–11445.

38. Rasmussen, S.G.F., H.-J. Choi, J.J. Fung, E. Pardon, P. Casarosa, P.S. Chae, B.T. DeVree, D.M. Rosenbaum, F.S. Thian, T.S. Kobilka, A. Schnapp, I. Konetzki, R.K. Sunahara, S.H. Gellman, A. Pautsch, J. Steyaert, W.I. Weis, and B.K. Kobilka. 2011. Structure of a nanobody-stabilized active state of the β2 adrenoceptor. Nature. 469: 175–180.

39. Schmidt, T., A. Bergner, and T. Schwede. 2014. Modelling three-dimensional protein structures for applications in drug design. Drug Discov. Today. 19: 890–897.

40. Skelton, N.J., C. Quan, D. Reilly, and H. Lowman. 1999. Structure of a CXC chemokine-receptor fragment in complex with interleukin-8. Structure. 7: 157–168.

41. Millard, C.J., J.P. Ludeman, M. Canals, J.L. Bridgford, M.G. Hinds, D.J. Clayton, A. Christopoulos, R.J. Payne, and M.J. Stone. 2014. Structural Basis of Receptor Sulfotyrosine Recognition by a CC Chemokine: The N-Terminal Region of CCR3 Bound to CCL11/Eotaxin-1. Structure. 22: 1571–1581.

42. Burg, J.S., J.R. Ingram, A.J. Venkatakrishnan, K.M. Jude, A. Dukkipati, E.N. Feinberg, A. Angelini, D. Waghray, R.O. Dror, H.L. Ploegh, and K.C. Garcia. 2015. Structural basis for chemokine recognition and activation of a viral G protein–coupled receptor. Science. 347: 1113–1117.

43. Pettersen, E.F., T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. 2004. UCSF Chimera-A visualization system for exploratory research and analysis. J. Comput. Chem. 25: 1605–1612.

44. Phillips, J.C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, and K. Schulten. 2005. Scalable molecular dynamics with NAMD. J. Comput. Chem. 26: 1781–1802.

45. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: Visual molecular dynamics. J. Mol. Graph. 14: 33–38.

46. MacKerell, A.D., D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins [†]. J. Phys. Chem. B. 102: 3586–3616.

47. Mackerell, A.D., M. Feig, and C.L. Brooks. 2004. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J. Comput. Chem. 25: 1400–1415.

48. Best, R.B., X. Zhu, J. Shim, P.E.M. Lopes, J. Mittal, M. Feig, and A.D. MacKerell. 2012. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi1$ and $\chi2$ Dihedral Angles. J. Chem. Theory Comput. 8: 3257–3273.

49. Klauda, J.B., R.M. Venable, J.A. Freites, J.W. O'Connor, D.J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A.D. MacKerell, and R.W. Pastor. 2010. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. J. Phys. Chem. B. 114: 7830–7843.

50. Lomize, M.A., I.D. Pogozheva, H. Joo, H.I. Mosberg, and A.L. Lomize. 2012. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 40: D370–D376.

51. Lippert, R.A., C. Predescu, D.J. Ierardi, K.M. Mackenzie, M.P. Eastwood, R.O. Dror, and D.E. Shaw. 2013. Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. J. Chem. Phys. 139: 164106.

52. R. Core Team. 2014. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

53. Van Rossum, G., and others. 2007. Python Programming Language. USENIX Annu. Tech. Conf. 41.

54. Grant, B.J., A.P.C. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S.D. Caves. 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 22: 2695–2696.

55. Skjærven, L., X.-Q. Yao, G. Scarabelli, and B.J. Grant. 2014. Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics. 15.

56. Wriggers, W., K.A. Stafford, Y. Shan, S. Piana, P. Maragakis, K. Lindorff-Larsen, P.J. Miller, J. Gullingsrud, C.A. Rendleman, M.P. Eastwood, R.O. Dror, and D.E. Shaw. 2009. Automated Event Detection and Activity Monitoring in Long Molecular Dynamics Simulations. J. Chem. Theory Comput. 5: 2595–2605.

57. Mills, J.E.J., and P.M. Dean. 1996. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. J. Comput. Aided Mol. Des. 10: 607–622.

58. Lanzarotti, E., R.R. Biekofsky, D.A. Estrin, M.A. Marti, and A.G. Turjanski. 2011. Aromatic–Aromatic Interactions in Proteins: Beyond the Dimer. J. Chem. Inf. Model. 51: 1623–1633.

59. McGaughey, G.B., M. Gagné, and A.K. Rappé. 1998. π-Stacking Interactions ALIVE AND WELL IN PROTEINS. J. Biol. Chem. 273: 15458–15463.

60. Sievers, F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, and D.G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7: 539.

61. Crooks, G.E., G. Hon, J.-M. Chandonia, and S.E. Brenner. 2004. WebLogo: A Sequence Logo Generator. Genome Res. 14: 1188–1190.

62. Schneider, T.D., and R.M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18: 6097–6100.

63. Wereszczynski, J., and J.A. McCammon. 2012. Nucleotide-dependent mechanism of Get3 as elucidated from free energy calculations. Proc. Natl. Acad. Sci. 109: 7759–7764.

64. Gasper, P.M., B. Fuglestad, E.A. Komives, P.R.L. Markwick, and J.A. McCammon. 2012. Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. Proc. Natl. Acad. Sci. 109: 21216–21222.

65. Pierce, L.C.T., R. Salomon-Ferrer, C. Augusto F. de Oliveira, J.A. McCammon, and R.C. Walker. 2012. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. J. Chem. Theory Comput. 8: 2997–3002.

66. Ott, T.R., A. Pahuja, S.A. Nickolls, D.G. Alleva, and R.S. Struthers. 2004. Identification of CC Chemokine Receptor 7 Residues Important for Receptor Activation. J. Biol. Chem. 279: 42383–42392.

67. Ott, T.R., F.M. Lio, D. Olshefski, X.-J. Liu, R.S. Struthers, and N. Ling. 2004. Determinants of High-Affinity Binding and Receptor Activation in the N-Terminus of CCL-19 (MIP-3β). Biochemistry (Mosc.). 43: 3670–3678.

68. Schwartz, T.W., T.M. Frimurer, B. Holst, M.M. Rosenkilde, and C.E. Elling. 2006. Molecular Mechanism of 7tm Receptor Activation—a Global Toggle Switch Model. Annu. Rev. Pharmacol. Toxicol. 46: 481–519.

69. Katritch, V., G. Fenalti, E.E. Abola, B.L. Roth, V. Cherezov, and R.C. Stevens. 2014. Allosteric sodium in class A GPCR signaling. Trends Biochem. Sci. 39: 233–244.

70. Angel, T.E., S. Gupta, B. Jastrzebska, K. Palczewski, and M.R. Chance. 2009. Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. Proc. Natl. Acad. Sci. 106: 14367–14372.

71. Angel, T.E., M.R. Chance, and K. Palczewski. 2009. Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors. Proc. Natl. Acad. Sci. 106: 8555–8560.

72. Yuan, S., S. Filipek, K. Palczewski, and H. Vogel. 2014. Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. Nat. Commun. 5: 4733.

73. Valentin-Hansen, L., T.M. Frimurer, J. Mokrosinski, N.D. Holliday, and T.W. Schwartz. 2015. Biased Gs versus Gq and β-arrestin signaling in the NK1 receptor determined by interactions in the water hydrogen-bond network. J. Biol. Chem. : jbc.M115.641944.

74. Altenbach, C., A.K. Kusnetzow, O.P. Ernst, K.P. Hofmann, and W.L. Hubbell. 2008. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. Proc. Natl. Acad. Sci. 105: 7439–7444.

75. Huang, W., A. Manglik, A.J. Venkatakrishnan, T. Laeremans, E.N. Feinberg, A.L. Sanborn, H.E. Kato, K.E. Livingston, T.S. Thorsen, R.C. Kling, S. Granier, P. Gmeiner, S.M. Husbands, J.R. Traynor, W.I. Weis, J. Steyaert, R.O. Dror, and B.K. Kobilka. 2015. Structural insights into μ-opioid receptor activation. Nature. advance online publication.

76. Wacker, D., C. Wang, V. Katritch, G.W. Han, X.-P. Huang, E. Vardy, J.D. McCorvy, Y. Jiang, M. Chu, F.Y. Siu, W. Liu, H.E. Xu, V. Cherezov, B.L. Roth, and R.C. Stevens. 2013. Structural Features for Functional Selectivity at Serotonin Receptors. Science. 340: 615–619.

77. Kruse, A.C., J. Hu, A.C. Pan, D.H. Arlow, D.M. Rosenbaum, E. Rosemond, H.F. Green, T. Liu, P.S. Chae, R.O. Dror, D.E. Shaw, W.I. Weis, J. Wess, and B.K. Kobilka. 2012. Structure and dynamics of the M3 muscarinic acetylcholine receptor. Nature. 482: 552–556.

78. Deupi, X., and J. Standfuss. 2011. Structural insights into agonist-induced activation of G-protein-coupled receptors. Curr. Opin. Struct. Biol. 21: 541–551.

79. Park, P.S.-H., D.T. Lodowski, and K. Palczewski. 2008. Activation of G Protein–Coupled Receptors: Beyond Two-State Models and Tertiary Conformational Changes. Annu. Rev. Pharmacol. Toxicol. 48: 107–141.

80. Smith, S.O. 2010. Structure and Activation of the Visual Pigment Rhodopsin. Annu. Rev. Biophys. 39: 309–328.

81. Rajagopal, S., S. Ahn, D.H. Rominger, W. Gowen-MacDonald, C.M. Lam, S.M. DeWire, J.D. Violin, and R.J. Lefkowitz. 2011. Quantifying Ligand Bias at Seven-Transmembrane Receptors. Mol. Pharmacol. 80: 367–377.

82. Legler, D.F., E. Uetz-von Allmen, and M.A. Hauser. 2014. CCR7: Roles in cancer cell dissemination, migration and metastasis formation. Int. J. Biochem. Cell Biol. 54: 78–82.

83. Shields, J.D., I.C. Kourtis, A.A. Tomei, J.M. Roberts, and M.A. Swartz. 2010. Induction of Lymphoidlike Stroma and Immune Escape by Tumors That Express the Chemokine CCL21. Science. 328: 749–752.

84. Stacker, S.A., S.P. Williams, T. Karnezis, R. Shayan, S.B. Fox, and M.G. Achen. 2014. Lymphangiogenesis and lymphatic vessel remodelling in cancer. Nat. Rev. Cancer. 14: 159–172.

85. Tutunea-Fatan, E., M. Majumder, X. Xin, and P.K. Lala. 2015. The role of CCL21/CCR7 chemokine axis in breast cancer-induced lymphangiogenesis. Mol. Cancer. 14: 35.

CHAPTER 3: COMPARISON OF FREE AND LIGAND-BOUND CONFORMATIONAL TRANSITIONS OF CC CHEMOKINE 7 (CCR7), USING MOLECULAR DYNAMICS SIMULATIONS

3.1 Introduction

Chemokine receptors are G protein-coupled receptors (GPCRs) that regulate the activity of leukocyte migration and their positioning within lymphoid organs and peripheral tissues to initiate an antigen-specific T cell immune response, stimulate wound healing, and regulate inflammatory responses (1–3). Regulating cell migration in the innate and adaptive immune response constitutes the central role of these receptors, making them susceptible to disease progression and thus, they are involved in many physiopathological disorders. For instance, chemokine receptors, CXCR4 and CCR5, constitute the main point of entry for human immunodeficiency virus type 1 infectivity (4); and CCR2 is involved in many autoimmune diseases where uncontrolled inflammatory responses exacerbate tissue damage and can lead to diseases such as multiple sclerosis, and rheumatoid arthritis (2). Additionally, cancer cells have the ability to subvert the chemokine system to contribute to the immunological tolerance and metastasis of cancer (5). Hence, chemokine receptors are validated therapeutic targets, against which, two drugs were approved by the US FDA and multiple drugs are ongoing Phase II trials (2).

Despite the modest success in drug discovery, designing small drug inhibitors targeting chemokine receptors remains challenging due to the promiscuity of the ligand-receptor interactions where several ligands share common receptors and vice versa (6). Given the interdependence between ligands and receptors, complete inhibition of the receptor will be detrimental to the patient due to undesirable side effects. Additionally, the size of the endogenous ligands can also be problematic when designing small molecule inhibitors that would compete with the large chemokine ligand at the orthosteric site. To address these challenges, allosteric modulators have been the key driver towards potent and selective inhibitors. Allosteric modulators bind at sites away from the orthosteric site, avoiding competitive binding with the orthosteric ligand. They are also able to fine-tune specific downstream signaling pathways, minimizing side effects resulting from complete inhibition of the receptor (2).

Allosteric modulators provide the opportunity to initiate the design of therapeutic drugs that target chemokine receptor deemed non-viable targets due to their vital role in many physiological processes. In particular, CC chemokine receptor 7 (CCR7) has a key role in orchestrating the adaptive immune response, where its ligands, CCL19 and CCL21, have distinct roles in the homing and functional compartmentalization of T cells and antigen-presenting dendritic cells to the secondary lymph nodes (7–9). Due to its critical involvement in the adaptive immune response, CCR7 has been excluded as a viable drug design target in the fear of altering the adaptive anti-tumor response even though CCR7 has been studied extensively and found to be the primary axis responsible for facilitating cellular migration of both cancer and host's immune cells. CCR7- and CCL21-expressing cancer cells were shown to promote tumor cell homing to the lymph nodes in gastric carcinoma, colorectal carcinoma, and breast cancer (7, 10) and induce lymphangiogenesis and lymph remodeling in breast cancer (5, 11). Furthermore, even though CCL21 is able to promote the migration of leukocytes to the tumor's microenvironment where immune cells induce antitumor immunity based on the tumor's antigen profile, many tumors are able to drive local immunological tolerance by manipulating CCL21-induced immune response (12). CCL21 was shown to mimic a lymphoid-like stroma that can recruit T regulatory cells and promote the differentiation of naïve T cells to active T regulatory cells to suppress effector functions in the developing tolerogenic tumor (12). Given the critical role of CCR7 in cancer cell metastasis and immunological tolerance, allosteric modulators can aim to fine-tune CCR7's biased signaling to suppress the undesired immune tolerance (activity and function of T regulatory cells) and promote an antitumor immune response.

To initiate CCR7's cellular function, CCL19 and CCL21 have been shown to selectively induce distinct signaling pathways in the cell (7). CCR7's diverse role could be modulated using allosteric drugs as to inhibit the immunological tolerance of the tumor environment and minimize any of the tumor's meddling with the adaptive immune response. Selective alteration of the receptor's physiological function is possible by disrupting the binding affinity or efficacy of its endogenous ligands using small molecule allosteric modulators. These modulators can induce slight modifications to the equilibrium state of the receptor that may profoundly alter its function by shifting the equilibrium state towards a more biased state of the receptor associated with a signaling pathway of interest. This requires deep understanding of the mechanics behind ligand binding and activation to fully exploit the potential of allosterically modulating the biased function of the receptor.

Structural insight into chemokine receptor dynamics remained limited due to the lack of available chemokine-bound receptor structures. However, recently determined crystal structures of vMIP-II-bound CXCR4 and CX3CL1-bound US28 (13, 14) have opened the door for dynamical studies of chemokine receptors bound to their endogenous chemokine ligands. Molecular dynamics (MD) simulations allow us to study the structural dynamics of CCR7 and characterize its behavior in its ligand-free (apo) and chemokine-bound forms. Our study demonstrates the importance of ligand-binding in coordinating the structural heterogeneity of the receptor. Despite the relatively high fluctuations in both apo and ligand-bound receptor, only ligand-bound CCR7 induces highly correlated motions within the receptor, while the apo receptor remains uncorrelated. A residue critical to initiating correlated conformational changes in the receptor, Y112$^{3.32}$, was observed to be in a different conformation in the apo receptor when compared to its bound counterparts. Understanding the structural behavior of the receptor in its different forms is important to isolate the structural features and receptor states that are critical to its function. These features could be targeted in the design of allosteric modulators to induce specific functional receptor states associated with a cellular signaling pathway of interest.

3.2 Materials and Methods

Nomenclature. Residues are stated as a one-letter amino acid code and a number corresponding to their order in the sequence. Additional sequence information is presented in the superscripts. Superscripts of receptor residues are numbered according to Ballesteros–Weinstein scheme and convey the helix number and position of each residue relative to the most conserved residue in the helix: N52$^{1.50}$ for TM 1, E80$^{2.50}$ for TM 2, R130$^{3.50}$ for TM 3, W159$^{4.50}$ for TM 4, P211$^{5.50}$ for TM 5, P254$^{6.50}$ for TM 6, and R294$^{7.50}$ for TM 7 (15).

System setup and molecular dynamics simulations. The apo receptor was modeled after the newly determined crystal structure of the chemokine receptor CXCR4 bound to a viral chemokine antagonist vMIP-II (PDB code 4RWS) following the procedure described in Chapter 2 (13). MD simulations of the generated model were performed following the protocol described in Chapter 2 using the Anton supercomputer (16).

Analysis protocols. Analysis of the MD trajectories was performed with in-house scripts using R programming language (17), Python (18), Chimera (19), and the Bio3D library (20, 21). In order to isolate the constituent motions of a system, we analyze the trajectory for correlated motions. We generate a DCC map for each of our CCR7 systems: (i) apo CCR7, (ii) CCL19-bound CCR7, and (iii) CCL21-bound CCR7. Analysis scripts to calculate DCC were developed using R and the Bio3d library (17, 20, 21). The fluctuations for each residue side chain are quantified using the maximum rmsf of all side chain atom's caluclated rmsfs. Rmsfs are calculated using Bio3d (20, 21). Atomic distances were calculated between non-hydrogen side-chain polar atoms of Y112$^{3.32}$, Y255$^{6.51}$, Q252$^{6.48}$. Interacting residue distance time series are determined by calculating the minimum distances of the set of atomic distance time series of two interacting residues using python (18).

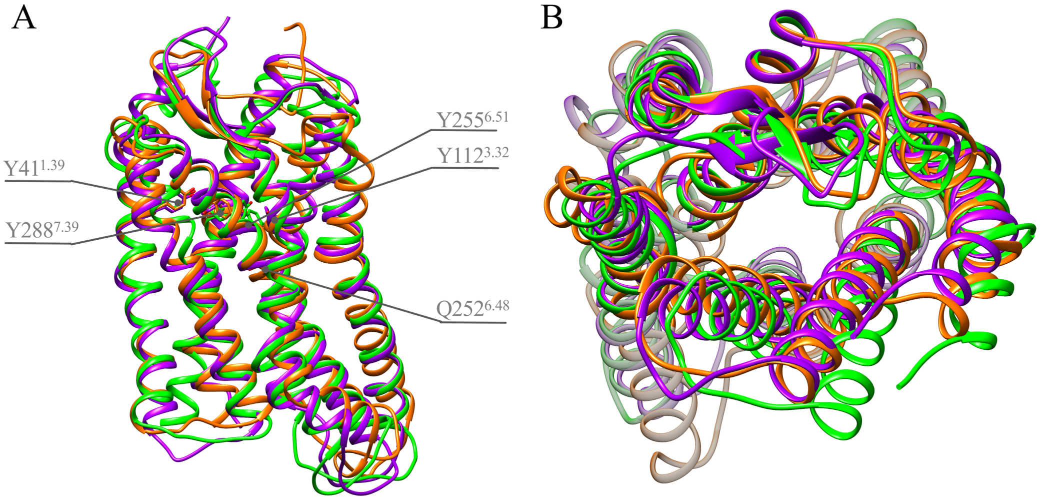


Figure 3.1 Molecular graphics of CCL19-bound (purple), CCL21-bound (green), and apo (orange) forms of CCR7 side view (A) and top view (B).

3.3 Results

Free and ligand-bound molecular dynamics simulations produce differences in their dynamic cross-correlation maps. On the basis of the recently solved structure of the chemokine-bound CXCR4 (13) (PDB code 4RWS), we modeled the complex structures of CCR7 in its ligand-free (apo, this work) and bound (CCL21 and CCL19) forms (Figure 3.1, see Chapter 2). Using the Anton supercomputer, we then performed seven μs MD simulations of the three structures of CCR7. Our study focuses on the transmembrane (TM) domain of the receptor to delineate the conformational states of the receptor and the role of the ligand in inducing the receptor's conformational changes. Previously, our group characterized molecular switches in CCR7, mediated by CCL19 and CCL21, and isolated different equilibrium states of the receptor in the ligand-bound simulations: a transitional state where the receptor experiences coordinated changes in its molecular switches and transmembrane helices, and an equilibrated state, where the receptor is stable and shows no changes in its molecular switches and TM helices (see Chapter 2). Hydrogen bond formation within the tri-tyrosine switch (Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$), in particular between Y112$^{3.32}$ and Y255$^{6.51}$, only occurs in the CCL21-bound CCR7 simulation and leads to global helical movements in the receptor's transmembrane helices that contribute to the transitioning of the receptor to distinct states.

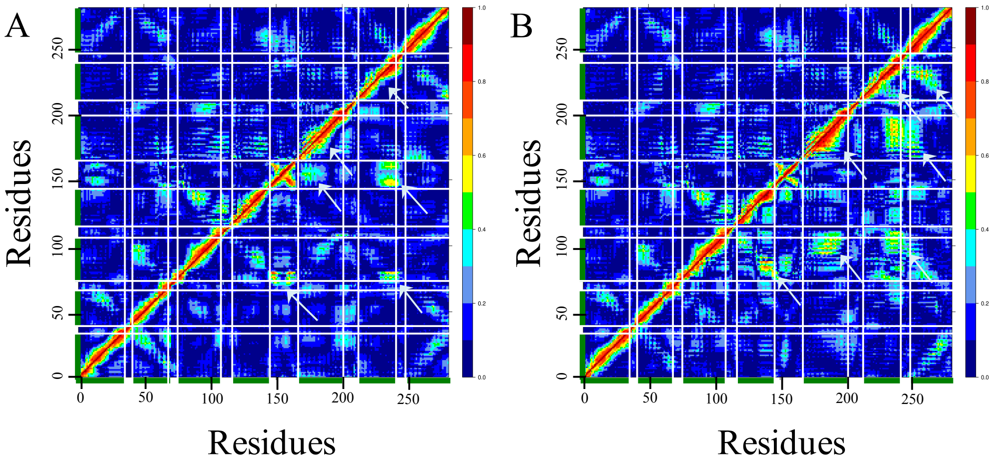


Figure 3.2 Dynamic cross-correlation map for CCL19-bound (A, lower triangle), CCL21-bound (B, lower triangle), and apo forms of CCR7 (upper triangles) calculated from the transitional states of the ligand-bound simulations (5.1-7 μs for CCL19, 1.9-5 μs for CCL21) and the full simulation of the apo receptor. The green bars denote the seven TM domains of CCR7. Areas of high correlation (> 0.5) are indicated with white arrows.

To evaluate the receptor's behavior and conformational changes in its free and bound forms, we compare the transitional states of both ligand-bound and apo receptor simulations through dynamic cross-correlation (DCC). DCC assesses the correlation of the receptor's backbone (C_α atoms) dynamics to detect large domain motions of the TM helices. DCC calculations are confined to the transitional state in order to assess the correlation of the receptor motions mediated by the ligand in the bound compared to the apo receptor simulations.

All three receptor forms exhibit different dynamical behaviors of their TM domains. The correlation maps of the free, CCL19-bound, and CCL21-bound simulations display different correlated protein motions (Figure 3.2). CCL21-bound simulation (Figure 3.2B, lower triangle) shows significantly more correlated motions when compared to the apo (Figure 3.2, upper triangles) and CCL19-bound simulations (Figure 3.2A, lower triangle), while the CCL19-bound simulation shows higher correlation than the apo receptor.

CCL19-bound CCR7 shows correlations between extracellular loop 2 (ECL2) and the extracellular (EXC) regions of TM3, TM5, and TM6. Additionally, correlated helical movements are manifested within helical domains TM6 and the EXC region of TM5, identified from the protrusions of high correlation coefficient on the DCC map diagonal in the CCL19-bound simulation side (Figure 3.2A, lower triangle) that are absent in the apo simulation (Figure 3.2A, upper triangle).

The cross-correlation map in Figure 3.2B exhibits a large number of correlated regions in the CCL21 DCC map (Figure 3.2B, lower triangle) compared to the apo map (Figure 3.2B, upper triangle). In CCL21-bound simulation, the EXC region of TM3 is correlated with the EXC regions of TM4 and TM6; and the INC region of TM3 is correlated with the INC region of TM5 and TM6. Additional correlations are present between the EXC regions of TM5 and TM6, between TM6 and TM7, and within the C_α making up TM5 and TM6.

It is also worth noting that the absence of correlation within TM1 and TM2 agrees with previously published work where both helices are shown to not be involved in the receptor's conformational changes (22).

CCR7 in its apo and bound forms displays differences in residue fluctuations. Protein function relies heavily on its dynamics and thermal fluctuations to mediate its concerted helical motions detected through DCC. Fast fluctuations at the picosecond and nanosecond timescales are essential to drive changes within CCR7's side chain interactions and drive the receptor to distinct states transitioning at the microseconds timescales and slower (23). To explain the differences in correlated helical motions between the receptor forms, we measure its fluctuations using the root-mean square fluctuation (rmsf) method to quantify the thermal fluctuations of each of residue, including backbone and side chain. According to rmsf calculations of our three simulations (apo, CCL19-bound, and CCL21-bound), CCR7 have similar fluctuation pattern, and exhibits its highest fluctuations at the loop domains, owed to the lack of secondary structure in the INC and EXC loops, and lowest at the TM domains, which are rich in secondary structure (Figure 3.3).

Despite the similarity in rmsf between the three receptor forms, Figure 3.3 shows differences in

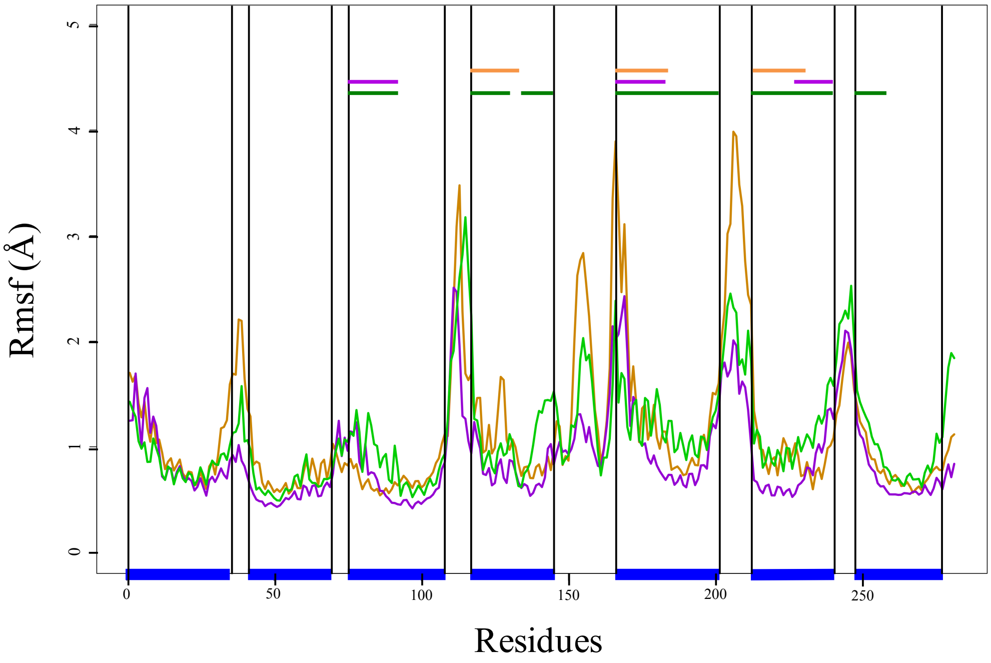


Figure 3.3 Rmsf of CCR7 in its CCL19-bound (purple), CCL21-bound (green), and apo (orange) forms calculated using the same states as in Figure 3.2. The blue horizontal bars at the bottom denote the seven TM domains of CCR7. Bars colored corresponding to the receptor colors indicate segments of high rmsf described in text.

fluctuations for each of the TM domains in the receptor. Fluctuations in the CCL19-bound receptor are relatively high in the EXC regions of TM3, TM5, and TM6, when compared to the CCL21-bound and apo receptors. CCL21-bound simulation shows relatively higher fluctuations in the majority of its TM domains: TM3 (EXC region), TM4, TM5, TM6, and TM7 (EXC region). High fluctuations in the ligand-bound simulations are expected due to the conformational changes characterized above in the DCC maps. In contrast, despite the absence of the ligand in the apo receptor, some domains still exhibit relatively high fluctuations in TM4 (INC region), TM5 (EXC region), and TM6 (INC region), when compared to the bound receptor forms.

The apo and bound forms of CCR7 induce different conformations in previously characterized molecular switches. To better understand receptor behavior, we look into the conformational states of previously determined molecular switches (see Chapter 2). We have previously demonstrated that ligand-induced coordinated fluctuations within specific residue interactions, in particular hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$, induce global helical changes (see Chapter 2). Thus, the conformational states of the molecular switches could explain the lack of correlated conformational changes in the apo receptor.

The correlation observed in the DCC map of the CCL21-bound simulation is a result of hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$ (Figure 3.2B). Conformational changes in the CCL21-bound receptor involve the majority of helical domains of CCR7 (TM3, TM4, TM5, TM6, and TM7) and constitute the largest conformational change seen in all receptor forms. Such conformational changes do not occur in the CCL19-bound and apo receptor due to the lack of hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$. Even though the hydrogen bond does not occur in the CCL19-bound receptor, the receptor still undergoes minor large helical changes portrayed in the DCC map (Figure 3.2A). These motions are localized to only a few helices (EXC regions of TM3, TM5, and TM6) and were previously shown to be associated with allosteric events occurring during the simulation such as: hydrogen bond formation between CCL19's N-terminus and Y41$^{1.39}$ and the reorientation of the χ_1-torsion angle in Y288$^{7.39}$ (see Chapter 2).

Correlated conformational changes arise from the dynamics of the receptor's side chain rearrangements. In this work, the apo receptor shows very little correlation in its DCC map (Figure 3.2). Consequently, we show that previously determined triggers of large conformational changes in the ligand-bound forms of the receptor remain unformed in the apo receptor (Figure 3.4). Specifically, hydrogen bond formation between Y112$^{3.32}$ and Y255$^{6.51}$, which is a critical component of CCR7's dynamics in the CCL21-bound simulation, is non-existing in the apo receptor (Figure 3.4). Y112$^{3.32}$ is hydrogen bonded to Q252$^{6.48}$ in the ligand-bound receptors before it transitions to form a hydrogen bond with Y255$^{6.51}$ (see Chapter 2). In contrast, Y112$^{3.32}$ in the apo receptor points away from the TMD of CCR7 and towards the ligand binding pocket (Figure 3.4). This orientation would otherwise be sterically hindered in the ligand-bound receptor due to the presence of the ligand N-terminal domain (residues 1-7 preceding the first two conserved cysteines) in the ligand binding pocket.

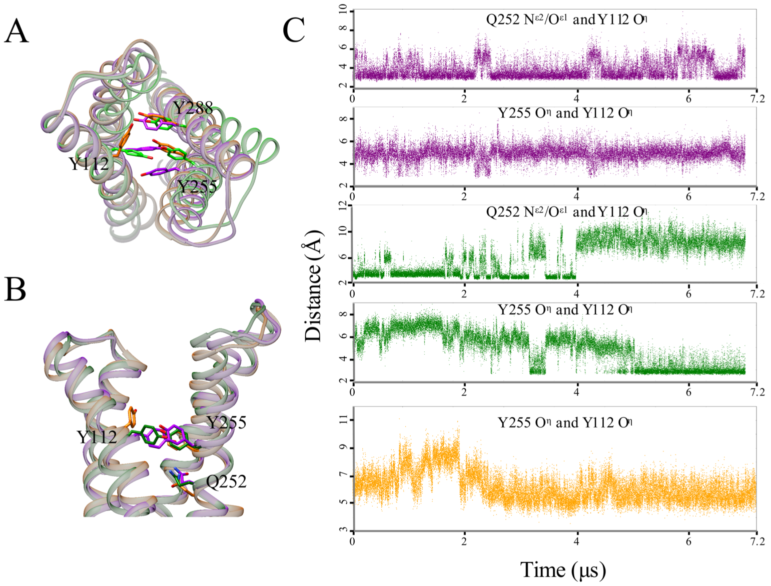


Figure 3.4 Comparison of Y112$^{3.32}$ interactions between CCL19-bound (purple), CCL21-bound (green), and apo (orange) forms of CCR7. (A) Molecular graphics of CCR7 top view illustrating the tri-tyrosine switch. (B) Molecular graphics of CCR7 side view illustrating Y112$^{3.32}$, Q252$^{6.48}$, Y255$^{6.51}$. (C) Polar interaction distance time series between Y112$^{3.32}$ and Q252$^{6.48}$ and Y255$^{6.51}$. Time series are labeled accordingly above each panel. Polar interaction distances are calculated as the minimum distance between side chain polar head group atoms.

3.4 Discussion

In this study, we highlight the importance of the ligand in coordinating the receptor's side-chain fluctuations to drive the conformational changes responsible for state transitions in the receptor. Despite the similar side chain fluctuation pattern in all receptor forms, only the ligand-bound receptors show substantial correlated conformational changes in the receptor. We conclude that the lack of correlation in the apo receptor is owed to the absence of a bound ligand capable of inducing conformational changes in the receptor's molecular switches. We have previously illustrated the role of the conformational changes in the molecular switches to coordinate the large helical motions in CCR7 (see Chapter 2). To that matter, apo CCR7 shows no changes in its molecular switches in the full 7-microsecond simulation, highlighting the importance of the ligand in transmitting information to the molecular switches and mediate the transitioning of the receptor between its different states to carry its function.

Root-mean square fluctuations and dynamic cross-correlation allow us to quantify receptor fluctuations and their correlations, respectively. As we contrast the apo and ligand-bound receptor simulations, we observe that each of the three receptor forms exhibit domains of higher fluctuations when compared to the remaining two simulations. However, only the ligand-bound simulations show high correlations associated with such domains. Correlation within the receptor in each of the CCR7 systems is assessed in the DCC maps. The EXC region of TM3, TM5, and TM6 of the CCL19-bound receptor were found to be correlated alluding to the coordination and directionality of these helical motions. Similarly, domains of high fluctuations in the CCL21-bound receptor are highly correlated in the DCC map. Correlated motions in the CCL19 and CCL21-bound simulations coincide with the higher fluctuation regions calculated by rmsf. In contrast, the apo receptor contains region of higher fluctuations when compared to the bound-receptors, however, these domains show low correlation, which indicate the critical role of the ligand in coordinating the receptor's helical motions.

The role of the ligand in coordinating the fluctuations within the TM domain is dependent on the orthosteric and allosteric changes induced by the ligand as it positions itself within the EXC region of the receptor. All three receptor simulations started from the same conformation of CCR7 to assess the role of the ligand or its absence on the receptor. Despite starting from the same conformation, both apo and bound receptors show different degrees of fluctuations, correlated motions, and orientation of $Y112^{3.32}$, a critical residue shown to induce large helical motions in CCR7. These differences originate from the ligands and prompt an induced fit in the receptor that predisposes it to the various changes in its molecular switches. Receptor plasticity plays a critical role in accommodating each ligand, which in turn induces specific orthosteric and allosteric changes capable of carrying specific receptor function. It is, then, critical to characterize the different receptor states in the EXC region of the receptor in order to initiate efforts in the rational design of therapeutic targets capable of modulating receptor function.

3.5 Acknowledgements

Anton computer time was provided by the National Center for Multiscale Modeling of Biological Systems (MMBioS) through Grant P41GM103712-S1 from the National Institutes of Health and the Pittsburgh Supercomputing Center (PSC). The Anton machine at PSC was generously made available by D.E. Shaw Research.

3.6 References

1. Pilkington, K.R., I. Clark-Lewis, and S.R. McColl. 2004. Inhibition of Generation of Cytotoxic T Lymphocyte Activity by a CCL19/Macrophage Inflammatory Protein (MIP)-3β Antagonist. J. Biol. Chem. 279: 40276–40282.

2. Allegretti, M., M.C. Cesta, A. Garin, and A.E.I. Proudfoot. 2012. Current status of chemokine receptor inhibitors in development. Immunol. Lett. 145: 68–78.

3. Zhu, L., Q. Zhao, and B. Wu. 2013. Structure-based studies of chemokine receptors. Curr. Opin. Struct. Biol. 23: 539–546.

4. Tan, Q., Y. Zhu, J. Li, Z. Chen, G.W. Han, I. Kufareva, T. Li, L. Ma, G. Fenalti, J. Li, W. Zhang, X. Xie, H. Yang, H. Jiang, V. Cherezov, H. Liu, R.C. Stevens, Q. Zhao, and B. Wu. 2013. Structure of the CCR5 Chemokine Receptor–HIV Entry Inhibitor Maraviroc Complex. Science. 341: 1387–1390.

5. Stacker, S.A., S.P. Williams, T. Karnezis, R. Shayan, S.B. Fox, and M.G. Achen. 2014. Lymphangiogenesis and lymphatic vessel remodelling in cancer. Nat. Rev. Cancer. 14: 159–172.

6. Koelink, P.J., S.A. Overbeek, S. Braber, P. de Kruijf, G. Folkerts, M.J. Smit, and A.D. Kraneveld. 2012. Targeting chemokine receptors in chronic inflammatory diseases: An extensive review. Pharmacol. Ther. 133: 1–18.

7. Förster, R., A.C. Davalos-Misslitz, and A. Rot. 2008. CCR7 and its ligands: balancing immunity and tolerance. Nat. Rev. Immunol. 8: 362–371.

8. Haessler, U., M. Pisano, M. Wu, and M.A. Swartz. 2011. Dendritic cell chemotaxis in 3D under defined chemokine gradients reveals differential response to ligands CCL21 and CCL19. Proc. Natl. Acad. Sci. 108: 5614–5619.

9. Schumann, K., T. Lämmermann, M. Bruckner, D.F. Legler, J. Polleux, J.P. Spatz, G. Schuler, R. Förster, M.B. Lutz, L. Sorokin, and M. Sixt. 2010. Immobilized Chemokine Fields and Soluble Chemokine Gradients Cooperatively Shape Migration Patterns of Dendritic Cells. Immunity. 32: 703–713.

10. Legler, D.F., E. Uetz-von Allmen, and M.A. Hauser. 2014. CCR7: Roles in cancer cell dissemination, migration and metastasis formation. Int. J. Biochem. Cell Biol. 54: 78–82.

11. Tutunea-Fatan, E., M. Majumder, X. Xin, and P.K. Lala. 2015. The role of CCL21/CCR7 chemokine axis in breast cancer-induced lymphangiogenesis. Mol. Cancer. 14: 35.

12. Shields, J.D., I.C. Kourtis, A.A. Tomei, J.M. Roberts, and M.A. Swartz. 2010. Induction of Lymphoidlike Stroma and Immune Escape by Tumors That Express the Chemokine CCL21. Science. 328: 749–752.

13. Qin, L., I. Kufareva, L.G. Holden, C. Wang, Y. Zheng, C. Zhao, G. Fenalti, H. Wu, G.W. Han, V. Cherezov, R. Abagyan, R.C. Stevens, and T.M. Handel. 2015. Crystal structure of the chemokine receptor CXCR4 in complex with a viral chemokine. Science. 347: 1117–1122.

14. Burg, J.S., J.R. Ingram, A.J. Venkatakrishnan, K.M. Jude, A. Dukkipati, E.N. Feinberg, A. Angelini, D. Waghray, R.O. Dror, H.L. Ploegh, and K.C. Garcia. 2015. Structural basis for chemokine recognition and activation of a viral G protein–coupled receptor. Science. 347: 1113–1117.

15. Ballesteros, J.A., and H. Weinstein. 1995. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In: Sealfon SC, editor. Methods in Neurosciences. Academic Press. pp. 366–428.

16. Shaw, D.E., R.O. Dror, J.K. Salmon, J.P. Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, K.J. Bowers, E. Chow, M.P. Eastwood, D.J. Ierardi, J.L. Klepeis, J.S. Kuskin, R.H. Larson, K. Lindorff-Larsen, P. Maragakis, M.A. Moraes, S. Piana, Y. Shan, and B. Towles. 2009. Millisecond-scale molecular dynamics simulations on Anton. In: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. . pp. 1–11.

17. R. Core Team. 2014. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

18. Van Rossum, G., and others. 2007. Python Programming Language. USENIX Annu. Tech. Conf. 41.

19. Pettersen, E.F., T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. 2004. UCSF Chimera-A visualization system for exploratory research and analysis. J. Comput. Chem. 25: 1605–1612.

20. Grant, B.J., A.P.C. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S.D. Caves. 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 22: 2695–2696.

21. Skjærven, L., X.-Q. Yao, G. Scarabelli, and B.J. Grant. 2014. Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics. 15.

22. Venkatakrishnan, A.J., X. Deupi, G. Lebon, C.G. Tate, G.F. Schertler, and M.M. Babu. 2013. Molecular signatures of G-protein-coupled receptors. Nature. 494: 185–194.

23. Henzler-Wildman, K., and D. Kern. 2007. Dynamic personalities of proteins. Nature. 450: 964–972.

CHAPTER 4: DETECTION OF MOLECULAR SWITCHES MEDIATING DOMAIN MOTIONS IN MOLECULAR DYNAMICS SIMULATIONS OF MEMBRANE PROTEINS

4.1 Introduction

Protein function is encoded into its dynamics as a large ensemble of conformations that can be grouped to different conformational states depending on protein function, free energy, and three-dimensional arrangement (1, 2). All accessible conformational states can be sampled by the protein regardless of any outside perturbation (ligand-binding, amino acid mutation, post translational modification, or environmental changes such as pH, ionic strength, temperature, etc.) (3). Conformational states are accessed at different equilibrium sampling probabilities depending on their energies, where, a ligand-free protein may still briefly sample its intermediate or active states despite favoring its inactive state (1). On the other hand, external perturbations, such as ligand-binding, result in an equilibrium shift where the protein favors its active state.

Allosteric function plays an important role in transmitting information between distant functional sites of the protein as a mechanism to regulate its transitions and sampling of conformational states upon external perturbation (1, 2, 4). To understand such a mechanism, we must ask how do the mechanics of protein structures emerge from the rearrangement of their constituent parts, specifically, side chain interactions within structured regions of proteins.

Molecular dynamics (MD) simulation is one of the many techniques used to study protein dynamics at atomic level (2). Several recent advances in enhanced sampling methods, simulation speed, and accuracy have allowed us to reach biologically relevant timescale that capture the transitioning of a protein between different states; and consequently, allow the study of allostery (2). In protein dynamics, large domain motions are sampled on the hundreds of nanosecond to microsecond timescales, which are now readily accessible by MD simulations (5–7). Accordingly, several studies have explored the folding mechanism of a number of fast folding proteins (8) and captured the transitioning of proteins between different states (9, 10). Protein MD simulations involve two major types of motions: large domain and side chain conformational changes. These motions constitute the dynamical components that facilitate the transmission of signals between distant sites in a protein in order to regulate its sampling of different states (1, 2).

Many MD analysis tools have been developed to systematically extract biologically relevant information encoded in large domain and local side chain motions of a protein. Widely used methods involve the detection of large domain conformational changes using principal component analysis (PCA) and dynamic cross correlation (DCC) applied to the three-dimensional coordinates of simulated proteins (11–13). Both methods focus on dominant protein motions and neglect the complex detail of the more intricate local motions at the side chain level, sampled by the protein to mediate its allosteric communication and state transitioning.

Other methods mainly revolve around detection of abrupt changes in spatiotemporal data comprising of inter-atomic distances or three-dimensional coordinate time series (14–16). The most recent method, SIMPLE, is designed to favor the detection of collective change points, depending on a sensitivity parameter (16). Despite the advances in event detection made possible by SIMPLE, this method still comes short in detecting functional molecular switches. Depending on the sensitivity parameter used, such motions can either be obscured by the large number of detected change-points when using a low sensitivity parameter, or omitted when using a high sensitivity parameter. Molecular switch detection in protein dynamics presents a challenging problem for the following reasons. First, functional molecular switches are subtle and manifest themselves as a single inter-residue interaction rearrangement that can be obscured by the several fluctuating inter-residue interactions. Change-point detection methods, such as SIMPLE, identifies all side chains rearrangements; however, several of which may not be functional. Second, molecular switches involve the more sporadic movement of amino acid residue side chains and could involve time delays and weak coupling to the larger domain movements of the protein. To overcome these challenges, our work focuses on extracting these molecular switches from all detected side chain rearrangements by assessing their correlation to the larger domain motions in the protein.

In this work, we reduce the protein dynamics to its constitutive dynamical components by screening for large domain motions and side chain rearrangements. We detect molecular switches using

Gaussian Mixture Models (GMM) and domain motions using DCC. Molecular switches and domain motions are then related through a DCC-based network and compartmentalized into different communities with similar dynamics. This is an efficient method to relate the local side chain motions (molecular switches) to the global domain motions of the protein. The different network communities comprise of side chain distance time series that are correlated (or anti-correlated) to the large domain motions of the protein. These dynamical components provide an understanding of how side chain rearrangements mediate the global motions of the protein, which eventually facilitate its transitioning between different protein functional states.

4.2 Materials and Methods

We introduce an approach to delineate the structural mechanism of allosteric regulation in protein dynamics (Figure 4.1). Our computational framework is designed to systematically extract side chain rearrangements mediating large domain motions from a protein's MD simulation trajectory. This is done by reducing the overall behavior of the protein to a set of coupled dynamical components, composed of side chain and backbone rearrangements.

Side chain rearrangements are often localized to a single inter-residue side chain interaction, which could be obscured by larger domain motions when extracted from a large MD data set of inter-atomic distance time series. Therefore, both dynamical components, side chain (Figure 4.1A) and backbone dynamics (Figure 4.1B), are extracted separately using different methods: GMMs and DCC, respectively.

Given the dynamic nature of proteins, only a tiny fraction of the protein's side chain dynamics is considered to behave as molecular switches that contribute to regulating the global protein dynamics. Therefore, extracted side chain rearrangements (Figure 4.1A) are further reduced by extracting those that are coupled to the large domain motions (Figure 4.1B). All dynamical components are projected into a DCC-based network and categorized into different communities, where large domain motions and side chain dynamics within the same community show correlated time series (Figure 4.1C).

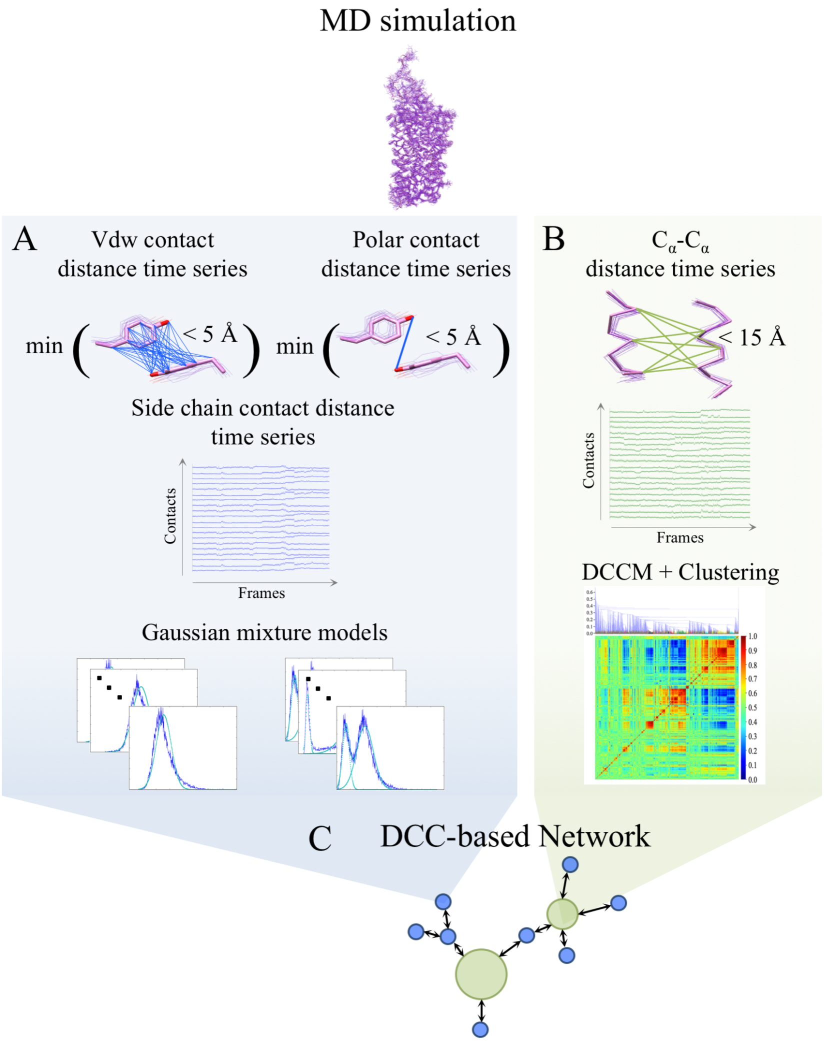


Figure 4.1 Schematic of our computational framework to detect molecular switches mediating large domain motions in proteins. (A) Van der Waals and polar interactions that sample a maximum distance of 5 Å during the simulation are used to calculate distance time series from the MD simulation 3-dimentional data. The minimum distance between all side chain or polar atoms are used to extract inter-residue side chain distance time series. Probability density of each time series are fitted to a GMM to extract side chain interactions that undergo rearrangements during the simulation. (B) C_α-C_α interactions that sample a maximum distance of 15 Å during the simulation are used to calculate the C_α-C_α distance time series. A DCC matrix of all pairwise C_α-C_α distance time series are clustered and clusters with a minimum coefficient of 0.95 are extracted as large domain motions of the protein. (C) Side chain rearrangements (blue nodes) and large domain motions (green nodes) of the protein are considered dynamical components of the protein and are inputted into a DCC-based network to relate both components. Network connections are based on the correlation coefficients of pairwise dynamical components which are calculated as the average DCC coefficient of the pairwise time series belonging to each component.

Detection of side chain contact rearrangements from MD simulations. Extracting all side chain rearrangements from MD simulations involves the identification of side chain interactions that experience abrupt changes in their distance time series. Abrupt changes in the inter-residue interactions indicate that these interactions are capable of switching between substates. We extract such inter-residue interactions by fitting a GMM to the probability density of each interaction distance timeline. GMMs are weighted sums of Gaussian densities and are used here as a parametric model of the probability density function of inter-residue time series (Gaussian densities are implemented in *scikit-learn*, a machine learning package in python) (17). Stable non-varying interactions show a unimodal distribution (Figure 4.2A), and multi-substate interactions show multi-modal distributions (Figure 4.2B). The optimal number of Gaussians was efficiently determined using the Bayesian information criterion using *scikit-learn* (17). GMM parameters are estimated using the iterative expectation-maximization algorithm, where the number of Gaussians is predetermined. This section of the computational framework is designed to extract all interactions that show contact formation and breakage at any point during the simulations, as such contacts can be deemed critical in mediating the larger domain motions. GMMs are fitted to all distance time series representing van-der-Waals (vdw) and polar interaction (listed below) distances between interacting side chain residues. Interacting residues used to calculate the distance time series are determined as the residues that came into contact (a distance of at least 5 Å between all non-hydrogen side chain atoms) at any point during the simulation. To ensure complete formation and deformation of the side chain contacts, we calculate the inter-residue side chain distance time series using the minimum distance between all non-hydrogen side chain atoms of each of the amino acids. Similarly, polar interactions are also calculated using the minimum distance between all non-hydrogen polar head group atoms of interacting polar amino acids (atoms C_{ζ}, N_{ε}, $N_{\eta1}$, or $N_{\eta2}$ for R; atoms C_{γ}, $O_{\delta1}$, or $N_{\delta2}$ for N; atoms C_{γ}, $O_{\delta1}$, or $O_{\delta2}$ for D; atom S_{γ} for C; atoms C_{δ}, $O_{\varepsilon1}$, or $N_{\varepsilon2}$ for Q; atoms C_{δ}, $O_{\varepsilon1}$, or $O_{\varepsilon2}$ for E; atoms CG, $N_{\delta1}$, $C_{\varepsilon1}$, $N_{\varepsilon2}$, or $C_{\delta2}$ for H; atom N_{ζ} for K; atom O_{γ} for S; atom $O_{\gamma1}$ for T; atom $N_{\varepsilon1}$ for W; atom O_{η} for Y). All distance time series are fit with a GMM to identify the number of substates that each interaction is sampling.

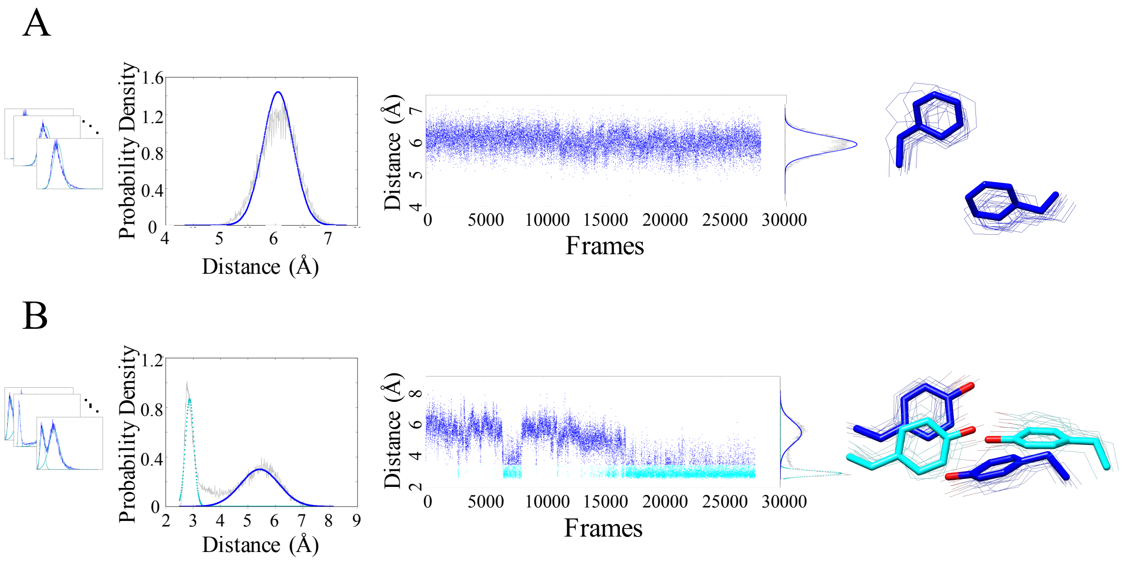


Figure 4.2 Examples of side chain distance probability densities fitted using GMM. (A) Side chain distance probability densities fitted by unimodal distributions show a stable inter-residue interaction through the majority of the simulation. (B) Side chain distance probability densities fitted by multimodal distributions represent inter-residue interactions that undergo rearrangements during the simulation.

Distance time series with unimodal GMMs are considered to be stable during the simulations, contributing to the structural stability (robustness) of the protein. On the other hand, multi-modal GMMs are amongst the dynamical components of the protein and contribute to the protein's conformational transitions between different functional states.

Detection of large domain motions through DCCM. Large domain motions in proteins involve the collective motion of backbone atoms and aid in the transitioning of the protein between different functional states. This part of the computational framework entails the detection of these motions as a collection of highly correlated inter-C_α distance time series.

All alpha carbon interactions within 15 Å at any point of the simulation are extracted, and all distance time series representing theses interactions are calculated. Pairwise dynamic cross-correlation of all distance time series are clustered based on their correlation coefficient and clusters with at least 0.95 correlation coefficient are extracted (Figure 4.3A, B). Each cluster is a set of time series that are localized to different protein sectors that exhibit different dynamical behaviors (Figure 4.3C).

The use of distance time series (rather than XYZ coordinates) presents various advantages in molecular dynamics simulation analysis. Apart from reducing the dimensionality of the data time series used (from three-dimensional XYZ coordinates to one-dimensional distance time series), the translation and rotation of the protein during the MD simulations can be ignored and therefore structure superimposition can be omitted. These improvements allow us to accentuate the changes in the global structure of the protein and attenuate the effects of atomic fluctuations seen in the XYZ coordinates. Thus, clusters with high DCC coefficient better portray the large domain dynamical behavior of the protein.

Network of the protein's dynamical components. The dynamical components of the protein are classified into different communities, using igraph (18). We create a DCC-based network comprised of the dynamical components of the protein (Figure 4.4), extracted in the previous sections and described in Figures 4.2 and 4.3. In the network of Figure 4.4, the blue and green nodes represent side chain and

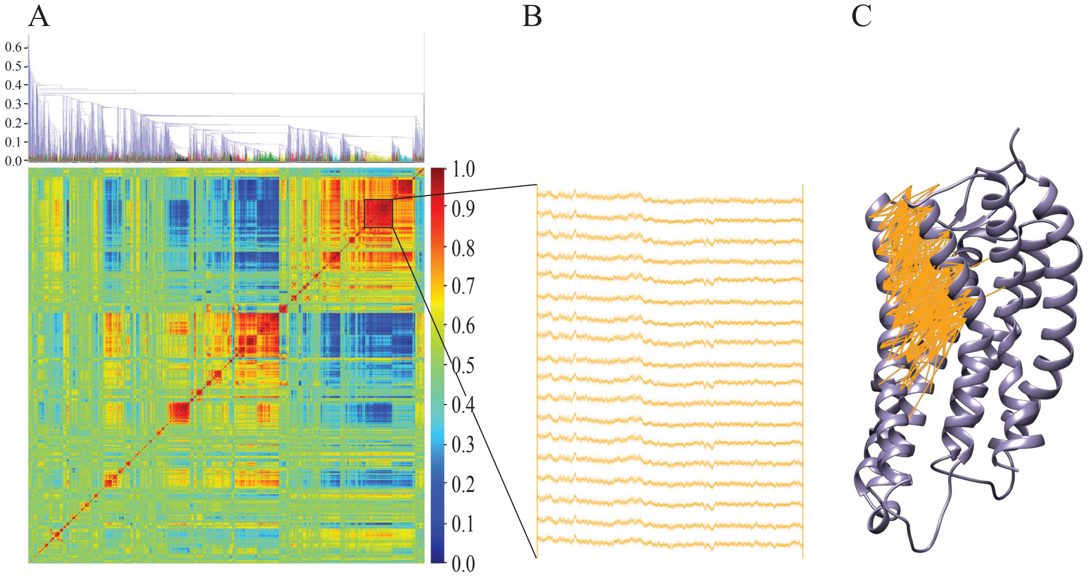


Figure 4.3 DCC heat map of pairwise C_α-C_α distance time series are clustered using hierarchical clustering. (A) The clustering dendrogram is reported above the DCC heat map. The DCC coefficient is used as the distance calculated between two clusters and shown as the y-axis of the dendrogram. Each color of the dendrogram represents a different cluster of time series that are correlated at a cutoff DCC coefficient of 0.95. (B) An illustration of the time series within the highlighted cluster in (A). (C) An example of molecular graphics demonstrating the interacting residues involved in the large domain motions illustrated in the highlighted cluster in (A). Each connection involves two C_α whose distance time series is within the highlighted cluster in (A).

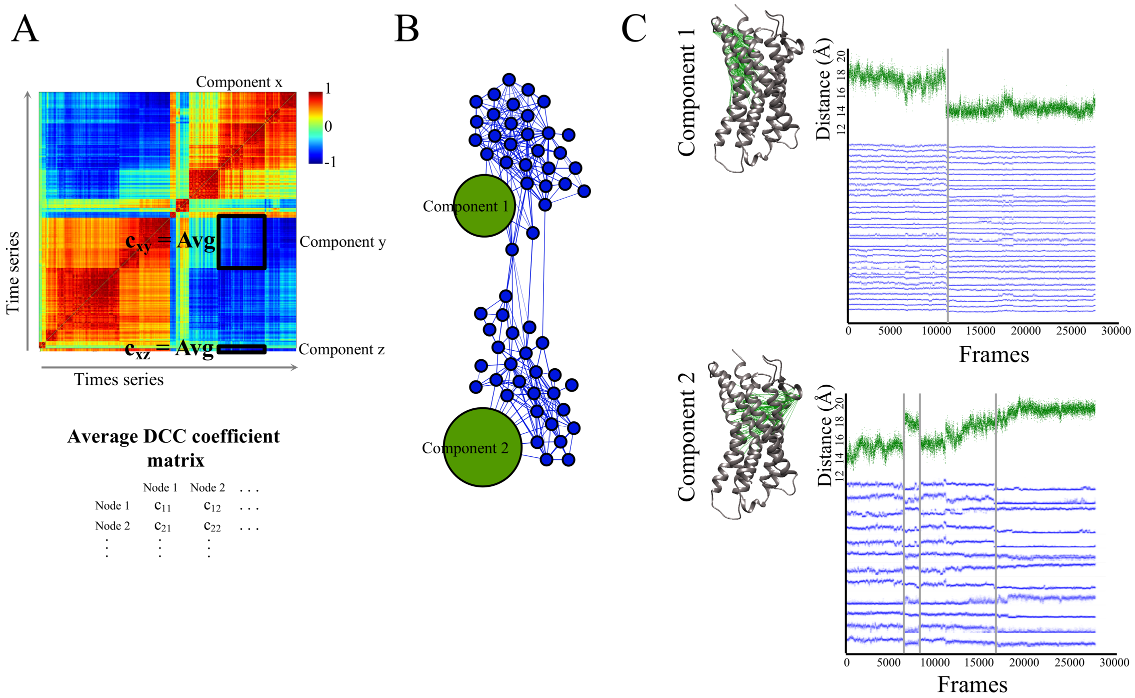


Figure 4.4 DCC-based network illustration of the protein's dynamical components. (A) Correlation coefficients of pairwise dynamical components are calculated as the average DCC coefficient of the pairwise time series belonging to each component as illustrated on a sample DCC heat map. Average correlation are calculated between pairwise large domain motions (components x and y), between pairwise side chain rearrangement time series (component z) and across both components. An average DCC coefficient matrix is generated for all dynamical components. (B) The network is built from a subset of the time series extracted from the MD simulation of CCL21-bound CCR7. The network is composed of two communities that are centered around large domain motions labeled as component 1 and component 2. Network nodes represent the dynamical components extracted from the subset time series data and are colored blue for side chain rearrangements and green for large domain motions. The size of each node is proportional to the number of time series the node represents. Edges connecting the dynamical components are based on the average pairwise DCC coefficient of the time series involved in each of the components. Edges are drawn between dynamical components of a minimum coefficient of 0.75. C_α-C_α distance. (C) Time series that comprise each of components 1 and 2 are projected into the molecular graphics of CCR7 and labeled accordingly. Components 1 and 2 represent large domain motions in a protein and are constituted of several highly correlated C_α-C_α distance time series. A sample time series from each of the large domain motion components is shown in green. Blue time series are side chain time series for each of the blue nodes within each of the communities centered around components 1 and 2. All time series show coupled abrupt changes within each of the large domain movements components highlighted in grey. The network was built using Gephi (19).

backbone interactions, respectively; and edges connect correlated components with a minimum correlation or maximum anti-correlation coefficients of 0.75 or -0.75 respectively. The correlation coefficient cutoff is defined by the user and can be adjusted to account for the faster behavior of side chain rearrangements (picosecond and nanosecond time scales) compared to the slower motions of larger protein domains (microsecond and millisecond time scales). While rearrangement in side chain interaction are manifested as abrupt changes in the distance time series, the larger domain motion experience more incremental changes that span hundreds of nanoseconds. The size of each node is proportional to the number of distance time series each node represents, where large domain dynamical components involve a large number of time series, while side chain rearrangements are characterized by one distance time series.

Correlation coefficients of pairwise dynamical components are calculated as the average DCC coefficient of the pairwise time series belonging to each component (Figure 4.4A). An average DCC coefficient matrix is generated for all the pairwise dynamical components. The produced matrix is projected into a network where components are connected based on a DCC coefficient cutoff (Figure 4.4B). Network communities are detected based on edge betweenness, where each community is composed of side chain and backbone dynamics that are correlated to each other. The extracted communities represent the large domain motions and the molecular switches that mediate the protein dynamics (Figure 4.4C).

Network community visualization using molecular graphics visualization tools. MD simulations provide an unsurmountable amount of dynamical information due to the high fluctuating and complex nature of protein dynamics. Here, the extracted communities, using igraph, were revealed to be useful in reducing the protein to its functional dynamical behavior. Each community is composed of molecular switches and large domain motions that are correlated. These communities can be outputted into a protein data bank file format to visualize the residues that make up the dynamical components of the community. The time series belonging to the dynamical components are outputted as an edge connecting two representative atoms of the time series' corresponding residues (Figure 4.3C).

4.3 Application to Molecular Dynamics Simulation Data

We apply our computational framework to previously published MD trajectories where we analyzed the simulations to understand the mechanism by which a ligand, CCL21 in our test case here, transmits information in CC chemokine receptor 7 (CCR7) (see Chapter 2). We have determined key conformational changes that act as molecular switches and facilitate the transitioning of the receptor between its different states by inducing large helical motions of its transmembrane domain (TMD). The simulation data of CCR7 was originally analyzed through manual and visual inspection of a large set of distance time series and generic summary measurement such as root mean square deviation (RMSD), principal component analysis (PCA), and comparison of the inter-residue mean distances between different time segments. Such non-systematic measures are very labor intensive and may not provide a complete analysis due to the overwhelming amount of the data outputted by the MD simulations. Nonetheless, we were able to detect a series of molecular switches that are mediated by various ligand-induced allosteric events. These molecular switches involve three tyrosine residues (Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$), three phenylalanine residues (F116$^{3.36}$, F208$^{5.47}$, and F248$^{6.44}$), and a polar interaction between Q252$^{6.48}$ and R294$^{7.45}$ in the TMD of CCR7 (see chapter 2). Molecular events within these switches are coupled with global helical movements in the receptor's TM helices and contribute to the transitioning of the receptor to distinct states.

Using a distance cutoff of 5 Å, a total of ~1200 inter-residue side chain distance time series were imported and fit to a GMM in order to systematically extract all multi-modal distance probability densities. The selected contacts reduced our data set to ~600 time series. However, the majority of these contacts comprise independent side chain rearrangements that do not contribute to the protein's major helical motions, and only a small fraction of these multi-modal contacts are considered to act as molecular switches in our previous analysis (see Chapter 2). The second part of our computational framework focuses on extracting the receptor's large domain motions using inter-residue C_α distance time series with a cutoff of 15 Å. A pairwise DCC matrix was generated for ~6000 distance time series, and then clustered at a DCC coefficient cutoff of 0.95. The high DCC cutoff generated clusters with highly correlated distance time series that involve structurally adjacent amino acids. This part of the computational framework generated ~1000 clusters which included multiple clusters of more than a hundred time series. After calculating all pairwise average DCC coefficients between all dynamical components and projecting our data onto a DCC-based network, all dynamical components were then reduced to a few communities with very similar dynamical behaviors, from which we have extracted coupled side chain and large domain motions.

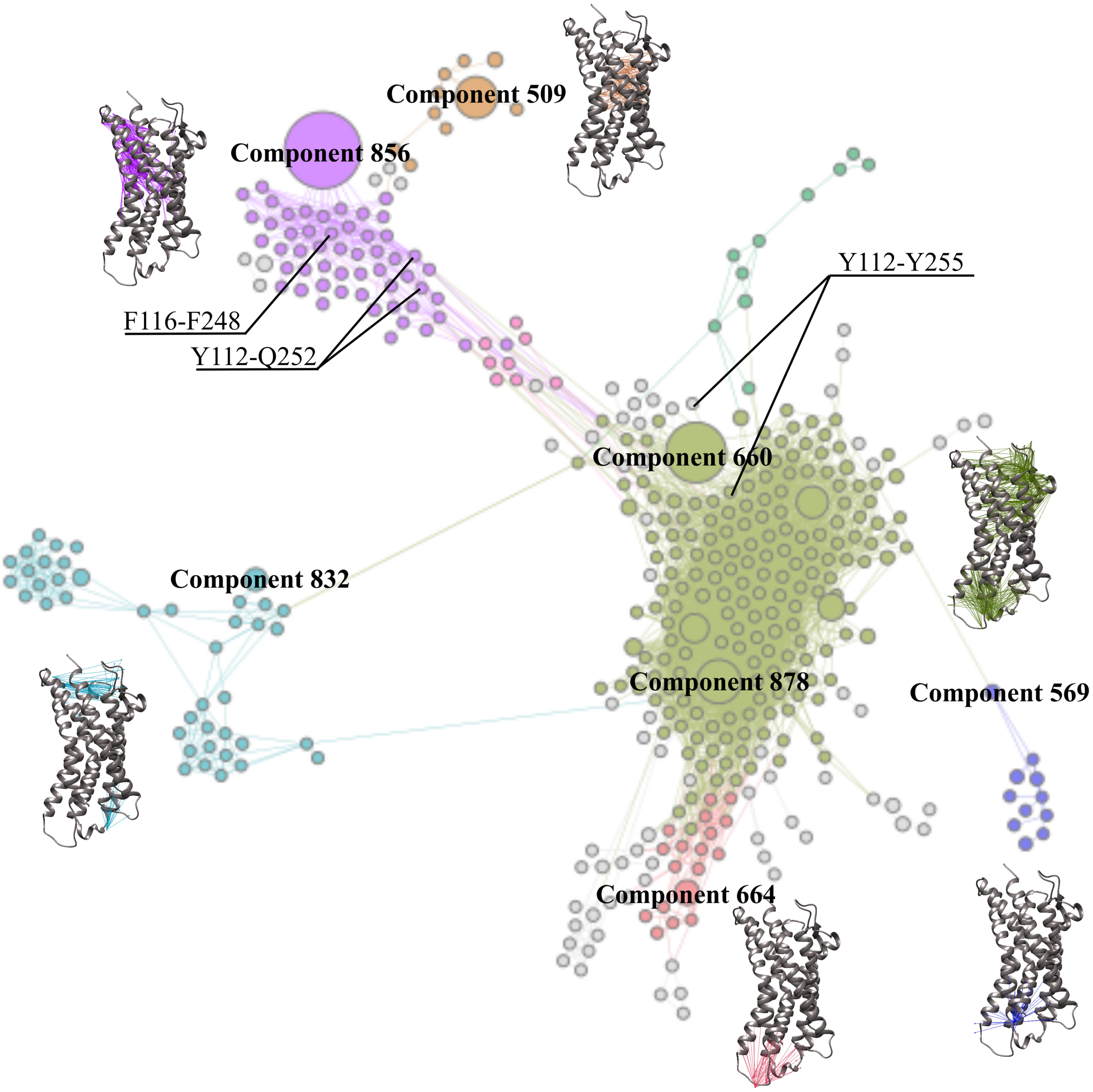


Figure 4.5 A DCC-based network of the full CCL21-bound CCR7 MD simulation dataset. Network communities are colored differently and dynamical components representing large domain motions are projected into a molecular graphics in which connections are colored according to the community they belong to. Previously determined molecular switches (F116-Q252, Y112-Q252, Y112-Y255) are labeled accordingly in the network (see Chapter 2).

Using our computational framework, we were able to systematically extract correlated molecular switches and large helical motions (Figure 4.5). In each of the network communities, the different nodes represent the large domain motions (large nodes) and the side chain rearrangements (small nodes) extracted from the MD simulation. Each of the five main communities is composed of a few nodes that represent the large domain movement of the receptor and several nodes that represent side chain rearrangements. Within the detected side chain rearrangements, Figure 4.5 illustrates the presence of previously determined molecular switches: F116-F248, Y112-Y255, and Y112-Q252. These molecular switches belong to different communities centered around large helical motions of the receptor, which demonstrates their coupling to different large dynamical components of the receptor. Our systematic approach has also identified new dynamical components of CCR7 that were overlooked in our previous manual and visual analysis (see Chapter 2). Specifically, large domain motions involving extracellular and intracellular domains of the receptor were extracted from the network and were found to either belong to independent communities (cyan, blue, and red communities in Figure 4.5) or coupled to other large helical movements in the TMD (green community in Figure 4.5).

4.4 Concluding Remarks

This computational framework focuses on linking the different dynamical components of a protein in order to extract side chain rearrangements that are coupled to global conformational changes. This is done through the detection of side chain contacts with multi-modal probability density function and large domain motions as clusters of highly correlated inter-residue C_α distance time series. Community detection in a DCC-based network of all extracted components correlate the side chain contacts to the large domain motions in order to pinpoint the different molecular switches that are coupled to the large domain motions detected in the protein dynamics.

As a proof of concept, this method was used to systematically detect the molecular switches responsible of mediating the large helical motions in CCR7 previously extracted (see Chapter 2). Ultimately, our computational framework reduces the overall behavior of the protein to a set of coupled dynamical components, composed of side chain and backbone rearrangements. This method provides a manageable dataset that is easily visualized in three-dimensional molecular graphics for visual analysis.

4.5 References

1. Motlagh, H.N., J.O. Wrabl, J. Li, and V.J. Hilser. 2014. The ensemble nature of allostery. Nature. 508: 331–339.

2. Henzler-Wildman, K., and D. Kern. 2007. Dynamic personalities of proteins. Nature. 450: 964–972.

3. Grant, B.J., A.A. Gorfe, and J.A. McCammon. 2010. Large conformational changes in proteins: signaling and other functions. Curr. Opin. Struct. Biol. 20: 142–147.

4. Popovych, N., S. Sun, R.H. Ebright, and C.G. Kalodimos. 2006. Dynamically driven protein allostery. Nat. Struct. Mol. Biol. 13: 831–838.

5. Salomon-Ferrer, R., A.W. Götz, D. Poole, S. Le Grand, and R.C. Walker. 2013. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. 9: 3878–3888.

6. Shaw, D.E., R.O. Dror, J.K. Salmon, J.P. Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, K.J. Bowers, E. Chow, M.P. Eastwood, D.J. Ierardi, J.L. Klepeis, J.S. Kuskin, R.H. Larson, K. Lindorff-Larsen, P. Maragakis, M.A. Moraes, S. Piana, Y. Shan, and B. Towles. 2009. Millisecond-scale molecular dynamics simulations on Anton. In: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. . pp. 1–11.

7. Miao, Y., F. Feixas, C. Eun, and J.A. McCammon. 2015. Accelerated molecular dynamics simulations of protein folding. J. Comput. Chem. 36: 1536–1549.

8. Lindorff-Larsen, K., S. Piana, R.O. Dror, and D.E. Shaw. 2011. How Fast-Folding Proteins Fold. Science. 334: 517–520.

9. Miao, Y., S.E. Nichols, P.M. Gasper, V.T. Metzger, and J.A. McCammon. 2013. Activation and dynamic network of the M2 muscarinic receptor. Proc. Natl. Acad. Sci. 110: 10982–10987.

10. Dror, R.O., D.H. Arlow, P. Maragakis, T.J. Mildorf, A.C. Pan, H. Xu, D.W. Borhani, and D.E. Shaw. 2011. Activation mechanism of the β2-adrenergic receptor. Proc. Natl. Acad. Sci. 108: 18684–18689.

11. Grant, B.J., A.P.C. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S.D. Caves. 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 22: 2695–2696.

12. Sethi, A., J. Eargle, A.A. Black, and Z. Luthey-Schulten. 2009. Dynamical networks in tRNA:protein complexes. Proc. Natl. Acad. Sci. 106: 6620–6625.

13. Stolzenberg, S., M. Michino, M.V. LeVine, H. Weinstein, and L. Shi. Computational approaches to detect Allosteric pathways in Transmembrane Molecular Machines. Biochim. Biophys. Acta BBA - Biomembr. .

14. Ensign, D.L., and V.S. Pande. 2010. Bayesian Detection of Intensity Changes in Single Molecule and Molecular Dynamics Trajectories. J. Phys. Chem. B. 114: 280–292.

15. Wriggers, W., K.A. Stafford, Y. Shan, S. Piana, P. Maragakis, K. Lindorff-Larsen, P.J. Miller, J. Gullingsrud, C.A. Rendleman, M.P. Eastwood, R.O. Dror, and D.E. Shaw. 2009. Automated Event

Detection and Activity Monitoring in Long Molecular Dynamics Simulations. J. Chem. Theory Comput. 5: 2595–2605.

16. Fan, Z., R.O. Dror, T.J. Mildorf, S. Piana, and D.E. Shaw. 2015. Identifying localized changes in large systems: Change-point detection for biomolecular simulations. Proc. Natl. Acad. Sci. 112: 7454–7459.

17. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12: 2825–2830.

18. Csardi, G., and T. Nepusz. 2006. The igraph software package for complex network research. InterJournal Complex Syst. 1695: 1–9.

19. Bastian, M., S. Heymann, and M. Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. In: Third International AAAI Conference on Weblogs and Social Media. .

CHAPTER 5: LYSINE ACETYLATION AND METHYLATION REGULATE KU70 FUNCTIONS

5.1 Introduction

Double-strand breaks (DSB) in DNA are one of the most dangerous forms of DNA break due to the intrinsic difficulty of its repair and its susceptibility to hazardous DNA errors (1, 2). DSBs can be repaired by two distinct repair mechanisms: homologous recombination (HR) and non-homologous end-joining (NHEJ). Both mechanisms are relatively accurate. However, the lack of a homologous template in NHEJ makes it error-prone and more susceptible to small sequence deletions (2, 3). Accumulating evidence links unrepaired or improperly repaired DNA DSBs to chromosomal translocation and abnormalities that lead to carcinogenic effects and health problems (2, 4).

High levels of recombination defects have been attributed to KU deficiency *in vitro* and *in vivo*, thus making KU70-KU80 heterodimer a critical complex in stable and accurate NHEJ (3). KU70-KU80 is the first protein recruited to the site of DNA break and deletion of either one of these KU proteins leads to impairment of DNA double strand break repair and sensitivity to radiation (5, 6).

Post-translational modifications (PTMs) are critical for proper DNA damage repair and have been widely studied in the context of NHEJ (7–9). Multiple KU70 (de)acetylation instances have been shown to regulate the NHEJ repair mechanism (7, 10–13). In chronic myeloid leukemia (CML) and prostate cancer cells, KU70 deacetylation promotes the cell's ability to repair DNA DSBs and its survival (7, 10, 11, 13, 14). Furthermore, KU70 deacetylation promotes acquisition of resistant BCR-ABL (CML oncoprotein) mutations in CML cells in association with its ability to stimulate aberrant NHEJ activity (11). However, in neuroblastoma cells, KU70 acetylation increases DNA repair activity when subject to ionizing radiation (12). Despite the paradoxical role of (de)acetylation in the three cell types, there is a clear role of acetylation in regulating proper DNA damage repair.

Our collaborators at the City of Hope, Dr. WenYong Chen's lab, have recently explored another form of PTMs on KU70, methylation, following a recent proteomic study that has identified lysine methylation on KU70 and several other DNA repair factors (15), but had not located specific methylation sites. Dr. Chen's lab has found that a lysine-specific demethylase 1 (LSD1) competes with SIRT1 for

KU70-binding and to regulate DNA repair and mutation acquisition in cancer cells, where LSD1 demethylates KU70 at three lysine residues (K9, K74 and K510) (WenYong Chen, unpublished data). Mutations of these lysines disrupt KU70 interaction with SIRT1, formation of KU70-KU80 heterodimer, and KU70's function in introducing genetic mutations (WenYong Chen, unpublished data).

Dysregulation of KU70-KU80 through acetylation and methylation of lysine residues has been shown in various cell lines to introduce DNA mutations through an error-prone NHEJ. However, no molecular mechanism describes how such PTMs might affect DNA-repair. Our objective is to assess the effects of acetylation and methylation on the interactions of KU70-KU80 with DNA, and its deacetylase and demethylase enzymes, SIRT1, and LSD1. Acetylation and methylation of KU70 are believed to alter the heterodimer's function through its dynamics by inducing various conformational changes in the protein (16–19). Molecular dynamics (MD) simulations of the heterodimer in its different PTM states and in the absence or presence of DNA suggest that methylation and acetylation of specific lysine sites produces electrostatic and long-range intra-molecular structural alterations within KU70 that may account for its functional changes. Our study sheds novel insight into the molecular mechanism governing the acquisition of genetic mutations and cancer drug resistance.

5.2 Materials and Methods

Construction of the KU70-KU80-DNA model. The crystal structure of KU70-KU80-DNA was obtained from the Protein Data Bank (PDB) using the PDB code 1JEY (3). The crystal structure of the complex was determined as KU70-KU80-DNA trimer without the KU70 N-terminal random coil (NRC: amino acids 1-33) or the linker-SAP domain (amino acids 535-609), therefore the missing coordinates were constructed. The SAP domain (amino acids 561-609) was obtained from the crystal structure of the DNA-free KU70-KU80 complex with PDB code 1JEQ (3). The missing NRC and the linker between the KU70 core and SAP domain were added using Modeller9.11 (20). Alignment of the KU70 sequences from www.uniprot.com and from the PDB file 1JEY was performed using the ClustalW2 server (21). Modeller9.11 was used to generate the structure from the alignment. Then, the SAP domain, extracted from

1jey, was attached to the rest of the modeled protein. The DNA three-way junction, used to force KU70-KU80 towards a single DNA binding mode for crystallographic purposes, was truncated to a 14-bp DNA duplex. Methyl groups were added using Chimera (22), by replacing hydrogen atoms of the side chain amine of K9, K74 and K510. K9 and K74 are dimethylated and K510 is monomethylated. All of the three lysines were computationally methylated or nonmethylated simultaneously. Acetyl groups were also added using Chimera to eight experimentally identified acetylated KU70 residues: K282, K317, K331, K338, K539, K542, K544, K553, and K556. All of these eight lysines were computationally acetylated or nonacetylated simultaneously. Four structures of KU70-KU80 were generated under nonmethylated and methylated conditions for a total of eight different structures. The four structures are the following complexes: KU70-KU80, KU70-KU80-DNA, acetylated state of KU70-KU80, and acetylated state of KU70-KU80-DNA.

Analysis of Electrostatic Similarities of Proteins (AESOP) framework. To delineate the role of charged residues in binding, calculations were performed using integrated Analysis of Electrostatic Similarities Of Proteins (AESOP) framework (23–27). The coordinates of the complex were obtained from the constructed KU70-KU80-DNA model. The calculations involved systematic mutation of the side chains of charged ionizable residues (Arg, Asp, His, Glu, and Lys) into alanine, one at a time, thus generating a family of mutant proteins. The program PDB2PQR was used to add atomic radii and partial charges to the atomic coordinate file using the CHARMM forcefield (28, 29). The electrostatic potentials were calculated by numerically solving the linearized Poisson-Boltzmann equation with the program Adaptive Poisson-Boltzmann Solver (APBS) (30). The molecular (dielectric boundary) and ion accessibility surfaces were determined using spherical probes and radii set to 1.4 and 2.0 Å respectively. The APBS calculations were performed on a grid resolution of ~1 Å for both 0 and 150 mM ionic strength. The dielectric coefficients for the protein interior and solvent for each complex were set to 20 and 78.54, respectively. Two electrostatic potential calculations were performed with different counter ion concentrations of 0 mM and 150 mM.

Pairwise similarities were calculated according to the electrostatic similarity distance (ESD) equation (1).

$$ESD = \frac{1}{N}\sum_{i,j,k}\frac{|\varphi_B(i,j,k)-\varphi_A(i,j,k)|}{\max(|\varphi_B(i,j,k),\varphi_A(i,j,k)|)} \qquad (1)$$

In Eq. (1), ϕ_A and ϕ_B refer to electrostatic potential at grid point (i, j, k) in proteins A and B, respectively, and N represents the total number of grid points at which electrostatic potential has been calculated. An ESD value of 0 denotes identical electrostatic potentials. As the ESD value increases, the dissimilarity in electrostatic potential increases (25, 27). The resulting complexes were hierarchically clustered in a dendrogram depending on their ESD value, using the linkage method.

Free energies of association were calculated based on the Coulombic potential and solvation energies of the complex. As described by equation (2), the solvation energies are calculated using a thermodynamic cycle in order to remove self-energies and grid artifacts described previously (24, 27, 31). Coulombic potentials are then calculated using nongrid-based Coulomb module in APBS and incorporated with the solvation energies to extract the free energies of association (equation (3)).

$$\Delta\Delta G^{solvation} = \Delta G_{AB}^{solvation} - \Delta G_{A}^{solvation} - \Delta G_{B}^{solvation} \qquad (2)$$

$$\Delta G^{association} = \Delta G^{coulombic} + \Delta\Delta G^{solvation} \qquad (3)$$

In these equations, AB refers to the protein complex KU70-KU80-DNA, A refers to free KU70-KU80, and B refers to free DNA. The electrostatic free energies of the mutants are represented relative to the parent protein as described in Eq. (4).

$$\Delta G^{binding} = \Delta G_{mutant}^{association} - \Delta G_{parent}^{association} \qquad (4)$$

Molecular dynamics simulations. MD simulations were performed in triplicate for the aforementioned eight complexes, for a total of 24 independently ran simulations. Two 20-ns trajectories and one 15-ns trajectory were generated for each of KU70-KU80 and acetylated state of KU70-KU80-DNA, while 15-ns trajectories were generated in all triplicate simulations for KU70-KU80-DNA and acetylated state of KU70-KU80. MD simulations were performed using NAMD, version 2.9 (32). Initial protein structure files were prepared using the PSFGEN utility in VMD (33) and the CHARMM27 forcefield (34) with CMAP terms (29). The forcefield parameters for methylated lysine (35) were kindly provided by Dr. Annick Dejaegere of the Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France. The protein complex was embedded into a water box using the VMD utility SOLVATE and the TIP3P model for the water molecules. The water box dimensions were 142 Å × 150 Å × 149 Å. The system was neutralized using sodium and chloride counterions at an ionic strength of 150 mM. NAMD was used to minimize the system using 2000 steps of conjugate gradient energy minimization, followed by an MD production run for 15 or 20 ns at 1 atm pressure and 310 K. All production run simulations were performed using periodic boundary conditions and particle-mesh Ewald electrostatics for long-range electrostatic interactions with a grid point density of 1/Å. Nonbonded van der Waals interactions and short-range electrostatic interactions were calculated with an interaction cutoff of 12 Å and switching distance of 10 Å. The SHAKE algorithm was employed to fix the length of all hydrogen-containing bonds, enabling the use of 2 fs integration time steps. Coordinates were sampled every 2 ps to generate a total of 7500 or 10000 snapshots for each trajectory.

Molecular dynamics simulation analysis. Analysis of the MD trajectories was performed with in-house scripts using R programming language (36), Chimera (22), and the Bio3D library (37). Hydrogen bonds were calculated with Chimera, using hydrogen bond criteria as described (22, 38). Charge-charge interactions were calculated using a cutoff value of 5 Å between the central atoms of the amino acid charged chemical groups (N_ζ for K, C_ζ for R, C_γ for D, C_δ for E, or any of the following atoms: C_γ, $N_{\delta1}$, $C_{\varepsilon1}$, $N_{\varepsilon2}$, and $C_{\delta2}$ for H). Hydrogen bonds and charge-charge interactions were used to generate contact map occupancies. The occupancies are calculated as percentages of the presence of each interaction within the MD trajectory (percent of MD snapshots in which the interaction is present). Contact maps were generated for each of the 12 simulated complexes. To evaluate the structural perturbations introduced by methylation, manifested as loss or gain of hydrogen bonding or charge-charge interactions, the difference in occupancies between nonmethylated and methylated maps was calculated. Since simulations were performed in triplicate, difference contact maps were calculated for all six pairwise combination differences between nonmethylated and methylated contact maps. The generated difference contact maps were compiled into one map, by taking the maximum difference from all six combination differences. This protocol was used to generate maximum difference contact maps for KU70-KU80 and KU70-KU80-DNA complexes. The effects of methylation may involve structural perturbations of multiple residues surrounding the methylation site, including relay effects on residues that are not directly in hydrogen bonding or charge-charge interaction contact with the methylation site. As an example, in the case of a hydrogen bond donor surrounded by four hydrogen bond acceptors, loss or gain of a hydrogen bond may take place with any of the four available hydrogen bond acceptors depending on the trajectory (compensatory effects for loss/gain may also be operative). The utility of the contact maps is to capture all residues affected by methylation and their different structural response on methylation among the various simulations.

Principal component analysis of molecular dynamics trajectories. Refined structural superposition and principal component analysis (PCA) was performed using Bio3D (39) to discern collective and global motions during an MD simulation from small-scale fluctuations and irrelevant noise. PCA was used to compare different MD trajectories of KU70-KU80 complexes in their methylated and nonmethylated states to gain insight into conformational differences associated with methylation. PCA reduces the dimensionality of a system by projecting the data on principal components, and is based on the covariance matrix. Each element of the matrix is calculated using the equation (5).

$$C_{ij} = \langle (r_i - \langle r_j \rangle) \times (r_j - \langle r_j \rangle) \rangle \tag{5}$$

where r_i … r_{3N} denotes the Cartesian coordinates of the C_α atoms. The eigenvectors and eigenvalues are then extracted from the covariance matrix representing a set of orthogonal vectors with the highest variance in atomic coordinates. The principal components, PC1 and PC2, were chosen to be the two eigenvectors with the largest eigenvalues. The eigenvalues represent the percentage of the mean-square fluctuations along the direction of the eigenvector. Projecting data on principal components of highest eigenvalues is a way of compressing the data to a few most relevant components, and thus reducing the dimensionality of the system.

5.3 Results

The KU70-KU80 heterodimer (Figure 5.1) is comprised of multiple functional subdomains that are involved in DNA binding and the orchestration of downstream effectors to trigger DNA repair (7, 12, 13, 40). Each KU monomer can be divided into three main subdomains: N-terminal random coil (NRC), core, and C-terminal subdomains (Figure 5.1). The core comprises of KU-binding domains that intertwine to form a ring-like structure and allow the DNA to cradle inside. The C-terminal domains of both monomers contain a flexible linker region and a α-helical region (3, 17, 41, 42). In KU70 monomer, the α-helical region is known as the SAP domain and was shown to bind the KU70-KU80 core and DNA independently (3, 43).

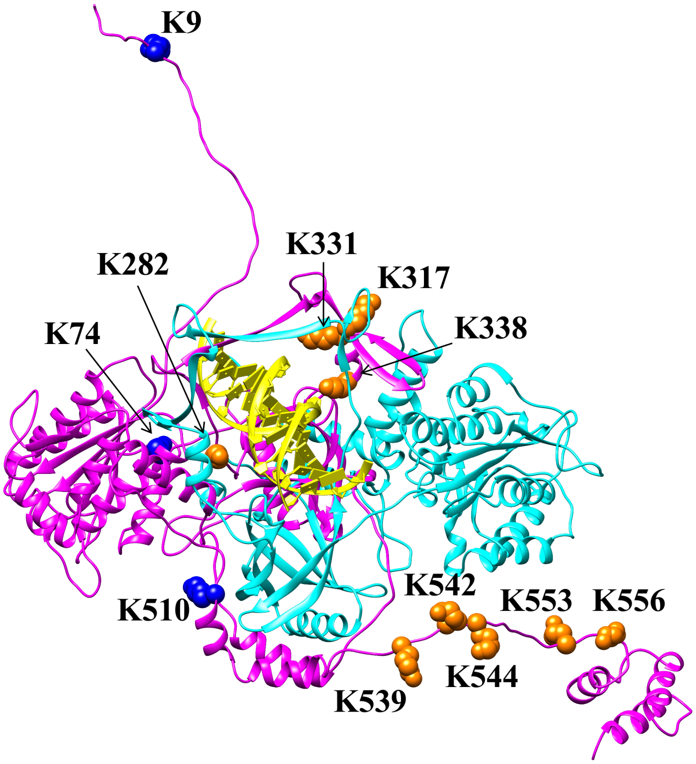


Figure 5.1 Modeled structure of the KU70 (magenta)-KU80 (cyan)-DNA (yellow) complex. The KU70 methylation sites are shown as sphere models in blue. The acetylation sites are also shown as sphere models in orange.

KU70 contains several acetylation and methylation sites that are believed to regulate its function through its dynamics which, consequently, affects the DNA repair process (16–19). Nine lysine residues have been identified as targets of (de)acetylation *in vivo*. Five of these residues (K539, K542, K544, K533, and K556) lie in the linker domain of KU70, while the remaining four (K282, K317, K331, and K338) are in the DNA binding region of the core domain (7, 10, 12, 40). Furthermore, our collaborators have identified several methylated lysine residues located on the KU70 core and NRC domain: K9, K74 and K510 (WenYong Chen, unpublished data). Here, we study the conformational changes and energetic contributions induced by all acetylated and methylated residues in the formation of the trimeric KU70-KU80-DNA complex, to delineate the role of the aforementioned PTMs in the stability of the complex and DNA repair.

Molecular analysis of the interaction between KU70's acetylation sites, and DNA. Given the high and negatively charged nature of the DNA, we focus on the role of electrostatics to assess the physicochemical basis of KU70's acetylated residues, and DNA interactions. Calculations are performed using our lab's AESOP framework to delineate the contribution of each charged residues in complex formation (25, 27). We systematically mutate positively charged residues in KU70 to alanine one at a time, mimicking the side chain charge-removal acetylation; and calculate electrostatic potential similarities and free energies of association with DNA (Figure 5.2). Two of the acetylated lysines (K282, K338) were shown to significantly diminish KU70-KU80's binding affinity to DNA when mutated to alanine due to their direct interaction with DNA (Figure 5.1). However, the remaining two acetylated lysines in the KU70 core domain (K317, K331), even though part of the DNA binding domain, do not significantly affect DNA binding to KU70-KU80. These results are in-line with experimentally determined binding affinities, where constitutive acetylation of K317 and K331 had no significant effect on DNA binding (10).

Five additional acetylation sites reside in the KU70 linker domain connecting the core and SAP domains (Figure 5.1). Electrostatic potential calculations of KU70's C-terminal tail demonstrate the presence of alternating regions of positively-charged patches, where the acetylation lysines reside, and negatively-charged patches (Figure 5.2C). In its charge-removing acetylated state, the linker domain loses its alternating patches and becomes predominantly negatively charged (Figure 5.2D). Preliminary short MD simulation of KU70's C-terminal domain (linker and SAP) has shown conformational fluctuations of the linker, attributed to the alternating positively and negatively charged patches. The observed formation and breakage of salt-bridges between the adjacent positive and negative amino acids illustrate the linker's conformational behavior, where the charge-removing acetylation of lysine residues present in the linker

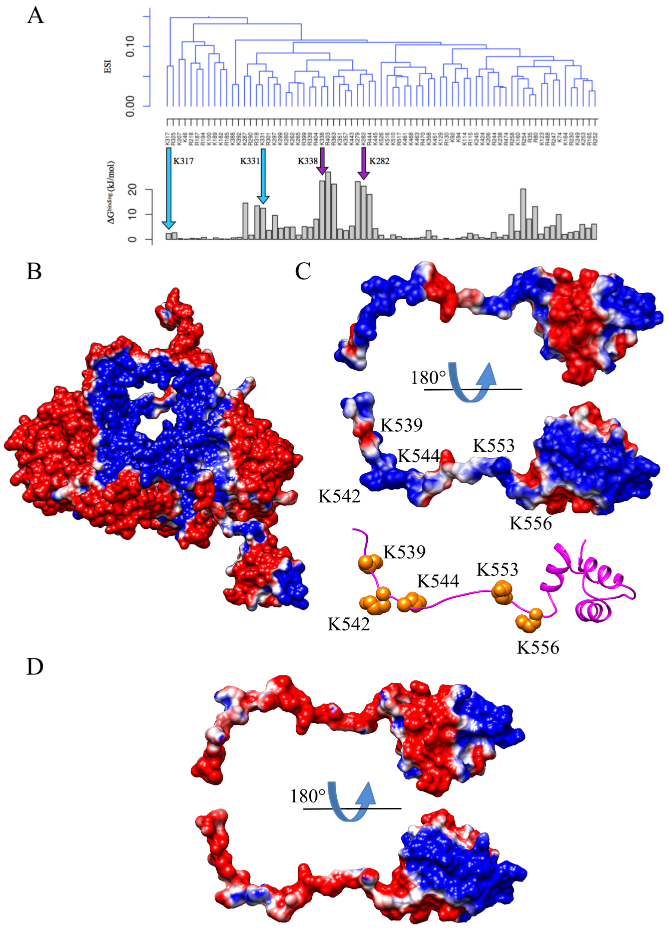


Figure 5.2 Electrostatic contribution of KU70's acetylation sites. (A) Top, Electrostatic clustering for KU70-KU80 mutants (x-axis). Bottom, binding energy results of KU70-KU80 and DNA complexes. The order of the horizontal axis is identical to top panel. Acetylated sites are indicated with purple and cyan arrows for mutations with significant and not significant effect on DNA binding. Mutated residues are positively charged residues in KU70. (B) Electrostatic potential projected onto the surface of KU70-KU80 complex. Molecular surfaces are colored based on electrostatic potential values, with a gradient from -1 kBT/e (red) to +1 kBT/e (blue). (C) Electrostatic potential projected onto the surface of linker-SAP domain of KU70. Molecular surfaces are colored based on electrostatic potential values, with a gradient from -1 kBT/e (red) to +1 kBT/e (blue). The linker-SAP domain electrostatic potentials are rotated about the horizontal axis (black line) by 180 degrees in the top and middle panel. A ribbon representation illustrating the acetylation sites is shown in the bottom panel. (D) Similar electrostatic representation as in (C) with linker-SAP domain in its acetylated state.

would result in the breaking of the salt-bridges and altering the linker's dynamical behavior that coordinate its function. Furthermore, the electrostatic potential surface of the SAP domain illustrates two oppositely charged surfaces on either side of the domain.

Both KU70 domains include several lysine residues susceptible to acetylation that directly or indirectly affect DNA binding. Acetylated residues in the DNA-binding domain neutralize the charged lysine residue, which are shown to affect binding to the DNA. Additionally, acetylation in the linker domain will alter the electrostatic profile and dynamical behavior of such domain which will in turn affect the SAP function in binding to the KU70-KU80 complex or the DNA (3).

Impact of lysine methylation on KU70 structure in its acetylated and nonacetylated states in the presence or absence of DNA. To gain structural insight into the roles of lysine methylation, we employed comparative computational modeling of KU70 with: K9, K74 and K510 in their methylated and nonmethylated states, K282, K317, K331, K338, K539, K542, K544, K533, and K556 in their acetylated and nonacetylated states, and in the presence or absence of DNA. We used the structure of the hybrid crystallographic and modeled KU70-KU80-DNA complex, described in *Materials and Methods*. The structure of the complex and topology of the methylated/acetylated lysines is shown in Figure 5.1. This structure includes KU70's core domain, N-terminal random coil (NRC), linker-SAP domain, and a 14-bp DNA duplex, and was used as the starting point for MD. We evaluated the effect of KU70 methylation by performing a series of MD simulations of KU70-KU80 in its different PTM conditions and with the presence or absence of DNA as listed in Figure 5.4C. Overall, the MD simulations showed a rather robust KU70-KU80 core, with the exception of the microenvironment of K74 (Figure 5.3A) and the significant flexibility of the long random coils of the NRC and the linker (Figure B.1). Among the three methylated lysines, K9 is part of the highly flexible NRC and the data did not detect any consistent differences between the nonmethylated and methylated states of the complex. On the other hand, K74 and K510 are part of the core domain and both lysines may induce local and long-range conformational changes in the KU70-KU80 complex upon methylation.

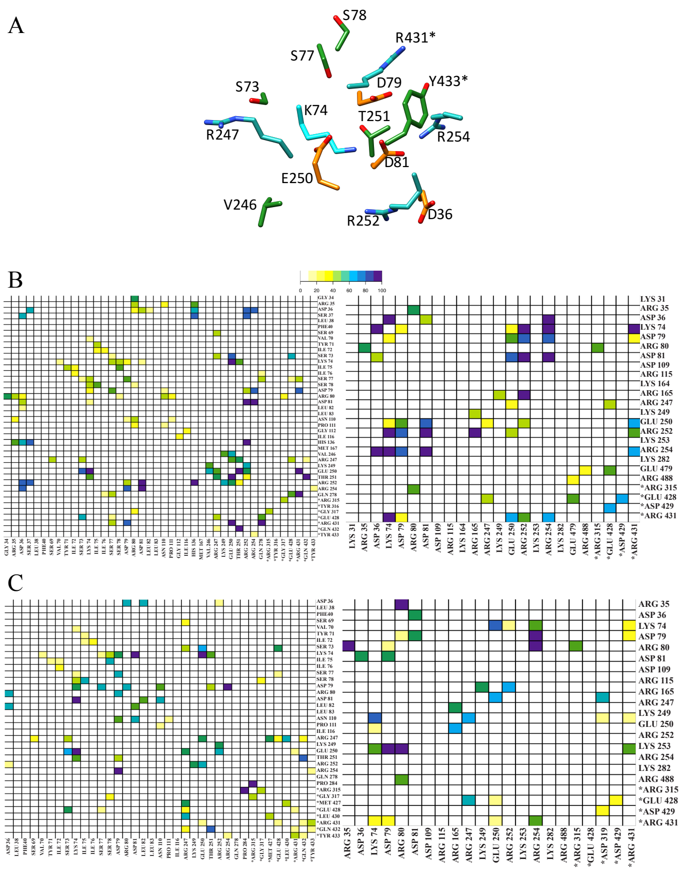
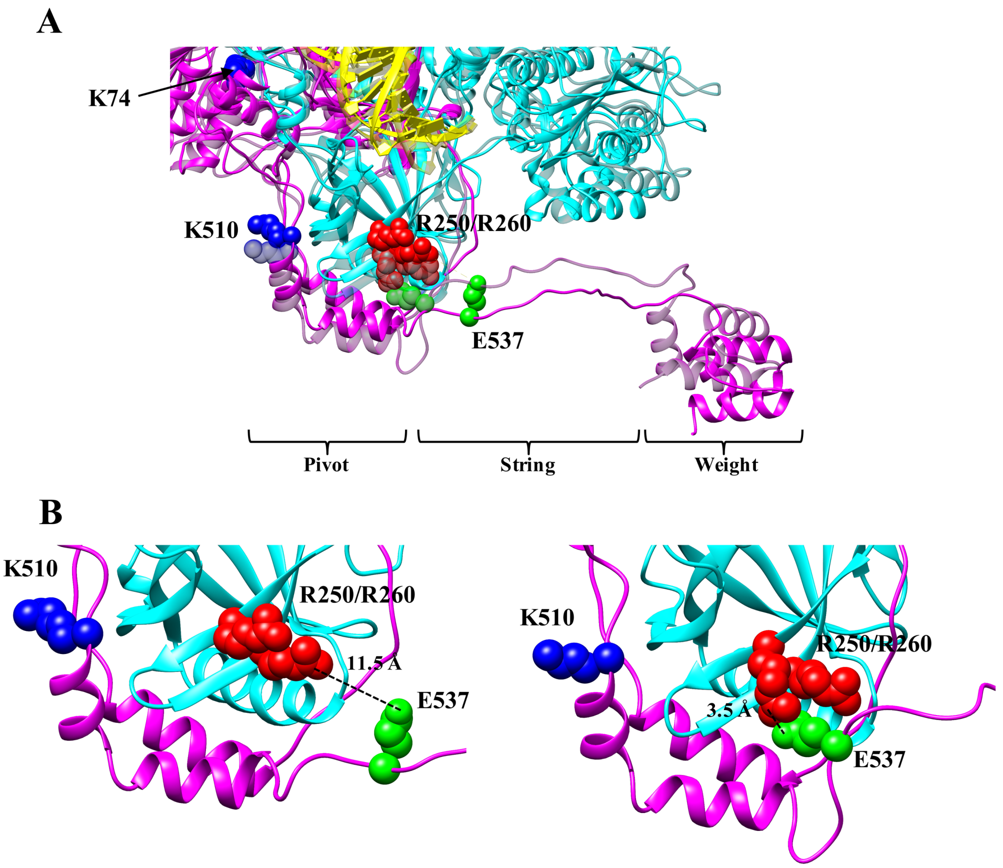


Figure 5.3 Effect of K74 lysine methylation on its microenvironment. (A) Molecular graphics representation of the K74 (cyan) microenvironment. The illustrated residues show more than 30% differences in all hydrogen bond and charge-charge interaction occupancy contact maps. For clarity, only amino acid side chains are shown, even though the calculations included both backbone and side chain hydrogen bonds. Residues (other than K74) are colored according to their charge properties: orange, negative; light blue, positive; and green, the rest. In addition, oxygen and nitrogen atoms are colored in red and blue, respectively. Asterisks denote KU80 residues. (B, C) Hydrogen bond and charge-charge (Coulombic) interaction occupancy changes in the vicinity of K74 between methylated and nonmethylated states of the complex in the MD simulations in its nonacetylated state. Methylated/nonmethylated lysines are K9, K74, and K510. The vicinity of K74 includes residues within 8 Å from the N^{ζ}-atom of K74 in the modeled KU70-KU80-DNA structure. Hydrogen bonds (left column) and charge-charge interactions (right column) were calculated between the residues selected in the vicinity of K74 for KU70-KU80 (B) and KU70-KU80-DNA (C) complexes. All possible hydrogen bonds between selected residues were calculated with Chimera. The cutoff value for a charge-charge interaction was 5 Å between the central heavy atoms of the amino acid charged chemical groups, and includes both favorable interactions between opposite charges (salt bridges) and unfavorable interactions between like charges. Occupancies were calculated as percentage of MD trajectory frames in which the interaction is present. By calculating the difference of occupancies between triplicate simulations of methylated and nonmethylated states of the KU70-KU80, the maximum observed change was extracted and compiled in one contact map (see Methods for more detailed explanation). The color code represents occupancy differences in the range of 0-100%, as indicated in the legend. Contacts maps pertaining to the acetylated states of the simulations are shown in Figure B.2.

K74 is located in a packed environment and its methylation introduced reorganization of the local microenvironment (Figure 5.1, and 5.3A). The differences between the nonmethylated and methylated lysines can be rationalized by the physicochemical properties of its side chain. The methyl groups replace terminal amine hydrogens, and unlike the case of acetylation, the amines retain their positive charge. Therefore, the side chain loses hydrogen bonding capabilities for each replaced hydrogen. Additional steric and hydrophobic effects are also operative because of the bulkier and hydrophobic methyl groups compared to hydrogens. Thus, we evaluated percent occupancy contact maps for hydrogen bonds and charge-charge interactions throughout the MD trajectories (Figure 5.3B, 5.3C, and Figure B.2). The microenvironment of K74 reveals 40% or more difference in occupancies of hydrogen bonds between K74 and at least two of the four nearby residues with hydrogen bonding capabilities (D79, D81, E250, and T251) (Figure 5.3B, 5.3C, and Figure B.2). The K74 microenvironment also reveals 40% or more difference in occupancies of favorable or unfavorable Coulombic interactions (defined as a distance < 5 Å between oppositely charged groups) between K74 and six nearby charged residues (D36, D79, E250, R252, and R254 of KU70, and R431 of KU80). Rearrangements around K74 propagate to nearby E250, which showed behavior similar to K74 in hydrogen bonding (with S73, V246, R247, T251, and R252 of KU70, and R431, R433 of KU80) and charge-charge interactions (with D79, D81, R247, and R252 of KU70 and R431 of KU80), with occupancy differences of 40% or more (Figure 5.3B, 5.3C, and Figure B2). Such local structural rearrangements may participate in the initiation of long-range motions that bring the protein to altered conformational and functional states as further detailed below.



C

	non-Acetylated			Acetylated		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
Ku70-Ku80 non-Methylated	93%	58%	98%	35%	95%	58%
Ku70-Ku80 Methylated	16%	0%	25%	0%	0%	1%
Ku70-Ku80 DNA non-Methylated	90%	10%	35%	12%	97%	96%
Ku70-Ku80 DNA Methylated	0%	31%	0%	1%	35%	0%

Figure 5.4 Effect of KU70 lysine methylation on the intra-molecular “pendulum.” **(A)** Superposition of representative MD structures of the KU70-KU80-DNA complexes in their methylated and nonmethylated (transparent model) states, focusing on the helical hinge domain, linker, and SAP domain. The formed clamp is shown in the nonmethylated complex (transparent), whereas the broken clamp is shown in the methylated complex. **(B)** Zoom-in illustrating the relative topology of K510 and the clamp residues in characteristic MD structures of methylated (left, broken clamp) and nonmethylated (right, formed clamp) states. (**C**) Intra-molecular salt bridge occupancies between the clamp residues E527 (KU70) and R250/R260 (KU80) from MD trajectories of methylated, nonmethylated, acetylated, and nonacetylated states. Occupancies were calculated from triplicate simulations of the complexes shown in the left column.

The KU70 C-terminal SAP domain also has a robust structure, but because it is attached to the linker it demonstrates significant mobility (Figure B.1). The last helical domain of the KU70 core (a helical hinge domain, residues 511-529), the linker (residues 530-560), and SAP domain (residues 561-609) resemble the pivot, string, and weight of a pendulum, respectively (Figure 5.4A). Monomethylation of K510 results in loss of one hydrogen atom (and the associated hydrogen-bonding capability of the lost methyl), while retaining the positive charge. Analysis of hydrogen bonding and charge-charge interaction occupancies in MD trajectories displayed no significant differences in the local moiety of methylated and nonmethylated K510, mainly because of the non-packed and highly solvated environment. However, with detailed analysis of the MD trajectories, we identified a bifurcated salt bridge between E537 (KU70) and R250 and R260 (KU80), located within the linker domain (Figure 5.4A, and 5.4B). These salt bridge interactions are persistent throughout our triplicate runs of the MD simulations with nonmethylated K74/K510, and absent in the MD simulations with methylated K74/K510 regardless of the acetylation state of KU70 or the presence of the DNA. Figure 5.4C shows high salt bridge occupancies of 93%, 58%, and 98% in KU70-KU80 with nonmethylated lysines, in contrast to low salt bridge occupancies of 16%, 0%, and 25% when lysines are methylated. The E537 (KU70)-R250/R260 (KU80) salt bridge acts as a clamp that restrains the pendulum motion and controls the mobility, and perhaps the function, of the linker-SAP domain. The clamp is similarly present in the KU70-KU80-DNA MD simulation and acetylated forms of the KU70-KU80 complex (Figure 5.4C). Therefore, it may be argued that methylation acts as a trigger that mediates the motion and function of the “structural pendulum”.

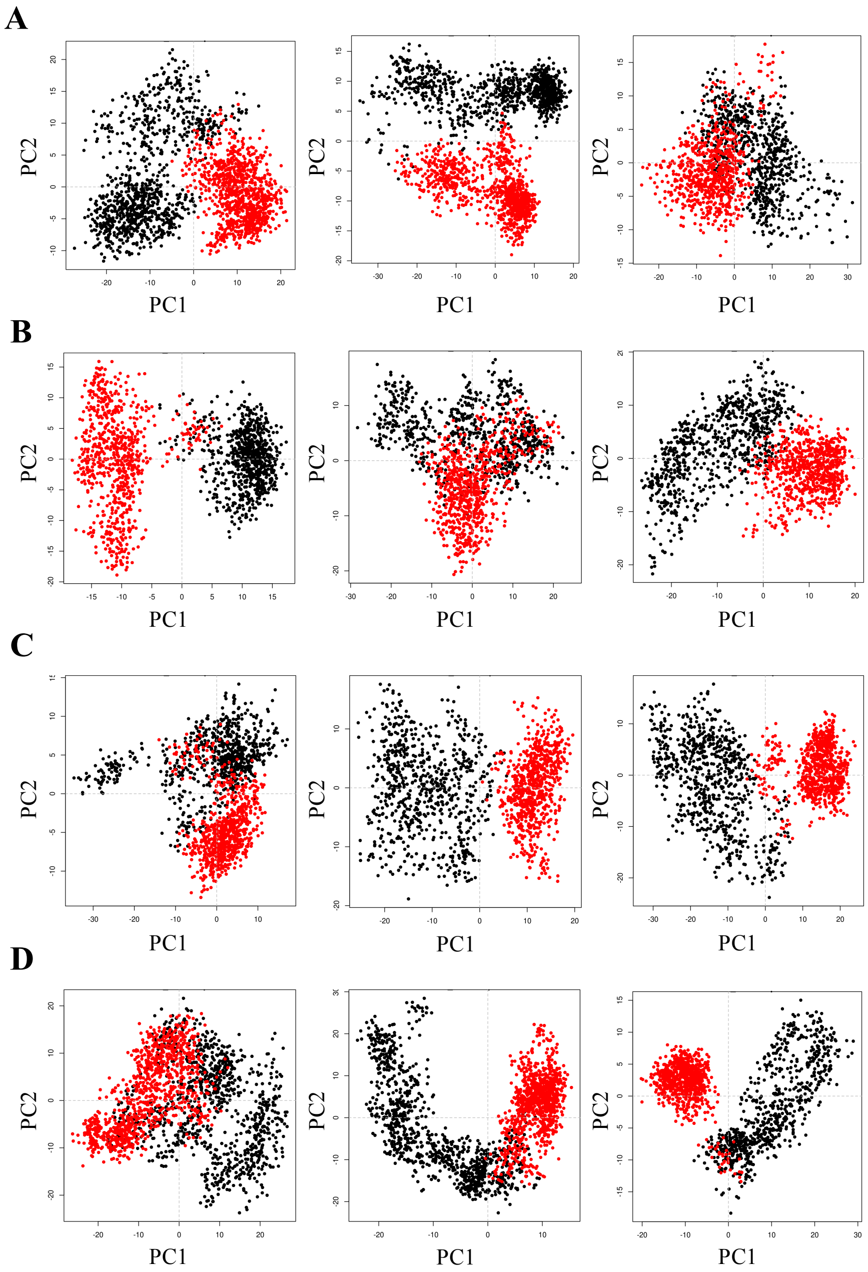


Figure 5.5 Principal component analysis of the helical hinge motion (K510-E537). Comparison of conformational sampling of the helical hinge domain (residues 510 to 537) using the first two principal components (PC1, PC2) in the following structures with methylated (black) and nonmethylated (red) states: (A) nonacetylated KU70-KU80, (B) nonacetylated KU70-KU80-DNA, (C) acetylated KU70-KU80, (D) acetylated KU70-KU80-DNA. Each panel illustrates comparison of the triplicate simulations.

We hypothesized that K74 and K510 may facilitate a long-range pendulum-like motion initiating at the helical hinge domain and propagating to the linker-SAP domain, when releasing the E537 (KU70)-R250/R260 (KU80) clamp upon methylation. We used principal component analysis (PCA) to characterize the motion responsible for propagating the methylation effect from K74 and/or K510 to the clamp residues located at the beginning of the linker, through the helical hinge domain (37, 39). PCA of the helical hinge domain minimizes the noise of local fluctuations at the atomic level and intensifies the global motions of such domains. Our analysis illustrates distinct conformational changes when comparing the methylated and nonmethylated structure simulations, evidenced by examination of the first two principal components (Figure 5.5). The first two principal components account for about 70% of the motion showing that the helical hinge domain undergoes distinct global motions in the methylated compared to the nonmethylated structures. PCA of all of our simulations shows that methylation of KU70 consistently induces different global motions of the helical hinge region regardless of its acetylation state and the presence or absence of the DNA (Figure 5.5).

5.4 Discussion

Three lysine residues, K9, K74, and K510, were identified by our collaborators, to undergo mono- and di-methylation in the NRC and core domains of KU70. Additional nine lysine residues were previously shown to be acetylated (7, 10, 12, 40). In this study, we identify the role of acetylation and methylation on KU70's dynamics, which consequently affects its binding function to its different binding partners, such as DNA, SIRT1, and LSD1. KU70's acetylation is shown to alter the DNA binding affinity by neutralizing the charge of four lysine residues that reside within the DNA-binding domain. Charge-removing acetylation of the remaining five lysine residues alters the electrostatic profile of the linker domain. Charge-removal contributes to a more negatively charged linker and disrupts the alternating positively and negatively charged patches in the linker, which will alter the linker's dynamics and the SAP domain's function of binding the KU70-KU80 complex or DNA (3).

Accordingly, we have also shown the effect of methylation on the local and global structures of KU70 protein using MD simulations. We propose that alterations in the local environment of methylated lysines propagate through correlated motions, inducing distal conformational changes in the KU70 linker-SAP domain. We have identified the E537 (KU70)-R250/R260 (KU80) salt bridge that acts as a clamp and is responsible for restricting the "pendulum-like" conformational space spanned by the linker-SAP domain. The linker domain starts with a multi-proline loop, PPDYNPE, ending with the clamp residue, E537. The presence of the rigid proline residues restricts the motion of this loop, and with the formation of the clamp in the nonmethylated structures the loop is locked as part of the helical hinge domain. Thus, the motion of the linker is further restricted in the presence of the clamp. Lysine methylation triggers long-range motions that are coupled to releasing the clamp and increasing the mobility (and available conformational space) of the linker-SAP domain. Crystal structures of nonmethylated KU70-KU80 (PDB code: 1JEQ) include the formation of the clamp in the absence of the DNA, which is associated with binding of the SAP domain to the KU70-KU80 core, while lost in the presence of DNA (3). This result confirms the role of the clamp in mediating the linker-SAP domain "pendulum" motion, hypothesized to affect its function (3).

The functional consequences of the conformational switch triggered by lysine methylation and acetylation in the DNA-binding and linker-SAP domains may be to facilitate interaction or inhibition of the KU70-KU80 complex moving along DNA to facilitate repair. This may be possible through regulating the DNA-binding affinity of the DNA-binding domain through acetylation; and increasing the motional amplitudes of the helical hinge domain and linker-SAP domain structure, known to bind DNA (3), upon rupture of the E537 (KU70)-R250/R260 (KU80) salt bridge. The linker-SAP domain dynamics is also altered through acetylation, neutralizing positively charged linker lysines, contributing to a more negatively charged linker and thus, further altering the linker-SAP domain dynamics.

Interplay between KU70 methylation and acetylation may be present, and regulated through different acetylation and methylation enzymes. Using lentiviral shRNA to knockdown LSD1 and SIRT1, our collaborators at City of Hope, Dr. WenYong Chen and colleagues, have identified that lysine deacetylase SIRT1 and demethylase LSD1 competitively bind to KU70 in response to chemotherapeutic agents and DNA damage, and consequently affect a cell's ability to repair broken DNA and acquire genetic mutations. Both SIRT1 and LSD1 strongly bind to the KU70 core domain, providing a molecular basis for such competition. Both effectors were shown to be regulated by KU70's NRC, linker, and SAP domains. Co-immunoprecipitation assays with KU70 domain truncations were used to determine the role of KU70's domains in regulating LSD1 and SIRT1 interactions. The SAP domain is a strong repressor motif for LSD1 interaction, while the NRC is a positive regulatory domain. On the other hand, the NRC is a strong repressor for SIRT1, the linker-SAP domain is a positive regulator for SIRT1 interaction and antagonized the repression from the NRC (WenYong Chen, data not shown). Given the role of the SAP domain in binding the KU70-KU80 core in addition to binding the DNA, the SAP domain and LSD1 may competitively bind to the KU70-KU80 core domain. Thus, since LSD1 and SIRT1 competitively bind KU70-KU80, the presence of the SAP domain will favor SIRT1 binding to KU70-KU80 by repressing LSD1 binding.

All three methylation sites, K9, K74 and K510, were mutated experimentally to arginines by our collaborators to mimic constitutive lysine nonmethylation. Co-immunoprecipitation of KU70's mutants demonstrate that arginine mutations did not affect KU70's interaction with LSD1 (lysine demethylase) but reduced its interaction with SIRT1 (lysine deacetylase) (WenYong Chen, data not shown). These results suggest the role of K9, K74, and K510 methylation in modulated the linker-SAP domain function in order to mediate SIRT1's interaction with KU70. The clamp residues may work as a molecular switch that regulates the SAP domain's interaction with KU70-KU80 which would mediate SIRT1 and LSD1 interaction with KU70 in order to regulate its acetylation and methylation state. Additional neutralizing and charge-reversing mutations of lysine residues on the SAP domain (Figure B.3), suggested by our electrostatic calculation to minimize the positively charged side of the SAP domain (K575, K595, and K596 to E or Q), reduced SIRT1 interaction but increased LSD1 interaction with KU70, while mutating the negatively charged side (E561, E583, D609 to K) increased SIRT1 interaction with KU70. This further demonstrates the role of the SAP domain in regulating interaction of LSD1 and SIRT1 with KU70. These mutations simulate the above SAP-truncation experiments by increasing LSD1 interaction.

Our results, along with those of our collaborators, shed novel insight into the role of methylation and acetylation in mediating KU70's interactions with LSD1 and SIRT1, which act as a feedback loop, regulating the methylation and acetylation levels in KU70. The acetylation and methylation pattern are hypothesized to act as a bar code determining KU70's functional domains' dynamics. This suggests that lysine methylation and acetylation may have a broader role in modulating DNA repair machineries for genome maintenance and cancer drug resistance.

5.5 Acknowledgements

We thank Dr. Annick Dejaegere (Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France) for providing the CHARMM forcefield parameters for methylated lysine. This study was supported by the City of Hope – UC Riverside Biomedical Research Initiative.

5.6 References

1. Lieber, M.R. 2010. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway. Annu. Rev. Biochem. 79: 181–211.

2. Khanna, K.K., and S.P. Jackson. 2001. DNA double-strand breaks: signaling, repair and the cancer connection. Nat. Genet. 27: 247–254.

3. Walker, J.R., R.A. Corpina, and J. Goldberg. 2001. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature. 412: 607–614.

4. Vignard, J., G. Mirey, and B. Salles. 2013. Ionizing-radiation induced DNA double-strand breaks: A direct and indirect lighting up. Radiother. Oncol. 108: 362–369.

5. Nussenzweig, A., C. Chen, V. da Costa Soares, M. Sanchez, K. Sokol, M.C. Nussenzweig, and G.C. Li. 1996. Requirement for Ku80 in growth and immunoglobulin V(D)J recombination. Nature. 382: 551–555.

6. Ouyang, H., A. Nussenzweig, A. Kurimasa, V. da C. Soares, X. Li, C. Cordon-Cardo, W. Li, N. Cheong, M. Nussenzweig, G. Iliakis, D.J. Chen, and G.C. Li. 1997. Ku70 Is Required for DNA Repair but Not for T Cell Antigen Receptor Gene Recombination In Vivo. J. Exp. Med. 186: 921–929.

7. Cohen, H.Y. 2004. Calorie Restriction Promotes Mammalian Cell Survival by Inducing the SIRT1 Deacetylase. Science. 305: 390–392.

8. Bennetzen, M., D. Larsen, C. Dinant, S. Watanabe, J. Bartek, J. Lukas, and J.S. Andersen. 2013. Acetylation dynamics of human nuclear proteins during the ionizing radiation-induced DNA damage response. Cell Cycle. 12: 1688–1695.

9. Polo, S.E., and S.P. Jackson. 2011. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. Genes Dev. 25: 409–433.

10. Chen, C.-S., Y.-C. Wang, H.-C. Yang, P.-H. Huang, S.K. Kulp, C.-C. Yang, Y.-S. Lu, S. Matsuyama, C.-Y. Chen, and C.-S. Chen. 2007. Histone Deacetylase Inhibitors Sensitize Prostate Cancer Cells to Agents that Produce DNA Double-Strand Breaks by Targeting Ku70 Acetylation. Cancer Res. 67: 5318–5327.

11. Wang, Z., H. Yuan, M. Roth, J.M. Stark, R. Bhatia, and W.Y. Chen. 2013. SIRT1 deacetylase promotes acquisition of genetic mutations for drug resistance in CML cells. Oncogene. 32: 589–598.

12. Subramanian, C., M. Hada, A.W. Opipari, V.P. Castle, and R.P.S. Kwok. 2013. CREB-Binding Protein Regulates Ku70 Acetylation in Response to Ionization Radiation in Neuroblastoma. Am. Assoc. Cancer Res. 11: 173–181.

13. Yuan, H., Z. Wang, L. Li, H. Zhang, H. Modi, D. Horne, R. Bhatia, and W. Chen. 2012. Activation of stress response gene SIRT1 by BCR-ABL promotes leukemogenesis. Blood. 119: 1904–1914.

14. Oberdoerffer, P., S. Michan, M. McVay, R. Mostoslavsky, J. Vann, S.-K. Park, A. Hartlerode, J. Stegmuller, A. Hafner, P. Loerch, S.M. Wright, K.D. Mills, A. Bonni, B.A. Yankner, R. Scully, T.A. Prolla, F.W. Alt, and D.A. Sinclair. 2008. SIRT1 Redistribution on Chromatin Promotes Genomic Stability but Alters Gene Expression during Aging. Cell. 135: 907–918.

15. Moore, K.E., S.M. Carlson, N.D. Camp, P. Cheung, R.G. James, K.F. Chua, A. Wolf-Yadlin, and O. Gozani. 2013. A General Molecular Affinity Strategy for Global Detection and Proteomic Analysis of Lysine Methylation. Mol. Cell. 50: 444–456.

16. Petrov, D., C. Margreitter, M. Grandits, C. Oostenbrink, and B. Zagrovic. 2013. A Systematic Framework for Molecular Dynamics Simulations of Protein Post-Translational Modifications. PLOS Comput Biol. 9: e1003154.

17. Hu, S., and F.A. Cucinotta. 2011. Computational studies on full-length Ku70 with DNA duplexes: base interactions and a helical path. J. Mol. Model. 18: 1935–1949.

18. Ulucan, O., O. Keskin, B. Erman, and A. Gursoy. 2011. A Comparative Molecular Dynamics Study of Methylation State Specificity of JMJD2A. PLOS ONE. 6: e24664.

19. Yang, S.-Y., X.-L. Yang, L.-F. Yao, H.-B. Wang, and C.-K. Sun. 2011. Effect of CpG methylation on DNA binding protein: Molecular dynamics simulations of the homeodomain PITX2 bound to the methylated DNA. J. Mol. Graph. Model. 29: 920–927.

20. Šali, A., and T.L. Blundell. 1993. Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol. 234: 779–815.

21. Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23: 2947–2948.

22. Pettersen, E.F., T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. 2004. UCSF Chimera-A visualization system for exploratory research and analysis. J. Comput. Chem. 25: 1605–1612.

23. Gorham, R.D., C.A. Kieslich, and D. Morikis. 2010. Electrostatic Clustering and Free Energy Calculations Provide a Foundation for Protein Design and Optimization. Ann. Biomed. Eng. 39: 1252–1263.

24. Gorham, R.D., C.A. Kieslich, and D. Morikis. 2011. Complement Inhibition by Staphylococcus aureus. Cell. Mol. Bioeng. 5: 32–43.

25. Gorham, R.D., C.A. Kieslich, A. Nichols, N.U. Sausman, M. Foronda, and D. Morikis. 2011. An evaluation of Poisson–Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. Biopolymers. 95: 746–754.

26. Kieslich, C.A., R.D. Gorham Jr., and D. Morikis. 2011. Is the rigid-body assumption reasonable?: Insights into the effects of dynamics on the electrostatic analysis of barnase–barstar. J. Non-Cryst. Solids. 357: 707–716.

27. Kieslich, C.A., D. Morikis, J. Yang, and D. Gunopulos. 2011. Automated computational framework for the analysis of electrostatic similarities of proteins. Biotechnol. Prog. 27: 316–325.

28. Dolinsky, T.J., J.E. Nielsen, J.A. McCammon, and N.A. Baker. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 32: W665–W667.

29. Mackerell, A.D., M. Feig, and C.L. Brooks. 2004. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J. Comput. Chem. 25: 1400–1415.

30. Baker, N.A., D. Sept, S. Joseph, M.J. Holst, and J.A. McCammon. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. 98: 10037–10041.

31. Gorham Jr., R.D., W. Rodriguez, and D. Morikis. 2014. Molecular Analysis of the Interaction between Staphylococcal Virulence Factor Sbi-IV and Complement C3d. Biophys. J. 106: 1164–1173.

32. Phillips, J.C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, and K. Schulten. 2005. Scalable molecular dynamics with NAMD. J. Comput. Chem. 26: 1781–1802.

33. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: Visual molecular dynamics. J. Mol. Graph. 14: 33–38.

34. MacKerell, A.D., D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. J. Phys. Chem. B. 102: 3586–3616.

35. Grauffel, C., R.H. Stote, and A. Dejaegere. 2010. Force field parameters for the simulation of modified histone tails. J. Comput. Chem. : NA–NA.

36. R. Core Team. 2014. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

37. Grant, B.J., A.P.C. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S.D. Caves. 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 22: 2695–2696.

38. Mills, J.E.J., and P.M. Dean. 1996. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. J. Comput. Aided Mol. Des. 10: 607–622.

39. Skjærven, L., X.-Q. Yao, G. Scarabelli, and B.J. Grant. 2014. Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics. 15.

40. Cohen, H.Y., S. Lavu, K.J. Bitterman, B. Hekking, T.A. Imahiyerobo, C. Miller, R. Frye, H. Ploegh, B.M. Kessler, and D.A. Sinclair. 2004. Acetylation of the C terminus of Ku70 by CBP and PCAF controls Bax-mediated apoptosis. Mol. Cell. 13: 627–638.

41. Zhang, Z., L. Zhu, D. Lin, F. Chen, D.J. Chen, and Y. Chen. 2001. The Three-dimensional Structure of the C-terminal DNA-binding Domain of Human Ku70. J. Biol. Chem. 276: 38231–38236.

42. Zhang, Z., W. Hu, L. Cano, T.D. Lee, D.J. Chen, and Y. Chen. 2004. Solution Structure of the C-Terminal Domain of Ku80 Suggests Important Sites for Protein-Protein Interactions. Structure. 12: 495–502.

43. Rivera-Calzada, A., L. Spagnolo, L.H. Pearl, and O. Llorca. 2007. Structural model of full-length human Ku70–Ku80 heterodimer and its recognition of DNA and DNA-PKcs. EMBO Rep. 8: 56–62.

APPENDIX A: SUPPORTING MATERIAL FOR CHAPTER 2

A.1 Discussion of Previous Mutagenesis Data

Experimental mutagenesis data obtained by Ott et al. present the differential effect of Q203$^{5.42}$ mutation to alanine, where the mutation impaired activation by 20 and 40 fold in response to CCL19 and CCL21 respectively. Even though Q203$^{5.42}$ is not involved in the characterized molecular switches, the mutagenesis data complies with our findings where Q203$^{5.42}$ is part of the residues carrying the global helical motion between TM4 and TM5 critical to the propagation of helical motions in the receptor (1). Additionally, tri-tyrosine residue Y112$^{3.32}$ mutation has shown a slight difference in activation measured by GTP-γS binding assays, in response to CCL19 and CCL21 (Figure 6 in (1)). This difference deemed insignificant by Ott et al. challenges the importance of Y112$^{3.32}$ supported by our findings (1). However, the fact that Y112$^{3.32}$ is involved in non-specific hydrogen bonding with Q252$^{6.48}$ and Y255$^{6.51}$ to stabilize both inactive and active conformations of the tri-tyrosine switch is compatible with the experimental observation that hydrogen bonding and the π-π interaction-abolishing mutation (Y112$^{3.32}$ to alanine) had nullifying effects on the activation rate of CCR7. Another mutation carried by Ott et al. shows that N281$^{7.32}$ mutation to alanine decrease the ability of CCL19 and CCL21 to activate CCR7 (1). Changes due to N281$^{7.32}$ mutation are minor (3.8 and 5.5 fold for CCL19 and CCL21 respectively) because neighboring hydrogen-bonding residues in the ligand compensates for the loss of hydrogen bonds carried by N281$^{7.32}$ in our simulations. The depicted mechanism and molecular switches provide a rationale behind previously published mutagenesis data and insight for future mutation that take advantage of the biased nature of CCL19 and CCL21.

A.2 Supplementary Figures

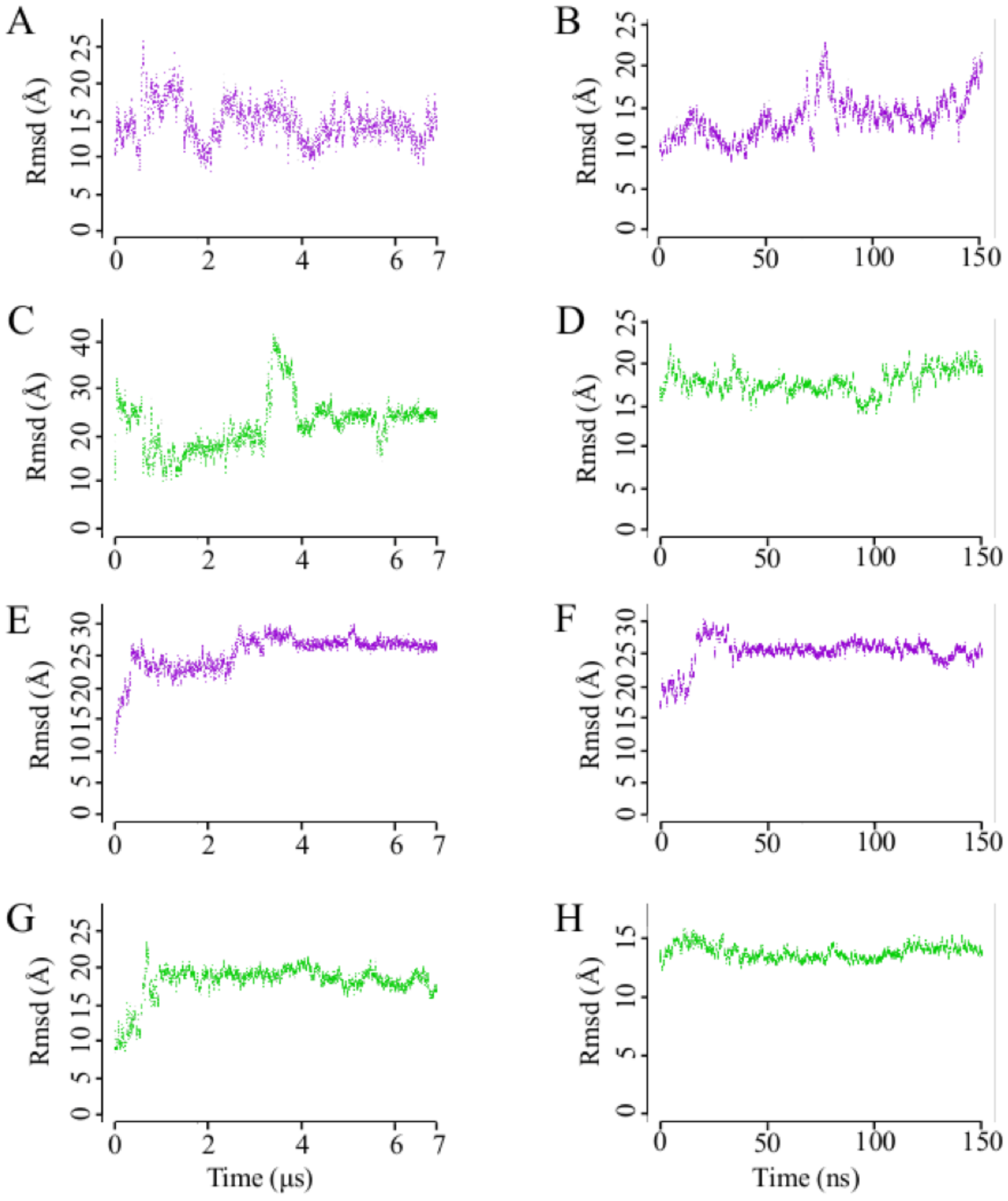


Figure A.1 Root mean square deviation (rmsd) of CCR7-NTD (A-D) and CCR7-CTD (E-H) calculated in both CCL19-bound (purple) and CCL21-bound (green) cMD (left column) and aMD (right column) simulations. Rmsds are calculated using Bio3D based on atomic coordinates.

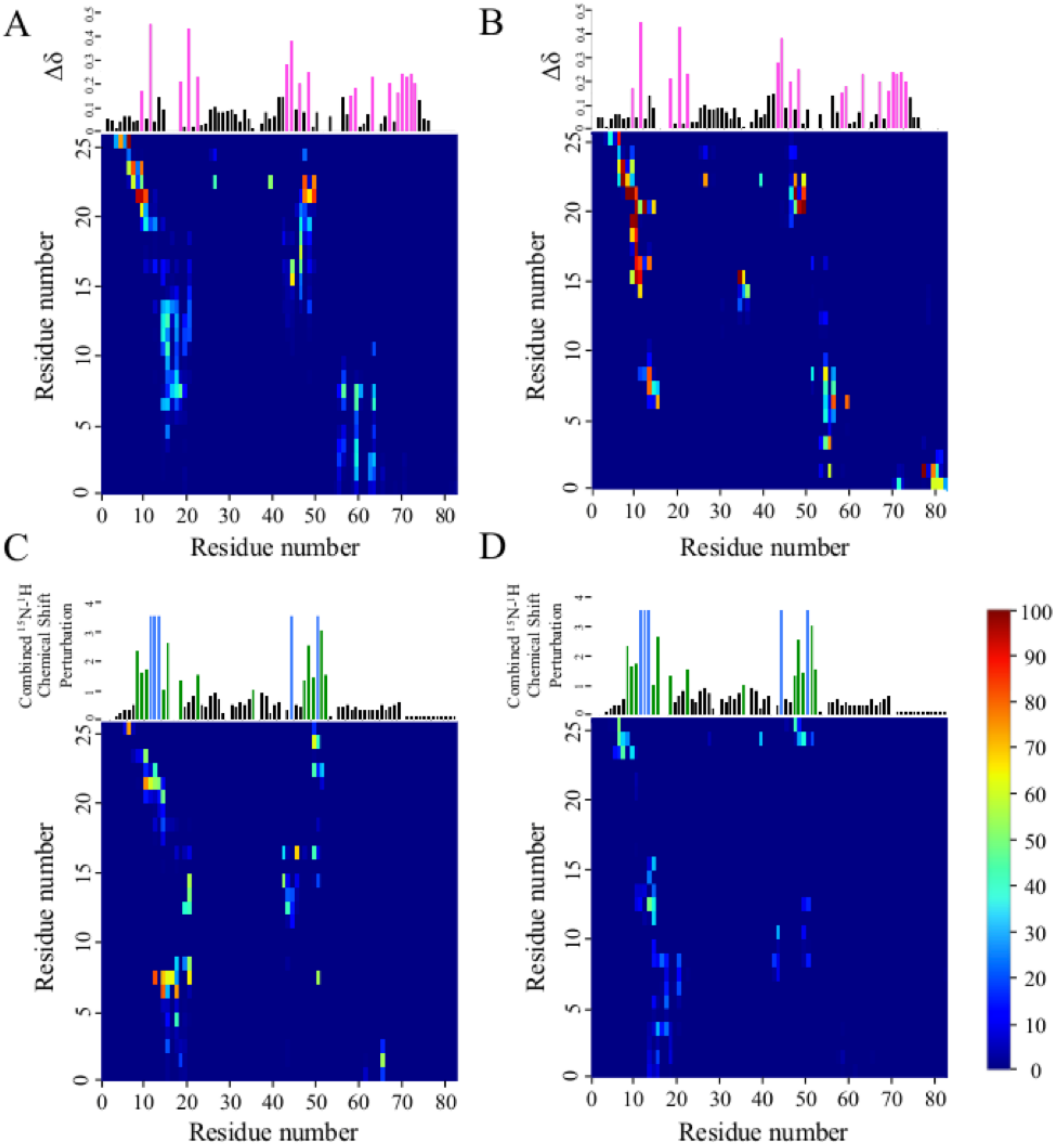


Figure A.2 Contact maps between CCR7-NTD and each of its ligands are in close agreement with published NMR data. Shown are the maps of color-coded percent occupancies of inter-residue contacts between CCR7-NTD and each of its ligands CCL19 (A: cMD simulation, B: aMD simulation) and CCL21 (C: cMD simulation, D: aMD simulation). Ligand and CCR7-NTD residues are indicated on the x- and y-axis, respectively. Published NMR chemical shift perturbations for each of the ligand's interaction with CCR7-NTD are reported as a bar plot above each contact map versus the ligand's residue numbering on the x-axis (2, 3). Residues with significant changes in chemical shifts are colored in magenta for CCL19 (A, B) and green and blue for CCL21 (C, D). Green indicates significant chemical shift perturbations (above 1.0 ppm), while blue indicates perturbations whose signal broadened beyond detection.

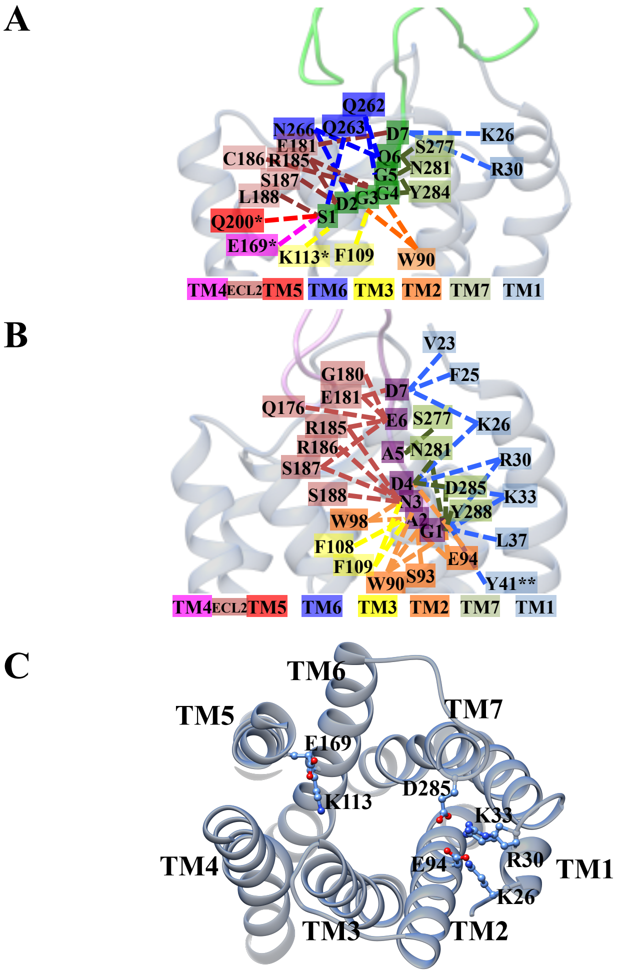


Figure A.3 CCL19 and CCL21 adopt different conformations in CCR7's binding pocket. Polar and non-polar interactions involved between CCL21 (green, panel A) and CCL19 (purple, panel B) N-terminal domains (ligand-NTD heptapeptides) and CCR7. CCL21 and CCL19 are shown in green and purple tubular rendering, respectively, and CCR7 is shown in ribbon rendering. Ligand-NTD residues are labeled with a one-letter amino acid code and sequence number, and are colored in green for CCL21 and purple for CCL19. Residues in contact with the ligand-NTDs are labeled, and are colored according to the TM and extracellular loop domain they belong to, with the same color code in both panels (A, B). Electrostatic patches in CCR7 (panel C) guide the ligand-NTD interactions within the binding pocket of the receptor. CCL19 interacts with TM7, TM1, TM2, TM3, and ECL2; while CCL21 interacts with all seven TM helices. Residues depicted are within 5 Å of the ligand-NTD with at least 50% occupancy in the equilibrated segment of the cMD simulations (the equilibrated segments are from 3 to 7 µs, and from 4.6 to 7 µs for CCL21 and CCL19, respectively, according to ligand rmsd time series in Figure A.5). Residues are marked with asterisks because they form important contact, despite the fact that their occupancies are below 50%. *$K113^{3.33}$, *$E169^{4.60}$, and *$Q200^{5.39}$ have 45%, 46%, and 27% occupancies respectively and form important interaction in stabilizing CCL21's pose. **$Y41^{1.39}$ has 18% occupancy within the equilibrated segment and is part of the CCL19-induced allosteric events (Figure 2.3). (C) CCR7's binding pocket contains two separate electrostatic interaction patches. The molecular graphic illustrates all charged residues within CCR7's binding pocket as stick models. One patch is formed between TM3 and TM4 (CCL21 contacts), and another is formed between TM1, TM2, and TM7 (CCL19 contacts).

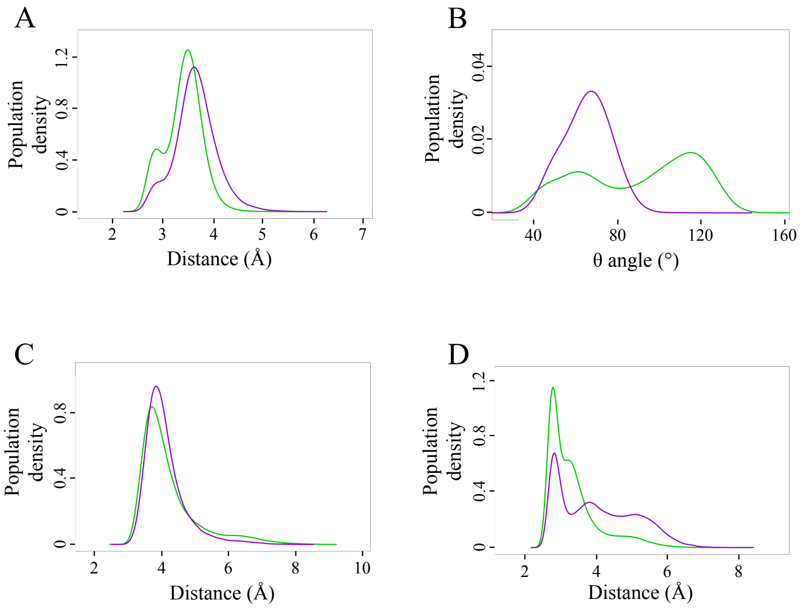


Figure A.4 Molecular switches in the CCR7 transmembrane domain adopt ligand-specific conformations within aMD simulations. All four panels, A, B, C, and D, show the same measures as in panels B, C, D, and E of Figure 2.1, respectively. These measures are calculated using the CCL19-bound (purple) and CCL21-bound (green) aMD simulations.

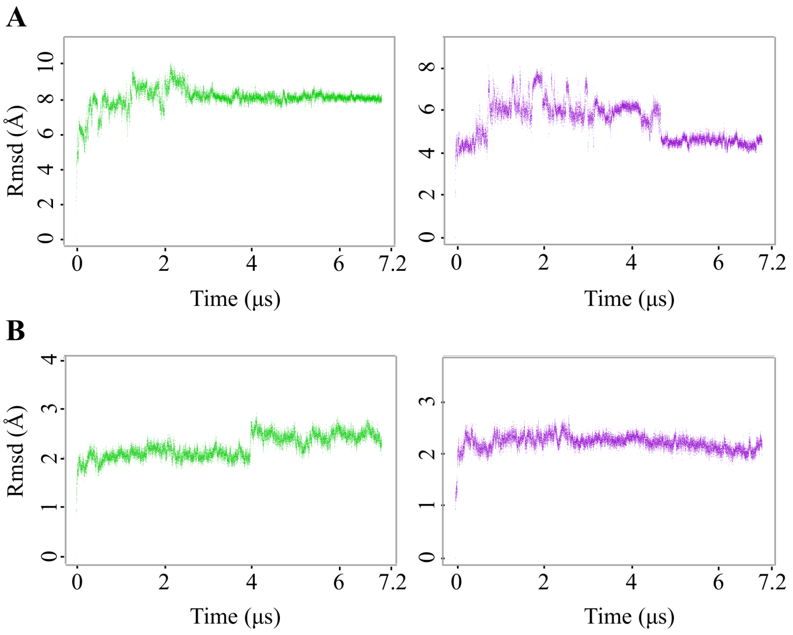


Figure A.5 Root mean square deviation (rmsd) calculated in both CCL19-bound (purple) and CCL21-bound (green) cMD simulations to assess the ligand-NTD stability and the receptor fluctuations. (A) Rmsd is calculated using intermolecular side-chain interactions between ligand-NTD and receptor within 5 Å. (B) Rmsd is calculated using intramolecular C_α-C_α interactions of the helical domain of CCR7 within 15 Å.

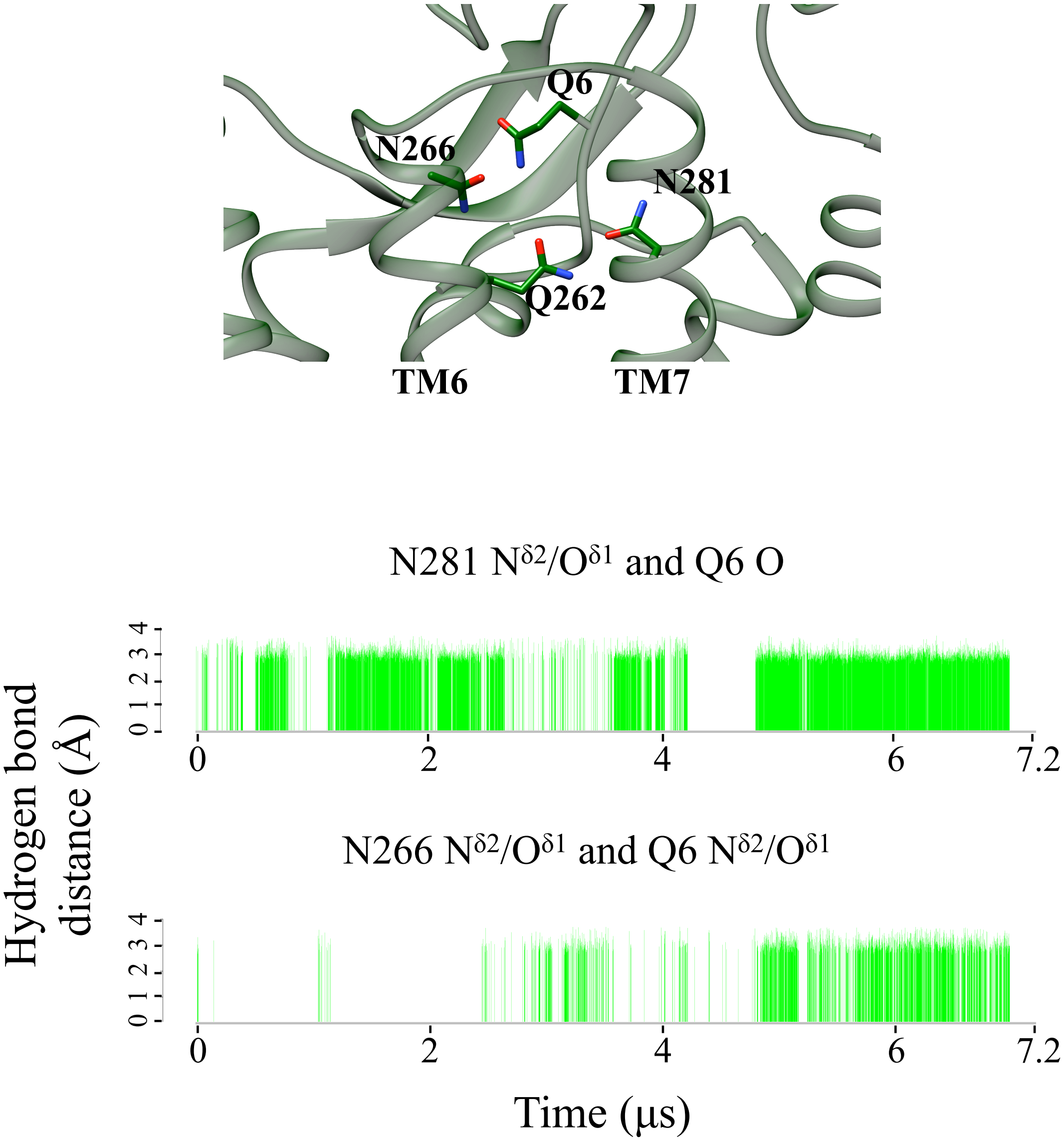


Figure A.6 CCL21 prompts a hydrogen bond network within TM6 and TM7 in cMD simulation. The molecular graphics is a representative structure of the dominant conformation showing a hydrogen bond between Q262$^{6.58}$ and N281$^{7.32}$ (4-7 μs). The hydrogen-bonding network comprises of Q262$^{6.58}$, N266$^{6.62}$, N281$^{7.32}$, and Q6^{CCL21}. Bar plots display the hydrogen bond distance time series between N281$^{7.32}$ side-chain and Q6^{CCL21} backbone and between N266$^{6.62}$ and Q6^{CCL21} side-chains. The atoms used to calculate the hydrogen bond distance are marked in the panel.

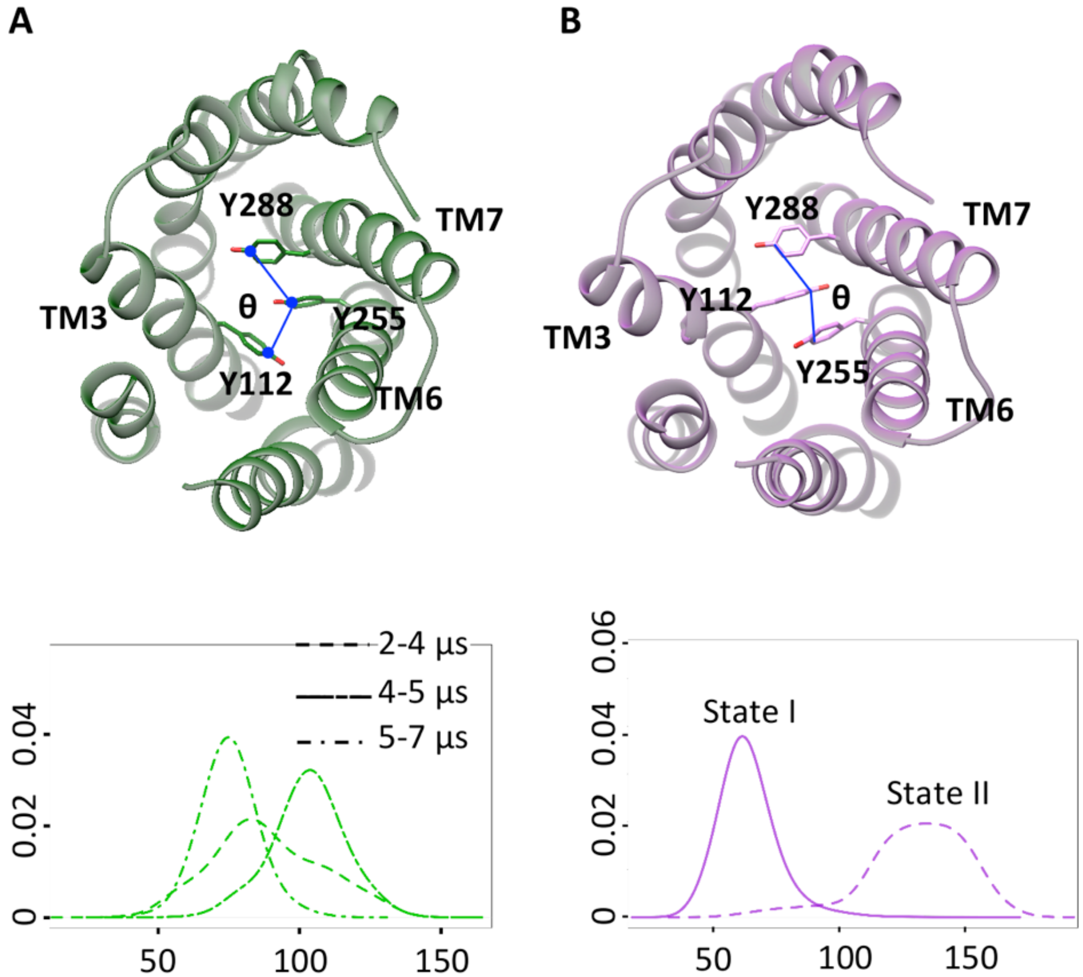


Figure A.7 π-stacking interactions in the tri-tyrosine switch. Molecular graphics represents a structure of the dominant conformation showing π-stacking interactions in the tri-tyrosine switch for CCL21-bound (A) and CCL19-bound (B) in cMD simulations. Y112$^{3.32}$, Y255$^{6.51}$, Y288$^{7.39}$ are shown as stick models and are labeled accordingly. θ is the angle formed by the C$_\zeta$ atoms of the three tyrosines. Distributions of the θ angle are plotted for the states indicated in the legend.

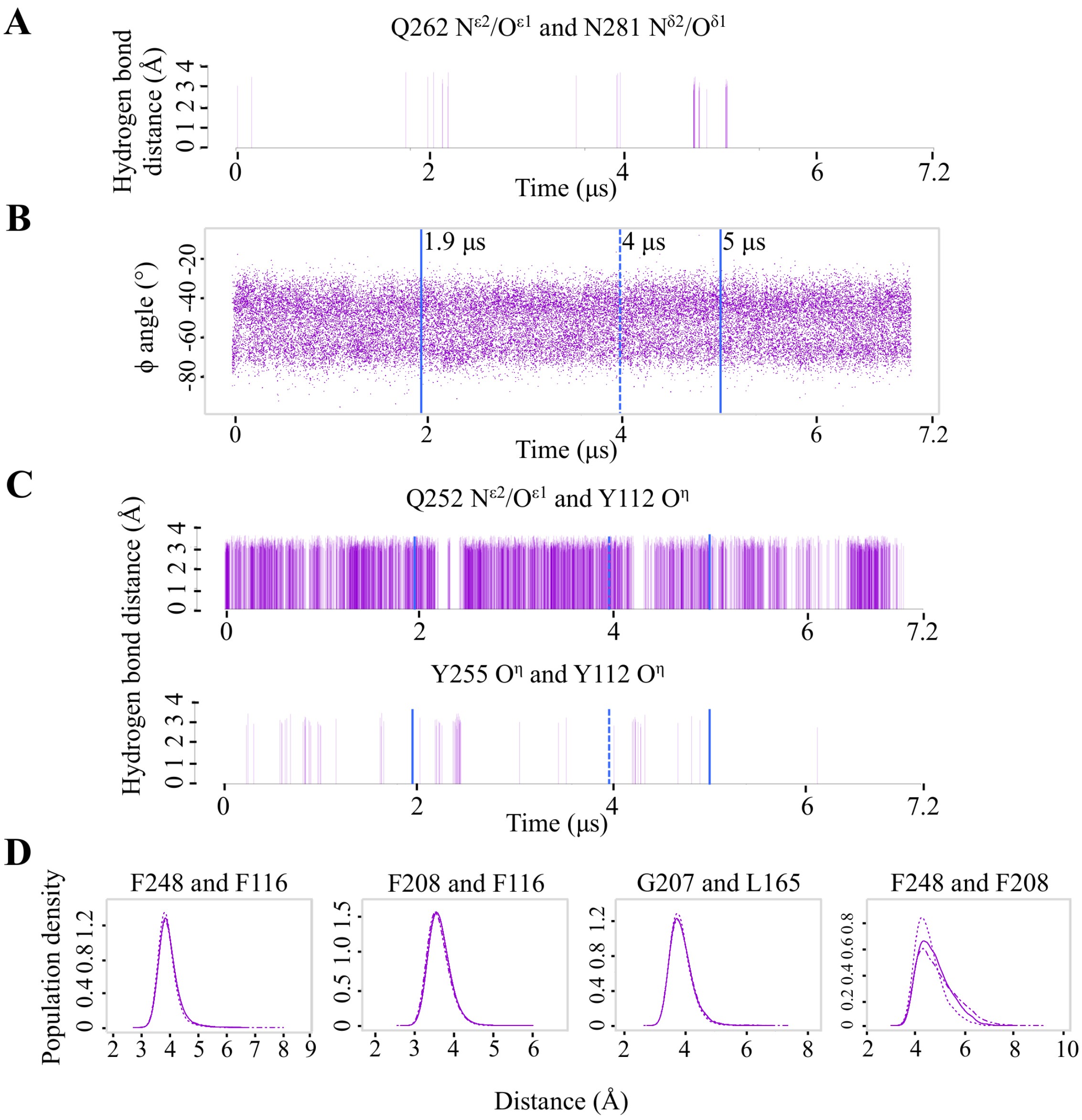


Figure A.8 Molecular switches characterized in CCL21-bound cMD simulation remain stable in the CCL19-bound cMD simulation. All four panels show the same measures as in panels B, C, D, and E of Figure 2.2. These measures are calculated using the CCL19-bound cMD simulation.

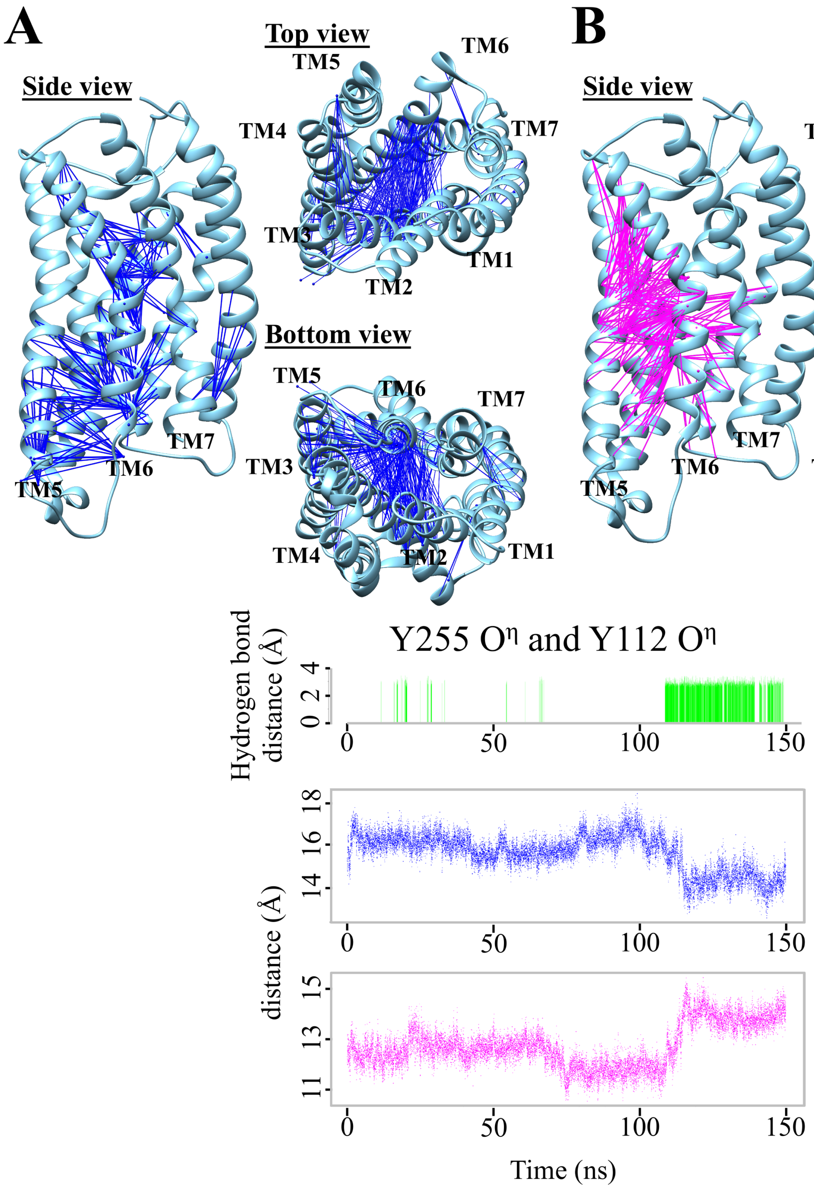


Figure A.9 During the aMD simulation, CCL21 induces global motions in CCR7 that show a decrease (A) and increase (B) in helical distances synchronized with the formation of hydrogen bond between Y112$^{3.32}$ and Y255$^{6.51}$. Side, top and bottom views of the receptor display the TM helices involved in the illustrated sets. Cross-correlation coefficient cutoff of 0.95 was used to cluster correlated Cα distance time series; and correlated time series that show a decrease (A) and increase (B) in helical distances that is coupled the formation of hydrogen bond between Y112$^{3.32}$ and Y255$^{6.51}$ were further grouped together. A representative distance time series from each of the sets is plotted with color code matching that in molecular graphics. The bar plot displays the hydrogen bond distance time series between Y112$^{3.32}$ and Y255$^{6.51}$ side-chains. The atoms used to calculate the hydrogen bond distance are marked in the panel.

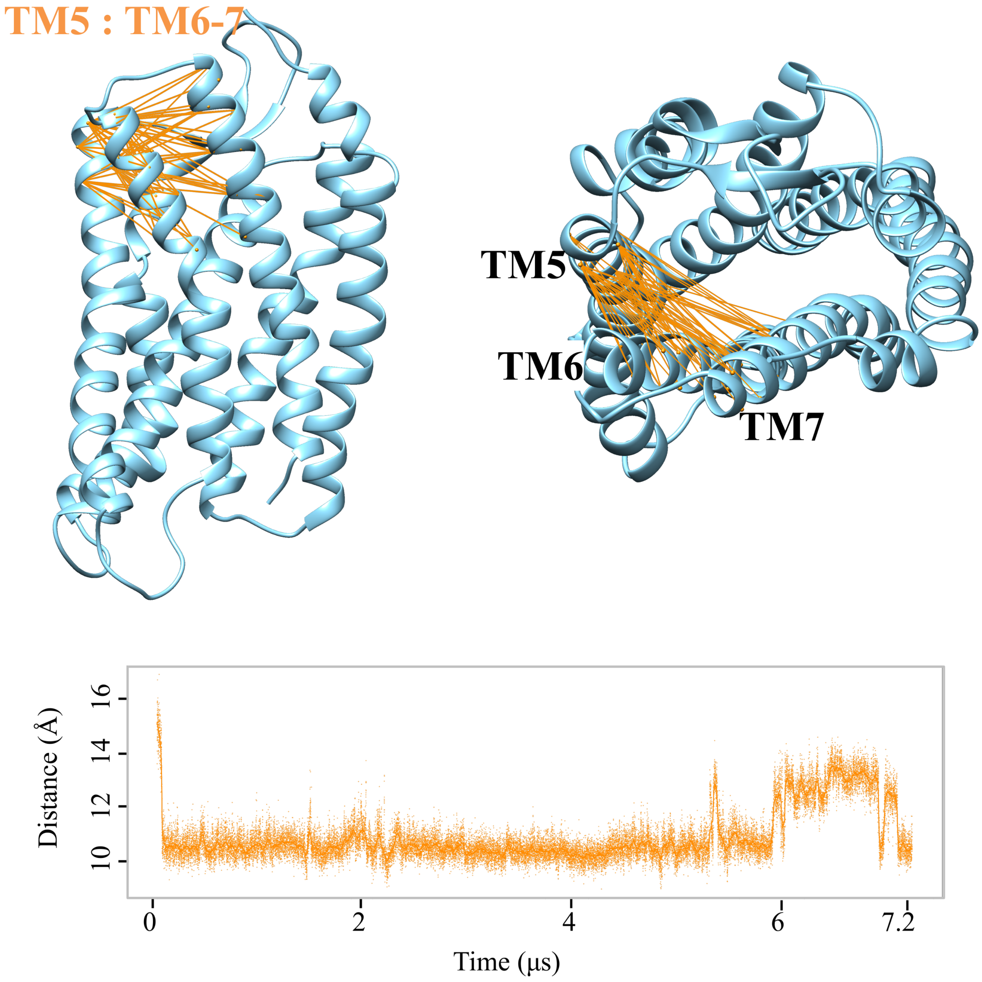


Figure A.10 During the cMD simulation, CCL19 binding promotes global helical motions in CCR7's EXC TM domains. Molecular graphics represents global helical motions in CCR7 as described in Figure 2.4. A representative distance time series from the set of correlated C_α-C_α distances is plotted with color-coding matching that in molecular graphics.

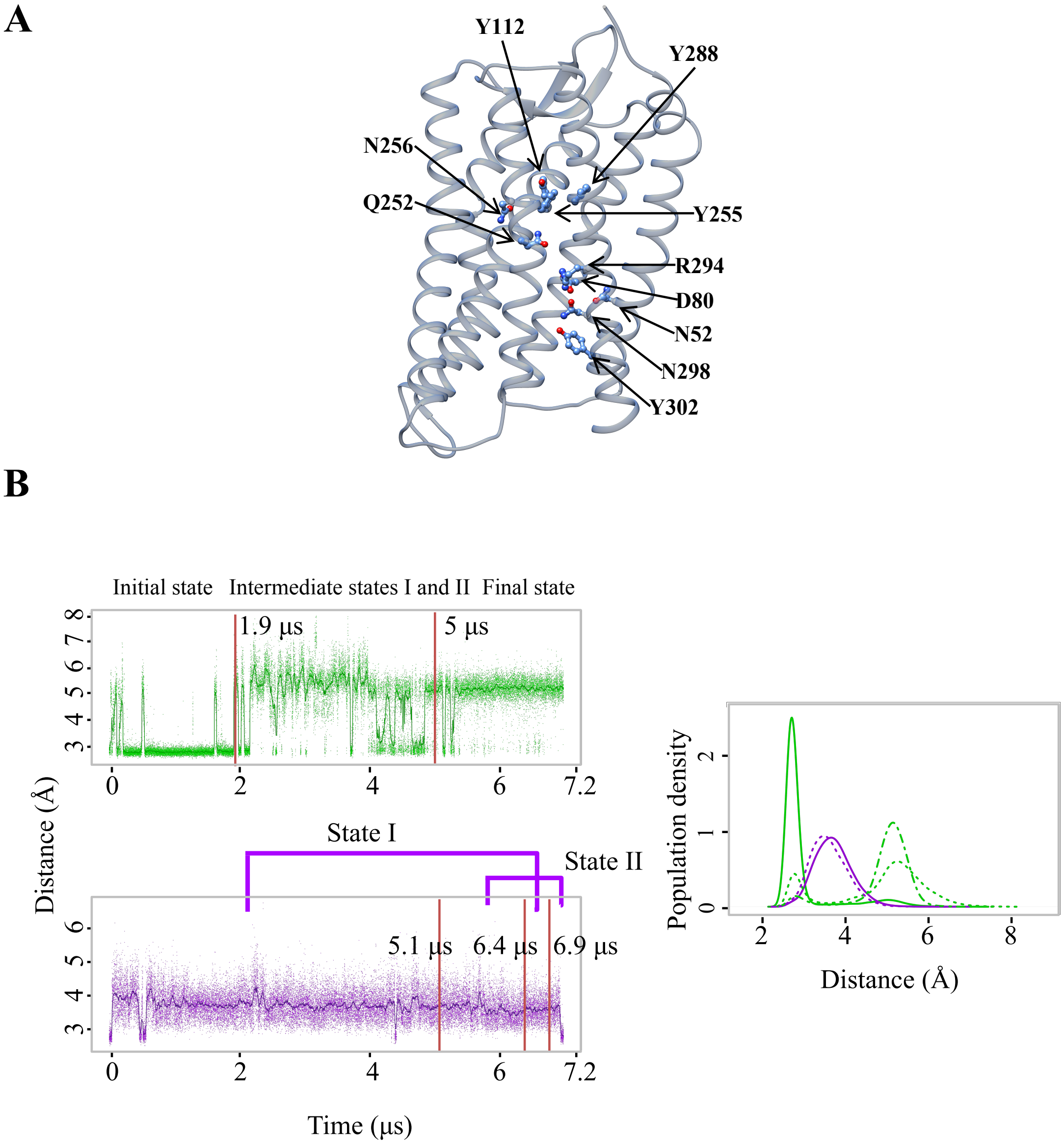


Figure A.11 CCL19 and CCL21 binding results in different arrangements in a polar interaction between TM6 and TM7 in cMD simulations. (A) Molecular graphics of CCR7 depicts polar residues involved in the characterized switches as stick models and labeled accordingly. (B) Side-chain distance between Q252$^{6.48}$ and R294$^{7.45}$ time series in CCL21 (green) and CCL19 (purple). The time series is broken down following the states in Figure 2.4 for CCL19 and Figure 2.2 for CCL21, and is labeled accordingly. Distributions of the distances are plotted for the following states in CCL21: solid line (initial state), dashed line (intermediate states I and II), and a dash-dot line (final state); and in CCL19: solid line (state I), dashed line (state II).

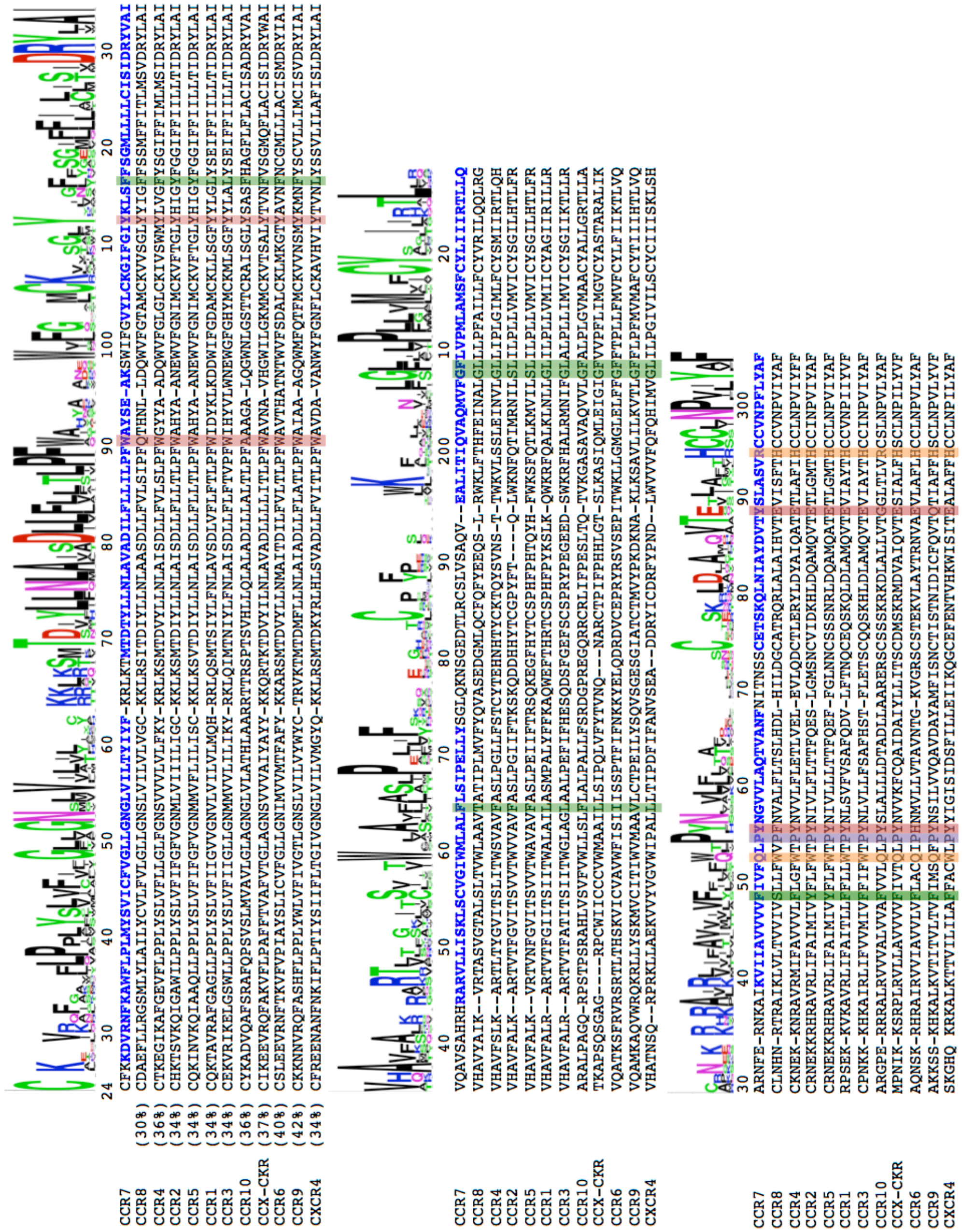


Figure A.12 Multiple sequence alignment of residues in CC chemokine receptor family and CXCR4 displays conserved motifs within the characterized switches. Percent identities of each chemokine receptor to CCR7 are reported between parentheses. The transmembrane helical domains are colored in blue in CCR7's sequence and CCR7's residue numbers are displayed above the alignment. Sequence logos are generated using the WebLogo3 (http://weblogo.berkeley.edu/logo.cgi). Residues involved in the characterized switches are highlighted in different colors: red (W90$^{2.60}$ and tri-tyrosine switch: Y112$^{3.32}$, Y255$^{6.51}$, and Y288$^{7.39}$), orange (Q252$^{6.48}$ and R294$^{7.45}$), purple (P254$^{6.50}$), and green (L165$^{4.56}$, G207$^{5.45}$ and tri-phenylalanine switch: F116$^{3.36}$, F208$^{5.47}$, and F248$^{6.44}$).

A.3 References

1. Ott, T.R., A. Pahuja, S.A. Nickolls, D.G. Alleva, and R.S. Struthers. 2004. Identification of CC Chemokine Receptor 7 Residues Important for Receptor Activation. J. Biol. Chem. 279: 42383–42392.

2. Love, M., J.L. Sandberg, J.J. Ziarek, K.P. Gerarden, R.R. Rode, D.R. Jensen, D.R. McCaslin, F.C. Peterson, and C.T. Veldkamp. 2012. Solution Structure of CCL21 and Identification of a Putative CCR7 Binding Site. Biochemistry (Mosc.). 51: 733–735.

3. Veldkamp, C.T., E. Kiermaier, S.J. Gabel-Eissens, M.L. Gillitzer, D.R. Lippner, F.A. DiSilvio, C.J. Mueller, P.L. Wantuch, G.R. Chaffee, M.W. Famiglietti, D.M. Zgoba, A.A. Bailey, Y. Bah, S.J. Engebretson, D.R. Graupner, E.R. Lackner, V.D. LaRosa, T. Medeiros, M.L. Olson, A.J. Phillips, H. Pyles, A.M. Richard, S.J. Schoeller, B. Touzeau, L.G. Williams, M. Sixt, and F.C. Peterson. 2015. Solution Structure of CCL19 and Identification of Overlapping CCR7 and PSGL-1 Binding Sites. Biochemistry (Mosc.). 54: 4163–4166.

APEENDIX B: SUPPORTING DATA FOR CHAPTER 5

B.1 Supplementary Figures

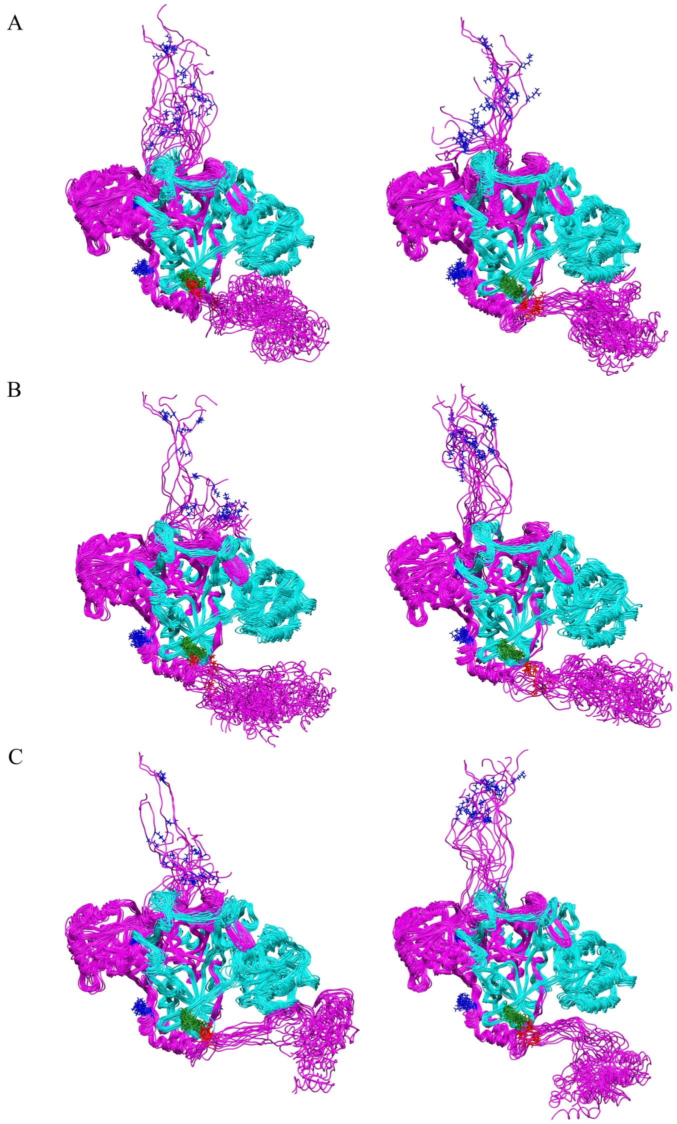


Figure B.1 Backbone representation of KU70 MD trajectories. Snapshots, corresponding to 1-ns intervals, from the MD of nonmethylated (left column) and methylated (right column) KU70-KU80 complex. Snapshots are superimposed on the initially modeled structure using C_α-atoms of the core of KU70-KU80 complex (excluding the NRC and the linker-SAP domains). The methylation (blue) and clamp residues (E537 {red}, R250/R260 {green}) are shown as stick models and are labeled accordingly. The panels represent the different triplicate runs.

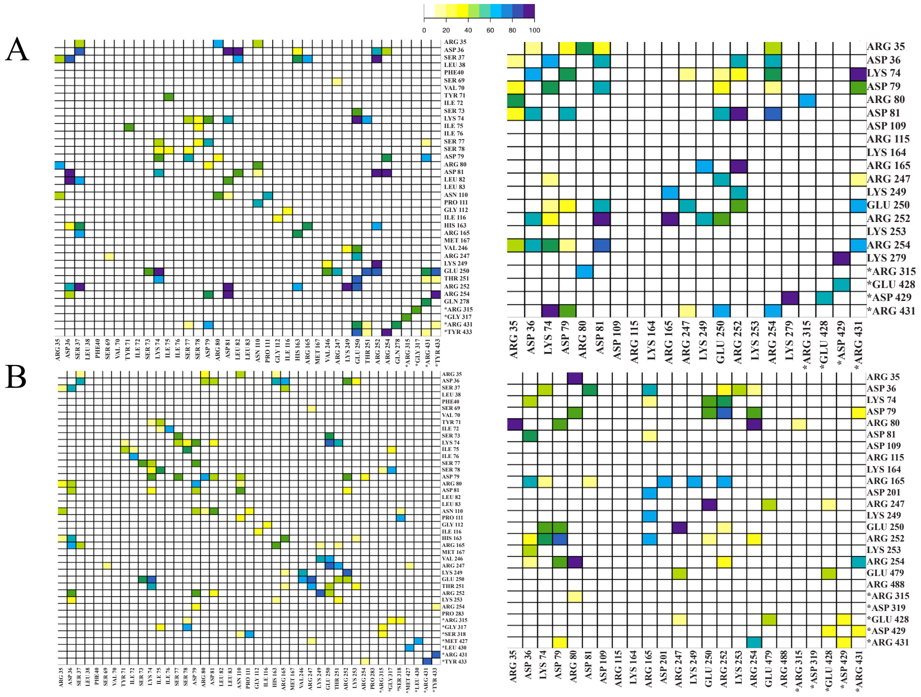


Figure B.2 Hydrogen bond (left column) and charge-charge (Coulombic) (right column) interaction occupancy changes in the vicinity of K74 between methylated and nonmethylated states of the complex in the MD simulations in its acetylated state. Occupancies are calculated as indicated in Figure 5.3B, and 5.3C.

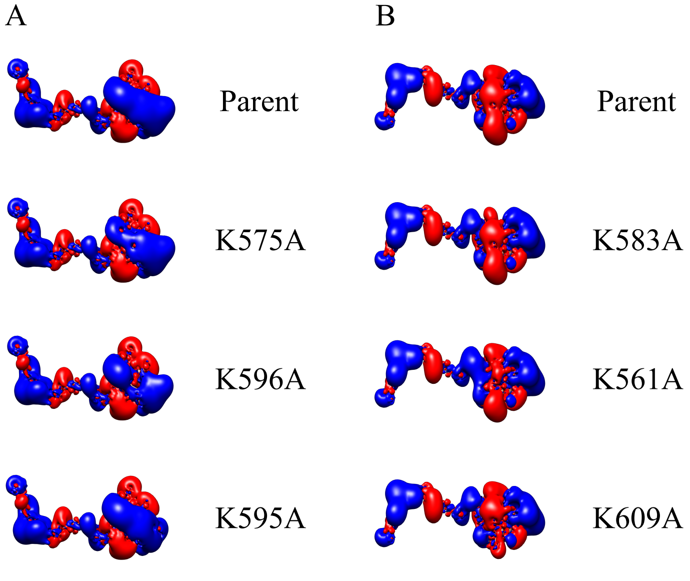


Figure B.3 Electrostatic potential of linker-SAP domain of KU70 illustrating the SAP domain positively-charged side (A) and negatively-charged side (B). Molecular isopotential contours are colored based on electrostatic potential values, -1 kBT/e in red and +1 kBT/e in blue. Electrostatic potentials are shown for the parent and alanine mutated linker-SAP domain as annotated on the figure.

APPENDIX C: ALGX BINDS ALGINATE THROUGH FINE-TUNED INTRA- AND INTER-MOLECULAR INTERACTIONS IN THE C-TERMINAL CARBOHYDRATE-BINDING MODULE

C.1 Introduction

Pseudomonas aeruginosa is the primary infection in patients with cystic fibrosis (CF) (1–3). Reprogramming of cell circuitry due to a mutation in the mucA gene converts *P. aeruginosa* into a fatal hyper alginate-producing strain (4). Such mutations have strong tendencies to occur in CF lungs, producing a thick mucus in the lungs, which, consequently, causes respiratory failure (5–7). Alginate is an exopolysaccharide consisting of acetylated mannuronic and guluronic acid with acetylation only occurring at the mannuronic acid residues (8). AlgX is a critical protein in alginate acetylation, although it's catalytic activity is not required for alginate biosynthesis (9, 10). Uronic acid polymers produced by an algX-deleted mutant, are primarily dimers and trimers (11), that are degraded in periplasm by AlgL, an alginate lyase (12). The crystal structure of AlgX revealed a protein that contains an N-terminal SGNH hydrolase-like domain and a C-terminal carbohydrate-binding module (CBM) (10), where, the mechanism and role of AlgX CBM binding to alginate have yet to be fully described.

In this study, we computationally assess the contribution of specific amino acid residues in alginate-binding by AlgX. Utilizing *in silico* docking and intra-molecular bond calculation studies, we identified the region and highlighted the inter- and intra-molecular interaction network in the CBM of AlgX responsible for alginate binding. Using alanine mutations, our collaborators in professor Neal Schiller's lab examined the *in vitro* and *in vivo* importance of our computationally-predicted amino acid residues in AlgX for alginate binding, biosynthesis, and alginate acetylation.

C.2 Materials and Methods

Structure modeling. The three-dimensional coordinates of the crystallographic structure of AlgX were obtained from the Protein Data Bank (PDB) using the PDB code 4KNC (10). AlgX was crystallized as a dimer with both molecules missing some residues due to the quality of electron density in the region. The missing residues from the protein are Glu-447 and Asp-250:Ser-251. We obtained the full protein sequence from UniProt (13) and aligned it to the sequence with missing residues obtained from the PDB, excluding the N-terminal and C-terminal loops that are not relevant for our study, using ClustalW2 (14). The alignment was used to generate a structure of AlgX with the missing residues using Modeller (15). Modeller has been shown to be very effective for short loops (16). The two missing segments, consisting of one and two amino acids, are located in solvent-exposed flexible loops, for which dynamic interconversion of multiple local conformations is expected; the conformations of the modeled three amino acids are representative within a locally optimized microenvironment. Overall, Modeller produced a structure with optimized side chain conformations, compared to the crystal structure.

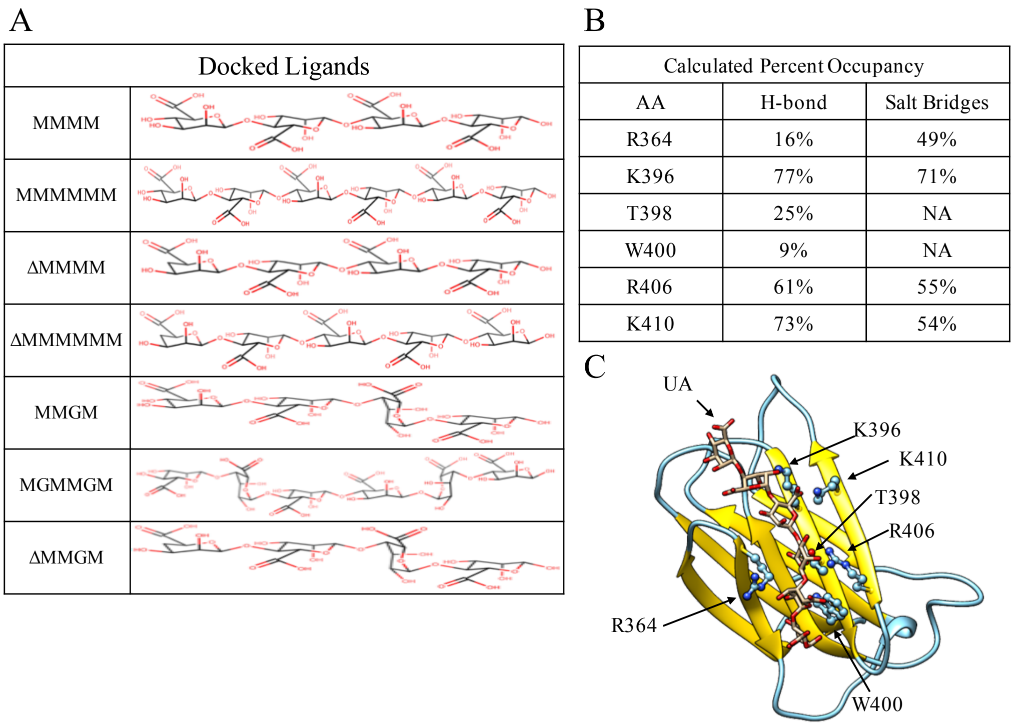
Crystal structures of UA polysaccharides, MMM, ΔMMM, and ΔMMGM, were extracted from PDB structures 2PYH, 1HV6, and 1Y3P, respectively (17–19). These structures were used as starting structures to generate different polysaccharide combinations of mannuronic (M) and guluronic (G) acid in the presence or absence of reducing ends (Δ). Polysaccharide combinations were generated using Chimera. Autodock Vina v1.1 (20) was used to dock our constructed alginate polymers into the carbohydrate-binding module of AlgX. We prepared the ligands to dock using AutoDockTools (21) by adding polar hydrogens to both the ligand and receptor, setting all single bonds of the ligand as rotatable, and saving both files as pdbqt files. The search space was reduced to the C-terminal carbohydrate-binding module of AlgX within a grid box of 27.5 x 41.75 x 34.3 Å.

After preparing the ligands and receptor and defining the binding site grid box, we used Autodock Vina to generate 20 models for each ligand at an exhaustiveness of 100 with a series of optimization tests provided by the Autodock Vina protocol. Models docked within the pinch point (10), see below, of the CBM were selected for further analysis. Resulting receptor-ligand bound complexes were analyzed using computational scripts to determine the percent occupancies of AlgX residues involved in alginate binding. The scripts were written in R (22) using the Bio3D v2.0 package and UCSF Chimera v1.8.1 (23). The occupancies represent the percentage of complexes that harbor a specific interaction between a residue in AlgX and a saccharide subunit in the docked ligand, from the complexes generated with Autodock Vina v1.1 with ligands bound at the pinch point. The complexes were examined to determine the presence of hydrogen bonds and salt bridges. Hydrogen bonds were extracted using the Chimera software and salt bridges were calculated using a cutoff value of 5 Å between the charged functional groups (23). Intra-molecular interactions were also calculated using a cutoff value of 8 Å to generate a network of ionic interactions illustrated in Figure C.2A.

Molecular dynamic (MD) simulations. MD simulations of AlgX were performed using NAMD, version 2.9 (24). Initial protein structure files were prepared using the PSFGEN utility in VMD (25) and the CHARMM27 forcefield (26) with CMAP terms (27). The protein complex was embedded into a water box using the VMD utility SOLVATE and the TIP3P model for the water molecules. The water box dimensions were 96 Å × 80 Å × 81 Å. The system was neutralized using sodium and chloride counterions at an ionic strength of 150 mM. NAMD (24) was used to minimize the system using 1000 steps of conjugate gradient energy minimization, followed by an MD production run for 10 ns at 1 atm pressure and 310 K. All production run simulations were performed using periodic boundary conditions and particle-mesh Ewald electrostatics for long-range electrostatic interactions with a grid point density of 1/Å. Nonbonded van der Waals interactions and short-range electrostatic interactions were calculated with an interaction cutoff of 12 Å and switching distance of 10 Å. The SHAKE algorithm was employed to fix the length of all hydrogen-containing bonds, enabling the use of 2 fs integration time steps. Coordinates were sampled every 2 ps.

C.3 Results

***In silico* docking of alginate polymers to the carbohydrate-binding module**. The crystal structure revealed that AlgX is a two-domain protein containing an N-terminal SGNH hydrolase-like domain and a C-terminal carbohydrate-binding module (CBM) (10). In the study conducted by Riley et al. AlgX CBM domain (10) was superimposed with CBM29-2 (PDB code 1GWK) in complex with a mannohexose ligand, indicating a set of four conserved amino acid residues (R364, T398, W400, and R406) dubbed the substrate recognition pinch point (SRPP). Given the differences in the architectural makeup of the alginate polymer compared to the mannohexose ligand, mainly by the presence of guluronic acid residues and acetylation in alginate, we sought to refine the conformational binding motif of AlgX and to provide additional insight into direct and indirect contributions from other amino acid residues. We docked a variety of alginate polymers using *in silico* modeling (Figure C.1A) to the CBM and calculated the interaction occupancies to provide the statistical significance of the interactions between specific AlgX amino acid residues and the alginate polymers (Figure C.1B). K396, R406, and K410 have calculated inter-molecular hydrogen-bonding percent occupancies of 77%, 61%, and 73%, respectively. R364, W400, T398 have a relatively low hydrogen-bonding occupancies of 16%, 9% and 25%, respectively. Salt bridge calculations demonstrated that R364, R406 and K410 have occupancies of 49%, 55% and 54% respectively, while K396 has the highest salt bridge occupancy of 71%. Our *in silico* data suggested that in addition to the previously proposed R364 and R406 of the putative SRPP, K396 and K410 are shown to be



Calculated Percent Occupancy		
AA	H-bond	Salt Bridges
R364	16%	49%
K396	77%	71%
T398	25%	NA
W400	9%	NA
R406	61%	55%
K410	73%	54%

Figure C.1 In silico docking of alginate ligands to AlgX carbohydrate binding module (CBM). (A) Variations of uronic acid polysaccharides composed of mannuronic (M) and guluronic (G) acids generated for the docking studies, some contain reduced ends (Δ). (B) Inter-molecular interaction occupancies of docked alginate ligands to AlgX CBM domain. NA (not applicable). (C) Molecular model of the CBM (in ribbon representation) with the amino acids examined for contributions to polysaccharide binding in complex with a representative uronic acid (UA) polysaccharide (MGMMGM).

significant contributors of inter-molecular interactions with the docked alginate polymers. Given our docking observations, we hypothesized that K396 and K410 also directly interact with the alginate polymer important for directing the polymer along the face of the CBM towards the hydrolase domain as previously mentioned by Riley et al. (10).

Conversely, T398 and W400 are not capable of forming salt bridges as they lack the anionic carboxylate or cationic amino functional groups. Thus, we have also calculated the aliphatic contribution of W400 and T398 to be 38% and 0% (28). Although T398 demonstrates some hydrogen-bonding and aliphatic contributions, and those of W400 may be negligible (Figure C.2B), T398 and W400 also contribute in inter- and intra-molecular interactions, the latter being important for the stability of the carbohydrate binding site. Intra-molecular analysis of the CBM revealed a polar environment with an elaborate network of salt bridges and hydrogen bonds, shown in Figure C.2A and Figure C.2B. The

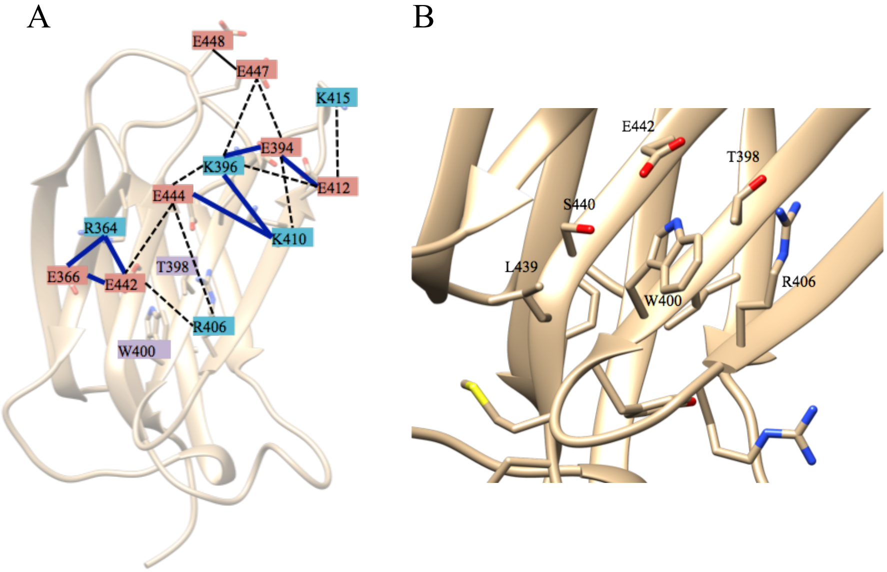


Figure C.2 Intra-molecular interaction network of the CBM. (A) Intra-molecular interaction network of charged residues in the CBM (in ribbon representation as a watermark). Red and blue boxes denote negatively and positively charged residues respectively. The black and dark blue lines are interactions between charged group heavy atoms of the residue side chains within 8 Å. The solid lines are interactions within 5.5 Å and the dashed lines are interactions between 5.5 Å and 8 Å. The thicker lines denote communities of charged interactions. The purple boxes denote non-charged residues. (B) Molecular model of the CBM (in ribbon representation) with the amino acids forming the polar cage around W400 shown in stick representation and labeled with residue letter code and number. The amino acid side chains shown are within 5 Å from W400.

analysis of the network suggests two communities of strong Coulombic interactions (within 5.5Å) in the alginate-binding site. T398 is situated within the two Coulombic communities, stabilizing the environment. Analysis of a 10-ns explicit-solvent molecular dynamics trajectory of the modeled AlgX structure revealed that T398 has the capacity to form polar/hydrogen bonds and non-polar intra-molecular contacts within 4 Å in the CBM, through its hydroxyl and methyl groups, respectively. Specifically, the hydroxyl and methyl groups of T398 are within 4 Å from polar or aliphatic groups, respectively, of the R364, W400, R406, E442 side chains (with more than 65% occupancies), stabilizing the CBM. Unlike T398, W400 is a bulky and hydrophobic tryptophan residue. Along with its inter-molecular contribution in aliphatic interactions to alginate binding, W400 is situated in a destabilizing polar cage environment formed by S440, E442, T398, R406, and R405 (Figure C.2B). Such a stressed environment might be associated with the functional dissociation of alginate.

C.4 Discussion

AlgX is a required component for alginate biosynthesis and its absence resulted in the loss of the mucoid phenotype due to the production of small Uronic acid polymers degraded by AlgL (11, 12). Here we elucidate the mechanism and role of AlgX CBM binding to alginate. Our *in silico* studies demonstrate that R364, K396, R406, K410, T398, and W400 are essential for alginate binding. These residues are further examined by our collaborators through site-specific alanine mutational studies.

Previously published work demonstrated the presence of a set of residues in the CBM of AlgX that could accommodate a single hexamannose polysaccharide (10). These residues form the (SRPP) and are composed of two basic (R364 and R406), an aromatic (W400), and a polar (T398) amino acid residues. Architecturally, the alginate polymer differs from the hexamannose polysaccharide, which is primarily a linear polymer. This is primarily due to the epimerization of mannuronic acid at C-5 position, by AlgG, into guluronic acid that causes a sharp bend in the alginate polymer (Figure C.1A) (29). To accommodate for this structural difference, we used various combinations of UA ligands composed of mannuronic and guluronic monomers as substrates for our docking studies. From our *in silico* modeling, we demonstrated that many of the amino acid residues that were proposed by Riley et al. (10) in the SRPP have high occupancies of hydrogen bonds and salt bridges with the docked alginate polymers, with the exception of T398 and W400 that were found to have low hydrogen bond occupancies.

Intra-molecular analysis of the CBM shows that W400 is situated in an unfavorable polar cage environment (Figure C.2B). We consider the presence of the bulky and hydrophobic tryptophan residue to be critical for providing steric hindrance in alginate binding. Although this steric hindrance contributes to reduction of binding affinity experimentally (28), it may be necessary to maintain a fine balance between association and dissociation in order to promote the alginate-shuttling mechanism of AlgX. Therefore, upon W400 mutation, this hindrance is eliminated and binding affinity is increased. In addition to W400, E444 and E447 show aliphatic interaction occupancies of 62% and 55%. These negatively charged residues form unfavorable interactions with the negatively charged carbohydrates, thus contributing to the driving force to pull alginate across the periplasm.

Experimental mutation of T398A resulted in a significant reduction in alginate binding (28). Given that T398 was shown to not interact with the alginate polymer, it is likely that T398 acts as a stabilizing factor between the two communities of strong Coulombic interactions (Figure C.2A) through polar and non-polar interactions, involving its side chain hydroxyl group and methyl group, respectively. The polar character of the hydroxyl group of T398 is a stabilizing factor for the surrounding strong Coulombic interactions, whereas the non-polar interactions from T398's methyl group stabilize the aliphatic groups of nearby side chains (including those of Arg, Lys, and Glu residues). Therefore, T398 may have an indirect role in binding by maintaining the integrity of the Coulombic network within the carbohydrate binding site. It is likely that mutation of T398 abolishes binding through a relay effect, because of the disruption of the Coulombic network, owed to the loss of the branched polar-methyl side chain moiety. Hence, T398 and W400 may play a concerted balancing role, with T398 stabilizing the local environment, and effectively stabilizing the Coulombic network that is important for binding, whereas W400 disfavors binding. Hence, disruption of these fine-tuned intra- and inter-molecular interactions, which are important for the alginate-shuttling mechanism, consequently affects alginate production.

In this work, we have demonstrated that AlgX binds alginate through a fine-tuned intra- and inter-molecular interactions network. In addition to the previously proposed R364 and R406 of the putative SRPP, we highlight the significance of K396 and K410 in contributing to the inter-molecular interactions with the docked alginate polymers. Furthermore, although T398 and W400 were found to form minimal inter-molecular interaction with alginate, they stabilize several intra-molecular interactions in the CBM and provide a balancing role responsible for the alginate-shuttling mechanism in alginate production.

C.5 References

1. Cheng, S.H., R.J. Gregory, J. Marshall, S. Paul, D.W. Souza, G.A. White, C.R. O'Riordan, and A.E. Smith. 1990. Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. Cell. 63: 827–834.

2. Gibson, R.L., J.L. Burns, and B.W. Ramsey. 2003. Pathophysiology and Management of Pulmonary Infections in Cystic Fibrosis. Am. J. Respir. Crit. Care Med. 168: 918–951.

3. Pier, G.B. 1985. Pulmonary disease associated with Pseudomonas aeruginosa in cystic fibrosis: current status of the host-bacterium interaction. J. Infect. Dis. 151: 575–580.

4. DeVries, C.A., and D.E. Ohman. 1994. Mucoid-to-nonmucoid conversion in alginate-producing Pseudomonas aeruginosa often results from spontaneous mutations in algT, encoding a putative alternate sigma factor, and shows evidence for autoregulation. J. Bacteriol. 176: 6677–6687.

5. Kerem, B., J.M. Rommens, J.A. Buchanan, D. Markiewicz, T.K. Cox, A. Chakravarti, M. Buchwald, and L.C. Tsui. 1989. Identification of the cystic fibrosis gene: genetic analysis. Science. 245: 1073–1080.

6. Riordan, J.R., J.M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J.L. Chou, and E. Al. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science. 245: 1066–1073.

7. Rommens, J.M., M.C. Iannuzzi, B.S. Kerem, M.L. Drumm, G. Melmer, M. Dean, R. Rozmahel, J.L. Cole, D. Kennedy, N. Hidaka, M. Zsiga, M. Buchwald, J.R. Riordan, L.C. Tsui, and F.S. Collins. 1989. Identification of the cystic fibrosis gene: chromosome walking and jumping. Science. 245.

8. Rehm, B.H.A., and S. Valla. Bacterial alginates: biosynthesis and applications. Appl. Microbiol. Biotechnol. 48: 281–288.

9. Baker, P., T. Ricer, P.J. Moynihan, E.N. Kitova, M.T.C. Walvoort, D.J. Little, J.C. Whitney, K. Dawson, J.T. Weadge, H. Robinson, D.E. Ohman, J.D.C. Codée, J.S. Klassen, A.J. Clarke, and P.L. Howell. 2014. P. aeruginosa SGNH Hydrolase-Like Proteins AlgJ and AlgX Have Similar Topology but Separate and Distinct Roles in Alginate Acetylation. PLOS Pathog. 10: e1004334.

10. Riley, L.M., J.T. Weadge, P. Baker, H. Robinson, J.D.C. Codée, P.A. Tipton, D.E. Ohman, and P.L. Howell. 2013. Structural and Functional Characterization of Pseudomonas aeruginosa AlgX ROLE OF AlgX IN ALGINATE ACETYLATION. J. Biol. Chem. 288: 22299–22314.

11. Robles-Price, A., T.Y. Wong, H. Sletta, S. Valla, and N.L. Schiller. 2004. AlgX Is a Periplasmic Protein Required for Alginate Biosynthesis in Pseudomonas aeruginosa. J. Bacteriol. 186: 7369–7377.

12. Monday, S.R., and N.L. Schiller. 1996. Alginate synthesis in Pseudomonas aeruginosa: the role of AlgL (alginate lyase) and AlgX. J. Bacteriol. 178: 625–632.

13. Consortium, T.U. 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43: D204–D212.

14. Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23: 2947–2948.

15. Webb, B., and A. Sali. 2014. Comparative protein structure modeling using Modeller. Curr. Protoc. Bioinforma. : 5–6.

16. Jamroz, M., and A. Kolinski. 2010. Modeling of loops in proteins: a multi-method approach. BMC Struct. Biol. 10: 5.

17. Rozeboom, H.J., T.M. Bjerkan, K.H. Kalk, H. Ertesvåg, S. Holtan, F.L. Aachmann, S. Valla, and B.W. Dijkstra. 2008. Structural and Mutational Characterization of the Catalytic A-module of the Mannuronan C-5-epimerase AlgE4 from Azotobacter vinelandii. J. Biol. Chem. 283: 23819–23828.

18. Yoon, H.-J., W. Hashimoto, O. Miyake, K. Murata, and B. Mikami. 2001. Crystal structure of alginate lyase A1-III complexed with trisaccharide product at 2.0 Å resolution1. J. Mol. Biol. 307: 9–16.

19. Momma, K., Y. Mishima, W. Hashimoto, B. Mikami, and K. Murata. 2005. Direct Evidence for Sphingomonas sp. A1 Periplasmic Proteins as Macromolecule-Binding Proteins Associated with the ABC Transporter: Molecular Insights into Alginate Transport in the Periplasm,. Biochemistry (Mosc.). 44: 5053–5064.

20. Trott, O., and A.J. Olson. 2010. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. 31: 455–461.

21. Morris, G.M., R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, and A.J. Olson. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem. 30: 2785–2791.

22. Grant, B.J., A.P.C. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S.D. Caves. 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 22: 2695–2696.

23. Pettersen, E.F., T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. 2004. UCSF Chimera-A visualization system for exploratory research and analysis. J. Comput. Chem. 25: 1605–1612.

24. Phillips, J.C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, and K. Schulten. 2005. Scalable molecular dynamics with NAMD. J. Comput. Chem. 26: 1781–1802.

25. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: Visual molecular dynamics. J. Mol. Graph. 14: 33–38.

26. MacKerell, A.D., D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins [†]. J. Phys. Chem. B. 102: 3586–3616.

27. Mackerell, A.D., M. Feig, and C.L. Brooks. 2004. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J. Comput. Chem. 25: 1400–1415.

28. Do, D.C. 2014. The Role of Bacterial Biofilms in Chronic Infections. UC Riverside: Biomedical Sciences. Retrieved from: http://escholarship.org/uc/item/30k1t548.

29. Jain, S., M.J. Franklin, H. Ertesvåg, S. Valla, and D.E. Ohman. 2003. The dual roles of AlgG in C-5-epimerization and secretion of alginate polymers in Pseudomonas aeruginosa. Mol. Microbiol. 47: 1123–1133.

APPENDIX D: ROLES OF APPBP1 AND UBA3 IN NEDD8 ACTIVATION

D.1 Introduction

Ubiquitin and Ubiquitin-like proteins (Ubls) are small regulatory proteins that are conjugated to various target proteins through a conserved but distinct enzymatic cascade (1). In general, this process requires three types of enzymes: activating enzymes E1s, conjugation enzymes E2s, and ligases E3s (2). In this study, we focus on a specific Ubl, NEDD8, which is part of the NEDDylation enzymatic cascade and whose activating enzyme E1 is the heterodimer protein APPBP1-UBA3 (3). We aim to characterize the molecular interactions between the subunits of E1 heterodimer and its function for NEDD8 activation. Ultimately, the goal of this study is to use computation to guide experimental mutagenesis studies in order to assess the molecular mechanism behind NEDD8 activation. We isolated three charged residues (E43A, D331A, and K506A) that are deemed important in the interaction between APPBP1 and UBA3 that our collaborators, professor Jiayu Liao's lab, validated experimentally. Aided with our analysis using the AESOP (Analysis of Electrostatic Similarities Of Proteins) computational framework, our collaborators were able to systematically dissected the kinetics, reaction intermediates, and non-covalent molecular interactions during NEDD8 activation. The residues produced were pivotal in assessing the role of APPBP1 as a scaffolding protein.

D.2 Materials and Methods

Computational alanine scan and electrostatic analysis. Calculations were performed using the computational framework AESOP (see Chapter 5 for method details) to delineate the role of charged amino acids in binding (4–8). This approach is useful for understanding the mechanism of protein-protein interactions, by performing computational alanine scans as a means of perturbation to quantitatively assess the impact of each mutation on the stability of the protein complex. The approach provides insight on the role of electrostatics in the formation of complexes between highly charged proteins, or proteins with localized charged patches. As the net charge of APPBP1, UBA3, and NEDD8 is -18, -9, and 0, respectively, we sought to elucidate the effects of charged amino acids on the stability of the APPBP1-UBA3-NEDD8 complex and pairwise complexes APPBP1-UBA3, UBA3-NEDD8, and APPBP1-NEDD8, in order to understand the mechanism of the specific protein-protein interactions.

The coordinates of the APPBP1-UBA3-NEDD8 complex were obtained from the Protein Data Bank (PDB code: 1R4N) (9). The program PDB2PQR (10) was used to add atomic radii and partial charges to the atomic coordinate file using the PARSE force field (11) for electrostatic potential calculations with

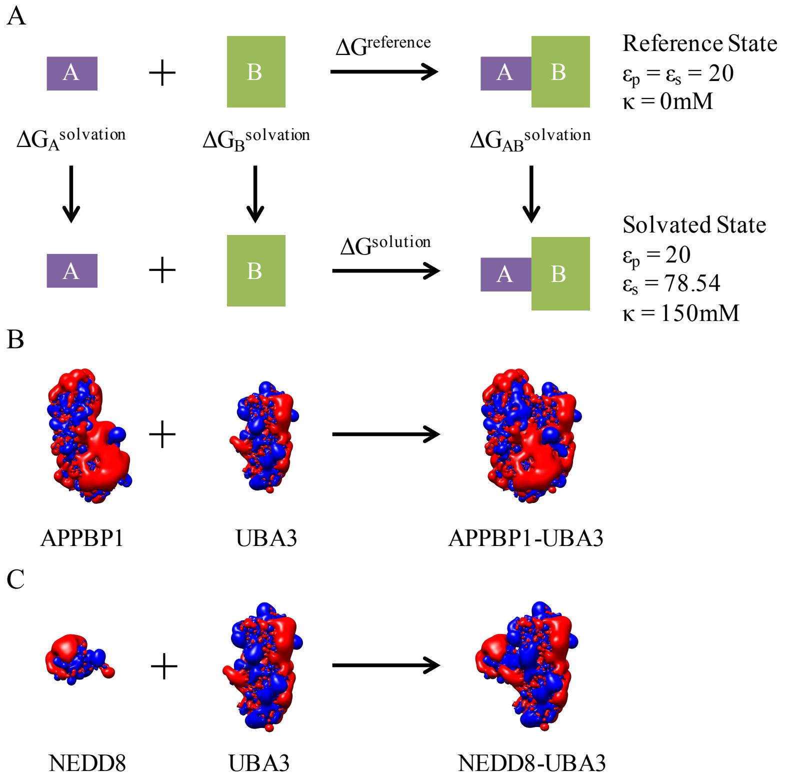


Figure D.1 (A) Thermodynamic cycle for calculation of electrostatic free energy of association and solvation. The top process models the reference state (dielectric coefficient of 20 for both proteins in interior and exterior and no counter ions), and the bottom process models the solvated state with dielectric coefficient of 20 for protein interior, 78.54 for protein exterior, and 150 mM concentration of counter ions. The three vertical processes show solvation of each free protein and complex. The iso-potential contours for A, B, and AB are shown as iso-potential contours in panels (B) and (C). (B) The iso-potential contours for APPBP1, UBA3, and APPBP1-UBA3 illustrated are the spatial distribution of electrostatic potential. Blue and red surfaces represent iso-values of ±1kBT/e, respectively, where kB is the Boltzmann constant, T is temperature, and e is the unit charge of an electron. (C) The iso-potential contours for NEDD8, UBA3, and NEDD8-UBA3.

the Adaptive Poisson-Boltzmann Solver (APBS) (12), or the stand-alone program COULOMB supplied by APBS.

The molecular (dielectric boundary) and ion accessibility surfaces were determined using spherical probes and radii set to 1.4 and 2.0 Å, respectively. The APBS calculations were performed on a grid consisting of 225 × 225 × 225 grid points with grid dimensions of 204 Å × 202 Å × 154 Å. For the clustering analysis and the calculation of electrostatic potentials in solution (bottom horizontal process in the thermodynamic cycle of Figure D.1) the dielectric coefficients for the solvent and protein were set to 78.54 and 20, respectively (5), and ionic strength was set to 150 mM. For the calculation of electrostatic potentials in the reference state (top horizontal process in the thermodynamic cycle of Figure D.1) the dielectric coefficient was set to 20 for both solvent and protein (5), and the ionic strength was set to 0 mM.

D.3 Results

The APPBP1-UBA3 heterodimer functions as E1 and catalyzes an adenylation reaction to produce an adenylated NEDD8 intermediate with ATP at its C-terminal glycine. In the E1-NEDD8-ATP crystal structure, both APPBP1 and UBA3 interact with NEDD8 (9).

The interaction of APPBP1 with UBA3 is not required for NEDD8 activation, but contributes to its rapid activation. As part of the study, our collaborators demonstrated that APPBP1 itself and the interaction of APPBP1 with NEDD8/ATP are not requited for NEDD8 rapid activation (13). However, APPB1 interaction with UBA3 may still play a role in the activation of NEDD8. Therefore, we examine the significance of the E1 heterodimer (interaction between APPBP1 and UBA3) interaction for NEDD8 activation.

To assess, whether this interaction was important, we performed a computational mutagenesis study to identify key amino acids for the formation of the APPBP1-UBA3 complex. Given that APPBP1 and UBA3 are highly charged, we focused on the effects of electrostatic potentials in association. We performed a systematic computational alanine scan analysis, in which we replaced every ionizable amino acid, one at a time, by alanine, followed by calculation and clustering of electrostatic potentials, as well as the calculation of electrostatic free energies of association, using the AESOP computational framework. Figure D.2 shows the results of the AESOP analysis, which allowed us to select APPBP1 mutants that were predicted to perturb the APPBP1-UBA3 interface, and to assess the effect of the APPBP1-UBA3 complex

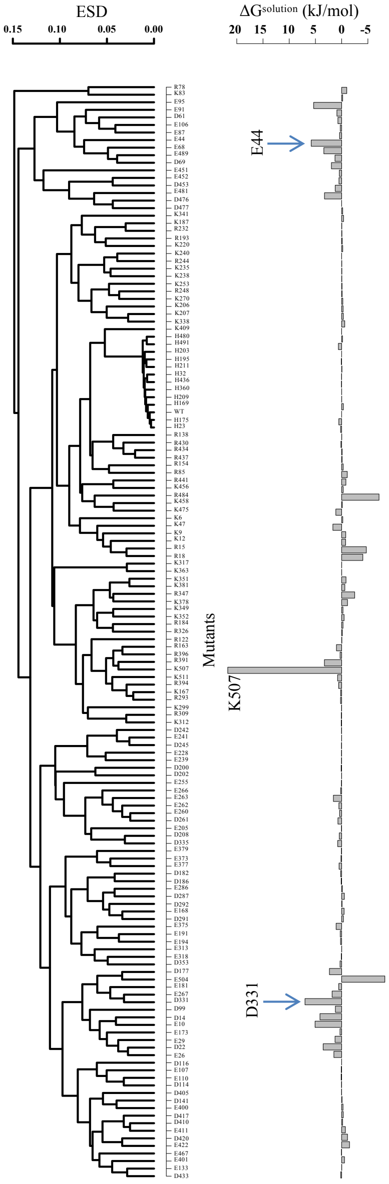


Figure D.2 Non-covalent interactions between APPBP1 and UBA3 are critical for NEDD8 activation and conjugation. Alanine scan electrostatic clustering and free energies of association of APPBP1 and NEDD8. Clustering dendrogram of the alanine scan mutants of APPBP1 using the average weighted difference ESD and the average linkage method. Free energies of mutants are ordered according to average weighted difference clustering (4-8).

on the activation of NEDD8. We selected mutations E43A, D331A, and K506A from three distinct dendrogram clusters with the most significant effects on loss of binding (Figure D.2), to construct single, double, and triple mutation combinations for experimental testing. Consistent with the computational analysis, all four APPBP1 mutants exhibited decreased interactions with UBA3, as illustrated by the quantitative FRET assay conducted by our collaborators (13). As expected by our AESOP analysis, APPBP1 triple mutant (D331A/E43A/K506A) showed the biggest reduction in affinity for UBA3.

D.4 Discussion

Our study provides a systematic and detailed investigation of the contribution of charged residues to complex formation between APPBP1 and UBA3. We isolate three charged residues that are deemed important in the interaction between APPBP1 and UBA3. These predictions were verified experimentally by our collaborators and showed reduced binding between APPBP1 and UBA3 upon alanine mutation. Additional enzymatic reaction kinetics studies were performed to evaluate the role of these residues and heterodimer formation on NEDD8 activation (13). Four mutant sets of APPBP1, D331A, E43A/D331A, D331A/K506A, and E43A/D331A/K506A, were generated by our collaborators and examined for NEDD8 activation. The single mutant, D331A, slowed the initial NEDD8 activation, but it was not able to completely abolish NEDD8 activation and conjugation. The double mutants, D331A/K506A and E43A/D331A, resulted in reduced kinetics and/or partial activation of NEDD8. Interestingly, the triple mutant, D331A/E43A/K506A, almost fully constrained NEDD8 activation to a great extent as compared to the NEDD8 activation in the absence of APPBP1 but presence of UBA3. The stronger impact of the triple mutations rather than the deletion APPBP1, led us to speculate that the D331A/E43A/K506A mutant may serve as a dominant negative mutant. The three mutations are located at the interface of APPBP1 and UBA3, and may introduce local destabilization effects. We speculate that the D331A/E43A/K506A mutant may serve as a dominant negative mutant that can no longer interact with UBA3 and instead competes for NEDD8 binding with UBA3, thus limiting UBA3-NEDD8 interaction and catalytic reaction. This implies that APPBP1 functions mainly as a scaffold protein to enhance molecular interactions and facilitate the catalytic reaction. Our studies provide mechanistic insights into the complex formation between APPBP1 and UBA3 and the role of the heterodimer in NEED8 activation.

D.5 References

1. Herrmann, J., L.O. Lerman, and A. Lerman. 2007. Ubiquitin and Ubiquitin-Like Proteins in Protein Regulation. Circ. Res. 100: 1276–1291.

2. Hochstrasser, M. 2000. Evolution and function of ubiquitin-like protein-conjugation systems. Nat. Cell Biol. 2: E153–E157.

3. Watson, I.R., M.S. Irwin, and M. Ohh. 2011. NEDD8 Pathways in Cancer, Sine Quibus Non. Cancer Cell. 19: 168–176.

4. Gorham, R.D., C.A. Kieslich, and D. Morikis. 2011. Complement Inhibition by Staphylococcus aureus. Cell. Mol. Bioeng. 5: 32–43.

5. Gorham, R.D., C.A. Kieslich, A. Nichols, N.U. Sausman, M. Foronda, and D. Morikis. 2011. An evaluation of Poisson–Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. Biopolymers. 95: 746–754.

6. Gorham, R.D., C.A. Kieslich, and D. Morikis. 2010. Electrostatic Clustering and Free Energy Calculations Provide a Foundation for Protein Design and Optimization. Ann. Biomed. Eng. 39: 1252–1263.

7. Kieslich, C.A., R.D. Gorham Jr., and D. Morikis. 2011. Is the rigid-body assumption reasonable?: Insights into the effects of dynamics on the electrostatic analysis of barnase–barstar. J. Non-Cryst. Solids. 357: 707–716.

8. Kieslich, C.A., D. Morikis, J. Yang, and D. Gunopulos. 2011. Automated computational framework for the analysis of electrostatic similarities of proteins. Biotechnol. Prog. 27: 316–325.

9. Walden, H., M.S. Podgorski, D.T. Huang, D.W. Miller, R.J. Howard, D.L. Minor Jr., J.M. Holton, and B.A. Schulman. 2003. The Structure of the APPBP1-UBA3-NEDD8-ATP Complex Reveals the Basis for Selective Ubiquitin-like Protein Activation by an E1. Mol. Cell. 12: 1427–1437.

10. Dolinsky, T.J., J.E. Nielsen, J.A. McCammon, and N.A. Baker. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 32: W665–W667.

11. Sitkoff, D., K.A. Sharp, and B. Honig. 1994. Accurate calculation of hydration free energies using macroscopic solvent models. J. Phys. Chem. 98: 1978–1988.

12. Baker, N.A., D. Sept, S. Joseph, M.J. Holst, and J.A. McCammon. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. 98: 10037–10041.

13. Chaudhry, M., and H. Kaur. 2014. Development of in vitro Sensitive FRET Technology to Dissect Reaction Dynamics and Intermediates in NEDD8 Conjugation Cascade. UC Riverside: Bioengineering. Retrieved from: http://escholarship.org/uc/item/2v50z8xd.

APPENDIX E: THE ROLE OF ZINC AND CALCIUM IONS IN MMP-14 AND TIMP-2 ASSOCIATION

E.1 Introduction

Matrix metalloproteinases (MMPs) are a family of Zinc dependent endopeptidases that play a key role in tissue remodeling (embryogenesis, growth, and wound healing) and migration of cells through the body. These enzymes are also involved in a number of diseases, such as arthritis, fibrosis, and tumor invasion (1). All MMPs illustrate the same catalytic domain sequence motif (HEXXHXXGXXHX) which contains the zinc ion coordination site. The catalytic zinc is coordinated by three histidine residues and a catalytic glutamate. Additionally, a second zinc and two calcium ions, found in MMPs, are catalytically inactive but structurally important (2). The catalytic site is regulated by a cysteine switch mechanism involved in MMP inhibition (3). The interaction of the thiol group and the N-terminal α-amino nitrogen with the catalytic zinc inactivates the MMP activity by stabilizing the positively charged catalytic site (4).

The role of MMP14 in tumor angiogenesis makes it one of the most crucial MMPs in both,

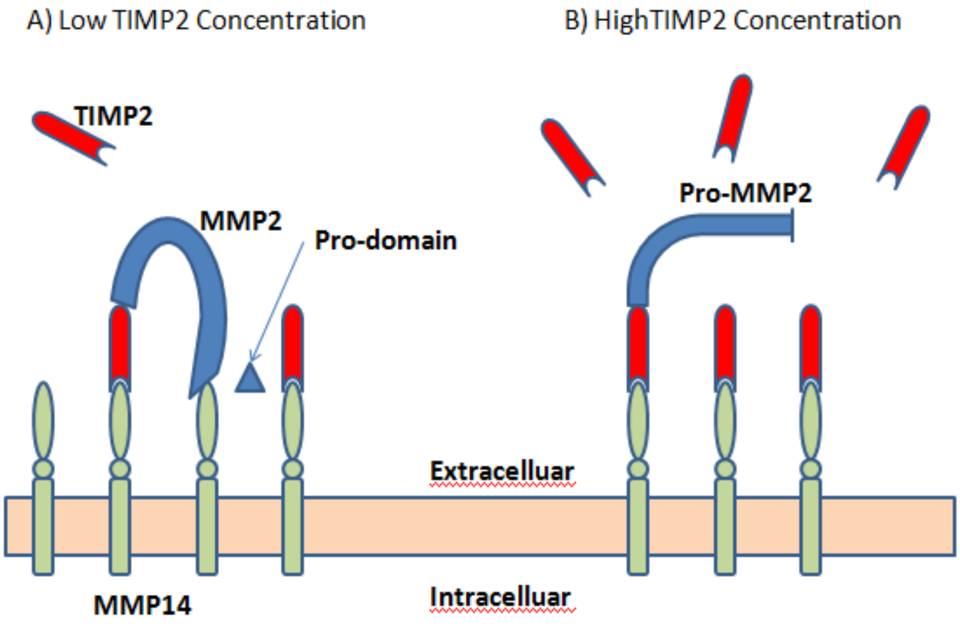


Figure E.1 Activation mechanism of Pro-MMP2.

development and invasion of tumors (5, 6). It has been shown that MMP14 contributes to the invasion of human gliomas and angiogenesis by regulating the activation of MMP2 and B (7). TIMP-2, MMP14's natural inhibitor, also plays an important role in antitumoral and antiangiogenic effects. It is involved in the regulation of matrix metalloproteinases and growth promoting activity. It was shown that overexpression of TIMP-2 was associated with a down-regulation of vascular endothelial growth factor (VEGF) expression and blood supply in the induced tumors (8).

The MMP14-TIMP2 complex is known to be involved in pro-MMP2 (progelatinase A) activation, acting as a receptor (Figure E.1). TIMP2's N-terminal domain binds to the catalytic domain of MMP14, which in turn exposes TIMP2's C-terminal domain to bind to pro-MMP2. As pro-MMP2 is attached to the complex, the cleavage of the prodomain is potentiated by active MMP14 at adjacent sites (6, 9). Overexpression of TIMP2 will induce the formation of additional MMP14-TIMP2 complexes which will increase the rate of activation of MMP2 until the activation is inhibited when TIMP2 is bound to all free MMP14s. The TIMP2 intervention in pro-MMP2 activation suggests that TIMP2 contributes concentrating and colocalizing pro-MMP2 by the active MMP14 proteins (9).

Discovery of MMP-14 inhibitors will play an important role for cancer treatment. A novel peptide (peptide G), capable to selectively inhibit the activity of MMP-14, has been reported (5). The development of MMP inhibitors, specifically antiproteolytic peptides, may have anticancer properties. Many inhibitors reaching phase I and II clinical trials have been unsuccessful because of their non-selective nature (5). Investigating the specificity of TIMP will enable the designing of selective inhibitors for therapeutic use (1).

A crystal structure of the MMP14-TIMP2 complex has been determined (10). In this paper we investigate the physicochemical basis of MMP14-TIMP2 interaction. Given the high and oppositely net charge of MMP14 and TIMP2, we focus on the role of electrostatics in binding specifically we analyze the role of ions in stabilizing the complex.

E.2 Materials and Methods

Calculations were performed using integrated Analysis of Electrostatic Similarities of Proteins (AESOP) framework to delineate the role of ions in binding (11). The coordinates of the complex were obtained from the Protein Data Bank (PDB code: 1BQQ). The calculations involved systematic removal of zinc and calcium ions, using 1, 2, 3, and 4 ions combinations. The program PDB2PQR was used to add atomic radii and partial charges to the atomic coordinate file using the CHARMM forcefield (12). The different combinations of ions of the complex consisted on having different combinations of the ions in the complex. These involve 16 complexes illustrated in Table E.1.

Complex	Name	Ions Number	Ions[a]
1	Native	4	Zn289;Zn290;Ca288;Ca291
2	Zn1Ca1Ca2	3	Zn289;Ca288;Ca291
3	Zn1Zn2Ca2	3	Zn289;Zn290;Ca291
4	Zn1Zn2Ca1	3	Zn289;Zn290;Ca288
5	Zn2Ca1Ca2	3	Zn290;Ca288;Ca291
6	Zn1Zn2	2	Zn289;Zn290
7	Zn1Ca2	2	Zn289;Ca291
8	Zn2Ca2	2	Zn290;Ca291
9	Zn2Ca1	2	Zn290;Ca288
10	Ca1Ca2	2	Ca288;Ca291
11	Zn1Ca1	2	Zn289;Ca288
12	Ca1	1	Ca288
13	Ca2	1	Ca291
14	Zn1	1	Zn289
15	Zn2	1	Zn290
16	MMP14	0	No ions

Table E.1 The 16 MMP14 complexes showing the different combinations of ions used. [a] The number corresponds to the ion number in the coordinate file.

The electrostatic potentials were calculated by numerically solving the linearized Poisson-Boltzmann equation with the program Adaptive Poisson-Boltzmann Solver (APBS) (13). The molecular (dielectric boundary) and ion accessibility surfaces were determined using spherical probes and radii set to 1.4 and 2.0 Å respectively. The APBS calculations were performed on a grid containing 129 x 129 x 129 grid points with grid dimensions of 130 Å x 122 Å x 132 Å for both 0 and 150 mM ionic strength. The dielectric coefficient for the solvent and protein interior for each complex were set to 2 and 78.54, respectively. Two electrostatic potential calculations were performed with different counter ion concentrations of 0 mM and 150 mM.

Pairwise similarities were calculated ccording to the following electrostatic similarity distance (ESD) equation,

$$ESD = \frac{1}{N}\sum_{i,j,k}\frac{|\varphi_B(i,j,k)-\varphi_A(i,j,k)|}{\max(|\varphi_B(i,j,k),\varphi_A(i,j,k)|)} \tag{1}$$

In Eq. (1), φ_A and φ_B refer to electrostatic potential at grid point (i, j, k) in proteins A and B, respectively, and N represents the total number of grid points at which electrostatic potential has been calculated. A 16 x 16 pairwise comparison of the distance matrix was generated. An ESD value of 0 denotes identical electrostatic potentials. As the ESD value increases, the dissimilarity in electrostatic potential increases (11).

The resulting complexes were clustered in a dendrogram depending on their ESD value. Free energies of association were calculated according to the thermodynamic cycle illustrated in figure E.2 (11). To

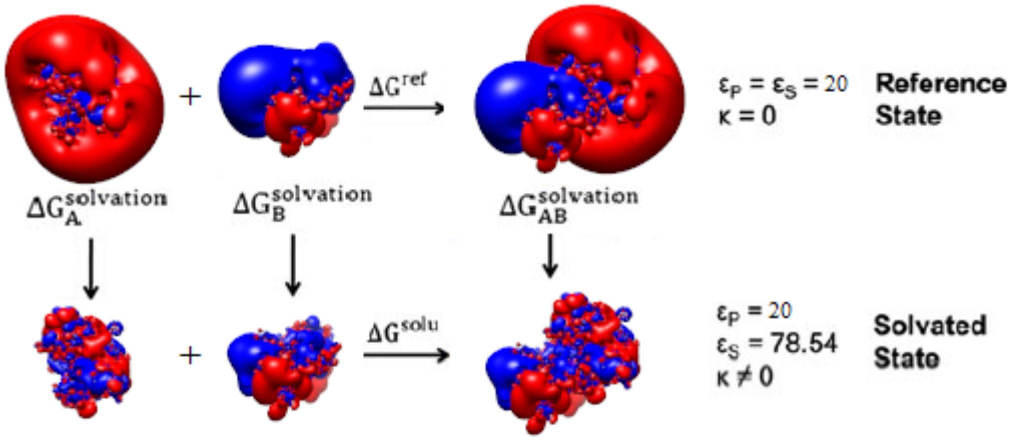


Figure E.2 Thermodynamic cycle for calculation of electrostatic free energy of association and solvation. The isopotential contours illustrated are the spatial distribution of electrostatic potential. Blue and red surfaces represent isovalues of $\pm 1.5\frac{k_B T}{e}$, respectively, where k_B is the Boltzmann constant, T is temperature, and e is the unit charge of an electron. The proteins A, B, and AB in the figure are MMP14, TIMP2, and MMP14-TIMP2 complex respectively. The top process models the reference state (dielectric coefficient of 20 for both proteins in interior and exterior and no counter ions) and the bottom process models the association step in a solvated state with dielectric coefficient of 20 for protein interior, 78.54 for protein exterior, and with or without counter ions. Three vertical processes show solvation of each free protein and the complex.

incorporate the effects of both the association and solvation, $\Delta\Delta G^{solvation}$ is calculated according to the following equation.

$$\Delta G^{solu} - \Delta G^{ref} = \Delta G_{AB}^{solvation} - \Delta G_{A}^{solvation} - \Delta G_{B}^{solvation} = \Delta\Delta G^{solvation} \qquad (2)$$

In these equations, AB refers to the protein complex MMP14-TIMP2, A refers to free MMP14, and B refers to the free inhibitor TIMP2.

E.3 Results and Discussion

We performed an electrostatic exploration of the MMP14-TIMP2 complex. Our aim was to determine the contribution of the MMP14 ions in binding with TIMP2. The sequence HELGHALGLEHS surrounding Zn289 contains three histidine residues coordinating the catalytic zinc ion and one glutamate residue, all of which are critical for catalysis (2). As shown in Table E.2 and Figure E.3, the three histidine residue distances to the catalytic zinc (zinc289) range from 1.8 to 2.4 Å and the glutamate distance is 4.8 Å

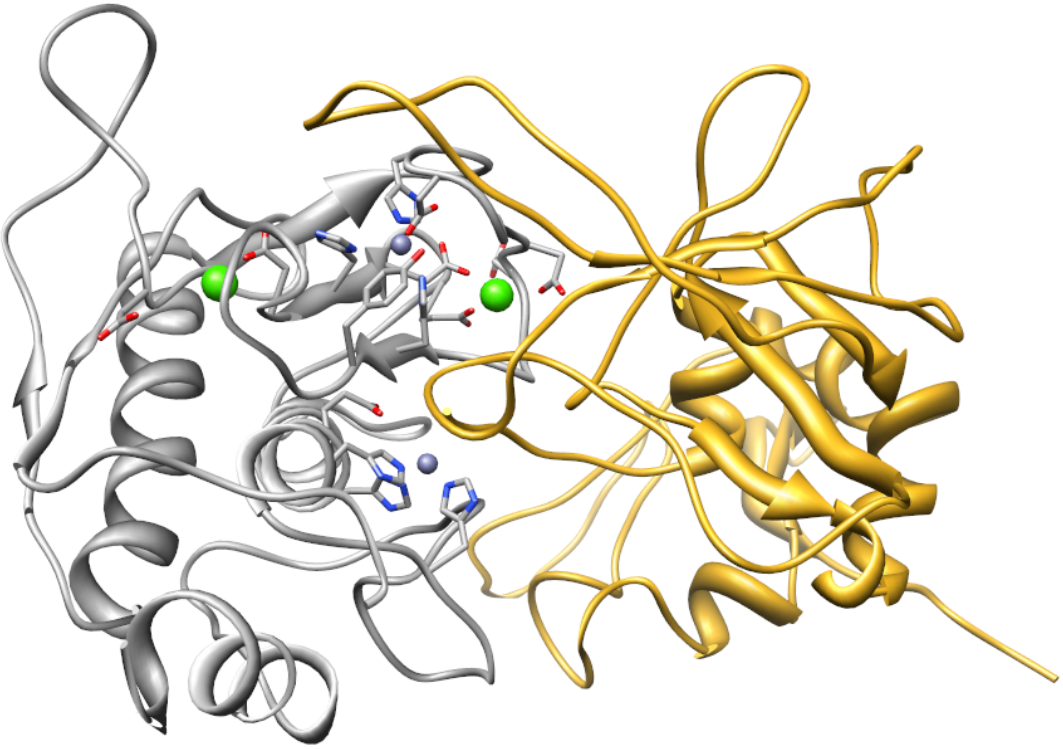


Figure E.3 Structure of MMP-14 TIMP-2 complex illustrating Zinc (blue) and Calcium ions (green).

away from the ion. The N-terminal α-amino nitrogen of the cysteine amino acid in TIMP2 also interacts with the catalytic ion at a distance of 2.9 Å (Table E.2). The structural zinc (Zn290) is tetracoordinated to three histidines and one aspartic acid in MMP14. Surrounding the structural zinc ion are three histidine amino acids at distances ranging from 1.8 to 2.6 Å and one aspartic acid at 2.9 Å. Two other calcium ions also contribute structurally to MMP14. The first calcium ion (Ca288) is coordinated by two aspartic acids 4.5 and 5.7 Å away from the ion and a tyrosine 7.2 Å away. The second calcium ion (Ca291) is coordinated by two aspartic acids 3.4 and 4 Å away and two glutamic acids 3.7 and 8.9 Å away.

Ions	amino acid	Distance (Å)
Ca288	ASP212	4.5
	ASP176	5.7
	TYR203	7.2
Ca291	ASP216	3.4
	GLU219	3.7
	ASP193	4.0
	GLU195	8.9
Zn289	HIS249	1.8
	HIS243	2.2
	HIS239	2.4
	GLU240	4.8
	CYS1001 N-term(TIMP2)[a]	2.9
Zn290	HIS186	1.8
	HIS201	2.2
	HIS214	2.6
	ASP188	2.9

Table E.2 The different ion distances to ionizable amino acids in vicinity with a cutoff distance of 9 Å. The amino acid number corresponds to the amino acid number in the PDB file code 1BQQ. [a] Zn298 interacts with TIMP2 CYS1001 N-terminus.

Using the PROPKA server, the apparent pK_a values of the ionizable amino acids were calculated and compared to the apparent pK_a of the ionizable residues in the complex without zinc and calcium ions (14-17). As illustrated in Table E.3, the apparent pK_as of the amino acids coordinating the catalytic zinc ion were lower compared to pK_a values in the complex without ions. This suggests the unfavorable coulombic interaction of the basic amino acids and the favorable coulombic interactions of the acidic amino acid with the zinc ion. Thus, the residues coordinating the catalytic zinc (Zn289) exits in stressed environment suggesting its catalytic capacity (18). Zn 289 is stabilized by the acidic amino acid GLU240. When the Zn289 is introduced to the molecule, the hisitidine pK_as drop unusually low because of the unfavorable coulombic interactions with the catalytic zinc. The coordinating histidines have unusually low pK_a values to assure neutrality at functional (physiological) pH (Table E.3).

	pK_a w/ ions	pK_a w/o Zn289	Model pK_a (PROPKA)
GLU 240	1.9	4.5	4.5
HIS 239	-1.3	2.9	6.5
HIS 243	-5	-0.8	6.5
HIS 249	-2.9	2.4	6.5

Table E.3 pK_a values of the residues coordinating the catalytic ion (Zn289) in the MMP14-TIMP2 complex with ions and without the catalytic zinc.

We performed an electrostatic clustering of the different MMP14 proteins and free energy calculations of the complexes. Figure E.4 shows that the MMP14 proteins with an equal number of ions are clustered together and have equal binding free energies when bound to TIMP2. The figure illustrates the change of binding free energy of the complexes as the number of ions in the complexes varies. The native complex, with all ions, has the highest $\Delta\Delta G^{solvation}$ suggesting loss of binding compared to the other complexes with lower number of ions. As the number of ions decreases, the $\Delta\Delta G^{solvation}$ decreases suggesting gain of binding. The gain and loss of binding can also be demonstrated by calculating the protein's net charge, which agrees with the results illustrated by the free energy calculations. The MMP14 native protein has a net charge of -7; as the number of ions in the protein decreases, the net charge decreases as well which results to a gain of binding to TIMP2 because of its positive net charge of +3. This shows that the presence of ions reduces the binding of MMP14 to TIMP2. As the number of ions in the MMP14 decreases, the free energy of binding decreases as well; illustrating a gain of binding. In fact, as the number of ions decreases the overall charge of MMP14 decreases making it more negatively charged which results in a gain of binding with the positively charged TIMP2. The ions create a less favorable environment for TIMP2 to bind. The significance of this finding is still unknown and will be the subject of future investigation.

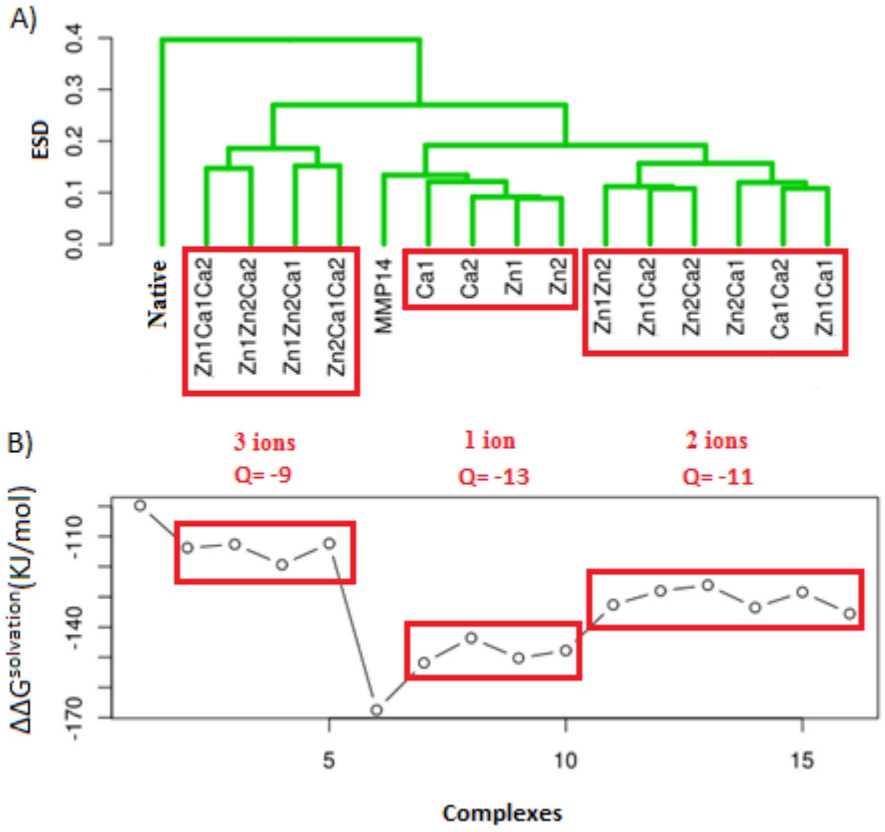


Figure E.4 (A) Electrostatic Clustering for MMP14 proteins. The dendrogram was labeled according to table 1. (B) Free energy results of MMP14-TIMP2 complexes. The order of the horizontal axis is identical to panel A. The boxes indicate complexes with the same number of ions. Net charge of each complex is shown.

Regardless of the number of ions present in MMP14 (Figure E.4A), the electrostatic potentials with the same number of ions are clustered together and their corresponding complexes have similar $\Delta\Delta G^{solvation}$. This suggests non-specific ion contribution in binding regardless of the ion's distance to the interface.

The fact that complexes with the same number of ions cluster together and have similar $\Delta\Delta G^{solvation}$ suggests that all ions contribute similarly to binding regardless of whether they are structurally or catalytically significant. In agreement with our data, TIMP2 is an inhibitor that regulates the catalytic mechanism of MMP14. Interpretation of the data of Figure E.4 suggests that upon binding of TIMP2 to MMP14 the catalytic zinc assumes a structural stability role, as the rest of the ions. This means that upon inhibition of MMP14 by TIMP2, the coordinating “catalytic” zinc loses its catalytic function and all ions become equally important by disarming the catalytic capacity of the catalytic zinc.

E.4 References

1. Butler, G. S., M. Hutton, B. A. Wattam, R. A. Williamson, V. Knauper, F. Willenbrock, and G. Murphy. 1999. The specificity of TIMP-2 for matrix metalloproteinases can be modified by single amino acid mutations. Journal of Biological Chemistry 274:20391-20396.

2. Manzetti, S., D. R. McCulloch, A. C. Herington, and D. van der Spoel. 2003. Modeling of enzyme-substrate complexes for the metalloproteases MMP-3, ADAM-9 and ADAM-10. J. Comput.-Aided Mol. Des. 17:551-565.

3. Parks, W. C., C. L. Wilson, and Y. S. Lopez-Boado. 2004. Matrix metalloproteinases as modulators of inflammation and innate immunity. Nature Reviews Immunology 4:617-629.

4. Browner, M. F., W. W. Smith, and A. L. Castelhano. 1995. Matrilysin-inhibitor complexes: common themes among metalloproteases. Biochemistry 34:6602-6610.

5. Suojanen, J., T. Salo, E. Koivunen, T. Sorsa, and E. Pirila. 2009. A novel and selective membrane type-1 matrix metalloproteinase (MT1-MMP) inhibitor reduces cancer cell motility and tumor growth. Cancer Biol. Ther. 8:2362-2370.

6. Sounni, N. E., M. Janssen, J. M. Foidart, and A. Noel. 2003. Membrane type-1 matrix metalloproteinase and TIMP-2 in tumor angiogenesis. Matrix biology : journal of the International Society for Matrix Biology 22:55-61.

7. Forsyth, P. A., H. Wong, T. D. Laing, N. B. Rewcastle, D. G. Morris, H. Muzik, K. J. Leco, R. N. Johnston, P. M. Brasher, G. Sutherland, and D. R. Edwards. 1999. Gelatinase-A (MMP-2), gelatinase-B (MMP-9) and membrane type matrix metalloproteinase-1 (MT1-MMP) are involved in different aspects of the pathophysiology of malignant gliomas. British journal of cancer 79:1828-1835.

8. Hajitou, A., N. E. Sounni, L. Devy, C. Grignet-Debrus, J. M. Lewalle, H. Li, C. F. Deroanne, H. Lu, A. Colige, B. V. Nusgens, F. Frankenne, A. Maron, P. Yeh, M. Perricaudet, Y. Chang, C. Soria, C. M. Calberg-Bacq, J. M. Foidart, and A. Noel. 2001. Down-regulation of vascular endothelial growth factor by tissue inhibitor of metalloproteinase-2: effect on in vivo mammary tumor growth and angiogenesis. Cancer research 61:3450-3457.

9. Butler, G. S., M. J. Butler, S. J. Atkinson, H. Will, T. Tamura, S. Schade van Westrum, T. Crabbe, J. Clements, M. P. d'Ortho, and G. Murphy. 1998. The TIMP2 membrane type 1 metalloproteinase "receptor" regulates the concentration and efficient activation of progelatinase A. A kinetic study. The Journal of biological chemistry 273:871-880.

10. Fernandez-Catalan, C., W. Bode, R. Huber, D. Turk, J. J. Calvete, A. Lichte, H. Tschesche, and K. Maskos. 1998. Crystal structure of the complex formed by the membrane type 1-matrix metalloproteinase with the tissue inhibitor of metalloproteinases-2, the soluble progelatinase A receptor. EMBO Journal 17:5238-5248.

11. Gorham, R. D., Jr., C. A. Kieslich, and D. Morikis. 2012. Complement Inhibition by Staphylococcus aureus: Electrostatics of C3d-EfbC and C3d-Ehp Association. Cellular and Molecular Bioengineering 5:32-43.

12. Dolinsky, T. J., J. E. Nielsen, J. A. McCammon, and N. A. Baker. 2004. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Research 32:W665-W667.

13. Baker, N. A., D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. U. S. A. 98:10037-10041.

14. Bas, D. C., D. M. Rogers, and J. H. Jensen. 2008. Very fast prediction and rationalization of pKa values for protein-ligand complexes. Proteins: Structure, Function, and Bioinformatics 73:765-783.

15. Li, H., A. D. Robertson, and J. H. Jensen. 2005. Very fast empirical prediction and rationalization of protein pKa values. Proteins: Structure, Function, and Bioinformatics 61:704-721.

16. Olsson, M. H. M., C. R. Sondergaard, M. Rostkowski, and J. H. Jensen. 2011. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. Journal of Chemical Theory and Computation 7:525-537.

17. Sondergaard, C. R., M. H. M. Olsson, M. Rostkowski, and J. H. Jensen. 2011. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. Journal of Chemical Theory and Computation 7:2284-2295.

18. Nielsen, J. E., and J. A. McCammon. 2003. Calculating pKa values in enzyme active sites. Protein Science 12:1894-1901.

APPENDIX F: INPUT FILES FOR SIMULATIONS AND SCRIPTS FOR DATA ANALYSIS

F.1 Side Chain Distance Time Series Calculation

```
import MDAnalysis
import MDAnalysis.analysis.distances
import numpy

cutoff = 5
start_frame=0
resID_difference = 3  ## means 4 or more
PDB = "/home/ziedgaieb/Documents/python_scripts_development/CCR7_CCL19_charged-N-
term_Xray_110-110-130_analysis/ccr7_ccl19.pdb"
DCD = "/home/ziedgaieb/Documents/python_scripts_development/CCR7_CCL19_charged-N-
term_Xray_110-110-130_analysis/ccr7_ccl19_7ms_wrapped_unwrapped.dcd"

u = MDAnalysis.Universe(PDB,DCD)


## get the contacts and make timeseries out of
selectionarg_sidechain_vdw_head = "((resname ALA and name CB) or \
(resname ARG and (name CZ or name NE or name NH1 or name NH2)) or \
(resname ASN and (name CG or name OD1 or name ND2)) or \
(resname ASP and (name CG or name OD1 or name OD2)) or \
(resname CYS and name SG) or \
(resname GLN and (name CD or name OE1 or name NE2)) or \
(resname GLU and (name CD or name OE1 or name OE2)) or \
(resname GLY and name CA) or \
(resname HIS and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSE and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSD and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSP and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname ILE and (name CG1 or name CG2 or name CD)) or \
(resname LEU and (name CG or name CD1 or name CD2)) or \
(resname LYS and name NZ) or \
(resname MET and (name SD or name CE)) or \
(resname PHE and (name CG or name CD1 or name CE1 or name CZ or name CE2 or name CD2)) or
\
(resname PRO and name CG) or \
(resname SER and name OG) or \
(resname THR and (name CB or name CG2 or name OG1)) or \
(resname TRP and (name CE2 or name CD2 or name CE3 or name CZ3 or name CH2 or name CZ2))
or \
(resname TYR and name OH) or \
(resname VAL and (name CB or name CG1 or name CG2)))"


selectionarg_sidechain_polar = "((resname ARG and (name CZ or name NE or name NH1 or name
NH2)) or \
(resname ASN and (name CG or name OD1 or name ND2)) or \
(resname ASP and (name CG or name OD1 or name OD2)) or \
(resname CYS and name SG) or \
(resname GLN and (name CD or name OE1 or name NE2)) or \
(resname GLU and (name CD or name OE1 or name OE2)) or \
(resname HIS and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSE and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSD and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSP and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname LYS and name NZ) or \
(resname SER and name OG) or \
(resname THR and name OG1) or \
(resname TRP and name NE1) or \
(resname TYR and name OH))"

selectionarg_sidechain_ele = "((resname ARG and (name CZ or name NE or name NH1 or name
NH2)) or \
(resname ASP and (name CG or name OD1 or name OD2)) or \
(resname GLU and (name CD or name OE1 or name OE2)) or \
(resname HIS and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSE and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSD and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSP and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname LYS and name NZ))"

selectionarg_sidechain_aromatic = "((resname HIS and (name CG or name ND1 or name CE1 or
name NE2 or name CD2)) or \
(resname HSE and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSD and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname HSP and (name CG or name ND1 or name CE1 or name NE2 or name CD2)) or \
(resname PHE and (name CG or name CD1 or name CE1 or name CZ or name CE2 or name CD2)) or
\
(resname TRP and (name CG or name CD1 or name NE1 or name CE2 or name CD2 or name CE3 or
name CZ3 or name CH2 or name CZ2)) or \
(resname TYR and (name CG or name CD1 or name CE1 or name CZ or name CE2 or name CD2)))"

selectionarg_sidechain = "not backbone"

selection_calpha = "name CA"

selectionarg = selectionarg_sidechain_polar

selectionarg1 = selectionarg + " and segid B and resid 26:304 and not name H*"
selectionarg2 = selectionarg + " and segid B and resid 26:304 and not name H*"
selection1 = u.select_atoms(selectionarg1)
selection2 = u.select_atoms(selectionarg2)
dmin                                                                                 =
MDAnalysis.analysis.distances.distance_array(selection1.get_positions(),selection2.get_po
sitions()) #selection1 rows selection2 col
loading = 0
print "> 0 percent of", u.trajectory.n_frames, "frames processed"
for ts in u.trajectory[start_frame:u.trajectory.n_frames]:
    d                                                                                =
MDAnalysis.analysis.distances.distance_array(selection1.get_positions(),selection2.get_po
sitions()) #selection1 rows selection2 col
    dmin = numpy.minimum(dmin, d)
    if        (((ts.frame*100)/u.trajectory.n_frames)%10        ==        0)        &
(((ts.frame*100)/u.trajectory.n_frames) != loading):
        print     ">",     (ts.frame*100)/u.trajectory.n_frames,     "percent     of",
u.trajectory.n_frames, "frames processed"
    loading = ((ts.frame*100)/u.trajectory.n_frames)
```

```
if selectionarg1==selectionarg2:
    dmin = numpy.triu(dmin)

index = numpy.where((dmin < cutoff) & (dmin > 0)) #row1 is rows (selection1) row2 is cols
(selection2)


#'''
## Forming an array with contact names: resid1 resnum1 atmname1 chain1 resid2 resnum2
atmname2 chain2
contact1 =
numpy.vstack((selection1[index[0]].resnames,selection1[index[0]].resids,selection1[index[
0]].names,selection1[index[0]].segids))
contact2 =
numpy.vstack((selection2[index[1]].resnames,selection2[index[1]].resids,selection2[index[
1]].names,selection2[index[1]].segids))
## create a contact log to write to file by creating two column with the following string
"resid1-resnum1-chain1"
contactsatomlog1 = [a + b + c + d + e + f + g for a, b, c, d, e, f, g in zip(
map(str,contact1[0]), ["-"]*numpy.shape(contact1[0])[0], map(str,contact1[1]), ["-
"]*numpy.shape(contact1[0])[0], map(str,contact1[2]), ["-"]*numpy.shape(contact1[0])[0],
map(str,contact1[3]))] #pasting the three rows with a seperation of "-"
contactsatomlog2 = [a + b + c + d + e + f + g for a, b, c, d, e, f, g in zip(
map(str,contact2[0]), ["-"]*numpy.shape(contact2[0])[0], map(str,contact2[1]), ["-
"]*numpy.shape(contact2[0])[0], map(str,contact2[2]), ["-"]*numpy.shape(contact2[0])[0],
map(str,contact2[3]))] #pasting the three rows with a seperation of "-"
contactsatomlog = numpy.vstack((contactsatomlog1,contactsatomlog2))

#'''

## Forming an array with contact names: resid1 resnum1 chain1 resid2 resnum2 chain2
contact1 =
numpy.vstack((selection1[index[0]].resnames,selection1[index[0]].resids,selection1[index[
0]].segids))
contact2 =
numpy.vstack((selection2[index[1]].resnames,selection2[index[1]].resids,selection2[index[
1]].segids))
contacts = numpy.vstack((contact1,contact2))
## create a contact log to write to file by creating two column with the following string
"resid1-resnum1-chain1"
contactslog1 = [a + b + c + d + e for a, b, c, d, e in zip( map(str,contact1[0]), ["-
"]*numpy.shape(contact1[0])[0], map(str,contact1[1]), ["-"]*numpy.shape(contact1[0])[0],
map(str,contact1[2]))] #pasting the three rows with a seperation of "-"
contactslog2 = [a + b + c + d + e for a, b, c, d, e in zip( map(str,contact2[0]), ["-
"]*numpy.shape(contact2[0])[0], map(str,contact2[1]), ["-"]*numpy.shape(contact2[0])[0],
map(str,contact2[2]))] #pasting the three rows with a seperation of "-"
contactslog = numpy.vstack((contactslog1,contactslog2))


## Removing self-interacting residue contacts eg: GLN 1 B interacting with GLN 1 B;
(these contacts exists cause we do not discriminate between two atoms within the same
residue interacting)
#indkeep_tmp = numpy.where((contactslog[0]==contactslog[1])==False)[0]
indkeep =
numpy.where(numpy.absolute(numpy.subtract(contacts[1,:].astype(numpy.float),contacts[4,:]
.astype(numpy.float)))>resID_difference)[0]
contacts = contacts[:,indkeep]
```

```
contactslog = contactslog[:,indkeep]
contactsatomlog =contactsatomlog[:,indkeep]

## Calculating Distance time series for each contact
## Readjusting the selection using the variables 'index' and 'indkeep'
selection11 = selection1[index[0]][indkeep] ## this selection includes repeats cause the
array of contact repeats atoms and residues (since one atoms can have multiple contacts)
selection22 = selection2[index[1]][indkeep] ## this selection includes repeats cause the
array of contact repeats atoms and residues (since one atoms can have multiple contacts)

distmatrix = []
loading = 0
print "> 0 percent of", u.trajectory.n_frames, "frames processed"
for ts in u.trajectory[start_frame:u.trajectory.n_frames]:
    d = MDAnalysis.analysis.distances.dist(selection11,selection22) #selection1 rows
selection2 col
    distmatrix.append(numpy.round(d[2,:],3))
    if (((ts.frame*100)/u.trajectory.n_frames)%10 == 0) &
(((ts.frame*100)/u.trajectory.n_frames) != loading):
        print ">", (ts.frame*100)/u.trajectory.n_frames, "percent of",
u.trajectory.n_frames, "frames processed"
    loading = ((ts.frame*100)/u.trajectory.n_frames)


distmatrix = numpy.transpose(numpy.asarray(distmatrix)) #rounding the matrix to save
memory
numpy.savetxt(fname='distance_matrix_intra_chainB_allFrames_resIDdiff_3_polar_ccl19.txt',
X=distmatrix,fmt='%.3f',delimiter=",")
residuecontactsatomlog = numpy.asarray([a + b + c for a, b, c in zip(
map(str,contactsatomlog[0]), ["::"]*numpy.shape(contactsatomlog[0])[0],
map(str,contactsatomlog[1]))])
residuecontactsatomlog = residuecontactsatomlog.reshape((1,len(residuecontactsatomlog)))
numpy.savetxt(fname='distance_matrix_intra_chainB_allFrames_resIDdiff_3_polar_ccl19.log',
X=residuecontactsatomlog,fmt='%s',delimiter=",")


## Finding the distmatrix per residue partners (minimum inter-atom distance of all atoms
per residue partners)

## Removing repeated contacts from contacts only (not contactslog) => to have only unique
residue-residue contacts
residuecontactslog = [a + b + c for a, b, c in zip( map(str,contactslog[0]),
["::"]*numpy.shape(contactslog[0])[0], map(str,contactslog[1]))]
unsort = numpy.unique(residuecontactslog,return_index=True)[1]
residuecontactslog = [residuecontactslog[i] for i in sorted(unsort)]
residuecontactslog = numpy.asarray([i.split('::') for i in residuecontactslog]) ## here
the contacts are switched to be organized in rows (unlike contactslog and contacts
variables)
contact1 = numpy.transpose([i.split('-') for i in residuecontactslog[:,0]]) ## now
they are switched to be organized in columns
contact2 = numpy.transpose([i.split('-') for i in residuecontactslog[:,1]])
contacts = numpy.vstack((contact1,contact2))


# this line is based on contactslog; this variable have been used cause it followed the
selection variable Readjusted using the variables 'index' and 'indkeep'
indicesmin = numpy.unique([a + b + c for a, b, c in zip( map(str,contactslog[0]),
["::"]*numpy.shape(contactslog[0])[0], map(str,contactslog[1]))],return_inverse=True)[1]
unsort = numpy.unique(indicesmin,return_index=True)[1]
distmatrixmin = []
loading = 0
counter=0
print "> 0 percent of", len(indicesmin[numpy.sort(unsort)]), "frames processed"
for ii in indicesmin[numpy.sort(unsort)]:
    counter+=1
    ind = numpy.where(indicesmin == ii)[0]
    distmatrixtmp = distmatrix[ind,:]
    distmatrixmin.append(distmatrixtmp.min(axis=0))
    if (((counter*100)/len(indicesmin[numpy.sort(unsort)]))%10 == 0) & (((counter*100)/len(indicesmin[numpy.sort(unsort)])) != loading):
        print ">", (counter*100)/len(indicesmin[numpy.sort(unsort)]), "percent of", len(indicesmin[numpy.sort(unsort)]), "contacts processed"
    loading=((counter*100)/len(indicesmin[numpy.sort(unsort)]))
    print ii

distmatrixmin = numpy.asarray(distmatrixmin)

numpy.savetxt(fname='distance_matrix_min_intra_chainB_allFrames_resIDdiff_3_polar_ccl19.txt',X=distmatrixmin,fmt='%.3f',delimiter=",")

contactslog1 = [a + b + c + d + e for a, b, c, d, e in zip( map(str,contact1[0]), ["-"]*numpy.shape(contact1[0])[0], map(str,contact1[1]), ["-"]*numpy.shape(contact1[0])[0], map(str,contact1[2]))] #pasting the three rows with a seperation of "-"
contactslog2 = [a + b + c + d + e for a, b, c, d, e in zip( map(str,contact2[0]), ["-"]*numpy.shape(contact2[0])[0], map(str,contact2[1]), ["-"]*numpy.shape(contact2[0])[0], map(str,contact2[2]))] #pasting the three rows with a seperation of "-"
contactslog = numpy.vstack((contactslog1,contactslog2))

residuecontactsatomlog = numpy.asarray([a + b + c + d  for a, b, c, d in zip( map(str,contactslog[0]), ["::"]*numpy.shape(contactslog[0])[0], map(str,contactslog[1]), ["_sidechain"]*numpy.shape(contactslog[0])[0])])
residuecontactsatomlog = residuecontactsatomlog.reshape((1,len(residuecontactsatomlog)))
numpy.savetxt(fname='distance_matrix_min_intra_chainB_allFrames_resIDdiff_3_polar_ccl19.log',X=residuecontactsatomlog,fmt='%s',delimiter=",")




##
```

F.2 Backbone Distance Time Series Calculation

```
import MDAnalysis
import MDAnalysis.analysis.distances
import numpy

cutoff = 15
start_frame=11000
resID_difference = 3  ## means 4 or more
PDB = "/home/ziedgaieb/Documents/python_scripts_development/CCR7_CCL21_charged-N-term_Xray_110-110-130_analysis/ccr7_ccl21.pdb"
DCD = "/home/ziedgaieb/Documents/python_scripts_development/CCR7_CCL21_charged-N-term_Xray_110-110-130_analysis/ccr7_ccl21_7ms_wrapped_unwrapped.dcd"
```

```
u = MDAnalysis.Universe(PDB,DCD)


##  get the contacts and make timeseries out of
selectionarg1 = "name CA and segid B and resid 26:304 and not name H*"
selectionarg2 = "name CA and segid B and resid 26:304 and not name H*"
selection1 = u.select_atoms(selectionarg1)
selection2 = u.select_atoms(selectionarg2)
dmin = MDAnalysis.analysis.distances.distance_array(selection1.get_positions(),selection2.get_positions()) #selection1 rows selection2 col
loading = 0
print "> 0 percent of", u.trajectory.n_frames, "frames processed"
for ts in u.trajectory[start_frame:u.trajectory.n_frames]:
    d = MDAnalysis.analysis.distances.distance_array(selection1.get_positions(),selection2.get_positions()) #selection1 rows selection2 col
    dmin = numpy.minimum(dmin, d)
    if (((ts.frame*100)/u.trajectory.n_frames)%10 == 0) & (((ts.frame*100)/u.trajectory.n_frames) != loading):
        print ">", (ts.frame*100)/u.trajectory.n_frames, "percent of", u.trajectory.n_frames, "frames processed"
    loading = ((ts.frame*100)/u.trajectory.n_frames)


if selectionarg1==selectionarg2:
    dmin = numpy.triu(dmin)

index = numpy.where((dmin < cutoff) & (dmin > 0)) #row1 is rows (selection1) row2 is cols (selection2)

#'''
## Forming an array with contact names: resid1 resnum1 atmname1 chain1 resid2 resnum2 atmname2 chain2
contact1 = numpy.vstack((selection1[index[0]].resnames,selection1[index[0]].resids,selection1[index[0]].names,selection1[index[0]].segids))
contact2 = numpy.vstack((selection2[index[1]].resnames,selection2[index[1]].resids,selection2[index[1]].names,selection2[index[1]].segids))
## create a contact log to write to file by creating two column with the following string "resid1-resnum1-chain1"
contactsatomlog1 = [a + b + c + d + e + f + g for a, b, c, d, e, f, g in zip( map(str,contact1[0]), ["-"]*numpy.shape(contact1[0])[0], map(str,contact1[1]), ["-"]*numpy.shape(contact1[0])[0], map(str,contact1[2]), ["-"]*numpy.shape(contact1[0])[0], map(str,contact1[3]))] #pasting the three rows with a seperation of "-"
contactsatomlog2 = [a + b + c + d + e + f + g for a, b, c, d, e, f, g in zip( map(str,contact2[0]), ["-"]*numpy.shape(contact2[0])[0], map(str,contact2[1]), ["-"]*numpy.shape(contact2[0])[0], map(str,contact2[2]), ["-"]*numpy.shape(contact2[0])[0], map(str,contact2[3]))] #pasting the three rows with a seperation of "-"
contactsatomlog = numpy.vstack((contactsatomlog1,contactsatomlog2))

#'''

## Forming an array with contact names: resid1 resnum1 chain1 resid2 resnum2 chain2
contact1 = numpy.vstack((selection1[index[0]].resnames,selection1[index[0]].resids,selection1[index[
0]].segids))
contact2 =
numpy.vstack((selection2[index[1]].resnames,selection2[index[1]].resids,selection2[index[
1]].segids))
contacts = numpy.vstack((contact1,contact2))
## create a contact log to write to file by creating two column with the following string
"resid1-resnum1-chain1"
contactslog1 = [a + b + c + d + e for a, b, c, d, e in zip( map(str,contact1[0]), ["-
"]*numpy.shape(contact1[0])[0], map(str,contact1[1]), ["-"]*numpy.shape(contact1[0])[0],
map(str,contact1[2]))] #pasting the three rows with a seperation of "-"
contactslog2 = [a + b + c + d + e for a, b, c, d, e in zip( map(str,contact2[0]), ["-
"]*numpy.shape(contact2[0])[0], map(str,contact2[1]), ["-"]*numpy.shape(contact2[0])[0],
map(str,contact2[2]))] #pasting the three rows with a seperation of "-"
contactslog = numpy.vstack((contactslog1,contactslog2))


## Removing self-interacting residue contacts eg: GLN 1 B interacting with GLN 1 B;
(these contacts exists cause we do not discriminate between two atoms within the same
residue interacting)
#indkeep_tmp = numpy.where((contactslog[0]==contactslog[1])==False)[0]
indkeep =
numpy.where(numpy.absolute(numpy.subtract(contacts[1,:].astype(numpy.float),contacts[4,:]
.astype(numpy.float)))>resID_difference)[0]
contacts = contacts[:,indkeep]
contactslog = contactslog[:,indkeep]
contactsatomlog =contactsatomlog[:,indkeep]

## Calculating Distance time series for each contact
## Readjusting the selection using the variables 'index' and 'indkeep'
selection11 = selection1[index[0]][indkeep] ## this selection includes repeats cause the
array of contact repeats atoms and residues (since one atoms can have multiple contacts)
selection22 = selection2[index[1]][indkeep] ## this selection includes repeats cause the
array of contact repeats atoms and residues (since one atoms can have multiple contacts)

distmatrix = []
loading = 0
print "> 0 percent of", u.trajectory.n_frames, "frames processed"
for ts in u.trajectory[start_frame:u.trajectory.n_frames]:
    d = MDAnalysis.analysis.distances.dist(selection11,selection22) #selection1 rows
selection2 col
    distmatrix.append(numpy.round(d[2,:],3))
    if (((ts.frame*100)/u.trajectory.n_frames)%10 == 0) &
(((ts.frame*100)/u.trajectory.n_frames) != loading):
        print ">", (ts.frame*100)/u.trajectory.n_frames, "percent of",
u.trajectory.n_frames, "frames processed"
    loading = ((ts.frame*100)/u.trajectory.n_frames)


distmatrix = numpy.transpose(numpy.asarray(distmatrix)) #rounding the matrix to save
memory
numpy.savetxt(fname='distance_matrix_intra_chainB_eq_resIDdiff_3.txt',X=distmatrix,fmt='%
.3f',delimiter=",")
residuecontactsatomlog = numpy.asarray([a + b + c for a, b, c in zip(
map(str,contactsatomlog[0]), ["::"]*numpy.shape(contactsatomlog[0])[0],
map(str,contactsatomlog[1]))])
residuecontactsatomlog = residuecontactsatomlog.reshape((1,len(residuecontactsatomlog)))
numpy.savetxt(fname='distance_matrix_intra_chainB_eq_resIDdiff_3.log',X=residuecontactsat
omlog,fmt='%s',delimiter=",")


## Finding the distmatrix per residue partners (minimum inter-atom distance of all atoms
per residue partners)

## Removing repeated contacts from contacts only (not contactslog) => to have only unique
residue-residue contacts
residuecontactslog = [a + b + c for a, b, c in zip( map(str,contactslog[0]),
["::"]*numpy.shape(contactslog[0])[0], map(str,contactslog[1]))]
unsort = numpy.unique(residuecontactslog,return_index=True)[1]
residuecontactslog = [residuecontactslog[i] for i in sorted(unsort)]
residuecontactslog = numpy.asarray([i.split('::') for i in residuecontactslog]) ## here
the contacts are switched to be organized in rows (unlike contactslog and contacts
variables)
contact1 = numpy.transpose([i.split('-') for i in residuecontactslog[:,0]])        ## now
they are switched to be organized in columns
contact2 = numpy.transpose([i.split('-') for i in residuecontactslog[:,1]])
contacts = numpy.vstack((contact1,contact2))


# this line is based on contactslog; this variable have been used cause it followed the
selection variable Readjusted using the variables 'index' and 'indkeep'
indicesmin = numpy.unique([a + b + c for a, b, c in zip( map(str,contactslog[0]),
["::"]*numpy.shape(contactslog[0])[0], map(str,contactslog[1]))],return_inverse=True)[1]
unsort = numpy.unique(indicesmin,return_index=True)[1]
distmatrixmin = []
loading = 0
counter=0
print "> 0 percent of", len(indicesmin[numpy.sort(unsort)]), "frames processed"
for ii in indicesmin[numpy.sort(unsort)]:
    counter+=1
    ind = numpy.where(indicesmin == ii)[0]
    distmatrixtmp = distmatrix[ind,:]
    distmatrixmin.append(distmatrixtmp.min(axis=0))
    if (((counter*100)/len(indicesmin[numpy.sort(unsort)]))%10 == 0) &
(((counter*100)/len(indicesmin[numpy.sort(unsort)])) != loading):
        print ">", (counter*100)/len(indicesmin[numpy.sort(unsort)]), "percent of",
len(indicesmin[numpy.sort(unsort)]), "contacts processed"
    loading=((counter*100)/len(indicesmin[numpy.sort(unsort)]))
    print ii

distmatrixmin = numpy.asarray(distmatrixmin)

numpy.savetxt(fname='distance_matrix_min_intra_chainB_eq_resIDdiff_3.txt',X=distmatrixmin
,fmt='%.3f',delimiter=",")

contactslog1 = [a + b + c + d + e for a, b, c, d, e in zip( map(str,contact1[0]), ["-
"]*numpy.shape(contact1[0])[0], map(str,contact1[1]), ["-"]*numpy.shape(contact1[0])[0],
map(str,contact1[2]))] #pasting the three rows with a seperation of "-"
contactslog2 = [a + b + c + d + e for a, b, c, d, e in zip( map(str,contact2[0]), ["-
"]*numpy.shape(contact2[0])[0], map(str,contact2[1]), ["-"]*numpy.shape(contact2[0])[0],
map(str,contact2[2]))] #pasting the three rows with a seperation of "-"
contactslog = numpy.vstack((contactslog1,contactslog2))

residuecontactsatomlog = numpy.asarray([a + b + c for a, b, c in zip(
map(str,contactslog[0]),                     ["::"]*numpy.shape(contactslog[0])[0],
map(str,contactslog[1]))])
residuecontactsatomlog = residuecontactsatomlog.reshape((1,len(residuecontactsatomlog)))
numpy.savetxt(fname='distance_matrix_min_intra_chainB_eq_resIDdiff_3.log',X=residuecontac
tsatomlog,fmt='%s',delimiter=",")




##
```

F.3 Detection of Side Chain Contact Rearrangements from MD Simulations

```
## Fitting gaussian distributions to the distance time series

def multipeak_gaussfit(oneDdata,num_bins = 500,outdir='.',filename='namewithoutext'):
    import matplotlib
    matplotlib.use('Agg')
    import matplotlib.pyplot as plt
    import matplotlib.mlab as mlab
    from scipy import optimize
    from scipy import signal
    from peakutils import peak
    from scipy import ndimage

    if not os.path.exists(outdir):
        os.makedirs(outdir)


    num_bins=500
    y_real, bins, patches = plt.hist(oneDdata, num_bins, normed=1, facecolor='green')
    plt.close()

    ##Peak detection using simple method after smoothing the curve (CWT from scipy was
not appropriate)
    #window_size=num_bins/100
    window_size=num_bins/50
    y_real_filtered = ndimage.filters.gaussian_filter(y_real, window_size)
    #window_size, poly_order = (num_bins/10)+1, 3
    #y_real_filtered = signal.savgol_filter(y_real, window_size, poly_order)
    nbre_peaks = len(peak.indexes(y_real_filtered,thres=0.05,min_dist=0))
    #peaks = (diff(sign(diff(y_real_filtered))) < 0).nonzero()[0] +1


    #Performing BIG score per Gaussian Mixture Model
    from sklearn import mixture
    test = oneDdata.reshape((len(oneDdata),1))

    n_components_range = range(1, nbre_peaks+4)
    #cv_types = ['spherical', 'tied', 'diag', 'full']
    cv_types = ['diag']
    lowest_bic = numpy.infty
    bic = []
    for cv_type in cv_types:
        for n_components in n_components_range:
            # Fit a mixture of Gaussians with EM
            gmm = mixture.GMM(n_components=n_components, covariance_type=cv_type)
            gmm.fit(test)
        bic.append(gmm.bic(test))
        if bic[-1] < lowest_bic:
            lowest_bic = bic[-1]
            best_gmm = gmm
            best_n_components = n_components
    clf = best_gmm
    clf.fit(test)


    y_est1 = clf.weights_[0] * mlab.normpdf(bins, clf.means_[0],
numpy.sqrt(clf.covars_[0]))

    filename1 =
filename+'_nbrePeaks_'+str(best_n_components)+'_weight_sum_GMM_minBIC.png'
    savepath = os.path.join(outdir, filename1)
    plt.plot(bins[1:501], y_real, label = 'Real Data')
    plt.plot(bins, y_est1, 'g.', label = 'Fitted')
    if best_n_components > 1:
        for i in range(best_n_components)[1:best_n_components+1]:
            y_est_tmp = clf.weights_[i] * mlab.normpdf(bins, clf.means_[i],
numpy.sqrt(clf.covars_[i]))
            y_est1 += clf.weights_[i] * mlab.normpdf(bins, clf.means_[i],
numpy.sqrt(clf.covars_[i]))
            plt.plot(bins, y_est_tmp, 'g.', label = 'Fitted')
    plt.plot(bins, y_est1, 'g.', label = 'Fitted')
    plt.xlabel('Smarts')
    plt.ylabel('Probability Density')
    plt.title('Histogram of contact ...')
    plt.savefig(savepath)
    plt.close()


    #Overlapping probability of two normal distribution with scipy
    #Organizing the distribution in a matrix (should I use weights or not?)
    from scipy import stats
    n_components = clf.n_components

    def solve(wgth1,wgth2,m1,m2,std1,std2):
        a = 1/(2*std1**2) - 1/(2*std2**2)
        b = m2/(std2**2) - m1/(std1**2)
        c = m1**2 /(2*std1**2) - m2**2 / (2*std2**2) -
numpy.log((std2*wgth1)/(std1*wgth2))
        return numpy.roots([a,b,c])

    overlapping_propabilities = numpy.zeros(shape=(n_components,n_components))
    for ii in range(n_components):
        for jj in range(ii,n_components):
            if ii == jj:
                continue
            m1, m2 = clf.means_[ii][0], clf.means_[jj][0]
            std1, std2 = numpy.sqrt(clf.covars_[ii])[0], numpy.sqrt(clf.covars_[jj])[0]
            wgth1 , wgth2 = clf.weights_[ii], clf.weights_[jj]
            max1, max2 = numpy.max(wgth1 * mlab.normpdf(bins, m1, std1)), numpy.max(wgth2
* mlab.normpdf(bins, m2, std2))
            intersect = numpy.unique(solve(wgth1,wgth2,m1,m2,std1,std2))
            r = intersect[numpy.logical_and((intersect>numpy.min(bins)),
(intersect<numpy.max(bins)))]
```

```
            if any(isinstance(t, complex) for t in r):
                overlapping_propabilities[ii,jj] = 100
                continue
            if len(r)==1:
                area = stats.norm.cdf(r,m2,std2) + (1.-stats.norm.cdf(r,m1,std1))
                if m1 > m2:
                    area = stats.norm.cdf(r,m1,std1) + (1.-stats.norm.cdf(r,m2,std2))
            else:
                if len(r)==2:
                    area        =        stats.norm.cdf(r[0],m2,std2)        +        (1-
stats.norm.cdf(r[1],m2,std2))          +          stats.norm.cdf(r[1],m1,std1)          -
stats.norm.cdf(r[0],m1,std1)
                    if max1 > max2:
                        area       =       stats.norm.cdf(r[0],m1,std1)       +       (1-
stats.norm.cdf(r[1],m1,std1))          +          stats.norm.cdf(r[1],m2,std2)          -
stats.norm.cdf(r[0],m2,std2)
                else:
                    print(len(r))
            overlapping_propabilities[ii,jj] = area*100

    ## write the percentages to a file and choose percentage cutoff to combine Gaussians
    overlap_area = 50
    inds_orig = numpy.asarray(numpy.where(overlapping_propabilities>overlap_area))
    inds = numpy.asarray(numpy.where(overlapping_propabilities>overlap_area))

    '''
    double check the areas calculated
    '''

    groups = []

    for i in range(inds.shape[1]):
        groups_tmp = [inds[0,0],inds[1,0]]
        inds = numpy.delete(inds,(0), axis=1)
        counter = 0
        inds_groups = []
        inds_groups_prev = [0]
        while not len(inds_groups_prev)==len(inds_groups):
            inds_groups_prev = inds_groups[:]
            for ii in range(len(groups_tmp)):
                if (groups_tmp[ii] in inds[0]):
                    inds_tmp = numpy.where(inds[0]==groups_tmp[ii])[0].tolist()
                    inds_groups.extend(inds_tmp)
                if (groups_tmp[ii] in inds[1]):
                    inds_tmp = numpy.where(inds[1]==groups_tmp[ii])[0].tolist()
                    inds_groups.extend(inds_tmp)
            inds_groups = numpy.unique(inds_groups).tolist()
            for ii in range(len(inds_groups)):
                groups_tmp.extend(inds[:,inds_groups[ii]].tolist())
            groups_tmp = numpy.unique(groups_tmp).tolist()
            counter =+1
        groups.append(groups_tmp[:])
        if not (len(inds_groups)==inds.shape[1]):
            inds = numpy.delete(inds,(inds_groups), axis=1)
        else:
            if ((len(inds_groups)+1)==inds.shape[1]):
                groups.append([inds[0,0],inds[1,0]])
                break
            else:
                break



    #combine gaussians with a common area of more than 50%

    gaussians_reduced = numpy.unique(numpy.hstack(inds_orig))
    n_components_reduced = n_components - len(gaussians_reduced) + len(groups)
    distribution_params = numpy.zeros(shape=(n_components_reduced,3))
    jj=0
    for ii in range(n_components):
        if not (ii in gaussians_reduced):
            distribution_params[jj,0] = clf.weights_[ii]
            distribution_params[jj,1] = clf.means_[ii]
            distribution_params[jj,2] = numpy.sqrt(clf.covars_[ii])[0]
            jj = jj + 1
        else:
            continue

    for ii in range(len(groups)):
        weights_tmp = 0
        means_tmp = 0
        covars_tmp = 0
        for zz in range(len(groups[ii])):
            weights_tmp = weights_tmp + clf.weights_[groups[ii][zz]]
            means_tmp = means_tmp + (clf.weights_[groups[ii][zz]] /
numpy.sum(clf.weights_[groups[ii][:]])) * clf.means_[groups[ii][zz]]
            covars_tmp = covars_tmp + (clf.weights_[groups[ii][zz]] /
numpy.sum(clf.weights_[groups[ii][:]])) * clf.covars_[groups[ii][zz]]
        distribution_params[jj,0] = weights_tmp
        distribution_params[jj,1] = means_tmp
        distribution_params[jj,2] = numpy.sqrt(covars_tmp)[0]
        jj = jj + 1



    y_est1 = distribution_params[0,0] * mlab.normpdf(bins, distribution_params[0,1],
distribution_params[0,2])

    filename1 =
filename+'_nbrePeaks_'+str(n_components_reduced)+'_weight_sum_GMM_minBIC_reduced.png'
    savepath = os.path.join(outdir, filename1)
    plt.plot(bins[1:501], y_real, color='0.75', label = 'Real Data')
    plt.plot(bins, y_est1, 'b.', label = 'Fitted1')
    if n_components_reduced > 1:
        for i in range(n_components_reduced)[1:n_components_reduced+1]:
            y_est_tmp = distribution_params[i,0] * mlab.normpdf(bins,
distribution_params[i,1], distribution_params[i,2])
            y_est1 += distribution_params[i,0] * mlab.normpdf(bins,
distribution_params[i,1], distribution_params[i,2])
            plt.plot(bins, y_est_tmp, 'b.', label = 'Fitted1')
    #plt.plot(bins, y_est1, 'b.', label = 'Fitted1')
    #plt.legend()
    plt.xlabel('Distances')
    plt.ylabel('Probability Density')
    #plt.title('Histogram of contact ...')
    plt.savefig(savepath)
    plt.close()

    ## Return all weights, means, and stdvs (17 Feb 16)
    ## RUN WITH 1 STDEV
    params = []
    for ii in range(distribution_params.shape[0]):
        params.append([distribution_params[ii,0], distribution_params[ii,1],
distribution_params[ii,2]])
        dist_frames = numpy.where((oneDdata <
distribution_params[ii,1]+1*distribution_params[ii,2]) & (oneDdata >
distribution_params[ii,1]-1*distribution_params[ii,2]))[0]
        params[ii].append(dist_frames)


    return params,n_components_reduced


    #Performing Dirichlet Process Gaussian Mixture Model (Explore more another time)
    from sklearn import mixture
    test = oneDdata.reshape((len(oneDdata),1))

    nbre_peaks = nbre_peaks + 3
    clf = mixture.DPGMM(n_components=nbre_peaks, alpha=100., n_iter=100)
    clf.fit(test)
    Y_ = clf.predict(test)

    y_est1 = numpy.zeros(num_bins+1).reshape(1,num_bins+1)
    if numpy.any(Y_ == 0):
        y_est1 += clf.weights_[0] * mlab.normpdf(bins, clf.means_[0],
numpy.sqrt(clf._get_covars()[0]))

    filename1 = filename+'_nbrePeaks_'+str(len(numpy.unique(Y_)))+'_weight_sum_DPGMM.png'
    savepath = os.path.join(outdir, filename1)
    plt.plot(bins[1:501], y_real, label = 'Real Data')
    plt.plot(bins, y_est1[0,:], 'g.', label = 'Fitted')
    if nbre_peaks > 1:
        for i in range(nbre_peaks)[1:nbre_peaks+1]:
            if not numpy.any(Y_ == i):
                continue
            y_est_tmp = clf.weights_[i] * mlab.normpdf(bins, clf.means_[i],
numpy.sqrt(clf._get_covars()[i]))
            y_est1 += clf.weights_[i] * mlab.normpdf(bins, clf.means_[i],
numpy.sqrt(clf._get_covars()[i]))
            plt.plot(bins, y_est_tmp[0,:], 'g.', label = 'Fitted')
    plt.plot(bins, y_est1[0,:], 'g.', label = 'Fitted')
    plt.legend()
    plt.xlabel('Smarts')
    plt.ylabel('Probability Density')
    plt.title('Histogram of contact ...')
    plt.savefig(savepath)
    plt.close()
```

```
##
## Fitting gaussian distributions to the distance time series
import os
import numpy

from numpy import genfromtxt
#from numpy import *
#from scipy.stats import norm


data =
genfromtxt("/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1a
pr16/side-
chain/1_contacts/distance_matrix_min_intra_chainB_eq_resIDdiff_3_polar.txt",delimiter =
",")
dataLog =
genfromtxt("/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1a
pr16/side-
chain/1_contacts/distance_matrix_min_intra_chainB_eq_resIDdiff_3_polar.log",dtype='S',del
imiter = ",")

start_frame = 0
timeseries_parameters = []
inds = []
n_distributions = []
for i in range(data.shape[0]):
    print(i)
    data_eg = data[i,start_frame:data.shape[1]]
    if numpy.min(data_eg) > 5:
        continue
    contact_name = dataLog[i]
    params,n_distributions_tmp = multipeak_gaussfit(oneDdata=data_eg, num_bins = 500,
outdir='./contact_polar_distribution_figures_eq_resIDdiff_cutoff50_eq_1stdev',
filename=str(i)+'_'+contact_name)
    timeseries_parameters.append(params)
    inds.extend([i])
    n_distributions.append(n_distributions_tmp)

numpy.savetxt(fname='contact_sidechain_distribution_parameters_eq_resIDdiff_cutoff50_fram
es_1stdev.log',X=dataLog[inds],fmt='%s',delimiter=",")


#saving parameters to a file
my_file =
open("contact_sidechain_distribution_parameters_eq_resIDdiff_cutoff50_frames_1stdev.txt",
"w")
for i in range(len(timeseries_parameters)):
    for j in range(len(timeseries_parameters[i])):
        if not (j == 0):
            my_file.write(",")
        my_file.write( "{:10.4f}".format(timeseries_parameters[i][j][0]) + "," +
"{:10.4f}".format(timeseries_parameters[i][j][1]) + "," +
"{:10.4f}".format(timeseries_parameters[i][j][2]) + "," + "-
".join(map(str,timeseries_parameters[i][j][3])))
    my_file.write( "\n")

my_file.close()
```

```
#extract the multimodal distributions as time series and log file
inds_multimodel = numpy.where(numpy.asarray(n_distributions)>1)[0]



numpy.savetxt(fname='distance_matrix_min_intra_chainB_eq_resIDdiff_3_sidechain_multimodal
.txt',X=data[inds,:][inds_multimodel,:],fmt='%.3f',delimiter=",")
residuelog                                                                          =
numpy.asarray(dataLog[inds][inds_multimodel]).reshape((1,len(dataLog[inds][inds_multimode
l])))
numpy.savetxt(fname='distance_matrix_min_intra_chainB_eq_resIDdiff_3_sidechain_multimodal
.log',X=residuelog,fmt='%s',delimiter=",")
##
residuelog                                                                          =
numpy.asarray(dataLog[inds][inds_multimodel]).reshape((1,len(dataLog[inds][inds_multimode
l])))



## Mapping into the PDB of each group

# Reading PDB
import MDAnalysis

PDB      =      "/home/ziedgaieb/Documents/python_scripts_development/CCR7_CCL21_charged-N-
term_Xray_110-110-130_analysis/ccr7_ccl21.pdb"

#Residues are ordered in alphabetic order following the charmm forcefield topology file
selectionarg1 = "((resname ALA and name CB) or \
(resname ARG and name CZ) or \
(resname ASP and name CG) or \
(resname ASN and name CG) or \
(resname CYS and name SG) or \
(resname GLN and name CD) or \
(resname GLU and name CD) or \
(resname GLY and name CA) or \
(resname HIS and name CG) or \
(resname HSE and name CG) or \
(resname HSD and name CG) or \
(resname HSP and name CG) or \
(resname ILE and name CG1) or \
(resname LEU and name CG) or \
(resname LYS and name NZ) or \
(resname MET and name SD) or \
(resname PHE and name CG) or \
(resname PRO and name CG) or \
(resname SER and name OG) or \
(resname THR and name CB) or \
(resname TRP and name CE2) or \
(resname TYR and name CG) or \
(resname VAL and name CB))"

#selection1 = u.select_atoms(selectionarg1)

dist_parameters_array = numpy.asarray(residuelog)
```

```
outdirS = './'
outdir                                                                                  =
'contact_sidechain_distribution_figures_eq_resIDdiff_cutoff50_eq_1stdev_multimodalPDB'
outdirgroup = os.path.join(outdirS, outdir)
if not os.path.exists(outdirgroup):
    os.makedirs(outdirgroup)
edges = numpy.asarray([jj.split("_")[0] for jj in dist_parameters_array[0]])
edges = numpy.asarray([jj.split("::") for jj in edges])
filename = '~molecular-switches_edges'+'.pdb'
savepath = os.path.join(outdirgroup, filename)
pdb = MDAnalysis.Writer(savepath, multiframe=True)
for j in range(edges.shape[0]):
    u = MDAnalysis.Universe(PDB)
    resnum1 = edges[j][0].split("-")[1]
    chain1 = edges[j][0].split("-")[2]
    resnum2 = edges[j][1].split("-")[1]
    chain2 = edges[j][1].split("-")[2]
    selection = selectionarg1 + " and ((resid " + resnum1 + " and segid " + chain1 + ")
or (resid " + resnum2 + " and segid " + chain2 + "))"
    protein = u.select_atoms(selection)
    if (abs(int(resnum1) - int(resnum2))==1):
        protein.set_resids((int(resnum1),(int(resnum2) + 1)))
    protein.set_names(("CA","CA"))
    pdb.write(protein)
pdb.close()
```

```
#
```

F.4 Detection of Large Domain Motions Through DCCM

```
##efficient reading of large files
import numpy as np

def generate_text_file(length=1e6, ncols=20):
    data = np.random.random((length, ncols))
    np.savetxt('large_text_file.csv', data, delimiter=',')

def iter_loadtxt(filename, delimiter=',', skiprows=0, dtype=float):
    def iter_func():
        with open(filename, 'r') as infile:
            for _ in range(skiprows):
                next(infile)
            for line in infile:
                line = line.rstrip().split(delimiter)
                for item in line:
                    yield dtype(item)
        iter_loadtxt.rowlength = len(line)

    data = np.fromiter(iter_func(), dtype=dtype)
    data = data.reshape((-1, iter_loadtxt.rowlength))
    return data

#generate_text_file()
#data = iter_loadtxt('large_text_file.csv')
```

```
## DCCM and Clustering
import os
import numpy

from numpy import genfromtxt
#from numpy import *
#from scipy.stats import norm



data_org = [numpy.array(map(float, line.split(","))) for line in
open('/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1apr16/c
-alpha/1_contacts/distance_matrix_min_intra_chainB_eq_resIDdiff_3.txt')]
data_org = numpy.asarray(data_org)
#data_org                                                                                =
numpy.loadtxt("/home/ziedgaieb/Documents/python_scripts_development/contact_distributions
_1apr16/c-alpha/1_contacts/distance_matrix_min_intra_chainB_eq_resIDdiff_3.txt",delimiter
= ",")
dataLog                                                                                  =
genfromtxt("/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1a
pr16/c-
alpha/1_contacts/distance_matrix_min_intra_chainB_eq_resIDdiff_3.log",dtype='S',delimiter
= ",")




#hierarchical clustering and dendogram
import matplotlib
matplotlib.use('Agg')
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster, cut_tree
from scipy.spatial.distance import pdist

## unecassary ==> remove when cleaning the code
data = data_org
#corrmatrix = numpy.corrcoef(data)
#numpy.fill_diagonal(corrmatrix, 0)

Z = linkage(data, 'single', 'correlation')

cutoff = 0.95
##group the data into clusters
##cutoff is at a correlation of 0.95
groups = fcluster(Z,t=1-cutoff,criterion='distance')
timeseries_number_cluster = numpy.unique(groups,return_counts=True)
corr_indices_groups = []
for group in timeseries_number_cluster[0][numpy.where(timeseries_number_cluster[1]>1)]:
    corr_indices_groups.extend(numpy.where(groups==group))

outdirS = './groups_'+str(cutoff)
if not os.path.exists(outdirS):
    os.makedirs(outdirS)
```

```
for i in range(len(corr_indices_groups)):
    outdir = 'group_'+str(i)
    outdirgroup = os.path.join(outdirS, outdir)
    if not os.path.exists(outdirgroup):
        os.makedirs(outdirgroup)
    for j in range(len(corr_indices_groups[i])):
        ## plotting the contact timeseries for each group
        filename = dataLog[corr_indices_groups[i][j]]+'_distr_'+str(j)+'.png'
        savepath = os.path.join(outdirgroup, filename)
        plt.figure(figsize=(40,12))
        plt.plot(data[corr_indices_groups[i][j],:], color= 'orange',
marker='o',markersize=3,markeredgecolor='none',ls='')
        plt.rc("font",size=32)
        plt.savefig(savepath,dpi=100)
        plt.close()
    filename = '~group_'+str(i)+'_DCCM'+'.png'
    savepath = os.path.join(outdirgroup, filename)
    colors = [(matplotlib.cm.jet(ii)) for ii in xrange(1,256)]
    new_map = matplotlib.colors.LinearSegmentedColormap.from_list('new_map', colors,
N=256)
    plt.figure(figsize=(30,25))
    plt.pcolor(numpy.corrcoef(data[corr_indices_groups[i],:]), cmap=new_map, vmin=0,
vmax=1)
    plt.colorbar()
    plt.savefig(savepath)
    plt.close()

cutoff = 0.95
##group the data into clusters
##cutoff is at a correlation of 0.95
groups = fcluster(Z,t=1-cutoff,criterion='distance')
timeseries_number_cluster = numpy.unique(groups,return_counts=True)
corr_indices = []
for group in timeseries_number_cluster[0][numpy.where(timeseries_number_cluster[1]>1)]:
    corr_indices.extend(numpy.where(groups==group)[0])


##Recalculating the dendogram with only the groups of interest
Z_G = linkage(data[corr_indices,:], 'single', 'correlation')

corrmatrix_G = numpy.corrcoef(data[corr_indices,:])
numpy.fill_diagonal(corrmatrix_G, 0)
#corrmatrix_G = corrmatrix[numpy.ix_(list(corr_indices),list(corr_indices))]

'''
#Calculating the number of groups for a series of cutoffs
group_numbers=[]
for cutoff in numpy.arange(0,1,0.05):
    group_numbers.append(len(numpy.unique(fcluster(Z,t=cutoff,criterion='distance'))))
'''

#writing the correlated groups to file
numpy.savetxt(fname='distance_matrix_min_intra_chainB_eq_resIDdiff_3_DCCM_0.95.txt',X=dat
a[corr_indices,:],fmt='%.3f',delimiter=",")
residuelog = numpy.asarray(dataLog[corr_indices]).reshape((1,len(dataLog[corr_indices])))
```

```
numpy.savetxt(fname='distance_matrix_min_intra_chainB_eq_resIDdiff_3_DCCM_0.95.log',X=res
iduelog,fmt='%s',delimiter=",")


#plotting the correlation matrix
colors = [(matplotlib.cm.jet(i)) for i in xrange(1,256)]
new_map = matplotlib.colors.LinearSegmentedColormap.from_list('new_map', colors, N=256)
plt.figure(figsize=(30,25))
plt.pcolor(corrmatrix_G, cmap=new_map, vmin=-1, vmax=1)
plt.colorbar()
plt.savefig("DCCM_Groups"+str(cutoff)+".png",dpi=100)
plt.close()



#import sys
#sys.setrecursionlimit(10000)
#dendrogram
plt.figure(figsize=(25,10))
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('sample index')
plt.ylabel('distance')
dend = dendrogram(Z_G, color_threshold = 0.05)
plt.savefig("figure00_G.png")
plt.close()



##plotting clustering alongside the corr matrix
import pylab
fig = pylab.figure(figsize=(100,100))
ax1 = fig.add_axes([0.09,0.1,0.2,0.6])
Z1 = dendrogram(Z_G, orientation='right', color_threshold = 0.05)
ax1.set_xticks([])
ax1.set_yticks([])

ax2 = fig.add_axes([0.3,0.71,0.6,0.2])
Z2 = dendrogram(Z_G, color_threshold = 0.05)
ax1.set_xticks([])
ax1.set_yticks([])

axmatrix = fig.add_axes([0.3,0.1,0.6,0.6])
idx1 = Z1['leaves']
idx2 = Z2['leaves']
corrmatrix_G_tmp = corrmatrix_G[idx1,:]
corrmatrix_G_tmp = corrmatrix_G_tmp[:,idx2]
colors = [(matplotlib.cm.jet(i)) for i in xrange(1,256)]
new_map = matplotlib.colors.LinearSegmentedColormap.from_list('new_map', colors, N=256)
im  =  axmatrix.matshow(corrmatrix_G_tmp,  aspect='auto',  origin='lower',  cmap=new_map,
vmin=-1, vmax=1)
axmatrix.set_xticks([])
axmatrix.set_yticks([])

axcolor = fig.add_axes([0.91,0.1,0.02,0.6])
pylab.colorbar(im, cax=axcolor)
fig.savefig("Dendogram_DCCM_Groups"+str(cutoff)+".png",dpi=200)
```

```
## Mapping into the PDB of each group

# Reading PDB
import MDAnalysis

PDB = "/home/ziedgaieb/Documents/python_scripts_development/CCR7_CCL21_charged-N-
term_Xray_110-110-130_analysis/ccr7_ccl21.pdb"

#Residues are ordered in alphabetic order following the charmm forcefield topology file

selectionarg1 = "((resname ALA and name CA) or \
(resname ARG and name CA) or \
(resname ASP and name CA) or \
(resname ASN and name CA) or \
(resname CYS and name CA) or \
(resname GLN and name CA) or \
(resname GLU and name CA) or \
(resname GLY and name CA) or \
(resname HIS and name CA) or \
(resname HSE and name CA) or \
(resname HSD and name CA) or \
(resname HSP and name CA) or \
(resname ILE and name CA) or \
(resname LEU and name CA) or \
(resname LYS and name CA) or \
(resname MET and name CA) or \
(resname PHE and name CA) or \
(resname PRO and name CA) or \
(resname SER and name CA) or \
(resname THR and name CA) or \
(resname TRP and name CA) or \
(resname TYR and name CA) or \
(resname VAL and name CA))"

#selection1 = u.select_atoms(selectionarg1)

dist_parameters_array = numpy.asarray(dataLog)

for i in range(len(corr_indices_groups)):
    outdir = 'group_'+str(i)
    outdirgroup = os.path.join(outdirS, outdir)
    if not os.path.exists(outdirgroup):
        os.makedirs(outdirgroup)
    edges = numpy.asarray([jj.split("::") for jj in
dist_parameters_array[corr_indices_groups[i]]])
    filename = '~group_'+str(i)+'_edges'+'.pdb'
    savepath = os.path.join(outdirgroup, filename)
    pdb = MDAnalysis.Writer(savepath, multiframe=True)
    for j in range(edges.shape[0]):
        u = MDAnalysis.Universe(PDB)
        resnum1 = edges[j][0].split("-")[1]
        chain1 = edges[j][0].split("-")[2]
        resnum2 = edges[j][1].split("-")[1]
        chain2 = edges[j][1].split("-")[2]
        selection = selectionarg1 + " and ((resid " + resnum1 + " and segid " + chain1 +
") or (resid " + resnum2 + " and segid " + chain2 + "))"
        protein = u.select_atoms(selection)
        if (abs(int(resnum1) - int(resnum2))==1):
            protein.set_resids((int(resnum1),(int(resnum2) + 1)))
        protein.set_names(("CA","CA"))
        pdb.write(protein)
    pdb.close()



##
```

F.5 Network of the Protein's Dynamical Components

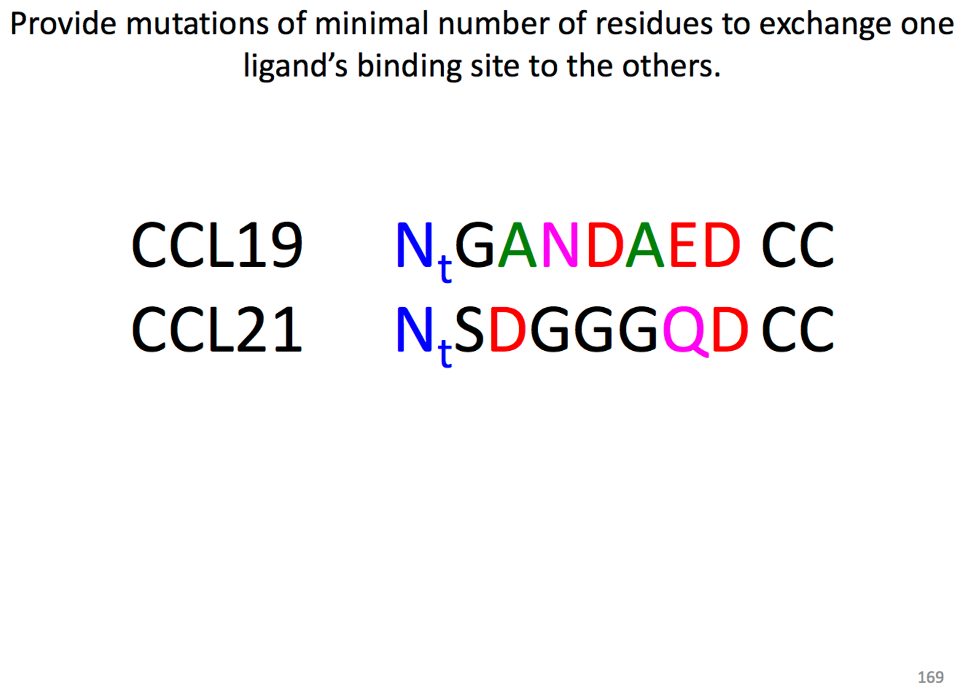
```
## DCCM and Clustering
import os
import numpy

from numpy import genfromtxt
#from numpy import *
#from scipy.stats import norm



data_sidechain = [numpy.array(map(float, line.split(","))) for line in
open('/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1apr16/s
ide-
chain/2_contacts_distributions/distance_matrix_min_intra_chainB_eq_resIDdiff_3_sidechain_
multimodal.txt')]
data_sidechain = numpy.asarray(data_sidechain)
dataLog_sidechain =
genfromtxt("/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1a
pr16/side-
chain/2_contacts_distributions/distance_matrix_min_intra_chainB_eq_resIDdiff_3_sidechain_
multimodal.log",dtype='S',delimiter = ",")

data_polar = [numpy.array(map(float, line.split(","))) for line in
open('/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1apr16/s
ide-
chain/2_contacts_distributions/distance_matrix_min_intra_chainB_eq_resIDdiff_3_polar_mult
imodal.txt')]
data_polar = numpy.asarray(data_polar)
dataLog_polar =
genfromtxt("/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1a
pr16/side-
chain/2_contacts_distributions/distance_matrix_min_intra_chainB_eq_resIDdiff_3_polar_mult
imodal.log",dtype='S',delimiter = ",")

data_calpha = [numpy.array(map(float, line.split(","))) for line in
open('/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1apr16/c
-alpha/2_contacts_DCCM/distance_matrix_min_intra_chainB_eq_resIDdiff_3_DCCM_0.95.txt')]
data_calpha = numpy.asarray(data_calpha)
dataLog_calpha =
genfromtxt("/home/ziedgaieb/Documents/python_scripts_development/contact_distributions_1a
pr16/c-
alpha/2_contacts_DCCM/distance_matrix_min_intra_chainB_eq_resIDdiff_3_DCCM_0.95.log",dtyp
e='S',delimiter = ",")
```

```
data = numpy.vstack((data_sidechain,data_polar,data_calpha))
dataLog = numpy.hstack((dataLog_sidechain,dataLog_polar,dataLog_calpha))

data_sidechain_polar = numpy.vstack((data_sidechain,data_polar))
dataLog_sidechain_polar = numpy.hstack((dataLog_sidechain,dataLog_polar))

#hierarchical clustering and dendogram
import matplotlib
matplotlib.use('Agg')
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster, cut_tree
from scipy.spatial.distance import pdist

## unecassary ==> remove when cleaning the code
#data = data_org
#corrmatrix = numpy.corrcoef(data)
#numpy.fill_diagonal(corrmatrix, 0)


Z = linkage(data_calpha, 'single', 'correlation')



cutoff = 0.95
##group the data into clusters
##cutoff is at a correlation of 0.95
groups = fcluster(Z,t=1-cutoff,criterion='distance')
timeseries_number_cluster = numpy.unique(groups,return_counts=True)
corr_indices_groups = []
for group in timeseries_number_cluster[0][numpy.where(timeseries_number_cluster[1]>1)]:
    corr_indices_groups.extend(numpy.where(groups==group))


groupdataLog = [a + b for a, b in zip(["group_"]*len(corr_indices_groups),
map(str,range(len(corr_indices_groups))))]
group_capacity = []
dcc_weights = numpy.zeros((len(corr_indices_groups),len(dataLog_sidechain_polar)))
counter = 0
for i in range(len(corr_indices_groups)):
    group_capacity.append(len(corr_indices_groups[i]))
    for j in range(len(dataLog_sidechain_polar)):
        if (counter%5000 == 0):
            print ">", counter, "out of",
str(len(corr_indices_groups)*len(dataLog_sidechain_polar)), "contacts processed"
        counter += 1
        ccc =
numpy.corrcoef(data_calpha[corr_indices_groups[i],:],data_sidechain_polar[j,:])
        ccc_average =
numpy.average(ccc[len(corr_indices_groups[i]),0:len(corr_indices_groups[i])])
        dcc_weights[i,j] = ccc_average


dcc_weights_groups = numpy.zeros((len(corr_indices_groups),len(corr_indices_groups)))
counter = 0
for i in range(len(corr_indices_groups)):
    for j in range(i,len(corr_indices_groups)):
        if (counter%5000 == 0):
            print ">", counter, "out of",
str(len(corr_indices_groups)*len(corr_indices_groups)), "contacts processed"
        counter += 1
        ccc = numpy.corrcoef(data_calpha[corr_indices_groups[i],:])
        ccc_average = numpy.average(numpy.tril(ccc,-1))
        dcc_weights_groups[i,j] = ccc_average
        dcc_weights_groups[j,i] = ccc_average

dcc_weights_switches = numpy.corrcoef(data_sidechain_polar)
numpy.fill_diagonal(dcc_weights_switches, 0)




## generate matrices to input to Gephi
## matrix of node1 node2 edgeWeight
## matrix with nodes nodeWeight
cutoff = 0.75
network = numpy.asarray(['node1','node2','weight'])
counter = 0
for i in range(dcc_weights.shape[0]):
    for j in range(dcc_weights.shape[1]):
        if (counter%5000 == 0):
            print ">", counter, "out of", str(dcc_weights.shape[0]*dcc_weights.shape[1]),
"contacts processed"
        counter += 1
        if abs(dcc_weights[i,j]) < cutoff:
            continue
        network =
numpy.vstack((network,[groupdataLog[i],dataLog_sidechain_polar[j],abs(dcc_weights[i,j])])
)
#           network = numpy.vstack((network,[groupdataLog[i],dataLog_sidechain_polar[j],-
numpy.log(abs(dcc_weights[i,j]))]))


counter = 0
for i in range(dcc_weights_groups.shape[0]):
    for j in range(i+1,dcc_weights_groups.shape[1]):
        if (counter%5000 == 0):
            print ">", counter, "out of",
str(dcc_weights_groups.shape[0]*dcc_weights_groups.shape[1]), "contacts processed"
        counter += 1
        if abs(dcc_weights_groups[i,j]) < cutoff:
            continue
        network =
numpy.vstack((network,[groupdataLog[i],groupdataLog[j],abs(dcc_weights_groups[i,j])]))
#                   network = numpy.vstack((network,[groupdataLog[i],groupdataLog[j],-
numpy.log(abs(dcc_weights_groups[i,j]))]))



counter = 0
for i in range(dcc_weights_switches.shape[0]):
    for j in range(i+1,dcc_weights_switches.shape[1]):
        if (counter%5000 == 0):
            print ">", counter, "out of",
str(dcc_weights_switches.shape[0]*dcc_weights_switches.shape[1]/2), "contacts processed"
```

```
        counter += 1
        if abs(dcc_weights_switches[i,j]) < cutoff or abs(dcc_weights_switches[i,j]) ==
1:
            continue
        network                                                                           =
numpy.vstack((network,[dataLog_sidechain_polar[i],dataLog_sidechain_polar[j],abs(dcc_weig
hts_switches[i,j])]))
#                                                                        network          =
numpy.vstack((network,[dataLog_sidechain_polar[i],dataLog_sidechain_polar[j],-
numpy.log(abs(dcc_weights_switches[i,j]))]))

network = numpy.delete(network,(0),axis=0)
## network matrix created with node1 node2 edge_weight

## generate the number of contact in each node
nodes = numpy.asarray(['node','size'])
for i in range(len(groupdataLog)):
    nodes = numpy.vstack((nodes,[groupdataLog[i],group_capacity[i]]))

for i in range(len(dataLog_sidechain_polar)):
    nodes = numpy.vstack((nodes,[dataLog_sidechain_polar[i],1]))

nodes = numpy.delete(nodes,(0),axis=0)
## nodes matrix created with node nbre_contacts

vertices = []
edges = network[:,0:2].tolist()
for line in edges:
    vertices.extend(line)


vertices = numpy.unique(numpy.asarray(vertices))

import igraph
G = igraph.Graph()
G.add_vertices(vertices)
G.add_edges(edges)
#option2: this option works on weight only and not any other edge attribute
#once the graph is weighted (as done in the next line), we can weight the edges as done
below
G.es['weight'] = 1
for i in range(len(edges)):
    line = edges[i]
    G[line[0],line[1]] = network[i,2]


#G.vs['nbre_contacts'] = 1
#for i in range(len(vertices)):
#    G.vs[i]['nbre_contacts'] = nodes[numpy.where(nodes[:,0]==G.vs[i]['name'])[0],1][0]

for i, v in enumerate(G.vs):
    v['nbre_contacts'] = str(nodes[numpy.where(nodes[:,0]==G.vs[i]['name'])[0],1][0])

##Writing Network into file
G.write(f="0_fullnetwork_nbreNodes_"+str(len(G.vs['name']))+".graphml",format="graphml")
#
```

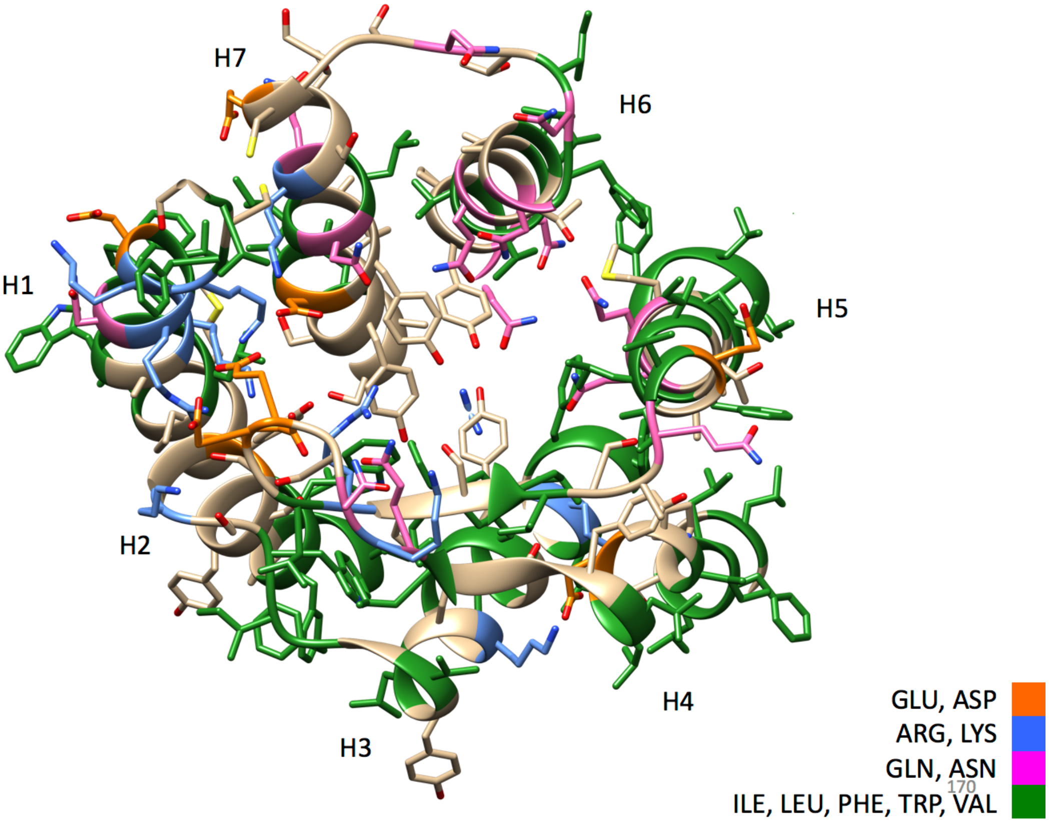
```
network_remaining = network
nodes_remaining = nodes


node_communities =  numpy.asarray(['node','community'])
membership_max_previous = -1
counter = 1
while counter <7:
    print(str(counter))
    vertices = []
    edges = network_remaining[:,0:2].tolist()
    for line in edges:
        vertices.extend(line)
    #
    vertices = numpy.unique(numpy.asarray(vertices))
    #
    import igraph
    G = igraph.Graph()
    G.add_vertices(vertices)
    G.add_edges(edges)
    #option2: this option works on weight only and not any other edge attribute
    #once the graph is weighted (as done in the next line), we can weight the edges as
done below
    G.es['weight'] = 1
    for i in range(len(edges)):
        line = edges[i]
        G[line[0],line[1]] = network_remaining[i,2]
    #
    #G.vs['nbre_contacts'] = 1
    #for i in range(len(vertices)):
    #                                                   G.vs[i]['nbre_contacts']            =
nodes[numpy.where(nodes[:,0]==G.vs[i]['name'])[0],1][0]
    #
    for i, v in enumerate(G.vs):
        v['nbre_contacts']                                                                  =
str(nodes_remaining[numpy.where(nodes_remaining[:,0]==G.vs[i]['name'])[0],1][0])
    ##extracting the largest fully connected network from the full network
    G_subnetwork = G.clusters().giant()
    ##Community Mapping
    #calculate dendrogram
    #communities  =  G_subnetwork.community_optimal_modularity()    ===>  this  crashed  my
large network
    #communities = G_subnetwork.community_optimal_modularity()
    communities = G_subnetwork.community_edge_betweenness(directed=False)
    #convert  it  into  a  flat  clustering  ===>  doesn't  work  unless  the  network  is  fully
connected
    clusters = communities.as_clustering()
    #get the membership vector
    membership = clusters.membership
    #Add the numer of communities from the previous network, So community membership dont
overlap with previous subnetwork
    membership = [x+(membership_max_previous + 1) for x in membership]
    G_subnetwork.vs['membership'] = membership
    membership_max_previous = numpy.max(membership)
    node_communities                                                                        =
numpy.vstack((node_communities,numpy.column_stack((G_subnetwork.vs['name'],
membership))))
    #
    ##Writing Network into file

G_subnetwork.write(f=str(counter)+"_subnetwork_nbreNodes_"+str(len(G_subnetwork.vs['name'
]))+".graphml",format="graphml")
    #
    ##Removing all the nodes used in the G_subnetwork from the network matrix
    for i in range(len(G_subnetwork.vs['name'])):
        indices = []
        node = G_subnetwork.vs['name'][i]
        indices.extend(numpy.where(node==network_remaining[:,0])[0])
        indices.extend(numpy.where(node==network_remaining[:,1])[0])
        network_remaining = numpy.delete(network_remaining,(indices),axis=0)
    #
    ##Removing all the nodes used in the G_subnetwork from the nodes matrix
    for i in range(len(G_subnetwork.vs['name'])):
        indices = []
        node = G_subnetwork.vs['name'][i]
        indices.extend(numpy.where(node==nodes_remaining[:,0])[0])
        indices.extend(numpy.where(node==nodes_remaining[:,1])[0])
        nodes_remaining = numpy.delete(nodes_remaining,(indices),axis=0)
    #
    counter += 1




## Outputting each community to a folder with: PDB file mapping of each node; Time series
of each node
## Using the following variables: node_communities and corr_indices_groups (each index i
contains a list of indices of timeseries belonging to group_i)
## and dataLog_calpha
membership = numpy.unique(node_communities[1:,1].astype(numpy.integer))
outdirSS = 'Network_communities_green_blue'
if not os.path.exists(outdirSS):
    os.makedirs(outdirSS)



#ii=membership[0]
for ii in membership:
    community                                                                            =
node_communities[numpy.where(node_communities[:,1].astype(numpy.integer)==ii)[0],:]
    outdirS = 'community_'+str(ii)
    outdirS = os.path.join(outdirSS, outdirS)
    if not os.path.exists(outdirS):
        os.makedirs(outdirS)

    ## Breaking the community to switches and groups
    groups_ind = numpy.where(['group' in x for x in community[:,0]])[0]
    community_groups = community[groups_ind,:]
    community_switches = numpy.delete(community,(groups_ind),axis=0)
    ## Mapping into the PDB of each group
    # Reading PDB
    import MDAnalysis
    PDB = "/home/ziedgaieb/Documents/python_scripts_development/CCR7_CCL21_charged-N-
term_Xray_110-110-130_analysis/ccr7_ccl21.pdb"
    #Residues are ordered in alphabetic order following the charmm forcefield topology
file
    selectionarg1 = "(name CA)"

    dist_parameters_array = numpy.asarray(dataLog_calpha)

    for i in range(community_groups.shape[0]):
        outdirgroup = outdirS
        group_number = community_groups[i,0].split("_")[1]
        edges = numpy.asarray([jj.split("::") for jj in
dist_parameters_array[corr_indices_groups[int(group_number)]]])
        filename = '~group_'+str(group_number)+'_edges'+'.pdb'
        savepath = os.path.join(outdirgroup, filename)
        pdb = MDAnalysis.Writer(savepath, multiframe=True)
        for j in range(edges.shape[0]):
            u = MDAnalysis.Universe(PDB)
            resnum1 = edges[j][0].split("-")[1]
            chain1 = edges[j][0].split("-")[2]
            resnum2 = edges[j][1].split("-")[1]
            chain2 = edges[j][1].split("-")[2]
            selection = selectionarg1 + " and ((resid " + resnum1 + " and segid " +
chain1 + ") or (resid " + resnum2 + " and segid " + chain2 + "))"
            protein = u.select_atoms(selection)
            if (abs(int(resnum1) - int(resnum2))==1):
                protein.set_resids((int(resnum1),(int(resnum2) + 1)))
            protein.set_names(("CA","CA"))
            pdb.write(protein)
        pdb.close()
        ## Outputting timeseries plot
        import matplotlib
        matplotlib.use('Agg')
        from matplotlib import pyplot as plt
        outdir = community_groups[i,0]
        outdirgroup = os.path.join(outdirS, outdir)
        if not os.path.exists(outdirgroup):
            os.makedirs(outdirgroup)
        for j in range(len(corr_indices_groups[int(group_number)])):
            ## plotting the contact timeseries for each group
            dataLog_ind = corr_indices_groups[int(group_number)][j]
            filename = dataLog_calpha[dataLog_ind]+'_timeseries.png'
            savepath = os.path.join(outdirgroup, filename)
            plt.figure(figsize=(40,12))
            plt.plot(data_calpha[dataLog_ind,:], "g.")
            plt.rc("font",size=32)
            plt.savefig(savepath,dpi=100)
            plt.close()

    ## PDB mapping of molecular switches
    #outdir = "molecular_switches"
    #outdirgroup = os.path.join(outdirS, outdir)
    #if not os.path.exists(outdirgroup):
    #    os.makedirs(outdirgroup)
    selectionarg1 = "((resname ALA and name CB) or \
    (resname ARG and name CZ) or \
    (resname ASP and name CG) or \
    (resname ASN and name CG) or \
    (resname CYS and name SG) or \
    (resname GLN and name CD) or \
    (resname GLU and name CD) or \
    (resname GLY and name CA) or \
    (resname HIS and name CG) or \
    (resname HSE and name CG) or \
    (resname HSD and name CG) or \
    (resname HSP and name CG) or \
    (resname ILE and name CG1) or \
    (resname LEU and name CG) or \
    (resname LYS and name NZ) or \
    (resname MET and name SD) or \
    (resname PHE and name CG) or \
    (resname PRO and name CG) or \
    (resname SER and name OG) or \
    (resname THR and name CB) or \
    (resname TRP and name CE2) or \
    (resname TYR and name CG) or \
    (resname VAL and name CB))"

    community_switches_tmp = [jj.split("_")[0] for jj in community_switches[:,0]]
    edges = numpy.asarray([jj.split("::") for jj in community_switches_tmp])
    filename = '~community_switches_edges'+'.pdb'
    outdirgroup = outdirS
    savepath = os.path.join(outdirgroup, filename)
    pdb = MDAnalysis.Writer(savepath, multiframe=True)
    for j in range(edges.shape[0]):
        u = MDAnalysis.Universe(PDB)
        resnum1 = edges[j][0].split("-")[1]
        chain1 = edges[j][0].split("-")[2]
        resnum2 = edges[j][1].split("-")[1]
        chain2 = edges[j][1].split("-")[2]
        selection = selectionarg1 + " and ((resid " + resnum1 + " and segid " + chain1 +
") or (resid " + resnum2 + " and segid " + chain2 + "))"
        protein = u.select_atoms(selection)
        if (abs(int(resnum1) - int(resnum2))==1):
            protein.set_resids((int(resnum1),(int(resnum2) + 1)))
        protein.set_names(("CA","CA"))
        pdb.write(protein)
    pdb.close()

    ## Outputting timeseries plot
    import matplotlib
    matplotlib.use('Agg')
    from matplotlib import pyplot as plt
    outdir = "molecular_switches"
    outdirgroup = os.path.join(outdirS, outdir)
    if not os.path.exists(outdirgroup):
        os.makedirs(outdirgroup)
    for j in range(len(community_switches[:,0])):
        ## plotting the contact timeseries for each group
        dataLog_ind = numpy.where(dataLog_sidechain_polar==community_switches[j,0])[0][0]
        filename = str(dataLog_sidechain_polar[dataLog_ind])+'_timeseries.png'
        savepath = os.path.join(outdirgroup, filename)
        plt.figure(figsize=(40,12))
        plt.plot(data_sidechain_polar[dataLog_ind,:], "b.")
        plt.rc("font",size=32)
        plt.savefig(savepath,dpi=100)
        plt.close()
##
```

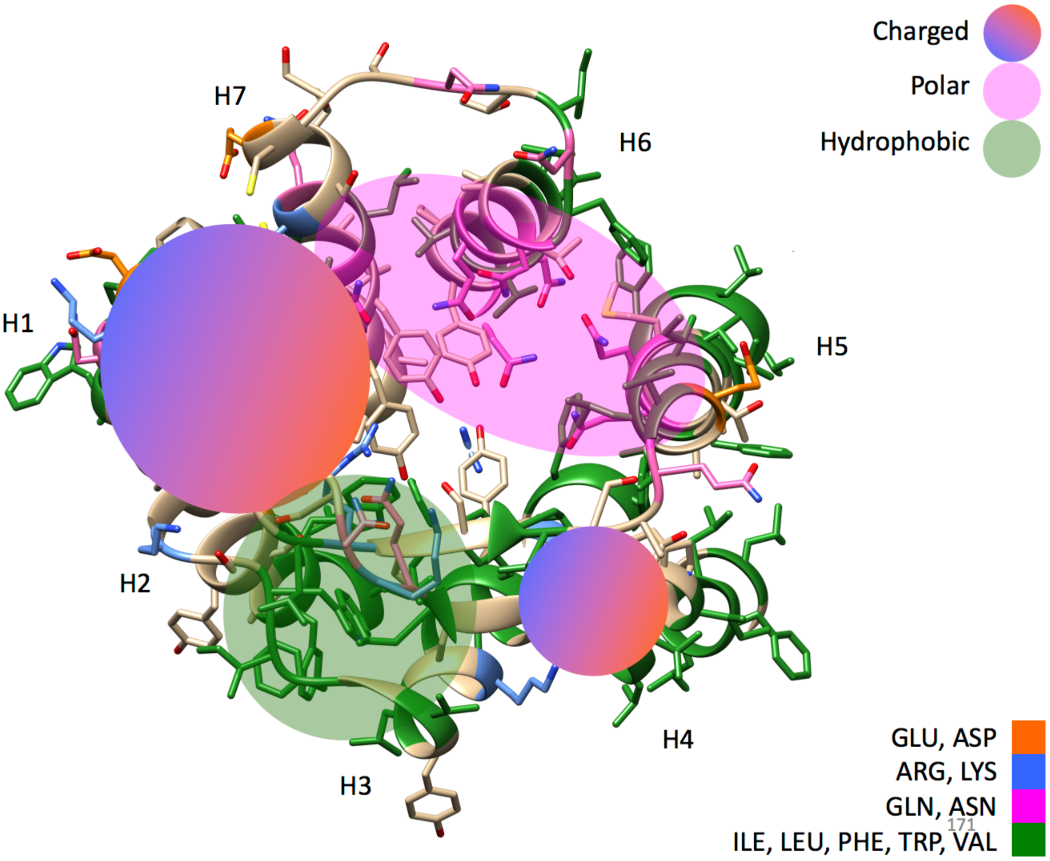
APPENDIX G: LIGAND DESIGN SLIDES

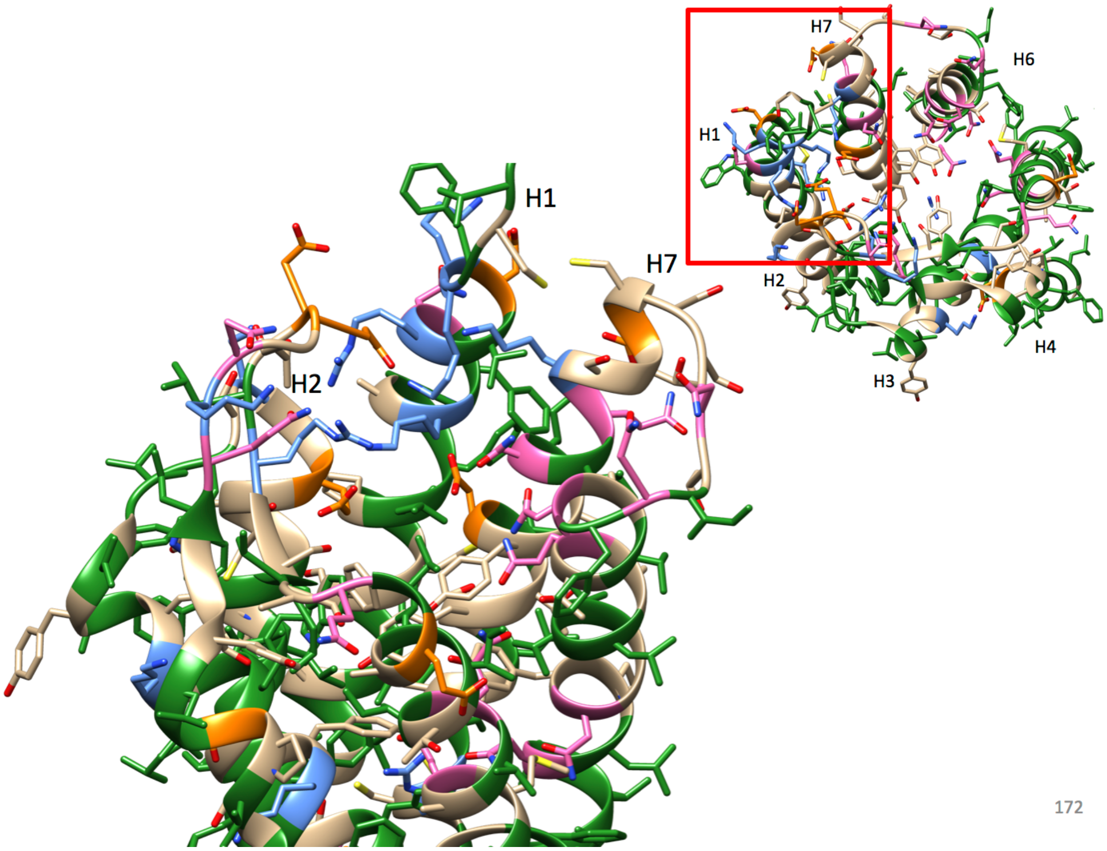


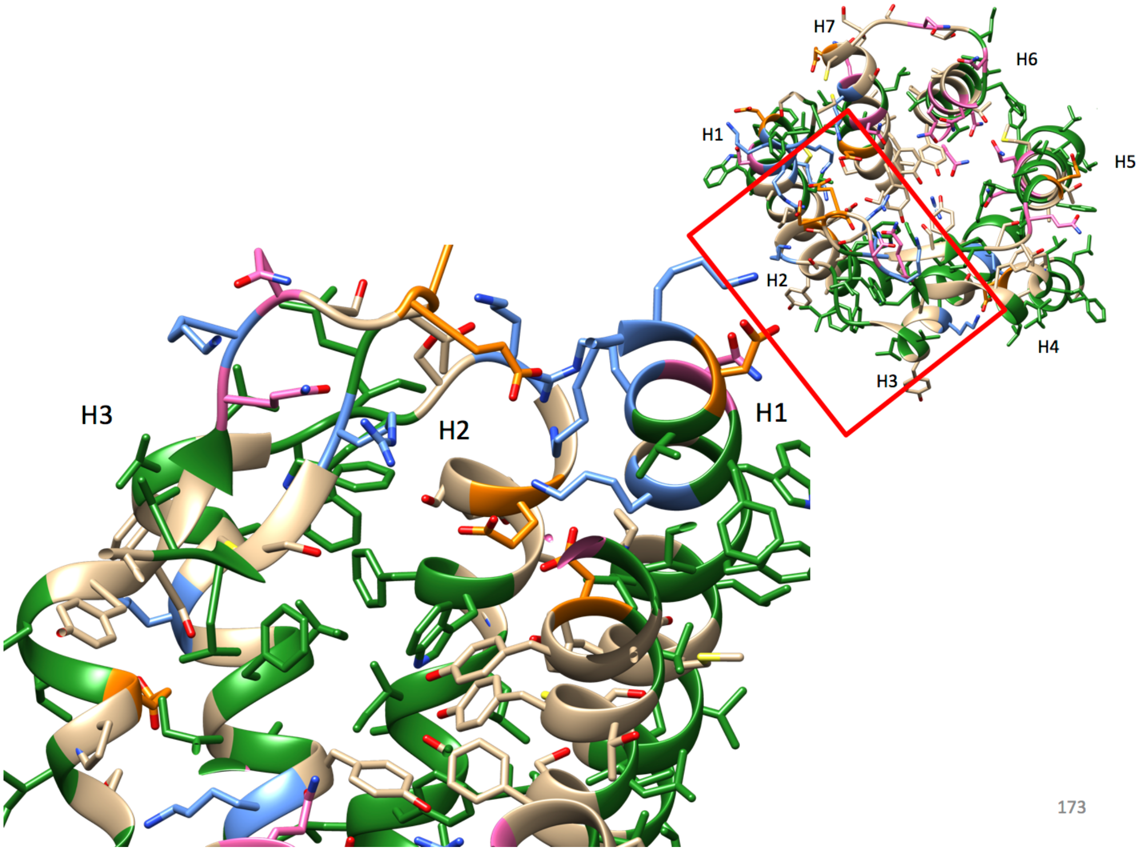
Provide mutations of minimal number of residues to exchange one ligand's binding site to the others.

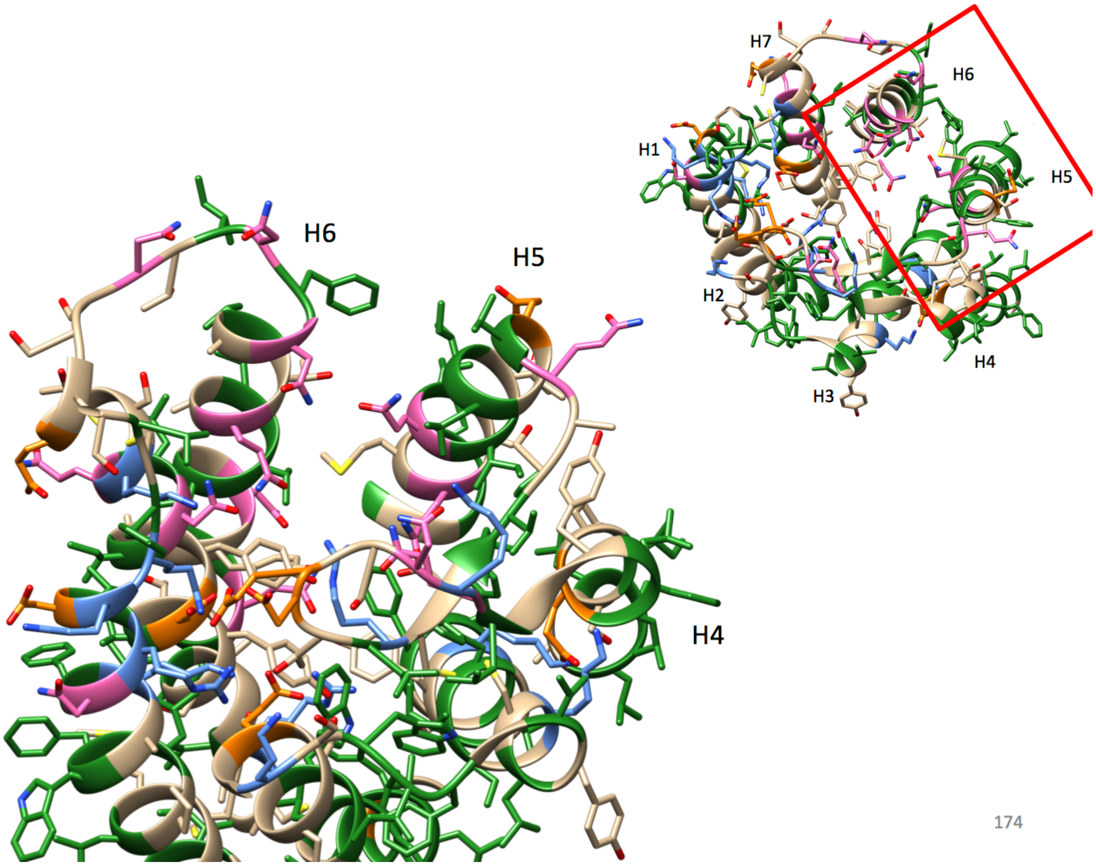
CCL19 N_tGANDAED CC
CCL21 N_tSDGGGQD CC

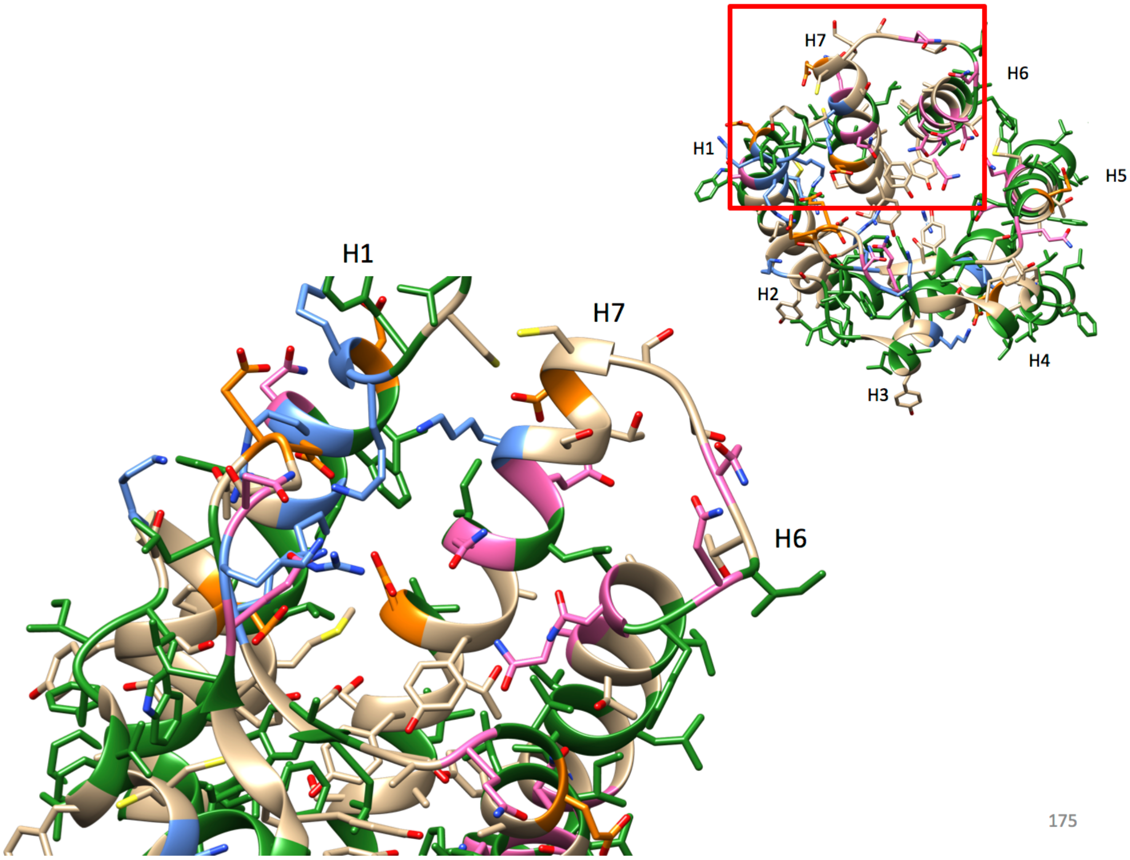

H7
H6
H1
H5
H2
H4
H3
GLU, ASP 🟧
ARG, LYS 🟦
GLN, ASN 🟪
ILE, LEU, PHE, TRP, VAL 🟩

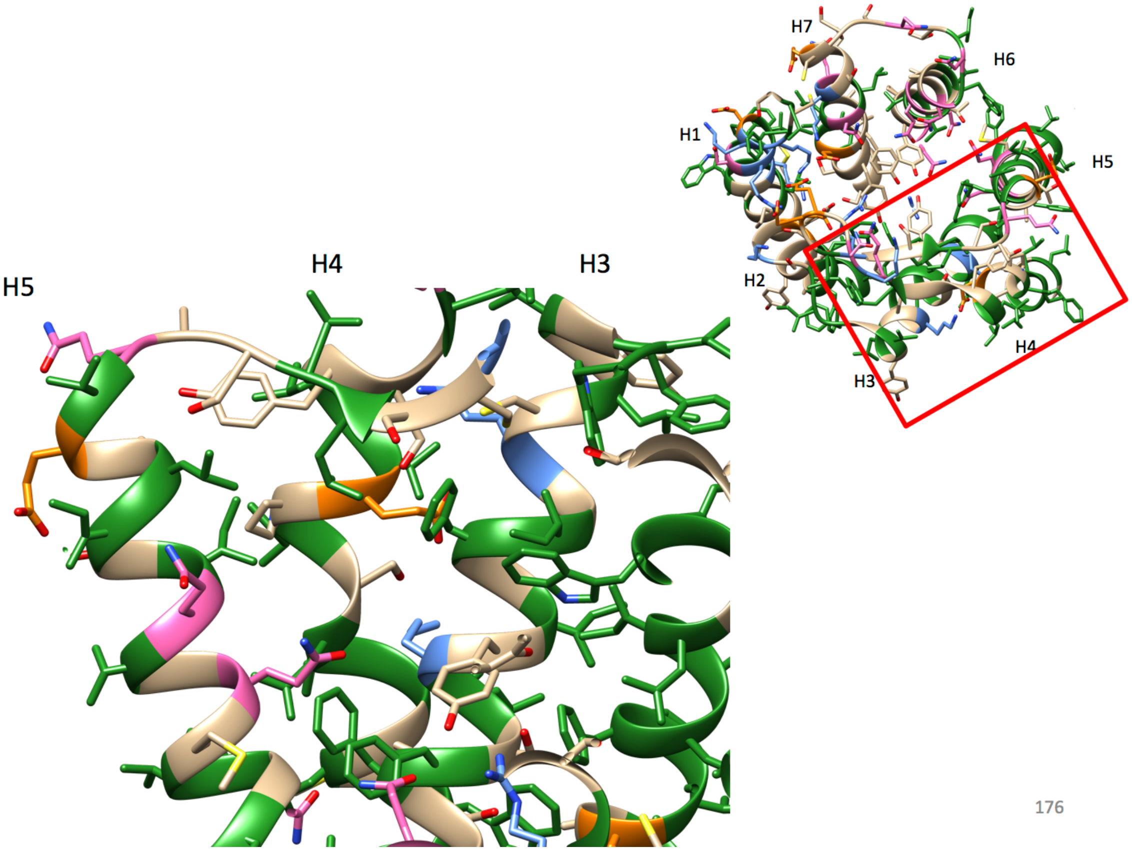

Charged
Polar
Hydrophobic
H7
H6
H1
H5
H2
H4
H3
GLU, ASP
ARG, LYS
GLN, ASN
ILE, LEU, PHE, TRP, VAL


H1
H7
H2
H7
H6
H1
H5
H2
H4
H3
172


H7
H6
H1
H5
H2
H4
H3
H3
H2
H1
173

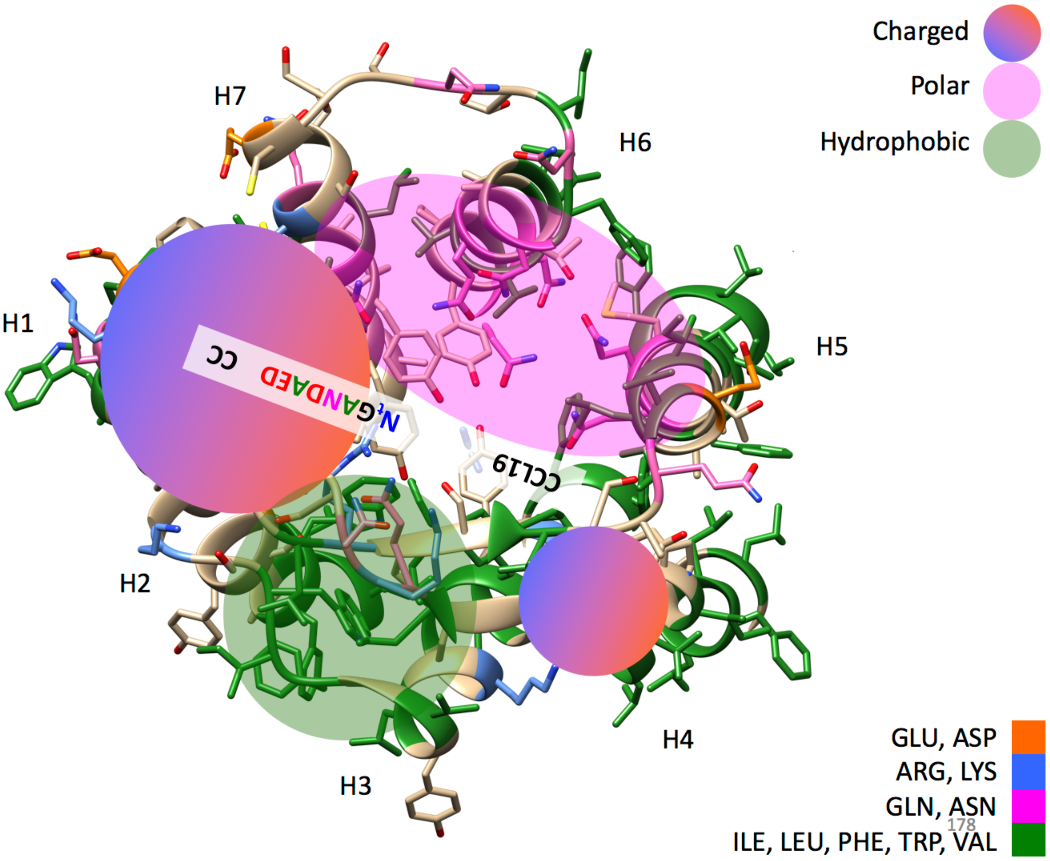

H6
H5
H4
H7
H6
H1
H5
H2
H4
H3
174

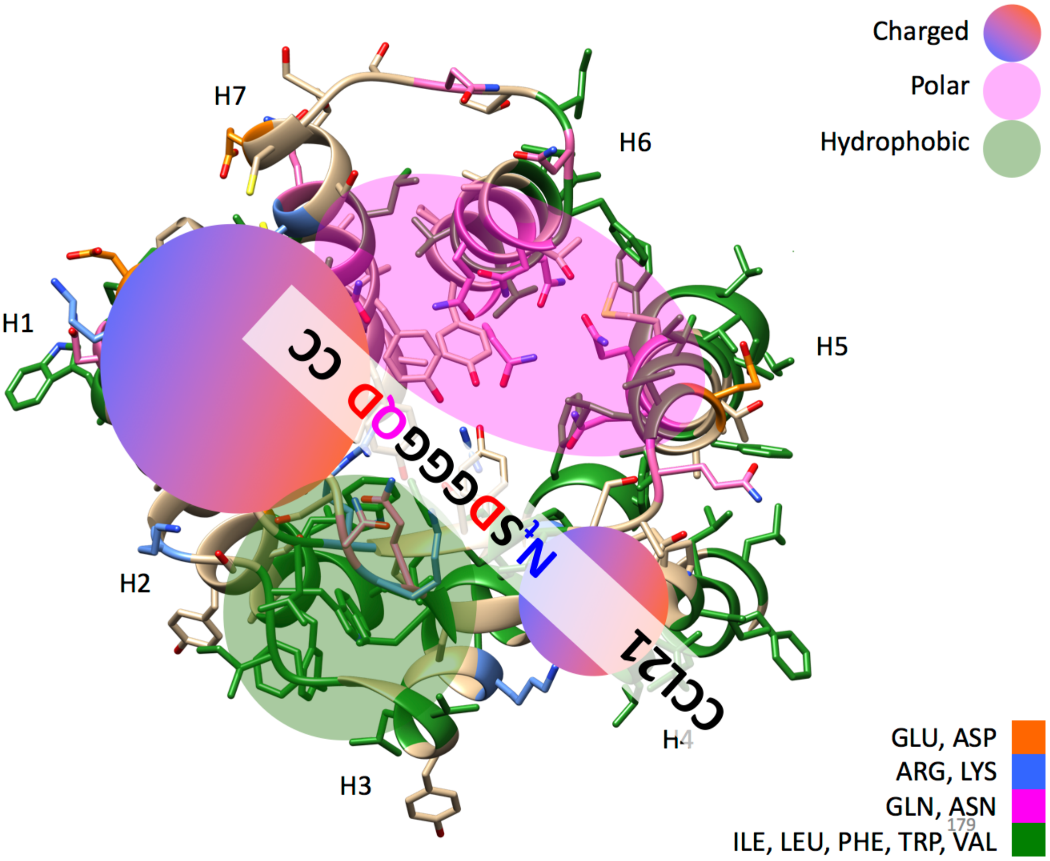

H1
H7
H6
H7
H6
H1
H5
H2
H4
H3

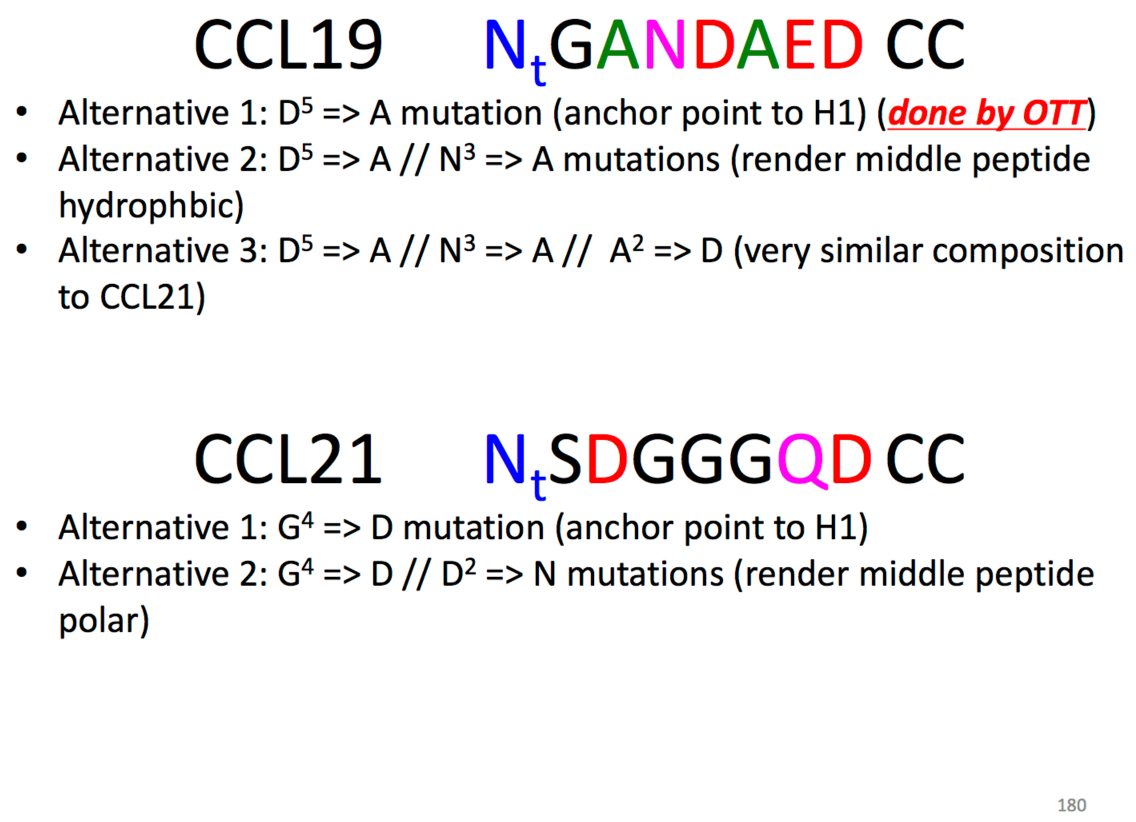

H5
H4
H3
H7
H6
H1
H5
H2
H4
H3

Provide mutations of minimal number of residues to exchange one ligand’s binding site to the others.

CCL19	N_tGANDAED CC
CCL21	N_tSDGGGQD CC


Charged
Polar
Hydrophobic
H7
H6
H1
H5
N$_t$GANDAED CC
CCL19
H2
H4
H3
GLU, ASP 🟧
ARG, LYS 🟦
GLN, ASN 🟪
ILE, LEU, PHE, TRP, VAL 🟩


Charged
Polar
Hydrophobic
H7
H6
H1
H5
H2
H3
H4
CCL21
GLU, ASP
ARG, LYS
GLN, ASN
ILE, LEU, PHE, TRP, VAL



CCL19 N_tGANDAED CC

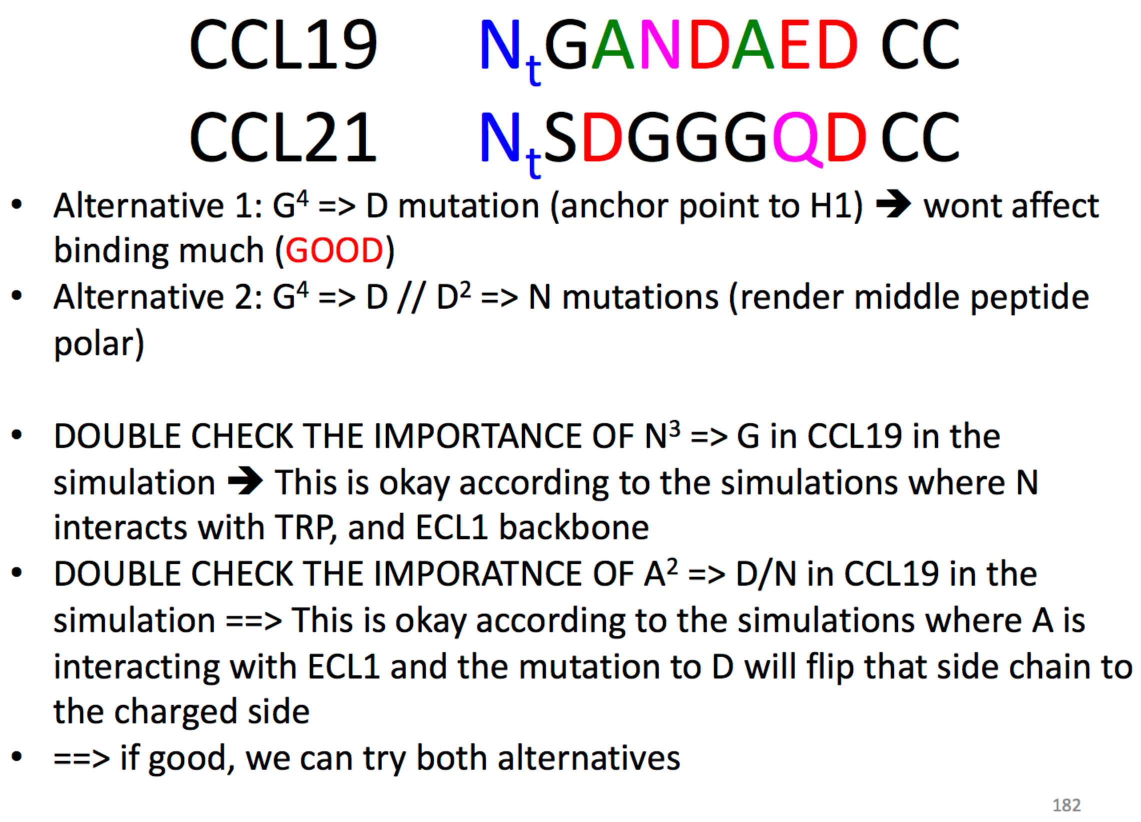
- Alternative 1: D^5 => A mutation (anchor point to H1) (***done by OTT***)
- Alternative 2: D^5 => A // N^3 => A mutations (render middle peptide hydrophbic)
- Alternative 3: D^5 => A // N^3 => A // A^2 => D (very similar composition to CCL21)

CCL21 N_tSDGGGQD CC

- Alternative 1: G^4 => D mutation (anchor point to H1)
- Alternative 2: G^4 => D // D^2 => N mutations (render middle peptide polar)

Experimental Data: CCL19 binding

- Mutations with intact backbone:
  - the backbone + D^4 gives the pose of the N-terminus (computational) ➔ CCL19 binding pose
  - mutation D^4 => A might change the binding pose which accounts for the slight decrease in binding affinity (4 fold) and G-protein activation (9fold) ➔ not very important
  - mutation N^3 => A important for binding (8fold) and activation (29 fold) ➔ in CCL19 binding pose
  - D^7 is extremely important for binding (60 fold) and activation (200 fold) ➔ in CCL19 binding pose
  - $T(A)^2$ and E^6 show (almost) no effect
- Mutations with non-intact backbone (truncations)
  - deleting 3 residues lower binding by 12 fold
  - deleting 4 residues lower binding by 30 fold (the extra bump is due to the (+) charged N-terminus
  - deleting 5 residues lower binding by 262 fold but when acetylated we get 12 fold
  - ➔ So deleting of 5 residues should lower binding up to 12 fold if all truncation were acetylated
  - deleting 6 residues with acetylation lower binding by 275 fold ➔ So E^6 is important for binding, but its mutation to A showed no importance ➔ So it is the length (backbone) of the N-terminus that matters here and should tell us that residues 6 is where we start leaving the pocket of activation and entering the binding site domain.
  - deleting 7 residues ➔ antagonist of the receptor
  - ➔➔ residues 1 – 5 important for activation, and not binding (unless we introduce a positive charge there)
  - ➔➔ residues 6 (not side chain) backbone and 7 important for binding and activation
  - ➔➔ residues 6 side chain is not important cause of residues 7 is charged as well.

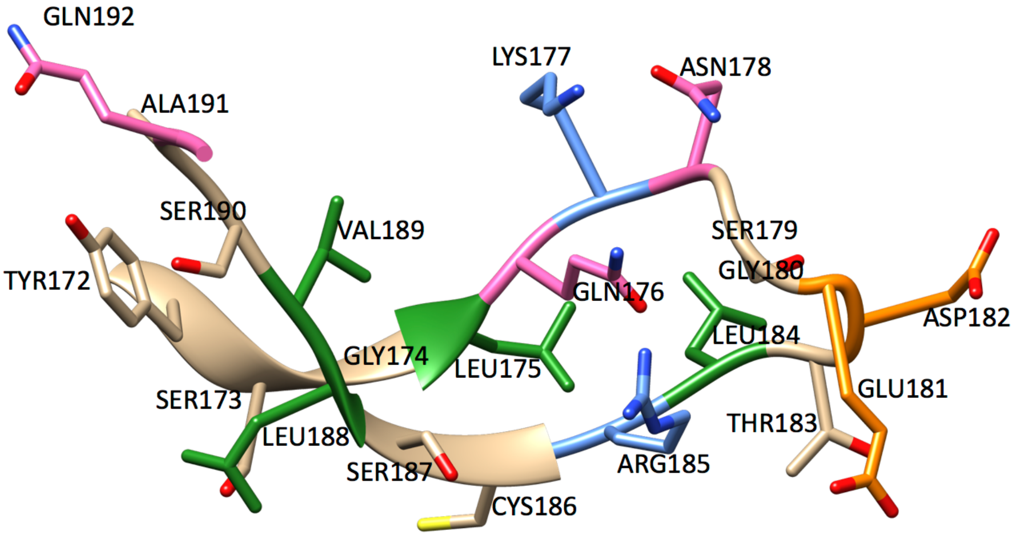


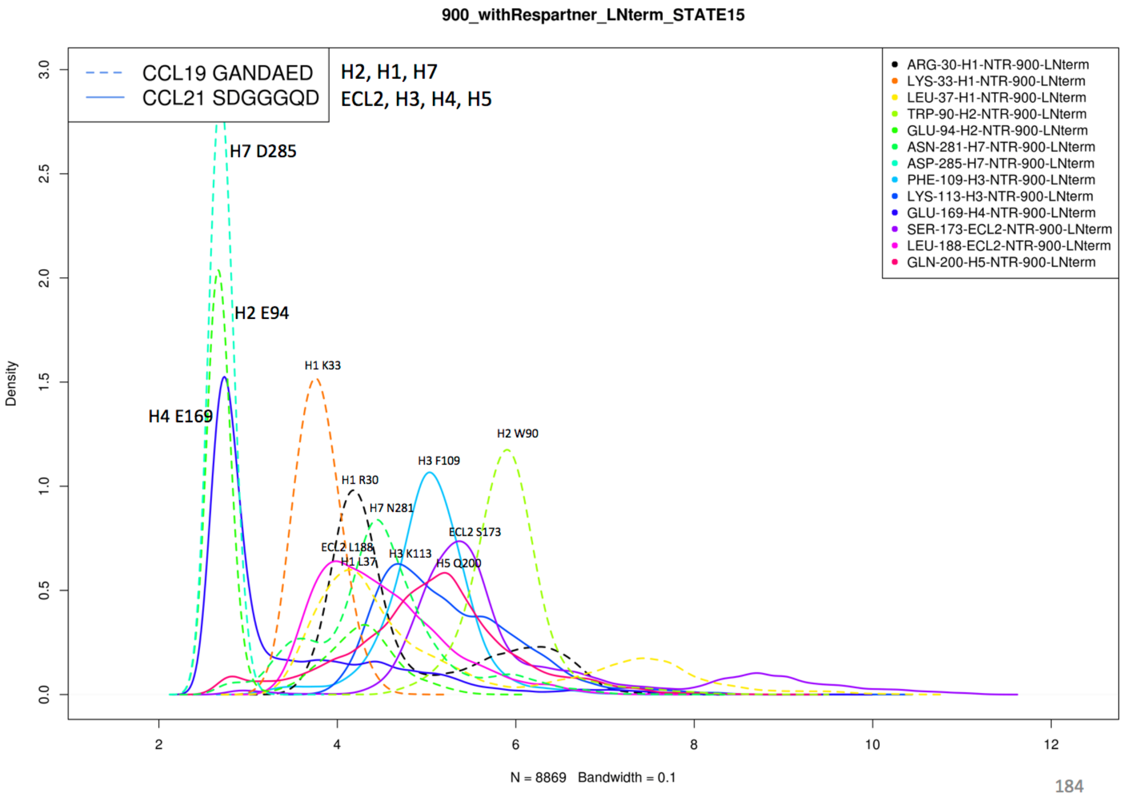
CCL19 N_tGANDAED CC

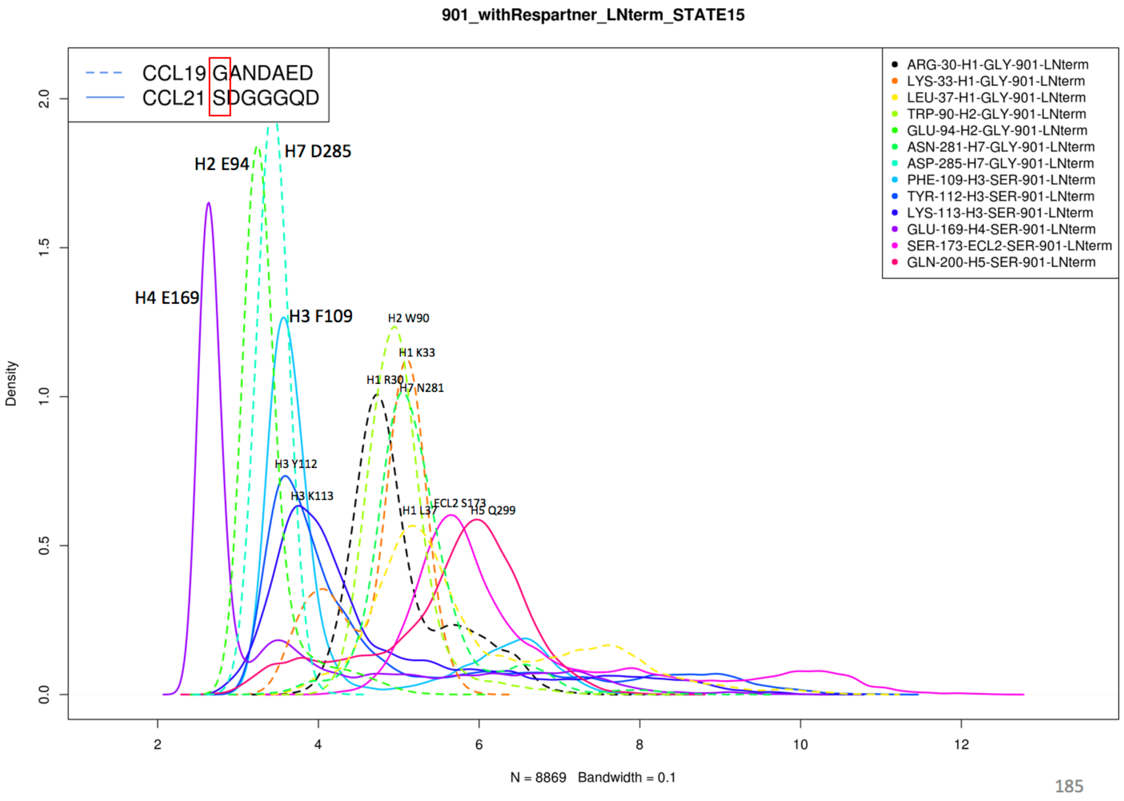
CCL21 N_tSDGGGQD CC

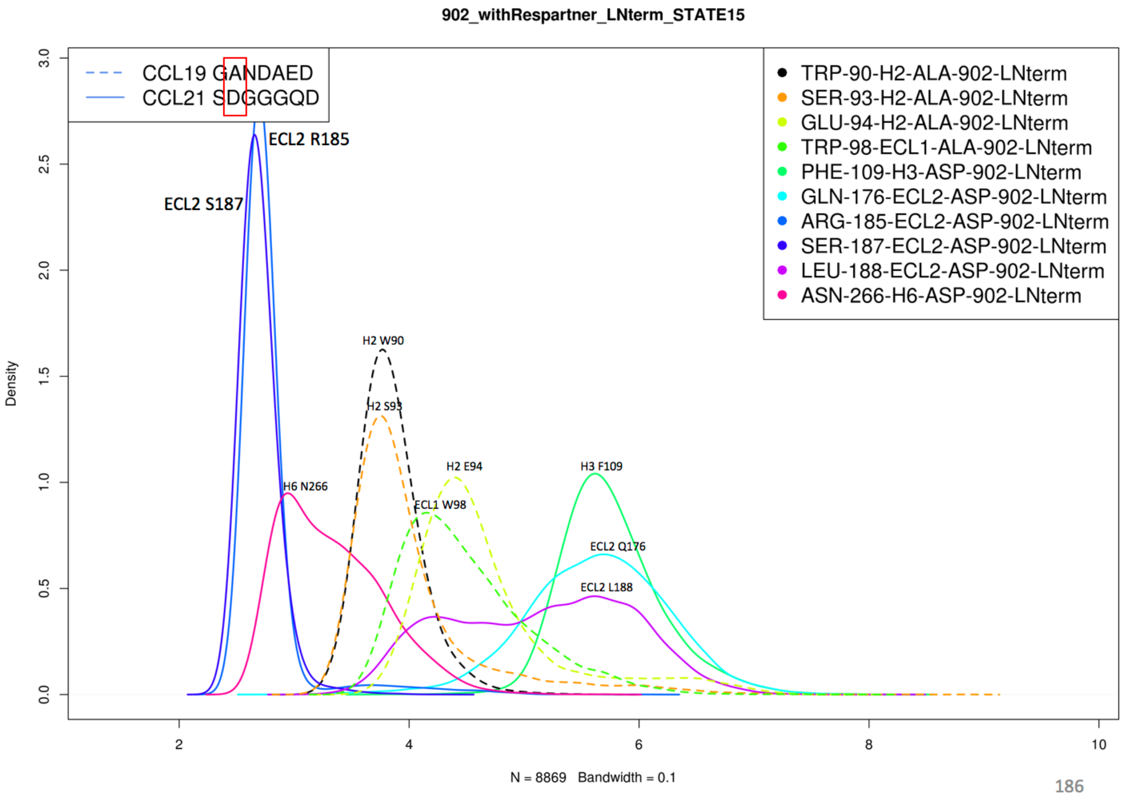
- Alternative 1: G^4 => D mutation (anchor point to H1) ➔ wont affect binding much (GOOD)
- Alternative 2: G^4 => D // D^2 => N mutations (render middle peptide polar)

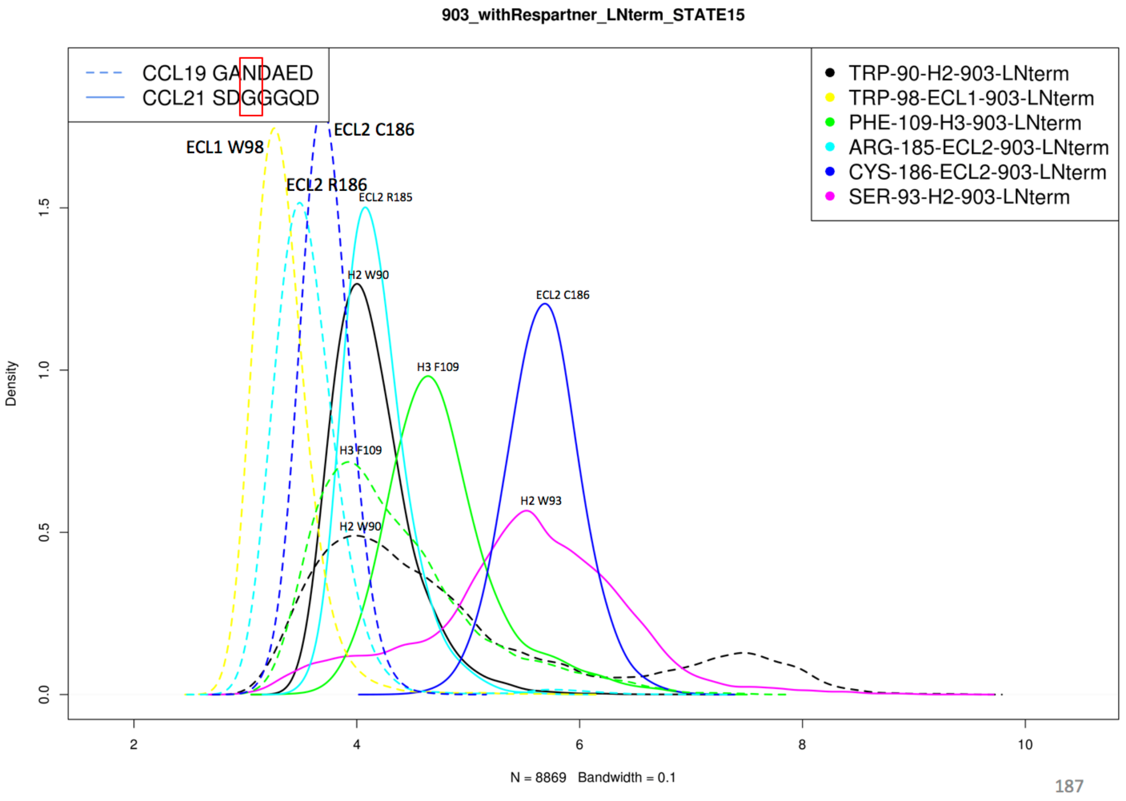
- DOUBLE CHECK THE IMPORTANCE OF N^3 => G in CCL19 in the simulation ➔ This is okay according to the simulations where N interacts with TRP, and ECL1 backbone
- DOUBLE CHECK THE IMPORATNCE OF A^2 => D/N in CCL19 in the simulation ==> This is okay according to the simulations where A is interacting with ECL1 and the mutation to D will flip that side chain to the charged side
- ==> if good, we can try both alternatives

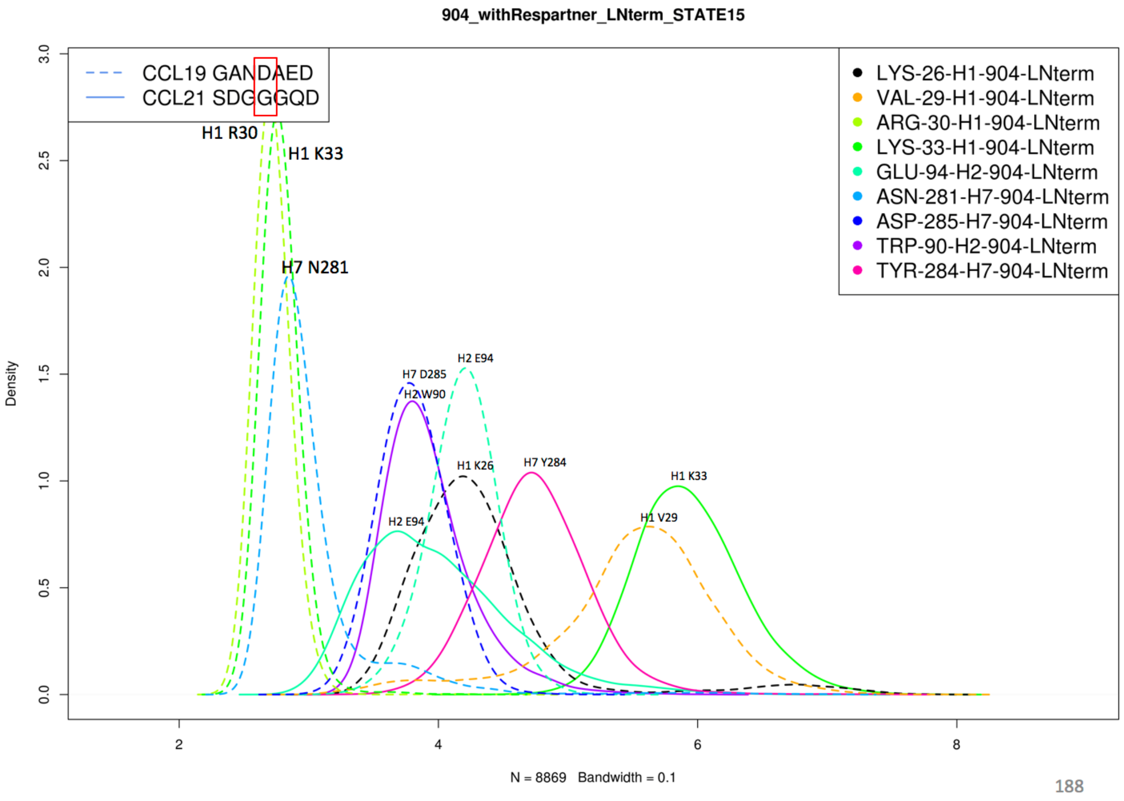

GLN192
LYS177
ASN178
ALA191
SER190
VAL189
SER179
TYR172
GLY180
GLN176
ASP182
LEU184
GLY174
LEU175
GLU181
SER173
THR183
LEU188
ARG185
SER187
CYS186

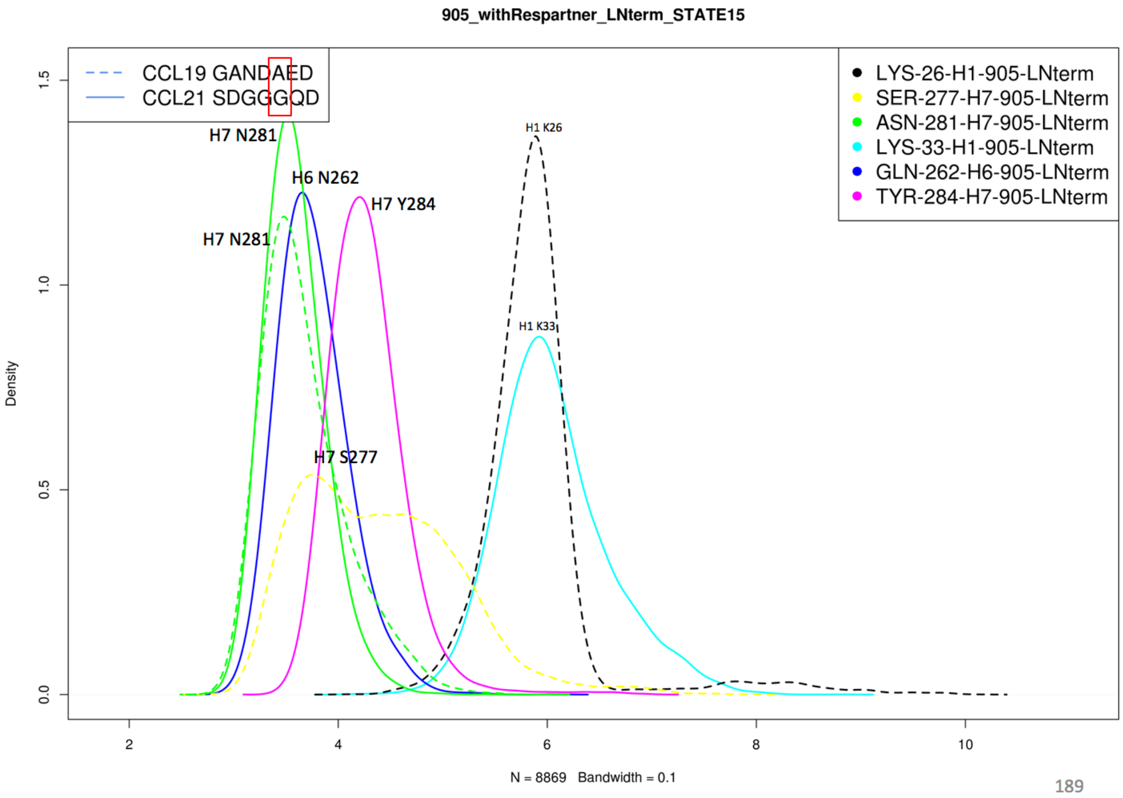

900_withRespartner_LNterm_STATE15
CCL19 GANDAED
CCL21 SDGGGQD
H2, H1, H7
ECL2, H3, H4, H5
ARG-30-H1-NTR-900-LNterm
LYS-33-H1-NTR-900-LNterm
LEU-37-H1-NTR-900-LNterm
TRP-90-H2-NTR-900-LNterm
GLU-94-H2-NTR-900-LNterm
ASN-281-H7-NTR-900-LNterm
ASP-285-H7-NTR-900-LNterm
PHE-109-H3-NTR-900-LNterm
LYS-113-H3-NTR-900-LNterm
GLU-169-H4-NTR-900-LNterm
SER-173-ECL2-NTR-900-LNterm
LEU-188-ECL2-NTR-900-LNterm
GLN-200-H5-NTR-900-LNterm
H7 D285
H2 E94
H4 E169
H1 K33
H2 W90
H3 F109
H1 R30
H7 N281
ECL2 S173
ECL2 L188
H1 L37
H3 K113
H5 Q200
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
10
12
N = 8869 Bandwidth = 0.1
184

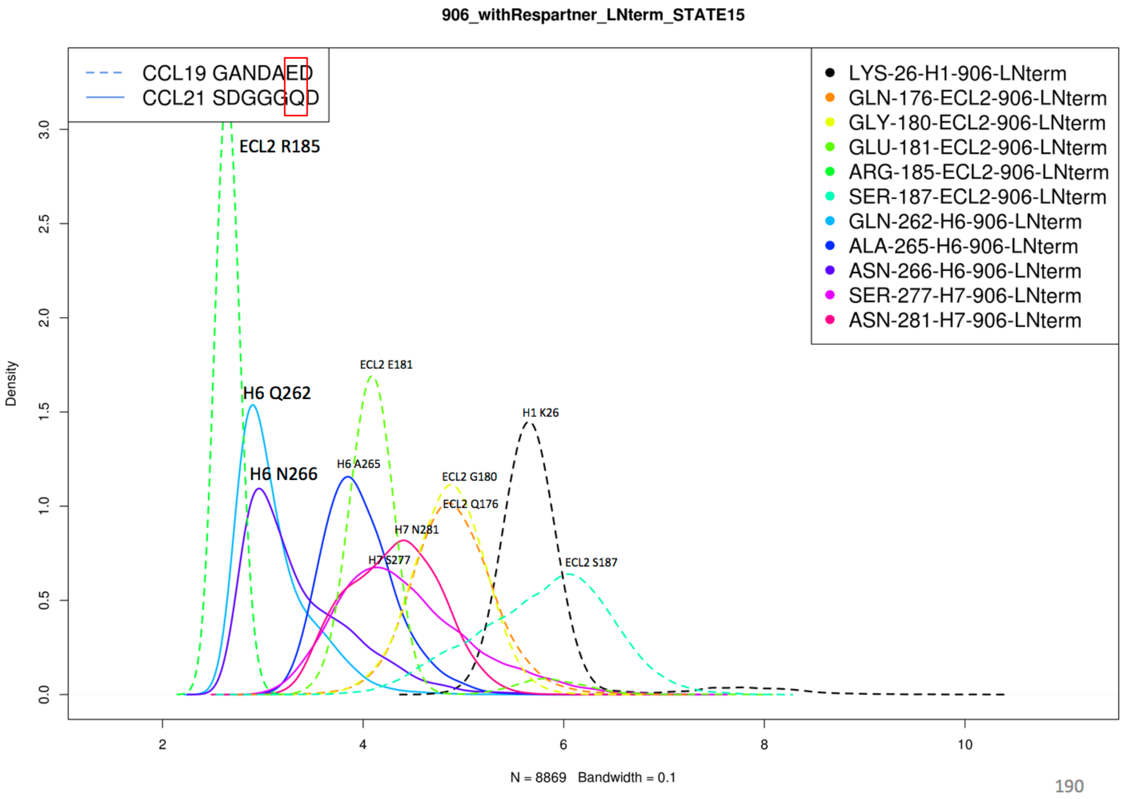

901_withRespartner_LNterm_STATE15
CCL19 GANDAED
CCL21 SDGGGQD
ARG-30-H1-GLY-901-LNterm
LYS-33-H1-GLY-901-LNterm
LEU-37-H1-GLY-901-LNterm
TRP-90-H2-GLY-901-LNterm
GLU-94-H2-GLY-901-LNterm
ASN-281-H7-GLY-901-LNterm
ASP-285-H7-GLY-901-LNterm
PHE-109-H3-SER-901-LNterm
TYR-112-H3-SER-901-LNterm
LYS-113-H3-SER-901-LNterm
GLU-169-H4-SER-901-LNterm
SER-173-ECL2-SER-901-LNterm
GLN-200-H5-SER-901-LNterm
H2 E94
H7 D285
H4 E169
H3 F109
H2 W90
H1 K33
H1 R30
H7 N281
H3 Y112
H3 K113
ECL2 S173
H1 L37
H5 Q299
Density
0.0
0.5
1.0
1.5
2.0
2
4
6
8
10
12
N = 8869 Bandwidth = 0.1
185

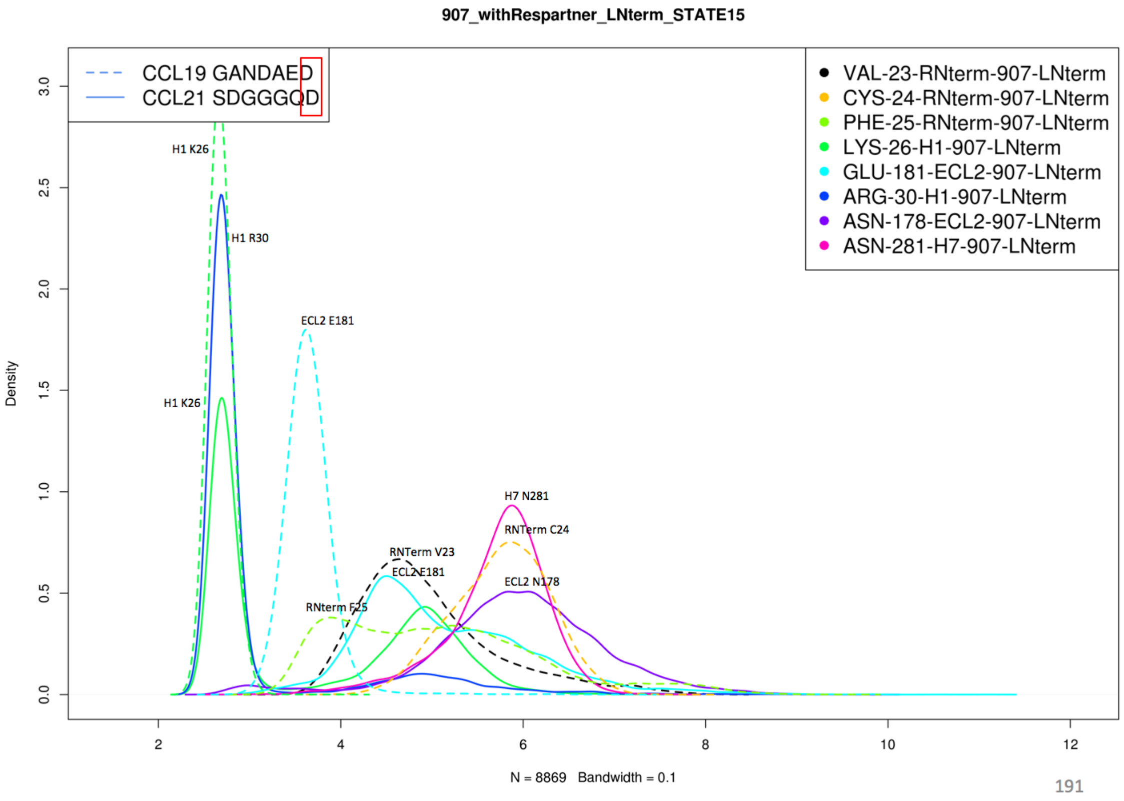

902_withRespartner_LNterm_STATE15
CCL19 GANDAED
CCL21 SDGGGQD
ECL2 R185
ECL2 S187
H2 W90
H2 S93
H2 E94
ECL1 W98
H6 N266
H3 F109
ECL2 Q176
ECL2 L188
● TRP-90-H2-ALA-902-LNterm
● SER-93-H2-ALA-902-LNterm
● GLU-94-H2-ALA-902-LNterm
● TRP-98-ECL1-ALA-902-LNterm
● PHE-109-H3-ASP-902-LNterm
● GLN-176-ECL2-ASP-902-LNterm
● ARG-185-ECL2-ASP-902-LNterm
● SER-187-ECL2-ASP-902-LNterm
● LEU-188-ECL2-ASP-902-LNterm
● ASN-266-H6-ASP-902-LNterm
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
10
N = 8869 Bandwidth = 0.1
186


903_withRespartner_LNterm_STATE15
CCL19 GANDAED
CCL21 SDGGGQD
TRP-90-H2-903-LNterm
TRP-98-ECL1-903-LNterm
PHE-109-H3-903-LNterm
ARG-185-ECL2-903-LNterm
CYS-186-ECL2-903-LNterm
SER-93-H2-903-LNterm
ECL2 C186
ECL1 W98
ECL2 R186
ECL2 R185
H2 W90
ECL2 C186
H3 F109
H3 F109
H2 W93
H2 W90
Density
0.0
0.5
1.0
1.5
2
4
6
8
10
N = 8869 Bandwidth = 0.1
187

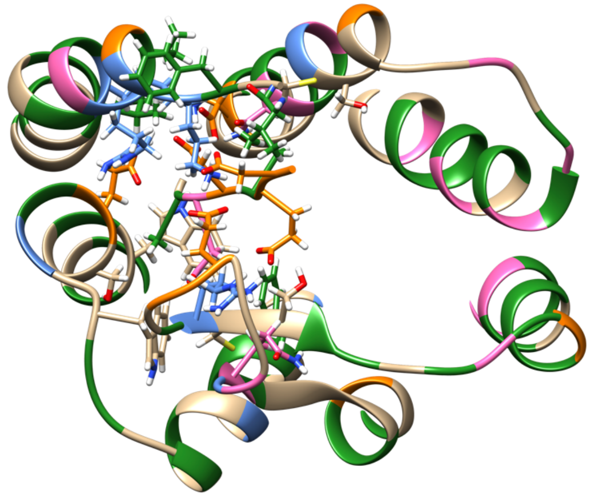

904_withRespartner_LNterm_STATE15
CCL19 GANDAED
CCL21 SDGGGQD
● LYS-26-H1-904-LNterm
● VAL-29-H1-904-LNterm
● ARG-30-H1-904-LNterm
● LYS-33-H1-904-LNterm
● GLU-94-H2-904-LNterm
● ASN-281-H7-904-LNterm
● ASP-285-H7-904-LNterm
● TRP-90-H2-904-LNterm
● TYR-284-H7-904-LNterm
H1 R30
H1 K33
H7 N281
H7 D285
H2 W90
H2 E94
H1 K26
H7 Y284
H1 K33
H1 V29
H2 E94
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
N = 8869 Bandwidth = 0.1
188


905_withRespartner_LNterm_STATE15
CCL19 GANDAED
CCL21 SDGGGQD
LYS-26-H1-905-LNterm
SER-277-H7-905-LNterm
ASN-281-H7-905-LNterm
LYS-33-H1-905-LNterm
GLN-262-H6-905-LNterm
TYR-284-H7-905-LNterm
H7 N281
H6 N262
H7 Y284
H7 N281
H1 K26
H1 K33
H7 S277
Density
0.0
0.5
1.0
1.5
2
4
6
8
10
N = 8869 Bandwidth = 0.1
189

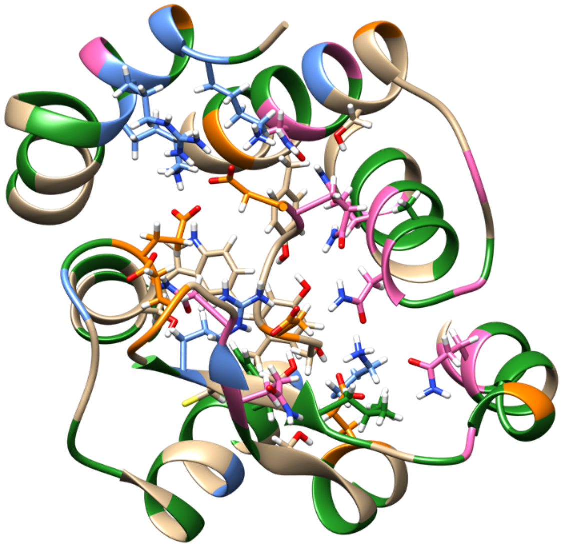

906_withRespartner_LNterm_STATE15
CCL19 GANDAED
CCL21 SDGGGQD
LYS-26-H1-906-LNterm
GLN-176-ECL2-906-LNterm
GLY-180-ECL2-906-LNterm
GLU-181-ECL2-906-LNterm
ARG-185-ECL2-906-LNterm
SER-187-ECL2-906-LNterm
GLN-262-H6-906-LNterm
ALA-265-H6-906-LNterm
ASN-266-H6-906-LNterm
SER-277-H7-906-LNterm
ASN-281-H7-906-LNterm
ECL2 R185
H6 Q262
H6 N266
ECL2 E181
H6 A265
H1 K26
ECL2 G180
ECL2 Q176
H7 N281
H7 S277
ECL2 S187
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
10
N = 8869 Bandwidth = 0.1
190


907_withRespartner_LNterm_STATE15
CCL19 GANDAED
CCL21 SDGGGQD
VAL-23-RNterm-907-LNterm
CYS-24-RNterm-907-LNterm
PHE-25-RNterm-907-LNterm
LYS-26-H1-907-LNterm
GLU-181-ECL2-907-LNterm
ARG-30-H1-907-LNterm
ASN-178-ECL2-907-LNterm
ASN-281-H7-907-LNterm
H1 K26
H1 R30
ECL2 E181
H1 K26
H7 N281
RNTerm C24
RNTerm V23
ECL2 E181
ECL2 N178
RNterm F25
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
10
12
N = 8869 Bandwidth = 0.1
191

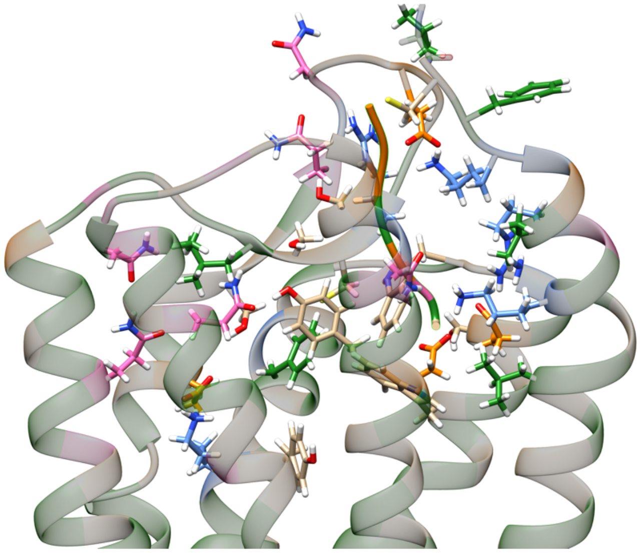
CCL19



CCL21

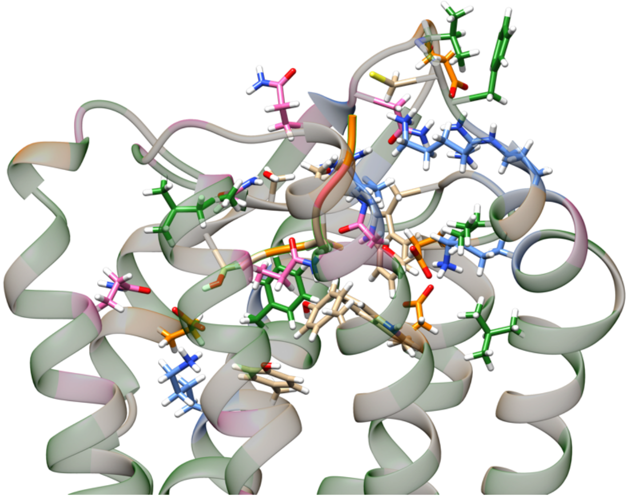


CCL19



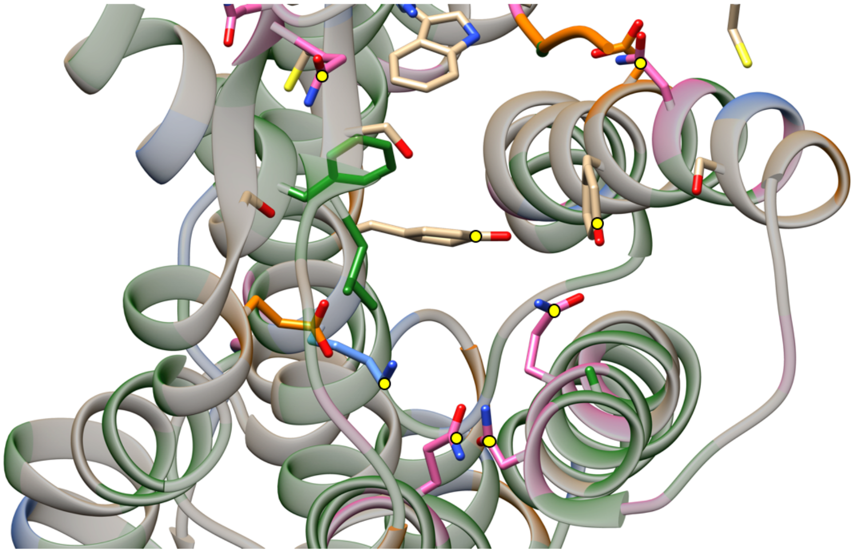
194

CCL21

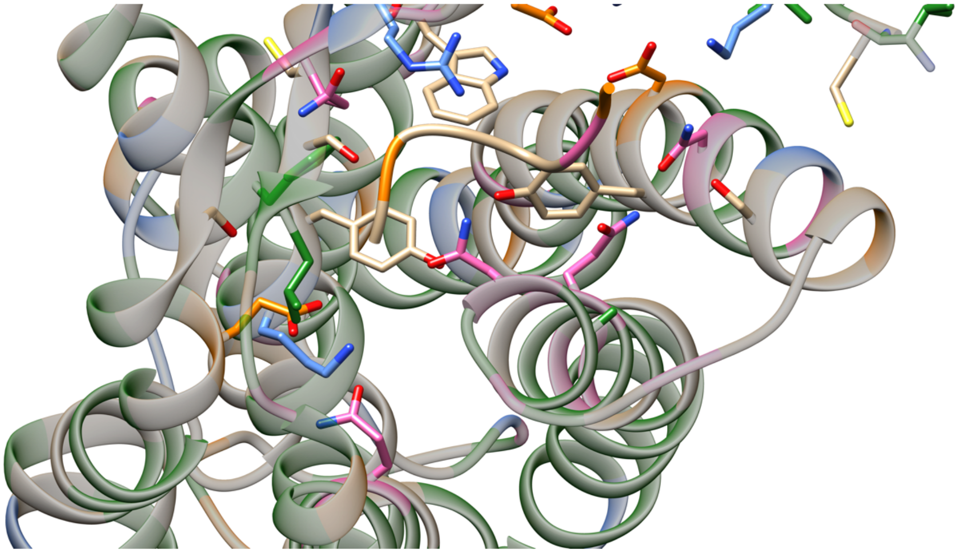


195

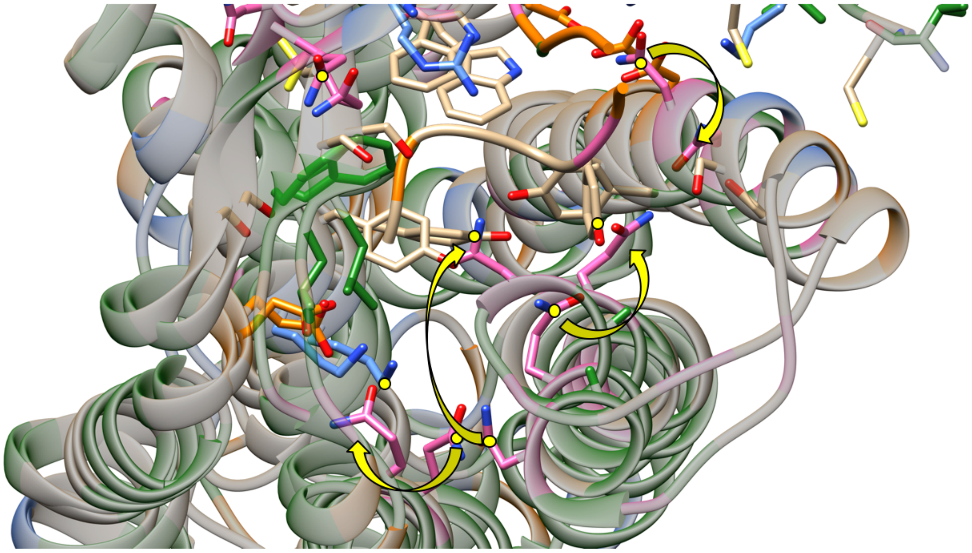
CCL19

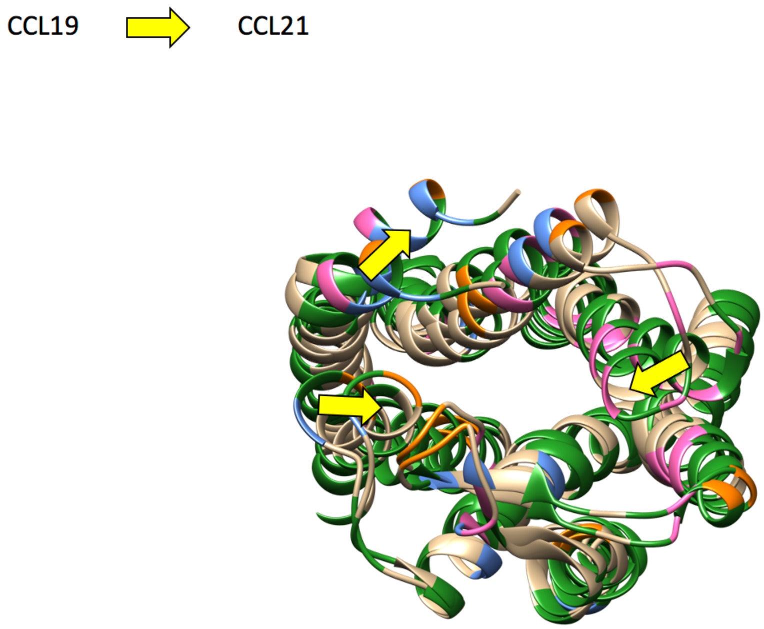


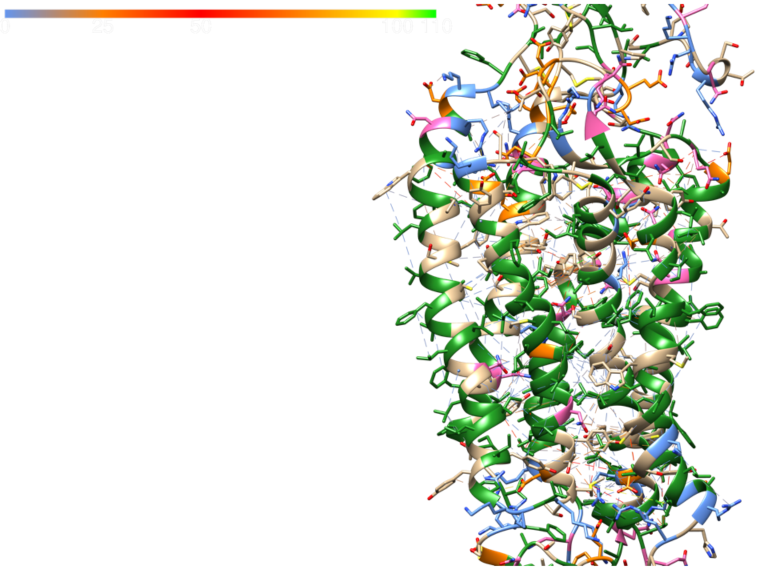
CCL21



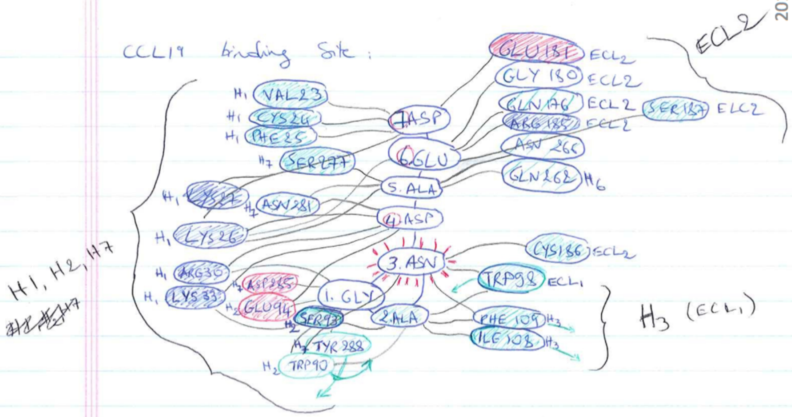
CCL19 ➡ CCL21




CCL19
CCL21

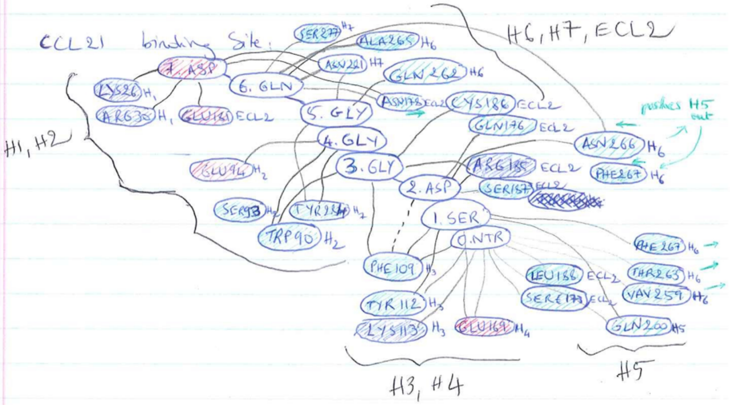

0
25
50
100
110

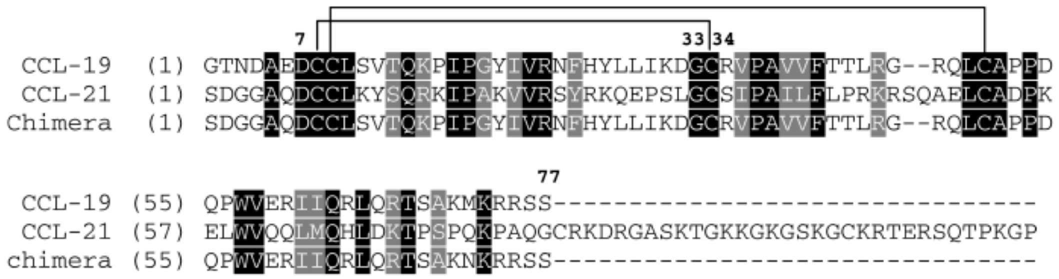
CCL19 binding Site:



mouse
CCL19 GANDAED CC
G T N D A E D (human)
CCL21 SDGGGQD CC
S D G G A Q D (human)

CCL21 binding Site:




7
33 34
CCL-19 (1) GTNDAEDCCLSVTQKPIPGYIVRNFHYLLIKDGCRVPAVVFTTLRG--RQLCAPPD
CCL-21 (1) SDGGAQDCCLKYSQRKIPAKVVRSYRKQEPSLGCSIPAILFLPRKRSQAELCADPK
Chimera (1) SDGGAQDCCLSVTQKPIPGYIVRNFHYLLIKDGCRVPAVVFTTLRG--RQLCAPPD
77
CCL-19 (55) QPWVERIIQRLQRTSAKMKRRSS--------------------------------
CCL-21 (57) ELWVQQLMQHLDKTPSPQKPAQGCRKDRGASKTGKKGKGSKGCKRTERSQTPKGP
chimera (55) QPWVERIIQRLQRTSAKNKRRSS--------------------------------

Zied GAIEB, Ph.D.

(937) 546-0392
ziedgaieb@gmail.com

Education

Ph.D. Bioengineering 2011-2016
University of California, Riverside
Dr. Dimitrios Morikis "*Molecular Switches Coordinate Dynamically Coupled Allosteric Networks in Protein Complexes.*"

Bachelor in Chemical Engineering 2007-2011
University of Dayton
Cum Laude

Research

Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7

Detected a series of molecular switches that are mediated by various ligand-induced allosteric events using molecular dynamics simulations

Investigated distinct structural changes (conformational states and correlated motions) between biased ligand binding to the receptor

Presented a clear connection between ligand binding and global transmembrane helical motions in CCR7 through a series of coupled allosteric events

Structural Mechanisms of Competitive Regulation of Ku70 Functions by Methylation and Acetylation

Quantified the biophysical effect of post-translational modifications (PTMs) on Ku70's dynamics

Investigated dynamic correlations between PTM sites and allosteric molecular switches

Identified global motions of protein dynamics coupled to electrostatics and charge removal by acetylation

AlgX Carbohydrate Binding Module Is Required for The Biosynthesis of Alginate in *Pseudomonas Aeruginosa*

Identified a set of residues essential for carbohydrate binding through docking studies of various alginate polymers into the carbohydrate binding module of AlgX

Predictions were confirmed through alginate binding, production, and acetylation in experimental studies conducted by collaborating group

Distinct Roles of E1 Heterodimer (APPBP1 and UBA3) in NEDD8 Activation

Successfully determined the requirement of a set of charged residues in protein-protein binding using a framework developed in our lab to study the role of electrostatics in protein-protein interactions

Predictions were tested experimentally to determine the requirement for an activation enzyme (E1), APPBP1, in the NEDD8 cascade conducted by a collaborating group

Undergraduate Student Researcher (University of Dayton)
Cloning, expressing, and characterizing cellulolytic enzymes

Successfully cloned and recombinantly expressed cellulolytic enzymes into Escherichia coli as potential bio-catalysts for the breakdown of lignocellulosic materials

Biochemically characterized the expressed enzymes (β–xylosidases) by determining the temperature and pH for optimum activity

Publications

Zied Gaieb, David D. Lo, Dimitrios Morikis, (2016) "Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7," Journal of Chemical Information and Modeling. Just Accepted Manuscript. doi:10.1021/acs.jcim.6b00367.

Zied Gaieb, Dimitrios Morikis, (2016) "Automated detection of molecular switches mediating large domain motions in membrane protein simulations" in preparation.

Zied Gaieb, Dimitrios Morikis, (2016) "Structural Insight into the Ligand-Free Behavior of CC Chemokine Receptor 7," in preparation.

Zied Gaieb*, Mendel Roth*, Teresa Hong, Markus Kalkum, Dimitrios Morikis, and WenYong Chen (2016) "Lysine Methylation Regulates KU70 Functions," awaiting experimental validation. *Shared first authorship.

Carl Z. Chen, Ronald D. Gorham Jr., **Zied Gaieb**, and Dimitrios Morikis, "Electrostatic Interactions between Complement Regulator CD46(SCR1-2) and Adenovirus Ad11/Ad21 Fiber Protein Knob," Molecular Biology International, vol. 2015, Article ID 967465, 15 pages, 2015. doi:10.1155/2015/967465.

Honors & Awards

NVIDIA GPU Award Finalist for Best GPU Poster 2015
Spring 2015 American Chemical Society National Meeting

Award for Computer Time on Anton 2014
National Research Council and National Academies of Science

Graduate Research Mentorship Fellowship 2014
University of California, Riverside

NBCR Best Poster Competition Winner 2012
NBCR (National Biomedical Computation Resource)
University of California San Diego

Dean's Distinguished Fellowship 2011-2012
University of California, Riverside

Dean's List 2008-2011
University of Dayton

Bro. William J. Wohlleben S.M. Memorial Scholarship 2008-2011
University of Dayton for academic excellence

International Scholarship 2007-2011
University of Dayton

Koehler Award Recipient 2008
University of Dayton for academic and leadership excellence

Oral Presentations

17th Annual UC Systemwide Bioengineering Symposium 2016
"Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7"
UC San Francisco, CA

3rd Annual CEPCEB Postdoc Symposium Program 2016
"Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7"
UC Riverside, CA

Biophysical Society Annual Meeting 2016
Protein-Dynamics and Allostery I Platform
"Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7"
Los Angeles, CA

(Invited Talk) **Department of Chemistry Seminar** 2016
"Analysis of Long Timescale Molecular Dynamics Simulations"
UC Riverside, CA

Department of Bioengineering Colloquium 2014
"Ligand-Induced Conformational Changes in CCR7"
UC Riverside, CA

13th Annual UC Systemwide Bioengineering Symposium 2012
"The role of Zinc and Calcium ions in MMP-14 and TIMP-2 Association"
UC Berkeley, CA

Poster Presentations

17th Annual UC Systemwide Bioengineering Symposium 2016
"Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7"
UC San Francisco, CA

1st SoCal TheoChem Symposium 2016
"Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7"
San Diego, CA

Spring 2016 ACS National Meeting 2016
"Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7"
San Diego, CA

Biophysical Society Thematic Meeting, 2015
Biophysics of Proteins at Surfaces: Assembly, Activation, Signaling
"Ligand-specific conformational changes in CCR7 coupled to selecting different signaling pathways upon CCL19 and CCL21 ligand binding"
Madrid, Spain

ACS National Meeting 2015
Best GPU Poster Finalist for the NVIDIA GPU Award
"Ligand-Specific Conformational Changes in CCR7"
Denver, Co

15th Annual UC Systemwide Bioengineering Symposium 2014
"The Structural Basis of DNA Repair Regulation by Ku70-Ku80 Acetylation"
UC Irvine, CA

14th Annual UC Systemwide Bioengineering Symposium 2013
"Computational Studies of Conformational Changes in CCR7 Coupled to Selecting Different Signaling Pathways upon CCL21 Ligand Binding"
UC San Diego, CA

National Biomedical Computation Resource 2012
"The role of Zinc and Calcium ions in MMP-14 and TIMP-2 Association"
UC San Diego, CA

Teaching Assistant

BIEN 001: Introductory Colloquium in Bioengineering 2015
Worked with the professor to organize the course, helped students develop presentation skills, and graded homework assignments

BIEN 135: Biophysics and Biothermodynamics 2012-2013
Led discussion sections on the following topics: the application of thermodynamic principles, biophysical properties of biomacromolecules, methods of characterizing protein properties and interactions

Worked with the professor to organize the course, graded homework assignments and reports, provided individual help to student groups

BIEN 165: Biomolecular Engineering 2014, 2016
Led hands-on projects on the following topics: modeling of biomolecules and biomolecular interactions, protein function, protein design, structure-based drug discovery.

Worked with the professor to organize the course, graded homework assignments and reports, provided individual help to student groups.

Research Mentor 2012-Present

Mentored graduate (Rohith Mohan, Reed Harrison, Nehemiah Zewde, Carl Chen) and undergraduate students (Ilya Lederman, Nehemiah Zewde) in various research related projects in the lab.

Outreach & Community Service

Inland Science and Engineering Fair 2016
Biophysical Society Award
Judged Science Fair Projects for Students in Grades 4-12
Bourns Technology Center

Science Olympiad projects 2012
Taught the principles and actively aided local middle school students in building a trebuchet and a tower for the Science Olympiad competition
Local middle school

Space Day 2012
Presented a science project at the local elementary school to demonstrate the basic physics of travelling in space
Jefferson Elementary school

10th Annual PossAbilities Triathlon 2012
Facilitated and organized a Triathlon tailored to people with physical disabilities
Loma Linda University

International Space Station Program 2012
Disseminated lectures at the Riverside Christian High School regarding data management. The program was intended to mentor students on designing an experiment that was to be run in the International Space Station
Riverside Christian High

Bioengineering Interdepartmental Graduate Student Association Treasurer 2012-2014
Maintained the organization's financial accounts
Worked in conjunction with Graduate Student Association and the Student Life financial team to thoroughly understand and implement financial policy
Produced official transitions report at the end of each term