UNIVERSITY OF CALIFORNIA
RIVERSIDE

# Multi-Scale Design of Surveillance and Regulation Within the Complement System

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

Doctor of Philosophy

in

Bioengineering

by

Reed Edward Harrison

September 2018

Dissertation Committee:

Dr. Dimitrios Morikis, Chairperson
Dr. Bir Bhanu
Dr. Chia-en A. Chang

The Dissertation of Reed Edward Harrison is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

# ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support and guidance of many individuals on a personal and professional level. Thanks to my advisor Dr. Dimitrios Morikis for his thoughtful and insightful mentorship, providing the necessary guidance and direction while allowing me to explore new approaches and develop new methods. Without his contribution, I would not have become the researcher that I am today, nor would I be planning to start a postdoctoral position at the University of California, San Diego within the coming month. Thanks also to my committee members Dr. Bir Bhanu and Dr. Chia-en A. Chang. Their advice and knowledge was invaluable to this dissertation and to my education.

I am grateful as well for the many organizations that have supported me and my research until now. Specifically, I acknowledge the University of California, Riverside (UCR) Chancellor's Fellowship, the National Science Foundation Integrative Graduate Education and Research Traineeship in Video Bioinformatics at UCR, the UCR Graduate Research Mentorship Program, the Whitaker International Summer Program, and the University of California President's Dissertation-Year Fellowship. The funding from these grants and fellowships allowed me to develop as a researcher and fostered new collaborations both locally and globally.

I also acknowledge the many contributions of our collaborators. I am grateful for the support of Dr. Alfonso Valencia and his lab members Juan Rodriguez, David de Juan, and Simone Marsili for sharing their experience and training me in methods of coevolution

(chapter 7). I am also grateful to Dr. Oscar Llorca and his lab member Andres Lopez-Perrote for providing an experimental perspective on the structure-function relationship of C3b (chapter 3).

I am grateful to current and former members of the Biomolecular Modeling and Design Lab (BioMoDeL) for their support and helpful discussions. Thanks to Dr. Ronald D. Gorham, Jr. for training me as a first-year graduate student on all the lab protocols, both experimental and computational, and providing helpful advice even after becoming a postdoctoral researcher at UMC Utrecht. Thanks to Dr. Zied Gaieb, Rohith Mohan, Nehemiah Zewde, and Rohaine Hsu for the many helpful conversations and feedback provided over the past few years. On a personal level, I am grateful for your friendship and I hope that we can keep in contact and communication for years to come.

Thanks to students and staff of the Bioengineering Department at UCR. I am grateful for the support and assistance of administrative staff who were always approachable and helpful. I am also grateful for the support from my classmates – their helpful discussions during classes and study sessions were always enjoyable and enhanced my education.

Finally, I would like to thank my family, especially my parents, for their continual support and understanding. It is thanks to their encouragement and support that I have been able to pursue a postgraduate degree.

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

1. Harrison, R. E. S.; Gorham, R. D.; Morikis, D. Energetic evaluation of binding modes in the C3d and Factor H (CCP 19-20) complex. *Protein Science.* **24,** 789–802 (2015).
2. López-Perrote, A., Harrison, R. E. S., Subías, M.; Alcorlo, M.; Rodríguez de Córdoba, S.; Morikis, D.; Llorca, O. Ionic tethering contributes to the conformational stability and function of complement C3b. *Molecular Immunology.* **85,** 137–147 (2017).
3. Harrison, R. E. S.; Mohan, R. R.; Gorham, R. D.; Kieslich, C. A; Morikis, D. AESOP: A Python Library for Investigating Electrostatics in Protein Interactions. *Biophysical Journal.* **112,** 1761–1766 (2017).

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

# DEDICATION

I dedicate this dissertation to

my parents and grandparents.

I am grateful for their continual

love, support, and encouragement.

ABSTRACT OF THE DISSERTATION

Multi-Scale Design of Surveillance and Regulation Within the Complement System

by

Reed Edward Harrison

Doctor of Philosophy, Graduate Program in Bioengineering
University of California, Riverside, September 2018
Dr. Dimitrios Morikis, Chairperson

As part of innate immunity, the complement system surveils biological tissues for foreign surfaces such as bacterial membranes and generates an immune response over a time period of minutes to hours. At a molecular level, surveillance of bacterial or foreign surfaces is accomplished by actions of complement component 3b (C3b). This protein covalently bonds to such surfaces and can interact with other complement proteins to form an enzyme that upregulates the concentration of C3b. This feed forward amplification allows for an immune response to be generated in a brief period of time, but necessitates some regulation to prevent off-target effects and states of autoimmunity. Of all complement regulators, Factor H (FH) is most influential modulator of C3b activity, binding C3b and downregulating the concentration of C3 convertase. Herein, we study the structure-function relationship within and between the proteins C3b and FH in order to understand molecular mechanisms of disease and to engineer new molecules that monitor or modulate

complement response. In these studies, we consider both static and dynamic structural features. From a static perspective, we have developed and applied a computational framework to predict electrostatic effects of mutations to study mechanisms of molecular interactions. From a dynamic perspective, we have leveraged biophysical methods to identify new functional conformational states and characterize conformational sampling of molecules. Importantly, these dynamical features range from harmonic fluctuations within a single conformational state that occur on the order of nanoseconds to domain motions that occur on the order of milliseconds. Given the range of time-scales that we study and the computational complexity associated with our methods, we use coarse grain methods to guide atomic simulations of protein dynamics for studies that involve domain rearrangements. Altogether, our studies contribute to a mechanistic understanding of complement structure and function, and we have engineered new molecules that target C3d and can be further optimized for therapeutic or diagnostic applications.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

# Introduction

## 1.1 Proteins Comprise Dynamical Systems

Like all matter, atoms that comprise amino acids are constantly in motion, vibrating harmonically in space over time.[1] As amino acids are polymerized into polypeptide chains, elements of secondary structure form. Alpha helices and beta strands emerge from stabilizing dipole-dipole interactions, hydrogen bonds.[2] Moreover, these elements of secondary structure adopt particular arrangements. While predicting the final structure of a protein given a particular sequence of amino acids is an ongoing area of research, experimental methods including nuclear magnetic resonance spectroscopy and X-ray crystallography have enabled determination of three-dimensional structures of proteins.[3] Hundreds of thousands of such structures have been deposited on curated databases such as the Protein Databank (www.rcsb.org) for ease of retrieval.[4] These structures, however, are only static snapshots of protein structure. Each protein will have some conformational freedom with its own distribution of conformers. This is especially notable in motor proteins that require conformational changes to walk across microtubules within cells; however, a wide range of proteins rely on conformational changes to open channels or to expose functional sites such as binding pockets that are essential to protein function.[5,6] Thus the structure of a protein is dynamical, changing over time and dependent on environmental factors such as solvent, ions, and the presence or absence of binding partners.

Understanding dynamical changes in protein structure is critical for developing new therapeutics that influence protein structure and function. To facilitate such studies, we adopt a theoretical approach validated by experiment to the degree possible. Specifically, we use standard force fields that parameterize bonded and non-bonded interactions to solve Newton's equations of motion and simulate atomic motion in proteins.[7,8] In these simulations, we solvate proteins of interest in a periodic box with physiological conditions. From observing protein behavior in these simulations, we can predict an ensemble of conformations that a protein may adopt and facilitate studies to modulate protein structure and function. Using such strategies, other researchers have been able to predict conformational changes associated with G-protein binding to G-protein coupled receptors and the opening and closing of chaperones.[6,9,10] Importantly, findings in these studies often correspond well to experimental evidence.[11]

### 1.2 Immune Surveillance and Regulation in the Complement System

The complement system is part of the innate immune system in humans, surveilling biological tissues for foreign surfaces and generating an immune response over a time period of minutes to hours, opsonizing bacterial membranes and attracting circulating leukocytes to promote phagocytosis of bacterial cells. Additionally, complement can form a membrane attack complex capable of lysing cells with susceptible membranes such as gram-negative bacteria.

An immune response in the complement system is initiated by the classical, mannose-binding lectin, and alternative pathways.[12] Both the classical and mannose-binding lectin

pathways recognize pathogen-associated molecular patterns (PAMPs); however, these pathways differ in recognition mechanisms.[12] The classical pathway relies on opsonization of foreign surfaces by antibodies, while the mannose-binding lectin pathway recognizes particular carbohydrate topologies associated with pathogens. In contrast to the classical and mannose-binding lectin pathways, the alternative path does not recognize PAMPs. Instead, the alternative pathway generates a continuous, low-grade immune response that is non-specific in nature, opsonizing both foreign and native tissues.[13] In healthy individuals, this response does not result in states of autoimmunity due to the presence of regulators on surfaces of self and in the fluid phase that prevent an immune response. The most important feature of the alternative pathway is the ability to amplify any immune response generated by the classical and mannose-binding lectin pathways leading to a substantial immune response over a time period of minutes to hours.[13] All three pathway converge to the formation of an enzyme called C3 convertase.[12] This enzyme participates in the alternative pathway to cleave serum complement component 3 (C3) into fragments C3a, an anaphylatoxin, and C3b, an opsonin. C3b may interact with other complement species to form additional C3 convertase resulting in feed-forward amplification of the immune response. As an anaphylatoxin, C3a attracts circulating leukocytes to sites of infection, while C3b covalently binds to surfaces of self and non-self via a thioester domain (TED). C3b is also capable of associating with C3 convertase to form C5 convertase. Similar to C3 convertase, C5 convertase is an enzyme that cleaves serum complement component 5 (C5) into C5a, an anaphylatoxin, and C5b. C5b initiates the formation of the

membrane attack complex. This complex forms a pore in cell membrane that leads to osmotic lysis of cells.

Due to the presence of a non-specific immune response, nature has evolved a number of complement regulators that prevent an immune response directed at surfaces of self.[12,14–16] Molecules such as Factor H (FH), decay-accelerating factor (DAF), membrane cofactor protein (MCP), and complement receptor 1 (CR1) act to control progression of an immune response. With the exception of FH, these regulators are expressed on the extracellular face of the cell membrane. FH, on the other hand, is not bound to any surface and diffuses through extracellular space. By recognition of particular patterns associated with native tissues, FH binds to membranes of self non-covalently and multivalently.[17] Some regulators such as CR1, DAF and FH regulate the complement system by promoting the degradation of convertases that forms, while other regulators such as CR1, DAF, MCP, and FH act to prevent the formation of convertases.[18] As these regulators are present to a higher extent near or on surfaces of self, regulation of the complement system is directed towards surfaces of self. Thus, both the basal response generated by the alternative pathway and any response generated by foreign surfaces such as bacteria do not result in states of autoimmunity.

Beyond recognition of foreign surfaces, generation of chemotactic gradient for circulating leukocytes, and opsonization of foreign surfaces, the complement system also acts as an adjuvant for adaptive immunity – promoting a humoral response.[19]

### 1.3 Genetic Variation Can Perturb Complement Surveillance and Regulation

Variation in genetic material can produce proteins with slightly different amino acid sequence. As a result, different isoforms of a protein can emerge in a population.[20,21] Each of these isoforms can be referred to as a polymorphism. Polymorphisms may confer risk for disease or result in disease so long as the isoform observed in a population.[21] Alternatively, variations in amino acid sequence that are not prevalent in a population are termed mutations. In the context of this dissertation, mutations of interest are typically involved in disease, though not all mutations affect function. Mutations are typically annotated by specific amino acid changes that differ from the canonical protein sequence. Despite differences in definition, both mutations and polymorphism arise from natural variation.

Variation in amino acid sequence does not always have effects on protein structure or function. Some changes may be silent, especially when located near non-functional regions of the protein.[21] Other changes may be immediately apparent and result in a protein with decreased functionality. Such changes in functionality may be subtle – perhaps lowering binding affinity for some target or ligand. In such cases, changes in functionality may not become apparent in biological systems until some environmental feature is perturbed by age, disease, trauma, or some other external factor.

Within the complement system, functional perturbations of protein amino acid sequence are of medical relevance.[16,22–24] Mutations and polymorphisms in molecules that initiate, propagate, or regulate complement response can result in diseases of the kidney, eye, and connective tissues. More recently, ongoing research is investigating the role of

complement proteins in diseases such as lupus and Alzheimer's.[22,25–27] While we will focus on relatively few mutations and polymorphisms in complement, we attempt to understand how exactly these perturbations in amino acid sequence affect protein structure – both static and dynamical – and protein function. Thus our research seeks to predict how protein structure is coupled to function in order to understand molecular etiology of disease and develop new diagnostic and therapeutic molecules to promote a healthy state.

### 1.4 *In Silico* Approaches to Study and Engineer Protein Structure

As protein dynamics occurs across multiple time-scales – from harmonic fluctuations of protein domains that occurs on the order of picoseconds to rearrangements of protein domains that occurs on the order of microseconds to milliseconds – a number of methods have been developed to model protein behavior with varying degrees of complexity.[8] As a result of improvements in computational hardware and with the advent of graphical processing unit computation, using *in silico* methods to study protein conformations and probe conformational dynamics has become increasingly tractable.

In particular, molecular dynamics (MD) simulations are routinely performed for simulations of biomolecules in aqueous solutions. In principle, MD simulations predict motions of atoms by first calculating the potential energy arising from bonded and non-bonded interactions within the simulation system according to

$$V(\vec{X}, A) = \sum_{bonds} k_{bond}(l - l_0)^2 + \sum_{angles} k_{angle}(\theta - \theta_0)^2 + \sum_{torsions} k_{torsion,n}[1 + \cos(n\varphi - \delta_n)] + \sum_{i<j} \left\{ \epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon r_{ij}} \right\} \quad (\mathbf{1.1})$$

where the parameters for every term of the equation are provided by an appropriate force field such as CHARMM.[8,28] In **Equation 1.1** $\vec{X}$ represents the positions of all atoms, and $A$ represents topological information for the system of interest such as bond connectivity. Together, $\vec{X}$ and $A$ provide all the information necessary to calculate $l$, $\theta$, $\varphi$, and $r_{ij}$ which represent bond lengths, bond angles, torsions, and pairwise distances for non-bonded interactions, respectively. The first such term models bond lengths as harmonic springs with some associated spring constant, $k_{bond}$, and an equilibrium bond length, $l_0$. The second term models bond angles in a Hookean manner with a spring constant, $k_{angle}$, and an equilibrium bond angle, $\theta_0$. The third term models torsion angles according to a Fourier series with a force constant of $k_{torsion,n}$ for each term, $n$, of the series. The final term models non-bonded interactions with Lennard-Jones and Coulombic potentials for all atom pairs with significantly non-zero contributions. The Lennard-Jones part uses parameters that describe the energy well depth, $\epsilon_{ij}$, and the distance between atoms $i$ and $j$ where the potential is zero, $\sigma_{ij}$. The Coulombic part uses parameters that describe the dielectric constant, $\varepsilon$, and atomic charges $q_i$ and $q_j$ on atoms $i$ and $j$, respectively. Though the unit elementary charge requires integer value, charges on individual atoms are typically

represented as partial charges. Partial charges are parameterized to approximate the overall charge distribution expected in monomers of biopolymers by forcefields such as CHARMM. Given the potential energy of the system, the force on an individual atom by the full system can be calculated according to

$$F_i(\vec{X}, A) = -\nabla_{x_i} \mathrm{V}(\vec{X}, A) \textbf{ (1.2)}.$$

Then, given forces on individual atoms, Newton's equations of motion are integrated to find a new coordinate position for atom after some sufficiently small time step, $\Delta t$, according to a Taylor series expansion

$$\vec{x}_i(t + \Delta t) = \vec{x}_i(t) + \frac{d\vec{x}_i(t)}{dt}\Delta t + \frac{d^2\vec{x}_i(t)}{dt^2}\frac{\Delta t^2}{2} + \cdots + \frac{d^m\vec{x}_i(t)}{dt^m}\frac{\Delta t^m}{m!} \textbf{ (1.3)}$$

of order $m$. The order in the expansion can be chosen to satisfy a specific error tolerance or computational efficiency. After successive integration cycles, time-series for positions of each atom in the system over the simulation time period are generated. These time-series are also referred to as trajectories and convey dynamical information about protein structure. On modern workstations, full-atom dynamics of small proteins (~100 amino acids) can be simulated at a rate of approximately 100 ns/day, facilitating studies of atomic fluctuations, side chain rearrangements, and loop motions over the course of a couple days or weeks. For significantly larger systems (~1000-2000 amino acids or more) or for studies of protein behavior that occurs on microsecond to low millisecond time scales such a domain motions, supercomputing resources are typically required.

When full atom simulations are not feasible, coarse grain methods can be used to study protein behavior. Such methods may involve MD with coarse grain force fields that parameterize groups of atoms rather than single atoms or involve simplified models that

have been shown to predict protein behavior despite ignoring some physical principles. Elastic network models (ENM) are examples of the latter type of coarse grain method and can be used to study harmonic motions in proteins that occur over longer time-scales. In ENMs nodes are specific atoms or groups of atoms from a protein structure, and edges are springs that connect all node pairs separated by a distance less than a cutoff value, usually around 10 Å. As edges represent springs, each edge can be parameterized with some associated force constant. This constant can be uniform across all edges or vary according to some additional criteria. A harmonic potential for an edge between nodes and can be described with Hooke's law as

$$V_{ij} = \frac{\gamma}{2}\left(\left[\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2 + \left(z_j - z_i\right)^2\right]^{\frac{1}{2}} - s_{ij_0}\right)^2 \quad \textbf{(1.4)}$$

where x, y, an z are Cartesian coordinates for the positions of nodes $i$ and $j$, $\gamma$ is the spring constant, and $s_{ij_0}$ is the equilibrium distance between nodes.[29] Typically, spring equilibria are assumed to be the distances observed in three dimensional structures of proteins. Given the potentials for all edges, a matrix of second derivatives can be organized into a Hessian matrix for each node pair according to

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 V_{ij}}{\partial x_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial x_i \partial y_j} & \frac{\partial^2 V_{ij}}{\partial x_i \partial z_j} \\ \frac{\partial^2 V_{ij}}{\partial y_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial y_i \partial y_j} & \frac{\partial^2 V_{ij}}{\partial y_i \partial z_j} \\ \frac{\partial^2 V_{ij}}{\partial z_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial z_i \partial y_j} & \frac{\partial^2 V_{ij}}{\partial z_i \partial z_j} \end{bmatrix} \textbf{(1.5)}.^{29}$$

Furthermore, the Hessian matrices for all node pairs can be organized into the Hessian super-matrix

$$H = \begin{bmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & \ddots & \vdots \\ H_{N1} & \cdots & H_{NN} \end{bmatrix} \textbf{(1.6)}$$

for $N$ number of nodes.[29] Decomposition of the Hessian super-matrix

$$H = M\Lambda M^T \textbf{ (Eq. 1.7)}$$

results in the eigenvectors, $M$, that describe the direction of vibrational modes and eigenvalues, $\Lambda$, that describe the frequencies of the vibrational mode for the associated eigenvector.[29] ENMs have close correspondence with thermal fluctuation factors from crystal structures of proteins[30] and have been used to explore conformational transitions in structures,[5,10,31,32] all from coarse grain information.

For studies of protein association, classical molecular dynamics simulations are often intractable. To study such behavior, Brownian dynamics (BD) simulations have been developed to simulate random-walk diffusion of molecules in solution. In BD atomic motion is simulated according to the Langevin equation which takes the form

$$m_i \frac{d^2\vec{x}_i}{dt^2} = F_i(\vec{X}, A) - \zeta_i \frac{d\vec{x}_i}{dt} + \varepsilon_i(t) \textbf{ (1.8)}$$

for a single atom, $i$.[8] Here, the force term, $F_i$, is the same as in classical MD (**Equation 1.2**), and the friction coefficient, $\zeta_i$, is set to a large value to facilitate random walk behavior.[8] The final term, $\varepsilon_i(t)$, represents random forces from water molecules that act on atom $i$. As BD is a stochastic method, thousands of trajectories must be acquired to predict meaningful associative behavior. To facilitate the large number of trajectories required, BD algorithms often make simplifying assumptions such as a rigid-body approximation for molecules being simulated.

Aside from dynamical studies, methods have been developed to quantify structural features of protein structures. These methods often rely on principles of biophysics to study the structure-function paradigm of proteins. For some proteins, electrostatic interactions play a significant biological role, accelerating the formation of encounter complexes in protein-protein interactions.[33] To study this behavior, the divergence of the electric displacement, $D = \varepsilon(r)\nabla\varphi(r)$, can be related to the distribution of charge via the Poisson-Boltzmann equation as

$$-\nabla \cdot [\varepsilon(r)\nabla\varphi(r)] = e\left[\sum_{i=1}^{F} z_i \delta(r - r_i) + \sum_{j=1}^{M} z_j n_j^0 e^{\frac{-ez_j\varphi(r)}{k_B T}}\right] \textbf{(1.9)}$$

where $\varepsilon(r)$ and $\nabla\varphi(r)$ are the dielectric coefficient and electrostatic potential, respectively, at position $r$. For applications in proteins, the distribution of fixed charges, $F$, in a protein is modeled as a set of Dirac delta functions where the charge for atom $i$ is $z_i$ at position $r_i$. Ions in solvent are modeled implicitly as mobile charges, $M$, according to a Boltzmann distribution where $z_j$ is the charge of ion $j$, $n_j^0$ is the bulk number of ions per unit volume for ion $j$, and $\varphi(r)$ is the electrostatic potential at position $r$. The numerator of the exponent, $-ez_j\varphi(r)$, then represents the potential of mean force required to bring the ion to position $r$ from infinity. While the Poisson-Boltzmann equation is non-linear, the equation can be linearized under the assumptions that the solution contains counterions of the same valence, here of ±1, and $ez_j\varphi(r) \ll k_B T$. After solving for electrostatic potentials, $\varphi(r)$, electrostatic free energies can be calculated according to

$$G_{electrostatic} = \frac{1}{2}\sum_i q_i \varphi(r_i) \textbf{(1.10)}$$

where $q_i$ is the charge of an atom at position $r_i$, and $\varphi(r_i)$ is the electrostatic potential at position $r_i$. Thus the Poisson-Boltzmann method can be used to quantify electrostatic properties of proteins, including salt-bridge interactions and electrostatic hot-spots on protein surfaces. In this dissertation, we apply such methods to thermodynamic cycles of protein-protein association to estimate how perturbations in protein structure affect free energies of binding.

In summary, *in silico* methods can be leveraged to study protein behavior in terms of static and dynamical properties. Researchers are actively involved in improving existing methods and developing new methods to improve computational toolboxes for understanding the protein structure-function paradigm. When investigating a particular aspect of protein behavior, both the types of biophysical property and associated time-scales that are relevant for the behavior of interest must be considered. These criteria are necessary to select appropriate *in silico* methods to study protein structure-and function.

### 1.5 Overview

Herein, we will leverage methods from computational chemistry and biophysics to understand the structure and function relationship of complement proteins. Furthermore, we will leverage such information to engineer new molecules with desired properties that have diagnostic and research potential. The following chapters will start by introducing AESOP, a computational framework with methods to study electrostatic structures of proteins. Electrostatics is a recurring theme in the complement system as a number of pattern recognition processes must take place in the context of the extracellular space.

Thus, long-range interactions may be key component of complement behavior.[34,35] Subsequently, we will discuss studies of immune surveillance in the complement system, focusing on the role of C3 and C3 cleavage fragments. We leverage these studies to design new peptides to serve as biomarkers of C3d deposition. Next, we discuss studies of immune regulation in the complement system, focusing on the role of FH and the effect of a particular single nucleotide polymorphism on FH structure and function. Finally, we summarize conclusions from each chapter in this dissertation.

## 1.6 References


1. Levitt, M. Molecular Dynamics of Native Protein. *J. Mol. Biol.* **168,** 621–657 (1983).

2. Dobson, C. M. Protein Folding and Misfolding. *Nature* **426,** 884–890 (2003).

3. Aloy, P. & Russell, R. B. Structural systems biology: Modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **7,** 188–197 (2006).

4. Rose, P. W. *et al.* The RCSB Protein Data Bank: Views of structural biology for basic and applied research and education. *Nucleic Acids Res.* **43,** D345–D356 (2015).

5. Zheng, W., Brooks, B. R. & Hummer, G. Protein conformational transitions explored by mixed elastic network models. *Proteins* **69,** 43–57 (2007).

6. Zheng, W. & Brooks, B. R. Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: Myosin versus kinesin. *Biophys. J.* **89,** 167–178 (2005).

7. MacKerell, J. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102,** 3586–3616 (1998).

8. Adcock, S. a & Mccammon, J. A. Molecular Dynamics : Survey of Methods for Simulating the Activity of Proteins Molecular Dynamics : Survey of Methods for Simulating the Activity of Proteins. **106,** 1589–1615 (2006).

9. Miao, Y. & McCammon, J. A. Mechanism of the G-protein mimetic nanobody binding to a muscarinic G-protein-coupled receptor. *Proc. Natl. Acad. Sci.* **115,** 3036–3041 (2018).

10. Tekpinar, M. & Zheng, W. Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model. *Proteins Struct. Funct. Bioinforma.* **78,** 2469–2481 (2010).

11. Noé, F. *et al.* Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 4822–7 (2011).

12. Mathern, D. R. & Heeger, P. S. Molecules great and small: The complement system. *Clin. J. Am. Soc. Nephrol.* **10,** 1636–1650 (2015).

13. Thurman, J. M. & Holers, V. M. The Central Role of the Alternative Complement Pathway in Human Disease. *J. Immunol.* **176,** 1305–1310 (2006).

14. E., M., A.P., H. & P.N., B. Functional anatomy of complement factor H. *Biochemistry* **52,** 3949–3962 (2013).

15. Martínez-Barricarte, R. *et al.* The molecular and structural bases for the association of complement C3 mutations with atypical hemolytic uremic syndrome. *Mol. Immunol.* **66,** 263–273 (2015).

16. K.Liszewski, M. & Atkinson, J. P. Complement regulators in human disease: Lessons from modern genetics. *J. Intern. Med.* **277,** 294–305 (2015).

17. Perkins, S. J., Nan, R., Li, K., Khan, S. & Miller, A. Complement Factor H-ligand interactions: Self-association, multivalency and dissociation constants. *Immunobiology* **217,** 281–297 (2012).

18. Zewde, N., Gorham, R. D., Dorado, A. & Morikis, D. Quantitative Modeling of the Alternative Pathway of the Complement System. *PLoS One* **11,** e0152337 (2016).

19. Elsen, J. M. H. Van Den. A Crystal Structure of the Complex between Complement Receptor 2 and its Ligand C3d. **608,** 608–612 (2011).

20. Forneris, F. *et al.* Regulators of complement activity mediate inhibitory mechanisms through a common C3b-binding mode. *EMBO J.* 1–17 (2016). doi:10.15252/embj.201593673

21. Bhattacharya, R., Rose, P. W., Burley, S. K. & Prlić, A. Impact of genetic variation on three dimensional structure and function of proteins. *PLoS One* **12,** 1–22 (2017).

22. Schramm, E. C. *et al.* Genetic variants in the complement system predisposing to age-related macular degeneration: A review. *Mol. Immunol.* **61,** 118–125 (2014).

23. Fremeaux-Bacchi, V. *et al.* Brief report Mutations in complement C3 predispose to development of atypical hemolytic uremic syndrome. **112,** 4948–4952 (2008).

24. Andersen, G. R. Mutations in complement C3 from aHUS patients. *Blood* **125,** 2316–2318 (2015).

25. Leffler, J., Bengtsson, A. a & Blom, A. M. The complement system in systemic lupus erythematosus: an update. *Ann. Rheum. Dis.* **73,** 1601–6 (2014).

26. Dos Santos, R. N., Morcos, F., Jana, B., Andricopulo, A. D. & Onuchic, J. N. Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Rep.* **5,** 13652 (2015).

27. Hong, S. *et al.* Complement and microglia mediate early synapse loss in Alzheimer

mouse models. *Science (80-. ).* **352,** 712–716 (2016).

28. Best, R. B. *et al.* Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J. Chem. Theory Comput.* **8,** 3257–3273 (2012).

29. Atilgan, A. R. *et al.* Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80,** 505–515 (2001).

30. Bahar, I., Atilgan, A. R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. *Fold. Des.* **2,** 173–181 (1997).

31. Gur, M., Madura, J. D. & Bahar, I. Global transitions of proteins explored by a multiscale hybrid methodology: Application to adenylate kinase. *Biophys. J.* **105,** 1643–1652 (2013).

32. Sen, T. Z., Feng, Y., Garcia, J. V., Kloczkowski, A. & Jernigan, R. L. The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models. *J. Chem. Theory Comput.* **2,** 696–704 (2006).

33. Kieslich, C. A., Gorham, R. D. & Morikis, D. 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 (2011).

34. Kieslich, C. A. *et al.* Exploring Protein-Protein and Protein-Ligand Interactions in the Immune System using Molecular Dynamics and Continuum Electrostatics. *Curr. Phys. Chem.* **2,** 324–343 (2012).

35. Gorham, R. D., Kieslich, C. A. & Morikis, D. Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Ann. Biomed. Eng.* **39,** 1252–1263 (2011).

CHAPTER 2

# AESOP – A Python Library for Investigating Electrostatics in Protein Interactions

## 2.1 Introduction

Electrostatics have been shown to play a pivotal role in guiding the association of a protein with its respective binding partner, forming the encounter complex.[1–4] Additionally, electrostatic interactions can act to thermodynamically stabilize a protein complex.[4] In attempts to improve protein activity, enhancing association rates is preferable to thermodynamic stabilization of the complex as interactions are typically diffusion limited.[1] In order to aid investigations into the electrostatics of protein interactions as well as facilitate attempts to re-engineer protein behavior, our lab previously developed AESOP (Analysis of Electrostatic Structures of Proteins) as a computational tool.[5–8] Here, we present an implementation of AESOP in Python[9] with increased functionality and capabilities for parallel processing. This framework may be used to both compare electrostatic similarity of protein families and dissect the electrostatic contribution of individual amino acids to the free energy of association.[5–8]

Though several other computational tools have been developed for similar purposes, AESOP is the only platform, to our knowledge, that is focused on protein electrostatics and offers multiple computational methods for both family-based and single-structure based analyses. In comparison, Surface Diver and PIPSA have been developed to compare electrostatic potentials across families of structurally homologous proteins. Surface Diver accomplishes this quantitative comparison without prior structural superpositioning

through spherical harmonic decomposition[10], whereas PIPSA requires superpositioning and instead allows for comparisons between common structural regions that are functionally similar.[11,12] Similar to PIPSA, AESOP requires superposed structures and quantitatively compares electrostatic potentials in a common grid space. PIPSA performs this comparison via calculation of a Hodgkin similarity index, a Carbo similarity index, or the average difference between electrostatic potentials in some specified spherical region or across the entire protein shell region[11,12], while AESOP uses a Reynolds linear similarity measure, discussed further in methods, for every grid point in the system.[7] A feature unique to AESOP, however, allows for comparisons to be made for a library of mutants from an input protein structure. This comparison is performed by mutating all ionizable amino acids within the protein structure to alanine such that each mutant within the library has a single mutation. The end result is a computational analysis to identify electrostatic hotspots within protein structure. Another computational tool named DrugScore$^{PPI}$ uses statistical potentials derived from a database of experimental mutations to predict effects of alanine mutations at the interface of protein complexes.[13] With this method, mutations may involve any type of amino acid. In contrast, AESOP uses Poisson-Boltzmann electrostatics and a thermodynamic cycle of solvation to predict effects of mutations involving ionizable amino acids on free energies of association relative to the parent structure. **Table 2.1** provides a brief comparison of features available across the computational tools discussed here and illustrates the flexibility of AESOP in studying the role of electrostatics in protein structure.

AESOP implements family-based comparisons of electrostatic potentials with the `ElecSimilarity` class and perturbation based analyses to gauge the contribution of

single amino acids with the `Alascan` and `DirectedMutagenesis` classes. Compared to previous iterations of the framework, the current implementation in Python has been coded in an object-oriented manner with each computational method implemented as a unique class. Furthermore, the current version of AESOP offers several additional features including energy minimization and structural superpositioning that are made possible through other python libraries such as Modeller. The end result is a more streamlined, efficient computational method for comprehensive investigations of protein electrostatics.

**Table 2.1:** Summary of features and methods across several computational tools

| Feature | AESOP | Surface Diver | PIPSA | DrugScore$^{PPI}$ |
|---|---|---|---|---|
| Alanine Scan | Y | N | N | Y |
| Directed Mutagenesis | Y | N | N | N |
| Electrostatic Similarity | Y | Y | Y | N |
| Local installation | Y | Y | Y | N |
| Cross-platform | Y | N | N | - |
| Parallelized implementation | Y | N | Y | - |
| Mutant generation | Y | N | N | - |
| Free energy predictions | Y | N | N | Y |
| Web server | alanine scan | N | Y | Y |

## 2.2 Theory

### *2.2.1 Electrostatic potentials*

Our framework utilizes the PDB2PQR[14,15] and APBS[16] software to generate grid-based electrostatic potentials for protein systems in solvated and reference states. AESOP uses a linearized adaptation of the Poisson-Boltzmann equation from APBS to calculate electrostatic potentials from unit or partial charges located within the protein, the electric permittivity of the protein and solvent, and ionic strength and accessibility of the solvent. Furthermore, AESOP avoids grid artifacts where possible by using consistent grid

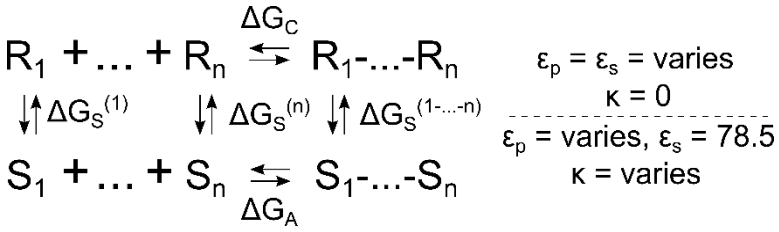



**Figure 2.1:** Thermodynamic cycle for binding of protein subunits in a reference (top) and solvated (bottom) state. In the reference state, the dielectric constant, ε, is assumed to be uniform in the protein, $\varepsilon_p$, and solvent, $\varepsilon_s$; however, in the solvated state, the uniform dielectric constants may differ between the protein and solvent. For the solvent, we set the dielectric coefficient to 78.5, the value for water at room temperature. A parameter for ion accessibility, κ, is held constant within each state. This variable is related to the ionic strength of the solvent.

specifications, canceling out artifacts according to the thermodynamic cycle of **Figure 2.1**.

*2.2.2 Electrostatic similarity*

Pairwise comparisons between grid-based potentials can be made according to

$$ESD(\varphi_A, \varphi_B) = \frac{1}{N}\sum_{k=1}^{N}\frac{|\varphi_A - \varphi_B|}{max(|\varphi_A|, |\varphi_B|)} \quad (\mathbf{2.1})$$

where $N$ is the number of grid points, $\varphi$ is an electrostatic potential from either system A or B at the same grid point $k$, and ESD is the electrostatic similarity distance.[7] An ESD of zero indicates identical electrostatic potentials at every grid point. Using this metric, we can compare a family of proteins on the basis of their electrostatic potentials in consistent grid dimensions.

*2.2.3 Thermodynamic cycle*

In order to predict the effect of mutations on the free energy of binding for some protein complex, our framework computes Coulombic and solvation free energies in each state of a thermodynamic cycle shown in **Figure 2.1**.[4–8] This task is accomplished by invoking APBS multiple times for each state of the thermodynamic cycle with consistent grid space

specifications. For each calculation the ionic strength, protein dielectric coefficient, and solvent dielectric coefficient are set according to the parameters of the solvated or reference state. Each state includes a set of protein subunits, $\{1, \dots, n\}$, and the resulting protein complex, $1 - \cdots - n$. Each protein subunit and the resulting complex is analyzed in a reference state, $\{R_1, \dots, R_n, R_{1-\cdots-n}\}$, and a solvated state, $\{S_1, \dots, S_n, S_{1-\cdots-n}\}$. While the end user is able to adjust ionic strength and dielectric coefficients, we suggest that the default values be used since we have previously parameterized calculations to agree with experimental data.[5]

Association free energies are calculated from mathematical operations on solvation and Coulombic free energies. First the free energies of solvation (vertical processes in **Figure 2.1**) are calculated according to

$$\Delta G_S = G_S - G_R \ (\mathbf{2.2})$$

where $\Delta G_S$ is the free energy of solvation calculated for each vertical process. Then, $\Delta\Delta G_S$ is calculated according to

$$\Delta\Delta G_S = \Delta G_S^{(1-\cdots-n)} - \sum_{k=1}^{n} \Delta G_S^{(k)} \ (\mathbf{2.3})$$

and represents the difference in solvation free energy between the complex and the sum of the individual subunits. The change in Coulombic free energy of complex formation, $\Delta G_C$ (upper horizontal process in **Figure 2.1**), is calculated similarly according to

$$\Delta G_C = G_C^{(1-\cdots-n)} - \sum_{k=1}^{n} G_C^{(k)} \ (\mathbf{2.4})$$

where $G_C$ represents the Coulombic free energy for the complex or an individual subunit.

Finally, the free energy of association, $\Delta G_A$ (lower horizontal process in **Figure 2.1**), is calculated according to

$$\Delta G_A = \Delta\Delta G_S + \Delta G_C \ (\mathbf{2.5}).$$

The resulting association free energy quantitatively describes the thermodynamic behavior of the protein system when considering only electrostatic forces, neglecting hydrophobic interactions and entropic effects.

## 2.3 Methods

AESOP is implemented in Python 2.7 and consists of a class for each method discussed here (3 in total) and a number of accessory functions. In the spirit of open access, the latest version of the framework is available to all users through PyPi[17] and GitHub[18]. Examples and walkthroughs of real-world biological analyses along with detailed documentation can be found on the AESOP webpage (aesop.readthedocs.io). AESOP primarily depends on APBS[16], PDB2PQR[14,15], Modeller[19,20], ProDy[21], NumPy[22], SciPy[23], NetworkX[24], and Matplotlib.[25] Notably, the ProDy library facilitates VMD-style selection strings to easily specify regions within a protein structure. In addition, AESOP currently uses the Python library multiprocessing[9] for parallel processing. The organization of AESOP is inspired by libraries such as SciPy that encapsulate computational methods with programming objects.[23]

### *2.3.1 Electrostatic similarity*

Within the AESOP framework, the `ElecSimilarity` class has been implemented to setup and perform an analysis of electrostatic similarities within any family of protein

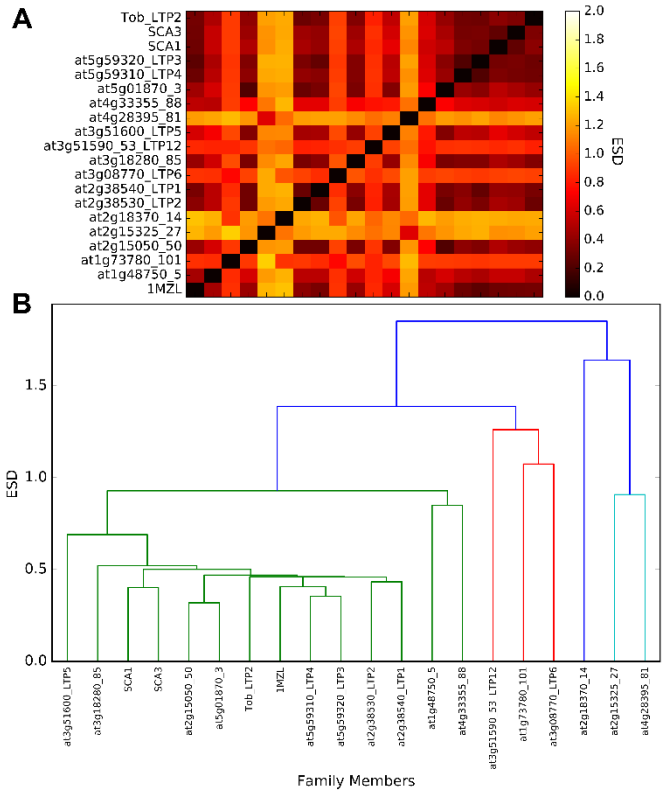



**Figure 2.2:** (A) Example output from the `plotESD` function. ESD values closer to zero are more similar in terms of electrostatics. (B) Example output from the `plotDend` function where differences in ESD between family members are shown in a hierarchical dendrogram.

structures. We advise the user to superpose these structures prior to running them through AESOP analysis; however, the `ElecSimilarity` class can superpose or center coordinates through the Modeller and ProDy libraries.[19–21] After the user provides structural files, the `ElecSimilarity` class calls PDB2PQR to add hydrogen atoms, place charges, assign van der Waals radii in each structure according to the PARSE forcefield, and convert the PDB file format to an APBS-readable PQR format.[14] In the case that only one structural file is provided, the `ElecSimilarity` class will generate a family of alanine mutants before converting structural files to PQR files. For each structure, electrostatic potentials are calculated in the same grid space and compared in a pairwise

manner by calculating the ESD (**Equation 2.1**). The results of this analysis can be returned as a distance matrix or a dendrogram (**Figure 2.2**). This analysis may take a significant amount of time to complete if many structures are being compared or if the grid space contains a large number of vertices. Electrostatic similarity studies are useful for comparison of global electrostatic properties of families of homologous structures[26,27] or structurally homologous repeat modules in protein structures.[28]

*2.3.2 Alanine scan mutagenesis*

To predict whether electrostatic interactions contribute favorably or unfavorably to association, AESOP implements a computational alanine scan with the `Alascan` class. This method accepts a single structure file and generates a library of mutants through side-chain truncation. While the mutation scheme should not generate new clashes in protein structure, the end-user has the option to perform minimization of all structures by implementing the conjugate gradient descent algorithm from Modeller. All mutations involve only ionizable amino acids, and each mutant structure contains a single mutation. Once generated, solvation and Coulombic free energies are calculated for each structure in the structure library, including the parent structure. Afterwards, the user will have the option to view electrostatic free energies of association in a bar graph, in a network if the NetworkX[24] Python library is installed, or return a matrix of free energies of association or solvation corresponding to the mutant identifiers (**Figure 2.3**). In the default bar graph and network, we typically report relative free energies of association where the parent free energy of association is subtracted from the association free energies of each mutant. Thus, the resulting value for the parent will be zero. Mutations that have more significant effects

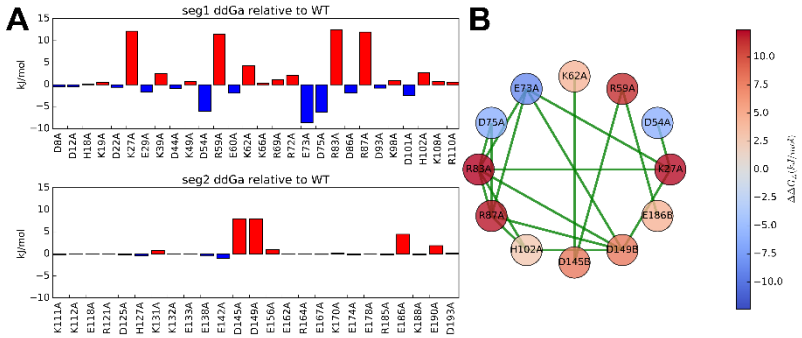



**Figure 3:** (A) Example output from `plotScan` for an alanine scan showing perturbations of free energy of association relative to wild type. Perturbations below zero indicate gain of binding mutations, and perturbations above zero indicate loss of binding mutations. Effects that are significant are typically greater or less than 2.5 kJ/mol compared to the parent. (B) Example output from `plotNetwork` for an alanine scan showing electrostatic interactions between residues with changes in free energy of association relative to parent outside thermal effects (±2.5 kJ/mol). Edges indicate the presence of an interaction; nodes represent individual amino acids; and node colors are scaled according to the free energy of association. Each amino acid label consists of a one letter amino acid code, the residue number, and the chain ID, in that order.

on complex formation will have larger magnitude deviations from the parent association free energy, observed as outliers outside expected thermal fluctuations ($kT \sim \pm 2.5$ kJ/mol). The alanine scan is a perturbation method to identify individual amino acids that have significant effect in the formation of the parent protein complex (loss of binding mutations). It can also identify mutations or sites of mutations that will enhance association (gain of binding mutations) in studies of design of protein-protein interfaces.[29–31]

*2.3.3 Directed mutagenesis*

Similar to alanine scan mutagenesis, the `DirectedMutagenesis` class facilitates calculation of free energies of association according to the same thermodynamic cycle. The difference is that the end user must specify a list of exact mutations to be performed, though only mutations involving ionizable amino acids should be considered. If AESOP will be performing a mutation to alanine, it is best to use the `Alascan` class as it will be more

efficient and use a simple side-chain truncation method. To support more complicated mutations, the `DirectedMutagenesis` class depends on Modeller.[19,20] As a result of the mutation process, backbone coordinates may be slightly perturbed around the mutated residue(s). Thus, this method requires extra calculations that the `Alascan` class does not require. Furthermore, each structure assessed with this method is minimized using the conjugate gradient descent algorithm in Modeller to remove clashes in protein structure. The `DirectedMutagenesis` class supports the same outputs as the `Alascan` class. The directed mutagenesis studies provide information on the significance of the mutated amino acids in the stability of the structure of the parent complex[31], and can also be used in the design of tailored protein-protein interfaces.[32]

*2.3.4 Web server portal*

The AESOP alanine scan method is also implemented as a web server (http://biomodel.engr.ucr.edu/software) using the Django web framework[33] to reduce technical barriers and allow for rapid analysis. Results are also displayed with interactive graphs to simplify interpretation without manual scripting. The modular implementation of AESOP and the high-scalability of Django facilitates the addition of new features into the framework and webserver while servicing a growing number of users.

## 2.4 Conclusion

The AESOP framework provides a simple interface for quantitative comparisons of electrostatic potentials generated by proteins and predicting electrostatic contributions of ionizable amino acids to protein association. AESOP also contains accessory functions to

generate figures for each computational method. As a result, the framework has potential to aid optimization of protein-protein interactions in redesigning protein interfaces, as well as investigate evolution of electrostatics across protein families. In past studies, AESOP has shown agreement with experimental mutagenesis studies involving protein interactions that are known to involve electrostatic interactions.[5] The developers of AESOP welcome and encourage collaboration through the AESOP GitHub repository (https://github.com/BioMoDeL/aesop/). Installation instructions and example cases are at the AESOP website (aesop.readthedocs.io), and an archived version of the source code is deposited on Zenodo (http://doi.org/10.5281/zenodo.269610).

## 2.5 References


1. McCammon, J. A. Darwinian biophysics: electrostatics and evolution in the kinetics of molecular binding. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 7683–4 (2009).

2. Zhou, H. X. Enhancement of protein-protein association rate by interaction potential: accuracy of prediction based on local Boltzmann factor. *Biophys. J.* **73,** 2441–2445 (1997).

3. Frembgen-Kesner, T. & Elcock, A. H. Striking effects of hydrodynamic interactions on the simulated diffusion and folding of proteins. *J. Chem. Theory Comput.* **5,** 242–256 (2009).

4. Elcock, A. H., Sept, D. & Mccammon, J. A. Computer Simulation of Protein− Protein Interactions. *J. Phys. Chem. B* 1504–1518 (2001). doi:10.1021/jp003602d

5. Gorham, R. D. *et al.* An evaluation of Poisson-Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. *Biopolymers* **95,** 746–754 (2011).

6. Gorham, R. D., Kieslich, C. A. & Morikis, D. Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Ann. Biomed. Eng.* **39,** 1252–1263 (2011).

7. Kieslich, C. A., Gorham, R. D. & Morikis, D. 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 (2011).

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

9. Python Software Foundation. Python Language Reference, version 2.7.

10. Długosz, M., Trylska, J., Długosz, M. & Trylska, J. Electrostatic similarity of proteins : Application of three dimensional spherical harmonic decomposition. *J. Chem. Phys.* **129,** 015103 (2008).

11. Blomberg, N., Gabdoulline, R. R., Nilges, M. & Wade, R. C. Classification of Protein Sequences by Homology Modeling and Quantitative Analysis of Electrostatic Similarity. *Proteins Struct. Funct. Genet.* **37,** 379–387 (1999).

12. Richter, S., Wenzel, A., Stein, M., Gabdoulline, R. R. & Wade, R. C. webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res.* **36,** 276–280 (2008).

13. Kruger, D. M. & Gohlke, H. DrugScore PPI webserver : fast and accurate in silico alanine scanning for scoring protein – protein interactions. *Nucleic Acids Res.* **38,** 480–486 (2010).

14. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32,** 665–667 (2004).

15. Dolinsky, T. J. *et al.* PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35,** 522–525 (2007).

16. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10037–41 (2001).

17. Python Software Foundation. Python Package Index. Available at: https://pypi.python.org.

18. GitHub Inc. GitHub. Available at: https://github.com/. (Accessed: 9th January 2016)

19. Fiser, A., Kinh Gian Do, R. & Sali. Modeling Loops in Protein Structures. *Protein Sci.* **9,** 1753–1773 (2000).

20. Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).

21. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27,** 1575–1577 (2011).

22. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Struture for Efficient Numerical Computation. *Comput. Sci. {&} Eng.* **13,** 22–30 (2011).

23. Eric, J., Travis, O., Pearu, P. & al., et. SciPy : Open source scientific tools for Python. (2001).

24. A. Hagberg, D. S. & Swart, P. Exploring network structure, dynamics and function using NetworkX. *Proc. 7th Python Sci. Conf.* 11–15 (2008).

25. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9,** 99–104 (2007).

26. López de Victoria, A., Kieslich, C. A., Rizos, A. K., Krambovitis, E. & Morikis, D. Clustering of HIV-1 Subtypes Based on gp120 V3 Loop electrostatic properties. *BMC Biophys.* **5,** 1–17 (2012).

27. Kieslich, C. A. & Morikis, D. The Two Sides of Complement C3d: Evolution of Electrostatics in a Link between Innate and Adaptive Immunity. *PLoS Comput. Biol.* **8,** e1002840 (2012).

28. Kieslich, C. A., Vazquez, H., Goodman, G. N., López de Victoria, A. & Morikis, D. The effect of electrostatics on factor H function and related pathologies. *J. Mol. Graph. Model.* **29,** 1047–1055 (2011).

29. Harrison, R. E. S., Gorham, R. D. & Morikis, D. Energetic evaluation of binding modes in the C3d and Factor H (CCP 19-20) complex. *Protein Sci.* **24,** 789–802 (2015).

30. Gorham, R. D., Rodriguez, W. & Morikis, D. Molecular analysis of the interaction between staphylococcal virulence factor Sbi-IV and complement C3d. *Biophys. J.* **106,** 1164–1173 (2014).

31. Mohan, R. R., Gorham Jr, R. D. & Morikis, D. A theoretical view of the C3d:CR2 binding controversy. *Mol. Immunol.* **64,** 112–122 (2015).

32. Liu, Y., Kieslich, C. A., Morikis, D. & Liao, J. Engineering pre-SUMO4 as efficient substrate of SENP2. *Protein Eng. Des. Sel.* **27,** 117–126 (2014).

33. Django Software Foundation. Django. (2013).

CHAPTER 3

# Intramolecular, Ionic Tethering in Complement Component 3b Contributes to Conformational Stability and Function

## 3.1 Introduction

Complement is a major component of innate immunity that participates in various functions such as fighting microbial infections, clearance of apoptotic cells, and modulation of adaptive immune response. Amplification of complement response in the alternative pathway (AP) takes place when large conformational changes occur in the C3 protein due to cleavage of C3 into C3a and C3b. This cleavage exposes regions of C3b required for binding surfaces and other proteins, especially complement regulators. Binding of Factor B (FB) forms the C3-pro-convertase (C3bB) complex. FB can then be cleaved by Factor D (FD) to assemble an active C3-convertase (C3bBb), a complex with a serine protease domain that catalyzes hydrolysis of a peptide bond in C3 to form C3b. This C3b can subsequently bind FB in the absence of regulation, eventually forming additional C3-convertase and initiating the amplification loop in the complement response.[1,2] The ability of C3b to bind FB relies on a significant conformational rearrangement after cleavage of C3. Following cleavage, the thioester-containing domain (TED) of C3 is translated and rotated approximately 60 Å, exposing FB binding sites on C3b and facilitating formation of C3 convertase. This process also exposes the internal thioester bond allowing binding to surfaces through a covalent bond.[3–5]

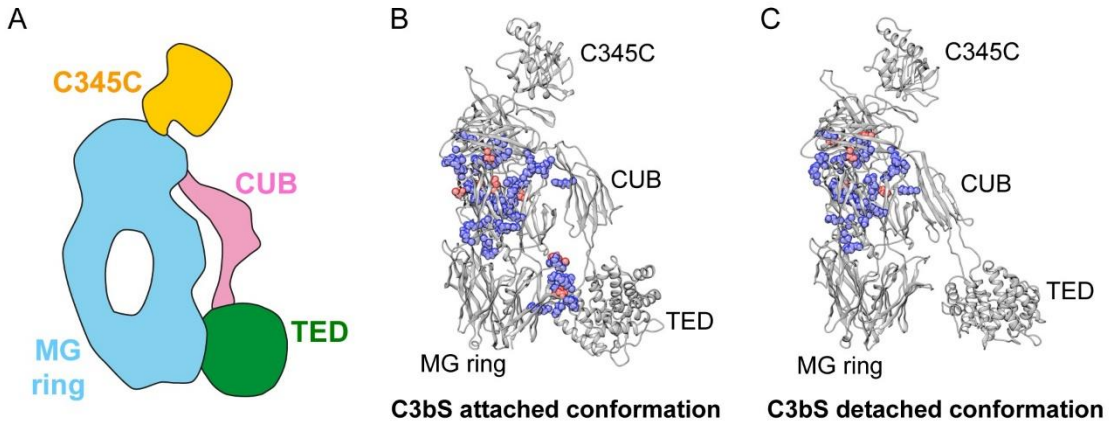



**Figure 3.1:** Electrostatic interactions in C3b maintain the attached conformation. (A) Schematic representation of C3b architecture based on the crystal structure (PDB ID 2I07). Domains of C3b, as presented in this manuscript, are color coded and labeled. Blue represents the MG ring; orange represents C345C; pink represents CUB; and green represents TED. CUB and TED comprise the swinging arm of C3b, and the arm is attached to the MG ring at CUB. (B) Crystal structure of C3b (PDB ID 2I07) highlighting residues predicted to be important for maintaining the structure of C3b in the attached conformation using electrostatic interactions. Residues that result in electrostatic energies of association (**Equation 3.2**) that differ from the parent by 2.5 kJ/mol or more are shown as colored spheres while those below the threshold are shown as colored regions of the licorice model for the structure of C3b. Red indicates that the electrostatic energy of association is more unfavorable in the mutant than in the parent, while blue indicates that the electrostatic energy of association is more favorable in the mutant than in the parent.

C3b consists of two polypeptide chains, α' and β, linked by a disulfide bridge and organized in multiple domains as revealed in several crystallographic structures.[3–6] A core of eight macroglobulin (MG1–MG8) domains (hereafter referred to as the MG ring) is connected to a C345C domain at one end of the molecule and to a long swinging arm. This arm contains both a "complement C1r/C1s, Uegf, Bmp1" (CUB) domain at the proximal end and a TED domain at the distal end. For simplicity, the structure of C3b can be described as a scaffold, comprising the MG ring and C345C domain, and a pendulum-like entity, comprising the CUB and TED domains (**Figure 3.1A**). The TED includes a reactive thioester moiety that anchors C3b to the surface of cells. The first four short consensus repeats (SCR) of Factor H (FH), the prototypic regulator of the alternative complement

pathway, interact with the MG ring and the TED domain of C3b.[7] This regulator accelerates the decay of the C3-convertase enzymes and acts as a cofactor for the Factor I (FI) mediated proteolytic inactivation of C3b.[1,2] The interaction between C3b and FH SCR 1–4 is only possible after a conformational change occurring during the transition from C3 to C3b which positions the TED domain near the MG1 domain.[7] Mutations in regions of C3b involved in binding FH are found in several diseases including atypical hemolytic uremic syndrome (aHUS).[1,2,8] Other regulators, such as membrane cofactor protein (MCP; CD46), decay accelerating factor (DAF; CD55) and complement receptor 1 (CR1; CD35), bind C3b at similar sites as FH SCR 1-4 despite variations in amino acid composition. In all such cases, a precise positioning of the TED domain contacting the MG1 domain from the macroglobulin core ring is required for sustaining the interaction between C3b and the regulators, as illustrated by recent crystal structures[6] and the mapping of mutations.[8]

It is currently known that C3b coexists in several conformations, and the positioning of the TED is not always adjacent to the MG ring, as in the crystal structures. Images of C3b obtained using single molecule electron microscopy (EM) revealed that the TED domain was detached from the position found in the crystal structures in some C3b molecules.[9,10] Time-resolved FRET measurements using a fluorophore attached to the TED domain[11] as well as the solution structure of C3b studied using X-ray and neutron scattering[12] further supported these findings. Also, proteins from pathogens such as the extracellular fibrinogen-binding protein (Efb-C) from S. aureus induce allosteric inhibition of complement by promoting conformational changes in C3b that perturb the positioning of the TED.[13] Importantly, Chen et al. suggest that C3b binds differently to other proteins

when the position of the TED domain is held away from the MG ring by Efb-C.[13] Hereafter, we will refer to C3b conformations as "attached" and "detached" when the TED region is held together or apart from the MG ring, respectively, though no covalent chemical bonds are formed between the TED and MG ring in the "attached" conformation. The use of this term in the manuscript implies non-bonded interaction, and the "attachment" can be reversible and/or transient. Our collaborators have previously obtained low-resolution models for the detached conformation using EM.[10] These models are just a sampling of several possible conformers where the TED no longer contacts the MG ring. In one such structure solved previously, the CUB domain remains in the proximity of the MG ring and the TED is extended. Modeling shows that the linker connecting CUB and TED is sufficient to account for the separation. This arrangement might not be the most general, but models obtained by X-ray and Neutron Scattering analyses also show a tendency of the CUB domain to remain closer to the MG ring when the TED domain extends away.[12] This separation could have been exacerbated in the electron microscopy analysis using negative staining, as observed in images obtained by other authors.[9] High-resolution models for a detached conformation have been obtained using X-ray and Neutron Scattering.[12] Recently, solution studies performed with the related C4b protein revealed similar conformational flexibility and its dependence of NaCl concentration.[14]

Here we have expanded our understanding of the conformational changes and associated functional changes in C3b by combining the predictions from an electrostatic model of protein structure with experimental imaging of the C3b conformations in single molecules to investigate the preference of C3b for the attached and detached

conformations. We describe an ionic tether holding the TED domain to the MG ring, which can be affected by ionic strength, mutations and polymorphisms, including the common R102G polymorphism. We find differences in the conformational sampling of the two forms of C3b – C3bS with R102 and C3bF with G102 – that contribute to an understanding of why the C3b isoforms behave differently in vivo. In addition, we find that assembly of the C3bB complex still occurs when the TED domain is displaced from the MG ring, indicating that subsequent formation of the active C3bBb enzyme and amplification of the immune response is compatible with such a conformational rearrangement. Overall, the electrostatic attachment of the TED domain to the MG ring and the conformational flexibility of C3b highlight dynamic structural changes of C3b in complement regulation.

## 3.2 Methods

### *3.2.1 Flexible Fitting of C3b to Electron Density*

All computational analyses were performed using structures with PDB identifiers 5fo7 and 4mrk.[6,12] The 5fo7 crystal structure describes the conformation of C3bS isoform when the TED is attached to the MG ring. The 4mrk model is a solution scattering structure for the C3bS isoform when the TED is detached from the MG ring. Since the solution scattering structure is a coarse grain model, the atomic positions were estimated by superposing domains of 5fo7–4mrk. Loops were then modeled using Modeller.[15,16] Full atom models for both 5fo7 and 4mrk with intra-chain disulfide bridges were then subjected to steepest descent minimization in Gromacs[17] until the greatest force was less than 1000 kJ/mol/nm or 50,000 steps were reached. Neighbor searching was performed using the

Verlet algorithm with a cutoff distance of 1.2 nm for van der Waals and Coulombic interactions. Electrostatic interactions were calculated using the particle mesh Ewald algorithm with the default grid spacing of 0.1 nm. Further minimization of C3b models with both inter and intra-chain disulfide bridges was performed in NAMD[18] with a conjugate gradient algorithm first without solvent and then with solvent until total energies stabilized (approximately 10,000 steps). In minimization and all subsequent simulations, the non-bonded cutoff distance was set to 1.25 nm, the switching distance was set at 1 nm and hydrogens covalently bonded to heavy atoms were constrained by the SHAKE algorithm. After initial minimization, a water box using the TIP3P model was constructed leaving a margin of 1.2 nm spacing on all sides of the protein to prevent interactions across periodic boundaries. Next sodium ions were added to neutralize the net charge of the system and additional sodium and chloride ions were added to bring the ionic strength of the system to 150 mM. After minimization, the solvated system was heated to 310 K step-wise using increments of 10 K over 64 ps with a 2 fs time step. The system was further equilibrated over several 50 ps runs with constraints for all protein atoms of 10, 5, 2, and 1 kcal/mol/Å$^2$, using a time step of 1 fs for the first run and 2 fs for all subsequent runs. A final equilibration run was performed with a time step of 2 fs constraining only backbone atoms of amino acids with a spring constant of 1 kcal/mol/Å$^2$. Finally, a production run was performed with a time step of 2 fs and backbone constraints of 1 kcal/mol/Å$^2$ for 100 ps. The final frame of these simulations provided an estimate of the full-atom structures for C3bS in both a TED-attached and TED-detached conformation. To estimate the corresponding conformations of C3bF, the Dunbrack rotamer library[19] was used to perform

the R102G mutation on the corresponding C3bS conformation using the `swapaa` command in UCSF Chimera.[20]

*3.2.2 Thermodynamic analysis of C3b*

An ensemble of mutant structures was built for each C3b structure by mutating one ionizable amino acid (R, K, H, D, E) per structure. Mutations were performed by side-chain truncation to form alanine in AESOP, and every ionizable amino acid is mutated to form a comprehensive ensemble of mutants. For each member of the ensemble, electrostatic (Coulombic and solvation) energies are calculated with the use of PDB2PQR[21,22] and APBS[23] according to previously established methods.[24–29] In all calculations, the maximum grid spacing was 1 Å; the ionic strength was varied between 0 and 150 mM of monovalent counterions. The dielectric constants used were 20 and 78.53 for the protein and solvent, respectively.[30] Using a previously described thermodynamic cycle[24–26,28,29], free energies of association are predicted according to

$$\Delta G_a = \Delta\Delta G_s + \Delta G_C \ (\mathbf{3.1}).$$

$\Delta\Delta G_s$ and $\Delta G_C$ represent the difference in solvation energies and Coulombic energies, respectively, between the whole C3b molecule and the individual α' and β chains of C3b. Typically, we report electrostatic energies of association relative to the value associated with the parent structure lacking mutations as

$$\Delta\Delta G_a = \Delta G_a^{mutant} - \Delta G_a^{parent} \ (\mathbf{3.2})$$

where all values less than zero represent gain of binding mutations and all values greater than zero represent loss of binding mutations. Moreover, any gain of binding mutation suggests that the parent residue is involved in electrostatic interactions with other residues

in the native structure in an overall energetically unfavorable manner, while loss of binding mutations suggest the parent residue was involved in electrostatic interactions that were overall favorable.

In order to isolate those mutations that are predicted to affect the electrostatic energy of conformational transition from TED attached to TED detached, we calculate electrostatic energies of association for each conformational state

$$\Delta G_{conformation} = \Delta G_a \ (\mathbf{3.3})$$

to calculate the electrostatic energy of conformational transition

$$\Delta\Delta G_{transition} = \Delta G^{attached}_{conformation} - \Delta G^{detached}_{conformation} \ (\mathbf{3.4}).$$

An electrostatic energy of conformational transition below zero favors the attached conformation, while a value above zero favors the detached conformation. Values for this electrostatic energy of conformational transition are then compared between F and S variants of C3b according to

$$\Delta\Delta\Delta G_{transition,\ S-F} = \Delta\Delta G^{S}_{transition} - \Delta\Delta G^{F}_{transition} \ (\mathbf{3.5})$$

in order to compare the energetics of the transition between the TED attached and detached conformations in the two isoforms. A value above zero suggests that the corresponding mutation causes the TED detachment to be more energetically favorable in the F isoform than in the S isoform. In contrast, a value below zero suggests that the corresponding mutation causes the TED detachment to be more energetically favorable in the S variant than in the F variant. Therefore, by examining $\Delta\Delta\Delta G_{transition,\ S-F}$ for individual residues, the relative importance of amino acids can be gauged in the transition process.

## 3.3 Results

### *3.3.1 Electrostatic analysis reveals two structural regions in C3b tethered by electrostatic interactions*

We performed an in silico alanine scan for every ionizable amino acid in the C3bS attached conformation, one at a time, and calculated the electrostatic energy of association for each alanine mutant as a difference from the electrostatic energy of association of the parent (wild type, see Materials and Methods). This computational analysis provided a list of residues within C3b where electrostatic interactions significantly contributed to hold the structure together. These included loss of binding mutations, where the replaced amino acid was important for favorable electrostatic interactions within C3b (**Figure 3.1B**, blue color) as well as gain of binding mutations, indicating that the replaced amino acid was making a net unfavorable electrostatic interactions in C3b (**Figure 3.1B**, red color). All the residues indicated in this manuscript are numbered according to C3 sequence nomenclature, which includes the signal peptide in the count and agrees with the numbering scheme of 5fo7. **Table 3.1** summarizes the list of significant positions, those where the absolute values of the differences between C3b and the alanine mutant were greater than the thermal energy (1 $k_BT$ = 2.5 kJ/mol). Interestingly, some of these computationally predicted residues have been described previously as mutations or polymorphisms that interfere with complement regulation. For instance, the R592W mutation associated with aHUS[31] results in lower binding of C3b to FH and CD46 and in resistance to regulation by MCP but not by FH.[29] Additionally, the R102G mutation results in decreased binding of C3b to FH, reduced sensitivity to FI cleavage, and conferred risk

for developing age-related macular degeneration (AMD) later in life.[27,28,32] The predicted residues clustered around two broad areas. Most residues were located in the MG ring and the adjacent CUB domain, probably contributing to its structural stability; however, we also found residues with electrostatic character at the interface of the MG ring and TED that likely promote association of the TED to the MG ring (**Figure 3.1B**).

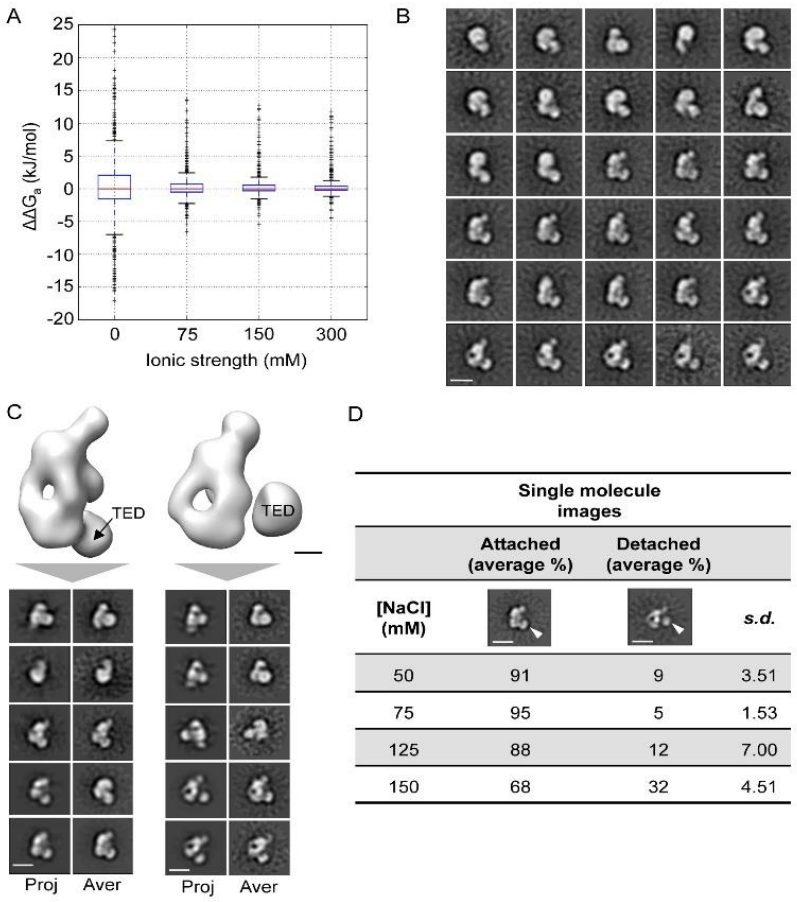
In order to compare the effects of mutations between C3bS in the attached and detached conformation, we performed another alanine scan for our theoretical model of the detached conformation based on PDB 2mrk.[12] Once again, the computational analysis provided a list of residues within C3b where electrostatic interactions significantly contributed to hold the structure together; however, in the detached conformation there are no significant mutations at the interface of the TED and MG ring at 150 mM ionic strength of NaCl (**Figure 3.1C**). This observation suggests that short-range, electrostatic interactions have been eliminated between these two domains in the detached conformation. **Table 3.1** displays the values for the free energy of association relative to the parent in the TED domain that are significant in the attached conformation but insignificant in the detached conformation. These mutations include D1029A, E1032A, E1035A, K1036A, and E1040A.

**Table 3.1:** Changes in free energy of association for α' and β chains of C3b relative the parent structure (S isoform). All mutations resulting in a change outside of thermal fluctuations (2.5 kJ/mol) for the attached conformation are displayed. Corresponding changes in free energy of association are displayed for both attached and detached conformations, and the domain where the residue is located is noted. Cells shaded blue indicate that the electrostatic energy of association is more unfavorable in the mutant than in the parent and outside thermal fluctuations, while red indicates that the electrostatic energy of association is more favorable in the mutant than in the parent and outside thermal fluctuations. $\Delta\Delta G_a$ is calculated according to **Equation 3.2**.

| | $\Delta\Delta G_a$ (kJ/mol) | | | | $\Delta\Delta G_a$ (kJ/mol) | | |
|---|---|---|---|---|---|---|---|
| **Mutant** | **Attached** | **Detached** | **Domain** | **Mutant** | **Attached** | **Detached** | **Domain** |
| E233A | -5.39 | -6.07 | MG | R258A | 4.39 | 8.93 | MG |
| D61A | -3.99 | -1.05 | MG | E586A | 4.70 | 4.19 | MG |
| D832A | -3.73 | -3.98 | MG | E822A | 4.86 | 9.55 | MG |
| E99A | -3.12 | -0.51 | MG | K927A | 4.97 | 4.37 | MG |
| E223A | -2.87 | -3.62 | MG | E769A | 5.00 | 4.34 | MG |
| D599A | -2.81 | -2.43 | MG | K66A | 5.02 | 0.77 | MG |
| D572A | -2.79 | -0.34 | MG | D1449A | 5.26 | 7.44 | MG |
| K1036A | -2.53 | -0.26 | TED | R148A | 5.33 | 4.70 | MG |
| K891A | 2.55 | 0.34 | MG | E1032A | 5.45 | 0.13 | TED |
| E294A | 2.55 | 4.07 | MG | D777A | 5.45 | 3.02 | MG |
| D337A | 2.55 | -0.80 | MG | D828A | 5.49 | 3.66 | MG |
| K608A | 2.61 | 2.00 | MG | E226A | 5.50 | -1.31 | MG |
| K783A | 2.69 | 0.48 | MG | E1040A | 6.06 | 0.04 | TED |
| K1436A | 2.75 | 3.35 | MG | R592A | 6.59 | 2.35 | MG |
| K65A | 2.90 | 0.72 | MG | E803A | 6.68 | 8.28 | MG |
| K100A | 2.98 | 0.29 | MG | R290A | 7.08 | 9.05 | MG |
| D1029A | 3.03 | 0.08 | TED | R937A | 8.01 | 1.82 | CUB |
| R848A | 3.21 | 4.69 | MG | K119A | 8.05 | 0.41 | MG |
| E197A | 3.23 | 3.47 | MG | R102A | 8.51 | -0.02 | MG |
| K607A | 3.35 | 4.85 | MG | K813A | 9.18 | 5.39 | MG |
| K615A | 3.37 | 1.47 | MG | R834A | 9.71 | 9.46 | MG |
| K1478A | 3.47 | 5.02 | MG | K104A | 10.24 | 1.00 | MG |
| K97A | 3.57 | 0.70 | MG | K136A | 10.74 | 6.98 | MG |
| D819A | 3.63 | 3.65 | MG | D588A | 10.82 | 3.31 | MG |
| R979A | 3.67 | 1.27 | CUB | K600A | 11.92 | 13.97 | MG |
| E1035A | 3.75 | 0.08 | TED | K610A | 12.01 | 9.69 | MG |
| K225A | 3.86 | 12.65 | MG | K930A | 12.41 | 8.51 | MG |
| K263A | 4.01 | 1.72 | MG | | | | |

*3.3.2 NaCl affects the tethering between TED domain and the MG ring*

The electrostatic component of contacts between the TED domain and the MG ring suggested that ionic strength could potentially affect the conformation of C3b. We first performed electrostatic energy calculations to examine the effect of several concentrations of monovalent counterions (0, 75, 150 and 300 mM) for the attached conformation of C3bS (**Figure 3.2A**). Increasing the counterion concentration decreased the energetic effects on the electrostatic energy of association (loss or gain of binding) in C3b mutants. As the ionic strength increases, the Debye length decreases, screening out the effects of point charge interactions greater than the Debye length. Assuming a solvent dielectric constant of 78.5 and a temperature of 300 K, the Debye length is 13.6, 8.6, and 7.9 Å at 50, 125, and 150 mM NaCl, respectively. This is consistent with the presence of more electrostatic interactions at lower ionic strengths predicted in the model. Consequently, we expect some long-range electrostatic interactions (7.9–13.6 Å) to be important for anchoring the TED to the MG ring, depending on the ion concentration. Notably, there were several loss-of-binding mutations at 150 mM that switched to gain-of-binding mutations at 0 mM ionic strength (data not shown), where free energies of association relative to the parent were outside thermal effects at both ionic strengths. These mutations include E197A, E226A, E294A, and E586A, and they are all located on the MG ring. E197 and E226, specifically, may participate in unfavorable, long-range Coulombic interactions (>8 Å) with the CUB at lower ionic strengths and in favorable, short-range (<8 Å) Coulombic interactions with the CUB at higher ionic strengths.




| Single molecule images | | | |
|---|---|---|---|
| | Attached (average %) | Detached (average %) | |
| [NaCl] (mM) |  |  | *s.d.* |
| 50 | 91 | 9 | 3.51 |
| 75 | 95 | 5 | 1.53 |
| 125 | 88 | 12 | 7.00 |
| 150 | 68 | 32 | 4.51 |

P=0.0020

**Figure 3.2:** Effect of ionic strength on the conformation of C3b. (A) Boxplot showing the effect of individual mutations on electrostatic energy of association (**Equation 3.2**) for all ionizable amino acids at ionic strengths of 0, 75, 150, and 300 mM NaCl in the C3bS structure with the TED attached to the MG ring. The red line indicates the mean; the blue rectangle indicates the bounds of the upper and lower quartile; the upper and lower whiskers indicate the range of the remaining quartiles; and '+' symbols indicate outliers, those mutations that have more significant effects. (B) 2D reference-free averages of C3b. A set of representative averages of one of the experiments is shown. Scale bar represents 10 nm. (C) Assignment of C3b views to attached or detached TED conformation. 2D reference-free averages from "B" were compared to two different 3D models of C3b (TED attached, left panel; TED detached, right panel) using the `multirefine` command implemented in EMAN. Representative pairs of theoretical projections of the models (Proj) and averages assigned to each class (Aver) are shown for each model. Scale bars represent 10 nm for the 2D-averages and 25 Å for the 3D models. The position of the TED domain in the 3D models is indicated. (D) Effect of ionic strength of the conformation of C3b (S variant), analyzed using electron microscopy. C3b images were classified as attached or detached conformations after image processing by comparison with typical C3b images observed before. The percentage of molecules counted for each conformation as an average from three independent experiments is indicated. The position of the TED domain is indicated by a white arrow. Scale bar represents 10 nm.

In order to assess our structural predictions, our collaborators observed the conformation of single C3b molecules to assess the involvement of NaCl concentration.[33] For this, C3bS purified from human plasma was dialyzed against several buffers with varying NaCl concentrations (ranging from 50 to 150 mM NaCl), and the conformation of the molecules was analyzed by EM. For these EM experiments, molecules were adsorbed on a support film and their conformation fixed using a staining agent prior to their observation inside the microscope. 2D-reference free averages of around 10000 molecules for each experiment were analyzed by image processing methods with the aim of classifying and averaging images with a similar conformation (**Figure 3.2B**, set of 2D averages obtained for a representative experiment is shown). Note that C3b preferentially interacts with the EM grid in a typical triangular side view, which represents the majority of the dataset, with some other less abundant orientations. To assign each particle to either the attached or detached conformation of C3b, each 2D average was compared with all the theoretical projections of 3D low resolution models of C3b with the TED attached or detached (**Figure 3.2C**). Each experiment was performed 3 times, and the number of molecules assigned to each conformation, including the statistical error, was calculated (**Figure 3.2D**). We found that increasing the concentration of NaCl in the preparation resulted in a larger ratio of C3b molecules where the TED domain was not interacting with the MG ring, and the differences between 50 mM and 150 mM were statistically significant ($P = 0.0020$), suggesting an effect of electrostatic interactions on the stability of C3b.

*3.3.3 Mutations in C3b can predispose detachment of the TED domain*

In order to compare the relative propensity for transitioning from a TED-attached to TED-detached conformation between S and F C3b isoforms, we calculated the relative electrostatic free energies of conformational transition ($\Delta\Delta\Delta G_{transition,\ S-F}$) as described in Materials and Methods. Using this strategy, we were able to predict that TED detachment is more favorable for the F isoform than the S isoform by 3.55 kJ/mol (**Figure 3.3A**). Additionally, we identified five mutants predicted to have significantly different effects on $\Delta\Delta G_{transition}$ of the detachment process between C3bS and F isoforms. Namely, E659A and K1036A are predicted to amplify the energetic difference for TED attachment between the S and F isoforms causing $\Delta\Delta\Delta G_{transition,\ S-F}$ to increase, while D996A, E1032A, and E1035A are predicted to diminish the energetic difference for TED-detachment between the S and F isoforms causing $\Delta\Delta\Delta G_{transition,\ S-F}$ to decrease. The existence of these mutations that have different effects depending on the isoform of C3b suggests a delicate balance of electrostatic interactions at the interface of the TED and MG ring. Multiple Coulombic interactions, both favorable and unfavorable, are shown in **Figure 3.3B** in the form of a distance heatmap of charged, side-chain atoms in ionizable amino acids of C3bS. Specifically, favorable, charge-charge interactions with a distance of separation less than 6 Å occur between K119-E1040, K104-E1032, R102-E1032, R102-E1035, and K97-E1032, while an unfavorable, charge-charge interaction with a distance of separation less than 6 Å occurs between R102-K1036. Thus, R102 participates in several favorable and unfavorable interactions, though we expect the net interaction to be favorable (**Figure 3.3C**). We predict that the R102G mutation reduces the magnitude of the

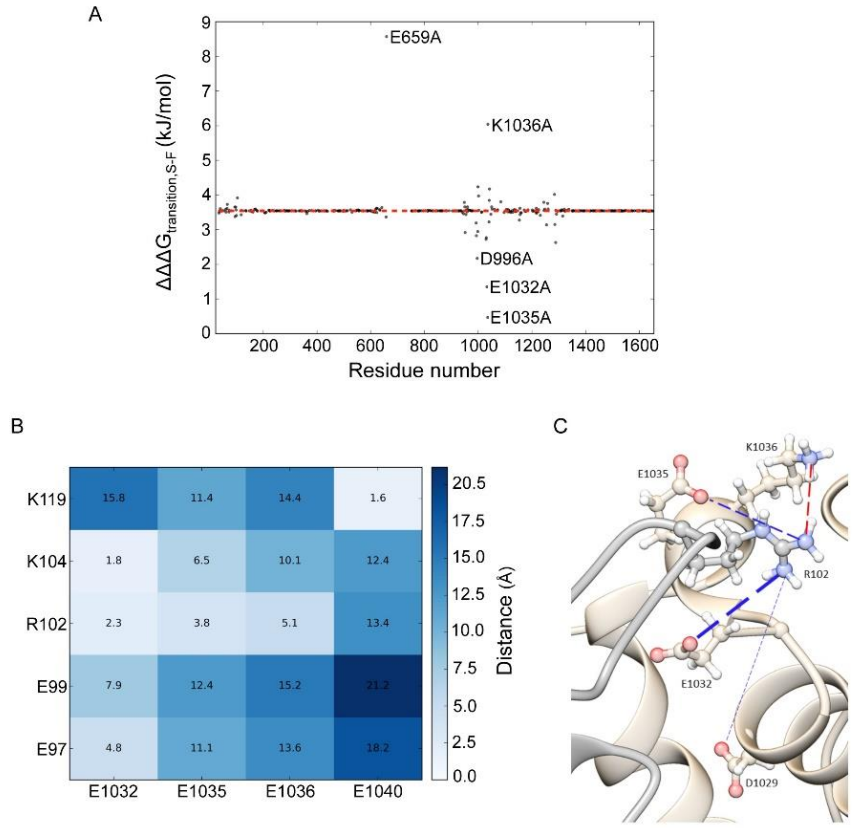
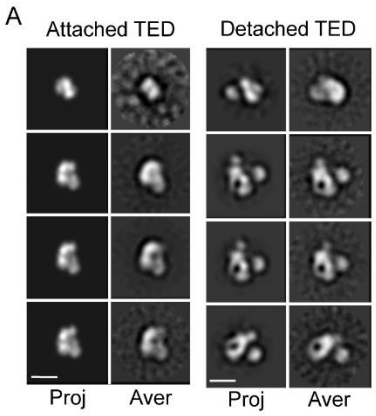



**Figure 3.3:** Prediction of mutations that affect the interaction of the TED domain and the MG ring. (A) At 150 mM NaCl, the transition electrostatic energy is more favorable for the F variant than the S variant of C3b as indicated by the dashed red line (calculated with **Equation 3.5**). Several mutations are predicted to cause deviations from this baseline expectation outside of thermal effects (2.5 kJ/mol). E95A and K104A are expected to make TED detachment more favorable in the F variant than in the S variant, while E1032A is expected to make the detachment process more favorable in the S variant than in the F variant. Effects of mutating R102 must be interpreted in a special manner, as this residue is a glycine in the F variant. After performing the R102A mutation, the transition electrostatic energy is expected to be similarly favorable in both S and F variants of C3b as the number and type of ionizable amino acids is identical after the mutation. (B) Minimum distance between charged side-chain atoms in key residues E95, R102, K104, and E1032, depicted in panel B. Higher distances (darker colors) indicate weaker Coulombic interactions. Interactions between oppositely charged ionizable amino acids that are separated by 8 Å or less are important at physiological ionic strength of 150 mM, but Coulombic interactions may occur at distances beyond 8 Å, at lower ionic strengths. (C) Crystal structure of C3b (PDB 2I07) in the attached conformation. Key residues that are expected to change the free energy of conformational transition differently between S and F C3b variants are shown as colored squares. As in panel A, these residues are predicted to cause deviation from the baseline difference in the electrostatic energy of conformational transition between S and F forms that is outside thermal effects (2.5 kJ/mol) when mutated to alanine.

net-favorable Coulombic interactions at the interface of the TED and MG ring. Additionally, the mutation is expected to alter the network of charge-charge interactions at this interface allowing for mutations that have different energetic effects depending on the C3b isoform of the parent structure (**Figure 3.3A**). To test the prediction that TED detachment is more energetically favorable for the F variant than the S variant, our collaborators analyzed the ratio of C3bS and F molecules in the attached and detached conformations when varying the concentration of NaCl (**Figure 3.4**).[33] C3b for each variant was obtained by standard purification methods from plasma of homozygous donors. Then, single molecules were observed and analyzed in the electron microscope as described previously, and the percentage of molecules in each conformation estimated after classifying the molecules by image processing. We performed three independent experiments and 22,000 molecule images were analyzed per each experiment. We followed the methodology described previously and compared each average image to computational projections of attached and detached models (**Figure 3.4A**). We did not find differences at 50, 75 and 125 mM NaCl, with around 20% of the molecules in the detached conformation for both isoforms (**Figure 3.4B**). When we reached 150 mM NaCl, the amount of molecules in the detached conformation increased to almost 50% for C3bF whereas C3bS stayed at 20%, and this difference was statistically significant ($P = 0.0023$). It cannot be ruled out that differences also occur at lower NaCl concentrations that could not be detected with our methodology. Overall, these results indicated that the F variant of C3b is more prone to a detached conformation than the S variant, at least at higher concentrations of ionic strength. This observation agrees with the computational prediction (at 150 mM) that




B

| Sample | [NaCl] (mM) | Attached (aver %) | Detached (aver %) | *s.d.* | |
|---|---|---|---|---|---|
| C3b F | 125 | 79 | 21 | 10.97 | |
| | 150 | 51 | 49 | 8.19 | |
| C3b S | 125 | 80 | 20 | 7.94 | P=0.0023 |
| | 150 | 81 | 19 | 7.86 | |

**Figure 3.4:** Ionic strength affects the conformation of C3b S and F variants differently. At least three independent experiments were performed to estimate the number of molecules for each variant in each condition (125 and 150 mM NaCl) using electron microscopy. The average number of molecule images per single experiment was around 22000, and the averages from 3 identical experiments are indicated.

the detachment process would be more energetically favorable for the F form than the S form.

### 3.4 Discussion

Current evidence shows that C3b coexists as a mixture of conformations.[9,10,12] Much of this intrinsic flexibility seems to affect the contacts of the TED domain to other parts of the molecule. Here we show that ionic tethering stabilizes the conformation of C3bS where the TED domain is in contact with the MG ring. We also show that the equilibrium distribution of attached and detached conformations can be affected by changes in ionic strength and the R102G polymorphism. In addition, we describe that the detached conformation is compatible with binding to FB and thus it is conceivable that the flexibility of the TED domain in the C3bBb convertase may affect regulation by FH under certain conditions. Our results are fully consistent with atomistic scattering models for C3b at different NaCl concentrations obtained using analytical ultracentrifugation and x-ray and

neutron scattering.[12] These models show that although the TED and the MG1 domains are in contact at 50 mm NaCl, the domains are separated at the physiological 137 mm NaCl concentration. Similar results had been recently described for the related complement protein C4b.[14]

The presence of the staphylococcal inhibitor Efb-C favors the detached conformation of C3b, and this conformation has a substantial impact on the rate at which FH binds C3b.[13] The interaction between Efb-C and C3d has been previously shown to be electrostatic in nature[34], and therefore Efb-C could disrupt the ionic tethering of TED on the MG ring. Other staphylococcal complement inhibitors, such as Ehp and Sbi-IV, have also been previously shown to electrostatically interact with C3d[25,34] and may have similar effects in disrupting the ionic tethering of TED.

In addition, we provide a proof of concept that the F variant of C3b is less stable in the attached conformation than the S variant, and this observation could rationalize the well-characterized functional differences between both forms of C3b.[32] To our knowledge, this is the first time that conformational differences have been observed in these two variants using purified protein. Our finding correlates well with the importance of the R102-E1032 salt bridge to maintain the interaction of C3d to immobilized C3c in surface plasmon resonance experiments.[12] In those experiments, the authors did not use the C3bS/F variants, whose mutation is in the C3d fragment, but tested the relevance of the salt bridge by mutating residue E1032. The AMD risk variant C3bF shows altered functionality, which is not due to differences in C3-convertase formation but rather to the susceptibility to regulation.[32] C3bF bound FH less efficiently compared with the S variant, causing

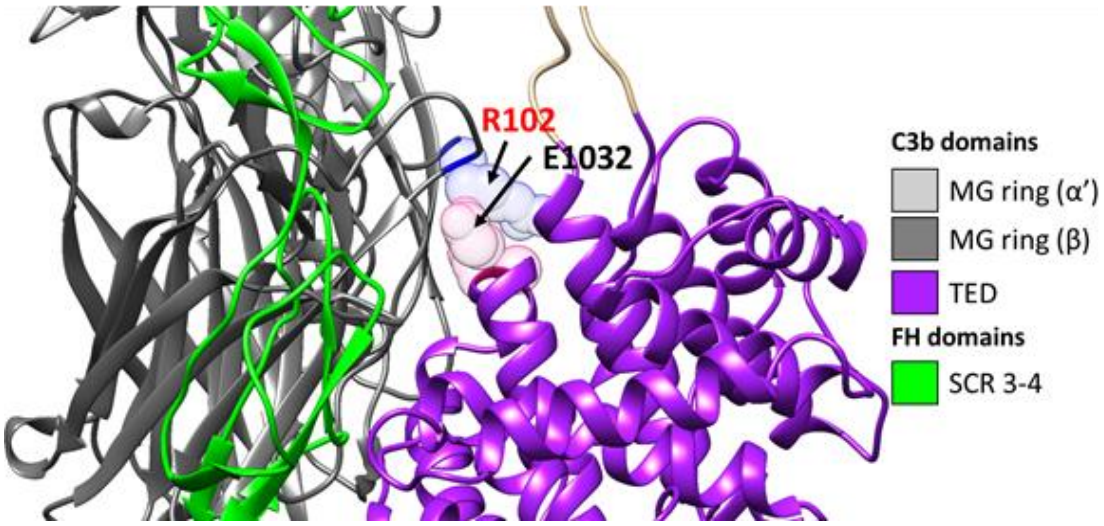



**Figure 3.5:** Ionic tethering between R102 and E1030 contributes to stability of C3b in the TED-attached conformation. This stability likely promotes robust interactions between FH SCR3-4 and C3b. Given the binding site of FH SCR3-4, detachment of the TED may perturb binding with FH. Domains are colored according to the inset key, and primary residues involved in the ionic tether are labeled.

decreased FI cofactor activity, extended convertase lifetime, and enhanced complement alternative pathway amplification. Our model suggests that subtle differences in conformational stability of each C3b variant could result in differences in FH binding (**Figure 3.5**). Finally, other mutations and polymorphisms could affect the conformational dynamics in C3b. In fact, we have recently described several mutations in patients that affect the conformation of C3b.[31] Some of the mutations predicted to affect the ionic tethering are located in the acidic concave surface on the TED and have been shown to contribute in accelerating the association of the TED (C3d) with its receptor, complement receptor2 (CR2), through long-range electrostatic interactions.[29] It is likely that this acidic concave surface is multifunctional, as it contributes to ionic tethering within C3b and also contributes to interactions with CR2[28] and FH.[26] Together, our findings extend previous information on the conformational flexibility of C3b in solution, highlighting the importance of dynamic structural changes of C3b in complement regulation.

## 3.5 References


1. Rodríguez de Córdoba, S. R., Harris, C. L., Morgan, B. P. & Llorca, O. Lessons from functional and structural analyses of disease-associated genetic variants in the complement alternative pathway. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1812,** 12–22 (2011).

2. Lea, S. M. & Johnson, S. Putting the structure into complement. *Immunobiology* **217,** 1117–1121 (2012).

3. Janssen, B. J. C., Christodoulidou, A., McCarthy, A., Lambris, J. D. & Gros, P. Structure of C3b reveals conformational changes that underlie complement activity. *Nature* **444,** 213–216 (2006).

4. Wiesmann, C. *et al.* Structure of C3b in complex with CRIg gives insights into regulation of complement activation. *Nature* **444,** 217–220 (2006).

5. Gros, P., Milder, F. J. & Janssen, B. J. C. Complement driven by conformational changes. *Nat. Rev. Immunol.* **8,** 48–58 (2008).

6. Forneris, F. *et al.* Regulators of complement activity mediate inhibitory mechanisms through a common C3b-binding mode. *EMBO J.* 1–17 (2016). doi:10.15252/embj.201593673

7. Wu, J. *et al.* Structure of complement fragment C3b-factor H and implications for host protection by complement regulators. *Nat. Immunol.* **10,** 728–733 (2009).

8. Schramm, E. C. *et al.* Mapping interactions between complement C3 and regulators using mutations in atypical hemolytic uremic syndrome. *Blood* **125,** 2359–2369 (2015).

9. Nishida, N., Walz, T. & Springer, T. A. Structural transitions of complement component C3 and its activation products. *Proc. Natl. Acad. Sci.* **103,** 19737–19742 (2006).

10. Alcorlo, M. *et al.* Structural insights on complement activation. *FEBS J.* **282,** 3883–3891 (2015).

11. Pechtl, I. C., Neely, R. K., Dryden, D. T. F., Jones, A. C. & Barlow, P. N. Use of time-resolved FRET to validate crystal structure of complement regulatory complex between C3b and factor H (N terminus). *Protein Sci.* **20,** 2102–2112 (2011).

12. Rodriguez, E., Nan, R., Li, K., Gor, J. & Perkins, S. J. A revised mechanism for the activation of complement C3 to C3b: A molecular explanation of a disease-associated polymorphism. *J. Biol. Chem.* **290,** 2334–2350 (2015).

13. Chen, H. *et al.* Allosteric inhibition of complement function by a staphylococcal immune evasion protein. *Proc. Natl. Acad. Sci.* **107,** 17621–17626 (2010).

14. Fung, K. W., Wright, D. W., Gor, J., Swann, M. J. & Perkins, S. J. Domain structure of human complement C4b extends with increasing NaCl concentration: implications for its regulatory mechanism. *Biochem. J.* **473,** 4473–4491 (2016).

15. Fiser, A., Kinh Gian Do, R. & Sali. Modeling Loops in Protein Structures. *Protein Sci.* **9,** 1753–1773 (2000).

16. Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).

17. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2,** 19–25 (2015).

18. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).

19. Shapovalov, M. V. & Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19,** 844–858 (2011).

20. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

21. Dolinsky, T. J. *et al.* PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35,** 522–525 (2007).

22. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32,** 665–667 (2004).

23. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10037–41 (2001).

24. Gorham, R. D., Kieslich, C. A. & Morikis, D. Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Ann. Biomed. Eng.* **39,** 1252–1263 (2011).

25. Gorham, R. D., Rodriguez, W. & Morikis, D. Molecular analysis of the interaction between staphylococcal virulence factor Sbi-IV and complement C3d. *Biophys. J.*

**106,** 1164–1173 (2014).

26. Harrison, R. E. S., Gorham, R. D. & Morikis, D. Energetic evaluation of binding modes in the C3d and Factor H (CCP 19-20) complex. *Protein Sci.* **24,** 789–802 (2015).

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

28. Mohan, R. R., Gorham Jr, R. D. & Morikis, D. A theoretical view of the C3d:CR2 binding controversy. *Mol. Immunol.* **64,** 112–122 (2015).

29. Mohan, R. R., Huber, G. A. & Morikis, D. Electrostatic Steering Accelerates C3d:CR2 Association. *J. Phys. Chem. B* **120,** 8416–8423 (2016).

30. Gorham, R. D. *et al.* An evaluation of Poisson-Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. *Biopolymers* **95,** 746–754 (2011).

31. Martínez-Barricarte, R. *et al.* The molecular and structural bases for the association of complement C3 mutations with atypical hemolytic uremic syndrome. *Mol. Immunol.* **66,** 263–273 (2015).

32. Heurich, M. *et al.* Common polymorphisms in C3, factor B, and factor H collaborate to determine systemic complement activity and disease risk. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 8761–6 (2011).

33. López-Perrote, A. *et al.* Ionic tethering contributes to the conformational stability and function of complement C3b. *Mol. Immunol.* **85,** 137–147 (2017).

34. Gorham, R. D., Kieslich, C. A. & Morikis, D. Complement Inhibition by Staphylococcus aureus: Electrostatics of C3d-EfbC and C3d-Ehp association. *Cell. Mol. Bioeng.* **5,** 32–43 (2012).

CHAPTER 4

# Large-scale, Functional Conformational Transitions in Complement Component C3

## 4.1 Introduction

As a part of innate immunity, serum and cell-bound proteins within the complement system discriminate between surfaces of self and non-self and generate an immune response towards non-self. Complement activation increases inflammation, enhances phagocytosis, lyses cells, and participates in hemostasis.[1,2] These responses occur over time scales from seconds to hours,[3] and are critical for responding to invading pathogens that penetrate the skin and mucosal tissues. In fact, many pathogens have developed strategies to modulate complement response and avoid death.[4]

The complement system involves feed-forward amplification to mount a significant immune response in a short period of time.[1,2] C3 convertase is an enzyme that participates in feed-forward amplification to generate an immune response. Within this amplification loop, C3 convertase cleaves serum complement component 3 (C3) into C3a, an anaphylatoxin, and C3b, an opsonin. C3b covalently bonds to nucleophiles on the surfaces of self and non-self via the thioester domain (TED). Bound C3b then associates with complement factor B circulating in serum. This complex is subsequently cleaved by complement factor D, also circulating in serum, to generate additional C3 convertase. The newly generated C3 convertase participates in the amplification loop, cleaving additional C3. As a result of this amplification loop, other complement species are generated that serve a biological purpose. For instance, C3a modulates inflammation, and C3b enhances

phagocytosis.[1,2] The positive amplification loop is critical for mounting an immune response over a short period of time; however, this feature imparts instability in biological systems. Left unchecked, this feed-forward amplification will indiscriminately attack foreign and native surfaces, destroying healthy cells along with foreign cells. For this reason, a number of native regulators exist to inhibit complement activation on surfaces of self.

A number of mutations and polymorphisms in complement proteins lead to complement dysregulation, complement deficiencies, or are risk factors for acquiring disease later in life. For C3 in particular, mutations may result in rare kidney diseases such as atypical hemolytic-uremic syndrome and C3 glomerulopathy, C3 deficiencies characterized by frequent bacterial infections, or increased risk for developing age-related macular degeneration.[5]

Despite the central role of C3 in the complement system, no FDA-approved therapeutics target C3. In fact, only two complement targets have FDA-approved therapeutics – C1 and C5.[5] C1 is a serine protease that initiates an immune response in the classical pathway of complement alternative, prior to involvement of C3.[5] C1 is inhibited by C1 inhibitor (C1INH), a natural protein important in regulation of complement. After this importance was recognized, C1INH was developed as one of the first complement therapeutics.[5] C5, the second protein target, is involved in the terminal pathway of complement and is inhibited by a monoclonal antibody named eculizumab.[5] Although inhibition of C5 regulates the formation of the lytic membrane attack complex and C5a-mediated inflammatory response, inflammation and opsonophagocytosis can still occur

through C3a and C3b, respectively. Furthermore, eculizumab is currently the most expensive drug on the market, costing approximately $442,000 per patient per year in the United Kingdom during the year 2015.[5] Eculizumab was developed to treat another rare disease, paroxysmal nocturnal haemoglobinuria; however, the therapeutic has also been approved to treat individuals with atypical hemolytic-uremic syndrome.[5] Given the lack of specific complement therapeutics and high cost of those that exist, there is a demand for less expensive therapeutics such as peptide and small molecule inhibitors.

After C3 is cleaved by C3 convertase to form C3b, the TED is translated a distance of 64.0 Å and rotated to expose the internal thioester.[6–8] This motion is accompanied by additional domain rearrangements that can be viewed in the crystal structures of C3 and C3b. Without this rearrangement, formation of the C3 convertase would not be possible.[8] Thus, the C3a domain of C3 stabilizes the overall conformation, inhibiting C3 activation. This stabilization is further supported by a coarse grain elastic network model (ENM) that predicts increases in square fluctuations in residues proximal to C3a and residues in the TED after cleavage of C3a (**Figure 4.1**). By understanding the structural dynamics of C3 and subsequent cleavage products, a more detailed understanding of C3b activation can be achieved. Moreover, it may be possible to design therapeutics to prevent activation of C3b, inhibiting complement response in certain cases of autoimmunity.

As C3 and C3b have a central role in propagating an innate immune response, we seek to understand the molecular mechanisms that govern large-scale conformational rearrangements that occur after C3a is cleaved from C3 at an atomic level. Since simulating such large-scale motions with molecular dynamics would normally be computational

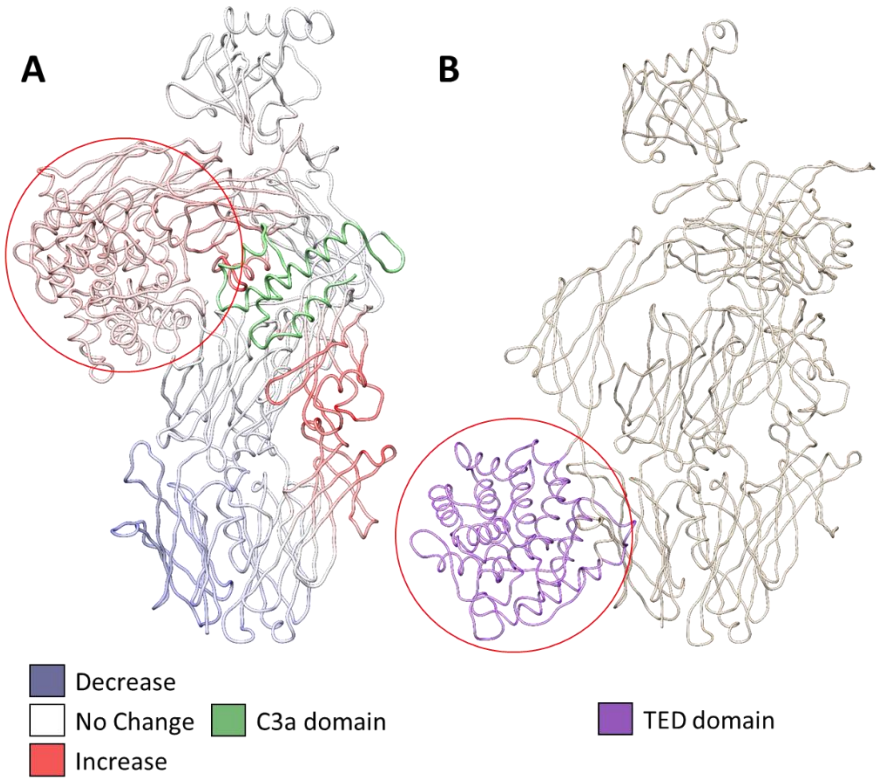



**Figure 4.1:** (A) Backbone structure of C3 shown with a licorice representation. The TED is circled in red and the C3a domain is colored light green. Otherwise, the backbone structure of C3 is colored blue, white, or red depending if the square fluctuations of the residue are expected to decrease, not change, or increase, respectively, after the cleavage of the C3a domain. (B) Backbone structure of C3b shown with a licorice representation. Note how the TED (circled in red and colored purple) has been rearranged with respect to other C3b domains.

intractable, we approach this problem by guiding molecular dynamics along an estimated conformational transition pathway. This pathway is modeled at a coarse level of granularity with a method based on ENMs. Specifically, we leverage crystal structures of the C3-like conformation of C3b and canonical C3b conformation to construct an interpolated ENM with two energy minima corresponding to the two reference structures at the ends of the pathway using the method of Tekpinar and Zheng.[9] From this model, we extract multiple conformers of C3b predicted to occur during C3 activation and measure distances between all pairwise combinations of domains within C3b. These distances are used to parameterize

harmonic restraints between all pairwise domains, and spring equilibria are changed according to a time-schedule during molecular dynamics simulation to accelerate domain rearrangements in C3b. Given the large number of atoms and extensive sampling required to model this system, we have leveraged simulation time on Anton2, a supercomputer built specifically for molecular dynamics, to make the project computationally tractable.

Here, we focus on dynamics of C3b during transition from a C3-like conformation to a C3b-like conformation as activation of C3b facilitates immune surveillance by exposing the internal thioester and facilitating opsonization of cell surfaces. We apply a hybrid method that uses a coarse grain estimate of the conformational changes in C3b during activation in order to accelerate progress along the activation reaction coordinate in atomistic molecular dynamics simulations. As a result, we identify intramolecular, electrostatic interactions that may facilitate a TED-attached conformation (as discussed in chapter 3) following C3 activation.

## 4.2 Methods

### *4.2.1 Preparation of C3b End States*

Structures for C3 and C3b were retrieved from the Protein Data Bank with identifiers 2a73[10] and 5fo7[7], respectively. Modeller[11,12] was used to model and refine loops in non-terminal regions, and UCSF Chimera[13] was used to mutate residues that deviate from the canonical C3 sequence using the `swapaa` command. Structures were trimmed to contain only residues from C3b leaving intact residues 23-665 from the β chain and residues 753-

1663 from the α' chain. Protonation states of histidines were assigned using the PROPKA method in PDB2PQR[14,15] at a pH of 7.4.

*4.2.2 Construction of an Interpolated Elastic Network Model for C3b Activation*

Using the prepared structures, the ENM-PATH method of Zheng et al was applied to our specific pathway in both the forward and reverse directions.[16] Specifically, the method was run once starting from the C3-like conformation of C3b and progressing towards the canonical conformation of C3b (forward), and then the method was run a second time starting at the canonical conformation of C3b and progressing towards the C3-like conformation (reverse). In each run, default parameters were implemented. Forward and reverse pathways were merged together at the conformer indices with lowest root-mean-square distance (RMSD) between $C_\alpha$ coordinates in order to model the transition from a C3-like conformation to the canonical C3b conformation. This method returned a total of 51 conformations between C3-like and C3b-like endpoints. Five representative frames from 51 were sampled to capture changes in C3d orientation (frames 6, 36, 41, 50, and 51) according to RMSD from the initial conformation. In each conformer, all pairwise distances between centers of masses of $C_\alpha$ in each C3b domain (MG1, MG2, MG3, MG4, MG5, MG6, MG7, MG8, C345C, CUB, and TED) were measured to determine equilibrium distances at each of the five pathway conformers. To ensure sufficient time was allowed for conformational changes, we simulated with each set of restraints from the conformational pathway to allow an RMSD change of no less than 3 Å/μs between consecutive conformers of the pathway. These simulation times were rounded up to the first multiple of simulation step size (200 ps). Spring constants for each domain pair were

set such that the maximum peak force associated with changing harmonic equilibria along the conformational pathway would produce a peak force of 50 pN, a force on the order of those typically used in atomic-force microscopy experiments.[17] This peak force was chosen in order to promote domain rearrangements without denaturation of individual domains.

*4.2.3 Molecular Dynamics of C3b Activation*

To prepare for simulations of C3b during activation, we first solvated, minimized, heated, and equilibrated a simulation system starting in a C3-like conformation (2a73). Each of these steps was carried out in NAMD[18] with the CHARMM36 forcefield.[19] C3b was solvated in orthorhombic, water box (tip3p model) with periodic boundary conditions and 150 mM NaCl. The size of the box was determined by using 15 Å padding in the positive and negative directions for each Cartesian basis beyond the minimum rectangle required to encapsulate C3b (rectangle edges were parallel to basis vectors). The resulting system was subject to minimization for 100 steps at a rate of 1fs/step before heating from 0 to 310 K in increments of 10 K for 1000 steps at 2 fs/step. During heating, protein was constrained with a spring constant of 1 kcal/mol/Å$^2$. Following heating, C3b was subject to a series of five equilibration simulations, each run for 50,000 steps at 2 fs/step. In each successive equilibration run, constraints were incrementally relaxed. In equilibration runs 1-4, constraints were applied with spring constants of 1, 0.5, 0.2, and 0.1 kcal/mol/Å$^2$ to all protein atoms. In the final equilibration run, only backbone atoms were constrained with a spring constant of 0.1 kcal/mol/Å$^2$. Following equilibration, a production run of 100 ns was performed.

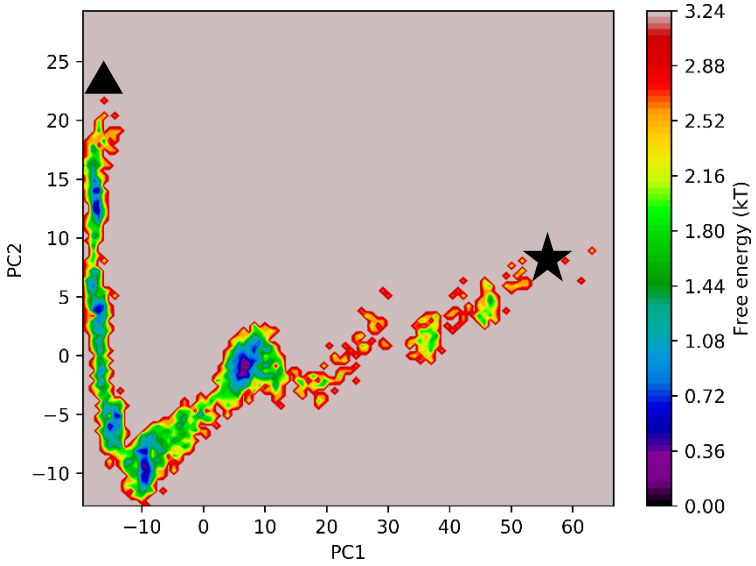



**Figure 4.2:** Free energy landscape for the activation of C3b as it proceeds from a C3-like conformation (triangle) to the canonical C3b conformation (star). This pathway is amenable to simulations that branch off at regular intervals without harmonic restraints to observe how the structure relaxes and to identify new conformational states of C3b.

Before running simulations on the Anton2[20] supercomputer, we re-solvated the final conformation from the 100 ns production run in a larger, cubic periodic box with an edge length of 180 Å. This process is necessary since simulations on Anton2 exhibit more diffusive motion than other conventional molecular dynamics software. In this process, we followed the same protocol as in simulation preparation except that the final production run was only 10 ns in total length due to increased computational complexity arising from the larger periodic cell.

Next, we used the coordinates, velocities, and periodic box from the end of the preceding trajectory to initialize an Anton2-compatible simulation system with CHARMM36 forcefield parameters. The simulation was carried out 310 K and 1 bar using the Nose-Hoover barostat and thermostat. The simulation was run for 8.7 μs with the

previously described restraint schedule where the simulation trajectory exhibited minimum RMSD from the canonical C3b conformation.

### 4.3 Results and Discussion

Simulations of C3b activation process were guided by an elastic network model, allowing preliminary sampling of the conformational pathway as shown in the free energy landscape of **Figure 4.2**. This landscape was constructed by superposing the first 8.6 μs of the ENM-guided molecular dynamics trajectory, performing principle component analysis on the Cartesian coordinates for all atoms in the time-series, and construction of a two dimensional histogram using a method from PyEMMA[21] where probabilities of conformations are expressed as free energies using a Boltzmann distribution. The hybrid method used to model conformational changes in C3b demonstrates the utility of coarse-grain models in guiding simulations of finer granularity. With additional simulation time, trajectories can branch of the currently described pathway to observe dynamics without any conformational bias imposed by the set of harmonic restraints. It should be noted, however, that restraints only apply forces based on the center of mass for a domain. Thus, simulation restraints primarily impose a translational bias that pulls apart or together domain pairs.

During the simulation of the activation process, we observed a number of high occupancy hydrogen-bond interactions that change according the restraint equilibria. In **Figure 4.3** we display the probabilities of observing hydrogen bonds that satisfy Baker-Hubbard criteria[22] over four different time windows of the activation pathway. Notably,

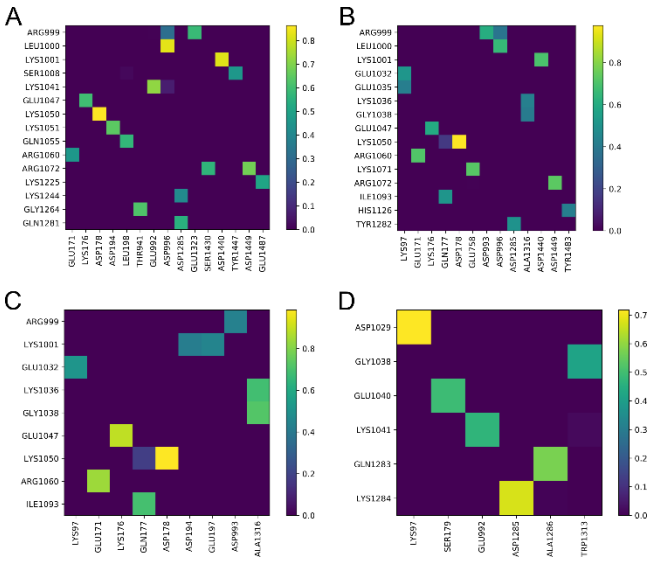



**Figure 4.3:** Hydrogen bond occupancies between restraint equilibria according to Baker-Hubbard criteria. Residues of C3d (TED) are located along the y-axis while residues of all other C3b domains are located along the x-axis. Heatmaps in panels A-D display probability of observing a hydrogen bond – the occupancy of the hydrogen bond interaction between a residue pair. Panels A, B, C, and D correspond to time windows from molecular dynamics simulations of 0-2.3 μs, 2.3-4.8 μs, 4.8-7.2 μs, and 7.2-8.7 μs, respectively.

many of these hydrogen bonds are also salt-bridges. As activation occurs, the number of high occupancy hydrogen bonds between C3d (TED) and other domains of C3b decreases until panel D where new, characteristically different hydrogen bond pairs are formed. Thus, as the TED pulls away from the MG ring of C3b during the activation process, stabilizing interactions appear to be lost. Electrostatic interactions may be essential to guide the TED into the canonical C3b conformation (TED-attached) following this detachment process.

## 4.4 Conclusion

Long timescale simulations of C3b guided by coarse-grain models of protein dynamics have led to an improved model for the activation of C3. This model currently suffers from bias imposed by pairwise, harmonic, domain restraints; however, this can be ameliorated

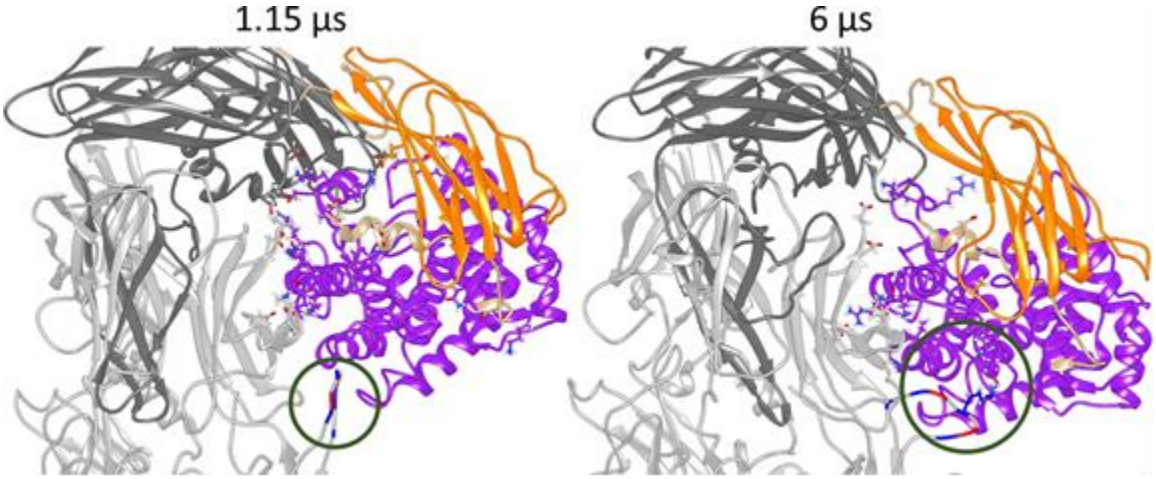



**Figure 4.4:** As C3b transitions from a C3-like conformation to a canonical C3b conformation, hydrogen bond interactions begin to form between the charged loop of the MG ring containing R102. Shown here are two conformations labeled according to the time index of the trajectory where they are observed. The structure is colored light grey, dark grey, orange, and purple to denote locations of the α' chain of the MG ring, the β chain of the MG ring, the CUB, and the TED, respectively. Residues on the charged loop containing R102 are surrounded by a green circle and colored red and blue to denote amino acids that are negatively or positively charged, respectively. Note that at later time indices, the TED begins to form hydrogen bonds with K97.

in the future by additional sampling of conformational dynamics in our initial trajectory without restraints implemented.

In our current model, we observe a number of salt-bridges between amino acids that stabilize the position of the TED (**Figure 4.4**). Though these specific residues do not appear to correspond to disease, it does suggest our previous studies of C3d detachment from the MG ring (chapter 3) may also apply to C3 activation.[23] Though a detached conformation of C3d is compatible with binding to Factor B, interactions with regulators may be perturbed. Thus, if electrostatic interactions are perturbed, the simplest strategy to combat states of autoimmunity may be to stabilize the C3-like conformation of C3d. Alternatively, if one wished to deregulate complement activation, it may be possible to disrupt electrostatic-mediated rearrangement of the TED.

## 4.5 References


1. Ricklin, D., Hajishengallis, G., Yang, K. & Lambris, J. D. Complement: a key system for immune surveillance and homeostasis. *Nat. Immunol.* **11,** 785–797 (2010).

2. Mathern, D. R. & Heeger, P. S. Molecules great and small: The complement system. *Clin. J. Am. Soc. Nephrol.* **10,** 1636–1650 (2015).

3. Zewde, N., Gorham, R. D., Dorado, A. & Morikis, D. Quantitative Modeling of the Alternative Pathway of the Complement System. *PLoS One* **11,** e0152337 (2016).

4. Ferreira, V. P., Pangburn, M. K. & Cortés, C. Complement control protein factor H: The good, the bad, and the inadequate. *Mol. Immunol.* **47,** 2187–2197 (2010).

5. Morgan, B. P. & Harris, C. L. Complement, a target for therapy in inflammatory and degenerative diseases. *Nat. Rev. Drug Discov.* **14,** 857–877 (2015).

6. Janssen, B. J. C., Christodoulidou, A., McCarthy, A., Lambris, J. D. & Gros, P. Structure of C3b reveals conformational changes that underlie complement activity. *Nature* **444,** 213–216 (2006).

7. Forneris, F. *et al.* Regulators of complement activity mediate inhibitory mechanisms through a common C3b-binding mode. *EMBO J.* 1–17 (2016). doi:10.15252/embj.201593673

8. Gros, P., Milder, F. J. & Janssen, B. J. C. Complement driven by conformational changes. *Nat. Rev. Immunol.* **8,** 48–58 (2008).

9. Tekpinar, M. & Zheng, W. Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model. *Proteins Struct. Funct. Bioinforma.* **78,** 2469–2481 (2010).

10. Janssen, B. J. C. *et al.* Structures of complement component C3 provide insights into the function and evolution of immunity. *Nature* **437,** 505–511 (2005).

11. Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).

12. Fiser, A., Kinh Gian Do, R. & Sali. Modeling Loops in Protein Structures. *Protein Sci.* **9,** 1753–1773 (2000).

13. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

14. Dolinsky, T. J. *et al.* PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35,** 522–525 (2007).

15. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32,** 665–667 (2004).

16. Zheng, W., Brooks, B. R. & Hummer, G. Protein conformational transitions explored by mixed elastic network models. *Proteins* **69,** 43–57 (2007).

17. Carrion-Vazquez, M. *et al.* Mechanical and chemical unfolding of a single protein: A comparison. *Proc. Natl. Acad. Sci.* **96,** 3694–3699 (1999).

18. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).

19. Best, R. B. *et al.* Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J. Chem. Theory Comput.* **8,** 3257–3273 (2012).

20. Shaw, D. E. *et al.* Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Commun. ACM* **51,** 91–97 (2008).

21. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11,** 5525–5542 (2015).

22. Baker, E. N. & Hubbard, R. E. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44,** 97–179 (1984).

23. López-Perrote, A. *et al.* Ionic tethering contributes to the conformational stability and function of complement C3b. *Mol. Immunol.* **85,** 137–147 (2017).

CHAPTER 5

# Multivalent Recognition of C3d by Complement Factor H (SCR 19-20)

## 5.1 Introduction

In the alternative pathway (AP) of the complement system, an immune response is facilitated through conversion of C3 to form C3b. Also referred to as tickover, this process consumes approximately one percent of C3 molecules per hour.[1] C3b can further interact with factor B (FB) to form C3 convertase. As a result, positive feedback occurs, and C3 convertase cleaves C3 to C3a, an anaphylatoxin, and C3b, an opsonin. Since C3b and C3 convertase are capable of binding to host and foreign surfaces via a thioester domain located within the C3d domain, the positive amplification loop occurs at these surfaces, generating an immune response.[1] While tickover is essential for immune response via the AP, overactivation and/or under-regulation can cause severe damage to native tissues due to the nonspecific nature of the interaction. In order to prevent a state of autoimmunity, native tissues express factor H (FH). This biomolecule competes with FB limiting the formation of C3 convertase, acts as a cofactor for the degradation of C3b to iC3b, and accelerates decay of C3 convertase.[2]

FH is a 155 kDa protein composed of 20 complement control protein (CCP) modules.[2] In order to explain negative regulation of AP amplification on host cells, several models that propose a mechanism to discriminate between self and non-self have been hypothesized that rely on functional CCP modules.[2–5] In short, C3b binding is believed to occur at CCP 1-4 and 19-20 while polyanion binding is thought to occur at CCP 1, 5, 7,

13, and 20.[4,6] Mutations in these modules have been linked to autoimmune disease. Of note for the present study, mutations in FH 19-20 have been linked to atypical hemolytic uremic syndrome (aHUS).[2,3,6] Moreover, mutating these aHUS-linked residues has been shown to affect C3d binding.[2,3] In order to understand pathogenesis of aHUS-linked mutations, studies have involved investigating relative binding affinity of FH mutants and characterizing the crystallographic structures of C3d in complex with FH CCP 19-20.[2,3] These structures suggest three different modes of binding as summarized in **Figure 5.1**. Protein Data Bank (PDB) crystallographic structures 3OXU[2] and 2XQW[3] contain multiple chain pairings of C3d with FH CCP 19-20 that are described as binding sites 1-3, herein. Site 1 involves a negatively charged surface of C3d, referred to as the acidic patch,[7] binding to positively charged FH CCP 20, while site 2 involves interaction between a ridge of C3d, not far from the acidic patch, and the negatively charged FH CCP 19. Located on the opposite face of C3d from sites 1-2, site 3 involves binding of the slightly positive, basic surface of C3d to the positively charged FH CCP 20.

In considering the physiological relevance of sites 1-3, a primary, C3b binding mode at site 2 has been suggested that does not involve the acidic patch or thioester bond of C3d.[3,5] To reach this conclusion, studies considered the relative accessibility of the C3d subdomain[3] within C3b and ensured that the proposed models enable discrimination between self and non-self.[2,3] Additionally, studies map functional domains of FH CCP 19-20 by experimentally characterizing binding of C3d with mutant FH.[2,3] Ultimately, these models emphasize the physiological importance of site 2; however, CCP 20 has a positive

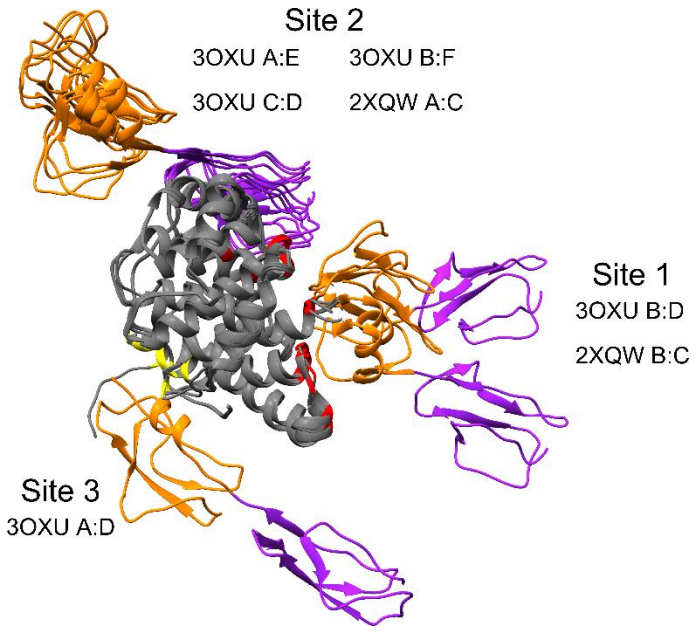



**Figure 5.1:** Molecular graphic representation of the competing binding modes from two crystallographic structures in the Protein Data Bank, 3OXU and 2XQW. Each binding site is described by one or more pairs of interacting chains (separated by a colon) from a specific crystallographic structure (PDB identifier of source noted). C3d is colored dim gray and specific residues associated with the acidic patch are colored in red while residues associated with the thioester domain are colored in yellow. FH is represented by CCP modules 19 and 20. FH CCP 20 is colored orange while CCP 19 is colored purple. At site 1, the binding configuration of C3d and FH varies between models (PDB 3OXU and 2XQW) suggesting apical flexibility.

charge that may interact with the negatively charged acidic patch of C3d at site 1.[6] One study does suggest a dual interaction of CCP 19 and 20 by 1 molecule of FH CCP 19-20 binding one C3b and either C3d or a host glycosaminoglycan (GAG)[3]; however, this conclusion is based on a mutated crystallographic structure to enable crystallization.[3] In light of recent debate regarding the physiological binding mode of CR2 to C3d,[8–13] we compare each binding site in terms of electrostatic character, structural stability, dissociative properties, and predicted free energy of binding to propose which binding mode or set of modes is likely to be physiologically relevant using computational methods.

## 5.2 Results

### *5.2.1 AESOP*

Using a computational framework developed in our lab and termed Analysis of Electrostatic Similarities of Proteins (AESOP), we investigate the electrostatic character of each binding site (**Figure 5.1**). Comprehensive, computational alanine mutagenesis of each ionizable amino acid, one at a time, in C3d revealed a high degree of electrostatic contributions in the binding modes of site 1 and site 2 as suggested by several mutations that cause loss or gain of binding outside the energetic fluctuations of thermal effects (up to 2.5 kJ/mol) (**Figure 5.2**). Additionally, mutations that affect binding at one site are generally observed to not affect other binding sites. This result is in agreement with each binding site being located at distinct surface areas of C3d. From AESOP analysis of FH CCP 19-20, sites 1 and 2 have the most electrostatic character, while site 3 generally lacks mutations with significant energetic perturbations. In general for sites 1 and 2, mutation of acidic residues in C3d and basic residues in FH CCP 19-20 result in loss of binding relative to wild type (WT), while mutation of basic residues in C3d and acidic residues in FH CCP 19-20 result in gain of binding. At site 3, there is notable deviation from the previous pattern of sites 1 and 2 where acidic residues of FH typically result in gain of binding mutations. Instead, mutation of acidic residues result in loss of binding and mutation of basic residues result in gain of binding which is in agreement with the basic surface of the thioester domain.[7]

To compare our comprehensive alanine scan to mutations mentioned in literature, we compiled a list of individual mutants from previous mutagenesis studies of C3d and FH

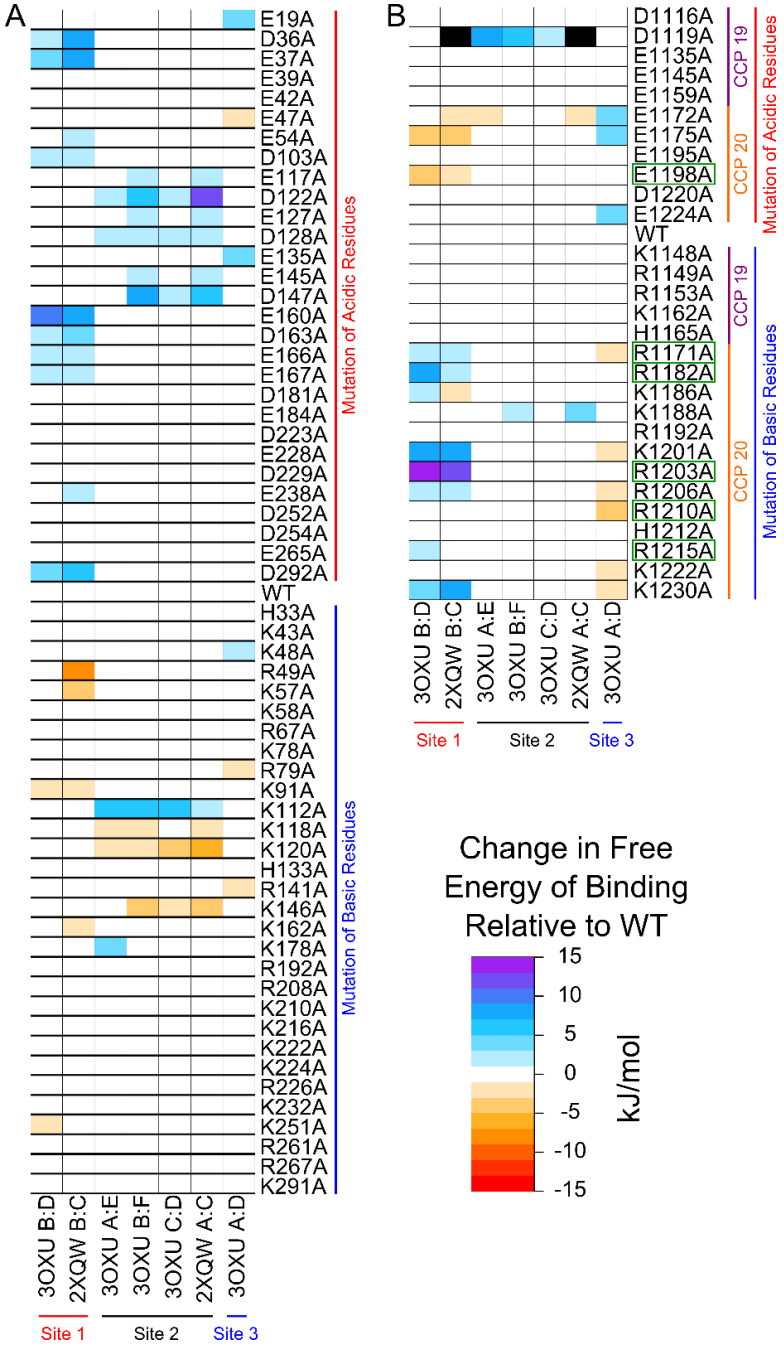



**Figure 5.2:** Change in free energy of binding compared to the wild type molecule resulting from computational alanine mutagenesis of (A) C3d and (B) FH. The heatmap scale is kept consistent between panels. Black for residue D1119 in panel B indicates no data since there is not an ionizable amino acid in this position within model 2XQW. Mutations boxed in green represent mutations that have been linked to aHUS. Negative values are gain of binding mutations while positive values are loss of binding mutations.

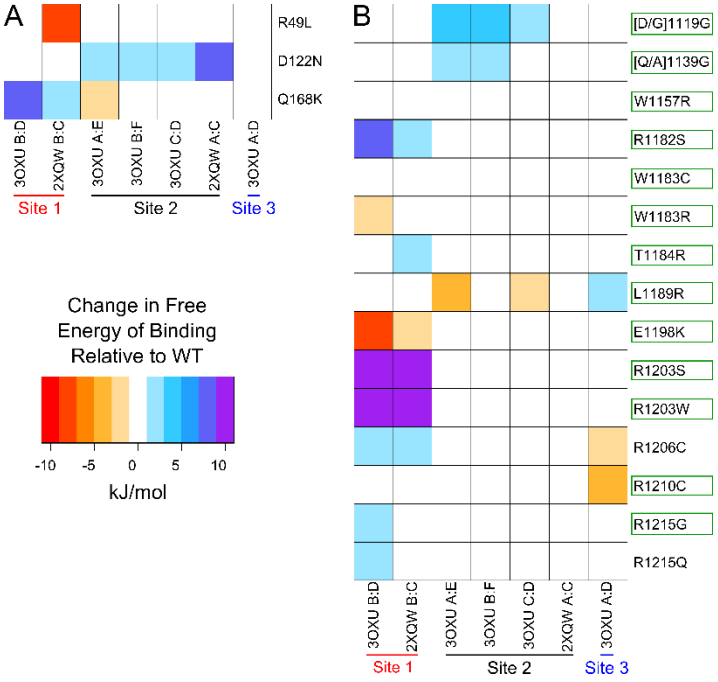



**Figure 5.3:** Change in free energy of binding compared to the wild type molecule resulting from computational site-directed mutagenesis for (A) C3d and (B) FH. The heatmap scale is kept consistent between panels. The first amino acid in brackets represents the residue in 3OXU, while the second is the residue in 2XQW. Mutations boxed in green represent mutations that have been linked to aHUS. Negative values are gain of binding mutations.

that involve residues with partial charges or ionizable amino acids to residues other than alanine.[2,3] Using the resulting list of mutations, we once again applied the method of AESOP. Mutations in C3d at residues D122 and Q168 have been shown to decrease binding. In our results (**Figure 5.3**), the D122N mutation corresponds to loss of binding at site 2, while Q168K corresponds to loss of binding at site 1. Shifting focus to FH, there are a number of mutations in literature that are linked to aHUS. Our results suggest (**Figure 5.3**) a number of aHUS linked mutations that correspond to loss of binding mutations at site 1 and 2. Despite general agreement, several mutations of aHUS-linked residues result in gain of binding mutations in our model at both sites 1 and 2.

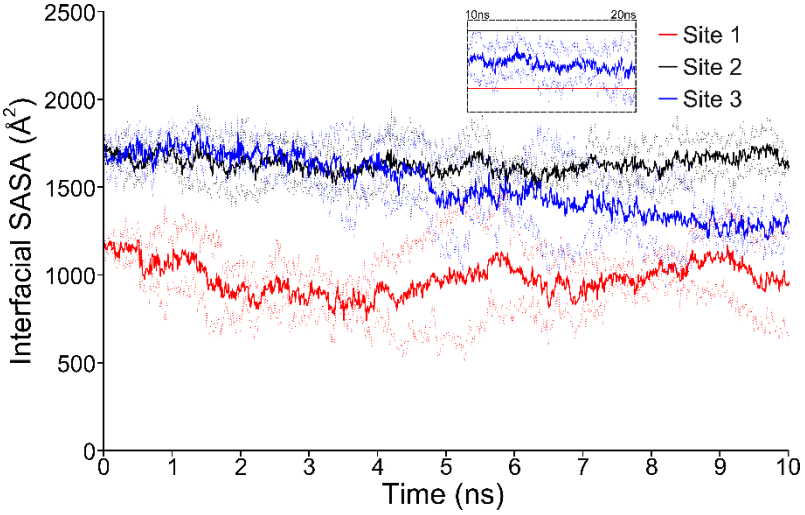



**Figure 5.4:** Interfacial SASA is plotted for each binding mode above and is colored according to binding site. Large, solid lines represent the mean values across a set of three MD trajectories, while smaller, dotted lines represent one standard deviation above or below the mean. Inset next to the key is the Interfacial SASA for site 3's extended MD trajectories from 10 to 20 ns (solid black and red lines here represent the extrapolated mean Interfacial SASA for each associated binding mode and are used to compare with the values for site 3).

### *5.2.2 Conventional Molecular Dynamics*

In order to characterize structural stability in terms of protein-protein interactions, we analyzed trajectories from molecular dynamics (MD) simulations. For each binding mode of C3d and FH CCP 19-20, root-mean-square deviation (RMSD) from the initial structure stabilizes relatively quickly (**Figure A.1-3**), though there is some moderate fluctuation about the mean value across the triplicate trajectories that may be attributed to the flexibility of the FH CCP 19 and 20 modules (**Figures A.2-3**). Solvent accessible surface area (SASA) lost upon binding (interfacial SASA) seems to be a better indicator of system stability for these systems and stabilizes within 10 ns for the binding modes at site 1 and site 2 (**Figure 5.4**). Site 3, on the other hand, experiences a steady decrease in interfacial SASA during the first 10 ns of the simulation, stabilizing only after extending the simulation to 20 ns. Fluctuations in the interfacial SASA at site 1 likely reflect the

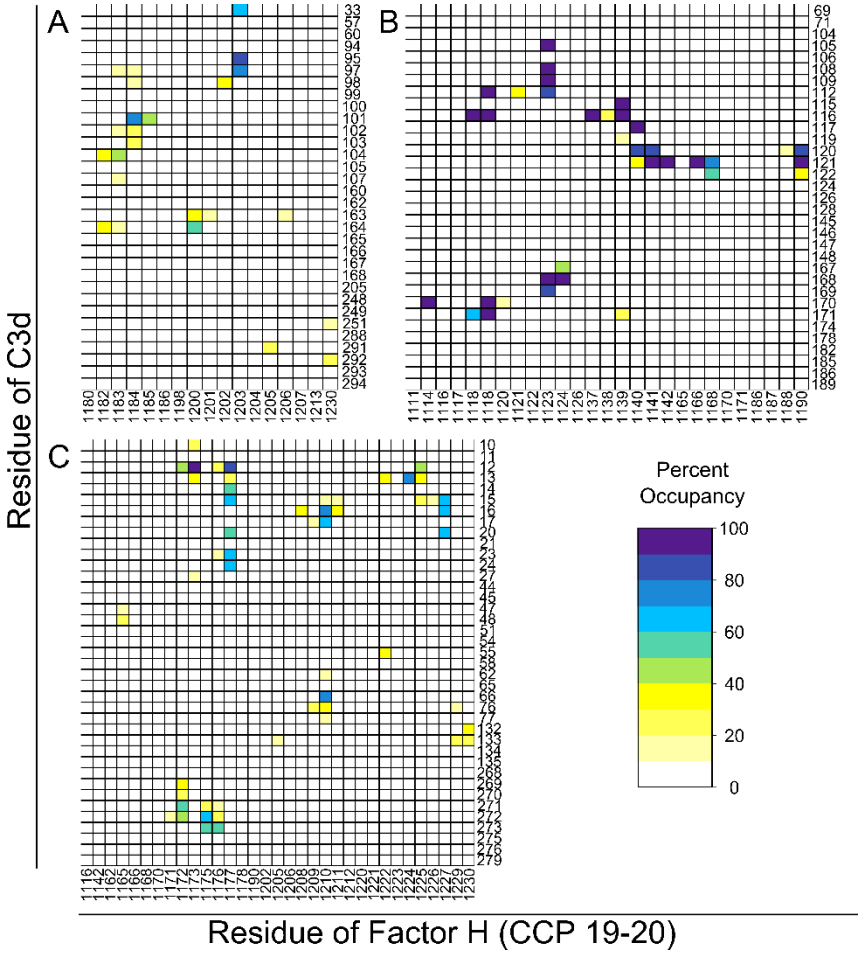



**Figure 5.5:** Percent occupancies of aliphatic interactions between C3d and FH at (A) binding site 1, (B) binding site 2, and (C) binding site 3. These values represent the percentage of snapshots throughout the triplicate trajectories for each binding site where a pair of functional groups with aliphatic features were observed to be within 4 Å in the C3d:FH complex.

flexibility of FH CCP 19-20 which is exacerbated by the end-to-end binding configuration that is represented in **Figure 5.1,** with the C-terminus of FH CCP 20 binding to residues of the acidic patch on C3d. This property is further illustrated by the conformational variation observed at binding site 1 between crystallographic structures (**Figure 5.1**).

Interaction profiles from all trajectories seem to vary drastically between binding modes. When considering potential aliphatic interactions with a cutoff distance of 4 Å (**Figure 5.5**), site 1 has the least amount of interactions, followed by site 3 with a higher

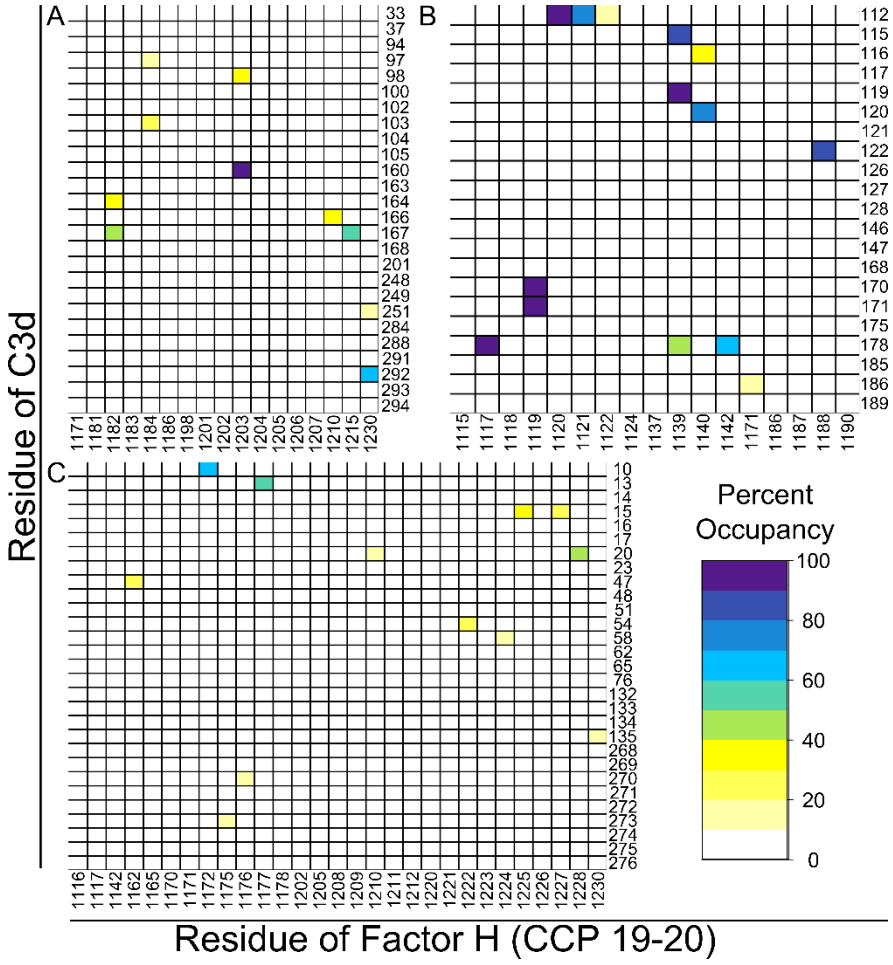



**Figure 5.6:** Percent occupancies of hydrogen bond interactions between C3d and FH at (A) binding site 1, (B) binding site 2, and (C) binding site 3. These values represent the percentage of snapshots throughout the triplicate trajectories for each binding site where a where a hydrogen bond is present in the C3d:FH complex.

number of interactions, and by site 2 with the most interactions. Also notable is the near time-invariant presence of many potential aliphatic interactions at site 2. In terms of hydrogen bonding (**Figure 5.6**), site 2 also exhibits the greatest amount of hydrogen bonds throughout all trajectories. Once again, many of these interactions are nearly time-invariant with occupancies approaching 100%. Site 1 exhibits one nearly time-invariant predicted hydrogen bond between FH residue R1203 and C3d residue E160 with three other high

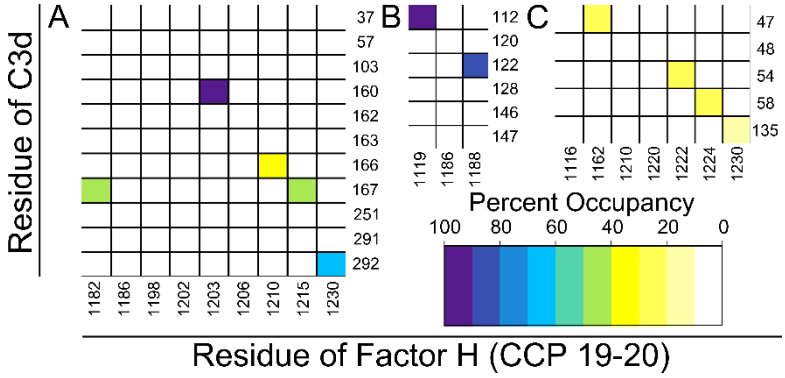



**Figure 5.7:** Percent occupancies of salt bridges between C3d and FH at (A) binding site 1, (B) binding site 2, and (C) binding site 3. These values represent the percentage of snapshots throughout the triplicate trajectories for each binding site where a pair of chemical groups capable of forming a salt bridge were observed to be within 5 Å from each other in the C3d:FH complex.

occupancy interactions. Site 3 exhibits the least amount of hydrogen bonding with only two moderate occupancy interactions. Considering potential salt bridges within a cutoff distance of 5 Å (**Figure 5.7**), site 1 and 2 exhibit two high occupancy interactions. In addition, site 1 has three lower occupancy predicted salt bridges. Site 3 has several lower occupancy salt bridges that appear to fluctuate over the course of the trajectories (**Figure A.4**). In two replicate simulations at this site, there is an average of four salt bridges near the end of the trajectories; however, in one replicate, the number of salt bridges drops to an average number of one salt bridge.

*5.2.3 Steered Molecular Dynamics*

During steered molecular dynamics (SMD) simulations, FH CCP 19-20 is pulled away from C3d at a constant velocity. As a result, binding modes exhibit similar peak forces with different lengths of time until intermolecular forces are essentially eliminated (**Figure 5.8**). For site 2, forces minimize after approximately 1.3 ns while site 1 and 3 took about 0.8 and

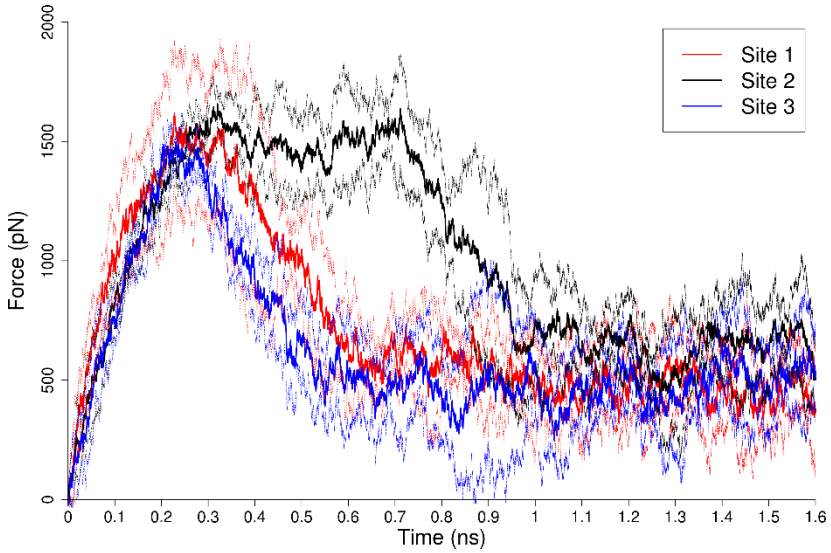



**Figure 5.8:** Forces required to move FH at a constant velocity along the surface normal vector are plotted over the length of time for each SMD simulation for each binding mode. Large, solid lines represent the mean values across five, independent trajectories for each binding mode. Smaller, solid lines represent one standard deviation above or below the mean value. Force-time responses are color coded according to the key in the top right corner.

0.6 ns, respectively. For site 1 the peak forces experienced greater variation, and the mean value seems to plateau slightly longer than site 3.

*5.2.2 Prediction of Association Rate Constants*

In order to characterize the associative properties of the C3d:FH(19-20) complex, we calculated association rate constants using the TransComp[14] computational method. For site 1, C3d and FH(19-20) are predicted to have rate constants of $3.53 \times 10^7$ $M^{-1}s^{-1}$, while site 2 and 3 are predicted to have rate constants of $3.26 \times 10^5$ and $6.05 \times 10^4$ $M^{-1}s^{-1}$, respectively. If electrostatics are neglected and only association from random diffusion is considered, sites 1, 2, and 3 were found to associate at rate constants of $5.48 \times 10^5$, $3.12 \times 10^5$, and $4.07 \times 10^5$ $M^{-1}s^{-1}$, respectively.

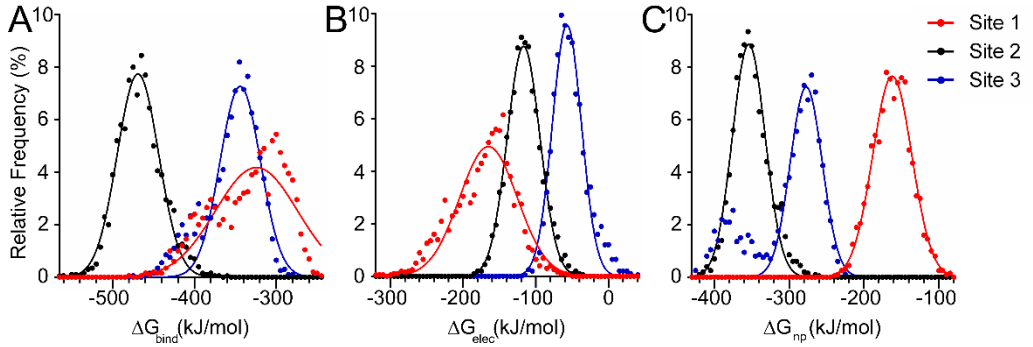



**Figure 5.9:** Histograms of energies from MM/GBSA calculations for each binding site are shown above with a bin size of 10 kJ/mol. (A) Distributions for change in the overall free energy of binding. (B) Distributions for change in free energy resulting from electrostatic interactions. (C) Distributions for change in free energy from nonpolar interactions. Energies were calculated for all but the first 100 ps of each MD trajectory, pooling replicates from the same binding modes.

### *5.2.5 Free Energy Calculations*

To assess the energetics of binding at sites 1-3, binding free energies are calculated using a molecular mechanics method with a generalized Born surface area approximation (MM/GBSA). In terms of free energy of binding between C3d and FH CCP 19-20, each binding mode is characterized by a distinct distribution of predicted energies that arise from different contributions by electrostatic and nonpolar interactions (**Figure 5.9**). Site 2 is predicted to have the lowest free energy that is largely based on nonpolar interactions in the protein complex. In addition to exhibiting the lowest free energy from nonpolar interactions of all binding modes, site 2 is predicted to have the second highest free energy of binding from electrostatic interactions. These results are in agreement with MD and AESOP analyses. In terms of overall predicted binding free energy, the distribution of energies overlap between site 1 and site 3, with site 1 exhibiting a much wider distribution of energies. Site 1 is predicted to have the largest contribution to free energy of binding from electrostatic interactions while having the lowest contribution from nonpolar

interactions. This reflects the low interfacial surface area expected from the end-to-end binding configuration and results from SMD analysis. Site 3, on the other hand, has the lowest electrostatic contribution to free energy of binding and the second highest nonpolar contribution. The bimodal distribution of energies from nonpolar interactions for site 3 likely reflects the gradually decreasing interfacial surface area that is present briefly during the last 10 ns of the MD simulations. In terms of statistical significance as described by Brown-Forsythe and Bartlett's tests, variances in distributions of overall binding free energies are significantly different across binding modes with a confidence level of 95%.

### 5.3 Discussion

Methods that assess the electrostatic nature of the interactions between C3d and FH note key differences in the basis of the interactions at each binding mode. AESOP suggests that site 1 has the most electrostatic character based on the greatest number of mutations in C3d and FH that are predicted to cause a change in binding. This result is also confirmed by MM/GBSA calculations. At site 2, AESOP suggests a lesser degree of electrostatic interactions based on the fewer number of mutations that result in predicted gain or loss of binding. Site 3, on the other hand, is suggested to interact primarily on the basis of nonpolar interactions based on the lack of mutations in AESOP that are predicted to increase or decrease binding in excess of energetic variations arising from thermal fluctuations. This lack of electrostatic character at site 3 is supported by results from the MM/GBSA calculations. Of key residues where mutations are linked to atypical hemolytic uremic syndrome (aHUS), site 1 exhibits the most mutations at these loci that result in loss of

binding (R1182, K1203). Additionally, most mutations of basic residues probably affect polyanion binding.

Given the variation in interfacial surface area that each binding mode exhibits, stability of the corresponding protein complexes is difficult to determine. . Without considering this variation, site 2 would certainly be the most stable from plotting interfacial SASA over time (**Figure 5.4**). Interfacial SASA for site 1 exhibits greater variation of interfacial SASA that may be caused in part by the end-to-end binding configuration between C3d and FH CCP 20. Interfacial SASA decreases at site 3 during the first part of the MD trajectory, suggesting the initial crystallized structure may not be physiological. This is not to say, however, that the binding mode is not physiological since crystallization of the protein complex could have induced structural perturbations. Rather the structure could be stable in a different configuration that approaches the structure towards the end of our MD simulation. As mentioned, interfacial surface area does not fully explain complex stability. In the same way, investigation of predicted protein interactions gives insight into the basis of the binding between C3d and FH while being skewed by varying interfacial surface areas. At site 2, we observed a large number of nearly time-invariant aliphatic interactions (**Figure 5.5**), suggesting a stable structure as well as a high degree of nonpolar interactions. Presence of several hydrogen bonds as well as two salt bridges (**Figure 5.6 and 5.7**) further implies stability of the site 2 complex. In contrast, site 1 lacks the number of stabilizing aliphatic interactions observed in other binding modes; however, it has a nearly time-invariant salt bridge (in combination with several other lower occupancy salt bridges and hydrogen bonds) that may promote binding of C3d and FH. This is not to say, however,

that the structure is unstable since this observation gives no insight into the resistance to external force in SMD simulations. Similar to site 2, site 3 has a large number of aliphatic interactions; however, these interactions tend to have lower occupancies than site 2. This result suggests slightly lower stability of the crystallized structure as one of three MD trajectories for site 3 decreases in number of aliphatic interactions present after 15 ns and does not stabilize. Moreover, there are relatively few predicted hydrogen bonds and salt bridges when compared to other binding modes.

From studying the dissociative properties throughout SMD simulations, binding modes seem to share similar stability despite variations in types of interactions. This conclusion is based on the similar peak force of 1500 pN that is observed across sets of triplicate SMD trajectories for each binding site. Despite their similarity, binding modes vary in the time required to separate the complex and eliminate protein interactions. Site 2 requires the longest amount of time to separate at 1.3 ns, suggesting a larger energetic barrier to dissociation. As mentioned in the results, sites 1 and 3 require similar amounts of energy to separate C3d and FH, with site 1 requiring slightly more energy on average.

As an alternative method to AESOP that accounts for both electrostatic and nonpolar interactions between C3d and FH CCP 19-20, MM/GBSA gives insights into the free energy of binding for each binding mode. Moreover, results from MM/GBSA are in agreement with our predicted outcomes from AESOP. As expected site 1 has the greatest electrostatic contribution to the free energy of binding, while site 3 has the least. Since AESOP does not account for hydrophobic interactions, we were unable to resolve the nonpolar contributions to the free energy of binding at site 2 from AESOP alone; however,

occupancy maps from MD analysis are in agreement with findings from MM/GBSA. Furthermore, MM/GBSA is in agreement with AESOP concerning the presence of electrostatic interactions at site 2 that are lesser in overall magnitude than the electrostatic contribution from site 1. When considering nonpolar interactions, MM/GBSA predicts that site 1 has the lowest contribution of nonpolar interactions to the free energy of binding while site 2 has the greatest nonpolar contribution. A discrepancy between analytical methods is found when comparing results from MM/GBSA and SMD. From SMD analysis, we expect site 1 to be slightly more thermodynamically stable than site 3; however, MM/GBSA predicts that site 3 has a slightly higher mean free energy of binding than site 1. This result may characterize some uncertainty associated with the overlapping distributions of binding free energies between sites 1 and 3 and with the larger distribution of energies from site 1.

When comparing our data from AESOP and protein interaction studies to mutations from literature that are considered to affect binding of C3d and FH, we observe the most agreement at residues within sites 1 and 2 (**Table A.1**). Specifically, mutations at residues D36, E160, D163, and Q168 on C3d and at residues R1182, E1198, R1203, and R1215 on FH suggest site 1 to be important to C3d and FH binding. Furthermore, mutations at residues E117 and D122 on C3d and at residues P1114, D1119, and Q1139 on FH suggest site 2 is also important for C3d binding. In contrast, site 3 lacks mutations that suggest a role in C3d and FH binding. There is one mutation mentioned in literature, however, that may result in increased binding between C3d and FH at site 3.

As expected, mutation of residues predicted to participate in salt bridge interactions between C3d and FH corresponds to loss of binding from AESOP. At site 1, alanine mutations of E160, E167, and D292 on C3d and of R1182, R1203, R1215, and K1230 on FH are predicted to cause loss of binding. Each of these residues participates in a high occupancy salt bridge interaction from our MD simulations. At site 2, alanine mutations of K112 and D122 on C3d and of D1119 and K1188 on FH are predicted to cause loss of binding. Once again, these residues participate in high occupancy salt bridge interactions from MD simulations. These results show agreement between AESOP and MD simulations. In general, we are not able to compare aliphatic or hydrogen bond interactions (that are not salt bridges) to results from AESOP, since our computational framework focuses on mutation of ionizable residues. In the directed mutagenesis studies, however, we did perform the mutation Q1139G which resulted in loss of binding at site 2. From MD analysis, Q1139 participated in a hydrogen bond interaction at site 2. Overall, each residue that participates in a high occupancy salt bridge at one binding site is observed as loss of binding mutation in AESOP at that corresponding site.

After comprehensive characterization of electrostatics, structural stability, dissociative properties, associative properties, and predicted free energy of binding, we suggest the physiological relevance of sites 1 and 2 and propose that FH may interact with C3b/d in a multivalent manner. This conclusion is in agreement with experimental studies and mechanisms proposed in literature.[3,5,15] We find that binding at site 3 is less likely to be physiological since simulations seem less stable, comparatively, and superposition of the C3d:FH(19-20) structure with the structure of C3b (PDB 2I07) reveals clashing of CCP 19

with a domain of C3b (**Figure A.5**). We cannot rule out binding at site 3, however, since intermodular flexibility could resolve the clashes with C3b. This same structural comparison suggests that binding of FH CCP 19-20 at site 1 or site 2 is feasible (**Figure A.5**). In at least one study, binding at site 1 was suggested to require a conformational change of C3b.[3] We suggest that the one clash observed from the superpositioning of FH CCP 19-20 with C3b of site 1 may be eliminated by flexibility in the end-to-end binding configuration. With this apically flexible structure, the complex may remain stable as a result of electrostatic interactions. We propose that sites 1 and 2 represent the major modes of binding between C3b and FH. Our results indicate that the complex at site 2 is very stable with a high degree of hydrophobic and electrostatic interactions and the most favorable binding free energy. Site 1, on the other hand, contains the negatively charged acidic patch of C3d and the positively charged FH (CCP 20) that may promote longer-range interactions of FH with C3d or C3b. This property, combined with insights from experimental studies (**Table A.1**), suggests that site 1 is potentially more important than site 2 for binding of FH to host surfaces.

In literature, two studies suggest competing modes of binding. Morgan *et al* suggested that the physiological mode of binding between C3d and FH CCP 19-20 occurs at site 3 with CCP 20 binding to the thioester domain and polyanionic surfaces simultaneously.[2] In this manner, the study hypothesized that FH CCP 19-20 can distinguish between self and non-self. While we cannot rule out this mode of binding, our computational study suggests that such a complex would be less stable and require significant bending at the intermodular domain of CCP 19 and 20. A second study by Kajander *et al* suggested a physiological

mode of C3b binding at site 2 through FH CCP 19, excluding binding at site 1 for steric reasons.[3] Furthermore, they hypothesized a dual interaction of FH CCP 19-20 with C3b at site 2 and either C3d at site 1 or host GAGs. Through this multivalent interaction, literature suggests avidity of FH to self surfaces is greatly increased and discrimination between self and non-self is facilitated.[3,5] Our study is in agreement that site 2 is important for C3b binding; however, we further suggest that site 1 may play an important role. Due to the electrostatic nature of the protein interaction, we hypothesize the apical binding conformation of FH CCP 20 is flexible, yet stable. Furthermore, literature mentions bacterial proteins that confer immune resistance bind to the acidic patch of C3d (site 1) in a crystallographic structure, while binding with nanomolar affinity to C3b.[16] Thus, site 1 is suggested to play some role in complement regulation and may be accessible in C3b. With regard to the dual interaction of CCP 19-20 with host surfaces, our study is unable to directly assess this behavior. However, structure superpositioning seems to indicate that CCP 20 is largely accessible when binding occurs at site 2 through CCP 19 (**Figure 5.5**).

Assuming that a dual interaction of FH CCP 19-20 does occur, we hypothesize that CCP 19 and 20 complement each other in such a way to increase the apparent affinity of FH for host surfaces. Kajander *et al* mention that mutations in CCP 20 significantly reduce the binding of FH(19-20) to C3b.[3] We propose that site 1 may be the initial binding mode that precedes binding at site 2 where there exists a higher energetic barrier to dissociation. The electrostatic nature of binding at site 1 may increase the apparent on rate of CCP 19 to site 2. This effect is likely more influential *in vivo* where CCP 20 is additionally capable

of binding to C3d or host GAGs. Additionally, the more thermodynamically stable binding at site 2 likely decreases the apparent off rate of CCP 20 from either C3d or host GAGs.

In conclusion, site 1 and 2 appear to be the most physiologically relevant binding modes. To assess the dual interaction of FH(19-20) with host surfaces, future mechanistic studies can investigate the structural stability of the resulting quaternary structure. Additionally, mechanistic studies may more directly assess feasibility of FH(19-20) binding at site 1 on C3b. Further exploration of protein interactions between FH CCP 19-20 and C3b at sites 1 and 2 is required for a deeper understanding of the physiological implications that are relevant to describing the pathophysiology of aHUS.

## 5.4 Methods

### *5.4.1 Structure Preparation*

Crystallographic structures of C3d in complex with FH were downloaded from the PDB using the identification codes 3OXU and 2XQW. From each structure all pairs of interacting protein chains were identified and superimposed based on the structure of C3d within the chain pair. In each sequence expression tags were removed and all chains were trimmed for consistency across all models. In effect, C3d contained residues 10 through 294, and FH contained residues 1109-1166 of CCP 19 and residues 1167-1230 of CCP 20. Note that the first residue of FH after the expression tag is residue 1107. For electrostatic analysis, each pair of interacting protein chains from each PDB model was examined. For subsequent molecular dynamics simulations, only one representative chain pair from 3OXU was used for each of the three binding modes. Site 1 consists of chains B and D;

site 2 consists of chains B and F, and site 3 consists of chains A and D as illustrated in **Figure 5.1**. Note that the 2XQW crystallographic structure mutated aspartic acid 1119 to glycine and glutamine 1139 to alanine in order to enable crystallization of the protein complex. From the important roles these residues play, the mutations pose significant threats to the integrity of the structural model; however, the 2XQW structure remains in general agreement with 3OXU.

*5.4.2 AESOP*

In order to analyze the energetic contributions of each individual amino acid to the binding of C3d and FH, our lab has developed and implemented a computational framework for predicted changes in free energy of binding using a mutagenesis perturbation approach.[17–21] This method substitutes alanine (or any other amino acid or combination of amino acids) for each ionizable amino acid and predicts the change in free energy of binding, according to

$$\Delta G_{bind} = \Delta G_{mutant} - \Delta G_{parent} \quad \textbf{(5.1)}$$

$$\Delta G_{association} = \Delta G_{Coulombic} - \Delta\Delta G_{solvation} \quad \textbf{(5.2)}.$$

The predicted electrostatic free energies of binding for the mutants are relative to the parent protein, given by **Equation 5.1**, where $\Delta G_{mutant}$ and $\Delta G_{parent}$ correspond to association free energies, calculated using a thermodynamic cycle that includes Coulombic and solvation effects as previously described[18,19] and shown in **Equation 5.2**. To prepare our collection of structures, a PQR file was generated for each structure of interacting chain pairs (PDB structure) using PDB2PQR[22,23] and the PARSE force field.[24,25] This file contained hydrogen atoms omitted in the PDB file and parameters necessary for

electrostatic calculations. APBS[26] was used for all electrostatic calculations using a solvent dielectric value of 78.54, protein dielectric constant of 20, and a grid spacing of 1 Å. Furthermore, ionic strength of the solvent was set to 150 mM. Electrostatic free energies of association were calculated as described previously.[21]

In addition to the comprehensive computational alanine scan, we performed computational site directed mutagenesis. All parameters from the previous method were used; however, SCWRL4[27] was used to replace the specified residues with another user-defined residue. Mutations were selected to compare against previous results found in literature[2,3]; however, we only performed mutations on residues that are ionizable or contain partial charges as this is a limitation of our computational method.

*5.4.3 Conventional Molecular Dynamics*

For each site of interaction on C3d with FH, three explicit solvent MD simulations were performed using NAMD[28] and the CHARMM27 force field.[29] Structures were initially minimized with the water molecules contained in the crystallographic structure and subsequently solvated in a cubic TIP3P water box leaving a minimum margin of 12 Å around the protein structure. Sodium and chloride ions were also added to the water box, bringing the ionic strength to 150 mM. Minimization of the solvated structure was performed for 25000 steps (50 ps) before heating to 310 K over 64 ps. Next, equilibration was performed where all protein atoms were constrained at 10 kcal/mol/Å$^2$ initially and then relaxed to 5, 2, and 1 kcal/mol/Å$^2$ before removing constraints altogether for the final equilibration run. Following equilibration, the MD simulation was run with periodic boundary conditions, Langevin temperature and pressure control, and particle mesh Ewald

(PME) electrostatics. Relevant parameters for the simulation included a nonbonded interaction cutoff of 12 Å and a switching distance of 10 Å. Hydrogen bonds were held constant according to the SHAKE algorithm as an integration time step of 2 fs was used. At least one of these simulations was run for 20 ns to establish the time required to reach structural stability within the simulation (i.e. a plateau of interfacial surface area and atomic RMSD throughout the trajectory). The remaining two simulations were run for 10 ns except in the case of site 3 where simulations required extension to 20 ns.

*5.4.4 Steered Molecular Dynamics*

Using the longest MD simulation of each binding site on C3d, snapshots from the stabilized portion of the trajectory were clustered according to the RMSD of the FH CCPs relevant for the corresponding binding mode. Specifically, RMSD was based on CCP 20 for site 1 and 3 but based on CCP 19 and 20 for site 2. The resulting dendrograms were split into 5 bins where the medoid of each set was reported as a representative structure for the corresponding binding site. In order to prepare the SMD simulation, each structure was oriented so the force was applied to FH in the z-direction. The z-direction also corresponds to the normal vector associated with the surface of C3d within 8 Å from FH. Each resulting structure was minimized without solvent for 5000 steps before solvation in a TIP3P water box with a minimum margin of 12 Å in all dimensions and a maximum margin of 36 to 46 Å in the z-direction to allow for a minimum separation of about 10 Å where interactions between protein chains were effectively eliminated. Addition of ions was performed as in MD simulations followed by minimization of 25000 steps (50 ps). Subsequent heating to 310 K was performed over 32 ps and was followed by one equilibration step with a

constraint constant of 10 kcal/mol/Å$^2$. In the actual SMD production run, all residues on C3d greater than 12 Å from FH were constrained with a constant of 10 kcal/mol/Å$^2$, and a constant velocity of 10 Å/ns was applied on the center of mass of FH. For site 1 and 3, the center of mass was calculated from all residues in CCP 20, and for site 2 the center of mass was calculated from all residues in CCP 19 and 20. For all stages in SMD simulations, force-field parameters were retained from the MD simulations, and periodic boundary conditions were only used in minimization with solvent and heating. Langevin temperature and pressure controls, PME grid electrostatics, and rigid hydrogen bonds were used in all steps except minimization without solvent. Furthermore, an integration time step of 1 fs was used for all steps after minimization with solvent.

*5.4.5 Free Energy Calculations*

Excluding the first 100 ps for each MD simulation trajectory of C3d and FH, we calculated free energies of binding for all remaining snapshots according to standard MM/GBSA methods[19,30] for the set of replicates corresponding to each binding mode, according to

$$\Delta G_{AB,n} = G_{AB,n} - G_{A,n} - G_{B,n} \qquad \textbf{(5.3)}$$

$$\Delta G_{bind} = \Delta G_{np} + \Delta G_{elec} \qquad \textbf{(5.4)}$$

$$\Delta G_{np} = \Delta E_{vdw} + \Delta G_{np,solv} \qquad \textbf{(5.5)}$$

$$\Delta G_{elec} = \Delta E_{Coul} + \Delta G_{elec,solv} \qquad \textbf{(5.6)}.$$

Changes in free energy are calculated according to **Equation 5.3**, where AB represents the protein complex and A or B individually represents a protein of the complex. Free energy of binding was calculated and reported according to **Equation 5.4** as described in

previous, published work.[19,30] The first terms on the right hand side of **Equation 5.5-5.6** are calculated using atomic radii and the CHARMM27 force field.[29] The second right hand term of **Equation 5.5** is calculated by multiplying a surface tension parameter ($\gamma$= 0.005 kcal/mol/Å$^2$) by the protein complex interfacial surface area, while the second right hand term of **Equation 5.6** is calculated from Generalized Born approximation within CHARMM.

*5.4.6 Association Rate Constant Prediction*

We used the TransComp[14,31,32] web server (http://pipe.sc.fsu.edu/transcomp/) to predict association rate constants with an ionic strength of 150 mM. For each binding site, we used the same chain pairs from PDB 3OXU as for MD simulations. Through a combination of force-free Brownian dynamics simulations and electrostatic free energy calculations, this method provides both the association rate constant from random diffusion and an overall association rate constant from both diffusion and long range electrostatic interactions.[14]

*5.4.7 Analytical Tools*

Analysis of computational data was largely performed within R[33] using the Bio3D[34] package as necessary. Visualization and analysis of protein interactions from the simulation trajectory snapshots was performed within UCSF Chimera.[35] In particular, Chimera was used to calculate hydrogen bonds. Aliphatic and salt bridges were calculated with custom scripts in R. VMD[36] was used translate atomic coordinates within proteins as well as calculate interfacial surface area. Interfacial surface area was calculated by subtracting individual SASA values for each protein from the SASA of the protein

complex. Our AESOP framework is available upon request (http://biomodel.engr.ucr.edu/software/index.html).

## 5.5 References


1. Thurman, J. M. & Holers, V. M. The Central Role of the Alternative Complement Pathway in Human Disease. *J. Immunol.* **176,** 1305–1310 (2006).

2. Morgan, H. P. *et al.* Structural basis for engagement by complement factor H of C3b on a self surface. *Nat. Struct. Mol. Biol.* **18,** 463–471 (2011).

3. Kajander, T. *et al.* Dual interaction of factor H with C3d and glycosaminoglycans in host-nonhost discrimination by complement. *Proc. Natl. Acad. Sci.* **108,** 2897–2902 (2011).

4. Schmidt, C. Q., Herbert, A. P., Hocking, H. G., Uhrín, D. & Barlow, P. N. Translational Mini-Review Series on Complement Factor H: Structural and functional correlations for factor H. *Clin. Exp. Immunol.* **151,** 14–24 (2008).

5. Gros, P. In self-defense. *Nat. Struct. Mol. Biol.* **18,** 401–402 (2011).

6. Kieslich, C. A., Vazquez, H., Goodman, G. N., López de Victoria, A. & Morikis, D. The effect of electrostatics on factor H function and related pathologies. *J. Mol. Graph. Model.* **29,** 1047–1055 (2011).

7. Kieslich, C. A. & Morikis, D. The Two Sides of Complement C3d: Evolution of Electrostatics in a Link between Innate and Adaptive Immunity. *PLoS Comput. Biol.* **8,** e1002840 (2012).

8. Szakonyi, G. *et al.* Structure of complement receptor 2 in complex with its C3d ligand. *Science (80-. ).* **292,** 1725–1728 (2001).

9. Isenman, D. E., Leung, E., Mackay, J. D., Bagby, S. & van den Elsen, J. M. H. Mutational analyses reveal that the staphylococcal immune evasion molecule Sbi and complement receptor 2 (CR2) share overlapping contact residues on C3d: implications for the controversy regarding the CR2/C3d cocrystal structure. *J. Immunol.* **184,** 1946–1955 (2010).

10. Elsen, J. M. H. Van Den. A Crystal Structure of the Complex between Complement Receptor 2 and its Ligand C3d. **608,** 608–612 (2011).

11. Shaw, C. D. *et al.* Delineation of the Complement Receptor Type 2-C3d Complex by Site-Directed Mutagenesis and Molecular Docking. *J. Mol. Biol.* **404,** 697–710 (2010).

12. Kieslich, C. A., Morikis, D., Yang, J. & Gunopulos, D. Automated computational framework for the analysis of electrostatic similarities of proteins. *Biotechnol. Prog.* **27,** 316–325 (2011).

13. Morikis, D. *F1000Prime Recommendation of [van den Elsen JM and Isenman DE, Science 2011 332(6029):608-11]*. (2011). doi:10.3410/f.10371956.14018054

14. Qin, S., Pang, X. & Zhou, H. X. Automated prediction of protein association rate constants. *Structure* **19,** 1744–1751 (2011).

15. Mohan, R. R., Gorham Jr, R. D. & Morikis, D. A theoretical view of the C3d:CR2 binding controversy. *Mol. Immunol.* **64,** 112–122 (2015).

16. Hammel, M. *et al.* A structural basis for complement inhibition by Staphylococcus aureus. *Nat. Immunol.* **8,** 430–437 (2007).

17. Kieslich, C. A. Development and Applications of a Computational Framework for Protein and Drug Design. *UC Riverside Electronic Theses and Dissertations* (University of California, Riverside, 2012).

18. Gorham, R. D., Kieslich, C. A. & Morikis, D. Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Ann. Biomed. Eng.* **39,** 1252–1263 (2011).

19. Gorham, R. D., Rodriguez, W. & Morikis, D. Molecular analysis of the interaction between staphylococcal virulence factor Sbi-IV and complement C3d. *Biophys. J.* **106,** 1164–1173 (2014).

20. Kieslich, C. A., Gorham, R. D. & Morikis, D. 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 (2011).

21. Gorham, R. D. *et al.* An evaluation of Poisson-Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. *Biopolymers* **95,** 746–754 (2011).

22. Dolinsky, T. J. *et al.* PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35,** 522–525 (2007).

23. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32,** 665–667 (2004).

24. Sitkoff, D., Sharp, K. A. & Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98,** 1978–1988 (1994).

25. Tang, C. L., Alexov, E., Pyle, A. M. & Honig, B. Calculation of pKas in RNA: On the Structural Origins and Functional Roles of Protonated Nucleotides. *J. Mol. Biol.*

**366,** 1475–1496 (2007).

26. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10037–41 (2001).

27. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins Struct. Funct. Bioinforma.* **77,** 778–795 (2009).

28. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).

29. MacKerell, J. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102,** 3586–3616 (1998).

30. Kieslich, C. A. *et al.* Exploring Protein-Protein and Protein-Ligand Interactions in the Immune System using Molecular Dynamics and Continuum Electrostatics. *Curr. Phys. Chem.* **2,** 324–343 (2012).

31. Alsallaq, R. & Zhou, H. X. Electrostatic rate enhancement and transient complex of protein-protein association. *Proteins Struct. Funct. Genet.* **71,** 320–335 (2008).

32. Qin, S. & Zhou, H. X. Prediction of salt and mutational effects on the association rate of U1A protein and U1 small nuclear RNA stem/loop II. *J. Phys. Chem. B* **112,** 5955–5960 (2008).

33. R Development Core Team. R: A Language and Environment for Statistical Computing. (2008).

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

35. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

36. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).

CHAPTER 6

# Factor H Inspired Design of Peptide Biomarkers of Complement Component 3d

## 6.1 Introduction

In prior chapters we have focused on the role of C3 and C3 cleavage fragments in complement system function. All three complement pathways – the alternative, classical, and lectin – converge to the formation of the C3 convertase which cleaves C3 into C3a and C3b.[1] As previously described, C3b is an opsonin and covalently bonds to surfaces of cells, both foreign and native.[1] Furthermore, this bound C3b can associate with other complement factors to form C3 convertase. Should such activation of C3 result in covalent attachment of C3b to host cells, a number of complement regulators act to prevent a state of autoimmunity by either preventing formation of C3 convertase or promoting the decay of any C3 convertase that forms. In the latter case, bound C3b will eventually be degraded into C3d.[1,2] C3d contains only the thioester domain of C3b and retains the covalent bond to the surface. As such, bound C3d is not susceptible to further degradation except by recycling of cell membrane.

Accumulation of C3d on membranes of tissues is a hallmark of complement activation. Though most tissues of self will be tagged with some small amount of C3d, an immune response from the complement system will result in more C3b that will bind non-specifically to foreign surfaces and native tissues alike. Complement regulators expressed by native cells then promote degradation of C3b to C3d, leading to accumulation of C3d on tissues of self. As such, C3d is an attractive biomarker to monitor complement activity.

Thus, molecules with fluorescent properties could be designed to bind C3d for imaging via confocal fluorescence microscopy.

To this end, our lab has previously conducted virtual screening studies to search for small molecules with intrinsic fluorescence that bind C3b.[2] While several such compounds were found, C3d binding affinities were generally moderate to low with dissociation constants between 1 to 500 μM.[2] Moreover, these molecules exhibited low solubility in aqueous solvents[2] and non-specific binding in cell imaging studies of the retinal pigmented epithelium (unpublished work). As peptides tend to be more soluble and easier to synthesize than small-molecules, we decided to design new, C3d-binding peptides with diagnostic potential.

To this end, we leveraged crystal structures of the C3d and Factor H (FH) complex to design a new generation of C3d-targeting peptides.[3,4] Given prior work (chapter 5) studying the putative binding modes of C3d and FH, we designed peptides that target both the site 1 and site 2 binding modes. Both free energy calculations and interactions from molecular dynamics simulations were leveraged to identify key residues that contribute to binding. These key residues participated in high-occupancy salt-bridge, hydrogen bond, or aliphatic interactions with C3d.[4] Subsequently, we designed sequences that contain 3 persistent salt-bridges or hydrogen bonds. For the initial site 1 peptide, template residues of FH included R1203, R1206, and K1230. For the initial site 2 peptide, template residues of FH included D1119, I1120, and Q1139. To promote the desired topology of interactions, we further relied on cyclization of peptide designs to constrain conformational space. Through iterative experimental binding assays and long time-scale molecular dynamics simulations,

we were able to design new diagnostic peptides with 62-2200 μM affinities and excellent solubility.

## 6.2 Methods

### *6.2.1 Structural Models of Peptides*

Peptide structures were generated from Protein Data Bank identifier 3oxu[3] at either putative binding site 1 or 2 with the use of homology modelling techniques. Specifically, 3oxu chains B and D were used to model peptides at site 1, and 3oxu chains B and F were used to model peptides at site 2. In the event template chains B, D, and F from 3oxu were missing atoms, we performed homology modelling using Modeller[5,6] to generate missing, non-terminal loops as in chapter 5. To model specific peptides, templates were trimmed to contain C3d and a subset of residues from FH. For site 1, template FH residues included residue numbers 1201-1206 and 1228-1230. For site 2, template FH residues included residue numbers 1117-1120 and 1136-1139. To generate peptide models, standard homology modelling protocols were applied to peptide sequences using Modeller[5,6] with patches for disulfide bridges that cyclize peptide designs.

### *6.2.2 Molecular Dynamics*

Proteins and peptides were simulated with OpenMM[7] using the amber14 ff14SB forcefield[8] parameters and the associated tip3p water model. Proteins and peptides were solvated into an explicit solvent with a 10 Å bounding box and 150 mM NaCl (neutral net charge). Protonation states of peptides and proteins were assigned according to a pH of 7.4, and system temperature was set to 310 K.

After system preparation, atomic motion of molecules was simulated with a Langevin integrator (friction coefficient 1 $ps^{-1}$), particle mesh Ewald electrostatics (10 Å cutoff, 0.0005 error tolerance), and a Monte Carlo barostat (1 atm). Initial velocities were randomly assigned from a Maxwell-Boltzmann distribution at 310K. Subsequently, systems were subject to energy minimization until convergence (10 kJ/mol). Finally, production molecular dynamics simulations were performed for a user-defined period of time, saving coordinates every 100 ps.

Analysis of molecular dynamics trajectories was performed with custom scripts in Python[9] leveraging the MDTraj[10] library. Calculations of root-mean-square fluctuations (RMSF) from the mean position for each atom were performed after imaging molecules (keeping protein-peptide complexes in the same frame of reference) and superimposing backbone atoms of C3d from all frames in the trajectory on the first frame. For plotting purposes, RMSF values for groups of atoms within a single residue were averaged together, though this only occurred if atomic selections include more than one atom in a least one residue. Construction of free energy landscapes of peptides were performed using the PyEMMA package.[11] For these figures, principle components of atomic positions in Cartesian space were found after superposing all frames in the peptide-only molecular dynamics trajectories.

*6.2.3 Binding Assays*

Microscale thermophoresis experiments[12–14] were conducted for all peptides to estimate dissociation constants for peptide-C3d complexes, if possible. After amine labeling of C3d (Complement Technology, Inc.) with an Alexa-type fluorophore (MO-

L002, NanoTemper Technologies, Inc.), C3d was transferred to MST buffer (50 mM tris-HCl pH 7.4, 150 mM NaCl, 10 mM $MgCl_2$, 0.05% Tween-20). Peptide was solvated separately with MST buffer at or near max solubility (2-10 μM). Next a 1:1 dilution series was generated for the peptide, starting from either max concentration or a lower value chosen to optimize sampling of the dose-response curve. Each dilution was mixed in a 1:1 manner with labeled protein and incubated for 5 minutes before being measured in the microscale thermophoresis device (Monolith NT.115, NanoTemper Technologies, Inc.). Data was processed with instrument software and fit to a previously described equation to estimate the dissociation constant.[12]

## 6.3 Results and Discussion

### *6.3.1 First Generation Peptide Designs*

Inspired by sites 1 and 2, we designed two primary peptides that mimic FH by preserving amino acids that participate in key interactions with C3d (**Figure 6.1**).[4] For naming peptides, we adopt a scheme that identifies the C3d binding site being targeted as well as the peptide number. According to this scheme S1P1 is the first peptide that targets site 1, and S2P1 is the first peptide that targets site 2. Due to the lack of complex stability at site 3, no peptides were designed based on this binding mode. S1P1 was designed to preserve interactions of residues 1201-1206 and 1228-1230 from FH with C3d. Thus, we settled on the sequence **[CKRGYRC]gpgAK**. Brackets denote cyclization while lower case denote amino acids (single-letter amino acid code) not present in the sequence of FH. Similarly, S2P1 was designed to preserve interactions of residues 1117-1120 and 1136-

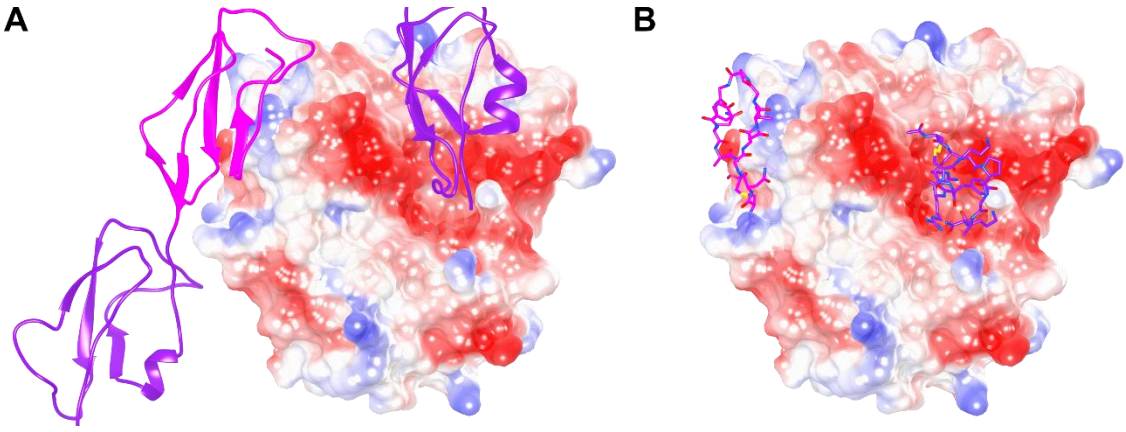



**Figure 6.1:** (A) Crystallographic complex of FH SCR19-20 in complex with C3d compared to (B) models for peptides S1P1 and S2P1 in complex with C3d. Coulombic potentials are projected onto the surface of C3d where red indicates a negative potential of -10 kcal/mol/e and blue indicates a positive potential of 10 kcal/mol/e. In panel A, SCR19 of FH is colored magenta and SCR20 is colored purple. In panel B, S1P1 is colored purple as it is inspired by FH SCR20, and S2P1 is colored magenta as it is inspired by FH SCR19.

1139 from FH with C3d. Thus, the sequence for S2P1 was **[cNGDIgpgYQaQc]**. Both S1P1 and S2P1 have acetylated N-termini, amidated C-termini, and cyclized designs (**Figure 6.1B**). For our purposes, cyclization was primarily used to constrain conformational space in an effort to promote binding with C3d.

Molecular dynamics over a time period of 1 μs shows that both S1P1 and S2P2 have structural minima near the desired conformation. Root-mean-square fluctuations (RMSF) in S1P1 are slightly higher than S2P1 on average (**Figure 6.2A-B**); however, this may not be important in the context of the C3d-peptide complex. In S1P1, the largest fluctuation observed is located in an uncyclized region at K12. Notably, simulations predict conformational energy minima that are proximal to the desired structure (**Figure 6.2C-D**).

We tested our initial designs using the technique of microscale thermophoresis to estimate dissociation constants for the C3d-peptide complexes. For S1P1, we found weak binding with a dissociation constant of 2.1 ± 1.6 mM (**Figure 6.3**). The dose-response

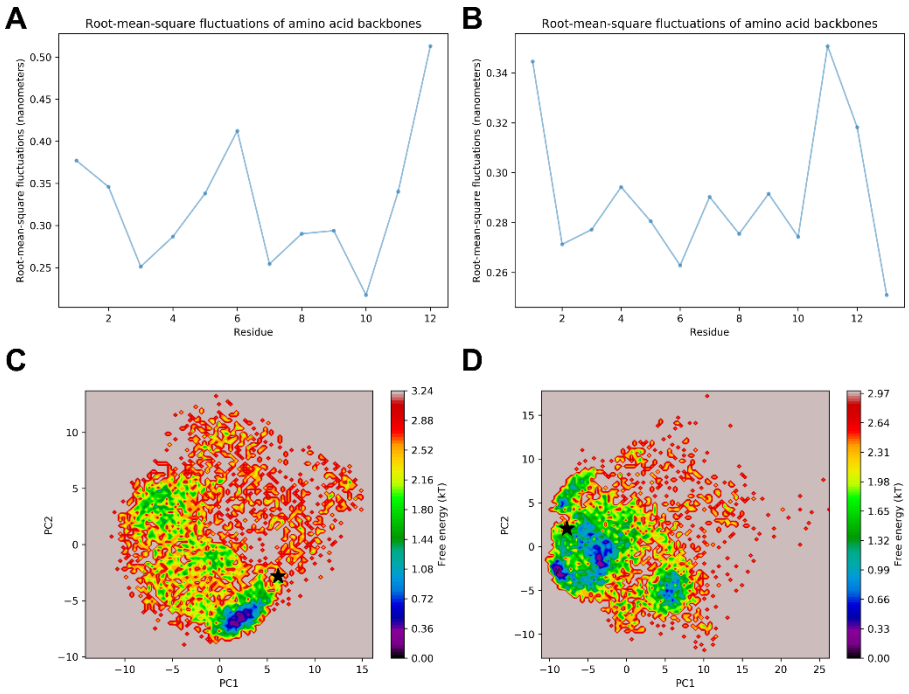



**Figure 6.2:** Structural stability of peptides was assessed with molecular dynamics. Panels A and B show root-mean-square fluctuations for backbone atoms on a per-residue basis for S1P1 and S2P1, respectively. In panels C and D the free energy landscapes for S1P1 and S1P2, respectively, are shown with respect to the desired conformation annotated with a star. In panels C-D, axes are the first and second principle components that account for the most variance of atomic fluctuations (separate state space for each peptide).

observed was not fully formed, contributing to higher uncertainty in the dissociation constant. In the case of S2P1, no binding was observed within the range of peptide concentrations tested. Given the experimental results, we selected S1P1 for further optimization.

*6.3.2 Second Generation Peptide Designs*

To guide optimization of S1P1, we conducted another molecular dynamics simulation of the C3d-S1P1 complex for 1 μs. Through visual observation of the trajectory, we observed that S1P1 began to dissociate from C3d at 50 ns before reassociating and binding in a different configuration at 75 ns. This new complex was relatively stable for the remainder of the simulation, except for the uncyclized tail that switched between positions.

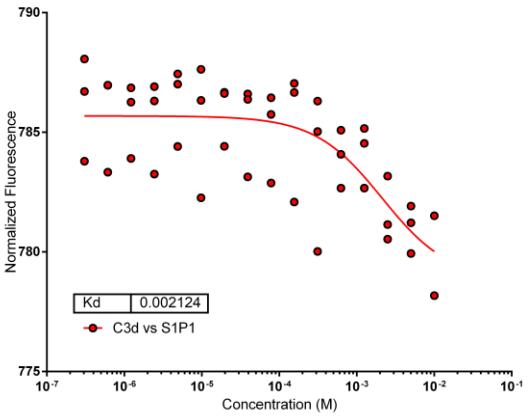



**Figure 6.3:** Thermophoretic response of fluorescently labeled C3d is observed at varying concentrations of S1P1. While not fully formed due to low affinity of S1P1 for C3d, the dose-response curve suggests binding. The dissociation constant is estimated at 2.1 mM (units of inset dissociation constant are M).

We recalculated RMSF for S1P1 from this simulation and observed large RMSF in $C_\alpha$ position at K12 (**Figure 6.4**). Thus, we hypothesized that the uncylized region (positions 8-12) may not contribute favorably to C3d binding.

In order to test our hypothesis, we designed a new peptide, S1P2, with the sequence **g[CKRGYRC]g**. In this construct we removed the flexible tail from S1P1 and added glycines to each end of the peptide. Since we ultimately plan to add fluorophores to the peptide, glycines act as placeholders to ensure that an amino acid conjugated to a

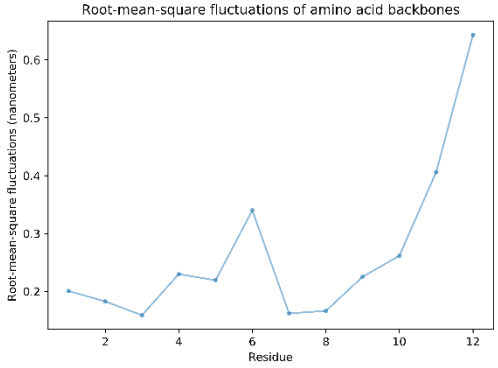



**Figure 6.4:** RMSF of backbone atoms in S1P1 during a simulation of the C3d-S1P1 complex. Note the increased RMSF at A11 and K12. These residues are located in the flexible tail of S1P1, and given the high fluctuations compared to the rest of S1P1, they may not contribute favorably to binding.

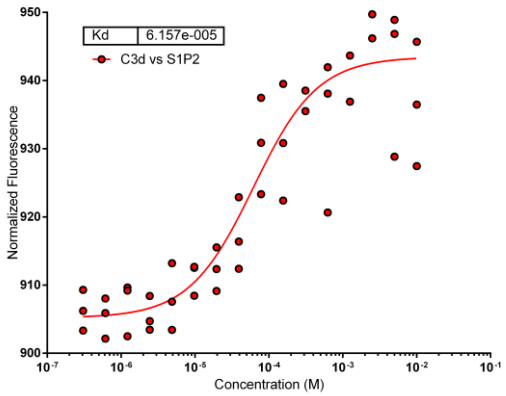



**Figure 6.5:** Thermophoretic response of fluorescently labeled C3d is observed at varying concentrations of S1P2. The dose-response curve is fully formed, and we estimate a dissociation constant of 61.6 μM (units of inset dissociation constant are M).

fluorophore could be located on either side. We used microscale thermophoresis to assess and quantify binding of S1P2 to C3d. Once again, a dose-response was observed (**Figure 6.5**); however, S1P2 was estimated to have a binding affinity of 61.6 ± 20.0 μM. Thus, S1P2 exhibits higher affinity for C3d than S1P1.

## 6.4 Conclusion

Leveraging known crystallographic complexes between C3d and FH and methods from computational chemistry, we set out to design new peptides to image deposition of C3d. Through optimizing our initial FH-inspired design with observations from molecular dynamics, we were able to design a novel peptide with low, micromolar affinity for C3d. Given the iterative nature of peptide design, S1P2 is a crucial step towards development of high-affinity peptides for imaging C3d deposition.

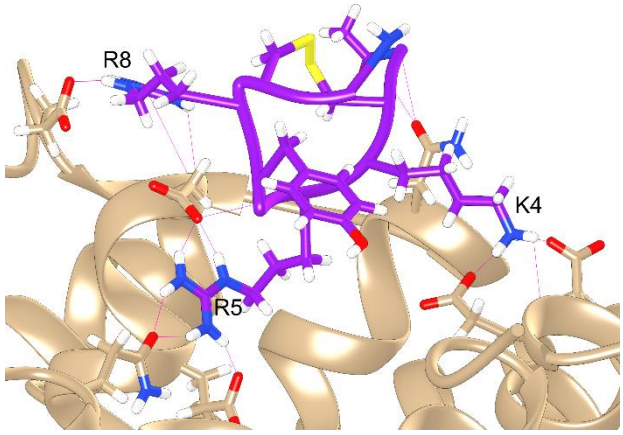



**Figure 6.6:** S1P2 (purple) is designed to bind the acidic patch of C3d (tan). From energy minimization of the S1P2 model superimposed on residues of FH from the site 1 binding mode, we predict salt-bridge interactions between residues of the acidic patch and peptide residues K4, R5, and R8. Hydrogen bonds are displayed as magenta lines.

Going forwards, we plan to further optimize S1P2 and introduce cyanine fluorophores. Optimization will once again be guided by molecular dynamics simulations. Those peptide positions with stable interactions with C3d will be retained, and those peptide positions that are less stable will be varied via a genetic algorithm. Given the favorable electrostatic interactions predicted to occur between S1P2 and C3d, positively charged residues from S1P2 such as K4, R5, and R8 will likely be retained (**Figure 6.6**). Following this optimization protocol, we will couple a cyanine dye to the peptide caps (terminal glycines in S1P2). As one of our long-term goals is to use a molecule based on S1P2 to image C3d deposition in primates, cyanine 7 can be used given an emission wavelength in the near infrared.

## 6.5 References


1. Mathern, D. R. & Heeger, P. S. Molecules great and small: The complement system. *Clin. J. Am. Soc. Nephrol.* **10,** 1636–1650 (2015).

2. Gorham, R. D. *et al.* Discovery of Small Molecules for Fluorescent Detection of Complement Activation Product C3d. *J. Med. Chem.* **58,** 9535–9545 (2015).

3. Morgan, H. P. *et al.* Structural basis for engagement by complement factor H of C3b on a self surface. *Nat. Struct. Mol. Biol.* **18,** 463–471 (2011).

4. Harrison, R. E. S., Gorham, R. D. & Morikis, D. Energetic evaluation of binding modes in the C3d and Factor H (CCP 19-20) complex. *Protein Sci.* **24,** 789–802 (2015).

5. Fiser, A., Kinh Gian Do, R. & Sali. Modeling Loops in Protein Structures. *Protein Sci.* **9,** 1753–1773 (2000).

6. Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).

7. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13,** 1–17 (2017).

8. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).

9. Python Software Foundation. Python Language Reference, version 2.7.

10. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109,** 1528–1532 (2015).

11. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11,** 5525–5542 (2015).

12. Dijkman, P. M., Lea, W. a, Gmbh, N. T., Lazic, A. & Gmbh, N. T. Interactions under Previously Challenging Conditions. **59,** 301–315 (2014).

13. Alexander, C. G. *et al.* Novel microscale approaches for easy, rapid determination of protein stability in academic and commercial settings. *Biochim. Biophys. Acta - Proteins Proteomics* **1844,** 2241–2250 (2014).

14. Wienken, C. J., Baaske, P., Rothbauer, U., Braun, D. & Duhr, S. Protein-binding assays in biological liquids using microscale thermophoresis. *Nat Commun* **1,** 100 (2010).

CHAPTER 7

# Molecular Mechanisms of Macular Degeneration Associated with the Complement Factor H Y402H Single Nucleotide Polymorphism

## 7.1 Introduction

High-throughput sequencing and genome-wide association (GWA) studies have identified both genetic hallmarks and risk factors for developing age-related diseases. How exactly genetic variation results in disease is often less well characterized. Such explanations are complex in nature and may involve perturbations in protein structure, disruption of interactions between biological molecules, or other environmental factors that affect expression levels and spatial concentration profiles. Other sequence-based methods can be used to provide further insight concerning how polymorphisms diverge from similar sequences. For example, genetic variation across orthologous protein sequences can be harnessed to identify coevolved pairs that are predictive of intramolecular contacts.[1,2]

Within the complement system, a part of innate immunity, such GWA studies have identified a missense mutation in the gene encoding factor H (FH) and the splice-variant factor-H-like protein 1 (FHL-1) that conveys risk for developing macular degeneration as an individual ages.[3] Specifically, this single nucleotide polymorphism (SNP) mutates tyrosine at position 402 to histidine (Y402H) and is associated with decreased binding affinity of particular FH and FHL-1 domains for heparin, C-reactive protein, and products of oxidative stress.[4–7] While the effect of age on the etiology of macular degeneration is

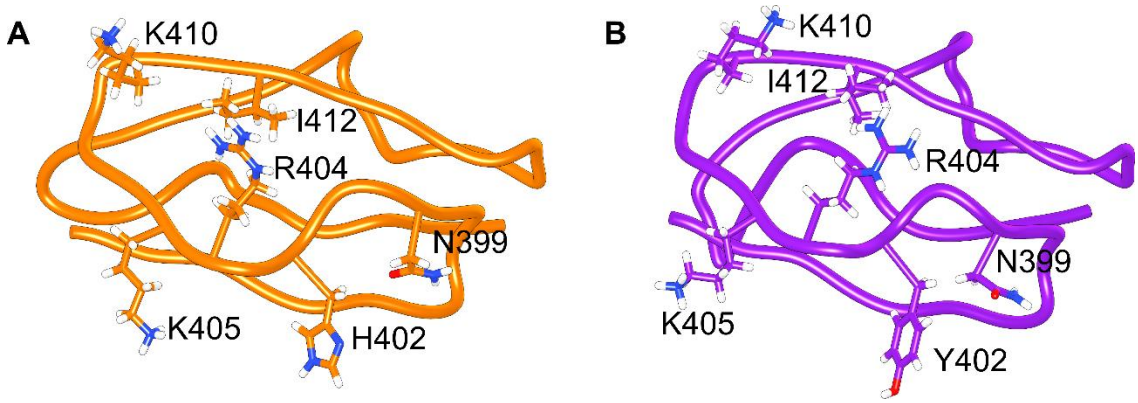



**Figure 7.1**: Structural representations of SCR7 are shown with key residues labeled. NMR HSQC experiments suggest chemical shift perturbations at labeled residues due to heparin binding. Aside from the Y402H polymorphism, structure is largely conserved between (A) $SCR7^{H402}$ from X-ray crystallography and (B) $SCR7^{Y402}$ from NMR spectroscopy. For all subsequent molecular graphics, we will display the same amino acid side chains and maintain a color scheme where $SCR7^{H402}$ is orange and $SCR7^{Y402}$ is purple.

still being studied, preliminary results have suggested that the extracellular matrix changes in terms of heparin density and composition with age.[8] Thus, the reduced affinity for heparin that is exhibited by the risk isoform, Y402H, may not result in disease until heparin density in the macula drops below a particular threshold or heparin composition changes significantly.

Structurally, 20 short consensus repeats (SCR) joined by 19 short linker sequences comprise FH. SCR1-4 interact with C3b and convertases to inhibit an immune response, SCR7 recognizes patterns in carbohydrates that are associated with surfaces of self, and SCR19-20 recognize both carbohydrate patterns of self and C3d, the thioester domain of C3 that is a hallmark of complement inactivation.[9–11] As a result, the pattern recognition properties of SCR7 and SCR19-20 direct the regulation of SCR1-4 towards surfaces of self, preventing an autoimmune response. The Y402H polymorphism occurs within SCR7, and has been shown to decrease the binding affinity of SCR6-8 for heparin though the SNP

in full-length FH does not significantly affect binding affinity.[4,5] As FHL-1 only contains the first seven SCR of FH, the regulator lacks the second recognition domain of SCR19-20. Thus, regulation by FHL-1 is more likely to be affected by the SNP. Within the field, it has been hypothesized that FHL-1 may impart more regulation in the macula than FH as the seven SCR of FHL-1 may diffuse more easily through Bruch's membrane than the 20 SCR of FH.[12] This effect may be more pronounced as Bruch's membrane thickens during early stages of AMD.[13]

Despite the known effects of the Y402H polymorphism on the association of SCR7 from FH and FHL-1 with heparin, no clear mechanism for how the SNP results in reduced affinity has been identified. Experimental studies have characterized the static structures of both SCR7 isoforms[4,14] and investigated interactions between $SCR7^{H402}$ and sucrose octasulfate.[14] However, all crystal structures from the Protein Databank are of $SCR7^{H402}$, and the nuclear magnetic resonance (NMR) structure of $SCR7^{Y402}$ does not contain any binding partners such as heparin. Indeed, no structural study of interactions between the non-risk isoform and heparin currently exists. Some speculations based on static structures hypothesize that Y402 sterically prohibits interactions with particular sulfation patterns in carbohydrates[14], but such hypotheses assume that the conformation of $SCR7^{Y402}$ in complex with heparin will adopt the solution structure that is mostly conserved for $SCR7^{Y402}$ and $SCR7^{H402}$. As protein sidechains are flexible and ligand binding often involves structural changes, such an assumption may not be warranted even though the solution structures of $SCR7^{Y402}$ and $SCR7^{H402}$ are similar (**Figure 7.1**).[4,5] Herein, we hypothesize that the Y402H mutation alters the preferred arrangement of key residues

known to be involved in heparin binding from heteronuclear single quantum coherence (HSQC) and mutagenesis experiments.[4,5,15] To assess the validity of this hypothesis, we leverage methods from bioinformatics and computational biophysics to assess the relevance and effects of the Y402H polymorphism. Our results show increased backbone fluctuations at position 402 and proximal residues as a result of disrupting a coevolved pair between Y402 and I412. Moreover, we describe differences in conformational sampling between SCR7 isoforms involving energy minima for sidechain positions 402 and 404 only exhibited by $SCR7^{Y402}$ where the molecule is able to associate with heparin significantly faster than $SCR7^{H402}$.

## 7.2 Results

### *7.2.1 Residue Y402 has coevolved with I412*

Leveraging large databases of protein sequences resulting from high-throughput sequencing methods, we constructed a multiple sequence alignment (MSA) for homologous SCR domains across 366 species ranging from viruses to humans. From these sequences, we computed co-evolutionary couplings between aligned columns using direct coupling analysis (DCA).[2] From this method pairs of positions in the MSA were considered strongly coupled if scores were greater than or equal to 0.7. Such couplings have been found previously to be strongly indicative of contacts in protein structure.[1,2] We found two strongly coupled positions with scores greater than 0.9 between positions in SCR7 separated by 3 or more residues. The coevolved pair with the highest score of 1.01 (p-

value: 4.0 x $10^{-8}$) occurs between residues V429 and P438 in SCR7. This predicted contact is observed to occur between two β-strands from structures of SCR7 deposited in the Protein Databank.[4,14] The second coevolved pair with a score of 0.91 (p-value: 8.8 x $10^{-7}$) occurs between residues Y402 and I412 in SCR7. Though these positions are spatially proximal in the structure of SCR7, no contacts are expected between the residues with a cutoff distance of 4 Å based on the NMR structure of $SCR7^{Y402}$.[4] Given the high coupling of Y402 with I412, we decided to investigate the conformational dynamics of both the $SCR7^{Y402}$ and $SCR7^{H402}$ isoforms to evaluate dynamical interactions between Y402 and I412 and to observe the structural effects of the Y402H mutation.

*7.2.2 The disruption of the coevolved pair destabilizes the structure of SCR7 proximal to position 402*

To simulate atomic motion in $SCR7^{Y402}$ and $SCR7^{H402}$ we performed multiple Molecular dynamics (MD) simulations for a total of 35 μs per isoform. Interestingly, $SCR7^{H402}$ exhibited an increase of approximately 1.5 Å in root-mean-square fluctuation (RMSF) of $C_\alpha$ carbons at residue number 402 compared to $SCR7^{Y402}$ (**Figure 7.2A**). We also observed increases in RMSF at positions 406-412, though smaller in magnitude with an average of 0.2 Å. SCR7 isoforms exhibited slightly different RMSFs for the centroids of residue sidechains (**Figure 7.2B**). Specifically, RMSFs were approximately 2 Å and 1.2 Å higher in $SCR7^{H402}$ than in $SCR7^{Y402}$ for N401 and [Y/H]402, respectively. These observations suggest that the Y402H polymorphism increases the flexibility of $SCR7^{H402}$ at residues including position 402 and proximal amino acids.

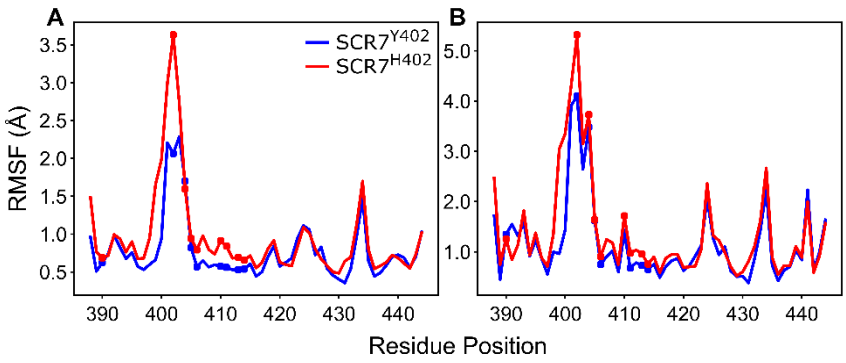



**Figure 7.2**: RMSF in (A) $C_\alpha$ positions and (B) sidechain centroids in SCR7$^{Y402}$ (blue) and SCR7$^{H402}$ (red) are shown for each residue position in SCR7. Circles are placed in RMSF plots to denote the location of residues Y390, [Y/H]402, R404, K405, F406, K410, D413, V414. Greater differences in RMSF between SCR7 isoforms are observed in $C_\alpha$ positions than in sidechain centroids. Additionally, the Y402H polymorphism is associated with increased RMSF for residues proximal to position 402.

### 7.2.3 *Conformational dynamics arising from slowest motions in SCR7$^{Y402}$ are more modular than in SCR7$^{H402}$*

To determine how increased fluctuations in SCR7$^{H402}$ result in conformational differences between SCR7$^{H402}$ and SCR7$^{Y402}$, we constructed Bayesian Markov models to describe conformational dynamics in each isoform.[16] These conformations are based on time-series of φ, ψ, $\chi_1$, and $\chi_2$ torsion angles that describe protein structure throughout each molecular dynamics simulation. Time-lagged independent component analysis[17] was used to decompose the SCR7$^{Y402}$ and SCR7$^{H402}$ torsional feature sets into projections along the five slowest eigenvectors for each isoform. As a result, each Markov model describes the five slowest, independent motions for the associated SCR7 isoform. Interestingly, by constructing a network from the probability transition matrix of each Markov model, we noticed a striking difference in modularity[18] between SCR7 isoforms. In SCR7$^{Y402}$ members of a community are more connected to members of the same community than members of other communities with a modularity of 0.745. SCR7$^{H402}$, however, has a lower

modularity of 0.660. This observation suggests that members of an individual community are more connected to members of another community in SCR7$^{H402}$ than in SCR7$^{Y402}$. This difference in modularity is clearly visible in **Figures 7.3-7.4** as a result of the force-directed layout of the graphs.[19]

After extracting representative structures from Markov chain states, we compared predicted $C_\alpha$ and $C_\beta$ chemical shifts to existing experimental data. Generally, observed chemical shifts from HSQC experiments deposited on the Biological Magnetic Resonance Data Bank (BMRB) fell within the 95% confidence intervals for the mean predicted chemical shifts, though we observed a few discrepancies. When our predictions for representative structures diverged from observed values, we typically observed that reference structures, the NMR and crystal structures for SCR7 isoforms from the Protein Databank (PDB), also diverged. In fact, our MD structures seemed to correct regions with unfavorable energy in the top NMR structure for SCR7$^{Y402}$, bringing predicted shifts closer to observed chemical shifts. Otherwise, reference predictions typically fell within 95% confidence intervals for mean shifts from representative structures from our Markov chains. To quantify how well representative structures agree with experimental evidence, we calculated root-mean-square deviation (RMSD) from observed chemical shifts for our representative structures and reference structures from the PDB (**Figure B.1-2**). For the $C_\alpha$ chemical shifts from our models, we expect RMSDs of 1.31 and 1.94 ppm for SCR7$^{Y402}$ and SCR7$^{H402}$, respectively. These values can be compared to reference RMSDs of 1.65 and 1.73 ppm for SCR7$^{Y402}$ and SCR7$^{H402}$, respectively. Similarly, we expect RMSDs of chemical shifts for $C_\beta$ of 1.51 and 2.19 ppm for SCR7$^{Y402}$ and SCR7$^{H402}$, respectively.

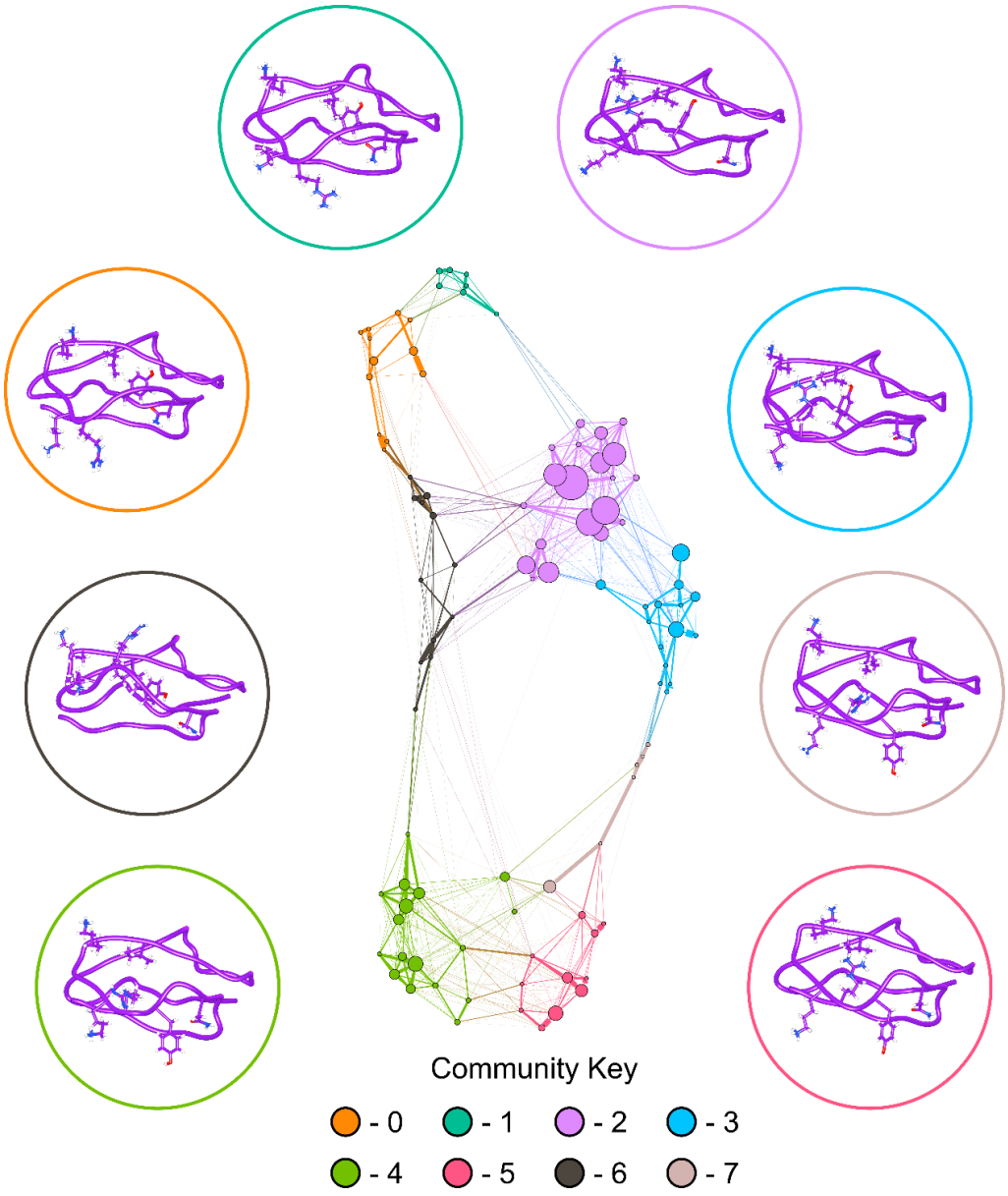



**Figure 7.3**: Graphical representation of conformational sampling of SCR7$^{Y402}$ is shown with representative structures for each community within the network. Nodes are conformational states from the discretized trajectories that are scaled in size according to the associated probability from the stationary distribution of the Markov chain. Edges represent transitions between conformational states and are scaled according to the associated probability from the Markov chain. Nodes are colored according to community membership, and a representative structure of SCR7$^{Y402}$ is shown for each community. Circles surrounding conformational states denote community membership of each structure, and amino acid side chains for residues N399, Y402, R404, K405, K410, and I412 are displayed. Interestingly, nodes within communities are more connected with other nodes in the same community than with nodes of other communities. This observation suggests a high degree of modularity within the network.

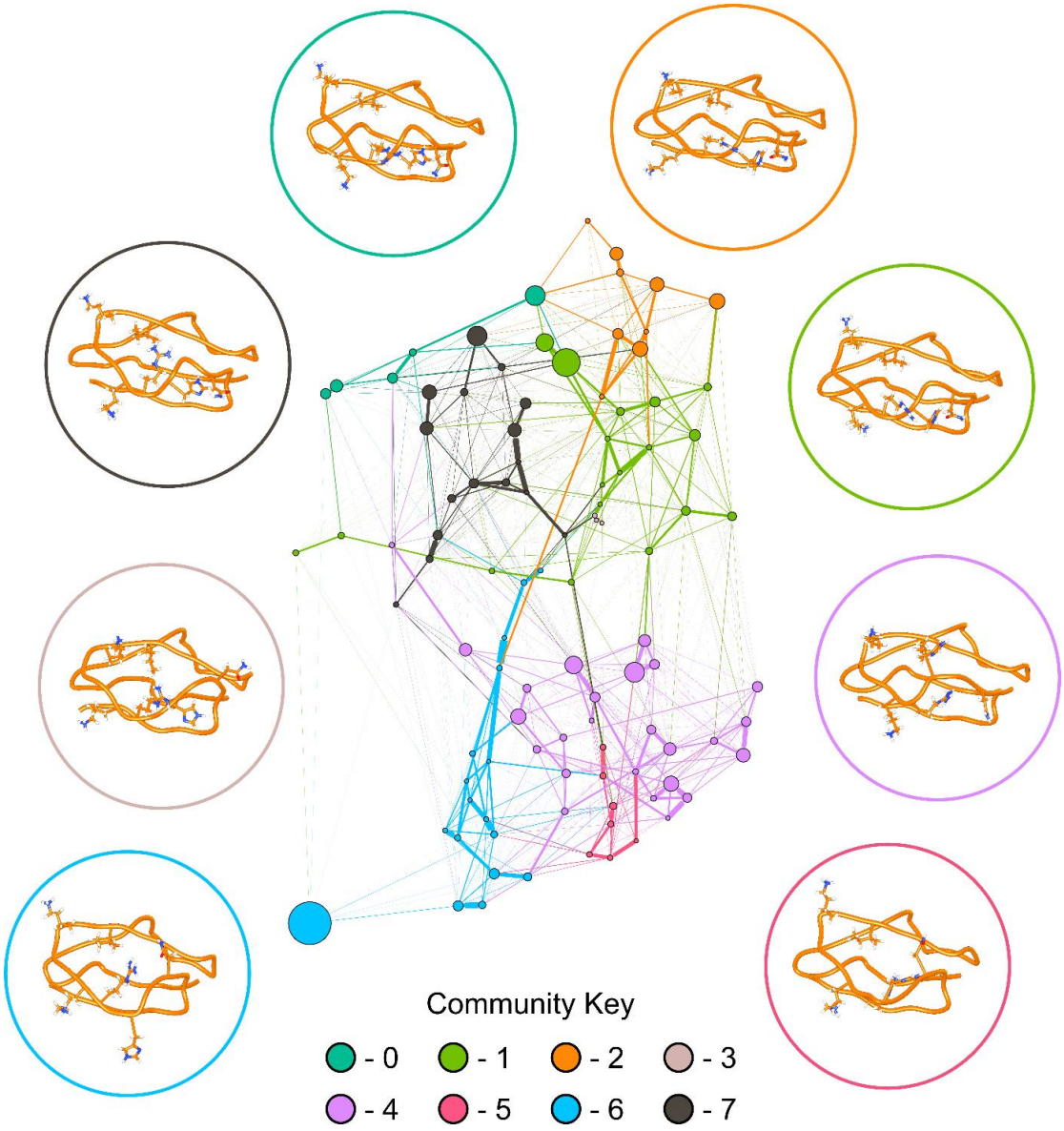



**Figure 7.4**: Graphical representation of conformational sampling of SCR7$^{H402}$ is shown with representative structures for each community within the network. Nodes are conformational states from the discretized trajectories that are scaled in size according to the associated probability from the stationary distribution of the Markov chain. Edges represent transitions between conformational states and are scaled according to the associated probability from the Markov chain. Nodes are colored according to community membership, and a representative structure of SCR7$^{Y402}$ is shown for each community. Circles surrounding conformational states denote community membership of each structure, and amino acid side chains for residues N399, H402, R404, K405, K410, and I412 are displayed. Compared to the network for SCR7$^{Y402}$ (**Figure 7.3**), the network for SCR7$^{H402}$ appears far less modular as there are more edges between nodes of different communities.

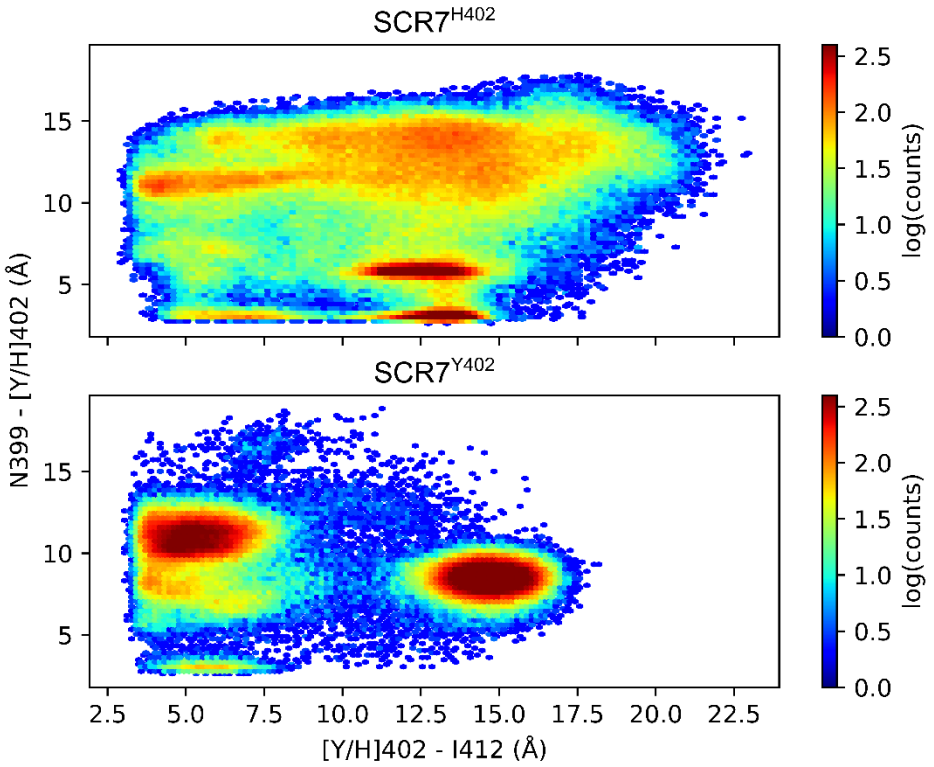



**Figure 7.5**: Free energy landscapes for distances between residue pairs [Y/H]402 – I412 and N399 – [Y/H]402 are displayed for each SCR7 isoform. The distance between [Y/H]402 and I412 is calculated from the positions of atom CE1 in Y402 or H402 and atom CD1 in I412, while the distance between N399 and [Y/H]402 is calculated from atom ND2 in N399 and atom ND1 or OH in H402 or Y402, respectively. Bins within the landscapes are colored according the logarithm of the number of conformations observed to fall within the associated bin. States with lower free energy are expected to be observed with higher frequency. Note how the free energy landscape for $SCR7^{Y402}$ is more constrained to two minima than $SCR7^{H402}$ that samples a larger area of the landscape. Also note how $SCR7^{Y402}$ is more likely to form a contact between Y402 and I412 with a cutoff distance of 4 Å than $SCR7^{H402}$, while $SCR7^{H402}$ is more likely to form a contact between N399 and H402 with a cutoff distance of 4 Å than $SCR7^{Y402}$.

Reference RMSDS for chemical shifts of $C_\beta$ are 1.82 and 1.93 for $SCR7^{Y402}$ and $SCR7^{H402}$, respectively.

### *7.2.4 Models for conformational dynamics suggest equilibrium differences between SCR7 isoforms in solution*

Network representations of conformational sampling in both SCR7 isoforms highlight structural differences. In $SCR7^{Y402}$ communities 0-3 and 6, Y402 prefers to interact with

I412, a residue with a side chain that is more buried in the structure of SCR7 (**Figure 7.3**). Communities 4-5 and 7, however, have Y402 in a conformation that is more solvent-exposed. Thus, Y402 switches between a conformation observed in the NMR structure and a contact predicted by a coevolved pair. In $SCR7^{H402}$ H402 appears to interact with I412 less frequently, preferring to interact with N399 or adopt a solvent exposed conformation (**Figure 7.4-7.5**). According to the Markov models for each isoform, the expected distance between position 402 and I412 is 4.0 Å in $SCR7^{Y402}$ and 7.0 Å in $SCR7^{H402}$. With a cutoff distance of 4 Å, the corresponding contact probabilities are 69.7% and 14.0% for $SCR7^{Y402}$ and $SCR7^{H402}$, respectively.

More broadly, the Markov chains for conformational sampling in each SCR7 isoform suggest expected differences between torsion angles. As shown in **Figure 7.6**, the largest differences involve changes in backbone dihedrals around A425, K424, and L422, though it is unclear how relevant these differences are for etiology of AMD. Other primary differences in backbone structure appear to occur at N401, N399, G403, Q400, and R404. Structural changes here could contribute to differences in binding affinity for heparin. Prior HSQC NMR experiments have identified a number of chemical shift perturbations that result when SCR7 binds heparin. These residues include Y390, Y402, R404, K405, F406, S411, I412, D413, and V414.[4] Other mutagenesis studies suggest the importance of R404, K405, and K410 for binding heparin.[5,15] Thus, disruptions to structure between residues N399 and V414 could disrupt the topology of interactions with heparin, leading to a lower binding affinity.

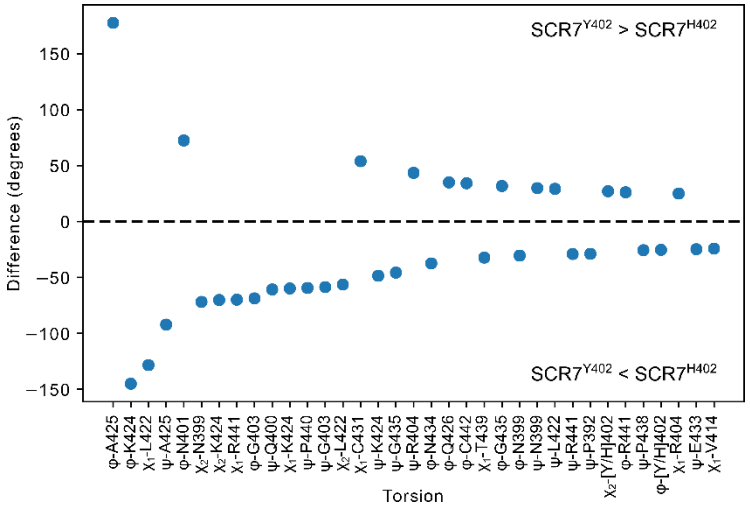



**Figure 7.6**: Expected differences of torsion angles between Markov chains for SCR7 isoforms are displayed. Angles larger in SCR7$^{Y402}$ are located above the dashed line, while angles larger in SCR7$^{H402}$ are located below the dashed line. Torsion types and residue membership are labeled along the x-axis in order of decreasing magnitude difference between isoforms. Expected torsions appear to be different between SCR7 isoforms at residues L422-A425, though these differences result in small sidechain rearrangements. Expected torsions also differ at residues N399-R404, resulting in more appreciable differences in sidechain arrangements.

### *7.2.5 SCR7 isoforms associate with heparin at different rates despite similar electrostatic potentials*

In order to evaluate how structural changes may affect association of SCR7 with heparin (**Figure 7.7**), we performed Brownian dynamics simulations with representative structures from each state within the Markov models for both SCR7$^{Y402}$ and SCR7$^{H402}$. SCR7$^{Y402}$ was found to associate with an expected rate constant of 6.89 (± 0.06) x $10^9$ while SCR7$^{H402}$ was found to associate with an expected rate constant of 6.75 (± 0.06) x $10^9$. Though these rates are of the same order of magnitude, SCR7$^{Y402}$ was found to associate significantly faster than SCR7$^{H402}$ (p-value < 0.001). Given the small magnitude difference between SCR7 isoforms and the similar electrostatic grid potentials (**Figure B.3**),

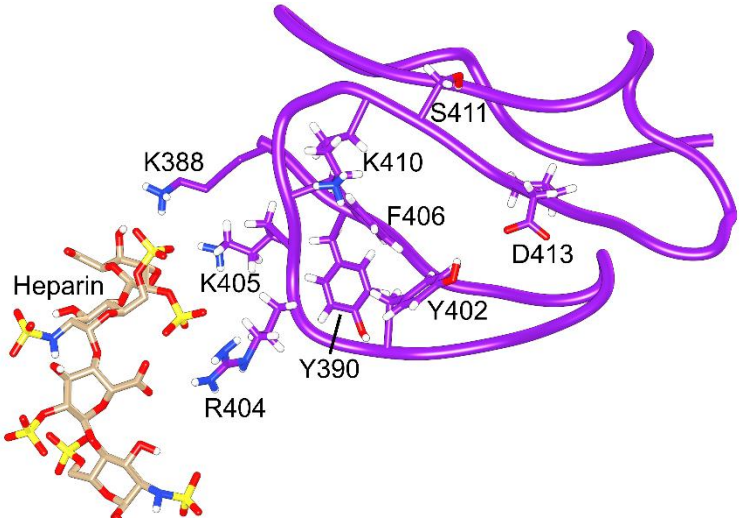



**Figure 7.7**: Example encounter complex between SCR7$^{Y402}$ and heparin (dp12) from a Brownian dynamics trajectory where binding occurs. All residues on SCR7 except K388 have been implicated in binding heparin by HSQC NMR and mutagenesis experiments.

topological differences in hydrogen bond donors and acceptors on the surface of SCR7 are likely the primary cause of differences in association rates.

### *7.2.6 Energy minima for positions of [Y/H]402 and R404 side chains explain differences in association rates*

Given the differences in association rate constants between SCR7 isoforms, we investigated how side chain conformations contribute to these differences. For each key residue identified from prior NMR experimental studies (Y390, Y402, R404, K405, F406, K410, S411, D413, V414), we calculated the first two principal components for the coordinates of a representative atom. In this way we constructed a free energy landscape for the position vector of each amino acid side chain in both SCR7 isoforms. Next, we annotated these landscapes with average association rates for nearby representative models from Markov chains for both SCR7 isoforms. The result can be viewed in **Figure 7.8** where

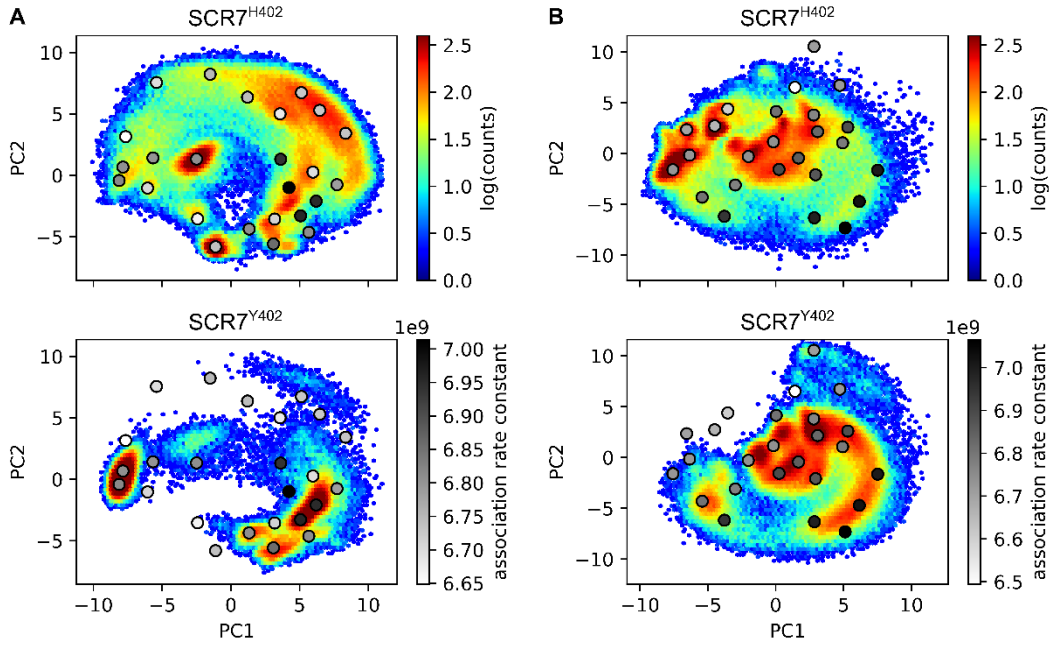



**Figure 7.8**: Free energy landscapes for a two-dimensional representation of sidechain orientations for residues (A) [Y/H]402 and (B) R404 are shown for SCR7 isoforms, where a larger value for the logarithm of the number of observations in a bin indicates a lower free energy. To describe the orientation of sidechains, position vectors are found from coordinates of atom NE2 in H402, OH in Y402, and CZ in R404 and decomposed into two dimensions with principal component analysis. Regions of the landscape are annotated with circles shaded according to the average association rate-constant observed for $SCR7^{Y402}$ and $SCR7^{H402}$ in the neighboring area. After analysis of variance, mean rate-constants for at least two regions are significantly different for both residues [Y/H]402 and R404 (p-values < 0.001). For orientations of both [Y/H]402 and R404, regions with high average association rate constants are more closely associated with energy minima of $SCR7^{Y402}$.

we show that $SCR7^{Y402}$ and $SCR7^{H402}$ prefer different sidechain conformations for residues [Y/H]402 and R404. Moreover, regions of the landscape with high association rates are more probable for $SCR7^{Y402}$ than $SCR7^{H402}$. Free energy landscapes for sidechain conformations of other key residues do not vary between SCR7 isoforms (**Figure B.4**).

### 7.3 Discussion

Despite structural similarities between SCR7$^{Y402}$ and SCR7$^{H402}$, the Y402H mutation in FH SCR7 is known to reduce the binding affinity of FH SCR6-8 for heparin. As Y402 appears to have coevolved with I412, the hydrophobic contribution of Y402 likely imparts stability to the conformation of SCR7. Even within the predicted ensemble of structures determined from NMR experiments for SCR7$^{Y402}$, there exists a structure where the Y402 sidechain is oriented towards the core of the protein instead of towards the solvent (though no contact is observed with I412 in particular). In the case of the Y402H polymorphism, H402 rarely interacts with I412. Instead, H402 samples a large range of conformations. One very probable conformation involves contacts (including hydrogen bonding) between H402 and N399. Thus, it is possible that the additional number of hydrogen bond donors and acceptors and the decreased hydrophobicity of H402 destabilize the heparin binding site of SCR7.

While our simulations suggest that residues fluctuate more near position 402 in SCR7$^{H402}$ than in SCR7$^{Y402}$, the conformational space between SCR7 isoforms overlaps. In fact, tertiary structure of SCR7 is largely conserved between isoforms. The two conserved disulfide bridges within SCRs cause any deviations in terms of tertiary structure to be unlikely. Side chain fluctuations and rearrangements in flexible loops, however, are still possible and may contribute to differences in binding affinity. As chemical shifts for SCR7$^{Y402}$ and SCR7$^{H402}$ are deposited on the BMRB, we compared predicted chemical shifts for our representative conformations from MD simulations to predicted chemical shifts for solved structures of SCR7 in the PDB. Representative structures from the Markov

model for SCR7$^{Y402}$ exhibited lower RMSD from experimentally determined chemical shifts than the NMR structure of SCR7$^{Y402}$ (**Figure B.1-2**). SCR7$^{H402}$, however, exhibited slightly higher RMSD from experimentally determined chemical shifts than the crystal structure of SCR7$^{H402}$ though predicted shifts of each residue for the crystal structure generally fell within the 95% confidence interval for predicted chemical shifts. Thus, MD simulations and the associated Markov chains appear to agree with experimental chemical shifts, even correcting regions with unfavorable energies from the NMR structure of SCR7$^{Y402}$.

Though the structure containing SCR7$^{H402}$ in complex with sucrose octasulfate suggests favorable contacts between H402 and sulfates[14], no structure for interactions between SCR7$^{Y402}$ and similar carbohydrates exists. Moreover, sucrose octasulfate contains more sulfates than does heparin for molecules with equal degree of polymerization. As a result, many hypotheses that attempt to explain the reduced binding affinity of the Y402H polymorphism are based on the structure of the risk isoform SCR7$^{H402}$. From experiments, however, there is a wealth of information on the key residues that mediate interactions of SCR7 with heparin. HSQC spectra suggest the involvement of Y390, Y402, R404, K405, F406, S411, D413, and V414 in heparin binding[4] while mutagenesis suggests the relevance of R404, K405, and K410.[5,15] By informing BD simulations with potential interacting residues and an ensemble of representative SCR7 conformations from MD, we were able to observe a lower heparin association rate constant for SCR7$^{H402}$ than for SCR7$^{Y402}$. Moreover, we were able to observe associations between side chain positions and heparin association rate constants that reveal a mechanism for the

decreased pattern recognition resulting from the Y402H SNP (**Figure 7.8**). Specifically, we observed that differences in association rates between SCR7 isoforms are best explained by side chain positions of [Y/H]402 and R404. Transient formation of contacts between Y402 and I412 appears to stabilize SCR7 in a conformation that associates with heparin at the highest association rates observed.

Disruption of the Y402-I412 coevolved pair in the risk-isoform $SCR^{H402}$ is observed to change the dynamical structure of SCR7. For example, the network of transitions between conformational states decreases in modularity. Moreover, the decrease in modularity is associated with an increase in square fluctuations proximal to position 402. As a result, sidechain arrangements of H402 and R404 sample a broader conformational space compared to $SCR7^{Y402}$. Though overlap in probable sidechain orientations at positions 402 and 404 exists between $SCR7^{Y402}$ and $SCR7^{H402}$, not all energy minima are shared. At position 402, the energy minima for sidechain orientation are distinct, and the preferred position of the Y402 sidechain is more constrained than H402. At position 404, both isoforms share a central energy minimum that corresponds to reference structures in the PDB (**Figure B.5**); however, $SCR7^{Y402}$ and $SCR7^{H402}$ have distinct secondary minima. Minima for $SCR7^{Y402}$ are associated with higher average rate-constants for association with heparin, while minima for $SCR7^{H402}$ are associated with lower average rate-constants. Thus, dynamical behavior of FH SCR7 isoforms provide a mechanism for the decreased binding affinity that results from the Y402H SNP. This mechanism is not apparent from the SCR7 structures deposited on the PDB but agrees with experimental chemical shifts from NMR experiments.

As protein structures from NMR experiments are models estimated from structural restraints that are heavily dependent on interproton distances, it is possible that structural models from NMR may lose important structural features, including contacts between coevolved pairs, depending on the observation of interproton contacts related to efficiency of magnetization transfer and spectral overlap. MD, however, can be used to identify dynamical changes in protein structure and extract new representative structures from energy minima. A Markovian approach characterizing conformational transitions of $SCR7^{Y402}$ was able to generate representative structures that are closer to expected conformational shifts than the corresponding NMR structure. These changes involve the transient formation of contacts between Y402 and I412. In comparison to the crystal structure of $SCR7^{H402}$, $SCR7^{Y402}$ residues Y402 and R404 are observed to sample different energy minima that correspond to faster association rates with heparin. $SCR7^{H402}$, on the other hand, samples conformations that are similar to the crystal structure, rarely sampling conformations in regions of the free energy landscape that correspond to fast association with heparin. Guided by existing experimental data and principles from bioinformatics, MD and BD can describe behavior in SCR7 that is difficult to achieve with experimental methods, resulting in a molecular mechanism for decreased heparin-binding affinity associated with the Y402H polymorphism.

## 7.4 Conclusions

Side chains of [Y/H]402 and R404 prefer different orientations in $SCR7^{Y402}$ and $SCR7^{Y402}$ leading to topological differences that contribute to a decreased rate constant for association with heparin. We also observe that $SCR7^{Y402}$ is more likely to preserve the coevolved contact between [Y/H]402 and I412 than $SCR7^{H402}$. Thus, the Y402H polymorphism appears to destabilize a portion of the SCR7 surface responsible for recognition of heparin. Likely, the increased number of possible dipole-dipole interactions and the decreased hydrophobicity of H402 compared to Y402 result in increased fluctuations at H402 and proximal residues in $SCR7^{H402}$. These fluctuations manifest in conformational sampling that is less modular and an energy landscape for the side chain orientation of H402 that is less constrained than Y402 from $SCR7^{Y402}$.

While conventional hypotheses based on interactions between H402 and sucrose octasulfate from $SCR7^{H402}$ state that Y402 may sterically prevent association with some carbohydrates that are not sterically hindered by H402, our data suggests a revised mechanism for decreased heparin binding. Specifically, the transient contacts between Y402 and I412 in $SCR7^{Y402}$ stabilize the molecule in a conformation that promotes association with heparin. This observation is in agreement with experiments that show SCR6-$8^{H402}$ eluting from a heparin column faster than SCR6-$8^{Y402}$.[4] With increasing age, both carbohydrate density and degree of sulfation have been shown to decrease in Bruch's membrane of the macula.[8] Additionally, the $FH^{Y402}$ has been shown to bind a wide range of sulfated carbohydrates, while $FH^{H402}$ has been shown to have more specificity and to require a high degree of sulfation in carbohydrates.[8] The two energy minima for the

orientation of Y402 in SCR7$^{Y402}$ likely impart robust recognition of a variety of sulfated carbohydrates. While one energy minimum of SCR7$^{H402}$ overlaps with SCR7$^{Y02}$, the orientation of H402 is less spatially constrained than Y402, and the unique minimum is not associated with a high rate constant for association with heparin.

To investigate structural effects of missense single nucleotide polymorphisms, it is not always possible to acquire crystal structures for pertinent isoforms. In the absence of complete structural information, our approach suggests that methods from bioinformatics and computational biophysics can be leveraged to explore these structural effects and provide structural insights from timescales that are not accessible from static NMR or crystallographic structures. Notably, discretization of conformations from molecular dynamics using independent component analysis appears to extract representative structures that are predicted to agree with experimental chemical shifts.

## 7.5 Methods

### *7.5.1 Direct coupling analysis*

The sequence of FH SCR7 was submitted to the `jackhmmer`[20] webserver with default parameters and subject to 3 iterations. The resulting 44,511 sequences from the `uniprotrefprot` database were passed through CD-HIT[21,22] with a clustering threshold of 95%, a word size of 5, a 90% alignment coverage relative to the longer sequence, and a length difference cutoff of 90%. MAFFT[23] was used to align the 22,825 sequences with the `dpparttree` method, 1000 parts, the BLOSUM62 matrix[24], a gap

opening penalty of 1.53, and a gap extension penalty of 0.123. The MSA was filtered such that only records with non-gap characters at positions corresponding to the four conserved cysteines forming two disulfide bridges, the conserved tryptophan, and Y402 were retained. Residues before the first cysteine of the first disulfide bridge and after the last cysteine of the last disulfide bridge were trimmed from the MSA, and columns with greater than 5% gap characters were removed. The final MSA of 4,940 sequences (including the record for FH SCR7) was passed through the DCA algorithm[2] with a reweighting factor of 90 and regularization factor of 0.01.

*7.5.2 Molecular dynamics*

To construct simulation systems, structures for the FH SCR7$^{Y402}$ and SCR7$^{H402}$ isoforms were retrieved from the Protein Databank with identifiers 2jgx and 2uwn, respectively. PDBFixer[25] was used to add missing heavy atoms to 2uwn and add hydrogens to 2jgx and 2uwn for the most common protonation state at pH 7.4. Structures were trimmed to contain only residues 388 through 444. PROPKA[26,27] was used to verify that protonation states are appropriate for pH 7.4, and we observed no differences in predicted protonation states. For FH SCR7$^{H402}$ in particular, all simulations used a neutrally charged H402 with hydrogen placed on atom $N_{\varepsilon 2}$. Each isoform was then solvated in a water box with a 10 Å buffer region around protein in the positive and negative x, y, and z directions. Water molecules were simulated with the TIP3P model, and sodium and chloride ions were added to each system such that the final ionic strength was 150 mM and the net charge of the system was neutral. For all molecular dynamics simulations, the CHARMM36 forcefield[28] was used.

Each simulation system was subject to steepest-descent minimization in Gromacs[29] with CHARMM36-compatible parameters. Specifically, these parameters require the Verlet cutoff scheme with cutoff distances of 12 Å for van der Waals and Coulombic interactions and a switching distance of 10 Å for van der Waals interactions. Particle mesh Ewald electrostatics (interpolation order of 4, Fourier spacing of 1.6 Å) was enabled while dispersion correction was disabled. These parameters are shared with all subsequent molecular dynamics simulations. For minimization in particular, an energy minimization step size of 0.1 Å was used, and minimization was terminated when the max force was less than 100 kJ/mol/Å.

After minimization, each system was equilibrated with short, 100 ps NVT and NPT simulations. NVT equilibration was performed first with a leap-frog integrator and 2 fs time step where initial velocities were randomly assigned from a Maxwell distribution. In this simulation heavy atoms were restrained and all bonds constrained with the LINCS algorithm (order of four, one iteration). A Berendsen thermostat was used to maintain a system temperature of 300 K with a time constant of 0.1 ps. Next, NPT equilibration was performed with the same integrator and time step, though velocities were preserved from the previous simulation. Similarly, this simulation restrained heavy atoms, constrained bonds with LINCS, and controlled temperature with the Berendsen thermostat. In addition, pressure was controlled with the Parrinello-Rahman barostat using a time constant of 2 ps, a reference pressure of 1 bar, and water compressibility of $4.5 \times 10^{-5}$ $bar^{-1}$, and isotropic coupling.

Following equilibration, a primary simulation of 10 μs and five secondary simulations of 5 μs were performed for each isoform. Thus, each isoform was simulated for a total of 35 μs. Primary simulations were continued from equilibration runs, while secondary simulations branched out from the 5 most infrequent state transitions observed in the primary simulations. Secondary simulations were initialized with random velocities drawn from a Maxwell distribution. For all primary and secondary molecular dynamics simulations, a leap-frog integrator with a 2 fs time step was used. Temperature and pressure were controlled with the Berendsen thermostat and Parrinello-Rahman barostat as in NPT equilibration, respectively. No restraints were imposed on any atoms, and only bonds between hydrogens and heavy atoms were constrained by the LINCS algorithm.

*7.5.3 Markov chain models for conformational dynamics*

Markov chain models that describe conformational sampling of FH SCR7$^{Y402}$ and SCR7$^{H402}$ were based on all data from primary and secondary molecular dynamics trajectories. Models for each isoform were built separately such that the state space is not shared between SCR7 isoforms; however, we aimed to keep parameters as similar as possible for each isoform to facilitate comparisons. Specifically, Cartesian coordinates for all atoms after every 100 ps were converted to a features set comprising sine and cosine components for the φ, ψ, $\chi_1$, and $\chi_2$ torsion angles. These features were decomposed with time-lagged independent component analysis using a lag time of 100 ps. This lag time was selected as it was observed to maximize the generalized matrix Rayleigh quotient implemented in MSMBuilder.[30] Subsequently, the first 5 independent components were clustered into 100 discrete conformational states using a mini-batch K-means algorithm

implemented in MSMBuilder.[30] A Bayesian Markov chain, implemented in PyEMMA[31], was built from these discretized trajectories using statistically uncorrelated state transitions and a lag time of 20 ns. This lag time was selected as it was the smallest value where relaxation timescales for the Bayesian Markov chain were not dependent on the selected lag time (**Figure B.6**). Markov chains were validated with Chapman-Kolmogorov tests (**Figure B.7**). Networks were constructed from the transition probability matrices of each Markov chain and visualized in Gephi[32] with a force-directed layout.[19] Network modularity was calculated in Gephi with a resolution of 1.0 and the randomize and edge weight parameters enabled.[18] Representative structures for each of the 100 states were identified by finding the conformation from the MD trajectories for the particular SCR7 isoform with the minimum distance to the cluster center from mini-batch K-means algorithm.

*7.5.4 Brownian dynamics*

Association of heparin to FH SCR7$^{Y402}$ and SCR7$^{H402}$ was investigated with Brownian dynamics (BD) simulations. For this method, we retrieved the structure of heparin (dp12) from the Protein Databank with identifier 1hpn (model 1).[33] 100 representative structures for each FH SCR7 isoform were extracted from the corresponding Markov chain as described previously. 10,000 BD trajectories were acquired for heparin associating with each representative conformation of each FH SCR7 isoform using BrownDye. [34] For every trajectory, binding was considered to occur if at least 3 separate hydrogen bond donor-acceptor pairs between heparin and SCR7 residues 390, 402, 404, 405, 406, 410, 411, 413, or 414 were separated by 4 Å or less during the simulation. No binding was considered to occur if a molecule did not bind within 1 x $10^6$ steps. PDB2PQR[35,36] was used to assign

partial charges and van der Waals radii from the CHARMM forcefield. Glycan reader[37] was used to convert heparin residues and atoms to CHARMM-compatible nomenclature, then partial charges and van der Waals radii were assigned manually from the CHARMM forcefield. APBS[38] was used to estimate a grid of electrostatic potentials for heparin and each SCR7 conformer using a solvent dielectric coefficient of 78.54, a protein dielectric coefficient of 20.0[39], a solvent with sodium and chloride ions for 150 mM ionic strength, and a Debye length of 8.08 Å in solvent. The average minimum distance moved during a time step was scaled by 0.2 using the `mimimum-dx` parameter of BrownDye.[34] Additionally, each simulation included desolvation forces and hydrodynamic interactions.

*7.5.5 Electrostatic similarity*

Electrostatic similarity for the 100 conformations of each FH SCR7 isoform used in BD simulations were compared using established methods in the AESOP python library.[40] All SCR7 conformers were superposed on backbone $C_\alpha$, and PDB2PQR[35,36] was used to assign partial charges and van der Waals radii to each atom at a pH of 7.4. APBS[38] was used to generate a grid of electrostatic potentials for each SCR7 conformer using a solvent dielectric coefficient of 78.54, a protein dielectric coefficient of 20.0[39], a 150 mM NaCl solution, and a grid resolution of 1 Å. As all grids shared common coordinates for vertices, an electrostatic similarity distance (ESD) was calculated pairwise for every pair of SCR7 conformers according to the standard method in AESOP. An ESD of 0 indicates electrostatic potentials are identical. An ESD value of 2 indicates electrostatic potentials that are equal in magnitude but opposite in polarity. Hierarchical clustering was performed on ESD values to generate a dendrogram comparing the electrostatic similarity of $SCR7^{Y402}$

and SCR7$^{H402}$ conformers. As most representative structures are within an ESD of 0.75, structures seem to have very similar electrostatic properties.

*7.5.6 Validation of representative SCR7 conformations*

Chemical shifts for $C_\alpha$ and $C_\beta$ atoms were predicted for all 100 representative states of each SCR7 isoform using the SPARTA+ method.[41] Expected, per-residue, chemical shifts were calculated according to the expectation operator

$$E[X] = \sum_{i=1}^{k} p_i x_i \quad (\mathbf{7.1})$$

where $p_i$ is the probability of state $i$ from the stationary distribution of the Markov chain for the particular SCR7 isoform, $k$ is the number of states (each isoform has 100 states), and $x_i$ is the observable for state $i$. RMSD from experimental chemical shifts is calculated according to

$$RMSD = \sqrt{\sum_{j=1}^{n} \frac{E[(\delta_{j,predicted} - \delta_{j,observed})^2]}{n}} \quad (\mathbf{7.2})$$

where $j$ denotes the residue index, $n$ is the total number of residues for which there exists predicted and experimental chemical shifts for the atom of interest, and $\delta_j$ is a chemical shift (predicted or observed as indicated by subscript) for residue $j$. Predicted chemical shifts are values returned from SPARTA+ for either conformations from our MD simulations or reference structures from the PDB, and experimental chemical shifts are values deposited on the BMRB for SCR7$^{Y402}$ or SCR7$^{H402}$. The expectation for the squared

deviations of chemical shift predictions from experimental values in **Equation 7.2** is calculated according to **Equation 7.1** for representative conformers of SCR7. For comparisons between MD and reference structures deposited in the PDB, the expectation operator simplifies to the mean shift for residue $j$. Note that RMSD values are not comparable across different atom types since chemical shifts for different atom types vary in magnitude.

### *7.5.7 Custom analysis scripts*

Most analyses were performed with custom Python scripts leveraging existing libraries such as MDTraj[42], PDBFixer[25], MSMBuilder[30], PyEMMA[31], SciPy[43], NumPy[44], Biopython[45], AESOP[39,40], pandas[46], NetworkX[47], and matplotlib.[48] Some shell scripts were also used to carry out coevolution methods, calling tools such as CD-HIT[21,22], MAFFT[23], and the DCA algorithm[2] kindly provided by the Valencia group.

## 7.6 References


1. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14,** 249–61 (2013).

2. Sutto, L., Marsili, S., Valencia, A. & Gervasio, F. L. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc. Natl. Acad. Sci.* **112,** 13567–13572 (2015).

3. K.Liszewski, M. & Atkinson, J. P. Complement regulators in human disease: Lessons from modern genetics. *J. Intern. Med.* **277,** 294–305 (2015).

4. Herbert, A. P. *et al.* Structure shows that a glycosaminoglycan and protein recognition site in factor H is perturbed by age-related macular degeneration-linked single nucleotide polymorphism. *J. Biol. Chem.* **282,** 18960–18968 (2007).

5. Clark, S. J. *et al.* His-384 allotypic variant of factor H associated with age-related macular degeneration has different heparin binding properties from the non-disease-associated form. *J. Biol. Chem.* **281,** 24713–24720 (2006).

6. Weismann, D. *et al.* Complement factor H binds malondialdehyde epitopes and protects from oxidative stress. *Nature* **478,** 76–81 (2011).

7. Sjöberg, A. P. *et al.* The factor H variant associated with age-related macular degeneration (His-384) and the non-disease-associated form bind differentially to C-reactive protein, fibromodulin, DNA, and necrotic cells. *J. Biol. Chem.* **282,** 10894–10900 (2007).

8. Keenan, T. D. L. *et al.* Age-dependent changes in heparan sulfate in human Bruch's membrane: Implications for age-related macular degeneration. *Investig. Ophthalmol. Vis. Sci.* **55,** 5370–5379 (2014).

9. Kieslich, C. A., Vazquez, H., Goodman, G. N., López de Victoria, A. & Morikis, D. The effect of electrostatics on factor H function and related pathologies. *J. Mol. Graph. Model.* **29,** 1047–1055 (2011).

10. Makou, E., Herbert, A. P. & Barlow, P. N. Functional anatomy of complement factor H. *Biochemistry* **52,** 3949–3962 (2013).

11. Harrison, R. E. S., Gorham, R. D. & Morikis, D. Energetic evaluation of binding modes in the C3d and Factor H (CCP 19-20) complex. *Protein Sci.* **24,** 789–802 (2015).

12. Schmidt, C. Q., Lambris, J. D. & Ricklin, D. Protection of host cells by complement regulators. *Immunol. Rev.* **274,** 152–171 (2016).

13. Ambati, J., Atkinson, J. P. & Gelfand, B. D. Immunology of age-related macular degeneration. *Nat. Rev. Immunol.* **13,** 438–451 (2013).

14. Prosser, B. E. *et al.* Structural basis for complement factor H–linked age-related macular degeneration. *J. Exp. Med.* **204,** 2277–2283 (2007).

15. Giannakis, E. *et al.* A common site within factor H SCR 7 responsible for binding heparin, C-reactive protein and streptococcal M protein. *Eur. J. Immunol.* **33,** 962–969 (2003).

16. Trendelkamp-Schroer, B., Wu, H., Paul, F. & Noé, F. Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.* **143,** (2015).

17. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139,** (2013).

18. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008,** (2008).

19. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9,** 1–12 (2014).

20. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11,** (2010).

21. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).

22. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).

23. Yamada, K. D., Tomii, K. & Katoh, K. Application of the MAFFT sequence alignment program to large data - Reexamination of the usefulness of chained guide trees. *Bioinformatics* **32,** 3246–3251 (2016).

24. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89,** 10915–10919 (1992).

25. Eastman, P. PDBFixer. (2016).

26. Søndergaard, C. R., Olsson, M. H. M. & Rostkowski Michałand Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J Chem Theory Comput* **7,** 2284–2295 (2011).

27. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p<italic>K</italic>$_a$ Predictions BT - Journal of Chemical Theory and Computation. *J. Chem. Theory Comput.* **7,** 525–537 (2011).

28. Best, R. B. *et al.* Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J. Chem. Theory Comput.* **8,** 3257–3273 (2012).

29. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2,** 19–25 (2015).

30. Harrigan, M. P. *et al.* MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **112,** 10–15 (2017).

31. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11,** 5525–5542 (2015).

32. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int. AAAI Conf. Weblogs Soc. Media* 361–362 (2009). doi:10.1136/qshc.2004.010033

33. Khan, S., Gor, J., Mulloy, B. & Perkins, S. J. Semi-Rigid Solution Structures of Heparin by Constrained X-ray Scattering Modelling: New Insight into Heparin-Protein Complexes. *J. Mol. Biol.* **395,** 504–521 (2010).

34. Huber, G. A. & McCammon, J. A. Browndye: A software package for Brownian dynamics. *Comput. Phys. Commun.* **181,** 1896–1905 (2010).

35. Dolinsky, T. J. *et al.* PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35,** 522–525 (2007).

36. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32,** 665–667 (2004).

37. Jo, S., Song, K. C., Desaire, H., MacKerell, A. D. & Im, W. Glycan reader: Automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *J. Comput. Chem.* **32,** 3135–3141 (2011).

38. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10037–41 (2001).

39. Gorham, R. D. *et al.* An evaluation of Poisson-Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. *Biopolymers* **95,** 746–754 (2011).

40. Harrison, R. E. S., Mohan, R. R., Gorham, R. D., Kieslich, C. A. & Morikis, D. AESOP: A Python Library for Investigating Electrostatics in Protein Interactions. *Biophys. J.* **112,** 1761–1766 (2017).

41. Shen, Y. & Bax, A. SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48,** 13–22 (2010).

42. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109,** 1528–1532 (2015).

43. Eric, J., Travis, O., Pearu, P. & al., et. SciPy : Open source scientific tools for Python. (2001).

44. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Struture for Efficient Numerical Computation. *Comput. Sci. {&} Eng.* **13,** 22–30 (2011).

45. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).

46. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* **1697900,** 51–56 (2010).

47. A. A. Hagberg, D. A. Schult & P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. *7th Python Sci. Conf.* **836,** 11–15 (2008).

48. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9,** 99–104 (2007).

CHAPTER 8

# Conclusion

## 8.1 Summary

Protein function is dependent on both static and dynamical aspects of protein structure. When structure is perturbed by mutations or polymorphisms, function can be affected. In our studies, we have explored how the R102G polymorphism affects the conformation of C3b and discussed how such structural effects can perturb interactions with complement regulators, conferring susceptibility for age-related macular degeneration. We have shown that mutations in FH19-20 (at site 1) perturb interactions with C3d (the TED from C3b) and are associated with age-related macular degeneration and rare diseases of the kidney, such as atypical hemolytic uremic syndrome. Furthermore, we have shown that the Y402H polymorphism that confers risk for age-related macular degeneration in FH7 disrupts pattern recognition of sulfated carbohydrates. Each of the studies characterizing these molecular mechanisms progress our understanding of the complement system, and in a more general sense contribute to an understanding of the machinery of cellular organisms – proteins.

For rational design of molecules with therapeutic and diagnostic potential, molecular mechanisms provide crucial information required to develop a new functional molecule. Initial molecules, however, are unlikely to perfectly meet design constraints. Successive rounds of optimization must be performed until the design criteria are met. From initial

ideation to optimization, computational methods can provide a framework to generate new hypotheses that may be tested with experimental techniques. As in the example of S1P2, initial peptide designs based on molecular mechanisms from molecular dynamics simulations can be improved through successive rounds of optimization. In our case, optimization involved computation to generate new hypotheses and experiment to evaluate the resulting hypothesis. As computational methods continue to be developed and improved, such workflows may further increase in efficiency and accelerate the design of new molecules with desired functional activities and tailored structural, physicochemical, and binding properties.

APPENDIX A

## Multivalent Recognition of C3d by Complement Factor H (SCR 19-20)

**Table A.1:** Comparison of observed computational results to literature-reported data. The table summarizes the amino acid being compared, the binding mode the amino acid was linked to as being important from our computational studies (AESOP or MD), and the result mutating the residue had on binding as reported in literature. "X" represents a topological relationship, "L" indicates loss of binding, and "G" indicates gain of binding. For AESOP results, there is considered to be a topological relationship if a mutation of the residue resulted in a predicted gain or loss of binding greater or less than 2.5 kJ/mol. For MD results, there is considered to be a topological relationship if the residue was involved in an interaction with an occupancy of 50% or greater.

| Key Amino Acid | | Implicated Binding Mode | | | AESOP Results | | MD Results | | | Literature Comparison | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prot[a] | Res[b] | Site 1 | Site 2 | Site 3 | Ala Scan[c] | Site-directed | Aliphatic | Hbond[d] | SB[e] | aHUS link | Loss of Binding | Gain of Binding |
| C3d | D36 | X | | | L | | | | | | X | |
| C3d | E117 | | X | | | | X | | | | X | |
| C3d | D122 | | X | | L | L | X | X | X | | X | |
| C3d | E160 | X | | | L | | | X | X | | X | |
| C3d | D163 | X | | | L | | X | | | | X | |
| C3d | Q168 | X | | | | L | X | | | | X | |
| FH19 | P1114 | | X | | | | X | | | X | | |
| FH19 | D1119 | | X | | L | L | | X | X | | X | |
| FH19 | Q1139 | | X | | L | | X | X | | | X | |
| FH19 | Y1142 | | X | | | | X | X | | X | | |
| FH20 | R1182 | X | | | | L | | X | X | X | X | |
| FH20 | W1183 | X | | | | | X | | | X | X | X |
| FH20 | T1184 | X | | | | L | X | | | X | X | X |
| FH20 | L1189 | | X | | | G | | | | X | | X |
| FH20 | T1190 | | X | | | | X | | | X | | |
| FH20 | E1198 | X | | | G | G | | | | X | | |
| FH20 | R1203 | X | | | L | L | X | X | X | X | X | |
| FH20 | R1210 | | | X | G | G | X | | | X | X | X |
| FH20 | R1215 | X | | | | | | X | X | X | X | |

[a]Protein
[b]Residue
[c]Alanine Scan
[d]Hydrogen bond
[e]Salt bridge

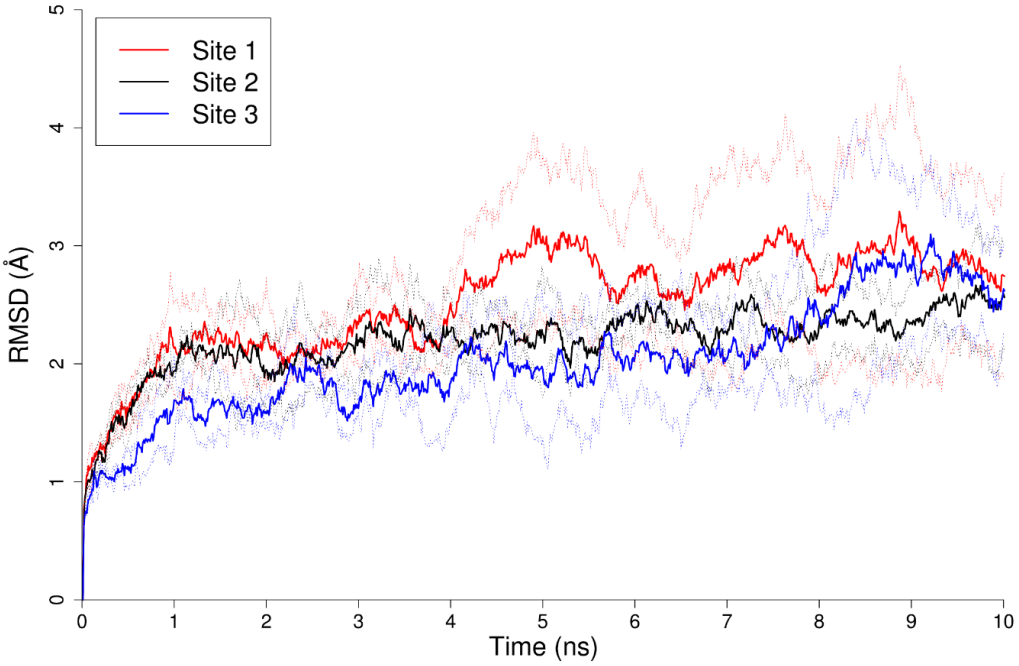



**Figure A.1:** Root-mean-square atomic deviation over the first 10ns of the MD trajectories. Solid lines represent mean values across triplicate trajectories, while dotted lines show one standard deviation above and below the mean. Profiles are colored according to the binding site investigated.

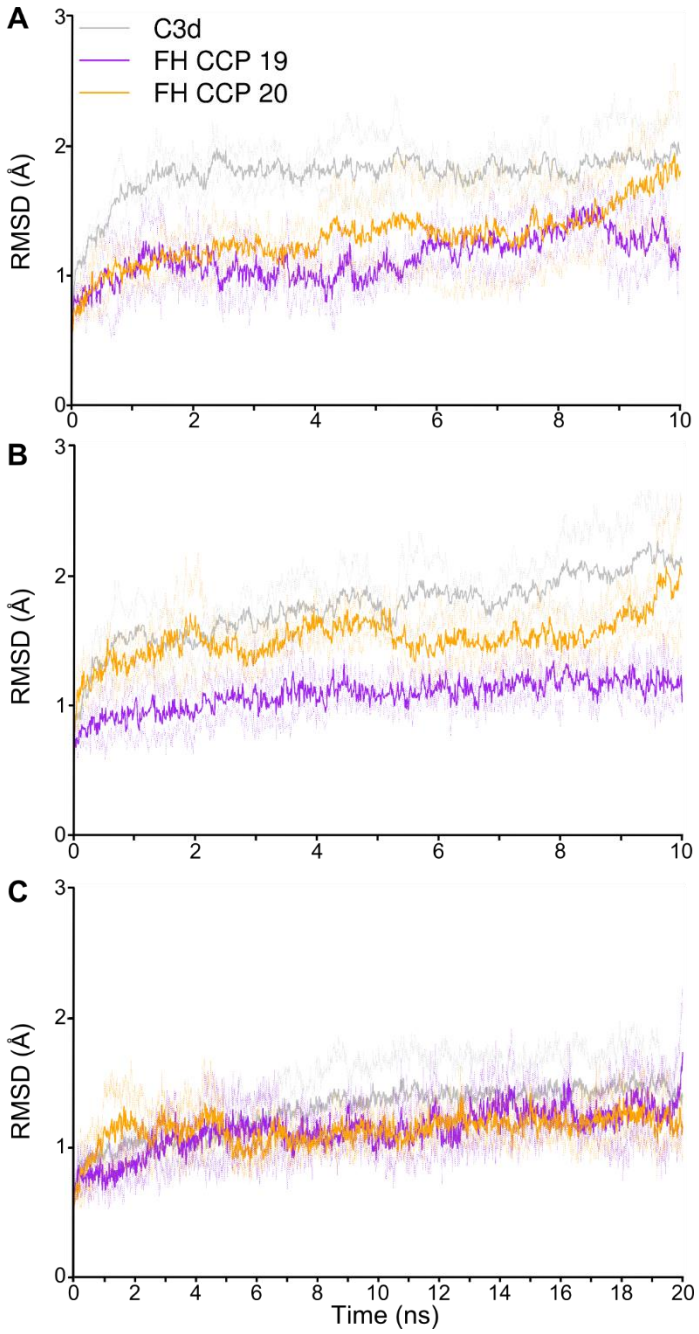



**Figure A.2:** Root-mean-square atomic deviation over full length triplicate MD trajectories for each subunit of the C3d:FH(19-20) complex at binding (A) site 1, (B) site 2, and (C) site 3. Mean values are represented by solid lines, while dashed lines represent one standard deviation above and below the mean.

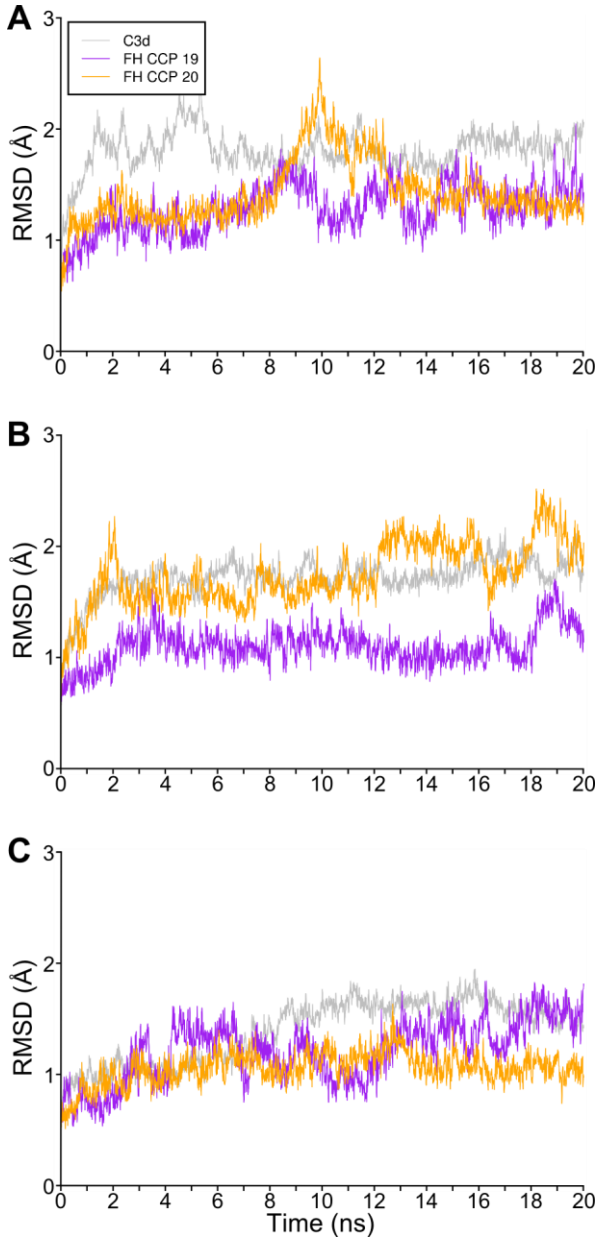



**Figure A.3:** Root-mean-square atomic deviation over the longest MD trajectory for each subunit of the C3d:FH(19-20) complex at binding (A) site 1, (B) site 2, and (C) site 3.

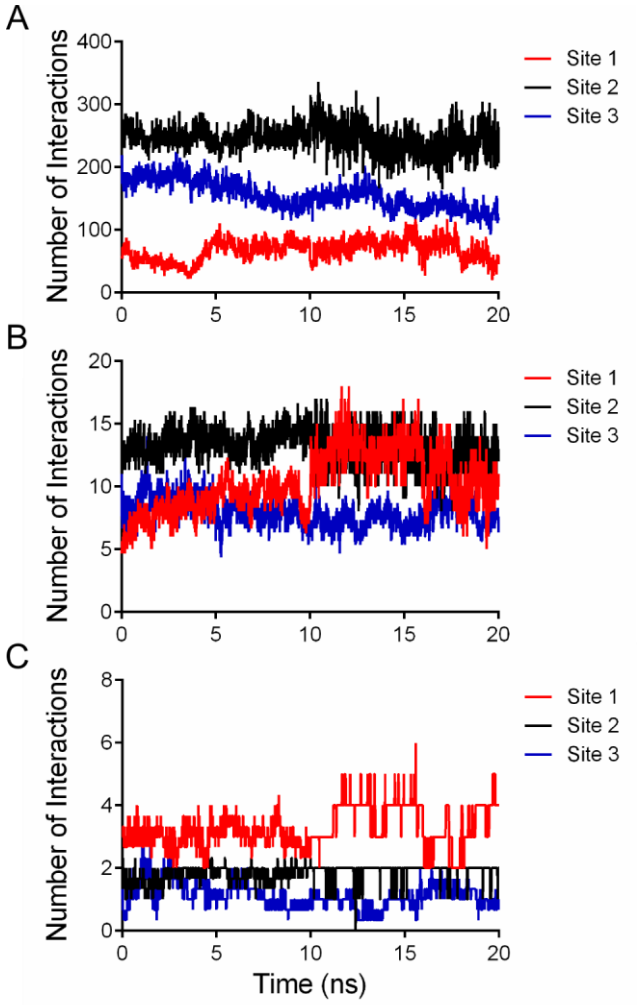



**Figure A.4:** Number of (A) aliphatic, (B) hydrogen bonds, and (C) salt bridges predicted to occur throughout MD simulations. For sites 1 and 2 the number of interactions is the mean value across triplicate trajectories from 0-10 ns and the value from the longest replicate from 10-20 ns. For site 3, the number of interactions is the mean value across triplicate trajectories from 0-20 ns.

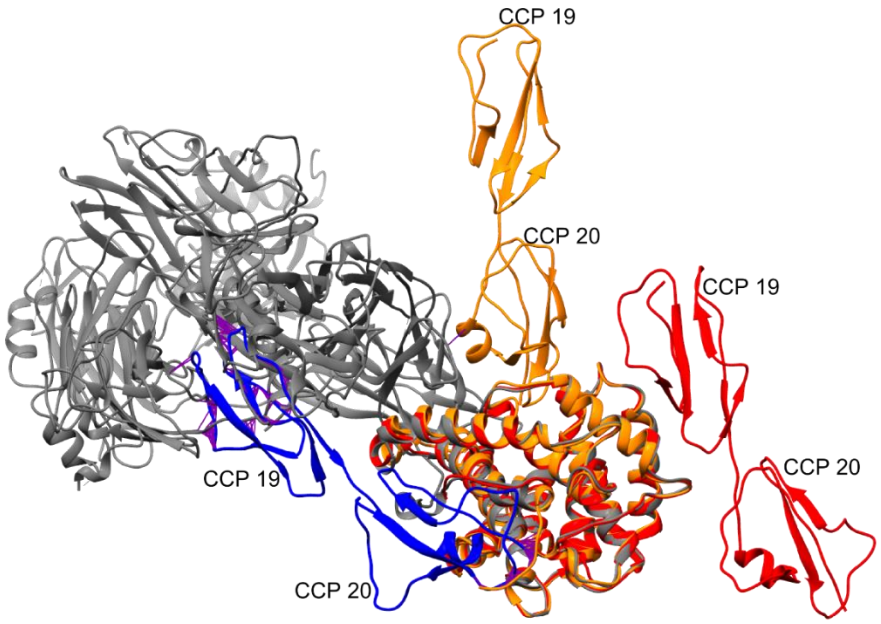



**Figure A.5:** Superpositioning of binding modes from PDB models 3OXU with the structure of C3b (model 2I07). The dim grey structure represents C3b. Site 1 (3OXU B:C), site 2 (3OXU A:E), and site 3 (3OXU A:D) are represented by the orange, red, and blue structures, respectively. Clashes between C3b and the binding modes are drawn in purple. Site one is only observed to have 1 clash, while site 3 has many more due to the overlap between CCP 19 and C3b.

APPENDIX B

# Molecular Mechanisms of Macular Degeneration Associated with the Complement Factor H Y402H Single Nucleotide Polymorphism

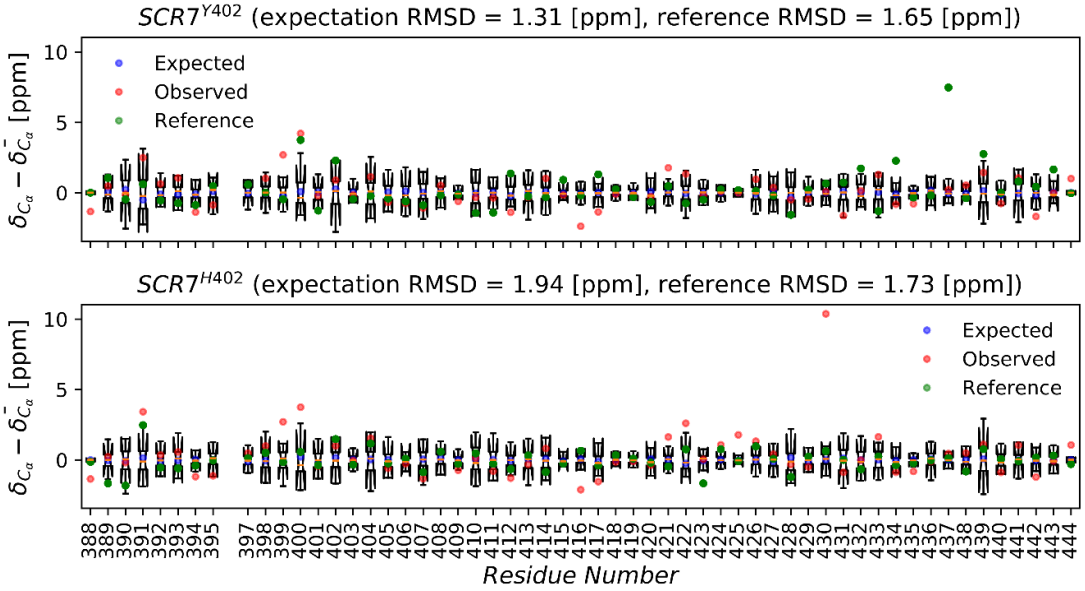



**Figure B.1**: Boxplots for predicted $C_\alpha$ chemical shifts from representative structures from Markov chains for $SCR7^{Y402}$ (top) and $SCR7^{H402}$ (bottom) are displayed above. Expected chemical shifts from the Markov chains for each SCR7 isoform are marked by blue circles, while predicted shifts for reference structures 2jgx and 2uwn from the PDB are marked by green circles. Known chemical shifts from experiments are annotated by red circles.

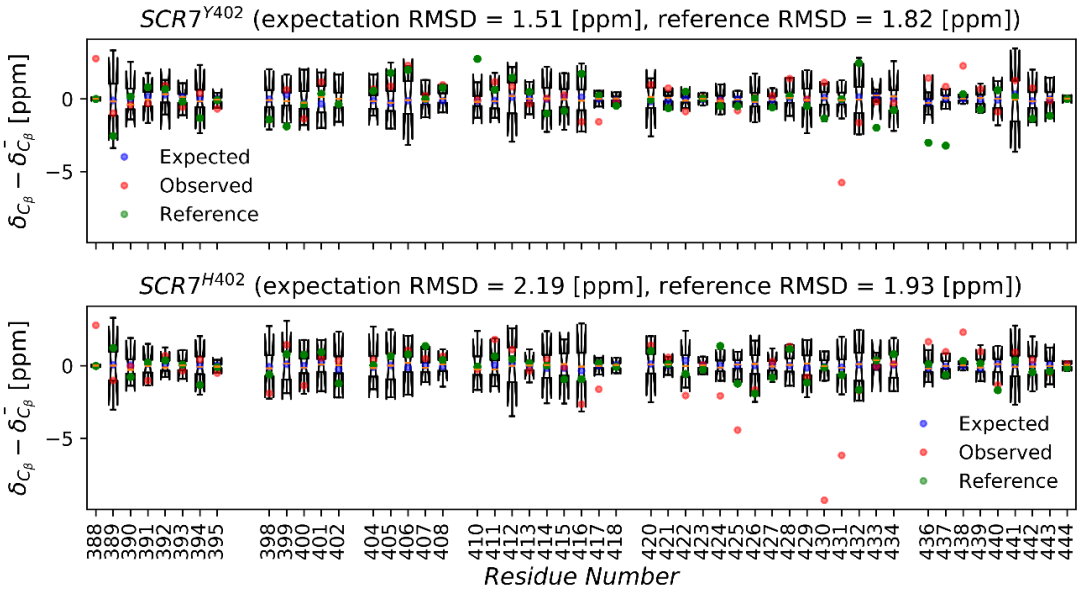



**Figure B.2**: Boxplots for predicted $C_\beta$ chemical shifts from representative structures from Markov chains for $SCR7^{Y402}$ (top) and $SCR7^{H402}$ (bottom) are displayed above. Expected chemical shifts from the Markov chains for each SCR7 isoform are marked by blue circles, while predicted shifts for reference structures 2jgx and 2uwn from the PDB are marked by green circles. Known chemical shifts from experiments are annotated by red circles.

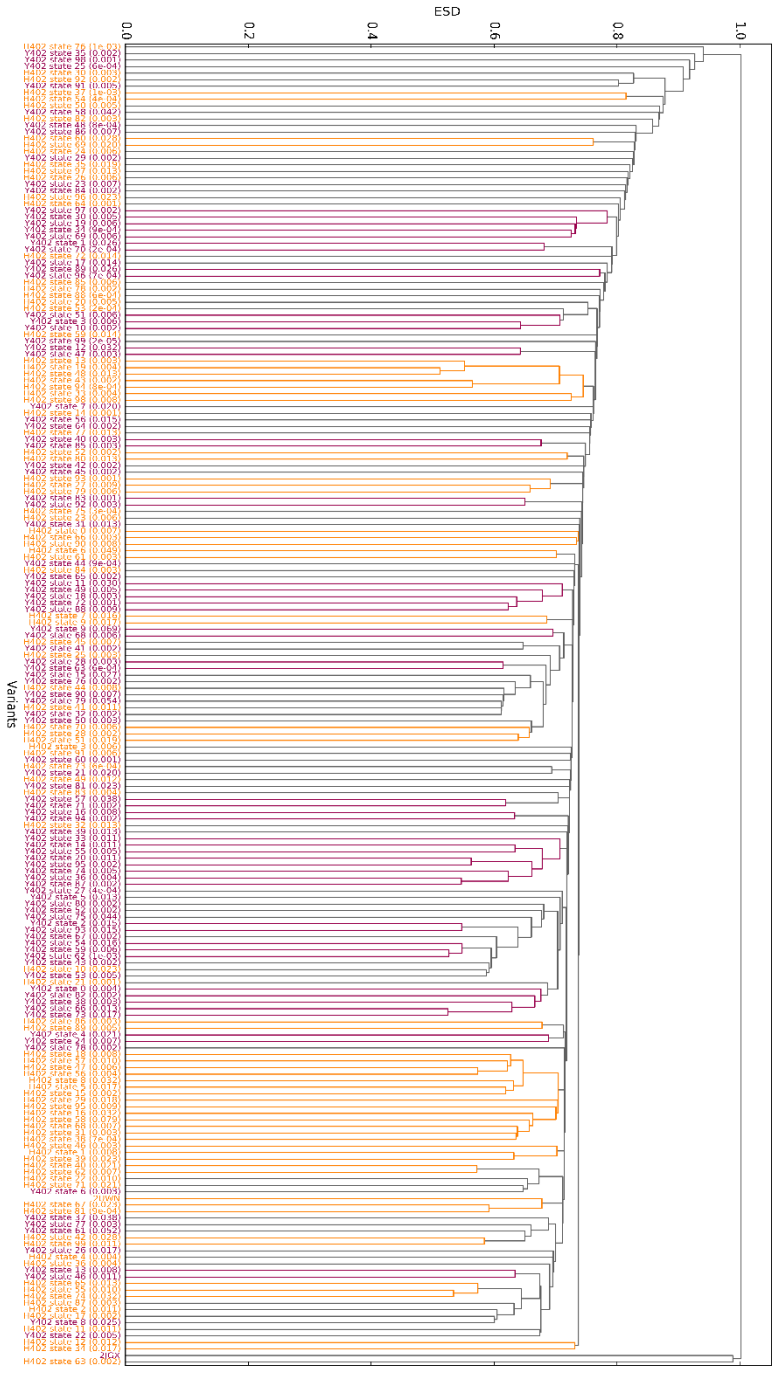



**Figure B.3**: Electrostatic similarity distance (ESD) is compared for each representative structure of SCR7 isoforms and for reference structures 2jgx and 2uwn from the Protein Databank. Most ESD values are below 0.7, indicating a large degree of similarity (an ESD of zero indicates identical potentials). The maximum ESD theoretically possible is 2.0.

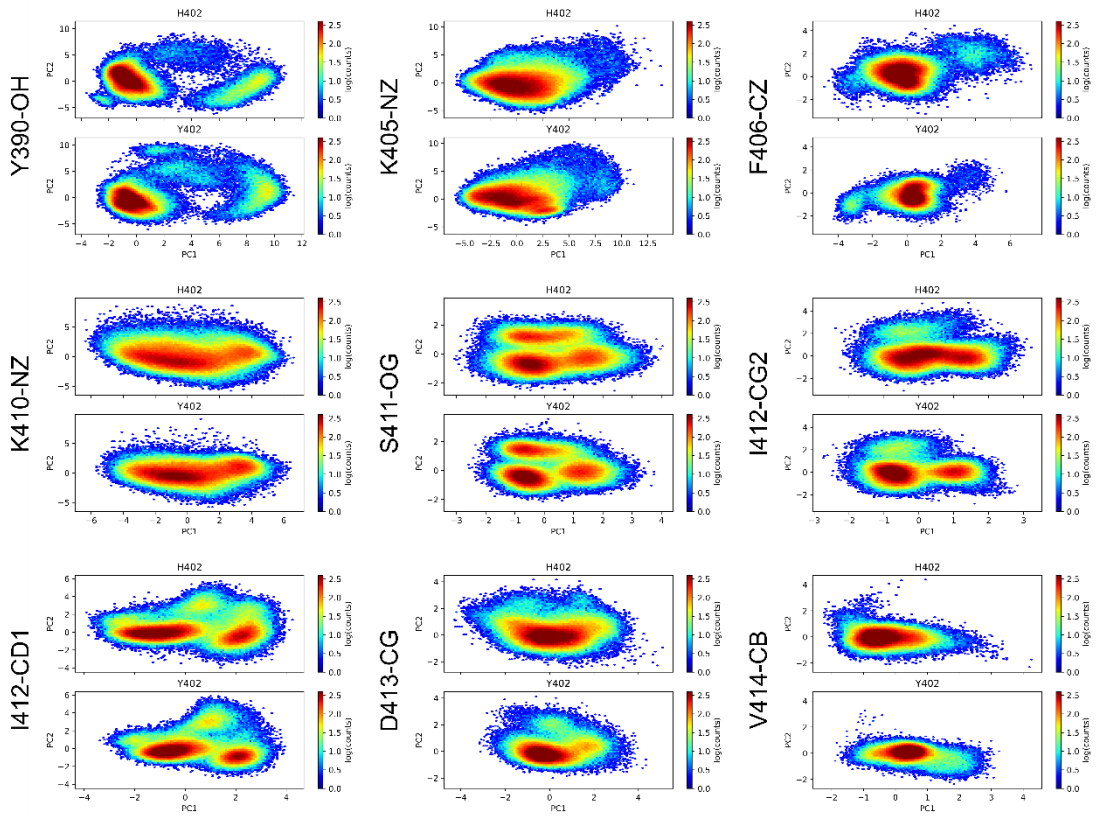



**Figure B.4**: Free energy landscapes for a two-dimensional representation of sidechain orientations for residues Y390, K405, F406, K410, S411, I412, D413, and V414 are shown for SCR7 isoforms, where a larger value for the logarithm of the number of observations in a bin indicates a lower free energy. To describe the orientation of sidechains, position vectors are found from coordinates of panel labels and decomposed into two dimensions with principle component analysis. For these residues, free energy landscapes between SCR7 isoforms are nearly identical.

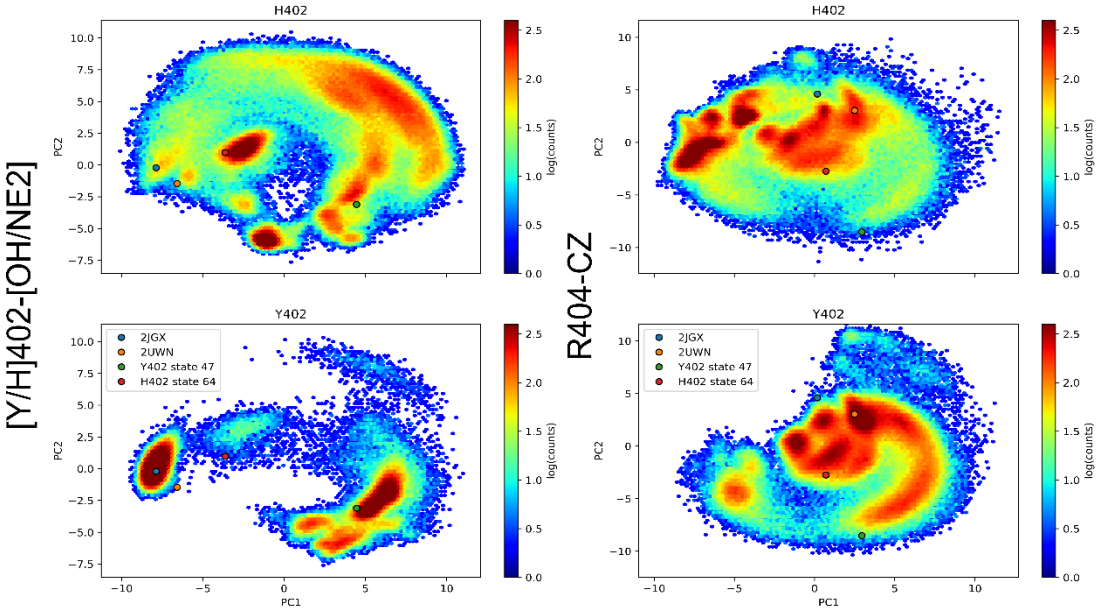



**Figure B.5**: Free energy landscapes for a two-dimensional representation of sidechain orientations for residues (A) [Y/H]402 and (B) R404 are shown for SCR7 isoforms, where a larger value for the logarithm of the number of observations in a bin indicates a lower free energy. To describe the orientation of sidechains, position vectors are found from coordinates of atom NE2 in H402, OH in Y402, and CZ in R404 and decomposed into two dimensions with principle component analysis. Regions of the landscape are annotated with circles to indicate the side chain orientations in references structures for SCR7$^{Y402}$ (2JGX) and SCR7$^{H402}$ (2UWN) and for states from each isoform where the maximum rate-constant for association with heparin is observed (Y402 state 47, H402 state 64).

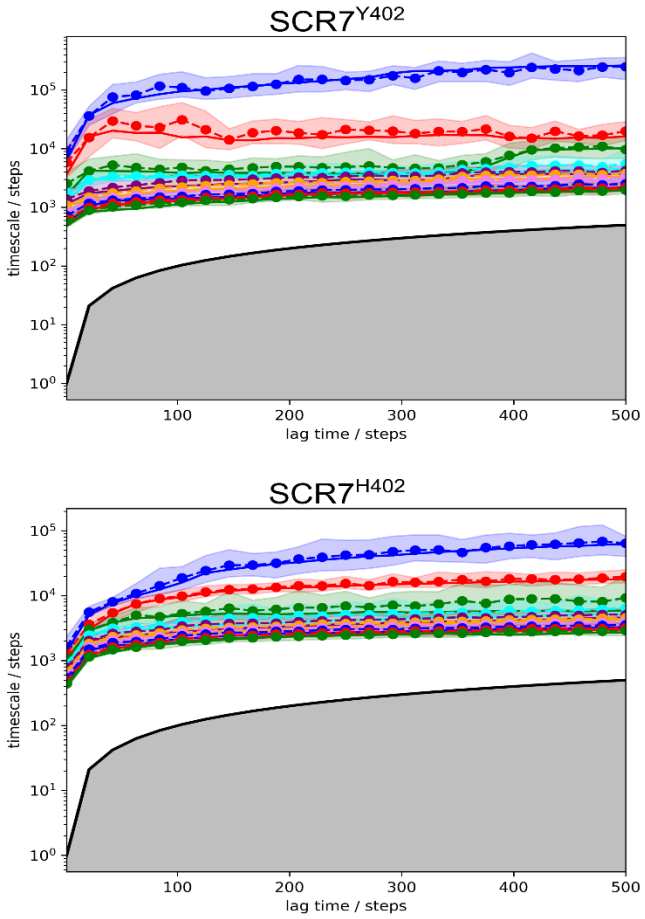



**Figure B.6**: Timescales calculated from the probability transition matrices from Markov chains for $SCR7^{Y402}$ (top) and $SCR7^{H402}$ (bottom) are plotted versus lag time used to construct the Markov chain. To satisfy the Markov property, a lag time should be selected where the timescale does not change with increasing values of the timescale. Different timescale responses are colored uniquely, and one step corresponds to 100 ps. We selected a lag time of 200 steps (20 ns).

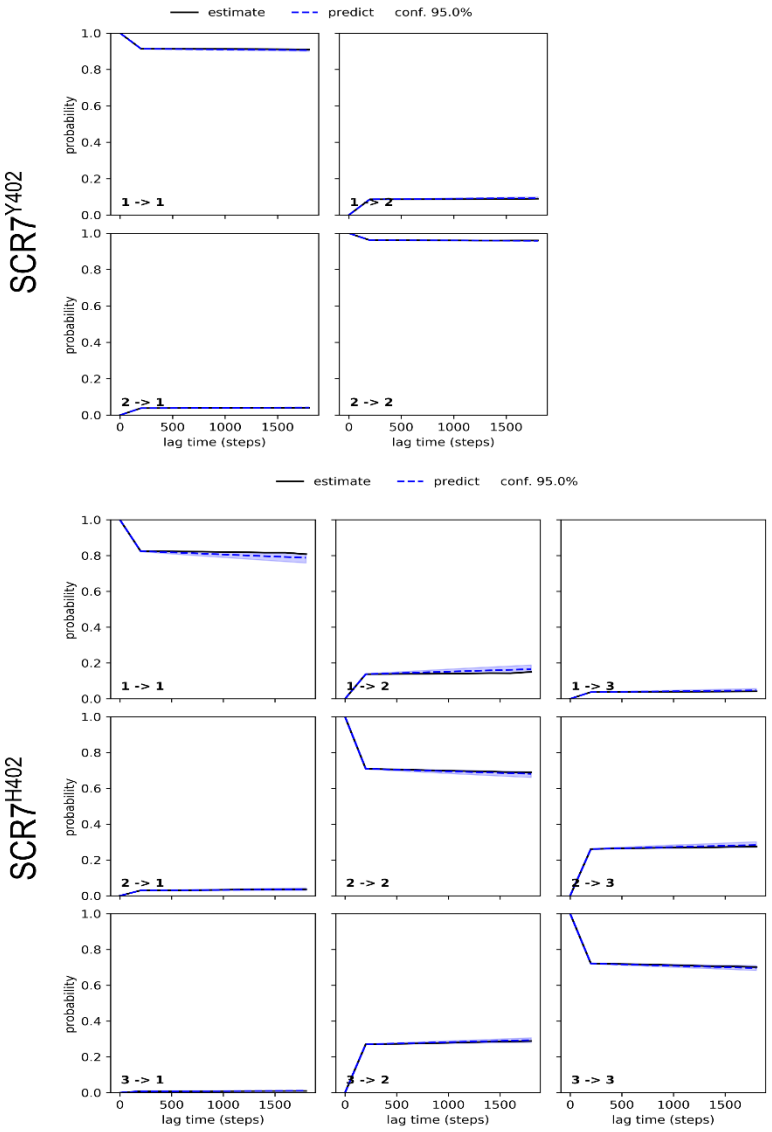



**Figure B.7**: Chapman-Kolmogorov validation of Markov chains for SCR7$^{Y402}$ (top) and SCR7$^{H402}$ (bottom) are show above. These panels suggest the probabilities of transitioning between metastable states by propagation of the Markov chain reproduces (within a 95% confidence interval) probabilities directly calculated from observed data. One step corresponds to 100 ps, and each subpanel describes the probabilities of a particular transition calculated at multiple lag times.

APPENDIX C

# Virtual Screening for PMX53 Analogs to Inhibit the Complement Component 5a Receptor

## I. Introduction

As a part of the innate immunity, the complement system comprises serum proteins and cell receptors that collectively act as the first line of defense against pathogens.[1] Through a complex balance of positive and negative feedback, the complement system recognizes and responds to foreign surfaces through the classical, lectin, and alternative pathways. Each pathway converges to the formation of the C3 convertase, an enzyme that cleaves C3 into C3a, an anaphylatoxin, and C3b, an opsonin. Furthermore, the opsonin C3b interacts with other complement proteins to form more C3 convertase, resulting in feed-forward amplification of the complement response. In the alternative pathway C3 convertase can subsequently bind to one or more additional molecules of C3b, forming C5 convertase. As a part of the terminal cascade, this enzyme cleaves C5 into C5a, a potent anaphylatoxin, and C5b, an opsonin that initiates the formation of the membrane attack complex (Scheme 1).

In healthy individuals, C5a is a key complement protein that induces inflammation and promotes chemotaxis of circulating leukocytes.[1] Overexpression of C5a, however, is linked to autoimmune conditions such as rheumatoid arthritis and neurodegenerative disease.[2] While the exact cause of the C5a imbalance is unclear, the need for specific complement inhibitors is evident. Currently, eculizumab, a monoclonal antibody, is the only

complement inhibitor of the terminal cascade approved by the FDA. In 2012, eculizumab was known as the most expensive drug costing an average of $409,500 per year[3], restricting the pharmaceutical to treating rare kidney diseases such as atypical hemolytic uremic syndrome.

In search of a selective inhibitor for the interaction of C5a with the C5a receptor (C5aR1), researchers designed a peptide named PMX53 with beta turn structure similar to the C-terminus of C5a.[2,4,5] While this peptide manifests potent inhibitory activity, PMX53 exhibits poor bioavailability *in vivo*. Compared to peptide or antibody therapeutics, small-molecules tend to be more inexpensive, available, and cell-permeable in addition to having more routes of administration.[6] Using the existing understanding of how PMX53 inhibits inflammation by C5a, we can select small molecules with similar physicochemical properties and desirable drug-like properties.

For the past 40 years, pharmacophore models have been used to describe and select compounds with desired properties.[7] Conceptually, pharmacophore models represent an intuitive way to represent a specified spatial arrangement of chemical features. Using this map of features, a wide variety of screening designs have been implemented to enhance the ability of pharmacophore models to select compounds with similar activity.[7] Of these strategies, pharmacophore screens that incorporate molecular dynamics (MD) simulations to account for protein flexibility have been shown to enrich compound hits with molecules more likely to exhibit the desired behavior.[7] Further enrichment has been demonstrated with docking and filtering protocols to select small-molecules that exhibit desirable properties.[7]

From previous studies[8], we have developed a pharmacophore model based on predicted key interactions of PMX53 with the C5a receptor (C5aR1). These predictions are in agreement with both theoretical[9] and experimental[5] findings. Key pharmacophore features involve key chemical groups in each residue of PMX53 (Figure 1). Of these residues, arginine is important for the formation of the ligand-receptor complex in terms of electrostatics, while tryptophan is primarily responsible for antagonism of C5aR1.[9]

Using multiple conformations of the C5aR1 receptor from MD simulations and ten nuclear magnetic resonance (NMR) structures of PMX53, we employ a screening protocol that accounts for flexibility in both receptors and ligands. In this ligand-based pharamacophore screen, each of ten pharmacophore models correspond to one NMR structure. After virtual screening using combinations of three to five pharmacophore features (Table 1), hits are further enriched by docking to each low energy C5aR1 structure. Ultimately, we present a library of small-molecules that bind C5aR1 and represent a starting point for developing specific inhibitors for the complement terminal cascade.

## II. Results and Discussion

### Virtual Screening

For ligand-based pharmacophore screening, a combination of chemical groups from each NMR structure of PMX53 comprise ten pharmacophore models (Figure 1). Each of these models is identical in the number and type of features contained; however, the relative spatial arrangements correspond to the unique side chain configurations of the ten NMR PMX53 structures. In each screening run, the center of each pharmacophore feature is

calculated from specific atoms within the corresponding residue; however the screening tolerance for each feature varies between screen 1 and screen 2-3 (Table 1). Screen 1 is intended as a coarse-grain screen, and the feature radii correspond to default radii of features in ZincPharmer except in the case of tryptophan where the radius is the minimum radius to encapsulate all backbone atoms of the indole ring. In screens 2-3, radii correspond to the average root mean square distance between the feature center and certain adjacent atoms across all PMX53.

In certain configurations of PMX53, the number of hits from screening varies drastically. Further inspection of these configurations suggests that variation in the side chain configuration of phenylalanine from PMX53 is restrictive in some NMR structures (Figure 1). To lessen the restrictivity of phenylalanine, the pharmacophore models were amended by replacing the aromatic feature from phenylalanine with a hydrophobic side chain and hydrogen bond acceptor from ornithine, a non-natural amino acid within PMX53. These new features are based on high occupancy interactions from a previously described C5aR1-PMX53 complex trajectory.[9]

From screening the DrugsNow[10] database in ZincPharmer[11] (screen 1) and an extended conformer database of 1.2 billion conformers from the DrugsNow database in Phase[12,13] (screens 2-3), we isolated approximately 7600 unique hits with similar spatial arrangement of chemical groups to PMX53.

**Docking**

Docking simulations predict the binding of each of the 7600 unique hits from virtual screening to the C5a receptor. Of these binding configurations, we isolated the top 200

highest affinity binding modes from unique compounds for visual inspection, prioritizing ligands that interact with key residues in C5aR1. These key residues are located C5a-binding pocket and participate in high occupancy interactions with PMX53 (Figure 2).[9] After a combination of visual inspection, filtering for hits that form hydrogen bonds, and root-mean-square distance clustering of binding configurations, we isolated twelve compounds predicted to bind to C5aR1. Each of these compounds have multiple binding modes binding near to the highest affinity binding mode on the clustering diagram (Figure 3), suggesting that the binding mode is favored. Favored binding modes for the selected twelve compounds also coincide with the C5a-binding pocket of C5aR1 (Figure 4), suggesting a similar binding locus to PMX53. In visualizing the binding of each top compound with C5aR1 (Figure 5), molecules 1, 3-5, 7, 11, and 12 are predicted to form hydrogen bonds with C5aR1. Of these compounds, only molecule 11 forms a hydrogen bond with a receptor residue that interacts with the arginine in PMX53. Molecules 1, 3-5, and 12 form hydrogen bonds with a residue that interacts with tryptophan in PMX53. Finally, molecules 1, 5, and 12 are predicted to form hydrogen bonds with residues of PMX53 that interact with multiple residues of PMX53. Together, these results indicate that virtual drug screening was able to select for small-molecules predicted to bind C5aR1 similar to PMX53. Moreover, these small-molecules have favorable drug-like properties as they agree with Lipinski's rule of five (Table 2).[14] These properties include a formula weight under 500 Da, 5 or less hydrogen bond donors, 10 or less hydrogen bond acceptors, and an octanol-water partition coefficient of 5 or less. Thus, the proposed small-molecules may have increased bioavailability.

## III. Conclusion

A combination of ligand-based pharmacophore screening and docking isolated twelve small-molecule analogues of PMX53. These compounds are expected to bind in the C5a-binding pocket of C5aR1, interact with residues of C5aR1 that have high occupancy interactions with PMX53, and agree with Lipinski's rules. Though binding has not been assessed for all compounds, molecule 11 has been shown to bind C5aR1 with micromolar affinity (Figure 6). Together, these twelve compounds represent a library of C5aR1 binding small-molecules that is an ideal platform for developing a novel small-molecules inhibitor of C5aR1. Future refinements and modifications of library compounds will optimize these small-molecules for treating autoimmune conditions linked to overactivation of the complement terminal cascade.

## IV. Materials and Methods

### Virtual Screening

Starting with ten conformers of PMX53, feature centers were calculated at specific locations in each associated PMX53 structure according to the atom selections in Table 1. In describing the radius for each feature, we used a combination of default features (screen 1), feature geometry (screen 1), and the average root-mean-square distance between the feature center and certain adjacent atoms across all PMX53 conformers (screen 2-3). In the case that the root-mean-square distance was larger than three angstroms, we used one half of the original distance value. Combinations of 4 to 5 pharmacophore features were used

to screen for similar compounds from an extended conformer database of small, drug-like compounds (Table 1). Since the arginine feature is expected to play a pivotal role, the feature was included in every set of pharmacophore features during the screening process. In screen 1, ZincPharmer[11] was used to screen for selected combinations of pharmacophore features for each NMR conformer of PMX53, finding hits from the ZINC DrugsNow database[10]. In screens 2-3, Schrodinger PHASE[12,13] was implemented to generate the extended conformer database of the publicly available ZINC DrugsNow database[10]. PHASE[12,13] was also used to screen for each combination of pharmacophore features for each NMR conformer of PMX53 and an overall PMX53 structure summarizing fluctuations across all conformers.

**Docking**

In order to isolate the most promising hits from virtual screening, a combination of docking based and cheminformatic methods were employed. Docking of hits from virtual screening was performed with Autodock Vina[15] using 4 protein structures from the characteristic trajectory that describes the current understanding of interactions between PMX53 and C5aR.[9] For each top affinity binding mode from docking, the mean number of predicted hydrogen bonds formed between the ligand and each of the four receptors were discerned. This descriptor was combined with the average predicted affinity of the top binding modes for each receptor and used to sort hits from virtual screening. For the purposes of visualization, hits that were not predicted to form hydrogen bonds with any receptor were excluded along with species with water-octanol partition coefficients that would have marginal solubility for subsequent experimental characterization. Prediction of

hydrogen bonding between ligands and receptors was performed in UCSF Chimera.[16] Additionally, all compounds were required to comply with Lipinski's rule of five. Chemical data was summarized using an Open Babel algorithm accessible via the ChemmineOB package in R.[17] While visually ranking top binding modes in UCSF Chimera, hits predicted to interact with key residues via hydrogen bonds and pi stacking were given the highest rank. Ligands that bound in the pocket of interest in a limited number of binding modes or formed few interactions with key residues were given secondary priority.

## V. References


(1) Mathern, D. R.; Heeger, P. S. Molecules Great and Small: The Complement System. *Clinical Journal of the American Society of Nephrology* **2015**.

(2) Woodruff, T. M.; Nandakumar, K. S.; Tedesco, F. Inhibiting the C5–C5a Receptor Axis. *Molecular Immunology* **2011**, *48*, 1631–1642.

(3) Shaughnessy, A. F. Monoclonal Antibodies: Magic Bullets with a Hefty Price Tag. *BMJ* **2012**, *345*, e8346–e8346.

(4) Zhang, L.; Mallik, B.; Morikis, D. Structural Study of Ac-Phe-[Orn-Pro-dCha-Trp-Arg], a Potent C5a Receptor Antagonist, by NMR. *Biopolymers* **2008**, *90*, 803–815.

(5) March, D. R. Potent Cyclic Antagonists of the Complement C5a Receptor on Human Polymorphonuclear Leukocytes. Relationships between Structures and Activity. *Molecular Pharmacology* **2004**, *65*, 868–879.

(6) Arkin, M. R.; Wells, J. A. Small-Molecule Inhibitors of Protein–protein Interactions: Progressing towards the Dream. *Nature Reviews Drug Discovery* **2004**, *3*, 301–317.

(7) Caporuscio, F.; Tafi, A. Pharmacophore Modelling: A Forty Year Old Approach and Its Modern Synergies. *Current Medicinal Chemistry* **2011**, *18*, 2543–2553.

(8) Kieslich, C. A. Development and Applications of a Computational Framework for Protein and Drug Design. PhD Thesis, University of California - Riverside: Department of Bioengineering, 2012.

(9) Tamamis, P.; Kieslich, C. A.; Nikiforovich, G. V.; Woodruff, T. M.; Morikis, D.; Archontis, G. Insights into the Mechanism of C5aR Inhibition by PMX53 via Implicit Solvent Molecular Dynamics Simulations and Docking. *BMC Biophysics* **2014**, *7*, 5.

(10) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* **2012**, *52*, 1757–1768.

(11) Koes, D. R.; Camacho, C. J. ZINCPharmer: Pharmacophore Search of the ZINC Database. *Nucleic Acids Research* **2012**, *40*, W409–W414.

(12) Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J Comput Aided Mol Des* **2006**, *20*, 647–671.

(13) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: A Novel Approach to Pharmacophore Modeling and 3D Database Searching. *Chemical Biology & Drug Design* **2006**, *67*, 370–372.

(14) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today: Technologies* **2004**, *1*, 337–341.

(15) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry* **2009**, NA – NA.

(16) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a Visualization System for Exploratory Research and Analysis. *J Comput Chem* **2004**, *25*, 1605–1612.

(17) Cao, Y.; Charisi, A.; Cheng, L.-C.; Jiang, T.; Girke, T. ChemmineR: A Compound Mining Framework for R. *Bioinformatics* **2008**, *24*, 1733–1734.

## VI. Figures

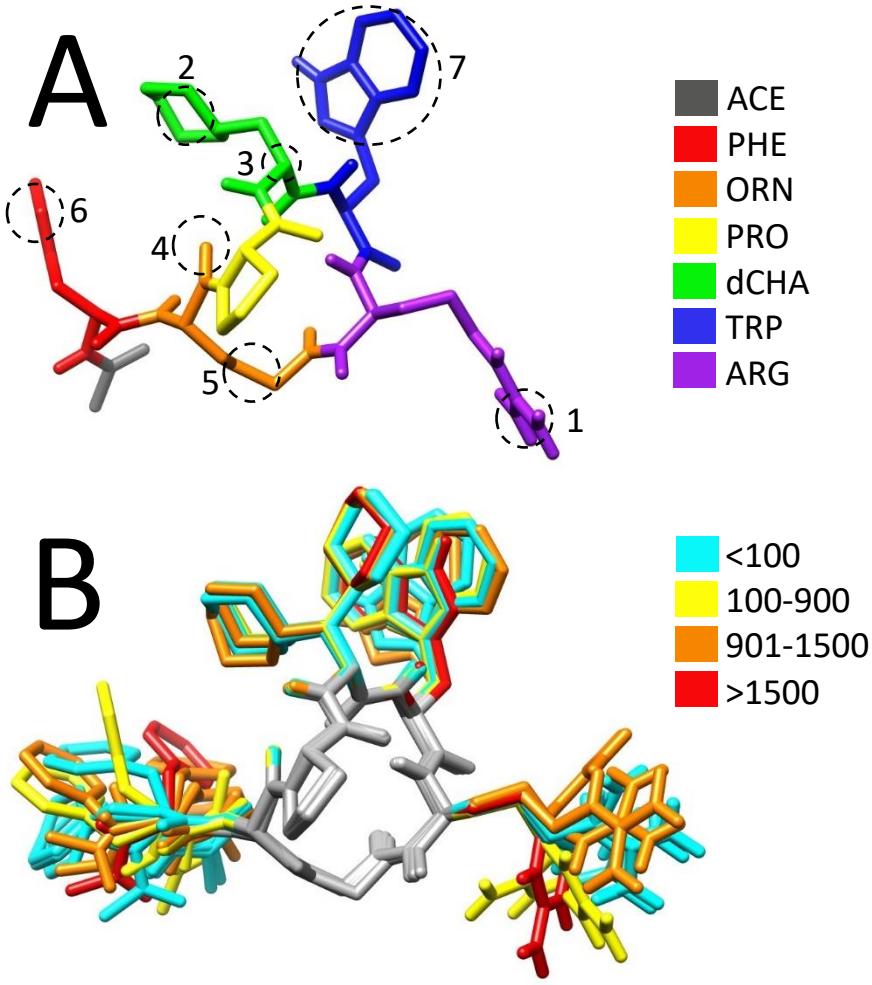



**Figure C.1**: NMR structures of PMX53 (hydrogens attached to nitrogen showing). (A) Location of pharmacophore features listed in Table 1 on the structure of PMX53. (B) Number of hits from initial pharmacophore screens with 10 NMR structures of PMX53. Low numbers of hits suggest the spatial arrangement of pharmacophore features is restrictive.

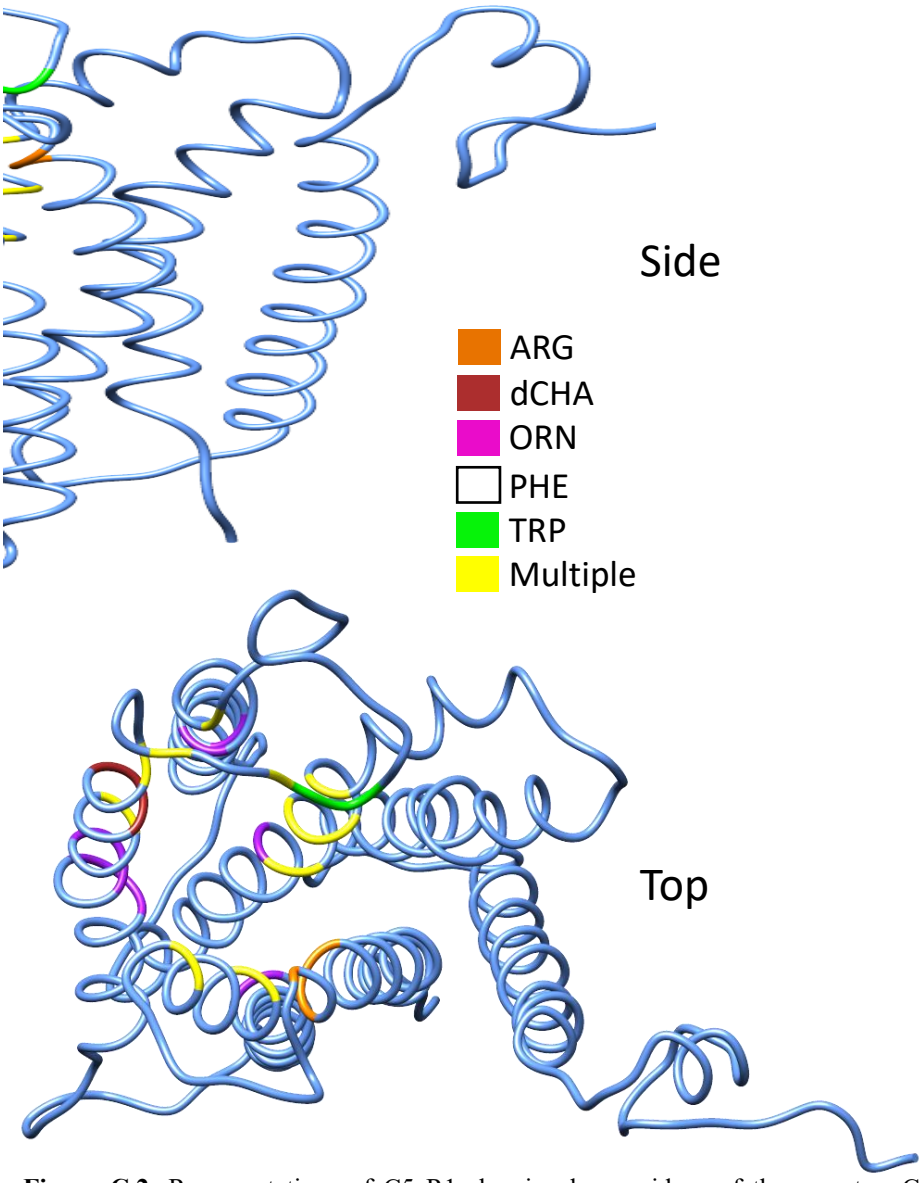



**Figure C.2**: Representations of C5aR1 showing key residues of the receptor. Colors correspond to residues of PMX53 that interact with the highlighted residues of C5aR1 in high occupancy interactions. The top view of C5aR1 illustrates the seven transmembrane alpha-helices, while the top view illustrates the C5a-binding pocket that is surrounded by key residues of C5aR1.

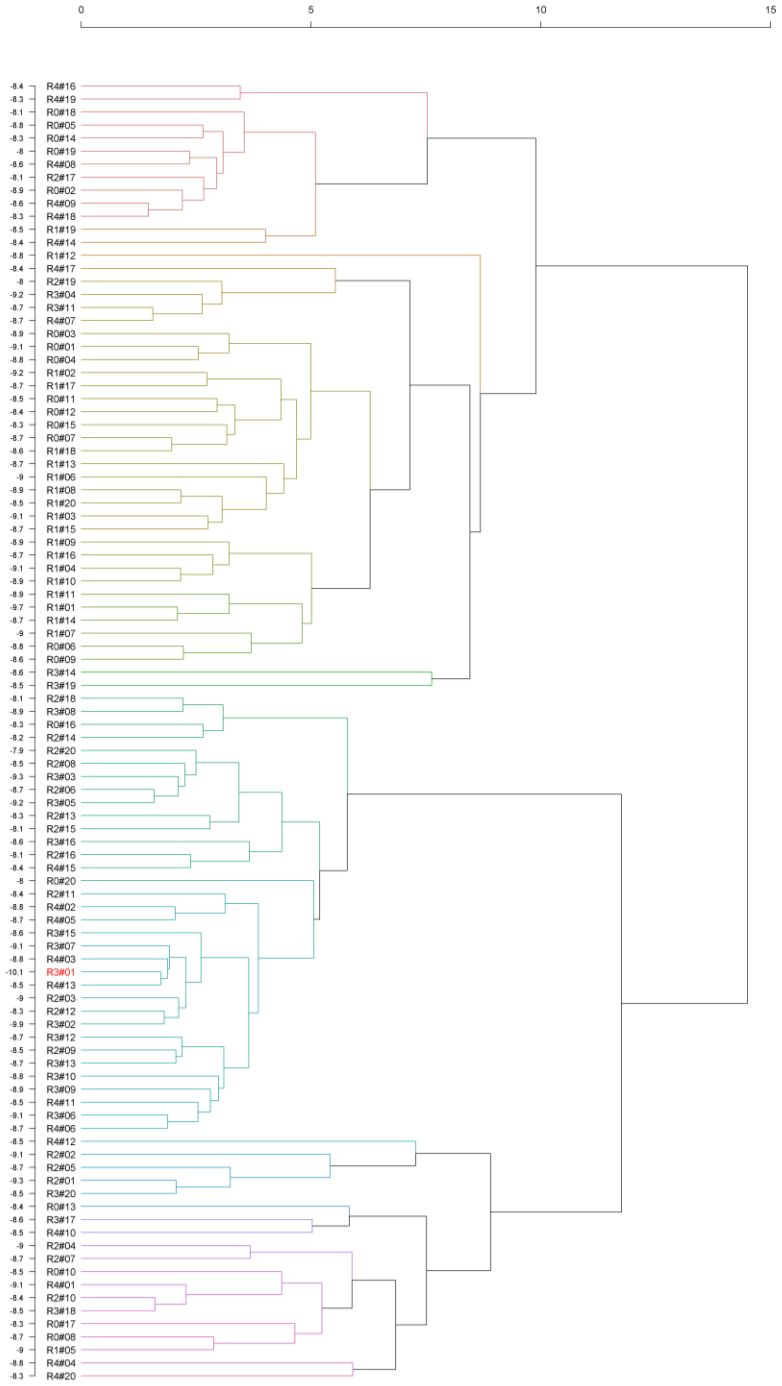



**Figure C.3**: Example of root-mean-square distance clustering of top binding poses for one compound on C5aR1. The number on the left hand axis represents the predicted binding affinity (kcal/mol), and the horizontal axis represents the root-mean-square distance between docked poses. Each pose is labeled according to Ri#j, where "i" represents the conformer of C5aR and "j" represents the number of the docked pose.

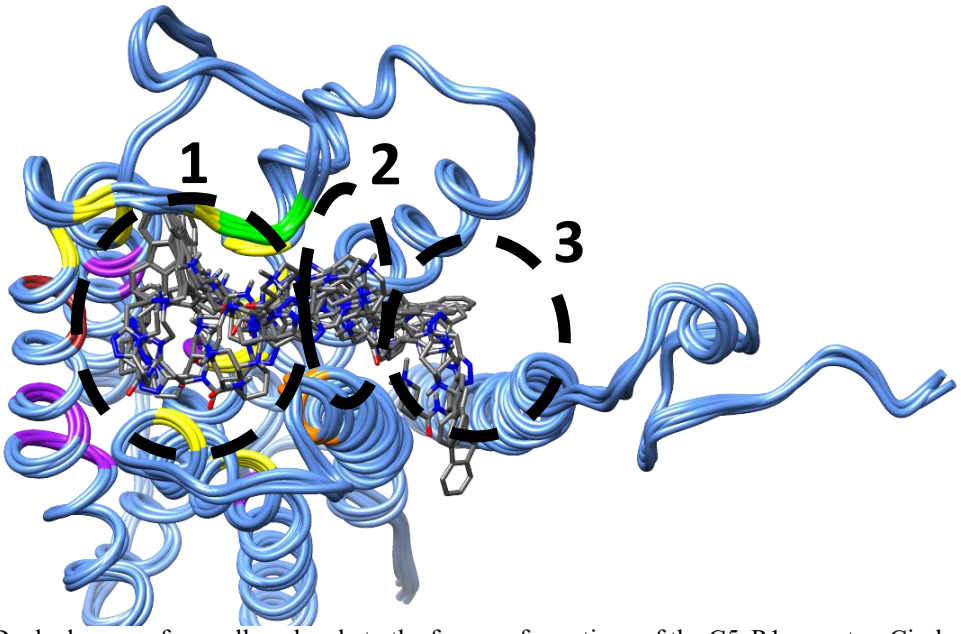



**Figure C.4**: Docked poses of a small-molecule to the four conformations of the C5aR1 receptor. Circled regions illustrate the (1) C5a-binding pocket, (2) bridge locus, and (3) allosteric pocket.

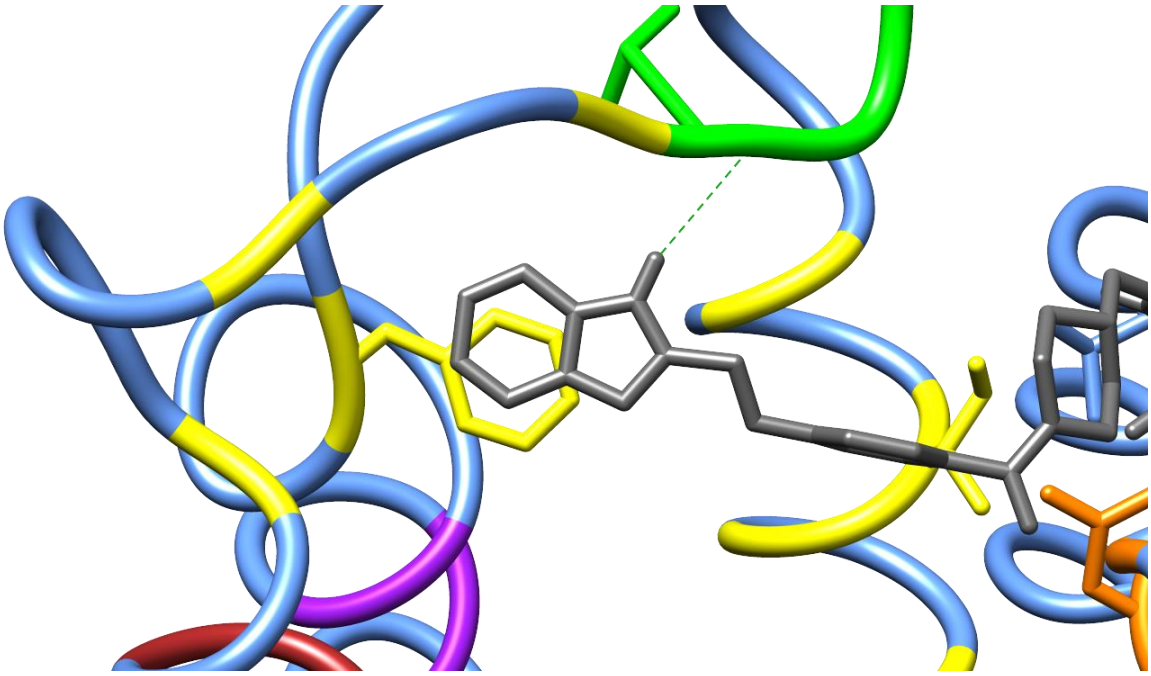


**Figure C.5**: Example of favorable interactions between key residues of C5aR1 with a small-molecule (grey) from the pharmacophore screen. Notable interactions include hydrogen bonding (dashed green line) and hydrophobic interactions between side chains.

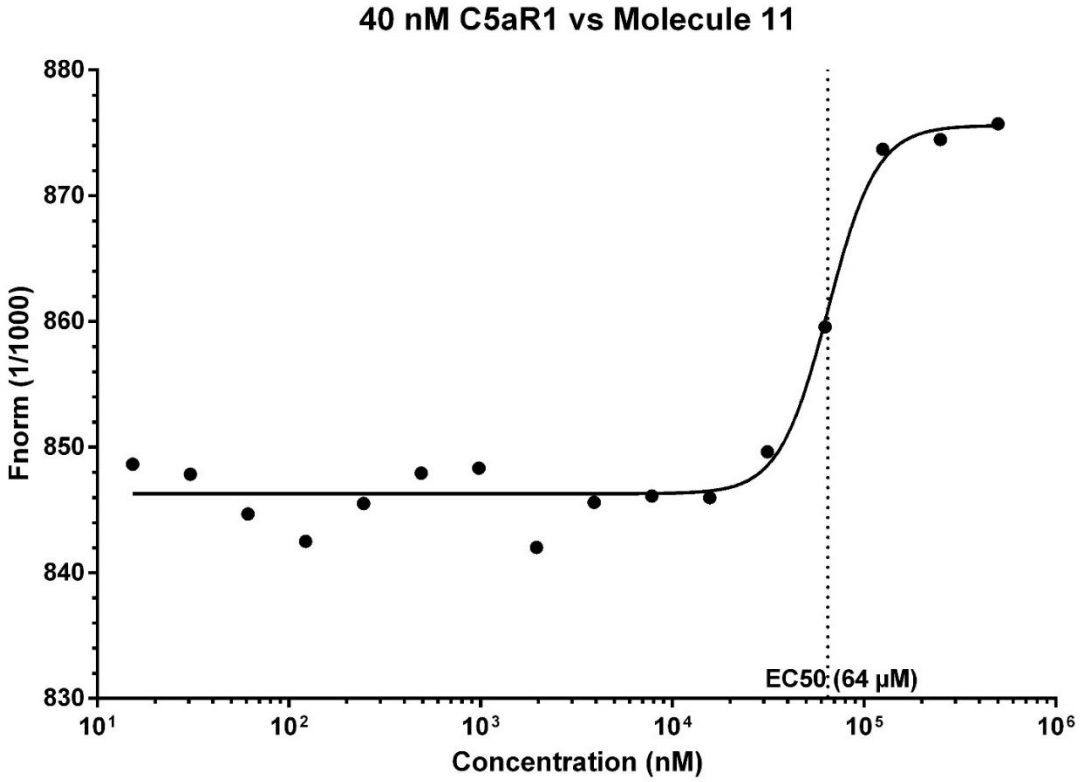



**Figure C.6**: Thermophoretic response of molecule 11 when titrated against 40 nM C5aR1 using microscale thermophoresis. Since C5aR1 is a G-protein coupled receptor, it is located in membrane preparations.

## VI. Tables

**Table C.1:** Ligand-based pharmacophore features and associated parameters

| Location Number[a] | Feature | PMX 53 Residue | Atoms of Feature[b] | Screening Protocol Tolerance (Å) | | |
|---|---|---|---|---|---|---|
| | | | | Screen 1 | Screen 2 | Screen 3 |
| **1** | Positive charge | Arginine | **CZ**, NH1, NH2 | 0.75 | 0.75 | 1.4 |
| **2** | Hydrophobic side chain | D-cyclohexylalanine | **CD1, CD2, CE1, CE2, CG, CZ** | 1 | 1 | 1.3 |
| **3** | Hydrophobic pivot point | D-cyclohexylalanine | **CG, CD, CB** | 0.75 | - | - |
| **4** | Hydrogen bond acceptor | Ornithine | **O**, C | - | - | 1.2 |
| **5** | Hydrophobic side chain | Ornithine | **CG, CD, CB** | - | - | 1.5 |
| **6** | Aromatic ring | Phenylalanine | **CD1, CD2, CE1, CE2, CG** | 1.1 | 1.1 | - |
| **7** | Aromatic ring | Tryptophan | **CG, CD1, CD2, NE1, CE2, CE3, CZ2, CZ3, CH2** | 1.9 | 1.9 | 1.9 |
| **7** | Hydrophobic side chain | Tryptophan | **CG, CD1, CD2, NE1, CE2, CE3, CZ2, CZ3, CH2** | 1.9 | 1.9 | 1.9 |

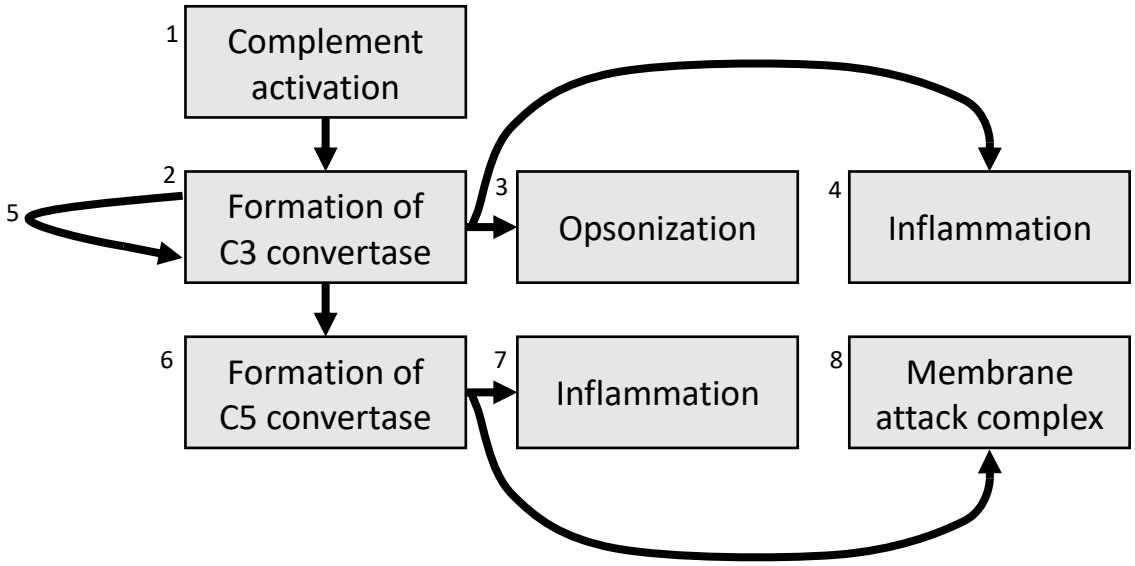
[a] Location of feature on PMX53, see Figure 1
[b] Centroid of bolded atoms determine the feature center

**Table C.2:** Chemical properties of small-molecule analogues of PMX53

| Compound ID | Open Babel Chemical Characteristics | | | | | | | | Docking Characteristics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | FW [a] | HBA1 [b] | HBA2 [b] | HBD [c] | logP [d] | MR [e] | nF [f] | TPSA [g] | Average Number of Hbonds | Average Binding Affinity | Average RMSD of Binding |
| **1** | 498.66 | 44 | 6 | 1 | 2.70 | 163.01 | 0 | 49.10 | 0.25 | -9.15 | 4.08 |
| **2** | 486.65 | 44 | 6 | 2 | 3.79 | 152.34 | 0 | 67.49 | 0.25 | -9.13 | 7.31 |
| **3** | 488.51 | 34 | 7 | 1 | 3.51 | 132.26 | 3 | 77.63 | 0.20 | -9.04 | 6.57 |
| **4** | 486.61 | 7 | 7 | 2 | 3.81 | 154.43 | 0 | 76.12 | 0.20 | -9.02 | 6.27 |
| **5** | 487.53 | 33 | 7 | 2 | 2.82 | 141.79 | 1 | 80.43 | 0.75 | -8.85 | 6.48 |
| **6** | 488.62 | 43 | 7 | 0 | 2.55 | 152.83 | 0 | 64.17 | 0.25 | -8.81 | 6.00 |
| **7** | 435.44 | 30 | 4 | 5 | 0.55 | 117.47 | 3 | 96.47 | 0.60 | -8.74 | 7.70 |
| **10** | 454.57 | 36 | 6 | 1 | 3.56 | 138.81 | 0 | 68.35 | 0.25 | -9.42 | 7.81 |
| **11** | 474.64 | 44 | 5 | 2 | 1.95 | 149.30 | 0 | 57.09 | 0.50 | -8.61 | 6.59 |
| **12** | 405.5 | 34 | 6 | 2 | 2.72 | 126.59 | 0 | 90.04 | 1.20 | -8.56 | 8.82 |

[a] Formula weight
[b] Hydrogen bond acceptor
[c] Hydrogen bond donor
[d] Octanol-water partition coefficient
[e] Molar refractivity
[f] Number of fluorine atoms
[g] Topological polar surface area

## VIII. Schemes




**Scheme C.1**: Complement activation from the alternative, lectin, or classical pathways (1) lead to the formation of the C3 convertase (2). This enzyme cleaves complement C3 into an opsonin (3) and inflammatory peptide (4). Furthermore, the opsonins that the C3 convertase creates can interact with other complement species to create more C3 convertase as a part of a positive amplification loop (5). In the terminal cascade, additional C3b interacts with existing C3 convertase to form C5 convertase (6). This enzyme cleaves complement C5 into a potent inflammatory agent (7) and a peptide that initiates formation of the membrane attack complex (8).

APPENDIX D

## Scripts and Methods

Certain scripts and methods are available from the BioMoDeL Github repository (https://github.com/BioMoDeL) or by request from the BioMoDeL website (http://biomodel.engr.ucr.edu). Methods are discussed below by the chapter.

Chapter 2

- Methods to perform electrostatic comparisons, computational alanine scans, and *in silico* directed mutagenesis are implemented in AESOP, an open source Python package (available at https://pypi.org/project/aesop/)

Chapter 3

- Free energies were calculated for multiple structures with AESOP
- Free energy components were compared according to equations described within the chapter

Chapter 4

- A single transition path for activation of C3b was estimated with the script "estimate_transition_path.py" after retrieving output files from the iENM protocol of Tekpinar and Zheng (https://enm.lobos.nih.gov/start_ienm.html)
- An Anton2-compatible ARK file was prepared using the script "generate_anton_restraints.py" after estimation and sampling of a model transition path (this method requires a base input ARK file and an XML file describing domains within the protein)
- The energy landscape for Anton2 simulations was constructed with the script "process_trj.py"
- Baker-Hubbard hydrogen bonds were calculated with the script "find_hbonds.py"

Chapter 5

- "mmgbsa_source_RH_31072014.r" was used as source code to run "mmgbsa_script_RH_30072014.r", calculating MM/GBSA energies and saving them to CSV files
- "smd.forces.r" was used to parse log files from steered molecular dynamics trajectories and save CSV files containing forces at each timestep of the trajectories
- The lab R source code ("analyze_md_140516.r") was used to process molecular dynamics trajectories, creating CSV datafiles required as inputs for the following analyses:
    - "cmp_md_interactions.r" calculates interaction occupancies across trajectories

- "cmp_md_interactions_vmd_hbonds.r" calculates hydrogen bond occupancies using the method implemented in VMD (Visual Molecular Dynamics)
- "cmp_md_rmsd_profile.r" calculates root-mean-square deviation of protein structure over a trajectory after superimposing on the initial protein complex
- "cmp_md_rmsd_subunit.r" calculates root-mean-square deviation of a protein structure over a trajectory after superimposing by domain subunits
- "cmp_md_sasa_profile.r" calculates solvent accessible surface area lost upon binding over molecular dynamics trajectories
- "cmp_mmgbsa_energy_time.r" plots MM/GBSA energy values over time from molecular dynamics trajectories
- "cmp_smd_force_profile.r" compares forces from steered molecular dynamics trajectories over time

Chapter 6

- "calc_pmf.py" was used to calculate energy landscapes for trajectories
- "calc_rmsd.py" was used to calculate root-mean-square deviation profiles for molecular dynamics trajectories over time
- "calc_rmsf.py" was used to calculate root-mean-square fluctuations for peptide residue positions
- "calc_sasa.py" was used to calculate per-residue solvent accessibilities
- "patch_ss.py" was used to patch disulfide bonds between specified cysteines, removing hydrogens as required

Chapter 7

- "xyz2torsions.py" was used to convert atomic Cartesian coordinates to a feature set of torsion angles
- "analyze_aesop_esd.py" was used to compare electrostatic similarity of FH SCR7 conformations for multiple isoforms
- "parameterize_model_h402.py" and "parameterize_model_y402.py" were used to initially parameterize time-lagged independent component analysis and Markov chain models for $SCR7^{H402}$ and $SCR7^{Y402}$, respectively
- "build_msm_h402.py" and "build_msm_y402.py" were used to manually refine Markov chain models for $SCR7^{H402}$ and $SCR7^{Y402}$, respectively
- Analysis and validation of Markov chains was performed for $SCR7^{H402}$ and $SCR7^{Y402}$ with "analyze_msm_h402.py" and "analyze_msm_y402.py", respectively
- Markov chains were compared with "compare_msm.py"
- Network files in GEXF format were generated using "generate_networks.py"
- "20180321_coevol_dist_final.py" was used to calculate distances between coevolved residues throughout molecular dynamics trajectories

APPENDIX E

**Curriculum Vitae**

# Reed Edward Shudde Harrison

## a. Professional Preparation

| | | |
|---|---|---|
| University of California, Riverside | Bioengineering | Ph.D., 2018 |
| University of Texas at Austin | Biomedical Engineering | B.S., 2013 |

## b. Appointments

2017: Instructor, Statistics and Programming GradQuant Bootcamp, University of California, Riverside
2017: Teaching Assistant, Integration of Computational and Experimental Biology, University of California, Riverside
2016: Teaching Assistant, Biophysics and Biothermodynamics, University of California, Riverside
2015: Doctoral Candidate, Department of Bioengineering, University of California, Riverside
2013: Graduate Research Assistant, Department of Bioengineering, University of California, Riverside

## c. Graduate Advisor

Professor Dimitrios Morikis, Department of Bioengineering, University of California, Riverside

## d. Awards and Honors

2018: Best Poster in Computational Bioengineering, 19th Annual University of California Systemwide Bioengineering Symposium
2017: University of California President's Dissertation-Year Fellowship, University of California, Riverside
2017: Outstanding Teaching Assistant in the Department of Bioengineering, University of California, Riverside
2016: Whitaker Summer Grant, Spanish National Cancer Research Centre, Madrid, Spain
2016: Graduate Research Mentorship Fellowship, University of California, Riverside
2014: NSF IGERT in Video Bioinformatics, University of California, Riverside
2013: Chancellor's Fellowship, University of California, Riverside

## e. Skills

**(i) Computational:** Molecular dynamics, protein engineering, docking, virtual screening, Poisson-Boltzmann electrostatics, molecular mechanics, image and video processing, bioinformatics, elastic network models, statistical methods, programming (MATLAB, R, Python, C#, C++)

**(ii) Experimental:** Protein binding assays (microscale thermophoresis, ELISA, red blood cell lysis)

**f. <u>Publications</u>** (* indicates primary authorship)


**Harrison, R. E. S.***; Morikis, D. (*in revision*). Molecular Mechanisms of Macular Degeneration Associated with the Complement Factor H Y402H Mutation. Biophysical Journal.

Rosenberg, E. M.*; **Harrison, R. E. S.***; Tsou, L. K.; Drucker, N.; Humphries, B; Rajasekaran, D.; Luker, K.; Wu, C.; Song, J.; Murphy, J. W.; Cheng, Y.; Shia, K.; Luker, G. D.; Morikis, D.; Lolis, E. J. (*in submission*). Functional Characterization, Dynamics, and Mechanism of CXCR4 Antagonists on a Constitutively Active Mutant. Cell Chemical Biology.

Ngo-Duc, T.*; Plank, J.; Chen, G.; **Harrison, R. E. S.**; Morikis, D.; Liu, Haizhou; Haberer, E. (2018). M13 Bacteriophage Spheroids as Scaffolds for Directed Synthesis of Spiky Gold Nanostructures. Nanoscale. doi: 10.1039/c8nr03229g

Winicki, N.*; **Harrison, R. E. S**; Perry, J.; Morikis, D. (2018). Identification of Targetable Surfaces of Cks1 for Development of New Cancer Therapeutics. UC Riverside Undergraduate Research Journal. url: https://escholarship.org/uc/item/14z8979k

Mohan, R. R.*; Wilson, M.; Gorham, R. D. Jr.; **Harrison, R. E. S.**; Morikis, V. A.; Kieslich, C. A.; Orr, A. A.; Coley, A. V.; Tamamis, P.; Morikis, D. (2018). Virtual Screening of Chemical Compounds for Discovery of Complement C3 ligands. ACS Omega. doi:10.1021/acsomega.8b00606

**Harrison, R. E. S.***; Mohan, R. R.; Gorham, R. D.; Kieslich, C. A.; Morikis, D. (2017). AESOP: A Python Library for Investigating Electrostatics in Protein Interactions. Biophysical Journal. doi:10.1016/j.bpj.2017.04.005

Lopez-Perrote, A*; **Harrison, R. E. S.***; Subias, M.; Alcorlo, M.; Rodriguez de Cordoba, S.; Morikis, D.; Llorca, O. (2017). Ionic tethering contributes to the conformational stability and function of complement C3b. Molecular Immunology. doi:10.1016/j.molimm.2016.12.015

Mohan, R. R.*; Cabrera, A. P.; **Harrison, R. E. S.**; Gorham, R. D.; Johnson, L. J.; Ghosh, K.; Morikis, D. (2016). Peptide redesign for inhibition of the complement system: Targeting age-related macular degeneration. Molecular Vision. <http://www.molvis.org/molvis/v22/1280/>.

Zhang, C*; Brown, M. Q.; van de Ven, W.; Zhang, Z.; Wu, B.; Young, M.; Synek, L.; Borchardt, D.; **Harrison**, **R. E. S.**; Pan, S.; Luo, N.; Huang, Y. M.; Ghang, Y.; Ung, N.; Li, R.; Isley, J; Morikis, D.; Song, J.; Guo, W.; Hooley, R.; Chang, C. A.; Yang, Z.; Zarsky, V.; Muday, G.; Hicks, G. R.; Raikhel, N. V. (2016). Endosidin2 targets conserved EXO70 to inhibit exocytosis. Proceedings of the National Academy of Sciences of the United States of America. doi:10.1073/pnas.1521248112

**Harrison**, **R. E. S.***; Gorham, R. D.; & Morikis, D. (2015). Energetic evaluation of binding modes in the C3d and factor H (CCP 19-20) complex: Energetic Evaluation of C3d:FH(19-20) Binding Modes. Protein Science. doi:10.1002/pro.2650

**Harrison**, **R. E. S.***; Criss, Z. K.; Feller, L.; Modi, S. P.; Hardy, J. G.; Schmidt, C. E.; Suggs, L. J.; Murphy, M. B. (2015). Mechanical properties of α-tricalcium phosphate-based bone cements incorporating regenerative biomaterials for filling bone defects exposed to low mechanical loads. Journal of Biomedical Materials Research Part B: Applied Biomaterials. doi:10.1002/jbm.b.33362

## g. Patents

Morikis, D.; Mohan, R.; **Harrison, R. E. S.**; Gorham, R. D. Jr.; Ghosh, K.; Cabrera, A. P. (provisional, 2016; application, 2017). Potent and Highly Soluble Pegylated Compstatin Peptides.

## h. Software

**Harrison, R. E. S.**; Mohan, R. R.; Morikis, D. (2016). AESOP. <https://github.com/biomodel/aesop>. doi: 10.5281/zenodo.269610

## i. Synergistic Activities

2017-2018: Mentor, Science Fair, Riverside Unified School District, California
2015-2016, 2018: Examiner and Grader in Protein Modelling, Inland Empire Science Olympiad, Ramona High School, Riverside
2017: Greeter, University of California – Riverside booth, 49th Annual Meeting of the Biomedical Engineering Society, Phoenix Convention Center, Arizona
2017-2018: Examiner and Grader in Microbe Mission, Inland Empire Science Olympiad, Ramona High School, Riverside
2017: Judge, Riverside County Science and Engineering Fair, Riverside, California.
2016: Member, Health Insurance Coordinator Search Committee, University of California, Riverside
2016: Member, Graduate Student Health Insurance Program Committee, University of California, Riverside
2015-2016: Graduate Student Association Representative, Bioengineering Interdepartmental Graduate Student Association, University of California, Riverside
2014-2015: President, Bioengineering Interdepartmental Graduate Student Association, University of California, Riverside
2014-2016: Undergraduate mentor, Bioengineering Mentorship Program, University of California, Riverside
2014: Examiner and Grader in Road Scholar category, Inland Empire Science Olympiad, University of California, Riverside

## j. Conferences

2018: Poster Presentation, Molecular Mechanisms of Macular Degeneration Associate with the Complement Factor H Y402H Single Nucleotide Polymorphism, Gordon Research Conference in Computational Chemistry, Mount Snow, West Dover, Vermont
2018: Poster Presentation, Missense Single Nucleotide Polymorphism Disrupts a Coevolved Pair and Adversely Affects Pattern Recognition in Complement Factor H, 19th Annual University of California Systemwide Bioengineering Symposium, Riverside Convention Center
2018: Poster Presentation, Molecular Basis for the Link between Macular Degeneration and a Single Nucleotide Polymorphism, 62nd Annual Meeting of the Biophysical Society, Moscone Center, San Francisco
2017: Poster Presentation, Molecular Basis for the Link between Macular Degeneration and a Single Nucleotide Polymorphism, 49th Annual Meeting for the Biomedical Engineering Society, Phoenix Convention Center, Arizona

2017: Oral Presentation, Ionic tethering contributes to the conformational stability and function of Complement C3b, 4th Annual Center for Plant Cell Biology Post Doc Symposium, University of California, Riverside

2016: Oral Presentation, Conformational Dynamics of Factor H are Altered in Age-related Macular Degeneration risk-variant (Y402H), University of California Video Bioinformatics NSF IGERT Retreat, UCLA Conference Center, Lake Arrowhead

2016: Poster Presentation, Synergy of Putative Binding Modes in the Factor H (CCP 19-20) and C3d Complex, 60th Annual Meeting of the Biophysical Society, Lost Angeles, California

2014: Poster Presentation, Research Fair for Engineers Without Borders, University of California, Riverside

2014: Poster Presentation, Virtual Screening Based Drug Discovery for Complement-Mediated Inflammation, University of California Systemwide Bioengineering Symposium

2014: Video Presentation, Role of Dynamics in Protein Interactions, University of California Video Bioinformatics NSF IGERT Retreat

**k. <u>Research Experience</u>**

**(i) Regulatory protein interactions between innate immune proteins C3d and Factor H**

a. Identify biological interactions from crystal structures by computing thermodynamic properties
b. Discern which putative binding modes are biologically relevant and which modes are artefactual
c. Apply knowledge to design peptide biomarkers of C3d for diagnostic applications
d. Quantify dissociation constants of designed peptide biomarkers

**(ii) Virtual drug screening for complement inhibitors**

a. Screen large database of small drug-like compounds for analogs of a known inhibitor
b. Combat limitations of known inhibitors by selecting for ideal physicochemical properties

**(iii) Computational framework for analysis of protein electrostatics**

a. Develop, implement, and test a computational framework in Python
b. Perform computational alanine scan mutagenesis to predict changes in free energy of association
c. Quantitatively compare electrostatic potentials resulting from protein structure

**(iv) Coevolution within regulators of the innate immune system**

a. Generate large library of sequences for complement control protein homologs with Sushi domains
b. Predict amino acid contacts to identify associations between contacts and protein function

**(v) Markov chain models for the effect of a risk-associated polymorphism in Factor H and FHL-1**

a. Estimate the potential of mean force from molecular dynamics for normal and risk isoforms
b. Build Markov chains describing the structural mechanism for pathogenesis of the risk isoform
c. Quantify association of heparin with multiple conformations of the FH risk and non-risk isoforms

**(vi) Structural features for activation in the G-protein coupled receptor CXCR4**

a. Sample structural changes in CXCR4 using accelerated molecular dynamics
b. Extract conformational features that distinguish between active and inactive states
c. Examine role of small-molecule IT1t in receptor inactivation

## l. International Experience

2016: Whitaker Summer Grant, Spanish National Cancer Research Centre, Madrid, Spain
2010: Maymester Study Abroad, Transport Phenomena in Biological Systems, University of Cambridge, England