UNIVERSITY OF CALIFORNIA
RIVERSIDE

Development and Applications of a Computational Framework for Protein and Drug Design

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

Doctor of Philosophy

in

Bioengineering

by

Chris Allen Kieslich

September 2012

Dissertation Committee:

- Dr. Dimitrios Morikis, Chairperson
- Dr. Thomas Girke
- Dr. David Lo

The Dissertation of Chris Allen Kieslich is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

This dissertation could not have been possible without the support and guidance of several key individuals whose aid allowed for the completion of these studies, and for this I will be forever grateful. I would especially like to express my gratitude to:

My dissertation committee, especially my adviser, Dr. Dimitrios Morikis, for their patience, guidance, and advice.

My labmates, especially Aliana López de Victoria and Ron D. Gorham Jr., for their support and all of the countless laughs that made our time together so enjoyable.

My funding sources, UCOP Tobacco Related Disease Research Program Dissertation Fellow, NSF/DAAD Central Europe Summer Research Institute Research Fellow, NSF East Asia and the Pacific Summer Institute Research Fellow.

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

Kieslich CA, Gorham RD, Morikis D (2011) Is the rigid-body assumption reasonable? Insights into the effects of dynamics on the electrostatic analysis of barnase-barstar. J Non-Cryst Solids 357: 707–716. doi:10.1016/j.jnoncrysol.2010.05.087.

Pyaram K, Kieslich CA, Yadav VN, Morikis D, Sahu A (2010) Influence of electrostatics on the complement regulatory functions of Kaposica, the complement inhibitor of Kaposi's sarcoma-associated herpesvirus. J Immunol 184: 1956–1967. doi:10.4049/jimmunol. 0903261.

Kieslich CA, Vazquez H, Goodman GN, López de Victoria A, Morikis D (2011) The effect of electrostatics on factor H function and related pathologies. J Mol Graph Model 29: 1047–1055. doi:10.1016/j.jmgm.2011.04.010.

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

DEDICATION

I dedicate this dissertation to my family whose love and support has made this all possible. I would especially like to thank my parents, for always believing in me and for pushing me to excel.

ABSTRACT OF THE DISSERTATION

Development and Applications of a Computational Framework for Protein and Drug Design

by

Chris Allen Kieslich

Doctor of Philosophy, Graduate Program in Bioengineering
University of California, Riverside, September 2012
Dr. Dimitrios Morikis, Chairperson

Interactions between biomolecules are essential to biological function, and fundamentally understanding the forces that drives these interactions is of great medical importance. Molecular modeling approaches represent a powerful set of tools for gaining a mechanistic perspective of biomolecular interactions at the atomic level. Due to the complexities of biological environments, a diverse set of molecular modeling methods are often needed to capture various aspects of biomolecular function. Here we present a computational framework that utilizes established molecular modeling methods, such as molecular dynamics, Poisson-Boltzmann electrostatics, and small-molecule docking, as well as includes novel tools for elucidating the role of electrostatics in protein association. Our novel computational tool, Analysis of Electrostatics Similarity Of Proteins (AESOP), utilizes theoretical mutagenesis, electrostatic clustering, and electrostatic free energies of association to evaluate the role of each charged residue in protein association. The AESOP framework has been applied to various biomolecular systems, including barnase-bartstar, the gold-standard system for protein electrostatics, as well as for the successful design of novel SUMO-4 substrate analogs. One biological system that is of key interest is the complement immune system, which is an ancient

component of innate immunity. The complement system is involved in the opsonization and clearance of foreign pathogens, and achieves its function through a cascade of protein-protein interactions largely driven by electrostatics. The key role of electrostatics in complement function is further evidenced by the existence of electrostatic “hot-spots” on the surface of complement proteins, and by the fact that pathogens utilize electrostatics in complement targeted evasion mechanisms. Over-activation of the complement system is implicated in numerous autoimmune and inflammatory diseases, and as a result anaphylatoxin receptor C5aR has become an important drug target. We have developed a conformationally-sampled pharmacophore model for known C5aR antagonists, which has utility in the design/evaluation of novel C5aR antagonists. We have also performed virtual screening based on a newly developed model of the interaction between C5aR and known potent antagonist PMX-53, and have identified a structurally diverse set of potential C5aR ligands. These studies define a computational framework that has utility in the analysis of many other protein targets.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF APENDICES

1. INTRODUCTION

1.1. BIOMOLECULAR INTERACTIONS DRIVE BIOLOGICAL FUNCTION

Biomolecular interactions are central to all biological function, from processes as fundamental as the absorption of oxygen by hemoglobin to as complex as cognitive perception. Biomolecules are dynamic in nature, and as Richard Feynman was famously quoted saying, “everything that living things do can be understood in terms of the jigglings and wigglings of atoms” [1]. The dynamics of biomolecules cover a large range of length and time scales, from motions as fast and short as bond vibrations up to large Brownian motions that occur during protein-protein association. Specificity is key for most biomolecular interactions [2], including biomolecules as small as serotonin and as well as large protein assemblies like the ribosome. Biomolecular specificity arises from structural and physicochemical complementarity, but is complicated by dynamic phenomena such as transient binding sites and conformational switching.

Understanding what drives interactions of biomolecules is of great medical importance due to its utility in studying disease pathogenesis and in the design of new therapeutics. One approach for studying biomolecular interactions is computational molecular modeling, which utilizes computer simulations to analyze the structure-dynamics-interactions-function relationships of biomolecules. One of the most accurate approaches for investigating biomolecular interactions is known as molecular dynamics (MD), and utilizes Newton’s laws of motion to simulate the thermal motions of proteins. However, due to the complexity and size of typical biomolecular systems, approximations must be made. Therefore, numerous other computational approaches have been developed to approximate specific aspects of the length

and time scales relevant to biomolecules [2-8]. Of these approaches, electrostatic calculations represent a powerful tool for elucidating complementarity in biomolecular interactions due to their efficiency and ability to be applied to large systems [3,9]. Docking calculations also investigate complementarity between biomolecules, but have the advantage of being able to predict the structure of the complex formed [4-6]. For most applications, a combination of computational approaches is needed to fully analyze the biomolecular interaction of interest, since each approach has its own advantages and limitations. Here we discuss a collection of work that takes advantage of the interplay between a combination of molecular modeling methods to address problems of biological/medical significance. The approach described henceforth represents a computational framework that can be applied to many other biomoelcular systems of interest.

1.2. MOLECULAR DYNAMICS SIMULATIONS OF PROTEIN DYNAMICS

Molecular dynamics (MD) is a type of computer simulation that utilizes classical mechanics to provide a time course trajectory for the thermal motions of a macromolecule [10]. These simulations approximate atoms as hard spheres and bonds as springs, where the spring constant would depend on the bond type. MD is dependent on a potential energy landscape, which is based on a force field that is composed of energy functions and parameters. Eq. 1-1 is an example of a potential energy function, *V(R)*, which is included in the CHARMM22 force field; here, the constants, *K*, represent the needed parameters.

$$V\left(\vec{R}\right) = \sum_{bonds} K_d\left(d - d_0\right)^2 + \sum_{angle} K_\theta\left(\theta - \theta_0\right)^2 + \sum_{dihedrals} K_\chi\left(1 + \cos\left(n\chi - \delta\right)\right) + \sum_{impropers} K_\phi\left(\phi - \phi_0\right)^2 + \sum_{nonbond}\left\{\varepsilon_{ij}\left[\left(\frac{R_{ij}^{\min}}{r_{ij}}\right)^{12} - \left(\frac{R_{ij}^{\min}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon_l r_{ij}}\right\} \quad \text{Eq. 1-1}$$

Where *d* is a bond distance, θ is a bond angle, χ is a dihedral (or torsion) angle, ϕ is an improper angle, ε_{ij} is the dielectric coefficient for atoms i,j, r_{ij} is the distance between atoms i,j, and q_i, q_j are the partial charges for atoms i,j respectively. Here K_d, K_θ, K_χ, K_ϕ, d_0, θ_0, ϕ_0, n, δ, and R_{ij}^{min} represent parameters and topologies that are specific to the type of atoms involved. This potential energy function can be converted into forces that would drive the motion of the atoms based on the first derivative of the potential by applying Eq. 1-2.

$$F_i = -\frac{dV}{dr_i} \qquad \text{Eq. 1-2}$$

The position of an atom after a finite time step, Δt, can be described by a Taylor series expansion as shown in Eq. 1-3.

$$x\left(t+\Delta t\right) = x(t) + \frac{dx(t)}{dt}\Delta t + \frac{d^2x(t)}{dt^2}\frac{\Delta t^2}{2} + \ldots \qquad \text{Eq. 1-3}$$

In this expression the first term refers to position, the second term refers to velocities resulting from kinetic energy, and the third term is acceleration and relates to the forces defined in Eq. 1-2. Several approaches have been developed for the iteration of Eq. 1-3 to generate MD trajectories, however currently the most widely used MD suite is Nanoscale Molecular Dynamics (NAMD) [11]. Not only does NAMD provide great scalability and speed but also includes a very useful graphical interface called Visual Molecular Dynamics (VMD) [12].

1.3. ELECTROSTATIC CALCULATIONS OF PROTEINS

When studying electrostatics, the fundamental expression that describes electrostatic potential, *V(r)*, is Coulomb's law, as shown by:

$$V(r) = \frac{1}{4\pi\varepsilon_0\varepsilon}\frac{q}{r} \qquad \text{Eq. 1-4.}$$

Coulomb's law defines electrostatic potential as a function of charge, *q*, distance, *r*, and a dielectric constant, ε, which is relative to vacuum permittivity, ε_0, and characteristic of the dielectric medium, as shown in Eq. 1-4. This expression is only appropriate for describing systems with a single dielectric medium, which is not the case when modeling proteins. Due to the complex nature of the protein environment in which there exists a hydrophobic core, with a dielectric constant near that of vacuum, surrounded by aqueous solvent, a more elaborate expression is needed. One such expression is the linearized Poisson-Boltzmann equation (LPBE), as shown in Eq. 1-5, which allows the calculation of electrostatic potential for systems of multiple dielectric constants, as well as accounts for the presence of mobile ions.

$$-\nabla\cdot\varepsilon(r)\nabla\phi(r) + \varepsilon_0\varepsilon(r)\kappa^2(r)\phi(r) = \frac{4\pi e^2}{\varepsilon_0 k_B T}\sum_{i=1}^{F} z_i\delta(r-r_i) \qquad \text{Eq. 1-5}$$

The LPBE defines electrostatic potential, ϕ, as a function of charge, *q*, dielectric coefficient, ε, and an ion accessibility function (κ), as described above. The charge term, *q*, accounts for fixed charges, such as point and partial charges, as well as mobile monovalent ions. The amount and type of mobile ions is controlled by the ion accessibility function, which is directly dependent on the ionic strength (*I*), as defined by:

$$I = \frac{1}{2}\sum_{i=1}^{M} z_i^2 n_i^0 \qquad \text{Eq. 1-6,}$$

$$\kappa^2(r) = \frac{4e^2 I}{\varepsilon_0\varepsilon k_B T} \qquad \text{Eq. 1-7.}$$

The current gold standard for calculating electrostatic potentials at atomic resolution using the Poisson-Boltzmann method is the Adaptive Poisson-Boltzmann Solver (APBS) [3]. The basic

strategy used by APBS is to place the protein in a three-dimensional grid, and then to assign the values for the charge, dielectric coefficient, and ionic strength function at every grid point. Two surfaces, known as the epsilon and kappa surfaces, are generated around the protein or protein complex to define the boundaries of the protein environment. The epsilon surface, which defines the boundary for the dielectric coefficient, is defined by rolling a sphere, typically the size of a water molecule, across the van der Waals surface of the protein and defining a new surface based on the center of this sphere. The kappa (or ion accessibility) surface is similarly defined, however a

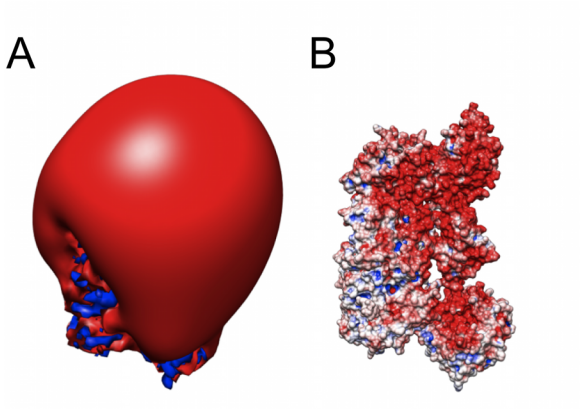


Figure 1-1 *Electrostatic potential representations of complement C3b:* (A) isopotential contour and (B) surface projection.

sphere the size of an ion is used rather than a water molecule. Following the solution of the LPBE, a three-dimensional array describing the spatial distribution of the electrostatic potential is obtained, which can be visualized using an isopotential contour (Figure 1-1). The LPBE electrostatic potentials can also be used to calculate electrostatic free energies of a specific state according to:

$$G_{electro} = \frac{1}{2}\sum q_i \phi_i \qquad \text{Eq. 1-8,}$$

such free energies can be used for the study of protein stability/association, including their pH dependencies.

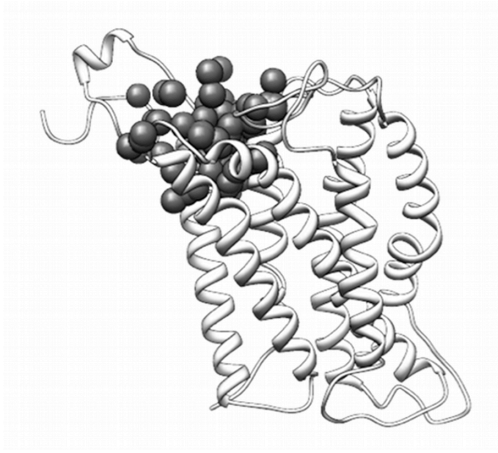


Figure 1-2 *Binding site of C5aR with representative spheres for DOCK6 docking calculations.*

1.4. SMALL MOLECULE DOCKING FOR DRUG DESIGN

The major question in computational drug design typically regards whether or not a specific ligand of interest will bind in a given binding pocket, and if so with what affinity. In the absence of available structural data for a given protein-ligand complex, the primary approach that can be used to answer such questions is referred as docking. Docking algorithms generate potential poses (orientations) of a protein-ligand complex by rotating and translating the ligand across the proposed binding site and then use a scoring function to rank the resulting poses. In order to dock a compound, the docking algorithm first generates a simplified representation of the binding site for efficiency, such as the spheres used by DOCK6 [4] as shown by Figure 1-2. Additionally, one of the major difficulties in docking that must be considered is how to handle receptor/ligand flexibility. Flexible receptor approaches, such as that employed in AutoDock [6], provide more realistic docking conditions, but can be computationally expensive. Another approach for incorporating receptor flexibility is to perform the docking procedure using

multiple structures from molecular dynamics simulations to sample the conformational space of the receptor [5].

Ligand flexibility is more widely used and as a result numerous algorithms have been developed, such as the "anchor-and-grow" algorithm of DOCK6 [4]. The DOCK6 algorithm first identifies the largest rigid substructure of the ligand, which serves as the "anchor" (often a ring), and then produces orientations for the "anchor" within the binding site. These orientations are generated by matching the heavy atoms of the anchor with the centers of as many spheres (Figure 1-2) as possible. In the DOCK6 approach, all possible orientations are scored using a grid-based molecular mechanics scoring function, which incorporates van der Waals and electrostatic components. The orientations are ranked according to their scores, spatially clustered with respect to heavy atom root mean squared deviations, and prioritized. The algorithm also identifies flexible bonds within the ligand structure, which are subsequently used to partition the remainder of the compound into rigid segments. The ligand is then grown from the best anchor orientations based on the addition of segments using the identified flexible bonds. From all of the anchor orientations the best scoring pose is selected.

Virtual screening approaches, which parallel high-throughput screening drug discovery, utilize docking methods to evaluate databases of compounds in order to identify potential new ligands for a given target. One popular ligand database for virtual screening studies is the ZINC database [13], which currently contains over 17 million purchasable compounds. In addition to the docking algorithm, the prime consideration that must be made when performing virtual screening is how to handle the immense computation required in testing such a large number of compounds; therefore, access to computer clusters or grid-computing [14] technologies is essential.

1.5. OVERVIEW

This work describes the design and applications of a computational framework for analyzing biomolecular interactions, which has utility for protein and drug design. We utilize a combination of established molecular modeling methods, including molecular dynamics, Poisson-Boltzmann electrostatics, and small-molecule docking, as well as develop novel electrostatic similarity methods. Section 2 describes the design of our computational tool, AESOP, which uses theoretical mutagenesis in combination with electrostatic clustering and free energy calculations, as well as ionization properties, to elucidate the role of electrostatics in protein association; this section includes the application of the AESOP framework to the test system barnase-barstar, as well as for the design of novel SUMO-4 substrate analogs. Section 3 utilizes the methods of the AESOP framework, but investigates the role of electrostatics in key proteins involved in the activation and regulation of the complement immune system. Section 3 also introduces a new methodological advance for the AESOP framework, which is used in identifying electrostatic "hot-spots" on the surface of complement fragment C3d in a study of the evolution of C3d to assume dual functions. Section 4 includes drug design studies targeting complement receptor C5aR, a key drug target for inflammatory and autoimmune diseases. We employ ligand-based approaches, by developing pharmacophore models for known C5aR antagonists, as well as perform ligand-based and receptor-based virtual screening.

2. DEVELOPMENT OF THE AESOP COMPUTATIONAL FRAMEWORK FOR ELECTROSTATICS-BASED PROTEIN DESIGN

2.1. ELECTROSTATIC CALCULATIONS IN PROTEIN INTERACTIONS

2.1.1. Two-step model of association

For studies focused on electrostatically-driven protein association, McCammon and coworkers have proposed a two-step association model that is used to deconstruct association, as well as the interactions that drive it [15,16]. In this two-step association model, the first step is known as recognition and consists of the initial collision of the two proteins free in solution through diffusive motion. Recognition is driven and or accelerated by long-range electrostatic interactions, and results in a weak non-specific encounter complex. This is then subsequently followed by the binding step, where short to medium range electrostatic interactions, van der Waals interactions, as well as entropic effects, drive the formation of a specific final complex. This model holds true for interactions between highly charged proteins and ligands, and is essential in understanding why mutations away from the binding interface can affect binding. An illustration of the two-step binding models is presented in Figure 2-1A. According to this model, long range electrostatic interactions are vital to the recognition or association phase of binding, and therefore, mutations that alter the spatial distribution of electrostatic potential can affect binding, even if the residue is away from the binding interface. This logic is different from traditional thinking, and is evidence for the need for a thorough electrostatic analysis.

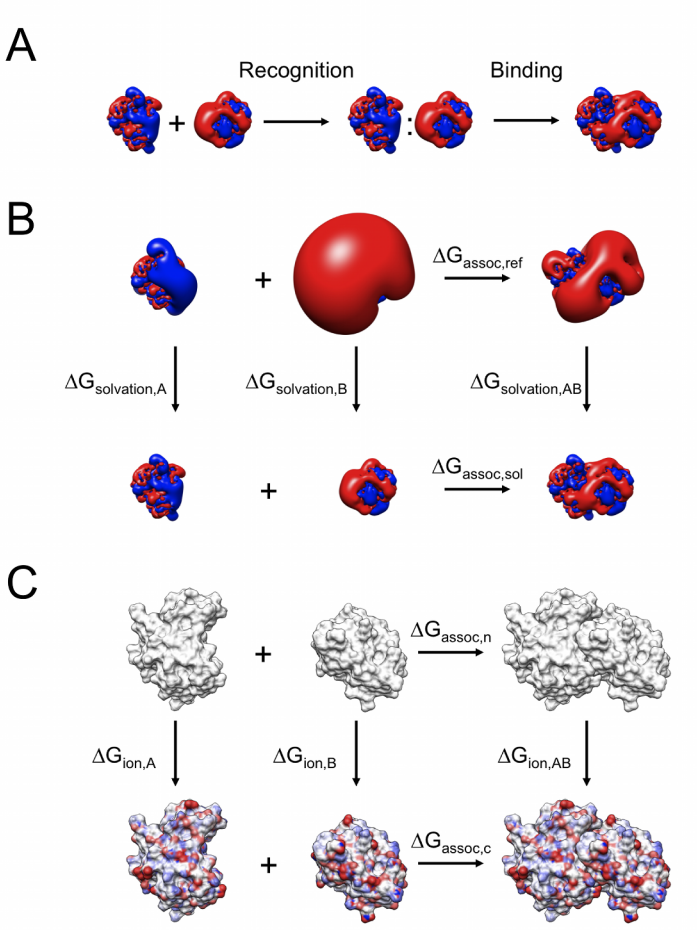


Figure 2-1 *Illustrations of the decomposition of the role of electrostatics in protein association.* (A) Schematic of the two-step model of association illustrated using isopotential contours of barnase-barstar. (B) Thermodynamic cycle for calculating solvation free energies of association illustrated using isopotential contours of barnase-barstar. The top process corresponds to association in the reference state and bottom process corresponds to association in the solution state (described in text). The color scheme of the isopotential contours is red for negative potential, and blue for positive potential. (C) Thermodynamic cycle for ionization free energies of association for barnase-barstar. White molecular surfaces indicate neutral proteins, while the red and blue surfaces are colored based on positive and negative atomic charge, respectively, indicating a charged protein state.

2.1.2. Electrostatic similarity calculations

Based on the two-step model for association, it is intuitive that proteins with similar electrostatic properties, or more specifically similar spatial distributions of electrostatic potential, are likely functional homologues. Methods for quantifying electrostatic similarity have

thus been of great interest for fields such as protein engineering and drug design. Early work in this field was centered in pharmacochemistry, where molecular similarities were used to compare the electron densities of small organic compounds based on quantum mechanical calculations. Carbo and coworkers [15,17] developed a similarity index (SI) for comparing molecular density functions, which has the advantage of being firmly grounded in quantum mechanics,

$$SI_{Carbo} = \frac{\int_V \rho_A \rho_B \, dV}{\left(\int_V \rho_A^2 dV\right)^{1/2} \left(\int_V \rho_B^2 dV\right)^{1/2}} \qquad \text{Eq. 2-1.}$$

Where, ρ_A and ρ_B represent density functions of the two molecules to be compared. Hodgkins and coworkers [18,19] were first to propose the use of a similarity measure for the comparison of molecular electrostatic potential or electric field. The Hodgkin electrostatic similarity index (ESI), is based on a dot-product and produces values from -1 to 1,

$$ESI_{Hodgkin} = \frac{2\sum \phi_A(i,j,k)\phi_B(i,j,k)}{\sum \phi_A(i,j,k)^2 + \sum \phi_B(i,j,k)^2} \qquad \text{Eq. 2-2.}$$

Where a value of 1 indicates identity, a value of -1 indicates anticorrelation, and 0 indicates no similarity. In this expression, the density functions have been replaced with the electrostatic potentials, ϕ_A and ϕ_B, and normalization is achieved using the sum of the self dot-products. Summations over all grid points *(i,j,k)* are performed, since space is discretized for numerical calculations of electrostatic potential. Two alternative measures, one by Reynolds and coworkers [20] and another by Petke [21], have also been proposed, both of which provide a linear relationship with respect to the proportionality of the compared electrostatic potentials,

$$ESI_{Reynolds} = \left[1 - \frac{1}{N}\sum_{N}\frac{|\phi_A(i,j,k) - \phi_B(i,j,k)|}{\max\left(|\phi_A(i,j,k)|, |\phi_B(i,j,k)|\right)}\right] \quad \text{Eq. 2-3,}$$

$$ESI_{Petke} = \frac{1}{N}\sum_{N}\frac{2\phi_A(i,j,k)\phi_B(i,j,k)}{\phi_A(i,j,k)^2 + \phi_B(i,j,k)^2} \quad \text{Eq. 2-4.}$$

Both the Reynolds and Petke ESIs utilize an average similarity that is locally normalized at each grid point. All three of these early ESIs (Hodgkins, Reynolds, and Petke) compared only potential values outside the van der Waals surface of the molecules. Wade and coworkers have since extended the ESI originally proposed by Hodgkins et al to the analysis of protein interactions [22,23]. However, Wade et al introduced the concept of a "skin" region, or a thin region of chosen thickness surrounding the protein, to account for the intricacies that arise when applying ESI methods to proteins. The skin region is used to focus the comparison to regions of functional importance, and to exclude large potential values that arise in the protein interior. ESI values can also be converted into electrostatic similarity distances (ESD), which allow for the application of clustering algorithms. ESD measures simply require that values of zero indicate identity, whereas increasing values indicate increasing dissimilarity. Three examples of ESD measures derived from Eq. 2-2, Eq. 2-3, and Eq. 2-4, are as follows:

$$DP = \sqrt{1 - \frac{2\sum\phi_A(i,j,k)\phi_B(i,j,k)}{\sum\phi_A(i,j,k)^2 + \sum\phi_B(i,j,k)^2}} \quad \text{Eq. 2-5,}$$

$$LD = \frac{1}{N}\sum_{N}\frac{|\phi_A(i,j,k) - \phi_B(i,j,k)|}{\max\left(|\phi_A(i,j,k)|, |\phi_B(i,j,k)|\right)} \quad \text{Eq. 2-6,}$$

$$LDP = \sqrt{1 - \frac{1}{N}\sum_{N}\frac{2\phi_A(i,j,k)\phi_B(i,j,k)}{\phi_A(i,j,k)^2 + \phi_B(i,j,k)^2}} \quad \text{Eq. 2-7.}$$

2.1.3. Electrostatic free energies of association

To quantify the electrostatic contributions to protein association, electrostatic free energies of association can ultimately be calculated, and such free energies have been shown to serve as good predictors of binding ability for highly and oppositely charged proteins [24-26]. When calculating electrostatic free energies of association, it is often of interest to incorporate solvation and other effects using a thermodynamic cycle such as that in Figure 2-1B. Solvation free energies of association, $\Delta\Delta G_{solv}$, which account for both solvation and association, can be calculated according to this thermodynamic cycle and the following expression,

$$\begin{aligned}\Delta\Delta G_{solv} &= \Delta G^{AB}_{solvation} - \Delta G^{A}_{solvation} - \Delta G^{B}_{solvation} \\ &= \Delta G_{assoc,sol} - \Delta G_{assoc,ref}\end{aligned} \qquad \text{Eq. 2-8.}$$

This thermodynamic cycle accounts for association in a uniform dielectric reference state, $\Delta G_{assoc,ref}$, without the presence of counterions, as well as a solvated state, $\Delta G_{assoc,sol}$. Additionally, the vertical processes represent solvation, $\Delta G_{solvation,(A,B,or\ AB)}$, of the two free components, as well as the complex, and aid in removing grid artifacts as has been discussed extensively [24].

2.1.4. Protein ionization

When performing electrostatic calculations extra consideration needs to be made in choosing protonation states of ionizable residues, since these charged residues contribute most to the electrostatic properties of proteins. Additionally, ionization properties of proteins drive pH dependent phenomena such as enzymatic catalysis, stability, and conformational switching. With this in mind, extensive research has gone into understanding the titration properties of proteins, which has been reviewed in great detail previously [27,28]. The protein ionization

state, or ionization vector, defines which titratable groups of the protein are in the charged or neutral state. This ionization state is dependent on pH, and since charged residues contribute to the electrostatic properties of a protein, ionization states are also very important for function and stability. Accurately determining ionizations states of protein is very important to understanding biological function, but calculations are quite complex. In a small molecule containing a single titratable site, ionization states can be calculated rather simply, based on acid-base dissociation. The protein environment is complex containing numerous titratable groups, with the protonation state of each group dependent on all other groups, resulting in an astronomical number of possible protein ionization states for even an average sized protein. To overcome this complex problem, methods that decompose the contributions of electrostatics in protein ionization and utilize statistical approximations to sample possible protein ionization states have been developed [27,28]. One such approach defines the free energy, *G*, of a particular ionization state based on contributions from two sets of free energies, self- and interaction energies, which are calculated based on the Poisson-Boltzmann method [27]:

$$G(x'_1, x'_2, ..., x'_M, pH) = 2.303RT\sum_{i=1}^{M} x'_i\gamma_i(pH - pK^m_{a,i}) + \sum_{i=1}^{M} x'_i\Delta\Delta G_{self,i} + \sum_{i=1}^{M-1}\sum_{j=i+1}^{M} x'_i x'_j \Delta\Delta G^{\text{int}}_{i,j} \quad \text{Eq. 2-9,}$$

$$\Delta\Delta G_{self,i} = \Delta G_{m\to p,c} - \Delta G_{m\to p,n} \quad \text{Eq. 2-10,}$$

$$\Delta\Delta G^{\text{int}}_{i,j} = \Delta G_{n\to c,i,j} - (\Delta G_{n\to c,i} + \Delta G_{n\to c,j}) \quad \text{Eq. 2-11.}$$

In Eq. 2-9 the ionization state of the protein is defined by *x'*, where a value of 1 indicates a titratable group is charged while 0 represents a neutral group. Additionally, the charge of each titratable group is defined by *γ.* Self-energies account for the penalty of placing each individual

titratable group into the protein environment, and are calculated according to the thermodynamic cycle of Figure 2-2 and Eq. 2-10. Interaction energies (Eq. 2-11), on the other hand, account for all possible Coulombic interactions between all possible pairs of ionizable groups within the protein.

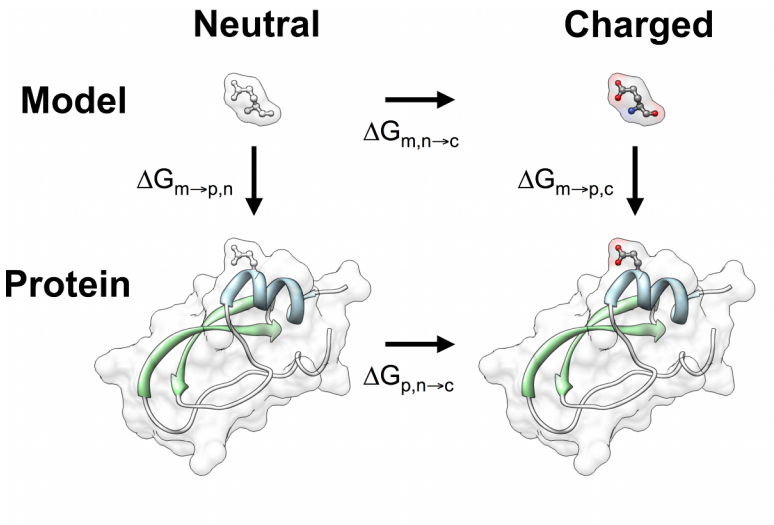


Figure 2-2 *Thermodynamic cycle for calculating ionization self-energies.* White ball-and-stick residue indicates ionizable residue in neutral state, while CPK coloring indicates the charged state.

Based on the self- and interaction energies, as well as Eq. 2-9, any ionization state of the system can be evaluated; however, a numerical approximation, such as clustering or Monte Carlo simulation, must used to identify a subset of lowest energy ionization states for a given pH [27]. From such calculations, protein titrations and accurate predictions of protein pK_as are possible [29-34], giving further insight into the electrostatic environment of proteins. Also, ionization free energies can be used to compute stability curves, such as pH-dependent free energies of association, as described by the thermodynamic cycle of Figure 2-1C.

2.2. PERTURBATIVE DESIGN OF KAPOSICA VARIANTS

2.2.1. Introduction

In order to overcome an immune response, viruses and other pathogens exhibit numerous strategies to evade immune recognition inside the body. One example includes immune system mimicry, in which a virus secretes molecules that function in a similar manner to native immune regulator proteins, thereby curtailing immune response against the virus itself. VCP, SPICE, and Kaposica are viral or viral-based proteins that consist of four CCP modules, which have similar structure to those of CR2 and FH, but with diverse sequences and physicochemical properties [35-37]. These three proteins are used by their respective viruses to trick the immune system by mimicking native regulators of the complement system and thus inhibiting the function of complement.

VCP and SPICE have high sequence identity, differing only by 11 amino acids out of 244, but their complement inhibitory activity differs by up to 1000-fold, with SPICE being more active. SPICE possesses higher net positive charge than VCP, as two glutamic acids in VCP are replaced by lysines. The location of these lysine residues in the second module (CCP2) results in a characteristic electrostatic potential distribution, which is distinct from that of VCP. An early qualitative study using Coulombic potentials demonstrated that systematically mutating ionizable amino acids in VCP in order to manipulate the spatial distribution of electrostatic potentials yields VCP-like proteins with variable binding and inhibitory activities [35]. A subsequent quantitative theoretical study used PB electrostatics and molecular dynamics simulations to provide a more complete picture on the interplay between dynamics and electrostatics towards developing a working hypothesis on the mechanism of function of VCP

and SPICE [36]. The correlation between the spatial distributions of electrostatic potentials and binding and inhibitory activities was excellent, demonstrating that CCP-containing regulators of the complement system can be designed by rationally manipulating charge. The most remarkable finding in these studies was that replacement of the two glutamic acids of VCP's CCP2 with lysine results in a protein with SPICE-like binding and inhibitory activities (with some subtleties discussed in [35,36]).

The evolution from the first VCP study [35] to the subsequent VCP/SPICE study [36] provided the transition from qualitative visual inspection to quantitative electrostatic and dynamic analysis, which paved the way for the design of complement regulators with tailored physicochemical/dynamic properties and inhibitory activities. Further quantification was initiated in the second study by introducing clustering analysis of electrostatic potentials [36], which later formed the basis for the development of the sophisticated AESOP framework [25,38]. Based on the success of the VCP/SPICE electrostatic analysis, this study aims to utilize these methods for computational design involving Kaposica, a viral protein-based complement inhibitor [37]. The basic design principle is to perturb the electrostatic character of various structural elements/domains of Kaposica to investigate the role of electrostatics in Kaposica's inhibition of the complement system.

2.2.2. Methods

Our atomic-level calculations are possible when three-dimensional structures are known at high resolution. However, the structure of Kaposica is yet to be solved; therefore, our initial step was the generation of a Kaposica homology model. The three-dimensional structure of VCP, a known homologue, has been experimentally determined [Protein Data Bank Code: 1G40 [39]]

and was used for the homology modeling of Kaposica. To generate the homology model for Kaposica, we first obtained the amino acid sequence alignment for the comparison of VCP, our template, and Kaposica, our target, using ClustalW [40]. The generated amino acid sequence alignment was manually optimized to enhance the level of sequence identity. The modified sequence alignment and template structure, 1G40, were imported into a DeepView, Swiss Institute of Bioinformatics (http://spdbv.vital-it.ch) project and submitted to the Project Mode of the SWISS-MODEL online server [41]. The generated Kaposica model was inspected for van der Waals clashes, secondary structure quality, and disulfide bond correctness. The Kaposica model was created with all four CCP modules intact. Models for each CCP module alone were extracted from the whole Kaposica model. In this study, the CCP module is defined as the sequence between and including the first and fourth conserved cysteines. The theoretical mutations were generated by importing coordinates, corresponding to either the entire Kaposica molecule or individual CCP modules, into Deep View and selecting the mutation site and type. Because multiple rotamers can be generated for each mutation, the rotamer that created the most favorable contacts, primarily hydrogen bonds, while minimizing the number of van der Waals clashes, was selected.

Electrostatic potential calculations were made with the Adaptive Poisson–Boltzmann Solver [3], which uses a grid-based method to solve the linearized Poisson–Boltzmann equation. Prior to electrostatic potential calculations, partial charges and van der Waals radii were assigned using the software PDB2PQR [42] and the PARSE force field [43]. A dielectric coefficient of 2 was used for the protein, whereas a dielectric coefficient of 78.5 was used for the solvent. The calculations were carried out at an ionic strength corresponding to a 0 mM concentration, assuming +1/-1 charges for the counterions. The dielectric surface was defined by selecting the

contact surface, using a sphere with a probe radius of 1.4 Å. The ion accessibility surface was defined using a sphere with a probe radius of 2.0 Å. Because calculations were performed on the entire Kaposica macromolecule, as well as individual CCP modules, the physical dimensions of the box used for each set of calculations were set to different values to ensure that the largest magnitude values were captured, while still providing maximum resolution. The grid size was 129 × 129 × 129 points. A box size of 130 Å × 120 Å × 190 Å was used for calculations of whole Kaposica, whereas a box size of 100 Å × 100 Å × 140 Å was used for calculations of individual CCP modules. The spatial distributions of electrostatic potential were visualized and plotted using isopotential contours using the molecular graphics program Visual Molecular Dynamics [[12]; Theoretical and Computational Biophysics Group, Urbana, IL; www.ks.uiuc.edu/Research/vmd].

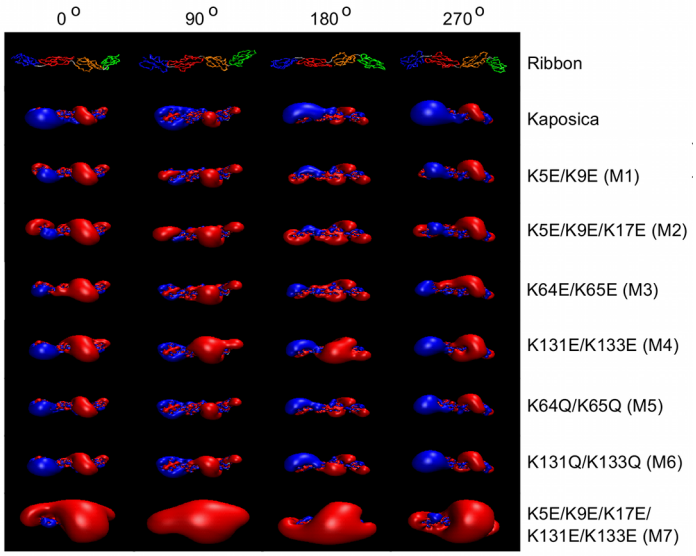


Figure 2-3 *Isopotential contours of Kaposica and its mutants, illustrating the spatial distribution of electrostatic potential.* The electrostatic potentials of Kaposica and seven mutants were calculated at 0 mM ionic strength and are presented in four orientations. The color code for the isopotential contours is blue for positive and red for negative electrostatic potential. Isopotential contours are plotted at ± 1 k_BT/e. A ribbon representation of Kaposica is also included for reference, and the four CCP modules are colored as follows: CCP1 is blue, CCP2 is red, CCP3 is orange, and CCP4 is green.

2.2.3. Results and Discussion

There is a long-standing view in the field of complement biology that ionic interactions play a vital role in complement regulation. This premise is primarily based on salt-dependent binding of complement regulators to C3b/C4b and mutagenesis data demonstrating that acidic residues on C3b/C4b, and basic residues on complement regulators, are important for the interaction between these proteins [44-47]. More recently, on the basis of mutagenesis and electrostatic modeling studies of VCP and SPICE, it has been proposed that overall positive potential on these viral regulators directs the initial recognition (Figure 2-1A) of C3b (termed step I). Recognition is then followed by binding, which involves electrostatic interactions (hydrogen bonds, salt bridges, and medium/weak Coulombic interactions), hydrophobic and van der Waals interactions, and entropic effects such as solvent exclusion and local structural rearrangements at the binding sites (termed step II) [35,36]. According to this association model, step I influences predominantly the k_{on} rate and the formation of a weak, nonspecific encounter complex, whereas step II influences predominantly the k_{off} rate and the formation of a strong, specific bound complex [29,36]. In the earlier studies of VCP/SPICE, correlation was found between the reduction of negative character of the CCP2/3 interface and C3b binding, inhibition of AP, and CFA [35]. Indeed, because of spatial cancellation of opposite electrostatic potentials, which is modulated by the dynamic character of VCP/SPICE, an increase of positive potential in one module may contribute to the decrease of negative potential in a nearby module, depending on potential magnitude [36].

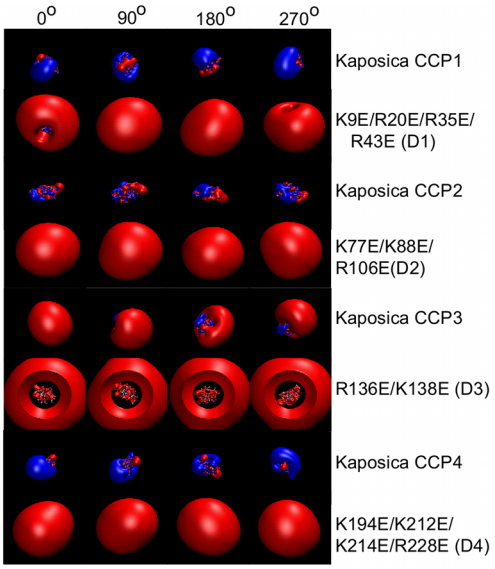


Figure 2-4 *Isopotential contours illustrating the spatial distributions of electrostatic potential of different CCPs of Kaposica and its mutants in isolation.* The electrostatic potentials of individual wild-type CCP domains and the corresponding charge-reversal mutants were calculated at 0 mM ionic strength and are presented in four orientations. The color code for isopotential contours is blue for positive and red for negative electrostatic potential. Isopotential contours are plotted at ± 1 k_BT/e.

We therefore asked whether the overall or localized electrostatic potential guides the complement regulatory activities of Kaposica. To answer this, we applied structural perturbations by mutating positively charged residues in a homology model of Kaposica, as illustrated in the top of Figure 2-3. We designed two sets of mutants; the first (Figure 2-3) was designed to reduce or eliminate positive potential in the whole molecule, and the second (Figure 2-4) was designed to delineate the role of electrostatic potential of the individual CCP modules. For the mutants of Figure 2-4, the minimal number of mutations for charge reversal was identified for CCP1, CCP2, and CCP4, and for charge enhancement for CCP3. Iterative design was also performed using a combination of computational and experimental data. After theoretical design, Kaposica and the designed mutants were then expressed in *E. coli*, and their activities were compared to determine the dependence of the activities on the electrostatic character of the molecule. The electrostatic calculations better depict binary interactions

between Kaposica and C3b/C4b, which can be more directly validated by SPR data than by CFA and DAA data [37]. When other molecules are involved (factor I or convertase components), additional complexity is introduced, which cannot be incorporated into our modeling at this time because of lack of three-dimensional structures of the multicomponent assemblies. For example, the SPR-based binding assay showed remarkable reduction in the binding of M7 mutant to C4b, which carries the most negative electrostatic potential (Figure 2-3) [37]. Significant decrease in binding to C4b was also observed in mutants with altered electrostatic potential in the linker region (M3–M7) and CCP1 (M2); M1, with two-residue replacement in CCP1, did not show any effect on C4b binding. Among domain electrostatic potential modulation mutants, D1–D3 showed significantly reduced binding, whereas D4 showed marginally reduced binding to C4b. These data together point out that positive potential around CCP1 and the linkers between CCPs 1–2 and 2–3 of Kaposica are vital for binding to C4b. These studies collectively illustrate examples of cross-talk between theory and experiment, in which theoretical calculations can be used in a predictive manner to guide the design of experiments.

2.3. DESIGN OF THE AESOP FRAMEWORK

2.3.1. Motivation

Based on the two-step association model (Figure 2-1A), mutations of ionizable residues away from the binding interface can affect protein association, by altering recognition. Therefore, from the point of view of protein design, mutations of ionizable residues provide an interesting advantage, since they can affect both the overall protein electrostatic potential (global), as well as more specific intermolecular interactions (local). Following the success of the VCP/SPICE/Kaposica studies [35-37], we desired to create a method, based on the perturbative design approach, which could systematically evaluate the role of each ionizable residue in protein association and stability. The goal was to develop a framework for identifying electrostatic "hot spots" and optimizable sites, in order to aid in the design of new protein analogs with customized electrostatic properties. The resulting computational framework is referred as Analysis of Electrostatic Similarities of Proteins (AESOP), and utilizes theoretical mutations, Poisson-Boltzmann electrostatics, and electrostatic similarity clustering to evaluate the role of electrostatics in protein association. Additionally, the AESOP framework also contains utilities for analyzing the pH dependent properties of proteins, which were developed in collaboration with Jan Antosiewicz of University of Warsaw.

2.3.2. Computational workflow of AESOP

The general workflow of the AESOP framework, as illustrated by Figure 2-5, starts with a protein complex and the generation of electrostatically perturbed protein analogs. The standard analysis involves the use of theoretical alanine-scan mutagenesis in which each ionizable

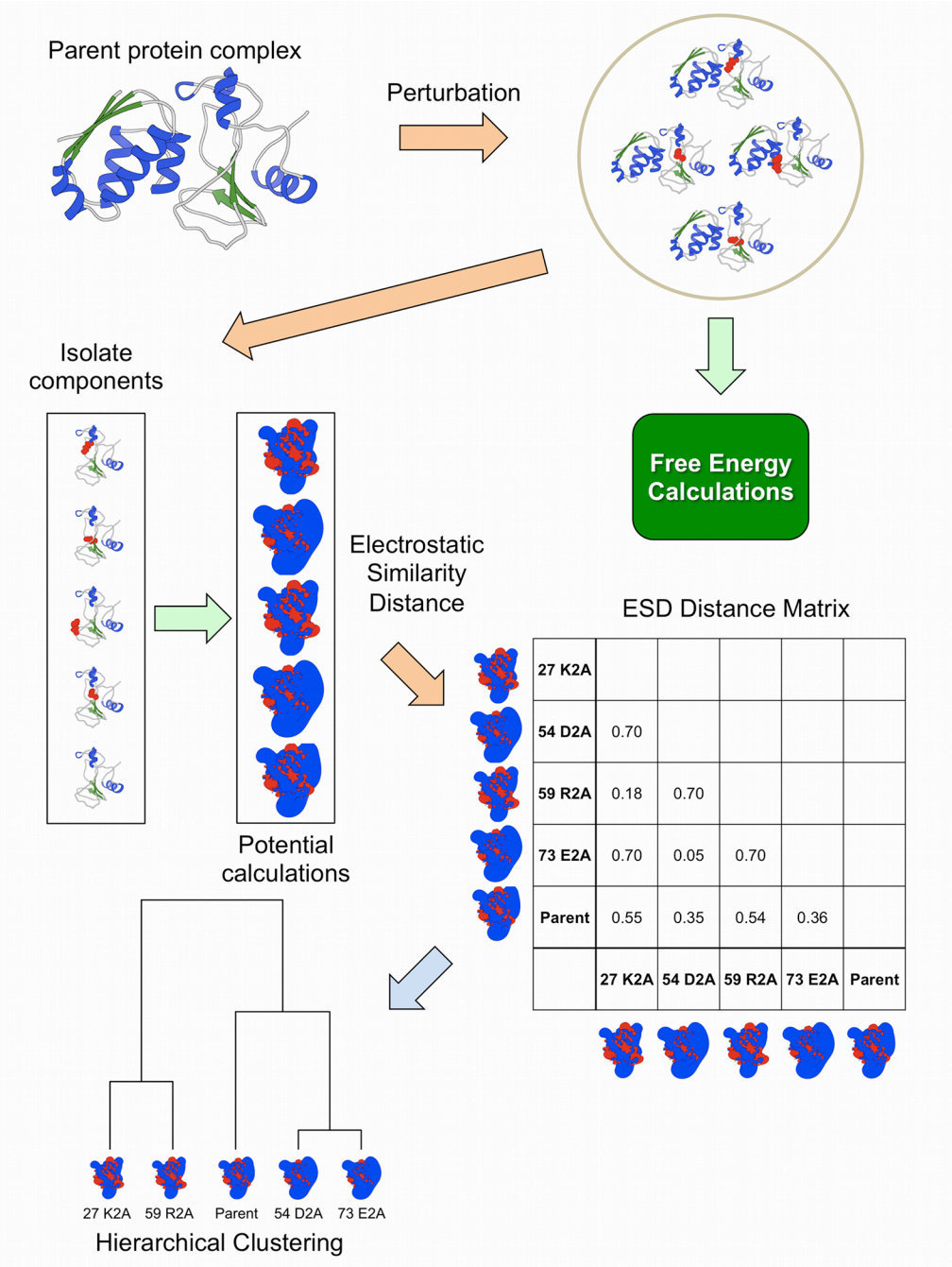


Figure 2-5 *Workflow of the AESOP Framework.* Arrow color indicates which utility is responsible for performing the various steps: orange, AESOP; green, APBS [3]; blue, R functions [48]. In ribbon model, red residues indicate perturbation site.

(charged) residue is replaced by alanine, one at a time. Additional perturbation methods, such as charge reversals and mutation permutations, are also included in the AESOP framework. Following the generation of perturbed mutant structures, there are two types of electrostatic calculations that are used to quantify the effects of the perturbations: electrostatic similarity clustering and electrostatic free energies of association. Electrostatic similarity clustering depicts the global effects of perturbations and relates to the recognition step of association, while electrostatic free energies of association capture both local and global effects, and therefore correspond to both recognition and binding.

To perform electrostatic similarity clustering, the mutated proteins are first isolated from the complex, and the spatial distributions of each mutant are calculated using the Poisson-Boltzmann method. A distance matrix is then populated by ESD values, according to Eq. 2-5, Eq. 2-6, or Eq. 2-7, quantifying the similarity between spatial distributions of electrostatic potential for each possible pair of mutants. Clustering methods are applied to the ESD distance matrix to classify the perturbations according to their effects on the spatial distributions of electrostatic potential; one such method, is hierarchical clustering [49], which produces a dendrogram tree as shown on the bottom of Figure 2-5. Electrostatic free energies of association are typically calculated based on a thermodynamic cycle that accounts for solvation effects by including association in both a solvated state as well as a reference state (Figure 2-1B). All electrostatic calculations in the AESOP framework (green arrows, Figure 2-5) are performed using the Adaptive Poisson-Boltzmann Solver (APBS) [3]. It should also be noted that the AESOP protocol could be applied in the absence of a protein complex, to families of either homologous proteins or alanine-scan mutants. In general, the AESOP was developed with flexibility in mind in order to allow for customized electrostatic analyses.

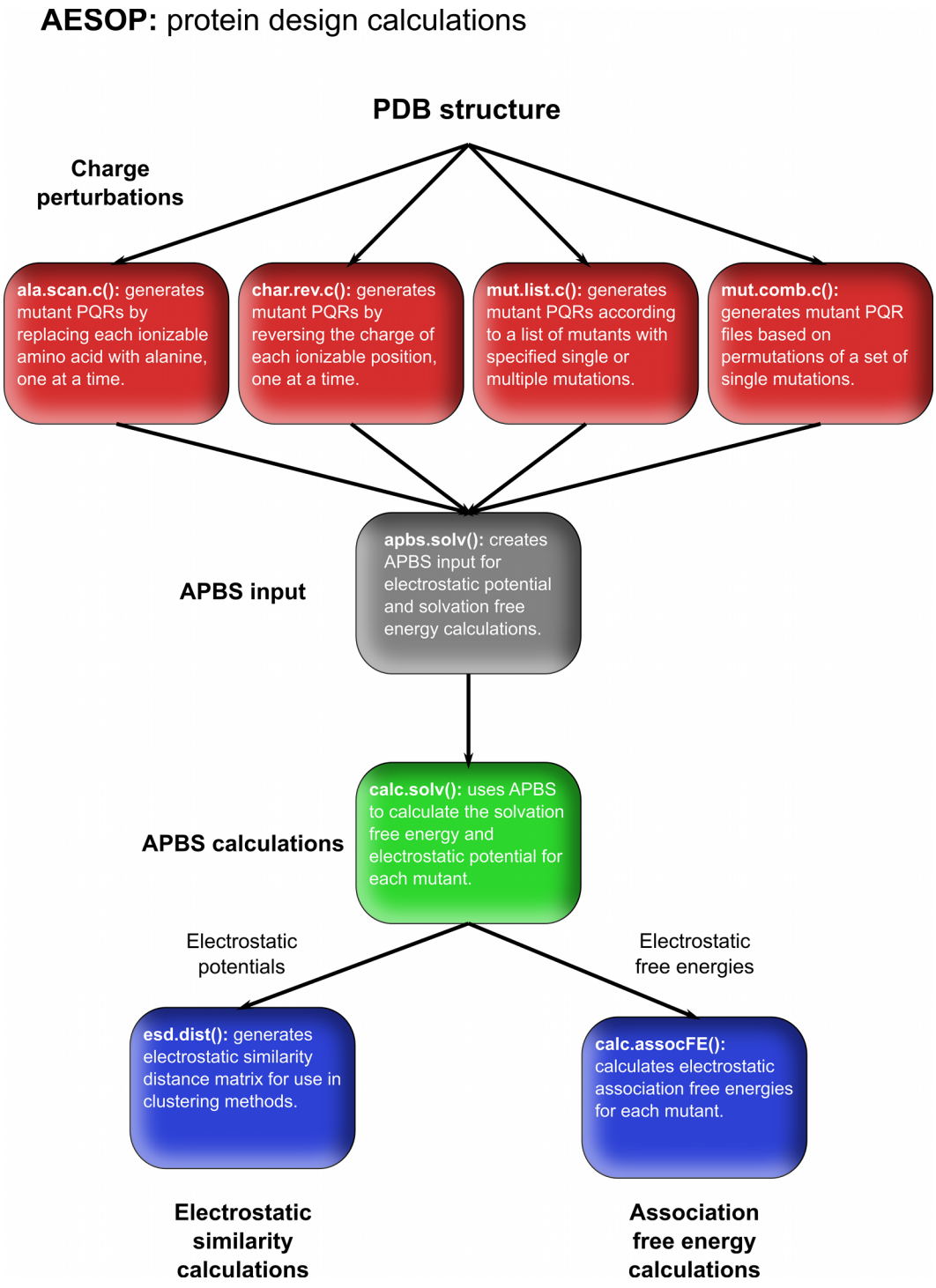


Figure 2-6 *Schematic of AESOP functions for electrostatics-based protein design.*

2.3.3. AESOP library structure and usage

The AESOP framework has been written using the R statistical language [48] and utilizes the Bio3D [50] library to handle PDB structure data. The AESOP framework relies on three external software packages: (i) SCWRL4 [51] for non-alanine mutations, (ii) PDB2PQR [42] for structure preparation and parameterization, and (iii) APBS [3] for all electrostatic potential and free energy calculations. The functions of the AESOP framework have been designed to be flexible and interchangeable, however, the framework contains two platforms, one for electrostatics-based protein design (Figure 2-6) and another for pH-dependent calculations (Figure 2-7). At this point in time, the two platforms have been developed independent of one another, but the combination of both platforms could have great utility for pH-dependent protein design in the future. As illustrated by Figure 2-5, all AESOP calculations begin with a PDB file which is read into the R environment using the read.pdb() function of Bio3D, which allows usage of Bio3D functions when making complex atom/residue selections and other manipulations based on a PDB file. The next step in all AESOP calculations, both design and pH-dependent, is the generation of PQR files, which contain the protein coordinates, including hydrogen atoms, as well as atomic radii and partial charge parameters. However, the AESOP function used in PQR generation is what mostly defines the type of analysis to be performed.

In the electrostatic-based design platform (Figure 2-6), the first primary step is the generation of electrostatically perturbed protein mutants. The AESOP framework includes 4 methods for introducing charge perturbations: ala.scan.c(), char.rev.c(), mut.list.c(), and mut.comb.c(). ala.scan.c() performs a theoretical alanine-scan by replacing each charged residue with alanine, one at a time. These alanine mutations are performed by an AESOP function called mut2ala(), which simply truncates the residue down to the C_β atom and generates the

appropriate hydrogen geometry. In addition to generating a directory containing the alanine-scan mutants, ala.scan.c() also returns a list of the mutants generated and the directories to which they were written. char.rev.c() is similar to ala.scan.c, however, instead of alanine mutations, the charge of each charged site is reversed one at a time by replacing every basic residue with glutamic acid and every acid residue with lysine. The charge reversal mutations are introduced using SCWRL4 [51], which optimizes the side chain rotamer upon mutation. The remaining two functions are somewhat different in that only specified positions are mutated, not every charged residue as in ala.scan.c() and char.rev.c(). The mut.list.c() function takes as input a list of specific mutants, which can contain single or multiple mutations, and generates the mutated PQR files using a combination of SCWRL4 and PDB2PQR. In contrast, mut.comb.c() accepts a list of single mutations and then generates PQR files for all permutations, using SCWRL4 and PDB2PQR, given a desired number of mutations per mutant. Caution should be used when using the mut.comb.c() approach, since the number of permutations can become extremely large for a relatively small list of single mutations depending on the number of mutations per mutant.

After generating PQR files for charge-perturbed structures, the next step is to setup the APBS parameters for the electrostatic potential and free energy calculations. The function apbs.solv() generates an object containing the APBS parameters based on the initial PDB and the location of perturbed PQR files. apbs.solv() will assign default values to parameters such as protein dielectric and will suggest grid lengths based on the coordinates of the protein, however, all parameters can be adjusted using the APBS keyword notation. It is imperative at this step that the parent structure is used in centering the electrostatic calculations to remove the possibility of grid artifacts. Once the APBS input parameters are initialized, the next step is

to perform the Poisson-Boltzmann electrostatic calculations. For efficiency, the calc.solv() function simultaneously calculates the solvation free energy and electrostatics potential of each mutated component/complex. Instead of calculating the six states of the thermodynamic cycle (Figure 2-1B), followed by an additional calculation for the electrostatic potential, calc.solv() calculates the vertical process of for the parent and each mutant and saves the electrostatic potential for solvated state.

The final steps of the AESOP protein design platform involve quantifying/comparing the effects of the perturbations on the electrostatic character. The first approach comparison is electrostatic similarity clustering, which requires the esd.dist() function to generate an electrostatic similarity distance matrix, containing all pair-wise comparisons. The esd.dist() function simply requires the name of the directory containing the electrostatic potentials to be compared, and returns a two-dimensional distance matrix that can used in clustering methods included in the R statistical language [48], such as hclust() for hierarchical clustering. The second step in quantifying the effects of the electrostatic perturbations is to calculate solvation free energies of association by combing the free energies of the three vertical processes according to Eq. 2-8.

The AESOP pH dependent platform follows a similar workflow to the AESOP design platform, as shown in Figure 2-7, but differs in the type of electrostatic calculations being performed. As discussed for the AESOP design platform, the pH dependent platform also begins with the generation of PQR files using PDB2PQR to add hydrogen atoms, as well as atomic radii and partial charge parameters. However, instead of generating perturbed PQR files, the get.pqrs() function is utilized to generate two PQR objects (from Bio3D) containing the same number of atoms but representing the fully charged or fully neutral states. Subsequently, the

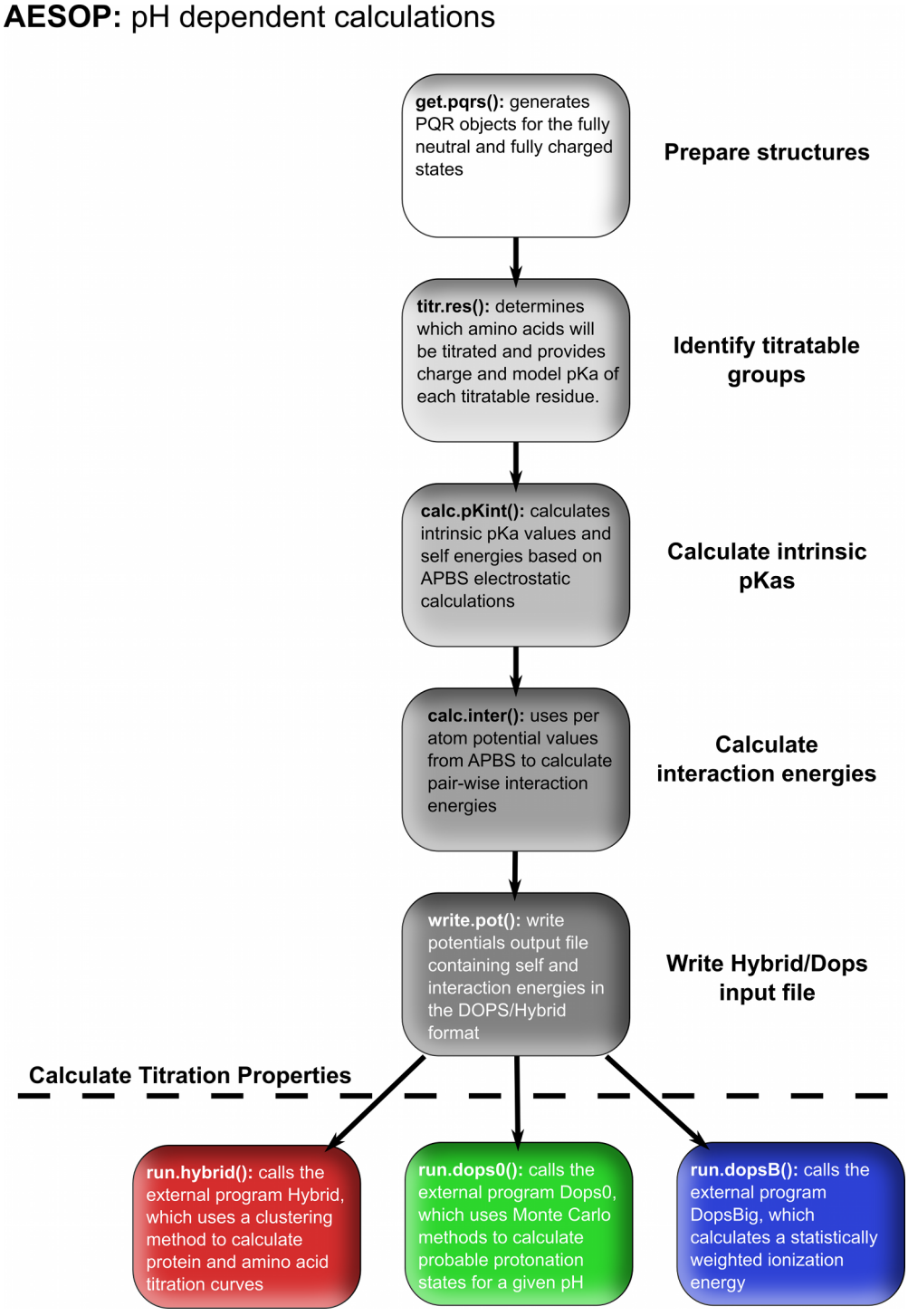


Figure 2-7 *Schematic of AESOP functions for pH dependent calculations.*

titr.res() function is used to identify all titratable groups/residues, and stores where each group/residue is located within the fully charged/neutral PQR files (e.g line numbers) for later use. At this step, specific residue types/numbers can be excluded/added to the titration. Also titr.res() saves the net charge and model pKa for each titratable group/residue, in addition to their locations in the PQR objects.

After obtaining the neutral/charged PQRs and the list of titratable groups, the next step is to calculate intrinsic pK_a values (self-energies) using the calc.pKint() function according to the thermodynamic cycle of Figure 2-2. Self-energies account for the penalty of placing a single charged group in the protein environment, and are calculated by considering a specific titratable group in both the charged/neutral states in both the model (free in solution) and protein environments. calc.pKint() uses the neutral/charged PQRs and the list of titratable groups/residues to generate the various states needed to calculate self-energies, and calls APBS to calculate the corresponding electrostatic free energies. Pairwise electrostatic interaction energies are then calculated between every possible pair of titratable groups. The ultimate step is to use the calculated self and interaction energies in combination with statistical approximation software to extract protein titration properties, such as protein titration curves and apparent pK_a values. Two such programs are HYBRID [52], developed by Michael Gilson, and DOPS [53], developed by Jan Antosiewicz. The HYBRID software uses a clustering approach to reduce combinatorial complexity, while DOPS utilizes Monte Carlo simulations to sample possible ionization states and minimizes the ionization free energy (Eq. 2-9). Therefore after calculation, the self and interaction energies are then written to a potentials file, in the format required by HYBRID and DOPS, using the write.pot() function. run.hybrid() calls the HYBRID software, which returns residue titration curves, apparent pK_a

values, and ionization free energies. run.dops0() calls a version of DOPS best for determining the lowest energy state for a given pH, while run.dopsB() uses more rigorous sampling.

2.3.4. Development and test of the pH dependence platform: pH dependent properties of barnase-barstar

As a proof of concept, the AESOP pH dependent platform was applied to the association of barnase barstar, a gold-standard system for protein electrostatics calculations [54-63]. The extensive experimental [64,65] data available in the literature for the pH dependence of barnase barstar association allow for validation of the AESOP pH dependent calculations. The procedure for the AESOP pH dependent platform, as described in section 2.3.3, was applied to the barnase barstar complex. Apparent pK_a values for barnase and barstar were calculated using HYBRID. Table 2-1 compares calculated apparent pK_a values, from both AESOP/HYBRID and PROPKA [66], to experimentally determined pK_a values for acidic residues of barnase [65]. The AESOP/HYBRID and PROPKA predictions perform comparably well with a root mean squared error (RMS) of ~1 when compared to the experimental values. It should be noted that the AESOP/HYBRID apparent pK_a values showed dependence on the parameters used for electrostatic calculations. More specifically, the predictions were dependent on the protein dielectric coefficient and type of dielectric boundary assumed, with a protein dielectric of 20 and a van der Waals surface providing the lowest RMS when compared to experiment (data not shown). PROPKA predictions have no such dependencies since the software uses an empirical approach rather than rigorous electrostatics calculations, as discussed for AESOP predictions thus far.

In addition to pKa values, pH-stability curves for the association of barnase-barstar were also calculated as illustrated by Figure 2-8. The AESOP predicted curve (red, Figure 2-8) was

calculated using the extensively sampled version of DOPS to identify the 50,000 lowest energy states, based on over 1 million Monte Carlo steps, for each pH value. A partition function was then applied to calculate the free energy at each pH value based on the energies of the 50,000

Table 2-1 *Predicted apparent pK_a values for acidic residues of barnase and comparison with experiment.*

Residue	pK_a^m	AESOP	PROPKA	EXPERIMENTAL[a]
A-ASP-8	4.0	3.8	3.8	3.1
A-ASP-12	4.0	4.1	3.5	3.8
A-ASP-22	4.0	3.6	3.9	3.3
A-GLU-29	4.4	4.6	3.7	3.8
A-ASP-44	4.0	4.0	4.2	3.6
A-ASP-54	4.0	3.9	2.7	2.2
A-GLU-60	4.4	4.3	4.6	3.4
A-GLU-73	4.4	4.4	5.4	2.1
A-ASP-75	4.0	3.0	2.2	3.1
A-ASP-86	4.0	4.2	3.3	4.2
A-ASP-93	4.0	3.6	2.2	2.0
A-ASP-101	4.0	4.0	1.0	2.0
A-CTER-110	3.8	3.7	1.4	3.3
RMS		**1.1**	**1.2**	

[a]Experimental pKa values were previously reported [65].

lowest energy states. The PROPKA curve (blue, Figure 2-8) in contrast was calculated based on changes in the predicted pK_a values upon association of barnase-barstar, according to:

$$\Delta G = -RT\sum_{m=1}^{M}\ln\left[\frac{[H^+]+K_{a,m}^{AB}}{[H^+]+K_{a,m}^{A\ or\ B}}\right]+const. \qquad \text{Eq. 2-12.}$$

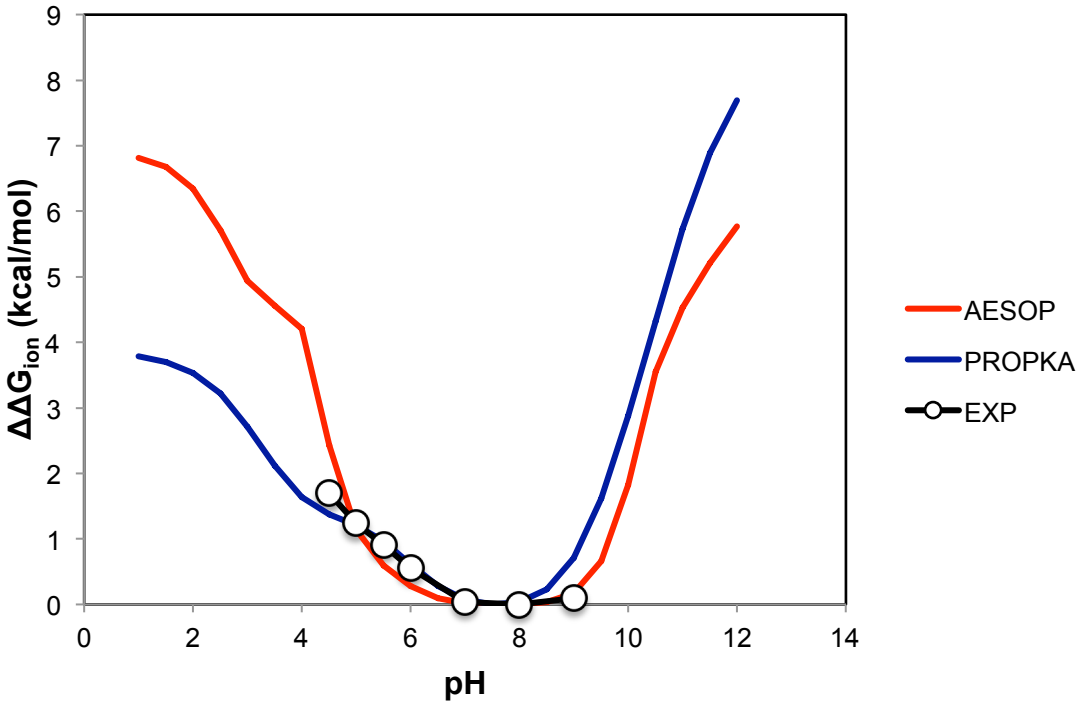


Figure 2-8 *Comparison of pH-stability curves for association of barnase and barstar.* Experimental data (EXP) was previously reported [64].

In Eq. 2-12 the pH and pK_a values are represented by the hydrogen concentration $[H^+]$ and acid association constants K_a, respectively, and RT represents the gas constant and temperature. In this expression, the pH-dependent stability (ΔG) is defined by a ratio of pK_a values, with the association constants for each titratable group in the bound complex (K^{AB}) in the numerator and association constants for each group in the free components ($K^{A \text{ or } B}$) are found in the denominator. The two predicted curves were compared to an experimentally derived pH stability curve (black/white, Figure 2-8) [64], and both show good agreement in the experimental pH range. As discussed for the pK_a predictions, the AESOP pH-stability also showed dependence on the protein dielectric and molecular surface definition used (data not shown). In general, PROPKA performs quite well when compared to experiment, both for pK_A values and

pH-stabilities, and is remarkably fast in comparison to the calculations of the AESOP framework. However, the AESOP pH-dependent platform is capable of incorporating nonstandard titratable groups and ion strength effects, which are not possible with the latest versions of PROPKA [66].

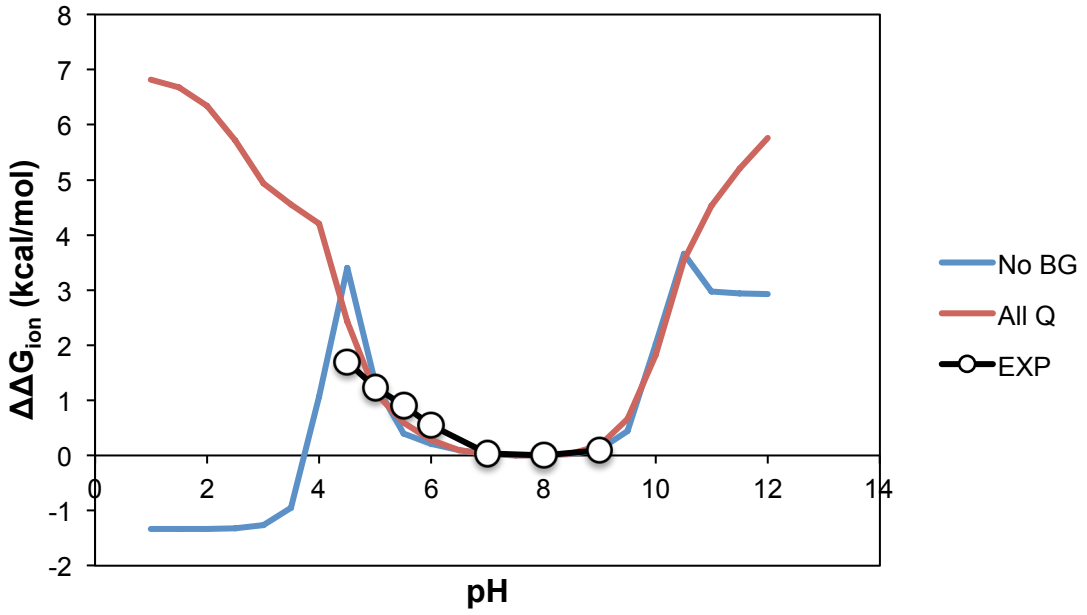


Figure 2-9 *Comparison of effects of background partial charges on predicted pH-stability curves.* Experimental data (EXP) was previously reported [64].

It should be noted that special consideration must be made regarding how/if to include background partial charges for atoms of the peptide backbone and polar-neutral side chains in the calculations of self and interaction energies using the AESOP framework. The AESOP framework contains approximations, which can substantially reduce computation time, but can cause discontinuities in resulting stability curves, as shown in Figure 2-9. If all background partial charges are ignored, it is possible to calculate the self and interaction energies based on the four calculations needed to calculate self-energies (Figure 2-2). This is achieved by saving the per-atom potentials for each titratable group in the charged and neutral states (bottom, Figure 2-2),

and determining the interaction energies based on how the potential produced by one group is felt by another. The resulting curve (No BG, Figure 2-9) contains discontinuities near pH of 4 and 10, but still agrees well with experiment (EXP). These discontinuities are not seen in the pH-stability curve generated using all background charges (All Q,); however, including all background charges requires explicit calculation of each interaction energy and takes at least five times more computational time, on a single CPU.

2.4. IS THE RIGID-BODY ASSUMPTION REASONABLE? INSIGHTS INTO THE EFFECTS OF DYNAMICS ON THE ELECTROSTATIC ANALYSIS OF BARNASE-BARSTAR

2.4.1. Introduction

In protein design, it is of great utility to determine the contributions of each amino acid residue to function, in order to identify which sites are indispensable for activity and which can be further optimized. Experimentally, researchers often perform alanine-scan mutagenesis in an attempt to identify such sites, but these methods can be both time consuming and costly. Therefore, computational methods with the ability to predict the effects of mutations in a high-throughput and efficient manner are in need. Electrostatic similarity methods can be used to computationally predict the effects of electrostatically relevant mutations on binding. To this end, we have developed a computational protocol, AESOP (Analysis of Electrostatic Similarities Of Proteins [25,38]), which utilizes ESD clustering to evaluate the effects of single charge perturbations that are introduced through alanine-scan mutations. Using a combination of the hierarchical ESD clustering and electrostatic free energies of association, AESOP provides

predictions of the effects of the alanine-scan mutation of ionizable amino acids on association driven biological function. Typically in Poisson-Boltzmann calculations, there is an underlying rigid body assumption that is utilized for simplicity. In this study, we assess the validity of the rigid body assumption for electrostatic calculations performed by incorporating dynamics from molecular dynamics (MD) simulations. The MD simulations are used to generate alternative conformations of the parent structure and to relax the structures of alanine scan mutants. Since various ESD metrics and comparative schemes have been proposed in the literature, we also examine their utility for the AESOP analysis. Finally, we discuss the correlations between the AESOP analyses and published experimental data.

For the current study we have chosen the barnase-barstar interaction as our test system to validate our methods. This is largely because there are experimental data for several alanine mutations in the literature [64,67-69], which form the gold-standard dataset in the literature of electrostatically driven association. Barnase is the extracellular ribonuclease of Bacillus amyloliquefaciens and barstar is its co-expressed inhibitor [70]. Electrostatics drives the rapid association and tight binding of barnase-barstar, inhibiting the potentially harmful activity of barnase inside the cell. The barnase-barstar complex is relatively small in size, with only 198 amino acid residues in total, and both proteins are excessively charged (The net charges are +2e for barnase and -6e for barstar.) Also, the barnase-barstar interaction has been the focus of multiple studies using various computational approaches, including free energy calculations, Brownian dynamics, etc., which allow for further analysis of our results [54-63].

2.4.2. Methods

The calculations of this study were performed at atomic resolution using an X-ray crystallographic structure of the barnase-barstar complex (PDB Code: 1BRS [70]). An AESOP analysis was performed using methods that have been previously described, with the crystal structure serving as the starting point or parent molecule [25,38]. In brief, WHATIF was first used to generate the alanine mutants based on the coordinate file of the barnase-barstar complex by replacing each ionizable amino acid with alanine, one at a time. Twenty-eight barnase alanine mutations were generated. The mutant complex files were subsequently split to obtain separate sets of coordinates for each component. PDB2PQR [42] was used to add hydrogen atoms, partial and whole charges, and van der Waals radii, according to the PARSE [43] force field, to the coordinates of each mutant. The Adaptive Poisson-Boltzmann Solver (APBS) was subsequently used to calculate the spatial distributions of electrostatic potential and electrostatic free energy differences corresponding to each alanine-scan mutant and parent protein. The probe radii for defining the dielectric and the ion accessibility surfaces were set to 1.4 Å and 2.0 Å, respectively. A protein dielectric constant of 2 was used throughout, and a solvent dielectric constant of 78.5 was used when relevant. Two sets of calculations were performed, which correspond to the presence or absence of monovalent counterions at 150 mM ionic strength. The APBS calculations were performed on a grid containing 129 × 129 × 129 grid points with grid lengths of 125 Å × 125 Å × 130 Å, providing a grid resolution of ~1 Å.

Electrostatic free energies of association were calculated according to a thermodynamic cycle, Figure 2-1B, which includes association in both a solvated, $\Delta G_{assoc,sol}$, and a vacuum-like reference state, $\Delta G_{assoc,refl}$. Based on this theoretical thermodynamic cycle, solvation free

energies of association, $\Delta\Delta G_{solv}$, were calculated according to Eq. 2-8. All steps discussed thus far were automated through the use of in-house Python scripts. Following the APBS calculations, ESD clustering was performed using three ESD distance metrics, including dot-product (DP), local dot-product (LDP), and average normalized difference (LD), which were calculated according to Eq. 2-5, Eq. 2-6, and Eq. 2-7. These ESD metrics are derived from the ESI metrics described by Eq. 2-2, Eq. 2-3, and Eq. 2-4. In addition, three comparison schemes were used in combination with the three ESD metrics, resulting in nine total electrostatic similarity methods. The utilized schemes are the whole-box (WB), shell, and skin, as seen in the schematic of Figure 2-10. For the shell scheme, all points inside the 2.0 Å ion accessibility surface were ignored, while for the skin method only grid points enclosed by two surfaces, at 2.0 Å and 7.0 Å from the protein surface, were included. In-house R scripts were used to generate distance matrices containing comparisons of the spatial distributions of electrostatic potential for each mutant-mutant and parent-mutant pair, based on one of the nine clustering method combinations. Hierarchical clustering was applied to the generated distance matrices using average linkage through the hclust function of the R-base package . R scripts were subsequently used to visualize the clustering results by generating dendrograms and plotting the electrostatic free energy differences according to the order of said dendrograms.

The software package NAMD (Nanoscale Molecular Dynamics) was used to perform a series of 8 MD simulations based on the structures of the parent complex and 7 barnase mutants (K25A, D52A, R57A, E58A, E71A, R85A, and H100A) found in the literature. These barnase alanine-mutants were generated as described above. Visual Molecular Dynamics (VMD) was used to generate a PSF file for each complex based on the CHARMM 22 forcefield . Complexes were placed in explicit water boxes, containing ~8000 water molecules, and NaCl

counterions were added at 150mM ionic strength, with the net charge of the system being neutral. Simulations were performed using PM Ewald electrostatics with 1 Å grid spacing, as well as Langevin constant temperature control to maintain the system temperature at 298 K. Constant pressure control was also implemented using a Langevin piston. A 2 fs time step was used in all MD simulations, which was allowed by rigid hydrogen bonds implemented through the SHAKE algorithm. Prior to each simulation 1000 steps of NAMD's conjugate gradient energy minimization were performed to remove any unfavorable contacts or geometries prior to MD equilibration. The initial simulation involving the parent complex was allowed to run for 10 ns, while the mutant complexes were ran for 5 ns.

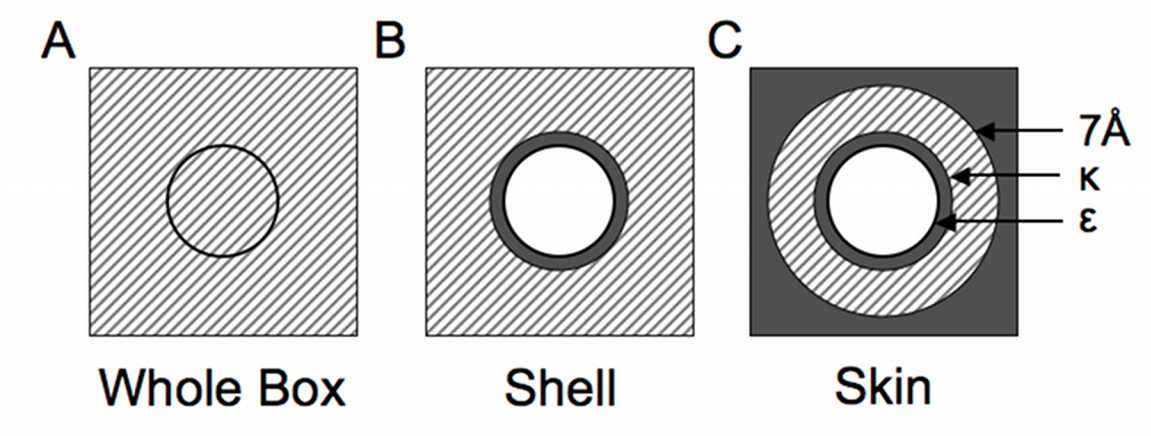


Figure 2-10 *Schematic of comparative schemes for ESD metrics.* Illustration showing which regions of the spatial distribution of electrostatic potential are considered by each comparative method: (A) whole box; (B) shell; (C) skin. The hashed area indicates grid points that are included, while the black and white areas indicate grid points excluded from the ESD calculations. The three surface boundaries (ε: dielectric, κ:ion-accessibility and 7 Å) are label in (C).

Following the MD simulations, two further complete AESOP analyses, as described above, were applied to the 5 and 10 ns snapshots from the parent complex trajectory. In order to determine if the ESD clustering methods provide similar clustering results, irrespective of the conformation of the parent, clustering similarities were calculated. Clustering similarity methods

require that objects, in our case alanine mutants, be partitioned in to a finite number of clusters, rather than the tree structure obtained from hierarchical clustering. To convert the generated dendrograms into the appropriate form, the tree cutting function, cutree, of R was used to extract a finite number of clusters, k. Since the ideal number of clusters to be extracted from each dendrogram is not obvious, tree cutting was performed for k values of 2 to 15 and similarities were calculated for each tree cutting result. Clustering similarities for each k value were calculated according to the Rand index,

$$R(C,K) = \frac{(a+d)}{(a+b+c+d)} \qquad \text{Eq. 2-13.}$$

In this expression, C and K represent two tree cutting results to be compared, while a is the number of pairs that cluster together in both C and K, b is the number of pairs that cluster together in C but not K, c is the number of pairs that cluster together in K but not C, and d is the number of pairs that are not joined into clusters in K and C. For the Rand index cluster comparison, a value of 0 indicates no agreement while 1 indicates identity.

2.4.3. Results

For this study, an initial 10 ns MD trajectory, using an explicit water box, was generated for the parent barnase-barstar complex. The 10 ns barnase-barstar MD trajectory was generated to monitor the extent of the fluctuations of the electrostatic properties of the parent complex. Figure 2-11 illustrates the fluctuations of the spatial distributions of electrostatic potential, together with $\Delta\Delta G_{solv}$ values, during the MD trajectory. The $\Delta\Delta G_{solv}$ values have a total range of ~150 kJ/mol, however, fluctuations are reduced towards the end of the trajectory where the solvation free energy begins to settle at around 1250 kJ/mol. Similar fluctuations and trends are visible in the isopotential contours of Figure 2-11B, but differences are more noticeable when

potentials are calculated at 150 mM ionic strength. Despite only slight changes in the protein backbone, as evident through comparison of the ribbon models, dynamics have a distinguishable effect on both the electrostatic association free energies and electrostatic potentials.

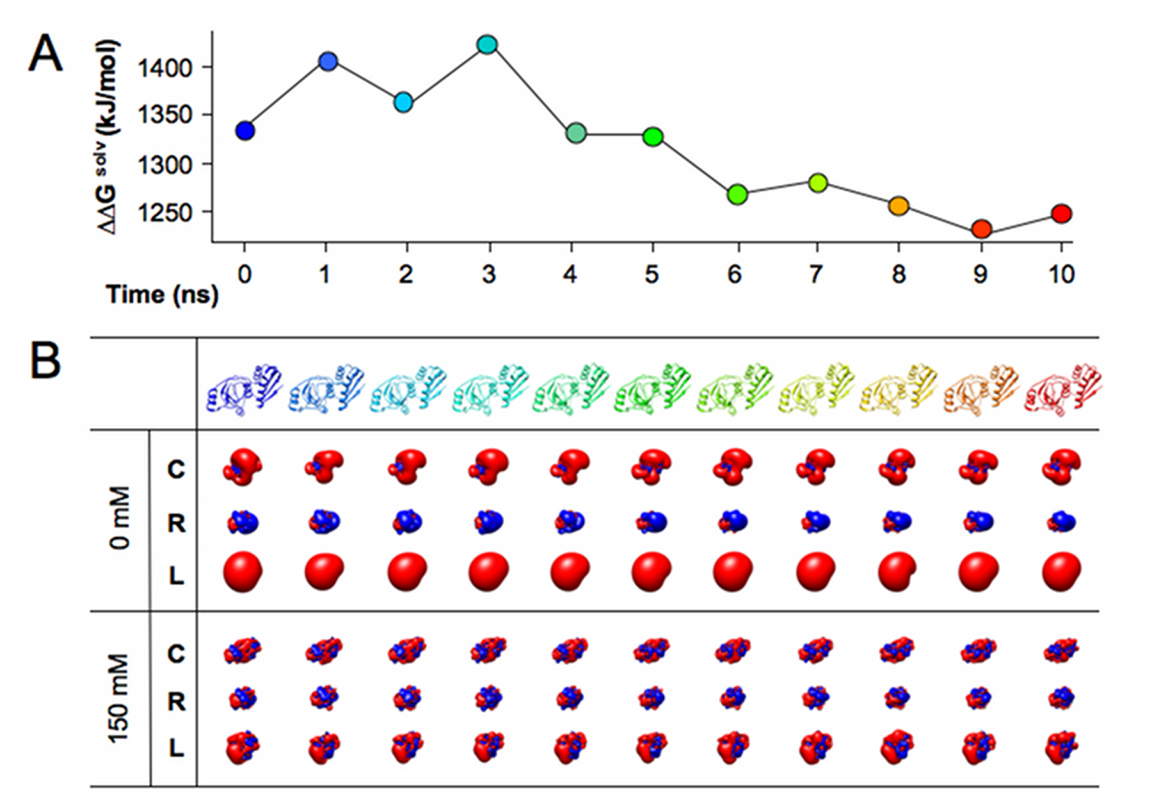


Figure 2-11 *Electrostatic fluctuations of barnase-barstar during a 10 ns MD trajectory.* (A) Plot of $\Delta\Delta G_{solv}$ for the crystal structure (0 ns) and 10 snapshots taken in 1 ns increments (1 - 10 ns). Data point color transitions from blue to red with time and is included to correlate with the color of the ribbon models of (B). (B) Ribbon models and isopotential contours depicting the conformation and corresponding electrostatic potential for the structures as described in (A). Isopotential contours are presented for barnase-barstar (C:complex), as well as barnase (R:receptor) and barstar (L:ligand) alone. Isopotential contours are shown for two ionic strengths, 0 mM and 150 mM. The color code is blue for +1 kT/e and red for -1 kT/e electrostatic potential.

The 10 ns MD trajectory for the parent complex also served the purpose of generating alternative conformations to be used in AESOP analyses. In total, three complete AESOP analyses were performed based on the 0, 5, and 10 ns barnase-barstar conformations. Each

AESOP analysis included the generation of 28 barnase alanine-scan mutations, and calculating the corresponding electrostatic potentials and $\Delta\Delta G_{solv}$, at 0 and 150 mM ionic strength according to the Poisson-Boltzmann method. This resulted in six sets of electrostatic potentials, each containing 29 spatial distributions, corresponding to parent barnase and barnase alanine mutants. Each set of potentials was then analyzed using each of the 9 comparative-scheme/ESD combinations, resulting in 54 sets of clustering results. The clustering results were individually plotted in dendrogram or tree form, and the solvation free energies of association were plotted in the order of the resulting dendrogram.

The results of all three AESOP analyses were quite comparable, producing correlations of > 0.96 when comparing the $\Delta\Delta G_{solv}$ values for the alanine-scan mutations generated from these snapshots (data not shown). The correlations between the clustering results and the electrostatic free energies of association, based on visual inspection, were similar for all clustering methods, with exception to WB-DP which performed noticeably worse. This was an expected result, since the WB-DP method is plagued by an inability to overcome extremely large and dominating potential values that are found in the protein core. An example dendrogram and free energy plot for AND clustering at 0 mM ionic strength are presented in Figure 2-12. As seen in Figure 2-12A, the AND clustering creates three distinct clusters, which separate the mutations according to the mutated residue type (acidic, basic, or neutral). Within each of these major clusters, smaller subclusters with distinct electrostatic properties exist, and this is further emphasized through visual correlations with $\Delta\Delta G_{solv}$ seen by comparing panels (A) and (B) of Figure 2-12. It should be noted that an increase in $\Delta\Delta G_{solv}$ is actually indicative of a favorable perturbation when using a low protein interior dielectric constant, such as the value of 2 used in this study.

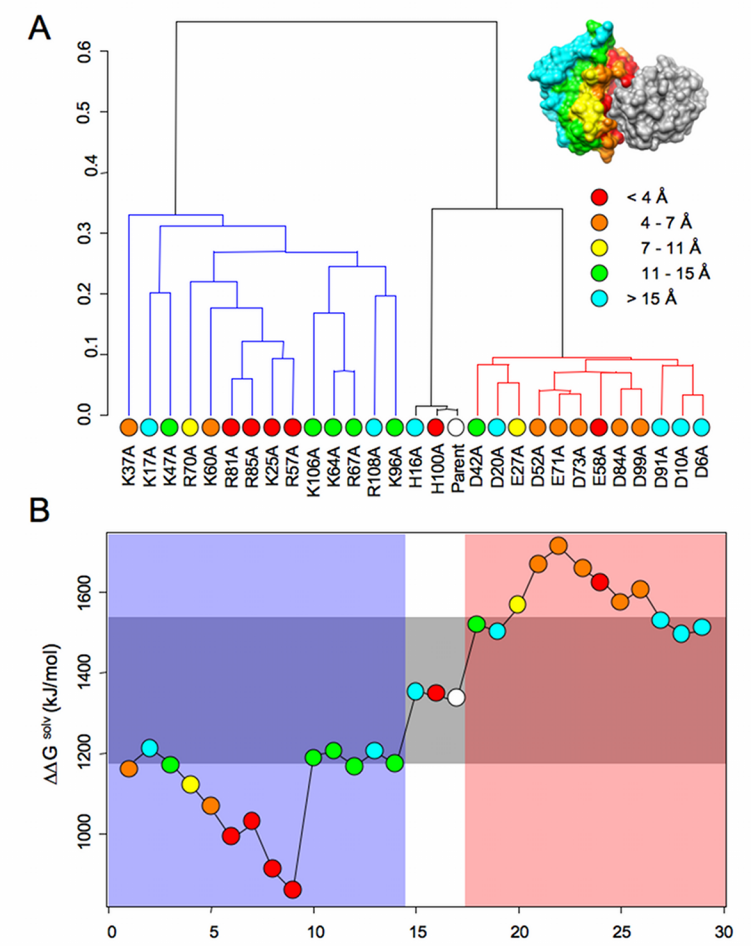


Figure 2-12 *AESOP analysis of barnase at 0 ns using WB-AND.* (A) Dendrogram corresponding to the AND electrostatic clustering of barnase alanine scan mutants using all grid points (WB-scheme). Colored circles indicate the distance of each mutated residue from the binding interface according to the legend and surface representation of the complex. In the surface representation, barnase residues are colored according to their distance from barstar, while barstar is in gray. The colored lines of the dendrogram are included to emphasize the three major clusters, which are on the type of the mutated residue (blue for basic, red for acidic, and black for neutral) (B) Plot of $\Delta\Delta G_{solv}$ for parent barnase and its alanine scan mutants according to the order of the dendrogram in (A). Colored circles follow the same coloring scheme as described in B. The three vertical boxes indicate the three primary clusters (blue for basic, red for acidic, and white for neutral) as described for (A). The dark grey box of B indicates mutations predicted to have little or no effect on binding. White circle indicates the parent protein in (A) and (B).

To quantitatively assess the effects of the starting structure on the AESOP analysis, correlations between our data for the 0, 5, and 10 ns structures and available experimental results were calculated. Table 1 provides relative-experimental ΔG values for seven barnase

alanine mutants seven (K25A, D52A, R57A, E58A, E71A, R85A, and H100A), calculated by dividing each experimental value by that of the parent. These relative experimental ΔG values were used in place of the actual values in order to be able to combine two different experimental data sets [64,67]. $\Delta\Delta G_{solv}$ values, calculated based on the three starting structures at both 0 and 150 mM ionic strength, are also presented for the seven mutants in Table 2-2. Correlation coefficients comparing each set of calculated ΔΔGsolv values to the relative-experimental ΔG data can also be found in Table 2-2. Inspection of these correlations suggests that both the ionic strength and the starting structure have little to no effect on the agreement with experimental data, since the coefficients range from 0.76 to 0.79.

Table 2-2 *Solvation free energies of association and correlations to experimental data.*

		$\Delta\Delta G_{solv}$ (kJ/mol)							
		0 ns[b]		5 ns[c]		10 ns[d]		Relaxed Mutants[e]	
Mutant	**Rel. ΔG_{exp}[a]**	*0 mM*	*150 mM*	*0 mM*	*150 mM*	*0 mM*	*150 mM*	*0 mM*	*150 mM*
Parent	1.00	1331.3	1344.4	1329.0	1342.5	1246.6	1259.2	1329.0	1342.5
K25A	0.72	912.8	919.0	914.8	920.7	904.3	910.6	1029.6	1036.6
D52A	1.05	1675.4	1694.6	1604.6	1624.1	1570.7	1590.4	1610.2	1627.5
R57A	0.73	854.7	861.2	856.1	863.0	804.8	810.9	1043.2	1049.4
E58A	1.02	1629.9	1649.6	1545.4	1564.7	1509.0	1527.9	1592.4	1610.4
E71A	0.90	1712.2	1732.1	1622.7	1643.5	1606.5	1626.9	1608.4	1626.5
R85A	0.70	1025.8	1033.4	966.3	974.0	903.7	911.2	1013.2	1019.0
H100A	0.67	1339.8	1353.0	1269.3	1282.5	1234.8	1247.7	1269.5	1280.6
Corr. Coef.[f]		0.76	0.76	0.79	0.79	0.77	0.77	0.82	0.82

[a]Relative experimental free energies from Schreiber et al [64,67].
[b]Alanine mutants based on the crystal structure of barnase-barstar
[c]Alanine mutants based on the 5ns MD snapshot of barnase-barstar
[d]Alanine mutants based on the 10ns MD snapshot of barnase-barstar
[e]Alanine mutants based on the crystal structure of barnase-barstar, and 5 ns of MD were applied following mutation.
[f]Correlation coefficients calculated using the Pearson method

Clustering similarities were also calculated according to the Rand index to quantify the effect of the starting structure on the clustering results for all nine of the methods described.

For this analysis a series of Rand indices were calculated corresponding to k values of 2 to 15. Table 2 contains the statistical analysis of the Rand indices resulting from the comparisons of the 0 and 5 ns and 0 and 10 ns clustering data. The clustering results for the different electrostatic similarity methods are consistent across the three parent conformations, which is evidenced by mean and median similarities of about 0.9 or higher, with an exception of WB-DP. However, slightly lower consistency, based on lower Rand indices, is seen between clustering results for potentials calculated at 150 mM ionic strength, when compared to 0 mM results. This suggests that the parent structure conformation has more effect on the results when considering higher ionic strengths.

Table 2-3 *Rand clustering similarity indices for the comparison 0, 5, and 10 ns clustering results.* The 0 ns & 5 ns similarities represent the comparison of the AESOP clustering dendrograms, where mutations were based either on the crystal structure (0 ns) or the 5 ns MD snapshot. A similar comparison between the results for the 0ns barnase-barstar structure and the 10 ns MD snapshot are also provided (0 ns &10 ns).

	0 ns & 5 ns						**0 ns & 10 ns**					
	0 mM			*150 mM*			*0 mM*			*150 mM*		
Method	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
WB-DP	0.77	0.80	0.12	0.81	0.80	0.06	0.84	0.88	0.10	0.81	0.79	0.07
WB-LDP	0.97	0.98	0.04	0.87	0.91	0.13	0.97	1.00	0.05	0.88	0.93	0.13
WB-AND	0.96	0.98	0.04	0.94	0.95	0.04	0.96	0.99	0.05	0.95	0.96	0.05
Shell-DP	0.98	0.99	0.03	0.93	0.94	0.06	0.96	0.98	0.04	0.96	0.97	0.03
Shell-LDP	0.97	0.98	0.04	0.87	0.91	0.13	0.97	1.00	0.05	0.88	0.93	0.13
Shell-AND	0.96	0.98	0.04	0.94	0.95	0.04	0.96	0.99	0.05	0.95	0.96	0.05
Skin-DP	0.96	0.97	0.03	0.94	0.95	0.05	0.97	0.97	0.03	0.94	0.94	0.03
Skin-LDP	0.95	0.96	0.05	0.91	0.92	0.04	0.94	0.96	0.06	0.77	0.79	0.12
Skin-AND	0.96	0.99	0.06	0.92	0.95	0.10	0.95	0.98	0.06	0.79	0.83	0.13

An additional seven MD trajectories, with 5 ns of simulation time each, were generated for seven barnase mutants (Table 2-2) in complex with barstar. AESOP analyses were performed using the 5 ns snapshots from the MD trajectories of the parent and seven relaxed barnase mutants. These analyses were used to investigate the effects of post-mutation relaxation on the

clustering and free energy results. The $\Delta\Delta G_{solv}$ values for these relaxed mutants and the corresponding experimental correlations are provided in Table 2-2. Only a small improvement in experimental correlation, from 0.76 to 0.82, was observed when considering the relaxation of the alanine scan mutations. ESD clustering was also performed to compare the spatial distributions of relaxed mutants. The clustering of the two sets of electrostatic potentials, before (0 ns, Table 2-2) or after relaxation (5 ns Relaxed Mutants, Table 2-2), individually resulted in very similar clustering dendrograms. This result, however, is probably owed to each set only containing 8 electrostatic potentials. Therefore, ESD clustering was applied to a combined set of potentials, containing the barnase parent and seven alanine mutations in two different conformations (0 ns and 5 ns Relaxed Mutants, Table 2-2). The clustering analysis was performed at two ionic strengths (0 and 150 mM) using all nine clustering methods to assess which methods are more sensitive to perturbation (mutation) affects rather than dynamic affects. Example dendrograms for WB-AND and Skin-DP at 0 and 150 mM are presented in Figure 2-13. Inspection of the 18 dendrograms showed that only the WB-AND at 0 mM (Figure 2-13A) clustering was able to group the mutations of the same residue together despite the structural differences. All other methods, including WB-AND at 150 mM, grouped mutations primarily based on whether or not relaxation was performed, as emphasized by the brackets in Figure 2-13B - D, instead of being based on electrostatic perturbation effects.

2.4.4. Discussion

We have developed a computational protocol, AESOP, for the electrostatic analysis of protein-protein interactions using ESD clustering [25,38]. The use of ESD clustering gives the AESOP protocol the unique capability of producing protein association predictions in the

absence of a structure for the protein complex. Through single ionizable residue mutations, AESOP analyzes the effect of each perturbation (charge removal) on the spatial distribution of electrostatic potential and hierarchically classifies the mutations based on these effects. Mutations of ionizable amino acids to alanine are not expected to alter global structure, because ionizable amino acids are typically solvent exposed and thus do not contribute to the stability of the protein core. We expect local structural changes to occur as a result of alanine scan mutagenesis, owed to the loss of an interacting partner for electrostatic interactions. Alanine scans are used to introduce perturbations into the physicochemical properties proteins, which in our calculations were the electrostatic properties of barnase-barstar. When electrostatics properties are important for binding, the alanine-scan perturbations will be manifested in the free energies and clustering dendrograms of the AESOP protocol.

In this study, we sought to determine if a rigid body assumption (e.g. using the experimental derived structure) is reasonable for the calculations of the AESOP protocol. In answering this question two secondary questions arose regarding the effects of dynamics on the electrostatic analysis of barnase-barstar: Does the conformation of the parent complex affect the analysis, and is it necessary to consider relaxation of each alanine-scan mutant? To answer these questions, we have examined the effects of pre- and post-mutation relaxation on the electrostatic of barnase-barstar, by coupling MD simulations with the AESOP analysis. To first elucidate the effects of conformation on the electrostatically-driven association of barnase-barstar, we generated a 10 ns MD trajectory based on the crystallographic structure of the complex. Examination of the backbone root mean squared deviations (RMSD) for the trajectory showed that equilibration occurred immediately, and that RMSD values steadily fluctuated between 1 – 2 Å (data not shown). This indicates that the crystal structure of barnase-barstar

was already well optimized prior to the MD simulation. Fluctuations in the recognition (isopotential contours, Figure 2-11B) and binding ($\Delta\Delta G_{solv}$, Figure 2-11A) characteristics of barnase-barstar were observed despite these small structural changes.

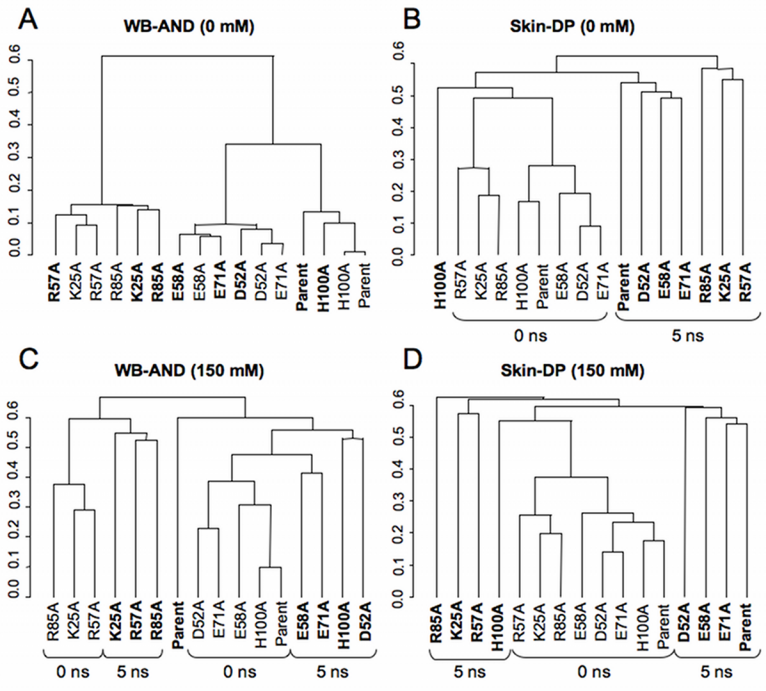


Figure 2-13 *Comparison of the effects of relaxation on ESD clustering results for barnase.* Dendrograms depicting the clustering results for electrostatic potentials of barnase and seven alanine mutants, before and after 5 ns of MD relaxation. Clustering was performed using: (A) WB-AND (0 mM); (B) skin-DP (0 mM); (C) WB-AND (150 mM); (D) skin-DP (150 mM). Bold labels indicate equilibrated structures. Brackets are included in (B), (C), and (D) to emphasize the clustering results.

The AESOP analysis provides a relative comparison of the effects of charge perturbations on electrostatics, therefore, the importance of the parent complex conformation, on which the alanine-scan mutations were based, was investigated. Three complete AESOP analyses were performed on the 0, 5, and 10 ns snapshots of the parent barnase-barstar complex using three ESD metrics (DP, LDP, and AND) and three comparative schemes (WB, shell, and skin). If the conformation were to have no effect on the analysis, we would expect the clustering results (dendrogram) and relative free energy trends to be identical when comparing

AESOP analyses from different barnase-barstar conformations. Rand similarity indices (Table 2-3) demonstrate that all nine methods are capable of reproducing similar clustering results irrespective of the parent complex conformation. We also observed that the relative free energy trends for the alanine-scan mutations were very similar for the three parent conformations (0, 5, and 10 ns), indicating that the conformation of barnase-barstar did not have a noticeable affect on the AESOP analysis. Also, little to no improvement in the correlation with experimental data (Table 2-2) was achieved by using an MD equilibrated structure (5 or 10 ns of pre-mutation relaxation). In addition, Wade et al reported very similar correlations (~0.8) with experimental data for their electrostatic binding free energies, which incorporated desolvation effects [59]. For their calculations, Wade et al performed 800 steps of energy minimization, but did not perform MD simulations. These results suggest, at least for the barnase-barstar complex, that the crystallographic structure is a sufficient starting point for electrostatic calculations, such as those performed by the AESOP analysis.

In addition to investigating the effects of dynamics on electrostatic calculations, we examined the utility of the nine electrostatic similarity methods for the rigid body assumption. Seven of the nine electrostatic similarity methods, excluding WB-DP and skin-LDP, produced similar clustering results showing good correlation with $\Delta\Delta G_{solv}$. The WB-DP method provided poor correlation with $\Delta\Delta G_{solv}$, owed to inability to overcome dominating values in the protein interior. Skin-LDP, on the other hand, produced unreadable dendrograms, since the resulting ESD values were very similar for all electrostatic potential comparisons. Of the three comparative schemes, the WB method is preferred, since the definition of the boundaries for the shell and skin method is computationally costly. For the WB comparative method to be proficient, an ESD metric that utilizes a local normalization, namely LDP or AND, must be used to

minimize the contributions of potential values in the protein interior. Here we propose the use of the WB-AND method for the clustering of electrostatic potentials for alanine-scan mutations. We have found that the WB-AND method, despite its simplicity, provides comparable results to the skin-DP method that has been proposed by Wade et al [22,23].

Theoretical amino acid mutations often introduce local and global conformational rearrangements requiring the use of MD simulations for optimization. Here we examined the benefits of post-mutation relaxation for electrostatic calculations of the barnase-barstar interaction, by performing a 5 ns MD simulation following the generation of alanine mutants found in the literature (Table 2-2 [64,67]). We found that only slight improvements in correlations with experimental data were obtained by performing 5 ns of post-mutation MD relaxation. This is most likely the result of minor local structural rearrangements. It is assumed that further structural rearrangements and improvements in experimental correlations would occur if the barnase mutations were allowed to relax in the unbound state. However, to eliminate to the possibility of introducing grid artifacts, which are an inherent problem in current Poisson-Boltzmann methods, only conformational changes of the bound state could be considered. These artifacts originate from self-energies introduced by charge discretization in nearest neighbor grid points, and are eliminated by using the theoretical thermodynamic cycle of Fig. 1 [24]. For this study, MD relaxation was performed for 7 alanine-scan mutations out of the 28 barnase ionizable amino acids, with each simulation requiring ~20 hrs. of computation on 40 processors. Given the small improvement that was observed compared to the computational effort used, the ~560 hrs. of computation required to relax all 28 barnase mutants is not warranted.

Throughout the calculations, we also considered the effects of ionic strength on the AESOP analysis, since the screening of coulombic interactions by counterions is an important feature of the protein environment. The modeled ionic strength (0 or 150 mM) had little to no effect on the relative trends of the electrostatic free energies, nor the correlations with experimental data, as seen in Table 2-2. Ionic strength did however have a noticeable effect on the electrostatic clustering results, especially when comparing multiple conformations of the same mutant (Figure 2-13). We have also observed this phenomenon, in which ionic screening affects clustering results but not electrostatic free energies, for other applications of the AESOP protocol (unpublished data). Surprisingly, only clustering performed at 0 mM ionic strength using the WB-AND method was able to properly classify mutations based on the effects of the perturbations, instead of dynamics. This result is unexpected, since WB-AND does not exclude potential values in the protein interior, and because 0 mM ionic strength is not believed to be an accurate representation of the protein environment. At 150 mM ionic strength, the presence of counterions emphasizes the effects of dynamics on electrostatic potential, since changes in conformation and the solvent accessible surface would affect the distribution of and screening by counterions. For electrostatic potential calculations of different conformations of the same protein, differences in ionic screening will result in unique spatial distribution of electrostatic potential, despite similar charge distributions. However, in the case of electrostatic free energies of association, differences in ionic screening effects are canceled out by the bottom horizontal process of the theoretical thermodynamic cycle (Figure 2-1B). Therefore, only magnitudes of $\Delta\Delta G_{solv}$ are affected, not the relative trends for all alanine mutations.

Along with validating our methods in light of the rigid body assumption, as discussed thus far, we have also produced a database of predictions for the mutation of each ionizable

amino acid of barnase to alanine. Of the 54 data sets generated for the electrostatic analysis of barnase, we prefer the AESOP analysis performed at 0 mM ionic strength with WB-AND (Figure 2-12), for its accuracy (e.g. good correlations with $\Delta\Delta G_{solv}$ and sensitivity to perturbation irrespective of dynamics) and computational efficiency. Figure 2-12 not only shows good correlation with experiment, but also provides predictions of the contribution of each mutated residue to the electrostatically-driven association of barnase-barstar. These results can be used alone to suggest new perturbative mutations to be tested experimentally, or can be used as a guide for combinatorial calculations, in which multiple alanine mutations would be introduced to design new barnase proteins with tailored electrostatic properties. This type of analysis can also be applied to other interactions with electrostatically-driven association, to gain understanding of the role electrostatics and for protein design.

2.4.5. Conclusion

In summary, we have performed electrostatic analyses of the barnase-barstar interaction, specifically electrostatic clustering and free energy calculations, using the AESOP protocol. We investigated the affects of pre- and post-mutation relaxation on electrostatic calculations through the incorporation of snapshots form MD simulations. Our perturbative calculations, based on theoretical alanine-scan mutagenesis, produced good correlations with existing experimental data, irrespective of the conformation of the parent complex. Additionally, only small improvements in the experimental correlations were observed when including 5 ns of post-mutation MD relaxation. These results suggest, at least for the barnase-barstar interaction, that the rigid body assumption is reasonable for the electrostatic calculations of the AESOP protocol. The data of this study show the utility of the AESOP protocol

for efficiently elucidating the contribution of each ionizable amino acid to association, and serve as an example of how these methods can be applied to the design on novel proteins with tailored electrostatic properties.

2.5. ELECTROSTATICS-BASED DESIGN OF SUMO4 SUBTRATE ANALOGS

2.5.1. Introduction

Sumoylation of cellular proteins by the ubiquitin-like members of the SUMO family has been found to be one of the essential regulation mechanisms in signal transduction and genome integrity [71]. One clear difference between the ubiquitin and SUMO pathways is the existence of up to four SUMO isoforms, with SUMO-1, -2, and -3 being the most prominent. Prior to sumoylation, a pre-SUMO peptide must first be matured through the cleavage of a C-terminal tail via an enzyme of the SENP family [71]. The removal of the C-terminal peptide tail exposes a Gly-Gly motif that is utilized during the sumoylation process. Following SUMO conjugation the SUMO peptide can also be removed from its substrate *in vivo*. As discussed for the SUMO family, multiple isoforms also exist in the SENP family, all of which have different specificities for the various SUMO peptides. Of these isoforms, SENP2 is responsible for the maturation of SUMO-1, -2, and -3 into their conjugatable forms, as well as the deconjugation of SUMO-1, -2, and -3 containing species. SENP2 therefore plays a significant role in the regulation of sumoylation [72].

Of the SUMO isoforms, SUMO-4 is still of mystery, since there are no known specific enzymes with SUMO4 as its substrate [73]. SUMO-4 exhibits high sequence similarity to SUMO-2, with only 14 amino acids different out of the total 95 (Figure 2-14) ([25,38,41]), but is still not processed by any known SUM0-spedific enzymes. One notable difference between SUMO-2 and

```
SUMO-2      MADEKPKEGVKTENNDHINLKVAGQDGSVVQFKIKRHTPLSKLMKAYCERQGLSMRQIRF
SUMO-4      MANEKPTEEVKTENNNHINLKVAGQDGSVVQFKIKRQTPLSKLMKAYCEPRGLSVKQIRF

SUMO-2      RFDGQPINETDTPAQLEMEDEDTIDVFQQQTGGVY 95
SUMO-4      RFGGQPISGTDKPAQLEMEDEDTIDVFQQPTGGVY 95
```

Figure 2-14 *Sequence alignment for SUMO-2 and SUMO-4.* Positions with differing amino acids are shown in red.

SUMO-4 is net charge, since the net charge of SUMO-2 is -3 while the net charge of SUMO-4 is 0. Based on this observation we propose that differences in the electrostatic character of SUMO-4, when compared to SUMO-2, prevents the formation of a stable bound complex with SUMO-specific enzymes, and as a result prevents catalysis. In this study, we aim to identify a minimum set of SUMO-4 mutations necessary to regain catalytic ability based on electrostatic perturbative design using the AESOP framework. Since SUMO-4 shares high sequence similarity to SUMO-2, which preferentially binds SENP2, we propose to use the SUMO-2:SENP2 interaction as a guide in the design of active SUMO-4 analogs. Ultimately the goal of this study is to use computation to guide experimental mutagenesis studies; therefore, top predicted analogs are to be experimentally validated using previously proposed methods.

2.5.2. Methods

Given that SUMO-4 does not form stable complexes with any SUMO-specific enzymes there is no available structural data for the SUMO-4:SENP2 interaction. Therefore, the initial step of this study was to generate an initial model of the SUMO-4:SENP2 interaction using a crystal structure of the SUMO-2:SENP2 complex (PDB Code: 2IO0,) as a template. The initial step utilized the homology modeling software MODELLER [74] to extend the N-terminal tail of

SUMO-2 in the SUMO-2:SENP2 complex to agree with an experimentally relevant sequence [75]. In order to ensure a reliable comparison, and since SUMO-2 and SUMO-4 are highly similar, the SCWRL4 package was used to introduce the 14 mutations necessary to form SUMO-4 based on SUMO-2 (Figure 2-14).

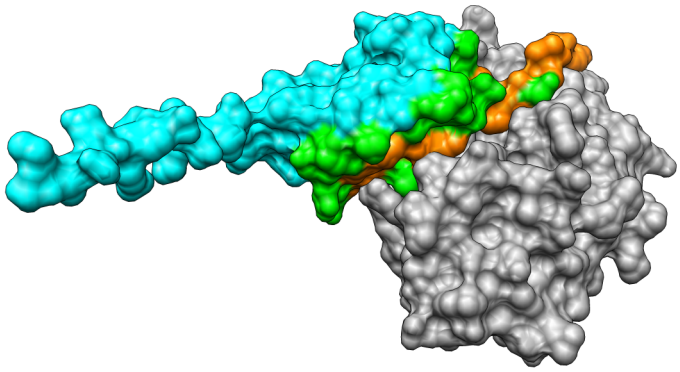


Figure 2-15 *Molecular graphic of SUMO-4:SENP2 interaction.* Both proteins are represented by a molecular surface with SENP2 in gray, while SUMO-4 residues are colored based on their distance from SENP2: orange, < 3.5 Å; green, > 3.5 Å and < 8 Å; cyan, > 8 Å.

All electrostatic calculations were performed using the AESOP framework, which has been discussed in great detail previously (section 2.3,). In brief, all alanine-scan mutations were performed using truncations by AESOP functions, while all non-alanine mutations were performed using the SCWRL4 package [51]. The PDB2PQR [42] utility was used to prepare all structures for electrostatic calculations by incorporating atomic radii and partial charges according to the PARSE forcefield . Electrostatic potential and free energies of association, according to the thermodynamic cycle of Figure 2-1B, were calculated using APBS [3] based on a grid with 129 × 129 × 129 grid points and coarse/fine grid lengths of 212 Å × 98 Å × 130 Å. The dielectric boundary was defined by a water-sized probe sphere with a radius of 1.4 Å, and the

dielectric coefficients of the protein and solvent environments were 20 and 78.54 respectively. All electrostatic calculations were performed using 0 mM of counterions.

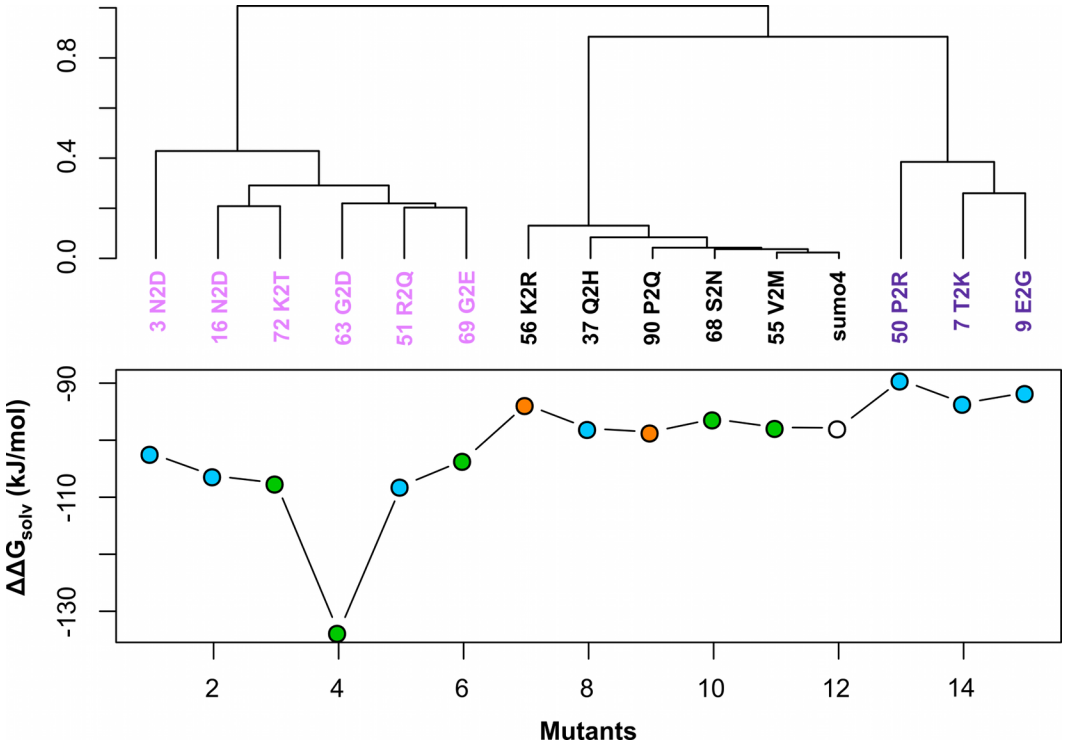


Figure 2-16 *Electrostatic clustering and free energies of association for SUMO-2 based mutations of SUMO-4.* Electrostatic clustering is illustrated as a dendrogram tree (top) with label color indicating the net charge of the resulting analog: pink, - 1; black, 0; purple, +1. The color of the free energy data points indicates the distance of the mutated residue from SENP2 as illustrated by Figure 2-15: orange, < 3.5 Å; green, > 3.5 Å and < 8 Å; cyan, > 8 Å (white point represents parent structure).

2.5.3. Results and Discussion

The initial step in identifying the electrostatic differences between SUMO-4 and SUMO-2 was to replace each position in SUMO-4, which differs from SUMO-2, with the corresponding SUMO-2 amino acid, one at a time. The effects of these mutations were quantified using electrostatic clustering and solvation free energies of association as shown in Figure 2-16. Of these 14 mutations, most had only minor effect on the solvation free energies of association, which have been shown to be predictive of binding ability . Only the mutation of SUMO-4

residue 63 from glycine to aspartic acid had a noticeable effect on the predicted binding ability (bottom of Figure 2-16), which favorably decreases the solvation free energy of association by approximately 40 kJ/mol. Despite having no effect on the electrostatic character of SUMO-4, a previous study has shown that mutating position 90 from proline to glutamine (90 P2Q) produces a SUMO-4 analog that is capable of being processed by SENP2 . It is believed that proline at position 90 introduces a kink in the C-terminal tail of SUMO-4 preventing hydrolysis and maturation; therefore, the 90 P2Q mutation was included in all designed analogs.

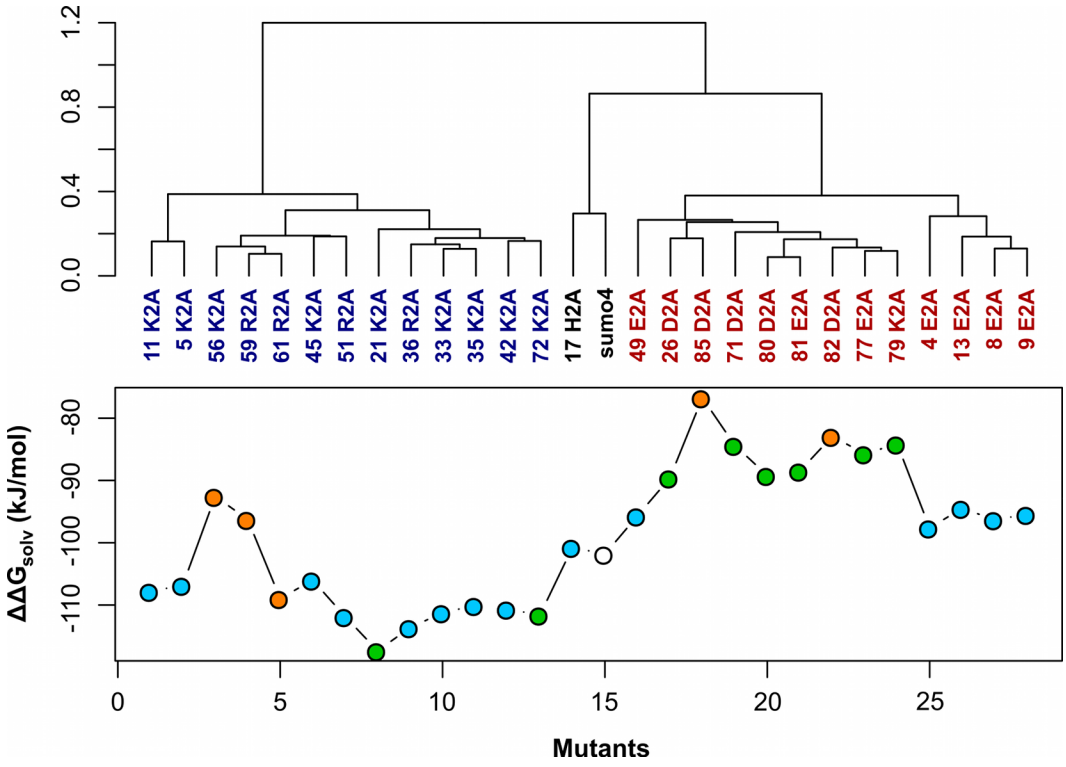


Figure 2-17 *Electrostatic clustering and free energies of association for SUMO-4 alanine-scan.* Electrostatic clustering is illustrated as a dendrogram tree (top) with label color indicating the type of amino acid being replace by alanine: red, acidic; blue, basic. The color of the free energy data points indicates the distance of the mutated residue from SENP2 as illustrated by Figure 2-15: orange, < 3.5 Å; green, > 3.5 Å and < 8 Å; cyan, > 8 Å (white point represents parent structure).

After analyzing the effects of each of the SUMO-2/SUMO-4 replacements, the next step was to elucidate the role of each charged residue of SUMO-4 in association. This was achieved

using alanine-scan mutagenesis in which each charged residue is replaced with alanine, one at a time. The electrostatic clustering and free energy results are presented in Figure 2-17. As is typical for most AESOP alanine-scan analyses , the dendrogram (top of Figure 2-17) exhibits three primary clusters containing mutations of only a single amino acid type: positive, negative, or neutral. On average, most mutations of positive residues are favorable, while most mutations of negative amino acids are unfavorable; however, the degree of these effects are dependent on the distance of the mutated residue from SENP2, as well as the surrounding protein/solvent environment. Of the alanine scan mutations, 21 K2A is predicted to be the most favorable according to its free energy value. A charge-reversal scan was also performed (data not shown), and resulted in comparable trends to the alanine scan mutations, but with the effects being exaggerated.

Based on the assumption that electrostatics is responsible for the poor association between SUMO-4 and SENP2, the net charge difference between SUMO-2 and -4 needs to be equalized in order to design a SUMO-4 analog with SUMO-2 like association. Therefore, a new SUMO-4 analog was designed by combining the most favorable mutation from SUMO-2 based mutations with the most favorable alanine-scan mutation. This new analog, referred henceforth as K21A/G63D/P90Q, has a net charge of -2 and requires one more mutation to reach the net charge of SUMO-2. A second alanine-scan was performed based on the K21A/G63D/P90Q analog to identify potential mutations that would result in binding similar SUMO-2. For this analysis the favorability of the mutations were not only ranked based on electrostatic free energies of association, but also based on electrostatic similarity to SUMO-2. This stems from the fact that analogs with comparable free energy values and similar electrostatic potentials are predicted to have similar recognition and binding, as defined by the two-step association model

(Figure 2-1A). Figure 2-18 contains the dendrogram and free energies corresponding to the K21A/G63D/P90Q alanine-scan calculations. As would be expected, SUMO-2 clusters with analogs containing an additional mutation of a basic amino acid, which results in the same net charge as SUMO-2. Additionally, the majority of all mutations predicted to be favorable are away from the binding interface, at > 8 Å from the nearest SENP2 atom.

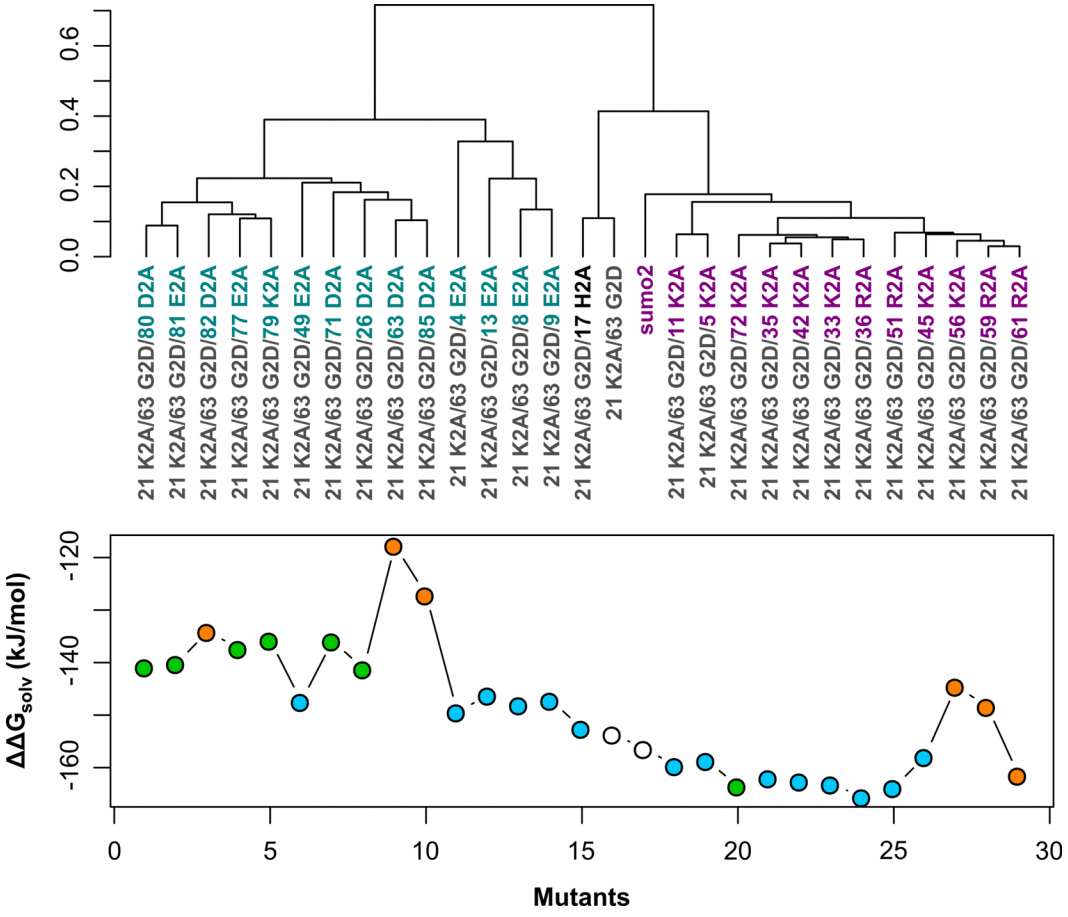


Figure 2-18 *Electrostatic clustering and free energies of association for the SUMO-4 K21A/G63D/P90Q alanine-scan.* Electrostatic clustering is illustrated as a dendrogram tree (top) with label color indicating the net charge of the analog: blue green, -1; gray, -2; purple, -3. The color of the free energy data points indicates the distance of the mutated residue from SENP2 as illustrated by Figure 2-15: orange, < 3.5 Å; green, > 3.5 Å and < 8 Å; cyan, > 8 Å (white point represents parent structure).

The final perturbative design approach that was applied in designing SUMO-4 analogs involved permutations of 5 single mutations identified from the SUMO-2 based mutations and

the SUMO-4 alanine scan: K21A, R51A, G63D, G69E, and K72A. Based on these 5 mutations 10 unique analogs are possible, and electrostatics clustering and free energies for these analogs are presented in Figure 2-19. These SUMO-4 analogs show slightly less similarity with SUMO-2, when compared to the analogs of Figure 2-18, with respect to both clustering and free energies. However, the analog G63D/G69E/K72A is of interest since it contains only mutations from the set of SUMO-2 based mutation.

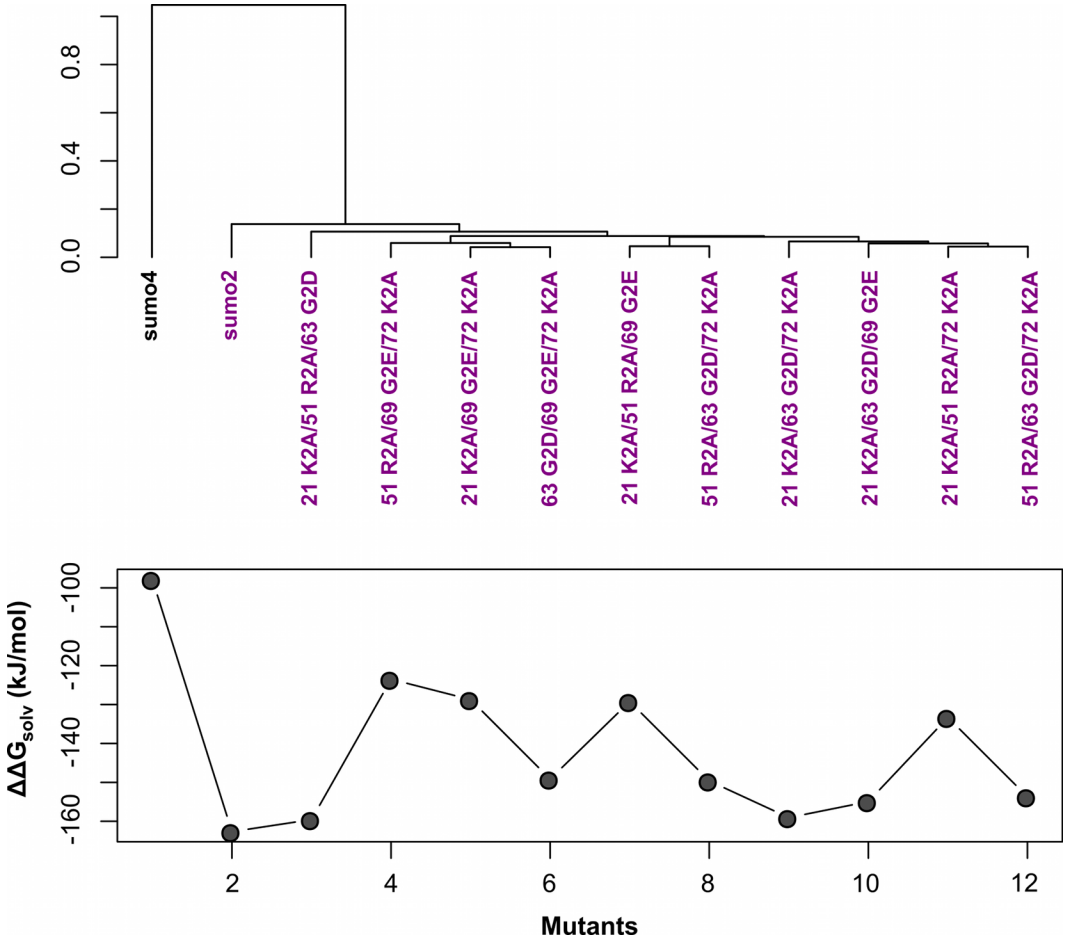


Figure 2-19 *Electrostatic clustering and free energies of association for SUMO-4 permutation analogs.* Electrostatic clustering is illustrated as a dendrogram tree (top) with label color indicating the net charge of the analog: black, 0; purple, -3.

Table 2-4 *List of predicted SUMO-4 analogs based on perturbative electrostatic design.*

#	Mutations	Net charge
1	P90Q	+0
2	P90Q/G63D	-1
3	P90Q/G63D/K21A	-2
4	P90Q/G63D/K21E	-3
5	P90Q/G63D/K21A/R36A	-3
6	P90Q/G63D/G69E/K72A	-3

Table 2-4 contains a list of proposed SUMO-4 analogs based on the four perturbative design approaches. P90Q has been proposed previously to improve SUMO-4 as a substrate for SENP2 [73], and therefore is included in all analogs. G63D was the most favorable mutation with respect to electrostatics (Figure 2-16), and therefore has also been included in most of the designed SUMO-4 analogs. Of the six proposed analogs of SUMO-4, three have been experimentally expressed and evaluated (Liu, Kieslich, Morikis, Liao – In progress). Table 2-5 contains Michaelis–Menten parameters for SUMO-2 and the three design SUMO-4 analogs as substrates for SENP2. The k_{cat}/K_M parameter is understood as the overall enzymatic efficiency of an enzyme and substrate, and was noticeably improved by all SUMO-4 mutations, since native SUMO-4 shows no catalytic activity with SENP2. The double SUMO-4 mutant P90Q/G63D shows the most improvement, with a k_{cat}/K_M value within about an order of magnitude of that of SUMO-2. However, the P90Q/G63D/K21A analog, which was predicted to be the most favorable of the experimentally evaluated analogs, resulted in a lower k_{cat}/K_M when compared to P90Q/G63D.

To gain further insight into the experimental results for the designed SUMO-4 analogs, we analyzed the inter/intramolecular Coulombic interactions that may contribute to the

improved/weakened association of the tested analogs. The strong favorability of G63D is easily understood after observing the placement of the aspartic acid between two basic amino acids forming a strong bifurcated intramolecular salt bridge (Figure 2-20A). Further analysis showed that a negative amino acid is conserved in all SUMO isoforms at position 63 (referred as position 67 in literature) [76]. Additionally, preliminary work elucidating the role of electrostatics in the interaction between SUMO-1:SENP2 also predicted that this conserved aspartic acid (63/67 D) is crucial to binding (Kieslich, Liao, Morikis – unpublished). Finally, comparison of the sequences of the SENP2 isoforms also shows that positive charge at the positions 456 and 459 is highly conserved; therefore, we propose that the G63D mutation, in combination with P90Q, not only makes SUMO-4 a good substrate for SENP2, but also for SENP1, 3, and 5. With regards to the lower k_{cat}/K_M for P90Q/G63D/K21A, replacement of lysine 21 with alanine removes intermolecular Coulombic interactions that stabilize local structure (Figure 2-20B). This local destabilization is most likely exacerbated by the fluorescent protein tags attached to the N and C-termini of SUMO-4 in the experimental system used.

Table 2-5 *Experimental enzyme kinetics parameters for designed SUMO-4 analogs.*

SENP2 Substrate	K_M (μM)	k_{cat}(s^{-1})	k_{cat}/K_M($M^{-1}\cdot s^{-1}$)
SUMO-2	0.054 ± 0.015	29.33 ± 2.571	(5.45 ± 1.590) x10^8
SUMO-4 P90Q	18.30 ± 3.670	15.34 ± 1.613	(8.38 ± 1.896) x10^5
SUMO-4 P90Q/G63D	1.65 ± 0.232	19.28 ± 0.520	(1.17 ± 0.168) x10^7
SUMO-4 P90Q/G63D/K21A	5.45 ± 0.837	21.20 ± 1.413	(3.89 ± 0.648) x10^6

Note: This data was collected and kindly provided by Yan Liu and Professor Jiayu Liao, Department of Bioengineering, University of California, Riverside.

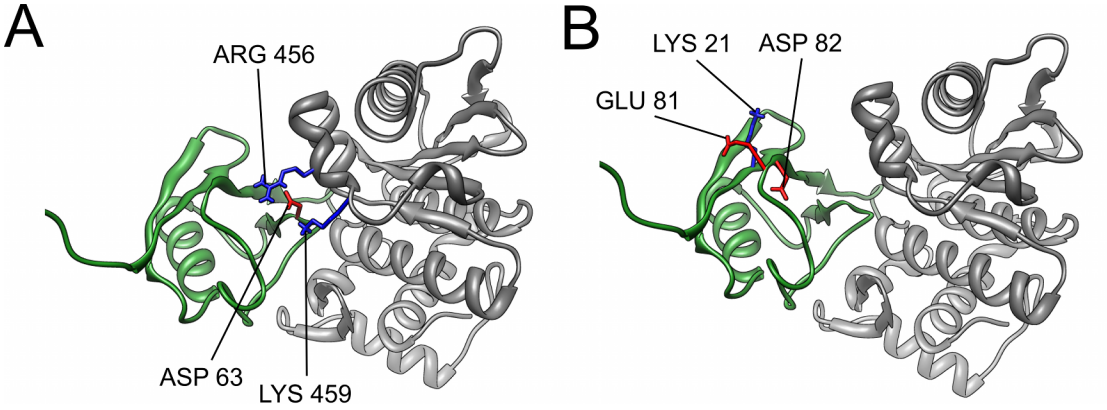


Figure 2-20 *Molecular graphic of SUMO-4 mutations.* Ribbon models are used to represent SUMO-4 (green) and SENP2 (gray). Charged residues of interest are in stick representation: red, negative; blue, positive.

2.5.4. Conclusion

In this study we applied electrostatic perturbative design using the AESOP framework to identify mutations of SUMO-4 that produce a suitable substrate for SENP2. Our predictions were experimentally validated using enzyme kinetics, and showed that the double mutant P90Q/G63D produced a SUMO-4 analog with a k_{cat}/K_M similar to the highly homologous SUMO-2. This result provides insight into the specificity of the SUMO/SENP families, since the charged residues involved in the gained interaction (namely SUMO-4 63, SENP2 456, and SENP2 459) are highly conserved. Additionally, this study marks the first application of the AESOP framework, in which the AESOP predictions were used to guide experimental studies, and therefore, serves as an example of the utility of the AESOP framework for electrostatic perturbative design.

3. THE ROLE OF ELECTROSTATICS IN THE COMPLEMENT IMMUNE SYSTEM

3.1. BACKGROUND

The complement system, a component of innate immunity and a link between innate and adaptive immunities, is a cascade of proteins whose proper function is vital to the body's ability to fight infection (as illustrated by Figure 3-1). Complement proteins attack foreign pathogens by covalently attaching to pathogen antigens, directly lysing pathogen surface membranes, and initiating inflammatory responses. The complement system functions through protein-receptor interactions and is tightly regulated by cell-bound or serum regulators. Complement system activation can be achieved by three distinct pathways: the classical pathway which involves the binding of the C1 complex to antigen-antibody complexes, the lectin pathway which involves the binding of mannose-binding-lectin (MBL) to carbohydrates located on pathogen surfaces, and the alternative pathway which involves the spontaneous activation-inactivation of C3, generation of C3b, and binding of C3b to pathogens (opsonization) [77]. All three pathways result in the generation of a C3 convertase whose function is the cleavage of C3 into C3a and C3b. A complex is then formed between the C3 convertase and C3b creating a C5 convertase, which is responsible for the cleavage of C5 in to C5a and C5b. The cleavage products C3a and C5a are collectively known as the anaphylatoxins, and are responsible for complement-mediated inflammation [77,78]. The complement fragment C5b, however, is ultimately involved in the formation of the membrane attack complex (MAC), which is responsible for bacterial cell lysis. A simple diagram illustrating the three pathways of complement activation, and their end results is shown in Figure 3-1.

Following activation, C3b becomes deposited on pathogen cell surfaces leading to formation of C3 convertatses and further activation of the complement system. In addition to C3b deposition, complement activation also leads to generation, as well as pathogen surface deposition, of cleavage product C3d, which is the thioester-containing domain of C3 [79]. C3d is a key component of complement activation, and upon binding to the B-cell receptor CR2 forms a link between innate and adaptive immunities [79]. The formation of B-cell receptor/co-receptor complexes (Figure 3-1) augments the adaptive immune response by increasing B-cell sensitivity to the antigen by 1,000 – 10,000 fold [79]. The ability of C3d to increase B-cell sensitivity to an associated antigen has led to numerous studies investigating the use of C3d as a vaccine adjuvant [79-82].

The complement system activation cascade is complicated further by the fact that C3b opsonization can occur on all cell surfaces, including host cells. Therefore, regulators, such as factor H, are recruited to prevent the formation of complement convertases on host cell surfaces (Figure 3-1), and lead to the degradation of C3b bound to host cells into an inactive form (iC3b). However, when this tight balance between activation and proper regulation fails, the complement system can attack own tissues typically leading to an autoimmune disease [77,78]. These instances of improper regulation are evidence of the need for complement-targeted therapeutics.

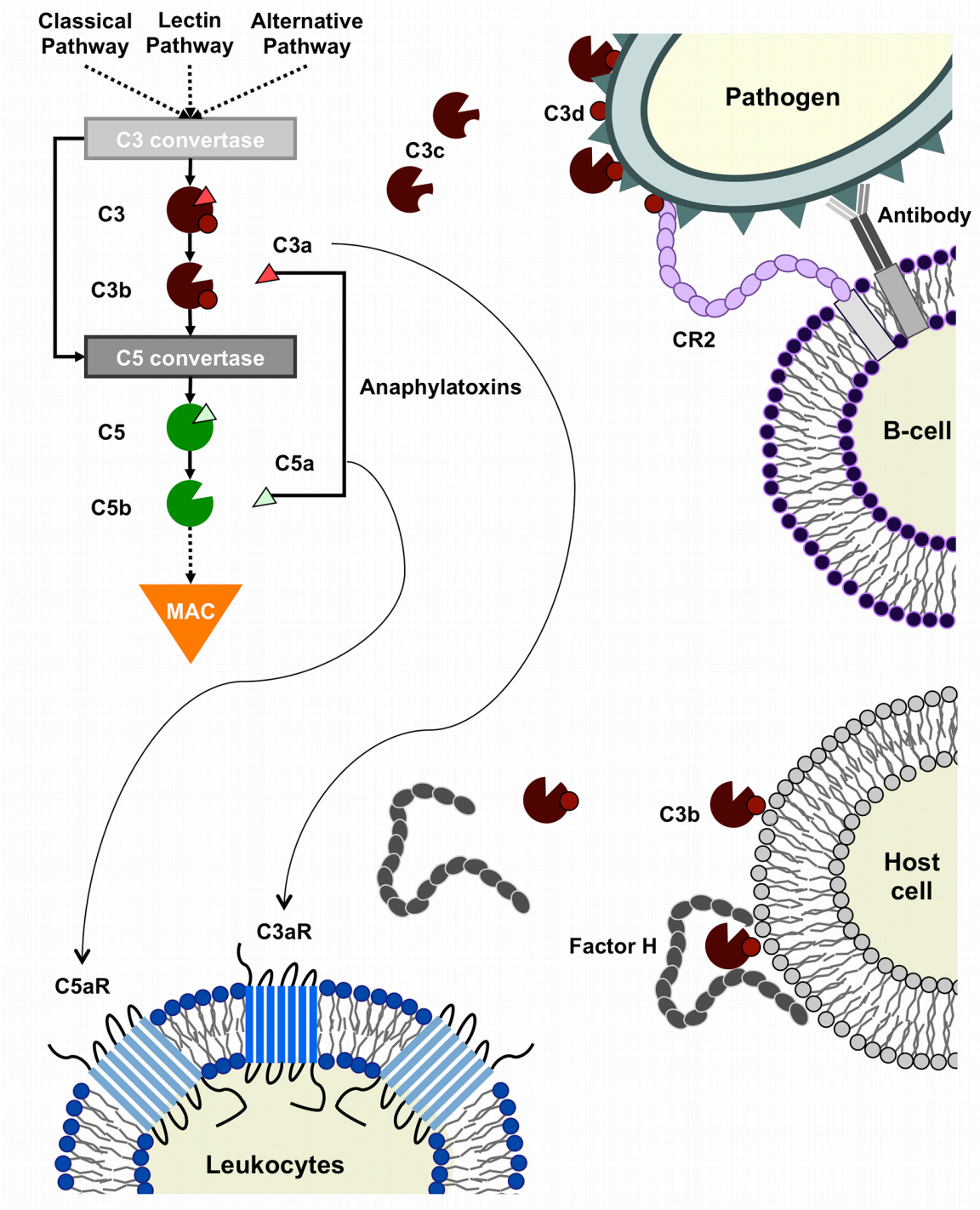


Figure 3-1 *Schematic of complement system activation, regulation, and outcomes.*

3.2. ELECTROSTATIC ANALYSIS OF FACTOR H

3.2.1. Introduction

Factor H (FH) is a regulator of the complement system [77,78,83] that has been implicated in a variety of diseases, such as age related macular degeneration (AMD) [84-88], atypical hemolytic uremic syndrome (aHUS) [89-92], and dense deposit disease (DDD, a.k.a. membranoproliferative glomerulonephritis Type II) [89,93]. The complement system is part of innate immunity and a link between innate and adaptive immunities, whose function is tightly regulated to avoid immune attack against own tissues. FH is an essential regulator of the complement system whose absence results in breaking down the prevention of complement attack against self-tissues [77,83]. FH functions by binding to complement protein fragment C3b and acting as cofactor to serine protease Factor I (FI), which cleaves and inactivates C3b. When C3b is bound covalently to pathogen surfaces, it is amenable to binding by FB which is subsequently activated and forms the convertase enzyme complex C3b-Bb. The convertase is responsible for amplified cleavage of C3 molecules and generation of additional C3b fragments that bind (opsonize) pathogen surfaces. A structural biology view of the C3 function has recently been presented [94-96].

The essence of FH regulation of the complement system depends on the charge composition of the surfaces it is capable of binding and on competition with FB [77]. FH binds with higher affinity to C3b than FB on host cell surfaces that are rich in polyanions, such as sialic acid and glycosaminoglycans (GAGs). On the contrary, FB binds with higher affinity to C3b than FH on pathogen surfaces that lack polyanion coating [77]. The capability of FH to interact with

polyanions implicates charge in its function and renders the study of the electrostatic properties of FH necessary to understand its function.

Charge plays a role in FH's association to disease through specific amino acids or groups of amino acids that form surface hotspots. Mutations of ionizable amino acids are associated with disease. FH is implicated in AMD through a single nucleotide polymorphism (SNP) that involves tyrosine/histidine change [84,86-88]. This SNP genetically predisposes part of the population to AMD. Polyanion-binding sites in FH have also been implicated in AMD and in aHUS (reviewed in [97,98]). FH is implicated in DDD through a lysine deletion [99].

FH has a chain-like structure consisting of 20 modules. These modules have the complement control protein (CCP) module topology and are called CCP modules [100]. They resemble compact ellipsoids that are connected with short and flexible loops. There are many high-resolution structures of FH CCP modules and one low-resolution structure of whole FH, which suggests a folded back topology that brings together the N- and C-termini [101]. In addition, the structure of the complex between the four N-terminal CCP domains of FH and C3b is available and provides the first insight on the specifics of the interaction [102]. Each FH CCP module is made of 51-62 amino acids and has a core stabilized by two disulfide bridges. Each loop is made of 3-8 amino acids. Despite similar structures and stabilities of FH CCP modules, there is a wide diversity in physicochemical properties, and most importantly in charge composition. The basic hypothesis of our study is that the charge (and electrostatic) diversity of CCP modules renders FH its binding properties to C3b and polyanion-coated surfaces. C3b is also a modular protein consisting of 8 macroglobulin (MG) domains and 5 other domains and linkers [102,103]. Despite their similar structures, the MG domains of C3b also have charge (and electrostatic) diversity.

In a recent review, the interactions of FH and their roles in complement regulation and disease have been discussed. However, the importance of charge and electrostatic interactions was not emphasized. Therefore, the goal of our study is to perform a comparative electrostatic analysis of all 20 CCP modules of FH and all 8 MG domains of C3b to identify electrostatic hotspots and predict potential binding sites. We also performed binding analysis of the structure of the FH(CCP1-4)-C3b and C3b-Bb complexes to gain insight on the C3b binding mechanism for competing proteins FH and Bb. The method we use is theoretical and is based on the calculation of electrostatic potentials using the Poisson-Boltzmann equation [3]. We have previously applied similar protocols in other proteins that contain CCPs and whose function is driven by electrostatic interactions, as evidenced by experimental data [24,29,30,35-37,104]. This is an atomic resolution study in which three-dimensional structures of the 20 CCP modules are needed. The basis of our study is the use of a clustering method to classify similarities or dissimilarities of the spatial distributions of electrostatic potentials of the various CCP modules of FH. We place special emphasis on CCP modules that are known from experimental data to bind to C3b and polyanion-coated surfaces, such as GAGs and sialic acid found in human cells and tissues. We also discuss those CCP modules that are implicated directly to disease. For example, the SNP and polyanion-binding hotspot implicated in AMD are located in CCP7. A polyanion-binding hotspot implicated in aHUS is in CCP20. A loss of charge mutation implicated in DDD is in CCP4. Finally, we make binding predictions for CCP modules of unknown function.

3.2.2. Methods

Our atomic-level calculations are possible when three-dimensional structures are known at high resolution. FH has been the focus of several structural biology studies, from which

several crystallographic or NMR structures for CCP modules have been determined. Currently the structures of 14 CCP modules are available, and in most cases multiple structures exist for single, double, triple, or quadruple modules (Appendix F). Modules with experimentally determined structures are: CCP1-8, 12, 13, 15, 16, and 19-20. Therefore, our initial step was the examination of all available CCP structures to (i) select those that are more suitable for our calculations and (ii) select those that are suitable to be used as templates to generate homology models for the 6 CCPs without experimentally-determined structures. Our criteria for (i) were quality (resolution), completeness of the structure with maximum amount of secondary structure, and lack of selenomethionine (MSE) amino acids. Appendix F and Appendix G provide lists of the experimentally derived structures used and reasoning. Our criteria for (ii) were quality of sequence alignment and percent identity. To generate homology models for CCP9-11, 14, 17, and 18 we first obtained the amino acid sequences from the SMART database and subsequently we used sequence alignments with the CCP modules of known structure. Here, the CCP module is defined as the sequence between and including the first and fourth conserved cysteines. Amino acid sequence alignment scores were produced for all possible target-template pairs using ClustalW [40]. The templates that produced alignments with the highest percent identity score, while still providing proper alignment of the four cysteines and a conserved tryptophan close to the C-terminus were chosen. The CCP module sequences and template structures, as listed in and, were imported into the Automodel module of the homology modeling package Modeller 9v5 . After generation, each model was visualized using Deep View [105] and inspected for van der Waals clashes, secondary structure quality, and disulfide bond correctness. The complete set of structures for all 20 FH CCP modules was superimposed using Deep View for comparison. Superimposition was based on C_{α}-atoms using CCP16 as reference,

because CCP16 contained the highest quality secondary structure. The coordinates of each CCP module were saved in the orientation of the superimposition. This was necessary to assure that all CCP modules had the same center and coordinate fit in the box used in subsequent Poisson-Boltzmann calculations.

Electrostatic potential calculations were performed using the Adaptive Poisson-Boltzmann Solver (APBS) [3], which utilizes a grid-based method to solve the linearized Poisson-Boltzmann equation. Prior to electrostatic potential calculations, partial charges and atomic radii were assigned using the software PDB2PQR and the PARSE force field . A dielectric coefficient of 2 was used for the protein whereas a dielectric coefficient of 78.5 was used for the solvent. Two sets of calculations were carried out at ionic strengths corresponding to 0 mM and 150 mM concentrations, assuming +1/−1 charges for the counterions. The dielectric surface was defined using a sphere with probe radius of 1.4 Å and selecting the contact surface, whereas the ion accessibility surface was defined using a sphere with probe radius of 2.0 Å. Since the amount of screening differs between the two ionic strength parameters, the physical dimensions of the box used for each set of calculations was set to different values to ensure that the largest magnitude values were captured, while still providing maximum resolution. The grid size was 129 × 129 × 129 points. A box size of 110 Å × 110 Å × 110 Å was used for the calculations corresponding to 0 mM ionic strength, whereas a box size of 60 Å × 60 Å × 60 Å was used for the calculations corresponding to 150 mM ionic strength. All CCP modules were centered on CCP16 in electrostatic potential calculations. The spatial distributions of electrostatic potential were visualized and plotted in the form of isopotential contours using the molecular graphics program Chimera [106].

After calculating the electrostatic potentials, the final step was the clustering of spatial distributions of electrostatic potentials of the 20 FH CCP modules. A distance matrix was generated based on an electrostatic similarity distance (ESD). An ESD is a measure of the amount of dissimilarity between the spatial distributions of electrostatic potential for two homologous proteins and has been used previously in different protein clustering studies [21,23,25,38,107-110]. The ESD used in this study is equivalent to an average difference, and was calculated by taking the average of the normalized error between all corresponding grid points according to Eq. 2-6. According to this scheme, two proteins (CCP modules here) are identical if they have ESD 0 and totally different if they have ESD 1. The distance matrix was produced by calculating the ESD between every CCP module pair, resulting in a 20 × 20 array. Distance matrix generation and hierarchal clustering using average linkage was performed using in-house R scripts that utilize functions of the R-base package [48]. The clustering results are presented as dendrograms using the R plot functions.

Clustering was also performed for the C3b MG domains using a similar procedure. The C3b structure with PDB Code 2I07 was used [102]. Since the MG domains exhibit more structural and sequence variability than the FH CCPs, a more elaborate alignment procedure was required. The “Match Maker” module of Chimera was utilized in this instance, since it includes a secondary structure score with a variable weight that provided optimal coordinate superposition. The MG domains were aligned to MG2 since this provided the best alignment. The grid size used for the MG domain calculations was 129 × 129 × 129 points. A box size of 135 Å × 140 Å × 130 Å was used for the calculations corresponding to 0 mM ionic strength, whereas a box size of 65 Å × 85 Å × 60 Å was used for the calculations corresponding to 150 mM ionic strength. All CCP modules were centered on MG2 in electrostatic potential calculations.

3.2.3. Results and Discussion

The 20 CCP modules of FH are structurally homologous, but have diverse sequences and physicochemical properties. Because of the striking charge diversity among individual FH CCP modules, we initiated this structural bioinformatics study to examine the electrostatic properties of CCP modules in terms of net charge, spatial distributions of charge and electrostatic potentials. Our goal is to classify CCP modules according to the spatial distributions of their electrostatic potentials and to make correlations with the binding properties and pathobiology of FH. These correlations are based on our clustering analysis of electrostatic potentials and knowledge from available experimental and clinical data.

The rationale for our study is that if electrostatics is responsible (or contributes) to binding, we can identify FH CCP modules with similar binding abilities. Besides electrostatic interactions with C3b, FH interacts with negatively charged surfaces because it has binding sites for sialic acid and GAGs, including heparin. Also, if electrostatics plays a known role in disease, through the presence of electrostatic hotspots in the surface of certain CCP modules, then other CCP modules with similar electrostatic profiles may also play a role in the specific disease.

FH is a long and flexible chain-like protein consisting of 20 modules, with each module having distinct properties and role in binding to PBS or C3b/C3c/C3d fragments or other molecules. Also, individual CCP modules have been associated to FH-mediated diseases, such as AMD, aHUS, and DDD. It is reasonable to decompose the binding and pathobiological properties of FH to respective properties of individual CCP modules. Although sequence identities are in the range of 5-40%, a major contributor to the formation of tertiary structures converged to the CCP module topology is the presence of two conserved disulfide bridges. (For clarification, 16 CCP modules have 30-40% identity with one or more CCP modules and 4 have 21-29% identities

with one or more CCP modules; Appendix A.) Our efforts are focused on the role of charge distributions within the tertiary structures and on classifying similarities/dissimilarities of spatial distributions of electrostatic potentials derived from charge. This work provides useful electrostatic analysis to form hypotheses and make predictions for interactions and biological function of FH.

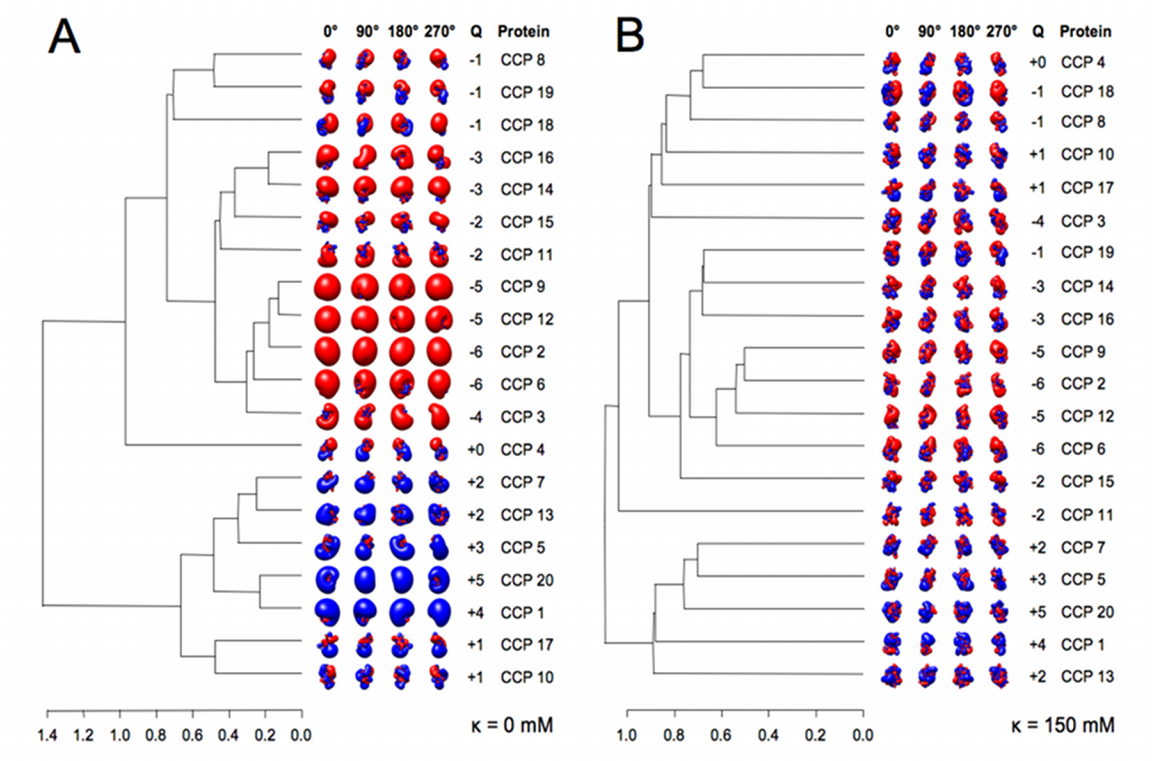


Figure 3-2 *Clustering diagram of the 20 CCP modules of FH according the spatial distributions of their electrostatic potentials.* Calculations of electrostatic potentials were performed using ionic strengths corresponding to 0 mM (A) and 150 mM (B) counterion concentration. Isopotential contours are plotted at ±1 k_BT/e, with blue and red corresponding to positive and negative electrostatic potential, respectively. Four orientations are shown corresponding to rotations about the vertical axis, as indicated. The net charge, Q, of each module is marked in the figure.

Figure 3-2 shows the clustering of the spatial distributions of electrostatic potentials for the 20 CCP modules of FH at ionic strengths of 0 mM and 150 mM. The figure depicts which CCP modules have similar electrostatic profile, by taking into account not only charge, but also how

the charge-generated electrostatic potential is distributed in space. Although CCP modules with similar charges tend to cluster together, it is the three-dimensional structure of each module and the charge topology within each structure that renders the specific spatial distributions of electrostatic potentials. There are seven CCP modules with positive net charge: CCP1, 5, 7, 10, 13, 17, and 20 (net charges 1 to 5; Figure 3-2). Twelve CCP modules have negative net charge: CCP2, 3, 6, 8, 9, 11, 12, 14, 15, 16, 18, and 19 (net charges −1 to −6; Figure 3-2). Only the CCP4 module has 0 net charge (Figure 3-2). Because of local cancellation or enhancement of electrostatic potentials in spatial proximity, CCP modules with similar charges may cluster differently, but within the same superclusters (Figure 3-2).

At 0 mM counterion concentrations we observe the formation of two superclusters including modules with net charges $Q > 1$ and $Q < -1$, whereas the module with $Q = 0$ forms its own branch on its own within the negative supercluster (Figure 3-2A). Finer clustering occurs according to similarities in localized electrostatic properties of modules within each supercluster. At 150 mM counterion concentration we also observe the formation of two superclusters including modules with net charges $Q > 2$ and $Q < -2$. However, modules with net charges 0 and ±1 cluster within the negative supercluster, with 5 out of 6 such modules forming their own cluster (Figure 3-2B). This demonstrates the importance of ionic screening of Coulombic interactions, which may reflect on protein function. These data also suggest that spatial distributions of electrostatic potentials may be better predictors for electrostatically-driven protein function than net charge alone.

Our analysis focuses on the FH CCP modules, defined with boundaries of the first and fourth conserved cysteines. In addition, the flexible linkers of consecutive CCP modules may mediate electrostatic interactions if they contain charged amino acids. Appendix B provides the

linker and N- and C-terminal sequences to indicate if they are amenable to electrostatic interactions. The basic idea behind performing electrostatic potential calculations is that the excess charge in individual CCP modules is responsible for driving FH interactions with target proteins or surfaces. Excess net charge is responsible for non-specific recognition between two proteins or a protein and a ligand carrying opposite net charges, based on long-range electrostatic interactions . Recognition is followed by specific binding, involving predominantly short-range interactions, such as hydrogen bonds, salt bridges, and hydrophobic. The electrostatic clustering of Figure 3-2 presents a quantitative classification of CCP modules with similar electrostatic properties and perhaps similar functions.

Our data suggest that the 5 predominantly positively charged (Q > 2) CCP modules are potential interaction sites with polyanions (Figure 3-2). CCP modules clustering together may have similar binding properties. These are the CCP1, 5, 7, 13, and 20. An earlier low-resolution structure of FH has proposed a folded-back dynamic structure that may enable multiple modules to recognize the same polyanionic surface. Polyanion-binding sites in CCP19-20 and CCP6-8 have been discussed in the past in view of experimental [111,112] and structural data. Also, heparin-binding sites have been previously discussed for CCP7, CCP12-14, and CCP19-20 [111,113]. Figure 3-2 suggests that it is impossible for CCP6 to be a polyanion-binding site since it is excessively negatively charged. Modules CCP8 and 19 have low probability for binding polyanions since their total charge is −1; however, they both have patches of positive electrostatic potential hotspots, which cannot exclude a localized ability to bind polyanions (Figure 3-2). It may also be possible that modules CCP8 and 19 act as scaffolds to orient neighboring modules CCP7 and 20, respectively, for polyanion binding. Previously not discussed,

to the best of our knowledge, CCP1 and CCP5 may deserve attention as potential polyanion-binding site, according to Figure 3-2.

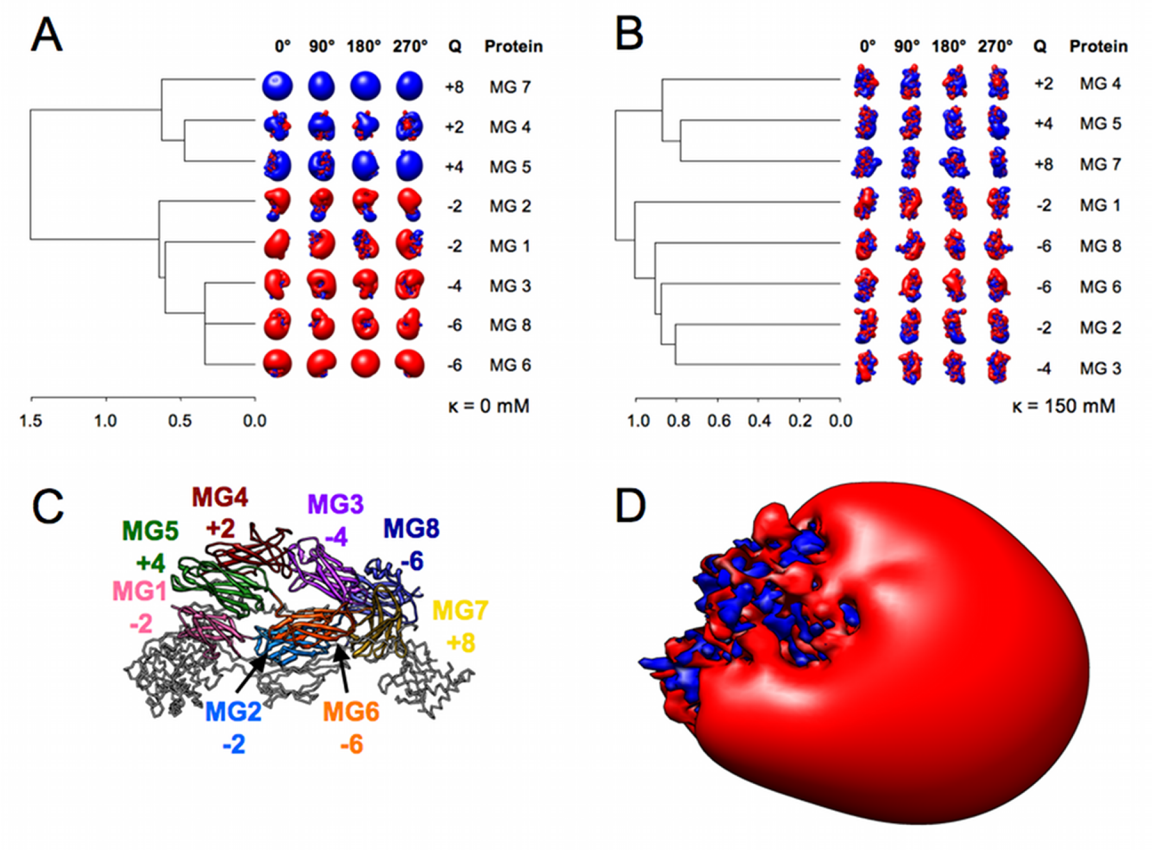


Figure 3-3 *Clustering diagram of the 8 MG domains of C3b (PDB Code 2I07) according to the spatial distributions of their electrostatic potentials.* Calculations of electrostatic potentials were performed using ionic strengths corresponding to 0 mM (A) and 150 mM (B) counterion concentration. Isopotential contours are plotted at ±1 k_BT/e, with blue and red corresponding to positive and negative electrostatic potential, respectively. Four orientations are shown corresponding to rotations about the vertical axis, as indicated. The net charge, Q, of each module is marked in the figure. (C) The 8 MG domains are shown in ribbon representations with different colors, as marked. The net charge of each MG domain is also marked. The amino acid ranges for each of the C3b domains are: MG1 (1-104), MG2 (105-209), MG3 (210-328), MG4 (329-426), MG5 (427-534), MG6 (535-577 and 746-806), MG7 (807-911), MG8 (1331-1474). (D) Isopotential contour plot for C3b at ±3 k_BT/e. The color code is blue for positive and red for negative electrostatic potential. The calculation of electrostatic potential was performed using ionic strength corresponding to 0 mM counterion concentration.

FH regulates complement activation by interacting with C3b. There are several experimental and structural studies that implicate CCP1-4 and 19-20 and possibly 7 and 12-14

(or 11-15) in binding to C3b [90,94,112,114-117]. CCP19-20 have also been proposed to bind to C3d and CCP12-14 have also been proposed to bind C3c , suggesting that different CCP modules may contact different C3b sites. Figure 3-3A,B shows the clustering of the spatial distributions of electrostatic potentials for the 8 MG domains of C3b, whose relative topology is shown in Figure 3-3C. As is the case with the CCP modules of FH, there is charge diversity in the MG domains of C3b, ranging from excess charge of −6 to excess charge of +8. Figure 3-3 depicts two superclusters for positive and negative CCP modules, with finer clustering within each supercluster based on the presence of localized charge hotspots and the inclusion or not of counterions in the calculations. There are other domains in C3b that are not homologous to MG domains and cannot be included in the clustering of Figure 3-3. Appendix I summarizes these domains, their structures and spatial distributions of their electrostatic potentials to indicate possible binding effects. The overall electrostatic potential of C3b is excessively negative (Figure 3-3D), with profound contributions from non-MG domains, such as a'Nt, C3d, CUB, LNK, and C345C with net charges −5, −4, −6, +2, and −12, respectively (Appendix I). Only 3 MG domains and LNK have positive net charges. Taking into account the dynamics of C3b, it is likely that there are transient openings that expose positively charged patches in C3b, which may be sites of interaction with negatively charged CCP modules of FH.

When we initiated this study, we expected, to a large extent, complementary electrostatic properties for interacting macrodipoles between C3b domains and FH modules. Although we recognize that the tertiary structure of C3b, which is rather compact, may introduce cancellations or enhancements in the electrostatic potentials at the vicinities of neighboring domains that contact each other, this may not be the case for FH whose modules form an extended chain-like structure. When the crystallographic structure of the complex

between FH(CCP1-4) and C3b was published [102], the following contacts were observed: (i) positive-negative net charge, CCP1-α'Nt and Linker(CCP1-CCP2)-α'Nt; neutral-negative net charge, CCP4-C3d and CCP4-MG1; positive-positive net charge, CCP1-MG7 and Linker(CCP1-CCP2)-α'Nt; negative-negative net charge, CCP2-α'Nt, CCP2-MG2, CCP2-MG6, CCP3-CUB, CCP3-MG1, and CCP3-MG2 (Figure 3-4). The types of contacts, decomposed to hydrogen bond and salt bridge (up to 3.5 Å), Coulombic (up to 8 Å), and hydrophobic interactions, are summarized in Appendix C, Appendix D, and Appendix E. In all cases, there are strong local pairwise interactions that stabilize the interacting FH(module)-C3b(domain) complexes. All interacting module-domain complexes involve at least one salt bridge and many hydrogen bonding, and Coulombic interactions. Hydrophobic interactions are more prominent in the case of CCP1 and 2 compared to CCP3 and 4. Figure 3-4B,C shows a surface projection of the electrostatic potential complementarity for the FH(CCP1-4)-C3b complex, depicting interacting positive-negative hotspots or positive/negative-neutral areas.

Based on our findings and guided by the crystal structure, we propose the following model for FH(CCP1-4)-C3b recognition and binding. The recognition process is driven by long-range interactions between the positive CCP1/Linker(CCP1-CCP2) and the negative α'Nt, which orient FH on the surface of C3b to favor specific short-range interactions and lock binding. Upon formation of the non-specific encounter complex, local structural rearrangements and solvent exclusion take place to accommodate favorable short-range interactions for binding. The binding step is dominated by the stronger and specific pairwise interactions in local areas of complementary charge hotspots or neutral-charged spots. In our analysis, above, long-range electrostatic interactions are depicted by the spatial distributions of electrostatic potentials and short-range electrostatic interactions are depicted by the surface projections of electrostatic

potentials. Our analysis shows that charge distributions and electrostatic potential distributions are more useful in understanding C3b-FH interactions than net charges alone. Our clustering analysis shows stronger ionic strength effect on neutral or slightly charged (+/– 1) modules/domains. The 0 mM clustering data of excessively charged modules/domains (|Q| ≥ 2), although non-physiological, contain less localized electrostatic hotspots, as would be expected by the inclusion of the dynamic character of amino acid side chains.

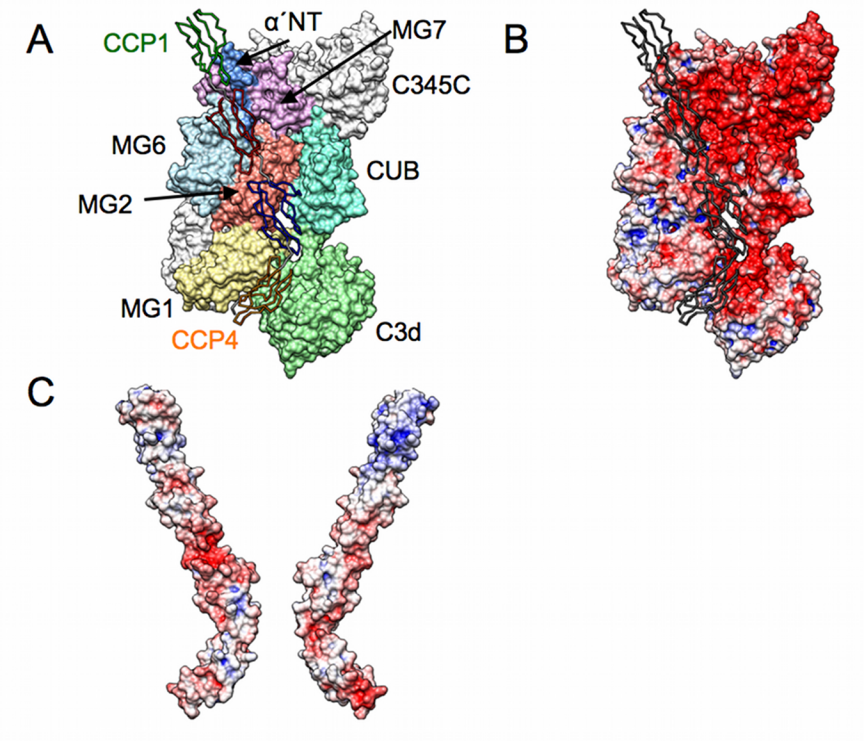


Figure 3-4 *Electrostatic nature of interaction between FH(CCP1-4) and C3b.* (A) FH(CCP1-4) in C$^{\alpha}$-trace representation and C3b in a surface representation. The 8 MG domains and the C3d, C345C, CUB, and α'Nt domains are shown in different colors and marked. CCP1 and CCP2 of FH are also marked. (B) FH(CCP1-4) in C$^{\alpha}$-trace representation and C3b in surface representation, colored by projection of its electrostatic potential. (C) FH(CCP1-4) in surface representation, colored by projection of its electrostatic potential. The right image of (C) is the contact surface of FH(CCP1-2) with C3b in the orientation of panel (B) and. Comparison of panels (B) and (C) depicts electrostatic potential complementarity. The color code in (B) and (C) is blue for positive and red for negative electrostatic potential.

Previous studies have shown that the FH module CCP7 is involved in AMD through the Tyr/His402 SNP, with His being the amino acid of the population at risk [84,86-88]. We have performed pKa calculations to investigate if Tyr402 (an acidic amino acid with model pKa of 9.6) or His402 (a basic amino acid with model pKa of 6.3) are found in conformationally strained environments. This would have been indicated by unusual shifts in pKa values; however this was not the case (data not shown). Indeed both Tyr402 and His402 are solvent exposed in the NMR structures [18], which suggests freedom of the amino acid in position 402 for involvement in specific side chain intermolecular interactions possibly with C3b. CCP7 has also a polyanion-binding hotspot on its surface [94] and overall positive electrostatic potential (Figure 3-2), which may be important for nonspecific interactions with negatively charged regions of C3b. Barlow and coworkers discuss the possibility of disruption of the ability of FH to bind GAGs upon mutation of Tyr402 to His402 [18]. They propose that the interaction of CCP7 with heparin or sulfated heparin oligosaccharides is driven by ring stacking interactions of the two tyrosine-rich regions of CCP7 (Tyr390, Tyr393, Tyr402, and 398, Tyr420) and sugar rings. This interaction may be perturbed upon removal of Tyr402 and replacement by His402 in the genetically predisposed population for AMD. A unit charge of the SNP amino acid, does not appear to play a role, as both Tyr402 and His402 are expected to be neutral at physiological pH since they are solvent exposed and our calculations do not suggest pKa perturbations. In essence, the model of Barlow and coworkers [18] implicates ring stacking in addition to overall charge complementarity in the case of heparin and GAGs in general. The same study provides a structural picture of these interactions using chemical shift perturbation mapping of NMR data and suggests the presence of two faces of interaction. Another study also suggests differential heparin binding for the Tyr402 and His402 variants of CCP6-8 [118]. Finally, the crystal structures of CCP6-8(His402)

(with the genetic variation His402) in complex with sucrose octasulfate demonstrates proximity of Tyr390 and His402 with sulfate groups [119].

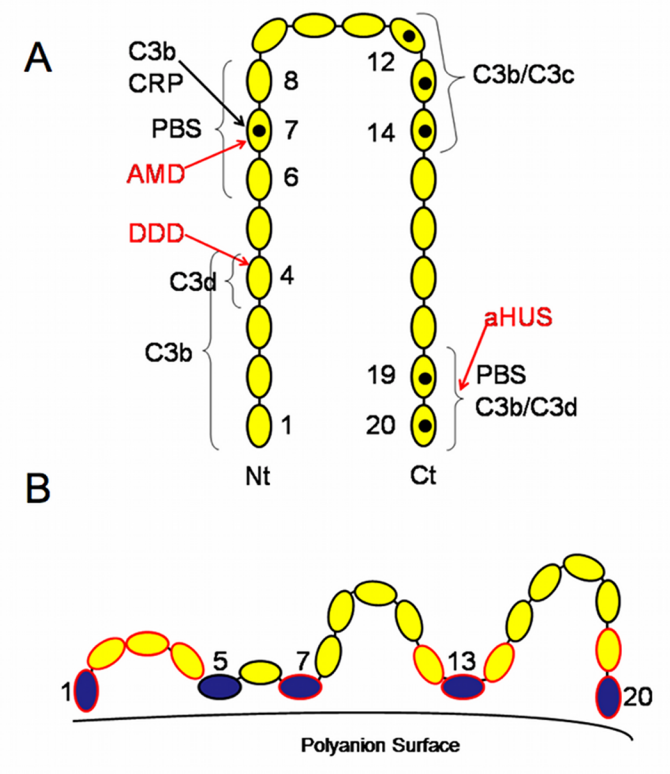


Figure 3-5 *Cartoon representing our model for FH interactions with polyanion-containing surfaces and C3b.* Other interaction sites are also marked. The CCP modules or module ranges discussed in text are numbered. (A) A simple model based on previous structural and mutagenesis data, which implicate CCP20 in anchoring polyanion-rich surfaces and CCP1-4, CCP7, 12-14 and 19-20 in interacting with C3b and its fragments. CCP modules that contain previously proposed polyanion-binding sites (PBS) are marked. The N- and C-termini are marked. The black dots indicate heparin binding sites. The CCP7 module interacting with C-reactive protein (CRP) is marked. CCP modules implicated in complement-mediated disease, such as AMD, DDD, and aHUS are marked in red. (B) A more elaborate model that incorporates electrostatic properties from our analysis. This model implicates CCP1, CCP5, CCP7, CCP13, and CCP20 in interactions with polyanion-rich sites. CCP modules colored in blue possess polyanion-binding sites and those with red borders are C3b-binding sites, in analogy with panel (A).

Module CCP4 is involved in DDD through deletion of Lys224 [99]. CCP4 is the only neutral module of FH, with balanced, and rather unusual, spatial distribution of electrostatic potential (Figure 3-2). According the FH(CCP1-4)-C3b crystal structure, CCP4 contacts the negatively charged domains C3d and MG1. It is likely that the deletion of the positively charged

Lys224 produces an imbalance in the spatial distribution of electrostatic potential of CCP4. This would favor excess negative electrostatic potential which may be disruptive for the binding of CCP4. It is possible that module CCP7 is also involved in DDD through the polyanion-binding hotspot [94]. Interestingly, there is a link between DDD and AMD because often patients have both diseases [120].

Modules CCP19-20 are involved in aHUS, because many aHUS-associated mutations are located in CCP20 [90,94,97,111,112,121,122]. Our data suggest that CCP20 is a polyanion-binding site because of its excess positive spatial distribution of electrostatic potential (Figure 3-2). On the contrary, CCP19, with net charge −1, has a somehow balanced spatial distribution of electrostatic potential (Figure 3-2). The role of CCP19 may be that of a scaffold to properly orient the structure of CCP20 or binding to C3b/C3d; however, a recent study suggests that CCP19 may also be a binding module to C3b/C3d [123], which may be possible through localized short-range interactions. It is likely that the observed mutations in CCP19-20 affect both functions of these modules, their polyanion-binding ability and their ability to bind C3b/C3d. A number of C3b mutations have also been reported to be associated with aHUS, however these mutations are enhancing or inhibiting mainly the interaction of C3b with membrane cofactor protein (MCP) [124]. Some of these mutations involve ionizable amino acids. Additional studies to identify similar mutations that affect C3b/C3d interactions with FH are necessary.

Figure 3-5A presents a previously published functional map of FH [117], including an additional C3d binding site for CCP4, identified by the crystal structure of FH(CCP1-4)-C3b [102]. These interactions are hypothesized to be responsible for regulation of the complement system in self-surfaces. The presence of a complex bending of FH is likely, to allow contact of the polyanion-binding site of CCP6-8 with cell surfaces; however, according to our data, only the

highly positively charged CCP7 of this triplet is capable of binding to polyanion sites. We propose a modified model which includes anchoring of FH on cell surfaces through polyanion-binding sites in the 5 excessively positively charged modules CCP1, 5, 7, 13 and 20. This binding allows for the following sets of interactions with C3b/C3d: (i) CCP1-4, 7, 12-14, and 19-20 with C3b; and (ii) CCP4 and 19-20 with the C3d portion of C3b. As C3d is the opsonization site, it must be in proximity to the cell surfaces, and so can be the positively charged modules CCP1, 5, 7, 13, and 20. This can happen in a 1:1 complex, which would bring CCP4 and 19-20 in a topology surrounding C3d, or in bivalent complexes. Recent studies have suggested that FH undergoes self-association in the presence of polyanions [121,123,125,126]. It is likely that the function of FH and implication to disease is more complex than presented here. However, the presented electrostatic arguments for association with polyanions and disease-related mutations are applicable for monomers and also for higher order self-associated complexes.

Our model proposes non-specific electrostatic interactions with polyanion surfaces and specific electrostatic (and non-electrostatic) interactions with C3b. Non-specific interactions with polyanions in self surfaces result in high concentration of FH molecules, which become available for C3b binding and subsequent C3b degradation. The number of C3b molecules is sparse in self surfaces because of lack of (or reduced) amplification loop of the complement system alternative pathway. On the contrary, in nonself surfaces there is less concentration of FH molecules and excess of the C3 convertase C3b-Bb. The latter is responsible for the amplification of C3 cleavage, which generates numerous copies of the C3b opsonin for pathogen surface coating, and subsequent elimination by phagocytosis. We can also factor in this model possible auto-inhibitory association interactions between positive and negative CCP modules (or

hotspots). This type auto-association may be present in free FH, which, in the presence of PBS, unwinds to accommodate binding to C3b.

Barlow and coworkers have previously presented a binding model involving anchoring of CCP20 to self-surface through "C3b/polyanion composite binding site", with CCP7 acting as a mediator and CCP1-4 disrupting the formation and stability of C3/C5 convertases [117]. The proposed mediating ability of CCP7 was called "proof-reading" to aid discrimination of self from non-self sulfation patterns [117]. The choice of the intermediate module CCP7 as the mediator may be based on the fact that it is part of the binding site of many bacteria known to interact with FH [89,100,111]; however, there are bacterial binding sites that do not include CCP7 [111]. The hypothesis of Barlow and coworkers is attractive and overall in line with our electrostatic data.

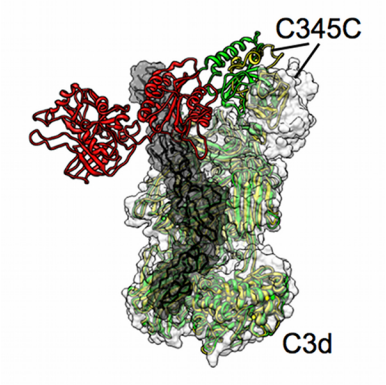


Figure 3-6 *Competition in the formation of C3b complexes.* Superposition of the C3b-FH(CCP1-4) and C3b-Bb (C3 convertase) complexes and free C3b. The structure of the C3b-FH complex (C3b in white, FH in black) is shown in surface representation, while the C3b-Bb complex is shown in ribbon representation (C3b in green, Bb in red) and free C3b is shown in yellow ribbon representation. Conformational transitions involving mainly the C345C domain of C3b are identifiable in the figure. The PDB Codes are 2WII for C3b-FH(CCP1-4), 2WIN for C3b-Bb, and 2I07 for free C3b.

A fundamental question rises on the physicochemical basis of the competition between FH and the C3 convertase C3b-Bb for C3b binding on self and non-self surfaces. Figure 3-6 shows

a superposition of the C3b-FH(CCP1-4) and C3b-Bb complexes, which demonstrates how FH(CCP1-4) binding can sterically hinder C3b-Bb access to its binding site [102]. According to our model FH(CCP1-4) binding is facilitated by non-specific electrostatic interactions between positive CCP modules and polyanion sites in self surfaces. FH(CCP1-4) binding is initiated by the electrostatic recognition of CCP1 and α'Nt. Upon FH binding, a conformational transition is observed in C3b involving moving away of the negatively charged C345C domain (Q = −12) (Figure 3-6), possibly owed to non-specific electrostatic repulsion with CCP2 (Q = −6) and CCP3 (Q = −4). Conformational transition in the opposite direction occurs upon Bb binding to C3b involving the C345C domain (Figure 3-6). In summary, it is likely that synergistic binding of the positive FH CCP modules to polyanion surfaces and specific C3b-FH contacts contribute to the formation of a stronger C3b-FH complex than a C3b-Bb complex on self surfaces. Although most of the binding studies are based on experimental mutagenesis studies or CCP module deletions, our electrostatic and clustering analysis together with structural data provide additional tools to guide future experiments that are needed for refining the proposed model of C3b-FH interaction. The electrostatic data are useful to establish mechanistic hypotheses for C3b-FH and C3b-Bb interactions.

3.3. THE TWO SIDES OF COMPLEMENT C3D: EVOLUTION OF ELECTROSTATICS IN A LINK BETWEEN INNATE AND ADAPTIVE IMMUNITY.

3.3.1. Introduction

Complement C3d is a domain and cleavage product of complement protein C3, which is the central protein involved in the activation and up regulation of the complement immune system [79,127]. As a domain of complement C3, C3d is involved in the opsonization of pathogens through a highly reactive thioester bond [79,127]. Additionally, when attached to pathogen cell surfaces complement fragment C3d interacts with complement receptor 2 (CR2), which is expressed on the surface of B-cells, to form B-cell co-receptor complexes, a link between innate and adaptive immunities [79,127,128]. The C3d-CR2 interaction increases B-cell sensitivity to an antigen by 1000 – 10000 fold, making the interaction a critical aspect in the initiation of an adaptive immune response [79,129]. Due to the significance of the C3d-CR2 interaction, extensive research has been performed investigating the nature of the interaction [29,30,130-135], as well as possible approaches for utilizing this interaction in the design of new therapeutics and vaccines [79,81]. As has been discussed for many complement proteins, electrostatic forces contribute significantly to the function of C3d [29,30]. Often as a result of evolution many proteins, especially those of the complement system, contain clusters of like-charged residues, which generate regions of high electrostatic potential that are often referred as electrostatic “hot-spots” [136,137]. These electrostatic “hot-spots” tend to correspond with functional sites, since they can result in acceleration of protein association, and can stabilize protein complexes [9,138]. When referring to the functional sites of C3d, two opposing surfaces have been described: 1) CR2-face, a highly acidic concave surface known to be the binding site

of several host/pathogenic ligands (Figure 3-7, top; see also Appendix J and Appendix K); and 2) thioester-face, a basic surface surrounding the thioester bond utilized in covalent attachment to pathogen cell surfaces (Figure 3-7, bottom). The acidic "patch" has been shown to be involved in recognition and binding during the association of C3d to CR2 [133,135], as well as to bacterial inhibitors of the complement system [96,139,140]. The basic surface however, accelerates the covalent attachment of C3/C3d to pathogenic cell surfaces.

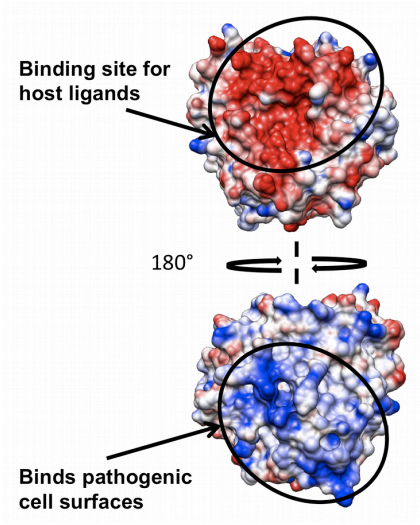


Figure 3-7 *Electrostatic potential of human C3d.* Electrostatic potential is projected onto the surface of C3d (PDB Code: 1C3D). The color transitions from red – white – blue when going from negative – neutral – positive electrostatic potential.

Evolutionarily, C3d is of great interest since it can be found in species that have been on the earth for 600 million years, while its interaction with CR2 is believed to have been gained after adaptive immunity, which first appeared in teleost fish [129]. Since electrostatics has been shown to play such a key role in the function of C3d, we propose that the electrostatic character of C3d has evolved to allow for optimal performance of both functions simultaneously. This follows from the logic that covalent attachment of C3d to pathogen cell surfaces, and the interaction between C3d and CR2 on the surface of B-cells, must both occur for this link

between innate and adaptive immunities to be activated. In this study, we investigate the presence of electrostatic "hot-spots" on complement C3d using a novel computational method involving perturbation of electrostatic properties of proteins. Additionally we probe the evolution of the electrostatic character of C3d, through the use of homology modeling, to gain insight into the role of electrostatics in the gained interaction with CR2, as well as the surface of the conserved thioester bond.

3.3.2. Methods

All calculations for human C3d, as well as all homology modeling, was based on the crystal structure of unbound human C3d (PDB Code: 1C3D) [135]. The sequence for human complement C3d was extracted from 1C3D and was used as a Blast query to identify C3d homologue sequences from the UniProt database [141]. The 23 C3d homologues were selected to optimize the range of sequence similarity when compared to human C3d (~37 – 85 % identity with human), but were chosen while keeping in mind that ~40% similarity is needed to ensure accuracy when performing homology modeling (Appendix P). As an initial comparison, a multiple sequence alignment comparing the 24 C3d homologues was generated using MUSCLE [142] and Bio3D [50] (Appendix L). The resulting alignment was used to populate a pairwise sequence identity matrix comparing the 24 C3d homologues. A second matrix comparing pairwise charge similarity was also generated by identifying the number of amino acid positions with a like charged amino acid for each pair of sequences, meaning the number of positions where both sequences have either K/R or D/E. The generated similarity matrices were used as input for hierarchical clustering using the R statistical language [48], and the results were

visualized using heatmaps with dendrograms. The functional regions for sequence clustering were defined based on the x-coordinates of the residues as illustrated by Appendix O.

Homology modeling was performed using the Automodel module of Modeller 9v9 [143] to produce structures for the 23 additional homologues of complement C3d. The crystal structure of unbound human complement C3d (1C3D) was used as the template structure. Sequence alignments between human C3d and the 23 C3d homologues were generated using MUSCLE in conjunction with the R package Bio3D, and supplied as input to Modeller. Each homology was subsequently superimposed onto 1C3D using Bio3D. The generated models were then prepped using PDB2PQR [42], in order to add atomic radii and partial charges according to the PARSE [43] forcefield to the coordinate files. Alanine-scan mutagenesis was performed using the coordinates of the 24 C3d homologues as parent structures and applying the AESOP computational framework, Analysis of Electrostatic Similarities of Proteins [25], which truncates each ionizable residue (Arg, Lys, His, Glu, and Asp) to alanine, one at a time.

A 20-ns explicit-solvent molecular dynamics simulation was configured using VMD [12] and performed using NAMD [11]. The crystal structure of human C3d (1C3D) was placed in a 82 Å × 82 Å × 82 Å water box with NaCl ions at 150 mM ion strength. The system was initially minimized using 1000 steps of energy minimization, to remove unfavorable torsions and contacts. The system was further relaxed by constraining the protein while allowing the solvent to equilibrate for 100 ps. Constraints were reduced linearly at 100 ps intervals, resulting in a total of 1 ns of equilibration. The simulation was performed using a 2 fs time step with rigid hydrogen bonds imposed by SHAKE.

All electrostatic potential calculations were performed using APBS [3], based on a grid with 129 × 129 × 129 grid points and length 98 Å × 116 Å × 116 Å. The solvent environment was

represented by a dielectric constant of 78.57 with a counterion concentration of 0 mM, while the protein dielectric constant was 20. Each electrostatic potential calculation was centered on 1C3D to ensure proper alignment of electrostatic potentials prior to similarity calculations. For each set of electrostatic potentials, cumulative distributions of electrostatic similarity (ESI) were calculated using the AESOP framework, according to the following expression:

$$ESI(i,j,k) = \frac{1}{N}\sum_{n=1}^{N} 1 - \frac{\left|\phi_A(i,j,k) - \phi_{B,n}(i,j,k)\right|}{max\left(\left|\phi_A(i,j,k)\right|, \left|\phi_{B,n}(i,j,k)\right|\right)} \quad \text{Eq. 3-1.}$$

Here, ϕ_A represents the electrostatic potential to which all other potentials are compared (parent), while $\phi_{B,n}$ represents the *N* members of the family of electrostatic potentials to be compared. The ESI is calculated at each grid point *(i,j,k)*, and normalized by *N*, the number of electrostatic potentials comparisons. For example, the ESI distribution for the C3d homologues (Figure 3-8B) was calculated by comparing the electrostatic potentials of the 23 homology models to the electrostatic potential of 1C3D. The surface projections of electrostatic similarity were generated using UCSF Chimera [106].

3.3.3. Results and Discussion

Through the course of evolution, in addition to sequence, electrostatic character is also often conserved. Conserved electrostatic potential can be responsible for acceleration and strengthening of protein-protein association, and is therefore indicative of the location of functional sites. Wade et al. have proposed computational methods utilizing homology modeling and Poisson-Boltzmann electrostatic calculations to quantitatively identify regions of conserved electrostatic character [22,109]. The approach calculates electrostatic similarity indices (ESI) to determine the cumulative spatial distribution of electrostatic similarity across a family of

homologous proteins. Similar ESI calculations were performed for 24 C3d homologues, based on homology models, as shown in Figure 2. The C3d homologues were chosen from a variety of species, at various evolutionary time points, and are diverse both in sequence (~36 – 84 % identity with human) and in net charge (-13 to +8). Isopotential contours for the electrostatic potentials of the C3d homologues further illustrate the diversity of their electrostatic character (Figure 3-8A). Despite large variations in the electrostatic potentials of the C3d homologues, the analysis identified two regions of high electrostatic similarity (Figure 3-8B; circled and indicated in red), which corresponded to the two functional sites of C3d. This conservation of the electrostatic potential surrounding the two functional sites further suggests a central role for electrostatics in C3d function; however, conserved electrostatic potential alone is not necessarily indicative of the existence of electrostatic "hot-spots".

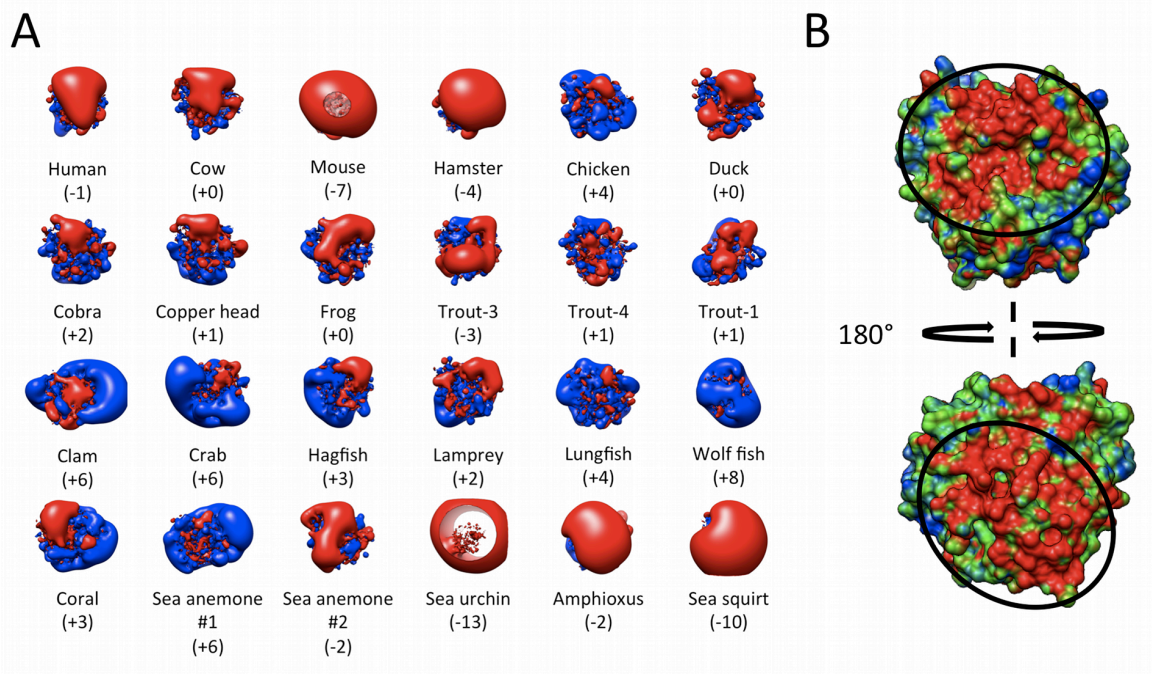


Figure 3-8 *Electrostatic similarity analysis of C3d homologues.* (A) Electrostatic potential distributions for 24 homologues of C3d [red, negative; blue, positive]. The net charge of each homologue is provided in the parentheses. (B) Cumulative electrostatic similarity distribution for 24 homologues projected onto the surface of human C3d [blue – green – red; low to high similarity].

If through the evolution of C3d, electrostatics has become crucial to function, it is plausible that the electrostatic nature of the protein would be resistant to perturbation. This stems from the assumption that a disruption in the electrostatic character of C3d would result in a reduction or loss of function. Furthermore, the necessity for electrostatic perturbation resistance is already suggested by the presence of clusters of like charged residues ("hot-spots"). To test this hypothesis, for the case of human C3d, we generated two sets of perturbed electrostatic potentials based on: i) dynamics and ii) mutations. For the dynamics, a 20 ns explicit-solvent molecular dynamics (MD) simulation was performed, from which 200 conformations of C3d were extracted. As for the mutations, a computational alanine-scan was performed using the AESOP framework [25], in which each charged residue was mutated to

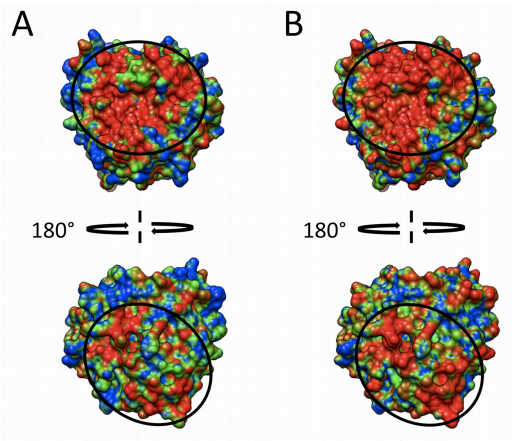


Figure 3-9 *Electrostatic similarity distributions for perturbed human C3d structures.* (A) Perturbation map for a 20 ns explicit-solvent MD simulation, based on 200 snapshots. (B) Perturbation map based on a theoretical alanine scan, consisting of 64 charged residue to alanine mutations. Color scheme is: blue – green – red; low to high similarity.

alanine, one at a time. For both sets of perturbed structures, the procedure used to compare the C3d homologues was also applied to generate ESI distributions to identify regions of high electrostatic similarity, or those regions least affected by perturbation. Surface projections of the ESI distributions, referred henceforth as perturbation maps, were generated for the two sets

of perturbed structures. Both perturbation maps, based on either dynamics or mutation, identify two similar regions with resistance to perturbation (Figure 3-9; circled and indicated in red). These regions correspond to the two functional sites of C3d (Figure 3-7), as well as the two regions of evolutionarily conserved electrostatic potential (Figure 3-8B). This resistance to electrostatic perturbation of the two functional sites, suggest the presence on an electrostatic "hot-spot" at each site, which compensate for changes in the character of these regions. The two electrostatic "hot-spots" are slightly larger and more distinct in the mutation-based perturbation map (Figure 3-9B), when compared to the dynamics-based (Figure 3-9A), and this arises due to the size of the perturbation. This is understandable since depending on the amount of conformational change, dynamics can have noticeable effect on electrostatic potential; however, both perturbation types are still quite modest, yet are able to identify these electrostatic "hot-spots".

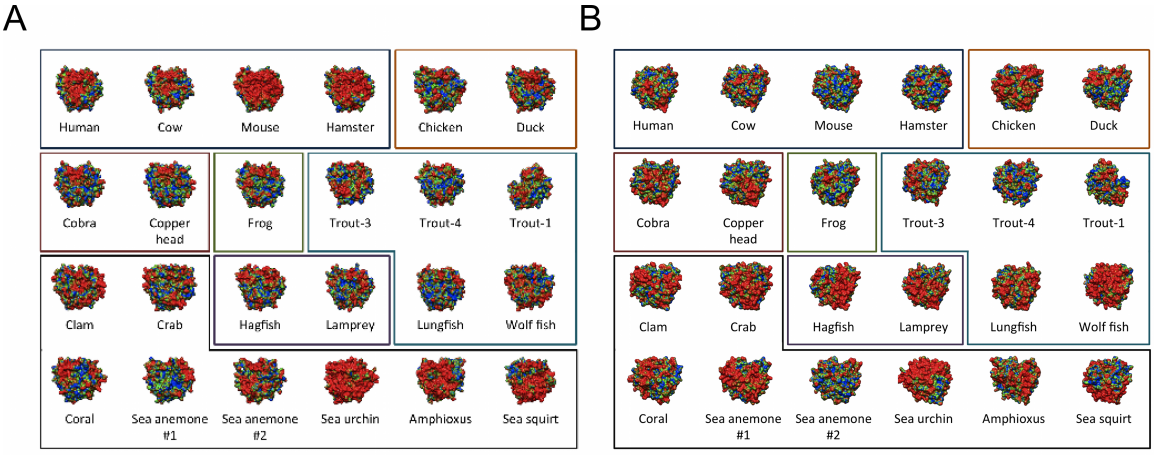


Figure 3-10 *Mutation-based perturbation maps for 24 C3d homologues.* The cumulative electrostatic similarity distributions for alanine-scan charge perturbations are projected onto the surface of each respective structure with two rotations: (A) CR2-face and (B) thioester-face. Perturbation map color scheme is: blue – green – red; low to high similarity. Colored boxes are used to group homologues from similar species. The box color code is as follows: mammals, dark-blue; birds, orange; reptiles, red; amphibians, green; fish, light-blue; jawless fish, purple; invertebrates, black.

Given the presence of perturbation resistance in the electrostatic character of human C3d, the question of how exactly did this characteristic come about arises. This question is even more interesting, when considering that C3d has gained the CR2 interaction, which is driven by electrostatics, over the course of evolution. It is possible that either the C3d electrostatic "hot-spot" has always been present and CR2 was simply opportunistic, or that the "hot-spot" has came to existence through co-evolution with CR2, which seems the most likely. To investigate whether the C3d electrostatic "hot-spots" are present in homologues of human C3d, we generated perturbation maps based on alanine-scan mutations for the remaining 23 C3d homologues (Figure 2-9). When comparing the perturbation maps for the CR2-face of the C3d homologues (Figure 2-9A), we find that the mammals (dark-blue box) are the only group of species in which all homologues exhibit the CR2 "hot-spot". This indicates that the electrostatic "hot-spot" of the CR2-face of C3d is something that has been gained through evolution. On the other hand, the electrostatic "hot-spot" on the thioester-face of C3d (Figure 2-9B) is much more predominate in lower species, such as the invertebrates (black box) and jawless fish (purple box), when compared to higher species like the mammals. The combination of these two results, the gain of the CR2 "hot-spot" and the reduction of thioester "hot-spot", is quite interesting, since it suggests a transition in the function of C3d. The two functions of C3d can be seen as opposing one another, and in order to optimize the new interaction with CR2, the conserved electrostatic "hot-spot" on the thioester face was reduced or lost, such is the case for mouse. There are exceptions to the mentioned trends, mainly the invertebrates (Figure 2-9; black box), which exhibit large diversity in net charge (ranging from -13 to +6) and electrostatic potential (Figure 3-8A). The invertebrates diverge first in the evolutionary tree, and therefore, have been evolving under their own pressures for much longer than any other group of species, which has

most likely been the cause of this increased diversity. Interestingly, this increased diversity has resulted in the C3d of amphioxus having very similar electrostatic character to human C3d, when comparing net charge and electrostatic potentials/"hot-spots", despite being separated by hundreds of millions of years in evolution. On a technical note, it should also be noted that due to the small size of the charge perturbations introduced by the alanine-scan mutations, electrostatic "hot-spots" could be overestimated in proteins with high net charge, which is most likely the case for homologues like the sea urchin (net charge -13).

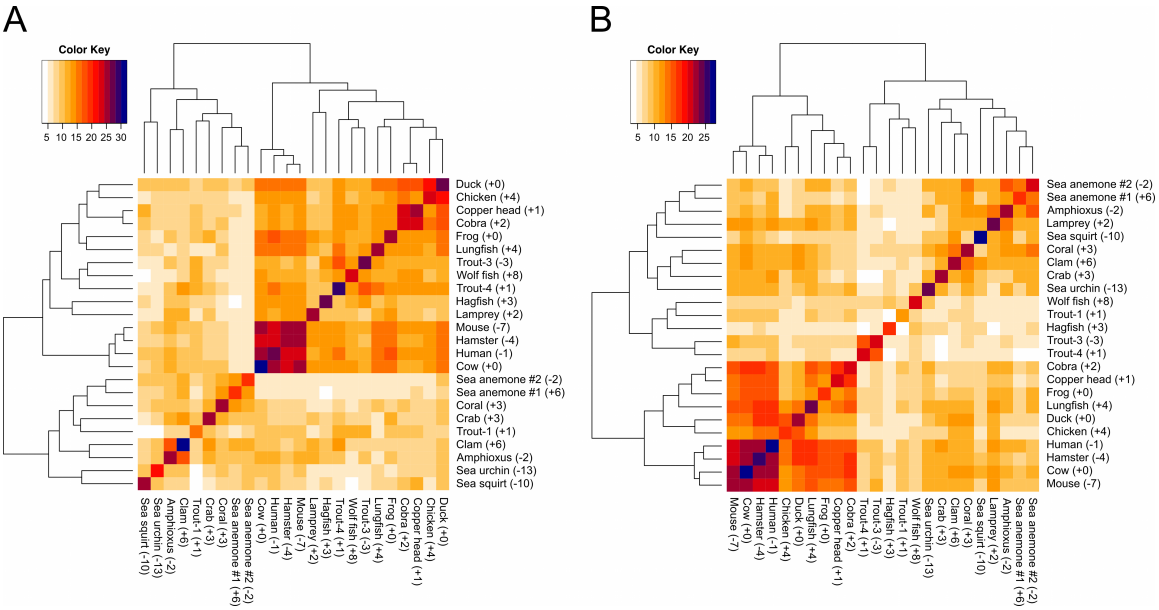


Figure 3-11 *Charge similarity clustering for the two sides of C3d.* Dendrograms with distance matrix heatmaps illustrate clustering of the 24 C3d homologues based on the number of positions with the same charge within the two functional regions: (A) CR2-face and (B) thioester-face. Net charge of each homologue is provided in parentheses.

Sequence-based approaches are typically used when analyzing the evolution of a protein, since conservation of amino acid positions can identify functionally important regions of sequence. As a comparison to our perturbation map approach, we have performed clustering for the 24 C3d homologues based on similarity of charged amino acid positions within the two

functional regions (Figure 3-11). The CR2-face charge clustering (Figure 3-11A) identified two primary clusters: (1) contains all species with cellular immunity [jawless fish, fish, amphibians, reptiles, birds, and mammals]; (2) all invertebrate species. The appearance of jawless fish (lamprey and hagfish) in cluster 1 of the CR2-face clustering, is quite interesting given that jawless fish could hypothetically be the first group of species to exhibit a C3d-CR2-like interaction, since they contain B-like and T-like cells [144]. This is in contrast to the current hypothesis that teleost fish are the first species to possess the C3d-CR2 interaction [129]. The charge patterns of the species of cluster 1 are noticeably more similar (darker region; Figure 3-11A), suggesting an emergence of a new role of charge in the function of C3d in the species of this cluster. Additionally, the mammals cluster separately from the other species of cluster 1 in the CR2-face charge clustering. Indicating that the CR2-face charge character of mammal C3d homologues is unique, which correlates with the perturbation map results (Figure 3-10A). In contrast, the thioester-face charge similarity clustering (Figure 3-11B) produced a similar classification of the C3d sequences as found using sequence percent identity (Appendix M and [145]). It should be noted that the net charge of the homologues has little to no effect on the charge similarity clustering (Figure 3-11), and it's the position of specific charged residues that distinguishes the clusters of C3d homologues, which is in agreement with the existence of electrostatic "hot-spots".

As has been discussed by McCammon [145], speed is often the main evolutionary driving force, even at the molecular level. Acceleration of biomolecular processes is achieved through long-range electrostatic interactions, which guide the formation of encounter complexes, increasing the diffusive rate [16]. Electrostatic "hot-spots", generated by clusters of like-charged residues, are frequent in nature, since they provide rapid association of

biomolecules satisfying the need for speed. However, there is a penalty paid in the form of a loss in local protein stability, since the presence of clusters of like-charged residues can result in numerous unfavorable intramolecular Coulombic interactions, as has been shown for the case of barnase-barstar [146]. Fersht and coworkers have proposed that nature often selects for function over stability [146], as is suggested by the presence of electrostatic "hot-spots" on many proteins. In the case of human C3d, electrostatic "hot-spots" have evolved surrounding the thioester bond and CR2 binding site, which when combined accelerate the formation of the B-cell co-receptor complex, a link between innate and adaptive immunity. The C3d-CR2 interaction greatly improves the immune response to an antigen, and as a result has been selected by nature as a target for immune evasion [96,139,140]. Structural evidence has shown that virulence factors of Staphylococcus aureus target the electrostatic "hot-spots" of human C3d (Appendix K). For example, Staphylococcus aureus secretes the highly cationic virulence factors EfbC and Ehp, which take advantage of the CR2 electrostatic "hot-spot" through the use of long-range, as well as short-range, electrostatic interactions (Appendix K-A) [139]. Additionally, domain IV of the Staphylococcal immunoglobulin-binding protein (Sbi) targets the thioester side electrostatic "hot-spot" of C3d (Appendix K-B), and in conjunction with Sbi domain III results in futile consumption of C3 through the formation of covalent adducts [140]. The electrostatic nature and binding sites of the Staphylococcus aureus virulence factors is further evidence for the key role of electrostatics in the function and evolution of complement C3d. When comparing the homologues of complement C3d, a weak CR2 "hot-spot" doesn't necessarily imply an absence of the C3d-CR2 interaction, but implies a less optimized interaction, specifically with respect to electrostatics. Therefore, based on our analysis we cannot conclude which homologues of C3d interact with a CR2 homologue, but we have

identified the onset of a new role of charge/electrostatics in the function of C3d after the divergence of jawless fish. We propose that this new role of charge corresponds with the appearance of the first multi-functional homologue of C3d. It should be noted that conserved electrostatic potential is not necessarily indicative of a conserved electrostatic "hot-spot", as can be seen when comparing the homologues of C3d (Figure 3-8 and Figure 3-10); however, our novel methods based on perturbation maps identified the electrostatic "hot-spots" of C3d and have potential utility in the identification of functional sites of other highly-charged biomolecular systems, as well as in drug design.

3.4. CONTROVERSY OF THE ASSOCIATION OF C3D-CR2

3.4.1. Introduction

The interaction between complement fragment C3d and B-cell receptor CR2 is an important step in the mounting of an immune response and helps form a link between the innate and adaptive immune systems [79]. As a result, the C3d-CR2 interaction has been a hot research topic for more than a decade. In the midst of these studies a controversy over the mode of interaction between C3d and CR2 arose. In 1998 the structure of free C3d was published, which revealed an acidic patch comprising of two hotspots of acidic amino acids [135]. In 2000, a mutagenesis study confirmed that mutations of acidic amino acids within the C3d patch decreased binding to CR2, whereas a mutation of a basic amino acid in the vicinity of the acidic patch increased binding to CR2 by two-fold [133]. It was then proposed that this acidic patch might be the site of interaction with CR2. However, in 2001 a structure of the C3d-CR2 complex was published, which showed that the binding interface was not at the acidic patch

[147]. In 2005 a mutagenesis study demonstrated that several mutants of basic amino acids in CR2 resulted in reduced binding to C3d, including one mutation (out of three possible) at the binding interface [132]. The results of that study were consistent with the binding interface of the crystal structure, and also showed that medium and long-range electrostatic effects are contributing to binding.

A more recent publication reported a mutagenesis study of C3d acidic amino acids, some of which were repeat mutations of the 2000 study but with a different assay (SPR versus Rosette) and some were new mutations [148]. Four new mutants were near the crystallographic binding interface and showed no effect on binding. The authors of that study questioned the physiologic relevance of the crystallographic structure. They proposed that the complex was forced to assume the observed binding interface by the crystallization conditions, which included the presence of zinc. Indeed, the crystallographic structure contains two non-physiologic zinc ions at the C3d-CR2 interface [147]. The same study [148] demonstrated using ELISA assays that C3d-CR2 binding in solution was abolished in the presence of zinc acetate. The possibility of a secondary or transient binding site has also been discussed, based on theoretical and experimental data [24,29,134,149,150]. We also came to similar conclusions regarding the physiological relevance of the co-crystal of C3d-CR2 following comparison of the latest mutagenesis data with new electrostatic calculations [38]. This controversy was brought to a closure upon the release of a new co-crystal structure of C3d-CR2 [151], which showed CR2 interacting with the acidic concave surface as was originally proposed [133].

In light of the new structural data for the C3d-CR2 interaction, the goal of this study is to reevaluate the role of electrostatic using computer simulation. In this study we apply our novel computational framework, AESOP, to elucidate the role of each charged residue through the use

of theoretical alanine-scan mutagenesis and electrostatic clustering/free-energy calculations. We discuss our results in view of all available experimental data and shed some light on the current state of the controversy over the mode of interaction between C3d and CR2.

3.4.2. Methods

Calculations of this study were performed using the latest co-crystal for C3d-CR2 (PDB Code: 3OED, [151]). Additionally, new calculations based on the previous C3d-CR2 co-crystal (PDB Code: 1GHQ, [147]) were also performed using the latest protocols for consistent comparison. All electrostatic calculations were performed using the AESOP framework, which has been discussed in great detail previously (section 2.3, [25,38]). In brief, all alanine-scan mutations were performed using amino acid truncations by AESOP functions. The PDB2PQR [42] utility was used to prepare all structures for electrostatic calculations by incorporating atomic radii and partial charges according to the PARSE forcefield [43]. Electrostatic potential and free energies of association, according to the thermodynamic cycle of Figure 2-1B, were calculated using APBS [3]. A grid with 129 × 129 × 129 grid points was used for both 3OED and 1GHQ, while coarse/fine grid lengths of 130 Å × 126 Å × 120 Å were used for 3OED and 168 Å × 128 Å × 140 Å for 1GHQ. The dielectric boundary was defined by a water-sized probe sphere with a radius of 1.4 Å, and the dielectric coefficients of the protein and solvent environments were 20 and 78.54 respectively. All electrostatic calculations were performed using 0 mM of counterions.

3.4.3. Results and Discussion

The theoretical alanine-scan mutagenesis for the 3OED structure produced 64 mutants of C3d and 23 mutants of CR2. The effects of the two sets of mutations (from C3d or CR2) were

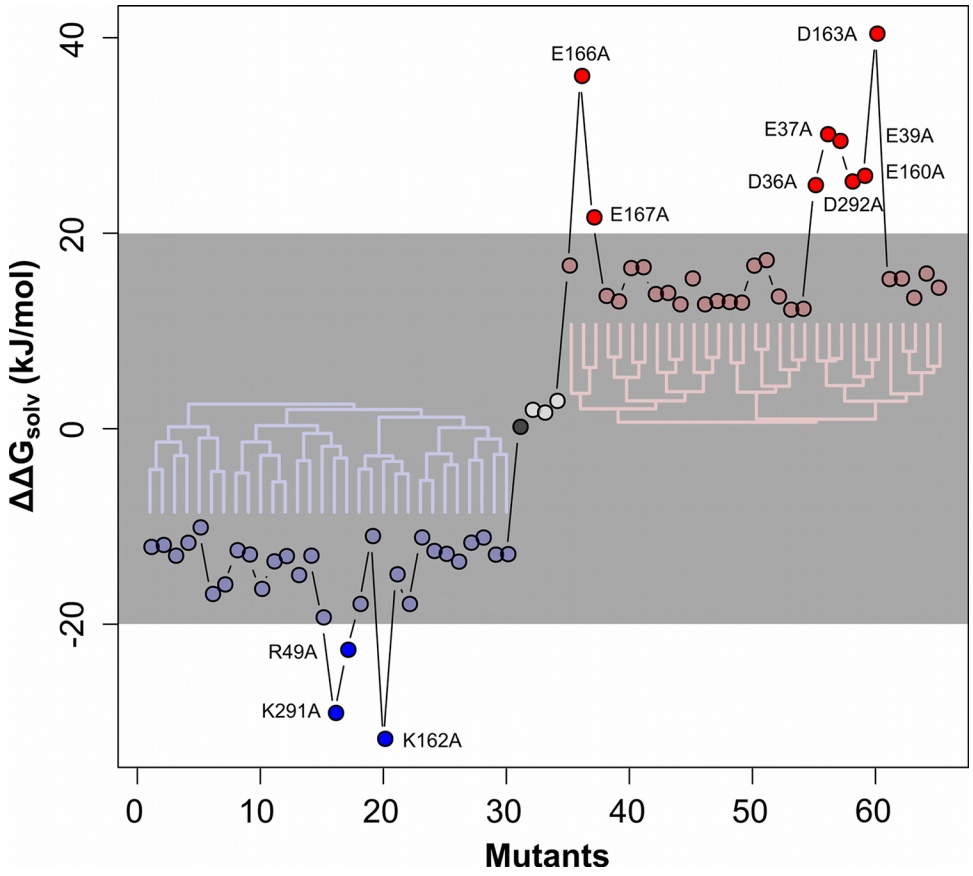


Figure 3-12 *Electrostatic free energies and clustering for the C3d alanine-scan.* Relative solvation free energies of association are plotted according to the electrostatic clustering, with the color of the circles indicating the type of amino acid mutated: red – acidic; blue – basic. The dark gray circle represents the parent structure, while the light gray circles indicate mutations of neutral histidines. The gray box illustrates the range of free energies containing mutations predicted to be benign (± 20 kJ/mol). Darker circles and labels indicate residues predicted to substantially affect binding. Clustering trees for the two primary clusters (acidic and basic) illustrate the electrostatic similarity within these clusters. Calculations were based on PDB 3OED.

quantified/compared using electrostatic free energies and clustering, as illustrated by Figure 3-12 and Figure 3-13. To focus the analysis on mutations of residues crucial to association, all mutations with free energies of association falling in the range of ± 20 kJ/mol (gray box of Figure 3-12 and Figure 3-13) are assumed to be to be benign. Figure 3-12 contains electrostatic free energies of association for each C3d mutant relative to the parent, as well as the electrostatic clustering for further classification. The electrostatic calculations identified two

clusters of acidic amino acids on C3d, which when mutated have notable effect on the electrostatic character of C3d that is predicted to be unfavorable for association. These two clusters of acidic amino acids correspond to the two groups of negatively charged amino acids previously identified on the acidic concave surface of C3d [133], and are composed of: (i) E166A and E167A; (ii) D36A, E37A, E39A, E160A, D163A, and D292A. Three basic residue mutations (R49A, K162A, and K291A) were also identified to have notable effect on the electrostatic character of C3d, but are predicted to be favorable for association, with R49A and K291A having similar effects since they cluster together. Of the 11 mutations predicted to affect association the most, all except R49A have been experimentally evaluated [133,148] and comparisons between the electrostatic free energies of association and their reported activities can be found Table 3-1.

Figure 3-13 contains the electrostatic free energies and clustering for the alanine-scan of CR2, which identifies five mutations of basic residues as being unfavorable for the association of C3d-CR2 (R13A, R28A, R36A, K41A, and K108A). Three of the five mutations (R13A, R36A, and K41A) have been experimentally evaluated [132] and their activities are provided in Table 3-1. No mutations with substantially favorable effects were identified for CR2, as indicated by the lack of bold data points below the gray box of Figure 3-13. In general, the mutations of CR2 are predicted to have less effect on association when compared to the mutations of C3d, as evidenced by a smaller range of relative association free energies (Figure 3-13).

Table 3-1 contains comparisons between experimental binding data for various alanine scan mutants and calculated solvation free energies of association based on the two available crystal structures (3OED and 1GHQ). Comparison of the two sets of solvation free energies of association (based on 3OED and 1GHQ) shows that on average the free energy values obtained

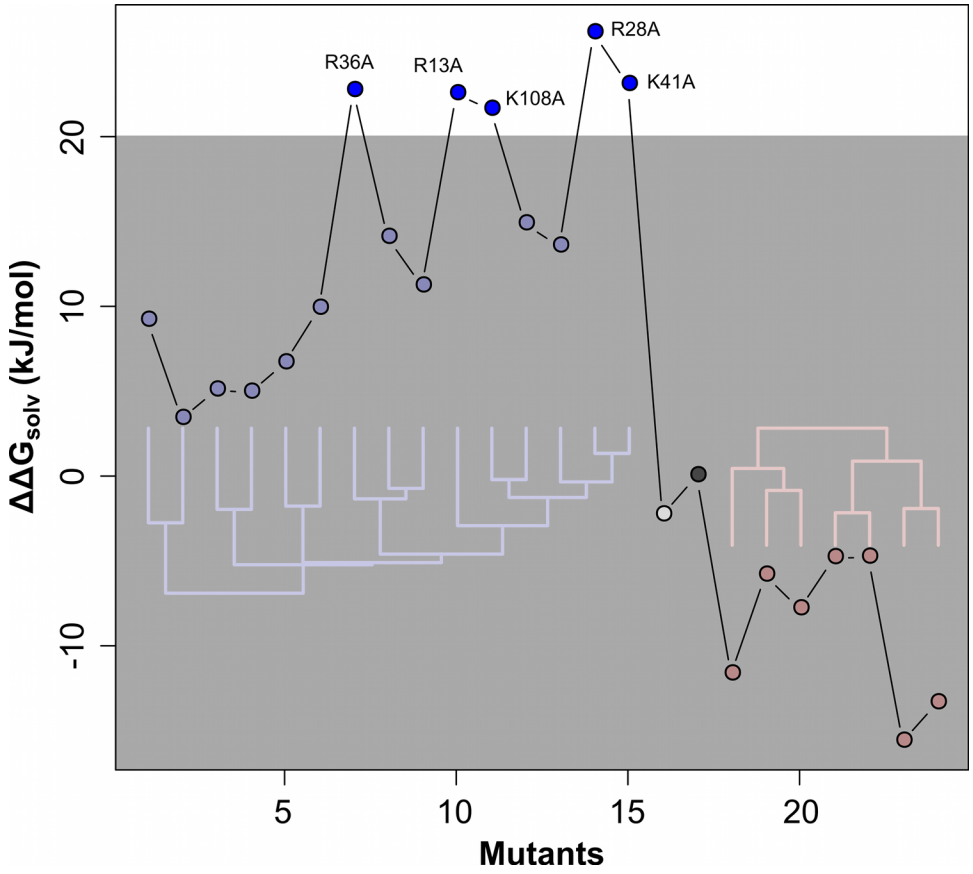


Figure 3-13 *Electrostatic free energies and clustering for the CR2 alanine-scan.* Relative solvation free energies of association are plotted according to the electrostatic clustering, with the color of the circles indicating the type of amino acid mutated: red – acidic; blue – basic. The dark gray circle represents the parent structure, while the light gray circles indicate mutations of neutral histidines. The gray box illustrates the range of free energies containing mutations predicted to be benign (± 20 kJ/mol). Darker circles and labels indicate residues predicted to substantially affect binding. Clustering trees for the two primary clusters (acidic and basic) illustrate the electrostatic similarity within these clusters. Calculations were based on PDB 3OED.

for the 30ED are noticeable more negative when compared to the 1GHQ based values. This suggests that the mode of interaction observed in the 3OED structure is more electrostatically favored in comparison to 1GHQ. Pearson correlation coefficients were used to quantify the agreement between the two sets of electrostatic free energies, based on the two crystal structures, and experimental binding data for mutations of C3d and CR2. As expected, the correlation between calculated electrostatic free energies and experimental data for mutations

Table 3-1 *Comparison of electrostatic free energies of association and experimental activities for alanine-scan mutants of C3d and CR2.*

C3d Mutants	$\Delta\Delta G_{solv}$		Relative $\Delta\Delta G_{solv}$		SPR[a,d]	Rosette[b,d]
	3OED	1GHQ	3OED	1GHQ		
Parent	-168.36	-43.83	0.00	0.00	4	4
D36A	-143.63	-35.21	24.73	8.62	2	4
E37A	-138.43	-34.67	29.93	9.16	1	2
E39A	-139.11	-36.05	29.25	7.77	1	2
E42A	-152.68	-37.09	15.68	6.74	3	4
E117A	-154.78	-18.40	13.58	25.43	4	
D122A	-154.97	-24.77	13.39	19.06	4	
D128A	-155.83	-24.93	12.53	18.90	4	
D147A	-155.53	-28.05	12.83	15.78	4	
E160A	-142.68	-32.32	25.68	11.50	0	1
K162A	-200.28	-56.51	-31.92	-12.68		5
D163A	-128.14	-31.76	40.22	12.07	0	0
E166A	-132.47	-30.20	35.89	13.63	0	0
E167A	-146.93	-28.36	21.43	15.47	3	3
K251A	-179.54	-52.55	-11.18	-8.72		2
K291A	-197.63	-52.63	-29.27	-8.80	3	4
D292A	-143.26	-35.11	25.10	8.72		3
SPR-COR	**-0.65**	**0.14**	**-0.65**	**0.14**		
Rosette-COR	**-0.68**	**-0.57**	**-0.68**	**-0.57**		

CR2 Mutants	$\Delta\Delta G_{solv}$		Relative $\Delta\Delta G_{solv}$		Rosette[c,d]
	3OED	1GHQ	3OED	1GHQ	
Parent	-168.36	-43.83	0.00	0.00	4
R13A	-145.86	-41.04	22.50	2.79	2
R36A	-145.67	-40.24	22.69	3.58	3
K41A	-145.31	-39.15	23.05	4.68	2
K50A	-164.99	-40.10	3.38	3.73	2
K57A	-153.52	-39.59	14.84	4.24	1
K67A	-154.32	-43.08	14.04	0.75	4
R83A	-154.83	-32.84	13.53	10.99	1
Rosette-COR	**-0.28**	**-0.78**	**-0.28**	**-0.78**	

[a]SPR data for C3d mutations were previously reported [148].
[b]Rosette data from C3d mutations were previously reported [133,148].
[c]Rosette data from CR2 mutations were previously reported [132].
[d]Experimental data are relative to the parent activity as follows: 0, 0% of parent activity; 1, 25% of parent activity; 2, 50% of parent activity; 3, 75% of parent activity; 4, 100% of parent activity; 5, 200% of parent activity.

of C3d were drastically improved for the 3OED structure. It would appear that there is a discrepancy in the correlations between the 1GHQ free energies and the two sets of experimental data, since the correlation with the SPR data is noticeably worse; however, this difference arises because the SPR dataset contains four mutations suggested to be at the binding interface according to the 1GHQ structure, which were found to have no effect on binding of C3d and CR2 (Table 3-1). This demonstrates the need for using a large number of mutations distributed evenly or in a representative way throughout the protein in order not to bias the correlation calculation. Additionally, the free energy correlations with the Rosette and SPR data for C3d mutations are comparable for 3OED structure.

Interestingly, the correlations between calculated electrostatic free energies and experimental binding data for CR2 mutations are substantially worse for 3OED when compared to 1GHQ. This phenomenon may have arisen due to sampling, since researchers have only performed mutations of CR2 residues near the binding interface identified by the 1GHQ structure. Correlations could potentially improve after collection of additional experimental data for new mutations, such as those identified in Figure 3-13. Lack of correlation for CR2 mutations with the 3OED-based calculations could also be further evidence for the existence of a secondary transient binding site, as has been previously suggested [24,29,134,149,150].

3.4.4. Conclusion

In this study, we employed our computational framework AESOP [25,38] to further analyze the controversy over the C3d-CR2 mode interaction in light of new structural data, which has recently been reported [151]. We utilize alanine-scan mutagenesis, as well as electrostatic clustering and free energy calculations, to evaluate the role of each charged

residue in association of C3d and CR2, and compare our results with available experimental data [130,132,133,148]. Surprisingly, only correlations between free energies and experimental data for mutations of C3d were improved in light on the new 3OED crystal structure, since better correlation was observed with calculations based on the 1GHQ structure for the mutations of CR2. This result further suggests the possibility of a transient binding site for CR2, and leaves the controversy over the mode of interaction between C3d and CR2 not fully resolved.

3.5. PERSPECTIVE

The complement system is an ancient line of defense, consisting of a complex cascade of protein-protein interactions responsible for attacking foreign pathogens, and when healthy, leaving host tissues unharmed [77,78]. As is the case for many biological functions, speed is of the essence for biomolecular interactions involved in immune defense, especially those of the complement system. Through evolution complement proteins have gained high charge content, which accelerates the recognition phase of association [16,145]. Electrostatic forces are essential to the opsonization of foreign pathogens (C3b/C3d) [activation], as well as factor H recognition of opsonized host cells [regulation]. Therefore, electrostatics plays a crucial role in the constant balance waged between the efficient clearance of foreign pathogens and harmful over activation. Transient complexes, such as complement convertases, are often formed though long-range non-specific electrostatic interactions, which provide the fastest response time, but are weak enough to prevent prolonged over-activation.

Complement proteins have evolved to include electrostatic “hot-spots”, which are regions containing large electrostatic potential generated by clusters of like charged residues. These electrostatic “hot-spots” accelerate the recognition and orientation phases of association,

and as a result are often indicative of functional sites. Interestingly, due to the complement system's constant balance between activation and regulation, electrostatic "hot-spots" of complement proteins are often the binding site for numerous proteins, as is illustrated for C3d in Appendix J. However, pathogens have evolved to utilize these electrostatic "hot-spots" in the evasion of the immune system, by producing virulence factors that are highly charged to compete with host ligands, as is illustrated by Appendix K.

Due to this central role of electrostatics, electrostatics-based design has great utility in developing therapeutics for complements system related disease. One key project that would benefit from our recent work is the design of C3d analogs as vaccine adjuvants. The design of C3d analogs could take advantage of the results of our recent studies investigating the C3d electrostatic "hot-spots", as well as the role of electrostatic in the C3d-CR2 interaction. A recent study has shown that mutations known to affect C3d-CR2 binding, also affect the effectiveness of C3d as an adjuvant [80]. Suggesting that electrostatically designed C3d analogs, which possess improved CR2 binding, could be utilized in the development of more efficacious vaccine adjuvants. As our recent study into the evolution of C3d electrostatics showed, human C3d has evolved to have a dual electrostatic character, optimized for binding CR2 and pathogenic cell surfaces, both crucial in the formation B-cell receptor/co-receptor complexes. However, since C3d is covalently attached to the antigen in designed vaccines, only binding to CR2 is necessary for adjuvant activity. Therefore, mutations of charged amino acids away from the acidic concave surface of C3d (CR2 binding site) can be utilized to design new single-function C3d analogs, with improved CR2 recognition and binding. An approach similar to that utilized in the design of new SUMO4 analogs, which was performed using the AESOP framework [25,38], should be employed for the design of new C3d analogs.

4. DESIGN OF COMPLEMENT SYSTEM INHIBITORS TARGETING C5AR

4.1. COMPLEMENT ANAPHYLATOXINS

Of the anaphylatoxins, C3a is found in the highest concentrations, since it is the cleavage product resulting from the activation of C3, the central complement protein that is involved in all three activation pathways. On the other hand, C5a is the most potent of the anaphylatoxins despite being produced in lower concentrations. The anaphylatoxins regulate the inflammatory response through interactions with their membrane-bound receptors, C3aR and C5aR, which result in the chemotaxis and activation of leukocytes [152,153]. C3aR and C5aR are members of the GPCR superfamily, all of which are comprised of seven transmembrane helices connected by intra- and extra-cellular loops. The 7-helical bundle structural motif and the existence of highly disordered loops that are found in all GPCR structures, allows for large amounts of conformational freedom. Since C5aR has a central role in complement-mediated inflammation, and GPCRs are popular drug targets, C5aR has been a key target for therapeutic design [154-158].

Due to the importance of C5aR as a drug target, there have been some studies that have modeled C5aR bound to ligands (C5a or peptidic analogs) with the intention of understanding experimental results. Nikiforovich et al. generated a C5aR:C5a complex using numerous molecular modeling tools including “enhanced homology modeling” and geometrical sampling [159]. This study focused on possible modes of interaction between C5aR and C5a, and used published mutagenesis results for validation. Another study, by Higginbottom et al., utilized homology modeling and docking methods to investigate important interactions between C5aR and analogs of peptidic antagonist cC5aR-pa (also known as PMX-53 as referred to below) [160].

Both the Nikiforovich and Higginbottom studies provided insight into C5aR interactions, but both studies can be improved upon, especially since neither of these studies included molecular dynamic simulations to optimize local and global conformations. Following the Higginbottom study, a new, more refined, solution structure of the cC5aR-pa was determined by Zhang et al. and deposited in the Biological Magnetic Resonance Bank [161]. Figure 4-1 contains molecular graphics of the Zhang et al. structure for cC5aR-pa. The new structure included more accurate force field parameters for the non-natural amino acids, and was performed at lower temperature, as allowed by the solvent. Therefore, it would be beneficial to perform similar docking studies with the new structure determined by Zhang et al. Also, inclusion of an explicit membrane during the modeling process would provide a more realistic environment, and therefore a better analysis of protein-ligand interactions.

4.2. CONFORMATIONALLY SAMPLED PHARMACOPHORE MODEL OF C5AR ANTAGONISTS

4.2.1. Introduction

Complement receptor C5aR (also termed CD88) is a member of the GPCR superfamily, and upon activation by complement anaphylatoxin C5a initiates a pro-inflammatory response [155]. Over-activation of C5aR has been implicated in several inflammatory/autoimmune diseases, such as lupus nephritis and sepsis [155], and therefore C5aR has been the focus of numerous drug design efforts [154-156,158,162]. Several groups reported initial success in the development of short linear peptide C5aR agonist in the early 1990's, but the first C5aR full antagonist, Me-FKPdChaWr, was later developed by researchers at Merck (as reviewed in [155]).

Structural studies using NMR spectroscopy observed that a salt-bridge between Lys 2 and the carboxylate of Me-FKPdChaWr stabilizes a cyclic-like structure [156]. This observation led to the development of a new family of cyclic hexapeptides, including the highly active analog Ac-Phe-[Orn-Pro-dCha-Trp-Arg] marketed as PMX-53 (Figure 4-1); Ac denotes the acetylated N-terminal end, Orn ornithine, dCha d-cyclohexaalanine and the square brackets indicate cyclization (analog 1, Table 4-1) [154,156,162].

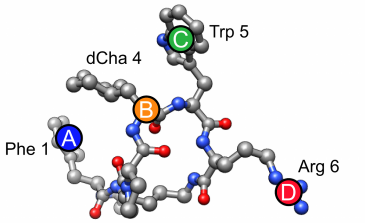


Figure 4-1 *Molecular graphic of PMX-53 illustrating the four pharmacophore points.* (A) center of benzene ring of Phe 1, blue; (B) C_α backbone atom of dCha 4, orange; (C) center of the indole ring of Trp 5, green; (D) C_ζ side chain atom of Arg 6, red.

Over 60 PMX-53 analogs have been experimentally evaluated [154,156,162] providing valuable structure-activity data, however only limited modeling has been performed to analyze the observed relationships. Since both linear and cyclized C5aR antagonists exhibit a well-defined structure, as observed by NMR studies [154,156,161,162], we propose that the dynamics of the C5aR antagonists free in solution contribute significantly to function and can be used to distinguish among various levels of affinity. Therefore, the aim of the current study is to develop for the first time a conformationally sampled ligand-based pharmacophore model for hexapeptide C5aR antagonists of the PMX-53 family, in order to analyze the relationship between structure, dynamics, and function.

Conformationally sampled ligand-based pharmacophore models utilize MD simulations to generate representative conformations for a set of related ligands in their unbound state and

extract properties, such as distances, angles, and torsion angles, describing the relative position of selected pharmacophore points [163-167]. Conformationally sampled pharmacophore models seek to identify pharmacophore descriptors whose statistical distributions describe conformations of ligands either unique to, or more frequently observed in, the most active ligands. Such models have been previously developed for ligands of δ-[163-165] and μ-opioid [166] receptors, also GPCRs, as well as complement system inhibitors of the compstatin-family [167]. In the current study, we extend previous methods by incorporating the anisotropic membrane/water environment in which the C5aR resides via a high quality implicit-membrane model [168]. We also include clustering methods to efficiently identify quality pharmacophore descriptors.

4.2.2. Methods

For this study we have selected 10 hexapeptide C5aR antagonists with IC_{50} values ranging from 0.25 – 1082 μM to compose our ligand set, and their sequences and reported experimental activities are provided in Table 4-1. Available structure-activity data has shown that Phe at position 1, an aromatic residue at position 5, and Arg at position 6 are all crucial for antagonist affinity [154]. Therefore, these three constraints were employed while selecting the analogs, and as a result the majority of substitutions affect position 4. The geometric centers of these key residues at positions 1, 5, and 6 were selected as pharmacophore points. Comparison of the sequences of the ligand set shows that positions 4 and 5 share some codependence with regard to affinity, suggesting that a potential interaction between positions 4 and 5 could have distinguishable effects on dynamics and function. Based on this observation, the backbone C_α of position 4 was also selected as a pharmacophore point since there is little consistency in amino

acids at this position. The definition of the four selected pharmacophore points is illustrated in Figure 4-1. The analogs have been grouped into three groups according to their experimental affinity as follows (Table 4-1): two highest affinity analogs (group I), two lowest affinity analogs (group III), and six analogs of varied affinity (group II). Of the 10 training analogs, only analog 5 is linear (Table 4-1). Analog 5 resembles very closely analog 3, the only difference being that analog 3 is cyclized. Also, analogs 3 and 5 are the only analogs with dArg at position 6.

Table 4-1 *List of C5aR hexapeptide antagonists with sequences and activities.*

#	Group	Sequence [a,b]	Reported receptor affinity (μM)	Reported antagonist activity (μM)	Ref. #
1	I	Ac-F-[OPdCha**F**R]	0.25	0.032	[154]
2		Ac-F-[OPdChaWR]	0.38	0.026	[154]
3	II	Ac-F-[OPdChaW**r**]	0.28/ 16.0	0.012/ 0.400	[156]/[162]
4		Ac-F-[OP**f**WR]	0.46	0.022	[154]
5		Ac-FOPdChaW**r**	0.80	0.200	[162]
6		Ac-F-[OP**I**WR]	1.13	N.D.	[154]
7		Ac-F-[OP**f****F**R]	5.20	5.210	[154]
8		Ac-F-[OP**w**WR]	32.0	N.D.	[154]
9	III	Ac-F-[OP**a**WR]	145	3.700	[154]
10		Ac-F-[OP**g**WR]	1082	N.D.	[154]

[a]Bold letters indicate amino acid postions that differ from the parent peptide (analog 1).
[b]Lowercase letters indicate D amino acids.

To sample the conformations of the training analogs (Table 4-1), we carried out implicit-solvent MD simulations starting from the lowest energy NMR structure of the parent (analog 2) hexapeptide [161]. All necessary mutations were introduced by the UCSF Chimera software [106]. MD simulations were carried out with the CHARMM program [169]. Standard amino acids were described by the CHARMM27 all-atom energy function including the CMAP correction [170,171]. The topology and parameters for non-standard groups (ornithine, cyclohexylalanine and the Orn2-Arg6 cyclization segment) were derived from CHARMM27 definitions for lysine, alanine/cyclohexane and the peptide group, respectively. The membrane environment was

modeled by the heterogeneous dielectric generalized Born variant (HDGB) of the Generalized Born Molecular Volume (GBMV) model [168], with default parameters (solvent dielectric ε=80, surface tension coefficient σ=0.015 kcal/Å^2, β=-12, S0=0.65, C0=3255, C1=1.085, D=-0.14). An 18-Å cutoff distance was used for non-bonded interactions. The lengths of covalent bonds containing hydrogen atoms were constrained by the SHAKE algorithm[172], and the equations of motion were solved with an integration time step of 1.5 fs [173]. The peptide was in contact with a Langevin heat bath at 300 K; a friction coefficient of 10 ps^{-1} was used for heavy atoms.

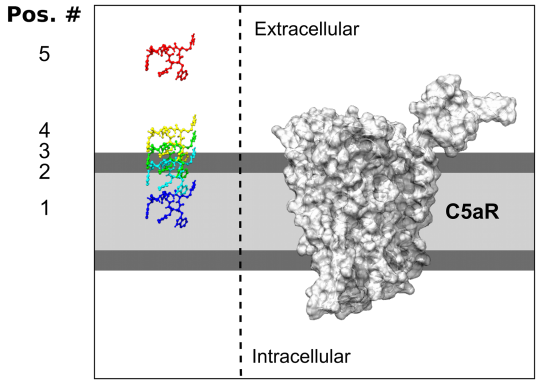


Figure 4-2 *Illustration of the five membrane positions used in the molecular dynamics simulations.* The membrane positions are as follows: (1) membrane center, blue; (2) 15 Å above center, cyan; (3) 20 Å above center, green; (4) 25 Å above center, yellow; (5) pure water, red. Dark gray bars depict polar regions of the membrane bilayer, while the light gray bar represents the hydrophobic membrane core. A molecular graphic of C5aR[159] is included for reference, but was not included in the simulations.

For all analogs we conducted five sets of simulations with varied environments. In the first four, the analogs were immersed in an (implicit) membrane and were restrained at distances of 0.0 Å, 15.0 Å, 20.0 Å and 25.0 Å from the membrane center, as illustrated by Figure 4-2; in the last they were immersed in pure water, represented by the GBMV II model [174]. Prior to each production run, the energy was minimized by 500 steps of steepest descent; subsequently, two 100-ps equilibration runs were performed, with backbone atoms harmonically restrained by 2 kcal/mol/Å^2 and 1 kcal/mol/Å^2 force constants. In the membrane

runs, the peptide center of mass was restrained to the desired position across the membrane by a weak, 0.5 kcal/mol/Å^2 force constant. The duration of each production run was 15 ns, with only the last 13.5 ns used in the analysis.

In order to generate the conformationally sampled pharmacophore descriptors we extracted 1,800 snapshots from each trajectory at 7.5-ps intervals. The four pharmacophore points (Figure 4-1) define 24 geometrical properties (6 distances, 12 angles, and 6 torsion angles), resulting in 276 pairs of properties. For all 276 pairs we generated two-dimensional histograms, with dimensions of 50×50 bins, using the 50 sets of available trajectories (10 analogs, simulated at 4 membrane positions and in pure water). To quantify the similarity between two analogs we compared the statistical distributions of their property pairs using a measure similar to that proposed for clustering of protein electrostatic potentials [25]. The similarity measure is defined as follows:

$$LD = \frac{1}{N}\sum_{i,j}\frac{\left|A(i,j) - B(i,j)\right|}{\max\left(A(i,j), B(i,j)\right)} \qquad \text{Eq. 4-1,}$$

where *A* and *B* are two-dimensional (50×50) histograms of a given property pair, generated by simulations of the peptides A and B in the same environment, *i,j* are the indices of the two-dimensional histogram, and N is the total number of histogram bins. The histograms will be referred to as pharmacophore descriptors from this point further. The 276 possible measurement pairs and the 5 environments (4 membrane positions and pure water) generate 1,380 pharmacophore descriptors. For each descriptor, we generated a 10×10 distance matrix, containing pairwise similarities for all peptide pairs in the 10-analog ligand set. Peptides were quantitatively classified using hierarchical clustering based on each of the generated distance matrices, and the classifications were visualized using dendrograms.

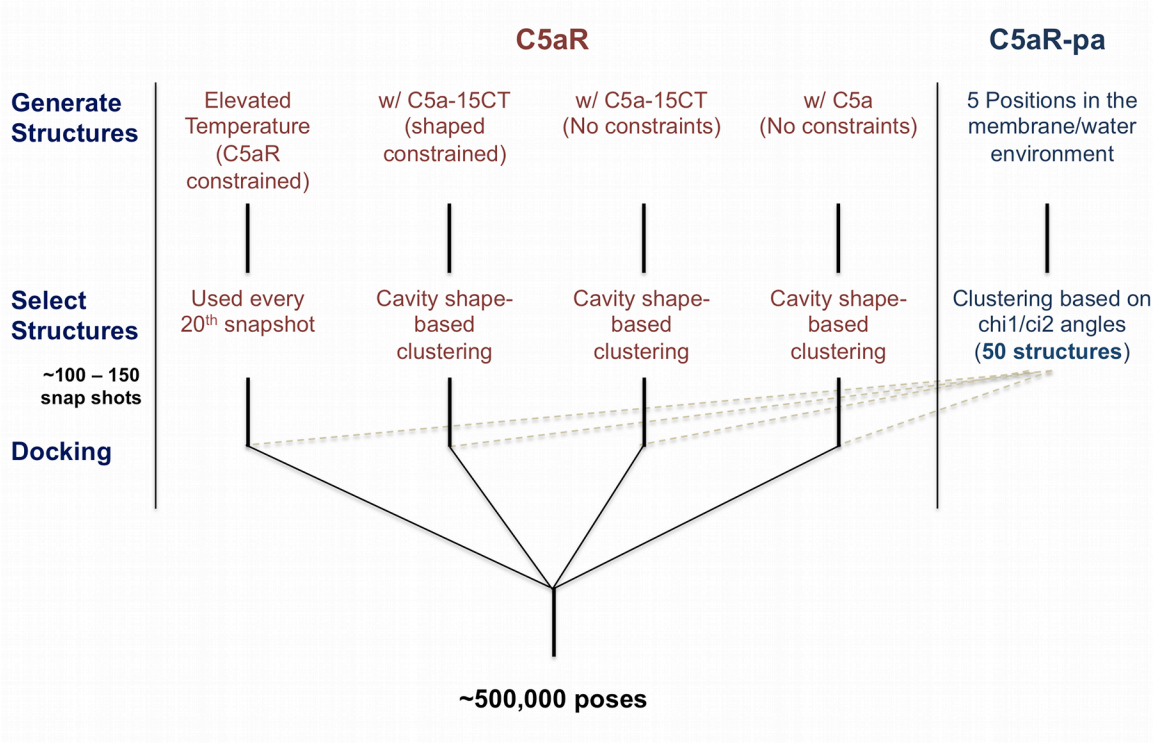


Figure 4-3 *Schematic of the molecular dynamics and docking procedure for PMX-53:C5aR model development.* Red text indicates procedures used to generate/select structures of C5aR, while light blue text corresponds PMX-53 (C5aR-pa). Dashed lines indicate the combination of the C5aR and PMX-53 structures using rigid-ligand docking.

For further analysis of the mode of interaction C5aR antagonists, a model of the PMX-53:C5aR interaction has also been developed in parallel. A novel procedure utilizing implicit membrane molecular dynamics and rigid ligand docking has been developed (as illustrated by Figure 4-3) were performed with hopes of extensively sampling the conformational space and generating a diverse set of binding sites for docking studies. PMX-53 conformations were also generated using the procedure discussed above. Based on the molecular dynamics simulations at the 5 membrane positions (Figure 4-2), 50 representative

PMX-53 structures were selected based on torsion angle clustering. (All molecular dynamics simulations for C5aR antagonists were performed by Dr. Phanourios Tamamis.)

Representative C5aR structures were selected from the MD simulations of the three C5aR complexes using a novel procedure based on receptor cavity shape. The cavity shapes were characterized by identifying water molecules that fill the void resulting from the removal of the C5aR ligand, and these calculations were performed with scripts written in the R statistical language [48] using the Bio3D library [50]. The underlying premise is that by ensuring that the water molecules are always in the same position relative to C5aR, it is possible to identify the regions of the C5aR cavity that are changing structurally by simply keeping track of which water molecules fit in the cavity. Implying that C5aR cavities containing similar water molecules have a similar shape. Cavity shape similarity was determined by first superimposing each MD snapshot onto the initial structure, which was centered in a 50 × 50 × 60 explicit water box, based on the C_{α}-atoms of the TM regions. Water molecules were selected by excluding all molecules within 2.4 Å of any C5aR atom, and by including only those molecules within 4 Å of the 15 C-terminal residues of C5a (a.a 60 -74). Lists identifying the water molecules that fit into each MD snapshot were created, and used to generate binary water fingerprints, where a value of 1 represents water molecules that were present in the cavity and a value of 0 indicates those that were absent. Distance matrices based on a binary similarity measure were generated to quantify the similarity between each pair of water fingerprints, which is indicative of the similarity between each pair of cavity structures. Hierarchical clustering was used to classify the MD snapshots into families based on water fingerprint similarity, and the clustering was visualized using dendrogram trees. The dendrogram trees were cut at a binary similarity value of 0.3 to extract clusters of MD snapshots containing structures that differ by fewer than 50 water

molecules. From each cluster, the cavity structure with the largest volume (most cavity water molecules) was selected as the representative structure. However, structures were extracted every 20 snapshots from the ligand-free elevated temperature trajectory. Approximately 150 structures were selected from each MD simulation to be used as input for the docking studies.

Rigid-ligand docking was performed based on the representative MD snapshots of C5aR and PMX-53 using the DOCK6 [4] docking program. Molecular surfaces were generated for each C5aR structure using DMS [175]. All possible spheres representing the potential binding sites of each C5aR structure were generated using the SPHGEN utility; however, only spheres within 7 Å of the 15 C-terminal residues of C5a were chosen to represent the putative PMX-53 binding site. A bump filter was employed to exclude docking poses with more than 5 receptor-ligand clashes, with a clash being defined as an atom-atom overlap of more than 50 percent. All orientations that passed the bump filter were saved as input for the CHARMM-based scoring procedure, since the docking program was not used to rank the C5aR:PMX-53 poses, due to the complexities of GPCR docking. The 51 selected conformations of PMX-53 were docked to the representative structures from each of the four C5aR MD simulations, resulting in approximately 500,000 docked poses. Following docking, an implicit membrane model in the CHARMM software was used to relax each pose using energy minimization, and subsequently score each pose using an association free energy. The top 40 poses, according to association free energy, were subjected to further optimization using implicit membrane molecular dynamics simulations to identify the lowest energy pose.

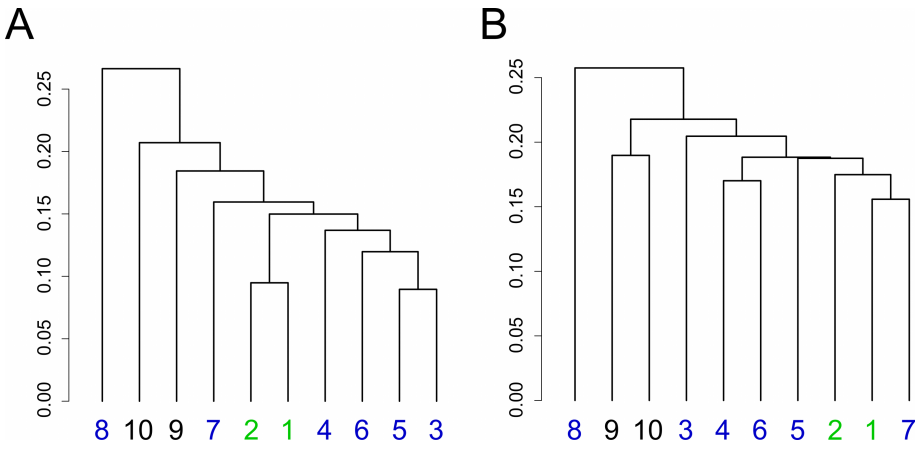


Figure 4-4 *Classification of C5aR antagonists.* Dendrograms illustrate the clustering of the 10 C5aR antagonists based on two pharmacophore descriptors at membrane position 1 (Figure 4-2): (A) ACD-CDB and (B) AD-CB. The coloring code is as follows: group I analogs are in green, group II analogs are in blue, and group III analogs are in black.

4.2.3. Results and Discussion

Based on the four chosen pharmacophore points, and the two-dimensional histogram analysis, there are 1,380 potential pharmacophore descriptors, which were visualized using dendrogram clustering trees. All 1,380 dendrograms were analyzed to identify pharmacophore descriptors able to distinguish between the three levels of affinity as defined by the three groups of analogs (Table 4-1). Of the 1,380 pharmacophore descriptors, only two were able to properly classify all training analogs, including the linear analog of group II (analog 5). The first identified pharmacophore descriptor is the pair of angles, ACD and CDB (ACD-CDB; Figure 4-4A), based on simulations performed at the membrane center (Figure 4-2; position 1). The second descriptor on the other hand is the pair of distances AD and CB (AD-CB; Figure 4-4B), also at the membrane center (position 1).

The ACD-CDB dendrogram (Figure 4-4A) shows a cluster containing analogs 1 through 7 (groups I and II), while analogs 8 through 10 compose their own clusters. Upon closer examination it also can be observed that analogs 1 and 2 (group I) form a subcluster indicative

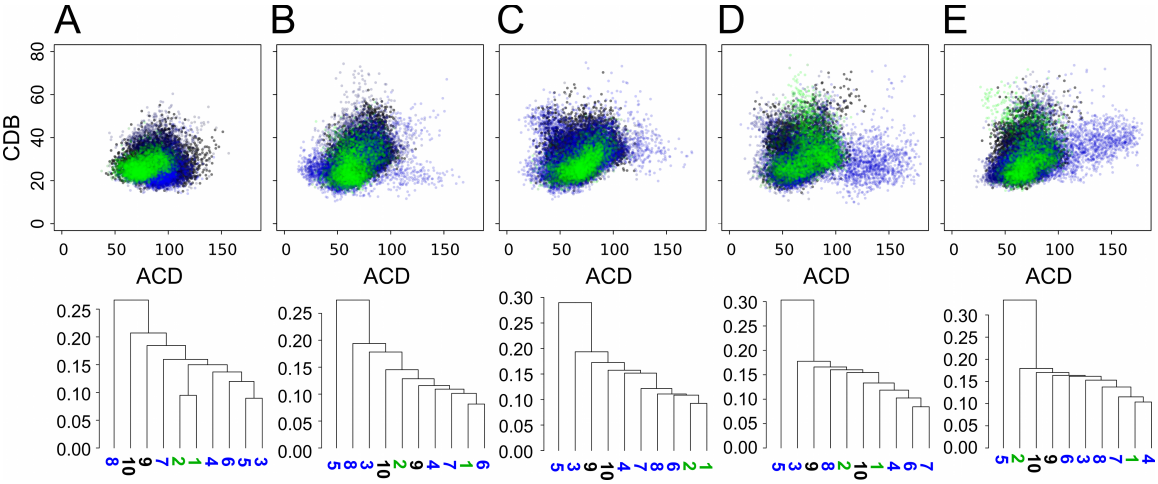


Figure 4-5 *Effects of the membrane environment of C5aR antagonists on ACD-CDB clustering*. Superimposed scatter plots and corresponding dendrogram of the ten analogs are presented for the five membrane positions: (A) position 1, (B) position 2, (C) position 3, (D) position 4, and (E) position 5. The coloring code is as follows: group I analogs are in green, group II analogs are in blue, and group III analogs are in black.

of higher similarity, which is mirrored in their comparable activities (Table 4-1). Cyclized analog 3 clusters closely to analog 5, which has identical sequence except that it is linear. We note that there is some controversy in the affinity of analog 3: One study reports an affinity of 0.28 µM [156], near the affinity (0.80 µM) of analog 5; another study reports a substantially lower affinity (16 µM) [162]. The AD-CB dendrogram (Figure 4-4B) also shows a cluster containing analogs 1 through 7 (groups I and II), however, the subclusters do not correspond as well with the analog activities as seen in the ACD-CDB clustering (Figure 4-4A). ACD-CDB and AD-CB describe very similar relationships between the pharmacophore points; however, ACD-CDB provides better distinction between the three levels of affinity represented by the analogs, and will be the primary focus of the remaining discussion. The scatter plot corresponding to the dendrogram for ACD-CDB at the membrane center is provided in Figure 4-5A. Inspection of the Figure 4-5A scatter plot shows that the ACD-CDB distribution is noticeably tight for high-affinity analogs (group I, green) and becomes more disperse for low-affinity analogs (group III, black).

Presumably, a tight distribution indicates a smaller loss in the conformational entropy of the corresponding analog upon binding, in line with its higher affinity. A larger ACD-CDB area could still be associated with high binding affinity, due to solvent-entropy contributions and/or a more enthalpically favored bound state.

It should also be noted that the ACD-CDB distributions, and the clustering results for that matter, are highly dependent on the peptide environment, or membrane position used when simulating the analogs, as shown by the five panels of Figure 4-5. When moving the analogs out of the membrane, from position 1 (membrane center) towards position 5 (pure water), the dielectric environment becomes increasingly polar, having dramatic effects on the dynamics of the peptide analogs. The effects of the membrane environment are most prominent in the case of the linear antagonist (analog 5); despite its theoretically higher flexibility, this analog contains a salt bridge between the basic side chain of Orn 2 and the carboxylate of Arg 6. The Orn 2/Arg 6 salt bridge stabilizes analog 5 in a conformation similar to that of the cyclic analog 3, and is most stable in the low dielectric environment of the membrane center (position 1). At other membrane positions, and in pure water**,** the distributions of the pharmacophore points of analog 5 are quite irregular with respect to all pharmacophore descriptors causing analog 5 to typically cluster separately from all other analogs. Thus, the membrane center is an optimum description of the surrounding environment for this ligand family, for the purpose of ligand-based pharmacophore modeling. This optimum location does not need to coincide with the exact binding site in the C5aR complex, which is predicted between the membrane center and surface by experimental and modeling studies [155].

Table 4-2 *Extended list of C5aR hexapeptide antagonists with sequences and activities.*

#	Group	Sequence [a,b]	Reported receptor affinity (μM)	Reported antagonist activity (μM)	Ref. #
11	IV	hydrocinnamate-[OPdChaWR]	0.22	0.031	[154]
12		Ac-F-[O**F**dChaWR]	2.43	N.D.	[154]
13		Ac-F-[OPdCha**w**R]	30.4	N.D.	[154]
14		Ac-F-[OP**f****Y**R]	69.2	N.D.	[154]

[a]Bold letters indicate amino acid postions that differ from the parent peptide (analog 1).
[b]Lowercase letters indicate D amino acids.

Further assessment of the ACD-CDB pharmacophore descriptor was achieved by studying an extended set of ligands, including four additional peptide analogs with varying affinity (Table 4-2). The additional analogs have substitution types/positions not observed in the initial ligand set, and therefore will test the overall predictive value of the ACD-CDB pharmacophore descriptor. The procedure followed for the initial ligand set was applied to generate MD snapshots for each additional analog at membrane position 1, and to produce ACD-CDB pharmacophore descriptors. Clustering comparing all 14 analogs based on the ACD-CDB pharmacophore descriptor is illustrated as a dendrogram in Figure 4-6. The Figure 4-6 dendrogram shows a primary cluster containing all higher affinity analogs with IC_{50} values of less than 3 μM, including analogs 11 and 12, which cluster with the group I analogs. Of the group IV analogs (orange, Figure 4-6), analog 14 appears to be an anomaly, since it clusters with analogs 3 and 5 despite having much lower affinity. It is likely that the lower binding affinity of analog 14 arises from factors not captured by the ACD-CDB pharmacophore descriptor, such as the transfer of the highly polar Tyr5 hydroxyl group to the complex interior, or unfavorable interactions of the same group with C5aR.

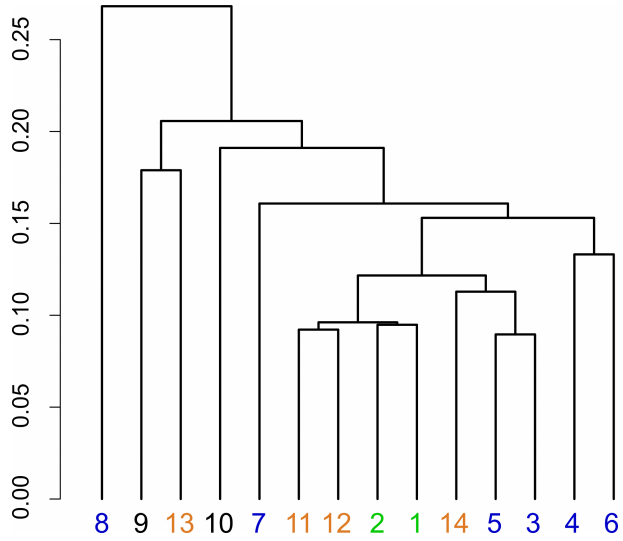


Figure 4-6 *Clustering of the extended set of C5aR antagonists based on the ACD-CDB pharmacophore descriptor at membrane position 1.* The coloring code is as follows: group I analogs are in green, group II analogs are in blue, group III analogs are in black, and group IV are in orange.

It should be noted that the probability distributions of useful free-ligand pharmacophore descriptors (such as the ACD-CDB combination) should not always be interpreted as representative of the actual ligand geometry in the complex. As a GCPR, C5aR exhibits high conformational freedom as it changes conformations to accommodate various sizes of ligands [155]. Also, the peptide analogs of the PMX-53 family are believed to bind C5aR in a buried cavity in the transmembrane region, which would likely induce large changes in the conformation and dynamics of the peptide. Comparison between the pharmacophore descriptors of free analogs and bound PMX-53 (analog 2) was performed in order to analyze the effects of binding to C5aR on PMX-53, and to ascertain the potential reasons of why the distributions of ACD-CDB and AD-CB are important for binding.

A docked complex of PMX-53:C5aR has recently been generated based on a novel docking procedure that incorporates rigid-ligand docking and CHARMM implicit-membrane MD simulations [in preparation]. A 20 ns implicit membrane molecular dynamics simulation, similar

to that described for the free peptide analogs, was carried out for the docked PMX-53:C5aR complex. The angles ACD and CDB, as well as distances AD and CB, were calculated for each of the bound snapshots, and scatter plots comparing the free and bound structures are shown in Figure 4-7. Interestingly, even though the most active analogs share a similar distribution of ACD-CDB in their unbound state, an induced-fit effect stabilizes a different PMX-53 conformation in the complex (Figure 4-7A, red points). On the other hand, the AD-CB distribution of bound PMX-53 (Figure 4-7B, red points) falls inside the distributions corresponding to the most active analogs (Figure 4-7B, green points). Thus, from the point of view of AD-CB, a conformational selection mechanism contributes to the binding of the peptide antagonists and is indicative of affinity.

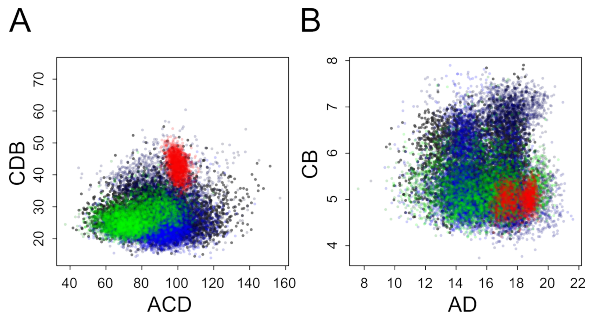


Figure 4-7 *Comparison of free group I – III analogs with PMX-53 bound to C5aR.* Superimposed scatter plots for the ten training analogs and a docked pose of PMX-53 (analog 2) are presented based on two measurement combinations: (A) angles ACD and CDB; (B) distances AD and CB. The coloring code is as follows: group I analogs are in green, group II analogs are in blue, group III analogs are in black, and bound analog 1 is in red.

In this study, we have developed a conformationally sampled pharmacophore model for C5aR peptide antagonists of the PMX-53 family. We have identified two pharmacophore descriptors, ACD-CDB and AD-CB, as indicative of peptide affinity. The proposed conformationally sampled pharmacophore model was evaluated using four peptide analogs not included in the initial ligand set, and showed predictive ability useful in the design of novel C5aR

antagonists. The novel methodological aspects of the conformationally sampled pharmacophore approach introduced here, specifically the inclusion of the C5aR environment by a high-accuracy implicit-membrane model, and the use of clustering analyses, proved to be critical to the success of this study. Additionally, these novel methods could have great utility in the study/design of ligands targeting GPCRs and other membrane receptors, as well as ligands for other protein targets.

4.3. VIRTUAL SCREENING TARGETING COMPLEMENT RECEPTOR C5AR

4.3.1. Introduction

Following the success of our conformationally sampled pharmacophore model, and the development of our PMX-53:C5aR model, the next logical step is design of new antagonists targeting C5aR. One popular approach is virtual screening [5], which utilizes docking algorithms to orient and rank entire databases of known compounds, with the goal of identifying potential small molecule binders. Virtual screening is much cheaper than traditional high-throughput screening approaches, since computation and structural knowledge are used to prescreen compounds prior to experiment. Virtual screening also has the advantage of identifying small molecules, which are often preferred to peptidic therapeutics due to their better bioavailability and cheaper manufacturing cost. Therefore, the goal of this study is to apply the knowledge gained from our conformationally sampled pharmacophore model and PMX-53:C5aR model development efforts to perform virtual screening studies of C5aR.

4.3.2. Methods

In this study, two approaches have been adopted to aid in the search for novel small molecule inhibitors of C5aR, one based purely on C5aR and a second based on the known potent antagonist PMX-53. All evaluated compounds, in both approaches, came from the ZINC database (UCSF) [13] comprised of commercially-available compounds for virtual screening. The first approach, C5aR-based, utilized the Drugs Now subset of the ZINC database, which currently contains ~6 million compounds. The Drug Now subset is filtered based on the Lipinski rules [176] for selecting compounds. For the C5aR-based approach we utilized a step-wise reduction of the entire ~6 million compounds based on docking each compound to C5aR using DOCK6 [4], with the complexity of the scoring increasing after each step. In this study, we used C5aR receptor structures and binding site representations for the C5aR-based screening that were generated during development of the PMX-53:C5aR model discussed in section 4.2.2. Five molecular dynamics snapshots of C5aR, which resulted in the best bound conformations of PMX-53 (specifically poses 22, 6, 1, 3, and 4), were selected for use in the docking calculations of this study. All of the selected C5aR structures resulted from the elevated temperature simulations, discussed in section 4.2.2, which provide better docking poses due to their more open binding site conformations.

The initial screen of the C5aR-based approach involved docking the ~6 million compounds of the Drugs Now subset to the initial (open) structure of pose 22 (best agreement with experimental data) from the PMX-53:C5aR docking study. The anchor-and-grow flexible ligand algorithm of DOCK6 [4] was used to orient the ligands in the C5aR binding site. A bump filter was employed to exclude docking poses/orientations with more than 3 receptor-ligand clashes, with a clash being defined as an atom-atom overlap of more than 50 percent. Simplex

minimization was applied to optimize ligand orientation, and resulting poses were scored using the grid-based scoring function (grid resolution of 0.3 Å), which accounts for van der Waals and electrostatics contributions to binding. The ~6 million compounds were ranked according the grid scores for the pose 22 based docking, and the top 50,000 compounds were selected for further analysis. These top 50,000 compounds were subsequently docked to 4 additional C5aR structures (initial structures from poses 6, 1, 3, and 4) in order to incorporate receptor flexibility into the virtual screening procedure. A second compound filtering was performed to identify the top 1000 compounds by ranking the compounds according to the minimum of the 5 grid scores, which are a result of docking each compound to the initial structures of poses 22, 6, 1, 3, and 4. The top 1000 compounds were subjected to a final round of optimization by rescoring the lowest energy pose of each compound using the Amber Score function of DOCK6. The Amber Score function includes energy minimization and molecular dynamics for the bound complex, as well as the free receptor and ligand, in order to incorporate flexibility, and scores the relaxed pose based on an association free energy using the AMBER forcefield. All atoms within 7 Å of the binding site were free during the 500 steps of molecular dynamics, as well as during the 100 steps of energy minimization before/after MD, while all other atoms were constrained.

The second virtual screening approach employed in this study utilizes a pharmacophore search performed using the online utility ZINCPharmer [177] to prescreen the ZINC database, prior to docking calculations. The ZINCPharmer webserver compares a pharmacophore query against the over 17 million purchasable compounds of the entire ZINC database, with the conformational space being extended by 10 structures per compound. The lowest energy structure from pose 6 of PMX-53 bound to C5aR was used to represent the optimum configuration of pharmacophore points. The pharmacophore query utilized in this study, which

was selected based on the results of our conformationally-sampled pharmacophore model for PMX-53 analogs (section 0), is illustrated by Figure 4-8. The pharmacophore query (Figure 4-8) identified 323 structures satisfying the pharmacophore model, which correspond to 141 unique compounds. The identified compounds were ranked using the Amber Score function of DOCK6, as described above; however, molecules were not oriented in the binding site of C5aR using DOCK6, instead superposition of pharmacophore points onto the bound model of PMX-53 was utilized. This approach forces identified ligands to be oriented in a manner similar PMX-53, in accordance with similar physicochemical characteristics. Initial conformations of superimposed ligands likely clash with C5aR, but the energy minimization and molecular dynamics should relax the system.

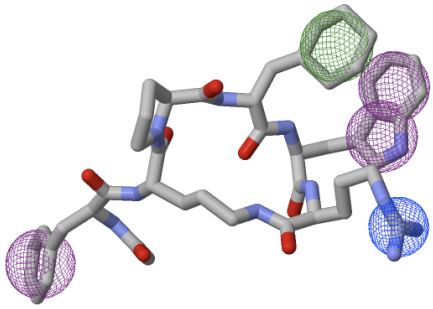


Figure 4-8 *Molecular representation of PMX-53 based pharmacophore query.* PMX-53 is represented by the stick CPK model, while the mesh spheres represent selected pharmacophore points: purple, aromatic; green, hydrophobic; blue, positive ion.

During the selection phase of the virtual screening, clustering based on 2D molecular similarity was performed using the ChemmineR [178] library to insure the selection of a structurally diverse set of compounds. ZINC compounds were converted from the MOL2 format to SDF format using iBabel and Open Babel [179]. Atom pairs for each compound were generated by importing the compounds in SDF format and using the sdf2ap() function of

ChemmineR. Binning clustering was performed using the cmp.cluster() function of ChemmineR based on a cutoff of 0.5.

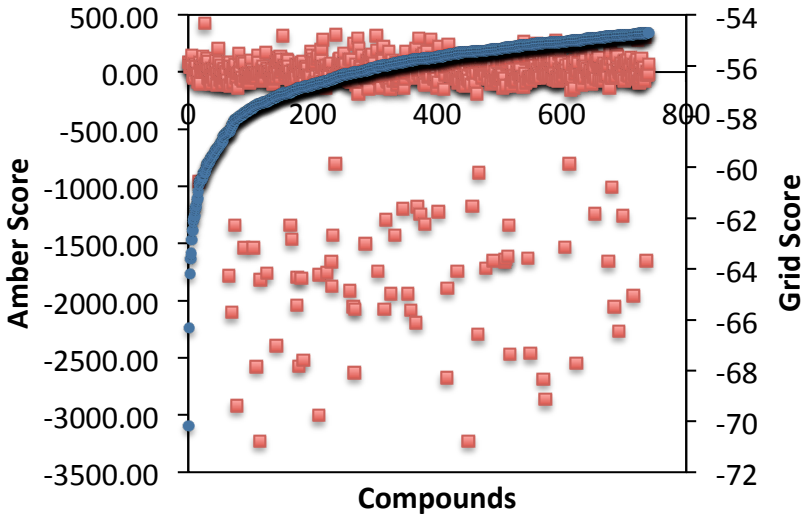


Figure 4-9 *Comparison between Grid and Amber Scores for the top C5aR compounds.* Compounds are ordered according to their Grid Score (blue circles), while Amber Scores (red squares) are also included for comparison.

4.3.3. Results and Discussion

In this study we have performed two virtual screens to identify antagonists of complement receptor C5aR, one based solely on C5aR and a second incorporating knowledge of a known potent antagonist of C5aR, PMX-53. The first approach filtered the ~6 million compounds of the Drugs Now ZINC database subset based on three rounds of docking with multiple C5aR structures. This resulted in ~6 million compounds being scored once, 50,000 of those compounds being score 4 additional times, and 1,000 compounds being scored a sixth and final time. The first two rounds of screening were performed using a grid-based score, which is rather efficient, while the final ranking was performed using the rigorous Amber Score function, which is based on the AMBER implicit solvent molecular mechanics forcefield. Due to difficulties

in parameterizing ligands for the Amber Score function, ~800 compounds were scored using AMBER, and comparison between the Grid Scores and corresponding Amber Scores are provided in Figure 4-9.

Inspection of Figure 4-9 makes clear that there is little correlation between the Grid and Amber Scores, with the best compounds based on the Grid Score resulting in positive Amber Scores. This is slightly concerning, given that the Grid Score was used to rank the compounds in the first two rounds of docking, but isn't unexpected since the Amber Score incorporates dynamics. Also, inspection of the Amber Score results shows that numerous compounds result in unrealistically low scores, which correspond to unrealistic ligand conformations. As a result, only Amber Scores in the range of -200 to 0 were considered in compound selection. As is the case with most virtual screening studies, the ultimate goal of this study is to identify compounds to be experimentally evaluated; therefore, two sets of ligands were selected based on the C5aR driven approach: (i) according to the Grid Score ranking (Table 4-3 and Figure 4-10); (ii) according to the Amber Score ranking (Table 4-4 and Figure 4-11). A third set of compounds were also selected based on Amber Scores for the compounds resulting from the pharmacophore search (Table 4-5 and Figure 4-12). All three sets were selected using a combination of molecular similarity clustering and score-based ranking in order to identify a diverse set of compounds.

Table 4-3 List of C5aR predicted ligands based on Grid Score rank.

#	ZINC ID	Grid Score
1	ZINC05288919	-80.3
2	ZINC27345053	-64.2
3	ZINC12770553	-63.6
4	ZINC12768157	-62.9
5	ZINC08635662	-62.5
6	ZINC09052093	-62.1
7	ZINC12573047	-62.0
8	ZINC24973120	-61.8
9	ZINC52627542	-61.7
10	ZINC12363678	-61.5
11	ZINC09335727	-60.9
12	ZINC20201915	-60.7
13	ZINC32967121	-60.6
14	ZINC33033451	-60.5
15	ZINC40412053	-60.5
16	ZINC20805647	-60.5
17	ZINC06949197	-60.2
18	ZINC08900797	-60.2
19	ZINC25326585	-60.2
20	ZINC12482524	-60.2

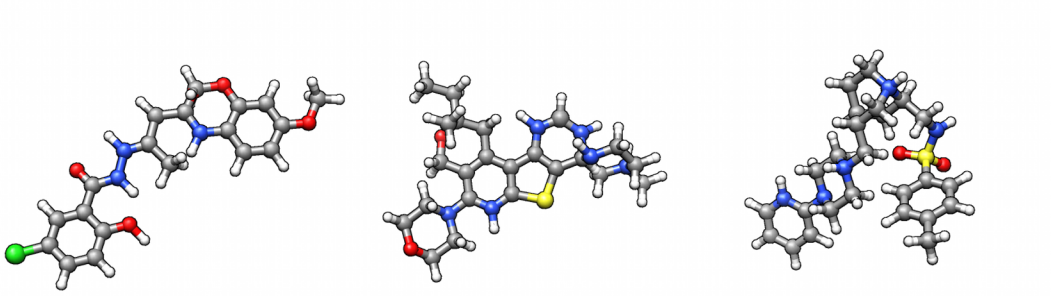


Figure 4-10 *Molecular representations of 3 examples of C5aR predicted ligands based on Grid Score rank.* Compounds are presented as ball and stick models with atoms being colored by element type.

Table 4-4 *List of C5aR predicted ligands based on Amber Score rank.*

#	ZINC ID	Amber Score
1	ZINC05445586	-187.4
2	ZINC22913610	-160.4
3	ZINC32895775	-147.9
4	ZINC20997469	-146.5
5	ZINC22922139	-146.4
6	ZINC20503850	-144.5
7	ZINC22909999	-143.9
8	ZINC11936238	-143.4
9	ZINC11874131	-139.2
10	ZINC09549560	-136.2
11	ZINC61722095	-135.9
12	ZINC11872853	-121.9
13	ZINC32547819	-119.8
14	ZINC08428537	-118.4
15	ZINC04954691	-117.4
16	ZINC20462720	-111.8
17	ZINC08610590	-111.5
18	ZINC12436369	-108.4
19	ZINC39412862	-106.6
20	ZINC12593746	-105.7

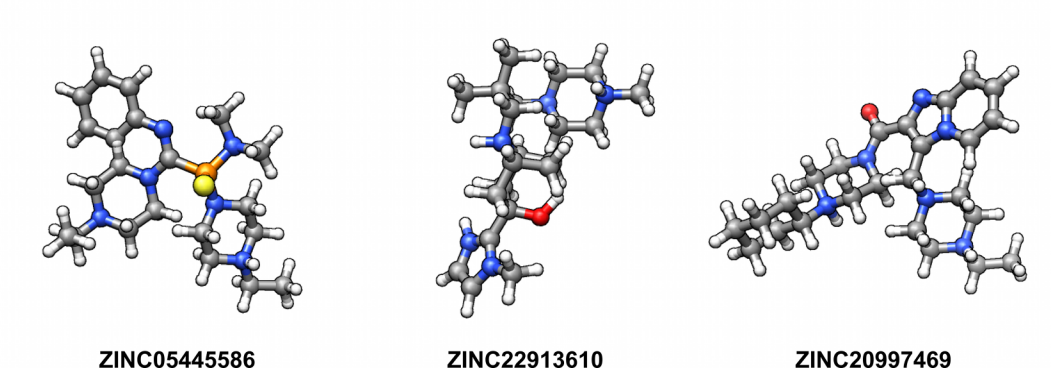


Figure 4-11 *Molecular representations of 3 examples of C5aR predicted ligands based on Amber Score rank.* Compounds are presented as ball and stick models with atoms being colored by element type.

Table 4-5 *List of C5aR predicted ligands based on pharmacophore search.*

#	ZINC ID	Amber Score
1	ZINC01771771	-75.0
2	ZINC12482083	-72.1
3	ZINC08432273	-70.6
4	ZINC12457584	-70.4
5	ZINC09892477	-65.0
6	ZINC12888898	-61.2
7	ZINC09380650	-60.9
8	ZINC14039645	-59.4
9	ZINC08386651	-58.6
10	ZINC60136689	-55.9
11	ZINC09508368	-55.4
12	ZINC12560926	-54.4
13	ZINC36557582	-53.9
14	ZINC12957734	-53.7
15	ZINC13863813	-52.5
16	ZINC12546616	-51.2
17	ZINC52627922	-48.9
18	ZINC62042084	-46.7
19	ZINC13694604	-46.0
20	ZINC40986324	-45.8

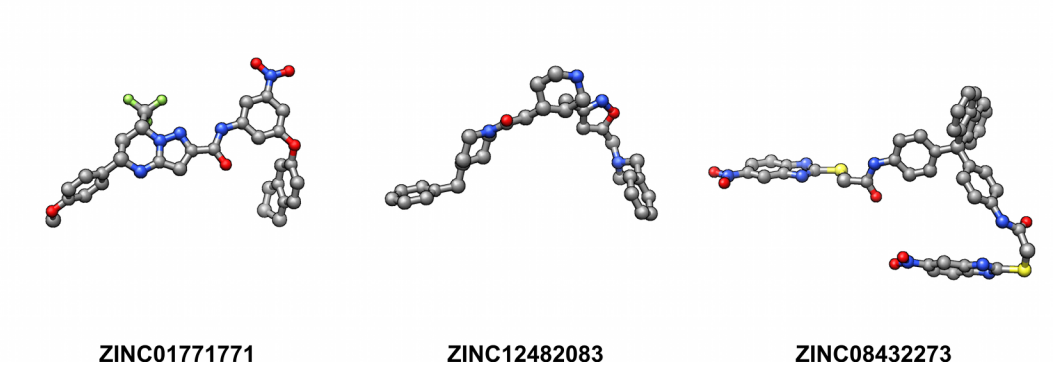


Figure 4-12 *Molecular representations of 3 examples of C5aR predicted ligands based on pharmacophore search.* Compounds are presented as ball and stick models with atoms being colored by element type.

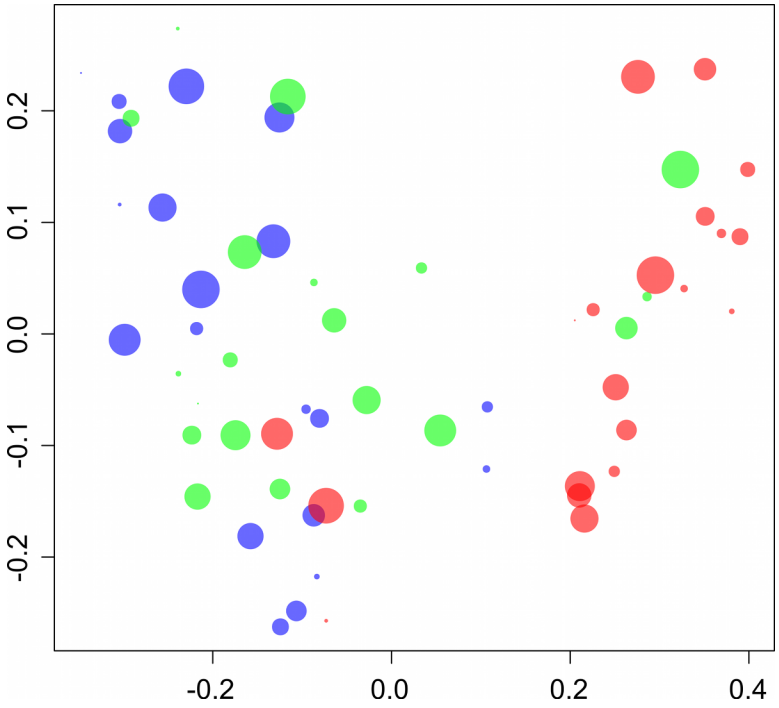


Figure 4-13 *Molecular similarity clustering comparing the three sets of selected compounds.* The color of the circles indicates the compound set: blue, Amber Score (Table 4-4); green, Grid Score (Table 4-3); red, pharmacophore search (Table 4-5). The size of the circle indicates the rank of the compounds in each respective set: large – small, 1 – 20, best – worst.

Molecular similarity clustering was also applied to compare the compounds of the three sets selected based on the virtual screening results (Table 4-3, Table 4-4, and Table 4-5). Clustering results were visualized using multi-dimensional scaling to identify the level of similarity between the compounds of the three selected sets (Figure 4-13). Here, the distance between points is proportional to similarity, and we can see that there is some crossover between the three sets of compounds. The pharmacophore search set (red, Figure 4-13) clusters mostly on its own, which is in accordance with the larger size of most of these compounds as seen in Figure 4-12. In contrast, there is a fair amount of similarity between the compounds of the Amber (blue, Figure 4-13) and Grid Score (green, Figure 4-13) sets. In closing, these compounds represent a diverse set of potential C5aR ligands ready for experimental evaluation.

5. CONCLUSION

Molecular modeling is a powerful tool for investigating the complex nature of biomolecular interactions, with utility in understanding the molecular basis of disease and in the design of novel therapeutics. Due to the immense complexity of biological environments, various methods have been developed to capture different aspects involved in biomolecular function [2-8]. A combination of modeling approaches can provide a more comprehensive view of a particular system, and can have great utility in protein/drug design.

Of the described molecular modeling methods, electrostatics calculations can efficiently provide insight into biomolecular recognition and binding, and therefore can be performed on a much larger scale than simulations such as molecular dynamics [3,9]. Our computational framework, Analysis of Electrostatic Similarities Of Proteins (AESOP) [25,38], takes advantage of the scalability of electrostatics calculations in order to design new protein analogs with tailored electrostatic character. The AESOP framework utilizes theoretical mutagenesis, electrostatic clustering, free energies of association, and calculations of ionization properties to elucidate the role of each charged residue in association. Initial application of the AESOP framework to the gold-standard system barnase-barstar showed the its predictive value in ascertaining the effects of mutations on protein association [25]. However, the design of new SUMO4 substrate analogs was the first application of AESOP in which in the predictions were used to successfully guide experiment (section 2.5). In addition to design, the AESOP framework has also been used to elucidate the role of electrostatics in association for various protein systems [26,38,139,180-183], and has potential utility for analyzing any other electrostatically-driven biomolecular interaction.

The complement system, an ancient component of innate immunity, is a cascade of protein interactions that are largely driven by electrostatic forces. Upon activation the complement system leads to opsonization of foreign pathogens, initiation of an inflammatory response, and bacterial cell lysis [77]; however, over-activation of the complement system can lead to disease [78]. The electrostatic character of complement regulators, such as factor H, drives the recognition of C3b bound to host cells and in healthy individuals to prevent further activation of the complement system [180]. Complement proteins, such as complement fragment C3d, have evolved to contain electrostatic "hot-spots" which guide recognition and often coincide with two distinct functional sites: an ancient site that is part of innate immunity and a newer site that is a link between innate and adaptive immunity (section 3.3). Additionally, we have also investigated the role of electrostatics in the C3d-CR2 interaction, which has been a subject of controversy over the last decade. Due to this crucial role of electrostatics complement system function, pathogens often utilize electrostatics in evasion tactics by inhibiting complement interactions [139] and mimicking native complement regulators, such as Kaposica (section 2.2) [37]. In addition to gaining insight into the role electrostatics in complement system function, these studies of factor H and C3d could also serve as springboards for the electrostatic-based design of complement inhibitors and vaccine adjuvants, respectively.

One of the more popular applications of molecular modeling is computational drug design, which utilizes simulation to identify potential new therapeutics. Various approaches have been developed, which utilize structural information of targets of interest, as well as known ligands of a given target. On the complement system, one of the prime drug targets is anaphylatoxin receptor C5aR, which has been implicated in numerous inflammatory diseases due to its role in chemotaxis and activation of leukocytes [155]. We have developed a ligand-

based conformationally-sampled pharmacophore model for antagonists of C5aR, which is capable of distinguishing the level of affinity of various analogs (section 4.2). Our pharmacophore model provides insight into the mode of action of C5aR antagonists, and could aid in the design/analysis of novel antagonists. Additionally, since small organic compounds are often preferred therapeutic candidates, we have also performed virtual screening of C5aR, based on a recently developed model of the C5aR-PMX-53 interaction. During the screening process we utilized structural knowledge of C5aR and known antagonist PMX-53, to identify a structurally diverse set of potential C5aR ligands. The next step in this project, to be carried out in the near future, involves the experimental evaluation of identified C5aR compounds by our collaborators at University of Queensland and University of Sheffield.

In general, the studies presented here represent a collection of methods and applications that can be utilized in the analysis of biomolecular interactions, as well as in protein/drug design. In addition to the direct findings of these studies, our work provides a blueprint of an approach that has broad application in protein/drug design. Our AESOP framework also will be made available to the public in the near future and should be of interest to researchers of the field. Moving forward, the most important aspect of the methods/approach developed here is its flexibility and adaptability, which allow the addition of new utilities to solve unforeseen problems and to account for new advances. Also, since the principles for understanding biomolecular interactions are ubiquitous, the developed approach can be applied in broad range of systems.

6. REFERENCES

1. Feynman R (2000) Feynman Lectures on Physics Vol. 1: 1–513.

2. Nooren IMA, Thornton JM (2003) Diversity of protein-protein interactions. EMBO J 22: 3486–3492. doi:10.1093/emboj/cdg359.

3. Baker N, Sept D, Joseph S, Holst M, McCammon J (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. P Natl Acad Sci Usa 98: 10037–10041.

4. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, et al. (2009) DOCK 6: combining techniques to model RNA-small molecule complexes. RNA 15: 1219–1230. doi:10.1261/rna.1563609.

5. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3: 935–949. doi:10.1038/nrd1549.

6. Morris G, Goodsell D, Halliday R, Huey R, Hart W, et al. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19: 1639–1662.

7. Ermak D, McCammon J (1978) Brownian dynamics with hydrodynamic interactions. J Chem Phys 69: 1352–1360.

8. Gilson MK, Zhou H-X (2007) Calculation of protein-ligand binding affinities. Annu Rev Bioph Biom 36: 21–42. doi:10.1146/annurev.biophys.36.040306.132550.

9. Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. Science 268: 1144–1149.

10. Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106: 1589–1615. doi:10.1021/cr040426m.

11. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26: 1781–1802. doi:10.1002/jcc.20289.

12. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. J Mol Graphics 14: 33–38.

13. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC - A free tool to discover chemistry for biology. J Chem Inf Model. doi:10.1021/ci3001277.

14. Levesque M, Ichikawa K, Date S, Haga J (2009) Design of a grid service-based platform for in silico protein-ligand screenings. 93: 73–82.

15. Schneider MC, Prosser BE, Caesar JJE, Kugelberg E, Li S, et al. (2009) Neisseria meningitidis recruits factor H using protein mimicry of host carbohydrates. Nature 458: 890–893. doi:10.1038/nature07769.

16. McCammon J, Northrup S, Allison S (1986) Diffusional dynamics of ligand receptor association. J Phys Chem-Us 90: 3901–3905.

17. Carbo R, Leyda L, Arnau M (1980) How similar is a molecule to another - An electron-density measure of similarity between 2 molecular-structures. Int J Quantum Chem 17: 1185–1189.

18. Herbert AP, Deakin JA, Schmidt CQ, Blaum BS, Egan C, et al. (2007) 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. doi:10.1074/jbc.M609636200.

19. Hodgkin E, Richards W (1987) Molecular similarity based on electrostatic potential and electric-field. Int J Quantum Chem: 105–110.

20. Reynolds C, Burt C, Richards W (1992) A linear molecular similarity index. Quant Struct-Act Rel 11: 34–35.

21. Petke J (1993) Cumulative and discrete similarity analysis of electrostatic potentials and fields. J Comput Chem 14: 928–933.

22. Wade R, Gabdoulline R, De Rienzo F (2001) Protein interaction property similarity analysis. Int J Quantum Chem 83: 122–127.

23. Blomberg N, Gabdoulline R, Nilges M, Wade R (1999) Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. Proteins 37: 379–387.

24. Cheung AS, Kieslich CA, Yang J, Morikis D (2010) Solvation Effects in Calculated Electrostatic Association Free Energies for the C3d-CR2 Complex and Comparison with Experimental Data. Biopolymers 93: 509–519. doi:10.1002/bip.21388.

25. Kieslich CA, Gorham RD, Morikis D (2011) Is the rigid-body assumption reasonable? Insights into the effects of dynamics on the electrostatic analysis of barnase-barstar. J Non-Cryst Solids 357: 707–716. doi:10.1016/j.jnoncrysol.2010.05.087.

26. Gorham RD, Kieslich CA, Nichols A, Sausman NU, Foronda M, et al. (2011) An evaluation of Poisson-Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. Biopolymers 95: 746–754. doi:10.1002/bip.21644.

27. Antosiewicz JM (2008) Protonation free energy levels in complex molecular systems. Biopolymers 89: 262–269. doi:10.1002/bip.20837.

28. Wu J, Morikis D (2006) Molecular thermodynamics for charged biomacromolecules. Fluid Phase Equilibr 241: 317–333. doi:10.1016/j.fluid.2005.12.044.

29. Zhang L, Mallik B, Morikis D (2007) Immunophysical exploration of C3d-CR2(CCP1-2) interaction using molecular dynamics and electrostatics. J Mol Biol 369: 567–583. doi:10.1016/j.jmb.2007.02.101.

30. Morikis D, Lambris J (2004) The electrostatic nature of C3d-Complement receptor 2 association. J Immunol 172: 7537–7547.

31. Morikis D, Elcock A, Jennings P, McCammon J (2001) Native-state conformational dynamics of GART: A regulatory pH-dependent coil-helix transition examined by electrostatic calculations. Protein Sci 10: 2363–2378.

32. Morikis D, Elcock A, Jennings P, McCammon J (2001) Proton transfer dynamics of GART: The pH-dependent catalytic mechanism examined by electrostatic calculations. Protein Sci 10: 2379–2392.

33. Morikis D, Roy M, Newlon M, Scott J, Jennings P (2002) Electrostatic properties of the structure of the docking and dimerization domain of protein kinase A II alpha. Eur J Biochem 269: 2040–2051. doi:10.1046/j.1432-1033.2002.02852.x.

34. Mallik B, Zhang L, Koide S, Morikis D (2008) pH dependence of stability of the 10th human fibronectin type III domain: A computational study. Biotechnol Progr 24: 48–55. doi:10.1021/bp0700915.

35. Sfyroera G, Katragadda M, Morikis D, Isaacs S, Lambris J (2005) Electrostatic modeling predicts the activities of orthopoxvirus complement control proteins. J Immunol 174: 2143–2151.

36. Zhang L, Morikis D (2006) Immunophysical properties and prediction of activities for vaccinia virus complement control protein and smallpox inhibitor of complement enzymes using molecular dynamics and electrostatics. Biophys J 90: 3106–3119. doi:10.1529/biophysj.105.068130.

37. Pyaram K, Kieslich CA, Yadav VN, Morikis D, Sahu A (2010) Influence of electrostatics on the complement regulatory functions of Kaposica, the complement inhibitor of Kaposi's sarcoma-associated herpesvirus. J Immunol 184: 1956–1967. doi:10.4049/jimmunol.0903261.

38. Kieslich CA, Morikis D, Yang J, Gunopulos D (2011) Automated computational framework for the analysis of electrostatic similarities of proteins. Biotechnol Progr 27: 316–325. doi:10.1002/btpr.541.

39. Ogata R, Ai R, Low P (1998) Active sites in complement component C3 mapped by mutations at indels. J Immunol 161: 4785–4794.

40. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948. doi:10.1093/bioinformatics/btm404.

41. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31: 3381–3385.

42. Dolinsky T, Nielsen J, McCammon J, Baker N (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res 32: W665–W667. doi:10.1093/nar/gkh381.

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

44. Oran AE, Isenman DE (1999) Identification of residues within the 727-767 segment of human complement component C3 important for its interaction with factor H and with complement receptor 1 (CR1, CD35). J Biol Chem 274: 5120–5130.

45. Liszewski MK, Leung M, Cui W, Subramanian VB, Parkinson J, et al. (2000) Dissecting sites important for complement regulatory activity in membrane cofactor protein (MCP; CD46). J Biol Chem 275: 37692–37701. doi:10.1074/jbc.M004650200.

46. Krych M, Hauhart R, Atkinson JP (1998) Structure-function analysis of the active sites of complement receptor type 1. J Biol Chem 273: 8623–8629.

47. Blom AM, Webb J, Villoutreix BO, Dahlbäck B (1999) A cluster of positively charged amino acids in the C4BP alpha-chain is crucial for C4b binding and factor I cofactor function. J Biol Chem 274: 19237–19245.

48. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. pp. Available:http://www.R-project.org/.

49. Jain A, Murty M, Flynn P (1999) Data clustering: A review. Acm Comput Surv 31: 264–323.

50. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22: 2695–2696. doi:10.1093/bioinformatics/btl461.

51. Krivov GG, Shapovalov MV, Dunbrack RLJ (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77: 778–795. doi:10.1002/prot.22488.

52. Gilson M (1993) Multiple-site titration and molecular modeling: Two rapid methods for computing energies and forces for ionizable groups in proteins - Gilson - 2004 - Proteins: Structure, Function, and Bioinformatics - Wiley Online Library. Proteins: Structure.

53. Pilat Z, Antosiewicz JM (2010) pK(a)'s of Ionizable Groups and Energetics of Protein Conformational Transitions. J Phys Chem B 114: 1393–1406. doi:10.1021/jp9040056.

54. Sheinerman F, Honig B (2002) On the role of electrostatic interactions in the design of protein-protein interfaces. J Mol Biol 318: 161–177. doi:10.1016/S0022-2836(02)00030-X.

55. Spaar A, Dammer C, Gabdoulline R, Wade R, Helms V (2006) Diffusional encounter of barnase and barstar. Biophys J 90: 1913–1924. doi:10.1529/biophysj.105.075507.

56. Spaar A, Helms V (2005) Free energy landscape of protein-protein encounter resulting from Brownian dynamics simulations of Barnase : Barstar. J Chem Theory Comput 1: 723–736. doi:10.1021/ct050036n.

57. Spaar A, Helms V (2006) Ionic strength effects on the association funnel of barnase and barstar investigated by Brownian dynamics simulations. J Non-Cryst Solids 352: 4437–4444. doi:10.1016/j.jnoncrysol.2006.03.117.

58. Vijayakumar M, Wong K, Schreiber G, Fersht A, Szabo A, et al. (1998) Electrostatic enhancement of diffusion-controlled protein-protein association: Comparison of theory and experiment on barnase and barstar. J Mol Biol 278: 1015–1024.

59. Wang T, Tomic S, Gabdoulline R, Wade R (2004) How optimal are the binding energetics of barnase and barstar? Biophys J 87: 1618–1630. doi:10.1529/biophysj.104.040964.

60. Gabdoulline R, Wade R (1997) Simulation of the diffusional association of Barnase and Barstar. Biophys J 72: 1917–1929.

61. Gabdoulline R, Wade R (1999) On the protein-protein diffusional encounter complex. J Mol Recognit 12: 226–234.

62. Camacho C, Weng Z, Vajda S, DeLisi C (1999) Free energy landscapes of encounter complexes in protein-protein association. Biophys J 76: 1166–1178.

63. Lee L, Tidor B (2001) Optimization of binding electrostatics: Charge complementarity in the barnase-barstar protein complex. Protein Sci 10: 362–377.

64. Schreiber G, Fersht A (1993) Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. Biochemistry-Us 32: 5145–5150.

65. Oliveberg M, Arcus V, Fersht A (1995) pK(a) values of carboxyl groups in the native and denatured states of barnase - the pk(a) values of the denatured state are on average 0.4 units lower than those of model compounds. Biochemistry-Us 34: 9424–9433.

66. Olsson MHM, Sondergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK(a) Predictions. J Chem Theory Comput 7: 525–537. doi:10.1021/ct100578z.

67. Schreiber G, Fersht A (1995) Energetics of protein-protein interactions - Analysis of the barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol 248: 478–486.

68. Schreiber G, Fersht A (1996) Rapid, electrostatically assisted association of proteins. Nat Struct Biol 3: 427–431.

69. Schreiber G, Frisch C, Fersht A (1997) The role of Glu73 of barnase in catalysis and the binding of barstar. J Mol Biol 270: 111–122.

70. Buckle A, Schreiber G, Fersht A (1994) Protein-protein recognition: Crystal structural analysis of a barnase-barstar …. Biochemistry-Us.

71. Kerscher O, Felberbaum R (2006) Modification of proteins by ubiquitin and ubiquitin-like proteins. Annu Rev Cell Dev ….

72. Reverter D, Lima CD (2006) Structural basis for SENP2 protease interactions with SUMO precursors and conjugated substrates. Nat Struct Mol Biol 13: 1060–1068. Available:http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17099700&retmode=ref&cmd=prlinks.

73. Owerbach D, McKay EM, Yeh ETH, Gabbay KH, Bohren KM (2005) A proline-90 residue unique to SUMO-4 prevents maturation and sumoylation. Biochem Biophys Res Commun 337: 517–520. doi:10.1016/j.bbrc.2005.09.090.

74. Fiser A, Sali A (2003) MODELLER: Generation and refinement of homology-based protein structure models Vol. 374. pp. 461–.

75. Liu Y, Song Y, Jiang L (2012) Frontiers in Biology, Volume 7, Number 1 - SpringerLink. Frontiers in Biology.

76. Reverter D, Lima CD (2004) A basis for SUMO protease specificity provided by analysis of human Senp2 and a Senp2-SUMO complex. Structure 12: 1519–1531. doi:10.1016/j.str.2004.05.023.

77. Walport M (2001) Advances in immunology: Complement (First of two parts). New Engl J Med 344: 1058–1066.

78. Walport M (2001) Advances in immunology: Complement (Second of two parts). New Engl J Med 344: 1140–1144.

79. Toapanta F, Ross T (2006) Complement-mediated activation of the adaptive immune responses. Immunol Res 36: 197–210.

80. Toapanta FR, DeAlmeida DR, Dunn MD, Ross TM (2010) C3d adjuvant activity is reduced by altering residues involved in the electronegative binding of C3d to CR2. Immunol Lett 129: 32–38. doi:10.1016/j.imlet.2009.12.022.

81. Dempsey P, Allison M, Akkaraju S, Goodnow C, Fearon D (1996) C3d of complement as a molecular adjuvant: Bridging innate and acquired immunity. Science 271: 348–350.

82. Knopf PM, Rivera DS, Hai S-H, McMurry J, Martin W, et al. (2008) Novel function of complement C3d as an autologous helper T-cell target. Immunol Cell Biol 86: 221–225. doi:10.1038/sj.icb.7100147.

83. Ferreira VP, Pangburn MK, Cortes C (2010) Complement control protein factor H: The good, the bad, and the inadequate. Mol Immunol 47: 2187–2197. doi:10.1016/j.molimm.2010.05.007.

84. Edwards A, Ritter R III, Abel K, Manning A (2005) Complement Factor H Polymorphism and Age-Related Macular Degeneration. Science.

85. Gehrs KM, Anderson DH, Johnson LV, Hageman GS (2006) Age-related macular degeneration--emerging pathogenetic and therapeutic concepts. Ann Med 38: 450–471. doi:10.1080/07853890600946724.

86. Hageman G, Anderson D, Johnson L, Hancox L, Taiber A, et al. (2005) A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. P Natl Acad Sci Usa 102: 7227–7232. doi:10.1073/pnas.0501536102.

87. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, et al. (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308: 419–421. doi:10.1126/science.1110359.

88. Klein R, Zeiss C, Chew E, Tsai J, Sackler R, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389. doi:10.1126/science.1109557.

89. Atkinson JP, Goodship THJ (2007) Complement factor H and the hemolytic uremic syndrome. J Exp Med 204: 1245–1248. doi:10.1084/jem.20070664.

90. Caprioli J, Noris M, Brioschi S, Pianetti G, Castelletti F, et al. (2006) Genetics of HUS: the impact of MCP, CFH, and IF mutations on clinical presentation, response to treatment, and outcome. Blood 108: 1267–1279. doi:10.1182/blood-2005-10-007252.

91. de Córdoba SR, de Jorge EG (2008) Translational mini-review series on complement factor H: genetics and disease associations of human complement factor H. Clin Exp Immunol 151: 1–13. doi:10.1111/j.1365-2249.2007.03552.x.

92. Fang CJ, Richards A, Liszewski MK, Kavanagh D, Atkinson JP (2008) Advances in understanding of pathogenesis of aHUS and HELLP. Br J Haematol 143: 336–348. doi:10.1111/j.1365-2141.2008.07324.x.

93. Appel GB, Cook HT, Hageman G, Jennette JC, Kashgarian M, et al. (2005) Membranoproliferative glomerulonephritis type II (dense deposit disease): an update. J Am Soc Nephrol 16: 1392–1403. doi:10.1681/ASN.2005010078.

94. Gros P, Milder FJ, Janssen BJC (2008) Complement driven by conformational changes. Nat Rev Immunol 8: 48–58. doi:10.1038/nri2231.

95. Janssen BJC, Gros P (2007) Structural insights into the central complement component C3. Mol Immunol 44: 3–10. doi:10.1016/j.molimm.2006.06.017.

96. Serruto D, Rappuoli R, Scarselli M, Gros P, van Strijp JAG (2010) Molecular mechanisms of complement evasion: Learning from staphylococci and meningococci. Nat Rev Micro 8: 393–399. doi:10.1038/nrmicro2366.

97. Saunders RE, Abarrategui Garrido C, Fremeaux-Bacchi V, Goicoechea de Jorge E, Goodship THJ, et al. (2007) The interactive factor H-atyplical hemolytic uremic syndrome mutation database and website: Update and integration of membrane cofactor protein and factor I mutations with structural models. Hum Mutat 28: 222–234. doi:10.1002/humu.20435.

98. Schmidt CQ, Herbert AP, Hocking HG, Uhrin D, Barlow PN (2008) Translational mini-review series on complement factor H: structural and functional correlations for factor H. Clin Exp Immunol 151: 14–24. doi:10.1111/j.1365-2249.2007.03553.x.

99. Licht C, Heinen S, Jozsi M, Loeschmann I, Saunders RE, et al. (2006) Deletion of Lys224 in regulatory domain 4 of Factor H reveals a novel pathomechanism for dense deposit disease (MPGN II). Kidney Int 70: 42–50. doi:10.1038/sj.ki.5000269.

100. Soares D, Barlow PN (2005) Complement control protein modules in the regulators of complement activation In: Morikis D, Lambris JD, editors. Structural Biology of the Complement System. Boca Raton, FL, USA: CRC Press/Taylor & Francis Group; 2005. p. 19-62.

101. Aslam M, Perkins SJ (2001) Folded-back solution structure of monomeric factor H of human complement by synchrotron X-ray and neutron scattering, analytical ultracentrifugation and constrained molecular modelling. J Mol Biol 309: 1117–1138. doi:10.1006/jmbi.2001.4720.

102. Wu J, Wu Y-Q, Ricklin D, Janssen BJC, Lambris JD, et al. (2009) Structure of complement fragment C3b-factor H and implications for host protection by complement regulators. Nat Immunol 10: 728–733. doi:10.1038/ni.1755.

103. Janssen BJC, Christodoulidou A, McCarthy A, Lambris JD, Gros P (2006) Structure of C3b reveals conformational changes that underlie complement activity. Nature 444: 213–216. doi:10.1038/nature05172.

104. Morikis D, Zhang L (2006) An immunophysical study of the complement system: Examples for the pH dependence of protein binding and stability. J Non-Cryst Solids 352: 4445–4450. doi:10.1016/j.jnoncrysol.2006.01.104.

105. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18: 2714–2723. doi:10.1002/elps.1150181505.

106. Pettersen E, Goddard T, Huang C, Couch G, Greenblatt D, et al. (2004) UCSF Chimera - A visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612. doi:10.1002/jcc.20084.

107. Schleinkofer K, Wiedemann U, Otte L, Wang T, Krause G, et al. (2004) Comparative structural and energetic analysis of WW domain-peptide interactions. J Mol Biol 344: 865–881. doi:10.1016/j.jmb.2004.09.063.

108. Stein M, Gabdoulline RR, Wade RC (2010) Cross-species analysis of the glycolytic pathway by comparison of molecular interaction fields. Mol Biosyst 6: 152–164. doi:10.1039/b912398a.

109. Wade R, Gabdoulline R, Ludemann S, Lounnas V (1998) Electrostatic steering and ionic tethering in enzyme-ligand binding: Insights from simulations. P Natl Acad Sci Usa 95: 5942–5949.

110. Zhang X, Bajaj CL, Kwon B, Dolinsky TJ, Nielsen JE, et al. (2006) Application of new multiresolution methods for the comparison of biomolecular electrostatic properties in the absence of global structural similarity. Multiscale Model Sim 5: 1196–1213. doi:10.1137/050647670.

111. de Cordoba S, Esparza-Gordillo J, de Jorge E, Lopez-Trascasa M, Sanchez-Corral P (2004) The human complement factor H: functional roles, genetic variations and disease associations. Mol Immunol 41: 355–367. doi:10.1016/j.molimm.2004.02.005.

112. Ferreira VP, Herbert AP, Hocking HG, Barlow PN, Pangburn MK (2006) Critical role of the C-terminal domains of factor H in regulating complement activation at cell surfaces. J Immunol 177: 6308–6316.

113. Blackmore T, Hellwage J, Sadlon T, Higgs N, Zipfel P, et al. (1998) Identification of the second heparin-binding domain in human complement factor H. J Immunol 160: 3342–3348.

114. Arlaud GJ, Barlow PN, Gaboriaud C, Gros P, Narayana SVL (2007) Deciphering complement mechanisms: The contributions of structural biology Vol. 44. pp. 3809–3822. doi:10.1016/j.molimm.2007.06.147.

115. Jokiranta T, Hellwage J, Koistinen V, Zipfel P, Meri S (2000) Each of the three binding sites on complement factor H interacts with a distinct site on C3b. J Biol Chem 275: 27657–27662.

116. Schmidt CQ, Herbert AP, Kavanagh D, Gandy C, Fenton CJ, et al. (2008) A new map of glycosaminoglycan and C3b binding sites on factor H. J Immunol 181: 2610–2619.

117. Sharma A, Pangburn M (1996) Identification of three physically and functionally distinct binding sites for C3b in human complement factor H by deletion mutagenesis. P Natl Acad Sci Usa 93: 10996–11001.

118. Clark SJ, Higman VA, Mulloy B, Perkins SJ, Lea SM, et al. (2006) 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. doi:10.1074/jbc.M605083200.

119. Prosser BE, Johnson S, Roversi P, Herbert AP, Blaum BS, et al. (2007) Structural basis for complement factor H linked age-related macular degeneration. J Exp Med 204: 2277–2283. doi:10.1084/jem.20071069.

120. Mullins RF, Aptsiauri N, Hageman GS (2001) Structure and composition of drusen associated with glomerulonephritis: implications for the role of complement activation in drusen biogenesis. Eye (Lond) 15: 390–395. doi:10.1038/eye.2001.142.

121. Jokiranta TS, Jaakola V-P, Lehtinen MJ, Pärepalo M, Meri S, et al. (2006) Structure of complement factor H carboxyl-terminus reveals molecular basis of atypical haemolytic uremic syndrome. EMBO J 25: 1784–1794. doi:10.1038/sj.emboj.7601052.

122. Richards A, Buddles MR, Donne RL, Kaplan BS, Kirk E, et al. (2001) Factor H mutations in hemolytic uremic syndrome cluster in exons 18-20, a domain important for host cell recognition. Am J Hum Genet 68: 485–490. doi:10.1086/318203.

123. Bhattacharjee A, Lehtinen MJ, Kajander T, Goldman A, Jokiranta TS (2010) Both domain 19 and domain 20 of factor H are involved in binding to complement C3b and C3d. Mol Immunol 47: 1686–1691. doi:10.1016/j.molimm.2010.03.007.

124. Fremeaux-Bacchi V, Miller EC, Liszewski MK, Strain L, Blouin J, et al. (2008) Mutations in complement C3 predispose to development of atypical hemolytic uremic syndrome. Blood 112: 4948–4952. doi:10.1182/blood-2008-01-133702.

125. Okemefuna AI, Li K, Nan R, Ormsby RJ, Sadlon T, et al. (2009) Multimeric Interactions between Complement Factor H and Its C3d Ligand Provide New Insight on Complement Regulation. J Mol Biol 391: 119–135. doi:10.1016/j.jmb.2009.06.013.

126. Pangburn MK, Rawal N, Cortes C, Alam MN, Ferreira VP, et al. (2009) Polyanion-Induced Self-Association of Complement Factor H. J Immunol 182: 1061–1068.

127. Sahu A, Lambris J (2001) Structure and biology of complement protein C3, a connecting link between innate and acquired immunity. Immunol Rev 180: 35–48.

128. Carroll MC (2004) The complement system in regulation of adaptive immunity. Nat Immunol 5: 981–986. doi:10.1038/ni1113.

129. Sunyer J, Boshra H, Lorenzo G, Parra D, Freedman B, et al. (2003) Evolution of complement as an effector system in innate and adaptive immunity. Immunol Res 27: 549–564.

130. Shaw CD, Storek MJ, Young KA, Kovacs JM, Thurman JM, et al. (2010) Delineation of the complement receptor type 2-C3d complex by site-directed mutagenesis and molecular docking. J Mol Biol 404: 697–710. doi:10.1016/j.jmb.2010.10.005.

131. Kovacs JM, Hannan JP, Eisenmesser EZ, Holers VM (2009) Mapping of the C3d ligand binding site on complement receptor 2 (CR2/CD21) using nuclear magnetic resonance and chemical shift analysis. J Biol Chem 284: 9513–9520. doi:10.1074/jbc.M808404200.

132. Hannan J, Young K, Guthridge J, Asokan R, Szakonyi G, et al. (2005) Mutational analysis of the complement receptor type 2 (CR2/CD21)-C3d interaction reveals a putative charged SCR1 binding site for C3d. J Mol Biol 346: 845–858.

133. Clemenza L, Isenman D (2000) Structure-guided identification of C3d residues essential for its binding to complement receptor 2 (CD21). J Immunol 165: 3839–3848.

134. Li K, Okemefuna AI, Gor J, Hannan JP, Asokan R, et al. (2008) Solution structure of the complex formed between human complement C3d and full-length complement receptor type 2. J Mol Biol 384: 137–150. doi:10.1016/j.jmb.2008.08.084.

135. Nagar B, Jones R, Diefenbach R, Isenman D, Rini J (1998) X-ray crystal structure of C3d: A C3 fragment and ligand for complement receptor 2. Science 280: 1277–1281.

136. Clackson T, Wells J (1995) A hot spot of binding energy in a hormone-receptor interface. Science 267: 383–386. doi:10.1126/science.7529940.

137. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. P Natl Acad Sci Usa 100: 5772–5777. doi:10.1073/pnas.1030237100.

138. Sheinerman FB, Norel R, Honig B (2000) Electrostatic aspects of protein-protein interactions. Curr Opin Struc Biol 10: 153–159.

139. Gorham R, Kieslich C, Morikis D (2011) Complement inhibition by Staphylococcus aureus: Electrostatics of C3d–Efbc and C3d–ehp association. Cell Mol Bioeng 5: 32–43.

140. Clark EA, Crennell S, Upadhyay A, Zozulya AV, Mackay JD, et al. (2011) A structural basis for Staphylococcal complement subversion: X-ray structure of the complement-binding domain of Staphylococcus aureus protein Sbi in complex with ligand C3d. Mol Immunol 48: 452–462. doi:10.1016/j.molimm.2010.09.017.

141. UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71–D75. doi:10.1093/nar/gkr981.

142. Edgar R (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797. doi:10.1093/nar/gkh340.

143. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Bioph Biom 29: 291–325.

144. Guo P, Hirano M, Herrin BR, Li J, Yu C, et al. (2009) Dual nature of the adaptive immune system in lampreys. Nature 459: 796–802. doi:10.1038/nature08068.

145. McCammon J (2009) Darwinian biophysics: Electrostatics and evolution in the kinetics of molecular binding. P Natl Acad Sci Usa 106: 7683–7684.

146. Schreiber G, Buckle A, Fersht A (1994) Stability and function: two constraints in the evolution of barstar and other proteins. Structure 2: 945–951.

147. Szakonyi G, Guthridge J, Li D, Young K, Holers V, et al. (2001) Structure of complement receptor 2 in complex with its C3d ligand. Science 292: 1725–1728. doi:10.1126/science.1059118.

148. Isenman DE, Leung E, Mackay JD, Bagby S, van den Elsen JMH (2010) 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. doi:10.4049/jimmunol.0902919.

149. Gilbert HE, Eaton JT, Hannan JP, Holers VM, Perkins SJ (2005) Solution structure of the complex between CR2 SCR 1-2 and C3d of human complement: an X-ray scattering and sedimentation modelling study. J Mol Biol 346: 859–873. doi:10.1016/j.jmb.2004.12.006.

150. Gilbert HE, Asokan R, Holers VM, Perkins SJ (2006) The 15 SCR flexible extracellular domains of human complement receptor type 2 can mediate multiple ligand and antigen interactions. J Mol Biol 362: 1132–1147. doi:10.1016/j.jmb.2006.08.012.

151. van den Elsen JMH, Isenman DE (2011) A Crystal Structure of the Complex Between Human Complement Receptor 2 and Its Ligand C3d. Science 332: 608–611. doi:10.1126/science.1201954.

152. Fernandez H, Henson P, Otani A, Hugli T (1978) Chemotactic response to human C3a and C5a anaphylatoxins. 1. Evaluation of C3a and C5a leukotaxis in vitro and under simulated in vivo conditions. J Immunol 120: 109–115. doi:Article.

153. Haas P-J, van Strijp J (2007) Anaphylatoxins - Their role in bacterial infection and inflammation. Immunol Res 37: 161–175. doi:Review.

154. March DR, Proctor LM, Stoermer MJ, Sbaglia R, Abbenante G, et al. (2004) Potent cyclic antagonists of the complement C5a receptor on human polymorphonuclear leukocytes. Relationships between structures and activity. Mol Pharmacol 65: 868–879.

155. Monk PN, Scola A-M, Madala P, Fairlie DP (2007) Function, structure and therapeutic potential of complement C5a receptors. Brit J Pharmacol 152: 429–448. doi:10.1038/sj.bjp.0707332.

156. Wong AK, Finch AM, Pierens GK, Craik DJ, Taylor SM, et al. (1998) Small molecular probes for G-protein-coupled C5a receptors: Conformationally constrained antagonists derived from the C terminus of the human plasma protein C5a. J Med Chem 41: 3417–3425.

157. Paczkowski N, Finch A, Whitmore J, Short A, Wong A, et al. (1999) Pharmacological characterization of antagonists of the C5a receptor. Brit J Pharmacol 128: 1461–1466.

158. Woodruff TM, Nandakumar KS, Tedesco F (2011) Inhibiting the C5-C5a receptor axis. Mol Immunol 48: 1631–1642. doi:10.1016/j.molimm.2011.04.014.

159. Nikiforovich GV, Marshall GR, Baranski TJ (2008) Modeling molecular mechanisms of binding of the anaphylatoxin C5a to the C5a receptor. Biochemistry-Us 47: 3117–3130. doi:10.1021/bi702321a.

160. Higginbottom A, Cain S, Woodruff T, Proctor L, Madala P, et al. (2005) Comparative agonist/antagonist responses in mutant human C5a receptors define the ligand binding site. J Biol Chem 280: 17831–17840. doi:10.1074/jbc.M410797200.

161. Zhang L, Mallik B, Morikis D (2008) Structural study of Ac-Phe-[Orn-Pro-dCha-Trp-Arg], a potent C5a receptor antagonist, by NMR. Biopolymers 90: 803–815. doi:10.1002/bip.21099.

162. Finch AM, Wong AK, Paczkowski NJ, Wadi SK, Craik DJ, et al. (1999) Low-molecular-weight peptidic and cyclic antagonists of the receptor for the complement factor C5a. J Med Chem 42: 1965–1974.

163. Bernard D, Coop A, MacKerell AD (2003) 2D conformationally sampled pharmacophore: A ligand-based pharmacophore to differentiate delta opioid agonists from antagonists. J Am Chem Soc 125: 3101–3107.

164. Bernard D, Coop A, Mackerell AD (2005) Conformationally sampled pharmacophore for peptidic delta opioid ligands. J Med Chem 48: 7773–7780.

165. Bernard D, Coop A, MacKerell AD (2007) Quantitative conformationally sampled pharmacophore for delta opioid ligands: reevaluation of hydrophobic moieties essential for biological activity. J Med Chem 50: 1799–1809. doi:10.1021/jm0612463.

166. Shim J, Coop A, MacKerell AD (2011) Consensus 3D model of μ-opioid receptor ligand efficacy based on a quantitative Conformationally Sampled Pharmacophore. J Phys Chem B 115: 7487–7496. doi:10.1021/jp202542g.

167. Mallik B, Morikis D (2005) Development of a quasi-dynamic pharmacophore model for anti-complement peptide analogues. J Am Chem Soc 127: 10967–10976. doi:10.1021/ja051004c.

168. Tanizaki S, Feig M (2005) A generalized Born formalism for heterogeneous dielectric environments: application to the implicit modeling of biological membranes. J Chem Phys 122: 124706. doi:10.1063/1.1865992.

169. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, et al. (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30: 1545–1614. doi:10.1002/jcc.21287.

170. Foloppe N, Mackerell AD (2000) All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. J Comput Chem 21: 86–104.

171. MacKerell AD, Feig M, Brooks CL (2004) Improved treatment of the protein backbone in empirical force fields. J Am Chem Soc 126: 698–699. doi:10.1021/ja036959e.

172. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. J Comput Phys 23: 327–341.

173. Chocholousová J, Feig M (2006) Implicit solvent simulations of DNA and DNA-protein complexes: agreement with explicit solvent vs experiment. J Phys Chem B 110: 17240–17251. doi:10.1021/jp0627675.

174. Lee MS, Feig M, Salsbury FR, Brooks CL (2003) New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J Comput Chem 24: 1348–1356. doi:10.1002/jcc.10272.

175. Richards FM (1977) Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng 6: 151–176. doi:10.1146/annurev.bb.06.060177.001055.

176. Lipinski CA (2000) Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods 44: 235–249.

177. Koes DR, Camacho CJ (2012) ZINCPharmer: pharmacophore search of the ZINC database. Nucleic Acids Res. doi:10.1093/nar/gks378.

178. Cao Y, Charisi A, Cheng L, Jiang T, Girke T (2008) ChemmineR: a compound mining framework for R. 24: 1733–1734.

179. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, et al. (2011) Open Babel: An open chemical toolbox. J Cheminform 3: 33–. doi:10.1186/1758-2946-3-33.

180. Kieslich CA, Vazquez H, Goodman GN, López de Victoria A, Morikis D (2011) The effect of electrostatics on factor H function and related pathologies. J Mol Graph Model 29: 1047–1055. doi:10.1016/j.jmgm.2011.04.010.

181. Chae K, Gonong BJ, Kim S-C, Kieslich CA, Morikis D, et al. (2010) A multifaceted study of stigma/style cysteine-rich adhesin (SCA)-like Arabidopsis lipid transfer proteins (LTPs) suggests diversified roles for these LTPs in plant growth and reproduction. J Exp Bot 61: 4277–4290. doi:10.1093/jxb/erq228.

182. Gorham RD, Kieslich CA, Morikis D (2011) Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. Ann Biomed Eng 39: 1252–1263. doi:10.1007/s10439-010-0226-9.

183. El-Assaad AM, Kieslich CA, Gorham RD, Morikis D (2011) Electrostatic exploration of the C3d-FH4 interaction using a computational alanine scan. Mol Immunol 48: 1844–1850. doi:10.1016/j.molimm.2011.05.007.

184. Hocking HG, Herbert AP, Kavanagh D, Soares DC, Ferreira VP, et al. (2008) Structure of the N-terminal region of complement factor H and conformational implications of disease-linked sequence variations. J Biol Chem 283: 9475–9487. doi:10.1074/jbc.M709587200.

185. Schmidt CQ, Herbert AP, Mertens HDT, Guariento M, Soares DC, et al. (2010) The central portion of factor H (modules 10-15) is compact and contains a structurally deviant CCP module. J Mol Biol 395: 105–122. doi:10.1016/j.jmb.2009.10.010.

186. Barlow PN, Steinkasserer A, Norman DG, Kieffer B, Wiles AP, et al. (1993) Solution structure of a pair of complement modules by nuclear magnetic resonance. J Mol Biol 232: 268–284. doi:10.1006/jmbi.1993.1381.

187. Norman DG, Barlow PN, Baron M, Day AJ, Sim RB, et al. (1991) Three-dimensional structure of a complement control protein module in solution. J Mol Biol 219: 717–725.

188. Herbert AP, Uhrin D, Lyon M, Pangburn MK, Barlow PN (2006) Disease-associated sequence variations congregate in a polyanion recognition patch on human factor H revealed in three-dimensional structure. J Biol Chem 281: 16512–16520. doi:10.1074/jbc.M513611200.

Appendix A *Percent identities for FH CCP modules.*

	CCP1	CCP2	CCP3	CCP4	CCP5	CCP6	CCP7	CCP8	CCP9	CCP10	CCP11	CCP12	CCP13	CCP14	CCP15	CCP16	CCP17	CCP18	CCP19	CCP20
CCP1	100	21	16	20	9	5	12	20	16	14	18	16	17	16	24	18	16	29	29	8
CCP2	21	100	19	16	14	15	18	21	21	20	25	9	11	24	24	22	20	12	25	14
CCP3	16	19	100	16	11	15	12	15	7	16	14	12	7	22	14	14	16	38	32	11
CCP4	20	16	16	100	32	32	28	28	35	18	18	16	19	20	28	26	22	26	22	16
CCP5	9	14	11	32	100	29	33	25	27	7	18	20	15	24	18	29	16	20	25	14
CCP6	5	15	15	32	29	100	29	20	19	16	23	25	15	16	24	27	12	14	16	8
CCP7	12	18	12	28	33	29	100	25	35	16	16	24	7	22	18	18	16	16	18	12
CCP8	20	21	15	28	25	20	25	100	30	12	18	22	11	27	24	25	21	23	25	13
CCP9	16	21	7	35	27	19	35	30	100	18	14	24	13	18	17	14	16	14	20	16
CCP10	14	20	16	18	7	16	16	12	18	100	32	25	5	16	18	27	23	16	12	10
CCP11	18	25	14	18	18	23	16	18	14	32	100	38	21	29	29	33	29	30	34	7
CCP12	16	9	12	16	20	25	24	22	24	25	38	100	13	25	31	33	24	24	31	12
CCP13	17	11	7	19	15	15	7	11	13	5	21	13	100	15	17	21	19	17	19	9
CCP14	16	24	22	20	24	16	22	27	18	16	29	25	15	100	37	27	22	31	35	22
CCP15	24	24	14	28	18	24	18	24	17	18	29	31	17	37	100	38	18	32	36	19
CCP16	18	22	14	26	29	27	18	25	14	27	33	33	21	27	38	100	31	31	29	9
CCP17	16	20	16	22	16	12	16	21	16	23	29	24	19	22	18	31	100	36	29	16
CCP18	29	12	38	26	20	14	16	23	14	16	30	24	17	31	32	31	36	100	40	14
CCP19	29	25	32	22	25	16	18	25	20	12	34	31	19	35	36	29	29	40	100	20
CCP20	8	14	11	16	14	8	12	13	16	10	7	12	9	22	19	9	16	14	20	100

Color in vertical columns corresponds to structure determination method:

NMR	
X-ray	
Both	
N/A	

Appendix B *Sequences of the flexible linkers of FH CCP modules and the N- and C-termini outside the first and last CCP modules and their net charges.* CCP modules are defined with boundaries the first and last conserved cysteine of each module.

Linker	Sequence	Length	Charge
NT	ED	2	-2
1-2	QKRP	4	+2
2-3	EVVK	4	+0
3-4	VEIS	4	-1
4-5	EEKS	4	-1
5-6	TLKP	4	+1
6-7	LRK	3	+2
7-8	IRVKT	5	+2
8-9	IKS	3	+1
9-10	YERE	4	-1
10-11	KEQVQS	6	+0
11-12	IVEEST	6	-2
12-13	VAIDKLKK	8	+2
13-14	SMAQIQL	7	+0
14-15	VEKIP	5	+0
15-16	EGLP	4	-1
16-17	IKTD	4	+0
17-18	RDTS	4	+0
18-19	KDSTGK	6	+1
19-20	LHP	3	+0
CT	AK	2	+1

Appendix C *List of hydrogen bonds formed between Factor H (CCP1-4) and C3b.*

Factor H			C3b				Distance (Å)
Residue Type	Residue Number	Atom	Residue Type	Residue Number	Atom	Domain	
ARG	39	NE	ASN	738	O	a'NT	3.87
ARG	39	NH2	ASN	738	O	a'NT	3.18
ARG	39	NH2	SER	900	OG	MG7	2.94
LEU	41	O	ILE	733	N	a'Nt	3.18
ARG	**60**	**NH2**	**ASP**	**732**	**OD1**	**a'Nt**	**2.75**
ARG	65	NE	GLU	737	O	a'Nt	3.74
ARG	65	NE	ASN	738	OD1	a'NT	3.77
ARG	65	NH2	GLU	737	O	a'NT	2.78
ASP	**72**	**OD2**	**LYS**	**774**	**NZ**	**MG6**	**2.69**
GLU	98	OE1	ARG	570	NH1	MG6	4.08
GLY	153	O	SER	159	N	MG2	2.97
GLN	154	NE2	ASP	156	OD1	MG2	3.66
ALA	155	O	SER	157	OG	MG2	4.10
ALA	155	N	SER	157	O	MG2	2.83
ARG	157	NE	GLU	189	OE2	MG2	3.09
ARG	**157**	**NH2**	**GLU**	**189**	**OE2**	**MG2**	**2.75**
GLU	171	OE1	THR	140	OG1	MG2	3.06
GLU	171	OE2	THR	140	OG1	MG2	3.02
GLN	205	N	GLU	1138	OE2	C3d	3.14
LYS	206	N	GLU	1138	OE2	C3d	3.60
ARG	214	NH1	ASN	1069	OD1	C3d	2.69
GLN	216	OE1	SER	1075	OG	C3d	2.88
GLU	**227**	**OE2**	**LYS**	**43**	**NZ**	**MG1**	**3.10**
ARG	228	NH1	VAL	1068	O	C3d	4.14
ARG	228	NH2	GLY	42	O	MG1	3.07
GLU	**235**	**OE2**	**ARG**	**72**	**NH2**	**MG1**	**3.14**

Colored blocks denote FH modules CCP 1 to 4.
The CCP1-CCP2 linker block is not colored.
Relaxed distance and angle criteria were used.
Bold face indicates salt bridges.

Appendix D *List of Coulombic interactions (up to 8 Å) formed between Factor H (CCP1-4) and C3b.*

Factor H			C3b				Unfavorable	Distance (Å)
Residue Type	Residue Number	Atom	Residue Type	Residue Number	Atom	Domain		
ARG	**60**	**CZ**	**ASP**	**732**	**CG**	**a'NT**		**4.69**
ARG	60	CZ	GLU	731	CD	a'NY		6.97
ARG	65	CZ	LYS	891	NZ	MG7	YES	7.73
ARG	65	CZ	GLU	737	CD	a'NT		5.11
HIS	**69**	**NE2**	**LYS**	**774**	**NZ**	**MG6**		**4.19**
HIS	69	ND1	ASP	775	CG	MG6		6.59
HIS	**69**	**ND1**	**GLU**	**744**	**CD**	**a'NT**		**3.90**
ASP	**72**	**CG**	**LYS**	**774**	**NZ**	**MG6**		**3.63**
ASP	72	CG	GLU	744	CD	a'NT	YES	6.58
GLU	98	CD	ARG	570	CZ	MG6		6.02
ASP	114	CG	GLU	737	CD	a'NT	YES	5.87
LYS	138	NZ	HIS	1290	ND1	CUB		7.66
GLU	145	CD	HIS	1287	ND1	CUB		5.57
HIS	151	NE2	LYS	154	NZ	MG2		6.52
HIS	151	NE2	ASP	156	CG	MG2		5.72
ARG	157	CZ	LYS	75	NZ	MG1	YES	6.34
ARG	**157**	**CZ**	**GLU**	**189**	**CD**	**MG2**		**3.83**
GLU	170	CD	ARG	72	CZ	MG1		5.33
GLU	**171**	**CD**	**GLU**	**189**	**CD**	**MG2**	**YES**	**4.33**
LYS	206	NZ	GLU	1137	CD	C3d		7.10
LYS	206	NZ	GLU	1138	CD	C3d		6.44
ARG	214	CZ	LYS	44	NZ	MG1	YES	7.91
ARG	214	CZ	ASP	1134	CG	C3d		6.94
GLU	**227**	**CD**	**LYS**	**43**	**NZ**	**MG1**		**3.89**
GLU	227	CD	LYS	1028	NZ	C3d		7.72
GLU	227	CD	HIS	38	NE2	MG1		5.39
ARG	**228**	**CZ**	**LYS**	**44**	**NZ**	**MG1**	**YES**	**4.61**
GLU	**235**	**CD**	**ARG**	**72**	**CZ**	**MG1**		**4.81**
GLU	235	CD	GLU	73	CD	MG1	YES	7.80
ARG	239	CZ	GLU	73	CD	MG1		6.86

Colored blocks denote FH modules CCP 1 to 4.
The CCP1-CCP2 linker block is not colored.
The central atom of the charged center was used to calculate distances.
Bold face indicates distance < 5 Å.
Underlining indicates distance < 4 Å.

Appendix E *List of all intramolecular interactions (up to 5 Å) formed between Factor H (CCP1-4) and C3b.*

Factor H			C3b				Distance (Å)
Residue Type	Residue Number	Atom	Residue Type	Residue Number	Atom	Domain	
TYR	38	O	ASN	738	OD1	a'NT	4.11
ARG	**39**	**NH2**	**SER**	**900**	**OG**	**MG7**	**2.94**
ARG	**39**	**NH2**	**ASN**	**738**	**O**	**a'NT**	**3.18**
ARG	**39**	**CD**	**PHE**	**898**	**CZ**	**MG7**	**3.40**
ARG	39	NH2	ILE	734	CG2	a'NT	3.77
ARG	39	NH2	ILE	739	CA	a'NT	4.20
SER	40	CA	PHE	898	CE1	MG7	3.78
SER	40	O	ILE	734	CD1	a'NT	3.83
SER	40	O	ILE	733	O	a'NT	4.18
SER	40	N	ASN	738	ND2	a'NT	4.40
SER	40	O	ALA	735	N	a'NT	4.56
LEU	**41**	**O**	**ILE**	**733**	**O**	**a'NT**	**2.74**
LEU	**41**	**CD2**	**HIS**	**897**	**O**	**MG7**	**3.40**
LEU	**41**	**CD1**	**PHE**	**898**	**CD1**	**MG7**	**3.47**
LEU	41	CD2	ASP	732	OD1	a'NT	3.65
LEU	41	CD2	HIS	896	C	MG7	3.88
LEU	41	N	ILE	734	CD1	a'NT	4.41
LEU	41	O	GLU	731	C	a'NT	4.60
LEU	41	O	ASP	730	O	a'NT	4.80
GLY	42	N	ILE	733	O	a'NT	4.81
ARG	**60**	**NH2**	**ASP**	**732**	**OD1**	**a'NT**	**2.75**
ARG	60	NH2	GLU	731	C	a'NT	4.94
LYS	61	O	PHE	898	CE1	MG7	4.98
GLN	63	CB	PHE	898	CZ	MG7	4.22
ARG	**65**	**NH2**	**GLU**	**737**	**O**	**a'NT**	**2.78**
ARG	65	NE	ASN	738	OD1	a'NT	3.77
ARG	65	NH2	VAL	740	CG2	a'NT	4.05
PRO	**66**	**O**	**VAL**	**740**	**CG2**	**a'NT**	**3.12**
PRO	66	CD	ASN	738	O	a'NT	4.42
CYS	67	CA	VAL	740	CG2	a'NT	4.37
GLY	**68**	**N**	**VAL**	**740**	**CG1**	**a'NT**	**3.37**
HIS	**69**	**CB**	**GLU**	**744**	**CG**	**a'NT**	**3.39**
HIS	69	CE1	LYS	774	NZ	MG6	3.83
HIS	69	CD2	LEU	773	O	MG6	4.30
HIS	69	CB	ASP	775	OD1	MG6	4.40
HIS	69	NE2	PHE	772	CB	MG6	4.53
GLY	71	CA	GLU	744	OE2	a'NT	3.63
ASP	**72**	**OD2**	**LYS**	**774**	**NZ**	**MG6**	**2.69**
ASP	72	N	GLU	744	OE1	a'NT	4.49
PHE	75	CE1	SER	170	OG	MG2	3.87
PHE	75	CE1	LEU	124	CD2	MG2	4.56
THR	77	CG2	ASN	752	OD1	MG6	3.64
PHE	78	CE1	PHE	772	CD2	MG6	4.28
PHE	78	CZ	PHE	772	CB	MG6	4.70
PHE	78	CZ	PHE	772	CD2	MG6	4.72
PHE	78	CE1	PHE	772	CE2	MG6	4.76
PHE	78	CE1	PHE	772	CG	MG6	4.79
PHE	78	CZ	PHE	772	CG	MG6	4.94
PHE	78	CE1	PHE	772	CB	MG6	4.97
LEU	80	CD2	PHE	772	CZ	MG6	3.52
ASN	**84**	**O**	**PHE**	**772**	**CE1**	**MG6**	**3.30**
ASN	**84**	**ND2**	**GLN**	**558**	**NE2**	**MG6**	**3.49**
VAL	85	N	PHE	772	CE1	MG6	4.73
PHE	**86**	**CZ**	**PHE**	**772**	**CD1**	**MG6**	**3.40**
PHE	86	CE2	GLY	556	O	MG6	3.60
GLU	98	OE1	GLN	163	OE1	MG2	3.66
GLU	98	OE1	ARG	570	NH1	MG6	4.08
GLU	98	OE2	ASN	162	O	MG2	4.12
TYR	100	OH	SER	170	OG	MG2	4.76
ASP	114	OD1	GLU	737	OE2	a'NT	3.91

Appendix E (cont.)

Factor H			C3b				Distance (Å)
Residue Type	Residue Number	Atom	Residue Type	Residue Number	Atom	Domain	
VAL	140	O	GLU	1292	OE2	CUB	4.57
SER	141	CB	GLU	189	OE2	MG2	4.65
SER	141	CB	SER	157	OG	MG2	4.77
SER	142	N	SER	157	CB	MG2	4.85
ALA	143	CB	GLN	155	O	MG2	3.94
ALA	143	CB	ASP	156	OD1	MG2	4.50
MET	**144**	**CE**	**SER**	**1293**	**OG**	**CUB**	**3.31**
MET	**144**	**CG**	**ARG**	**1298**	**NH1**	**CUB**	**3.37**
MET	144	CE	GLN	155	OE1	MG2	3.86
MET	144	SD	ARG	1288	O	CUB	4.12
MET	144	CE	LEU	1296	CD2	CUB	4.33
MET	144	SD	HIS	1290	CD2	CUB	4.65
GLU	**145**	**OE2**	**THR**	**1286**	**O**	**CUB**	**3.49**
GLU	145	OE2	HIS	1287	CE1	CUB	4.38
PRO	146	CG	ARG	1288	CG	CUB	4.23
PRO	146	CD	THR	1286	CG2	CUB	4.36
HIS	151	CE1	ASP	156	OD2	MG2	4.54
PHE	**152**	**CE2**	**GLN**	**163**	**NE2**	**MG2**	**3.40**
PHE	152	O	LEU	158	CD2	MG2	4.80
GLY	**153**	**O**	**SER**	**159**	**N**	**MG2**	**2.97**
GLY	**153**	**O**	**LEU**	**158**	**CA**	**MG2**	**3.18**
GLY	153	O	SER	157	O	MG2	4.20
GLN	**154**	**CA**	**SER**	**157**	**O**	**MG2**	**3.20**
GLN	154	CG	ASP	156	OD1	MG2	3.55
GLN	154	CA	LEU	158	CA	MG2	4.21
GLN	154	C	SER	159	N	MG2	4.95
ALA	**155**	**N**	**SER**	**157**	**O**	**MG2**	**2.83**
ALA	155	N	LEU	158	CA	MG2	4.00
ALA	155	CB	SER	159	N	MG2	4.20
ALA	155	CB	THR	140	OG1	MG2	4.28
ARG	**157**	**NH2**	**GLU**	**189**	**OE2**	**MG2**	**2.75**
GLU	**170**	**OE2**	**ARG**	**72**	**NH1**	**MG1**	**2.89**
GLU	**171**	**OE2**	**THR**	**140**	**OG1**	**MG2**	**3.02**
GLU	**171**	**OE1**	**GLU**	**189**	**OE1**	**MG2**	**3.16**
HIS	**173**	**CE1**	**GLN**	**161**	**OE1**	**MG2**	**3.34**
HIS	173	ND1	SER	159	OG	MG2	3.74
HIS	173	NE2	ASN	162	ND2	MG2	4.87

Appendix E (cont.)

Factor H			C3b				Distance (Å)
Residue Type	Residue Number	Atom	Residue Type	Residue Number	Atom	Domain	
ILE	203	CD1	GLN	1076	OE1	C3d	3.77
ILE	203	CG2	GLN	1139	OE1	C3d	4.28
ILE	203	CD1	SER	1075	CB	C3d	4.63
ILE	203	O	GLU	1138	OE1	C3d	4.79
SER	**204**	**CB**	**GLU**	**1138**	**OE1**	**C3d**	**3.36**
GLN	**205**	**N**	**GLU**	**1138**	**OE2**	**C3d**	**3.14**
LYS	206	N	GLU	1138	OE2	C3d	3.60
ASN	212	O	ASP	1134	OD2	C3d	4.41
ARG	**214**	**NH1**	**ASN**	**1069**	**OD1**	**C3d**	**2.69**
ARG	214	CB	ASP	1134	O	C3d	3.64
ARG	214	CD	ILE	1135	CG1	C3d	3.82
ARG	214	NH1	VAL	1068	CG2	C3d	3.96
GLN	**216**	**OE1**	**SER**	**1075**	**OG**	**C3d**	**2.88**
GLN	216	NE2	ILE	1135	CG2	C3d	3.57
GLN	216	CG	GLN	1139	OE1	C3d	4.33
GLN	216	NE2	ASP	1134	O	C3d	4.54
TYR	225	CD2	ALA	1072	CB	C3d	4.03
GLU	**227**	**OE2**	**LYS**	**43**	**NZ**	**MG1**	**3.10**
GLU	227	O	ALA	1072	CB	C3d	3.61
GLU	227	OE2	GLN	87	OE1	MG1	4.10
GLU	227	OE2	HIS	38	CE1	MG1	4.57
GLU	227	O	ILE	1071	C	C3d	4.89
ARG	**228**	**CD**	**VAL**	**1068**	**O**	**C3d**	**2.97**
ARG	**228**	**NE**	**ASN**	**1069**	**O**	**C3d**	**3.03**
ARG	**228**	**NH2**	**GLY**	**42**	**O**	**MG1**	**3.07**
ARG	228	NH2	LYS	44	NZ	MG1	3.82
ARG	228	CD	ILE	1071	O	C3d	4.20
ARG	228	CD	LEU	1070	N	C3d	4.21
ARG	228	C	ALA	1072	CB	C3d	4.29
ARG	228	NH2	LYS	43	O	MG1	4.60
GLY	229	N	ALA	1072	CB	C3d	4.08
GLY	229	N	ILE	1071	O	C3d	4.95
ASP	230	OD1	VAL	1068	O	C3d	4.04
GLU	**235**	**OE2**	**ARG**	**72**	**NH2**	**MG1**	**3.14**
SER	236	OG	ARG	72	NH2	MG1	4.07
SER	236	OG	GLU	73	OE2	MG1	4.41
TRP	238	O	LYS	44	O	MG1	4.98
ARG	239	NE	LEU	45	CD2	MG1	3.79
ARG	239	CG	LYS	44	O	MG1	3.94
PRO	**240**	**CA**	**LYS**	**44**	**O**	**MG1**	**3.37**
PRO	240	CB	LYS	43	O	MG1	3.97
LEU	241	N	LYS	44	O	MG1	4.39
LEU	241	CD1	VAL	46	CG1	MG1	4.87

Colored blocks denote FH modules CCP 1 to 4.
The CCP1-CCP2 linker block is not colored.
Bold face indicates distance < 3.5 *Å*.
Underlining indicates distance < 3 *Å*.

Appendix F *List of available experimentally determined structures of Factor H modules.*

PDB Code	Module(s)	Experimental method	Reference
2RLP	1-2	NMR	[184]
2RLQ	2-3	NMR	[184]
2WII	1-4	X-ray	[102]
-	5	NMR	[100]*
2W80	6-7	X-ray	[15]
2W81	6-7	X-ray	[15]
2JGW	7	NMR	[18]
2JGX	7	NMR	[18]
2UWN	6-8	X-ray	[119]
2V8E	6-8	X-ray	[119]
2KMS	12-13	NMR	[185]
1HFH	15-16	NMR	[186]
1HFI	15	NMR	[186]
1HCC	16	NMR	[187]
2BZM	19-20	NMR	[188]
2G7I	19-20	X-ray	[121]

*Coordinates from: http://www.bionmr.chem.ed.ac.uk/bionmr/public_html/FH~05.pdb

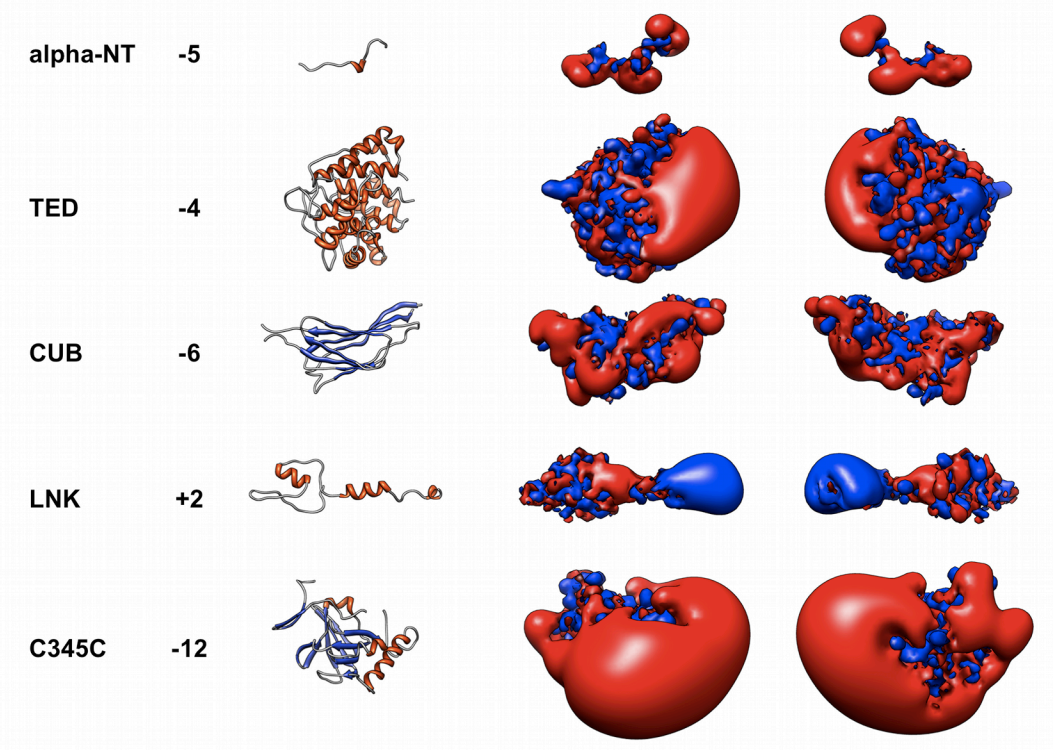
Appendix G *List of experimental coordinate files used for Factor H electrostatics and reasoning.*

Domain	PDB Code	Reason for selection
CCP1	2RLP	Only available structure
CCP2	2RLQ	More secondary structure
CCP3	2RLQ	Only available structure
CCP4	2WII	Only available structure
CCP5	Online[b]	Only available structure
CCP6	2UWN	No MSE[a] residues
CCP7	2UWN	No MSE residues
CCP8	2UWN	No MSE residues
CCP12	2KMS	Only available structure
CCP13	2KMS	Only available structure
CCP15	1HFI	More secondary structure
CCP16	1HCC	More secondary structure
CCP19	2BZM	X-ray structure is missing side chains
CCP20	2BZM	X-ray structure is missing side chains

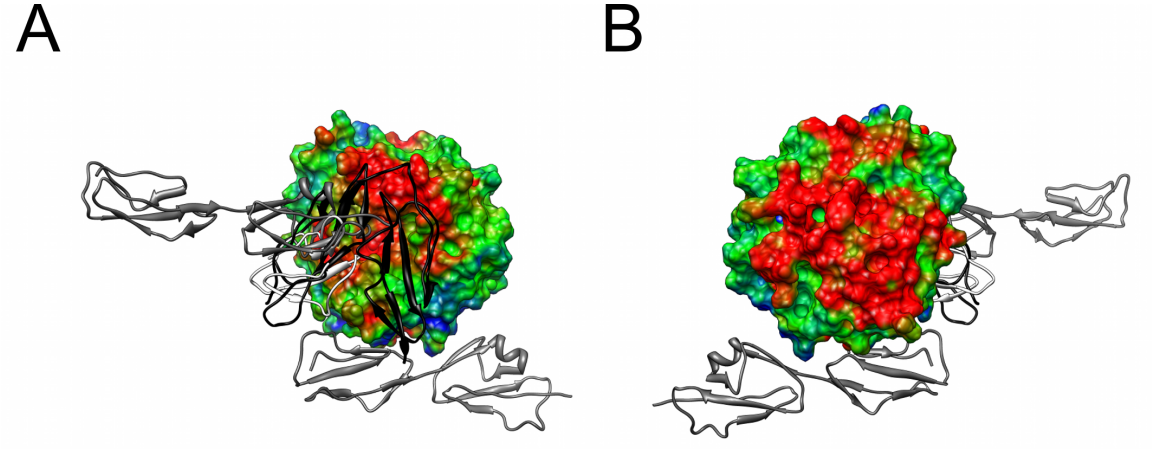
[a]MSE: selenomethionine.
[b]from http://www.bionmr.chem.ed.ac.uk/bionmr/public_html/FH~05.pdb.

Appendix H *Template modules used for homology modeling of Factor H and percent identities.* The coordinate files for the template structures are as in Appendix G.

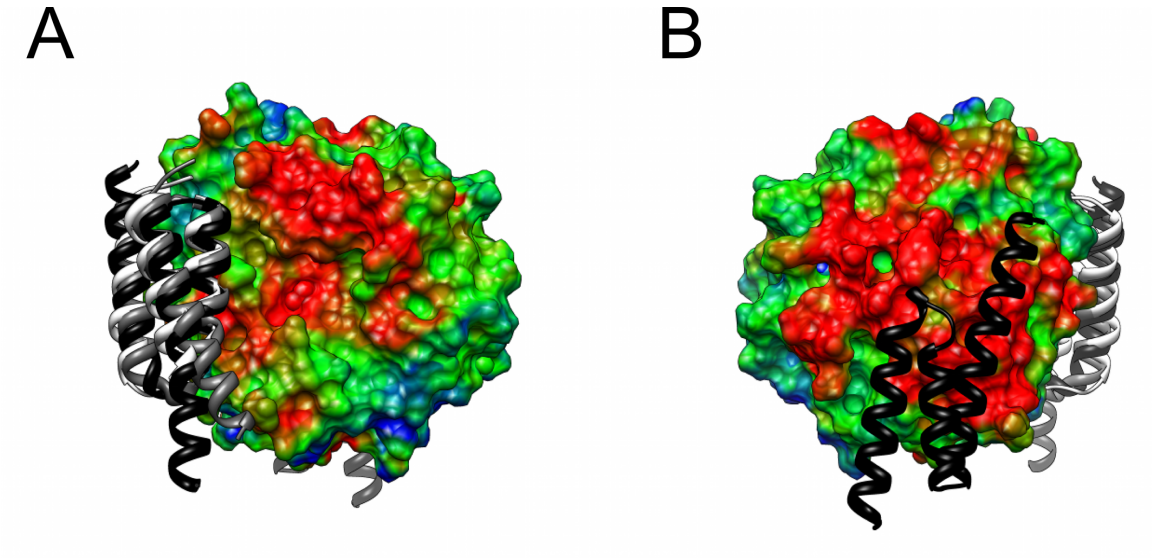
Model	Template	% ID
CCP9	CCP7	35
CCP10	CCP16	27
CCP11	CCP19	34
CCP14	CCP15	37
CCP17	CCP16	31
CCP18	CCP19	40



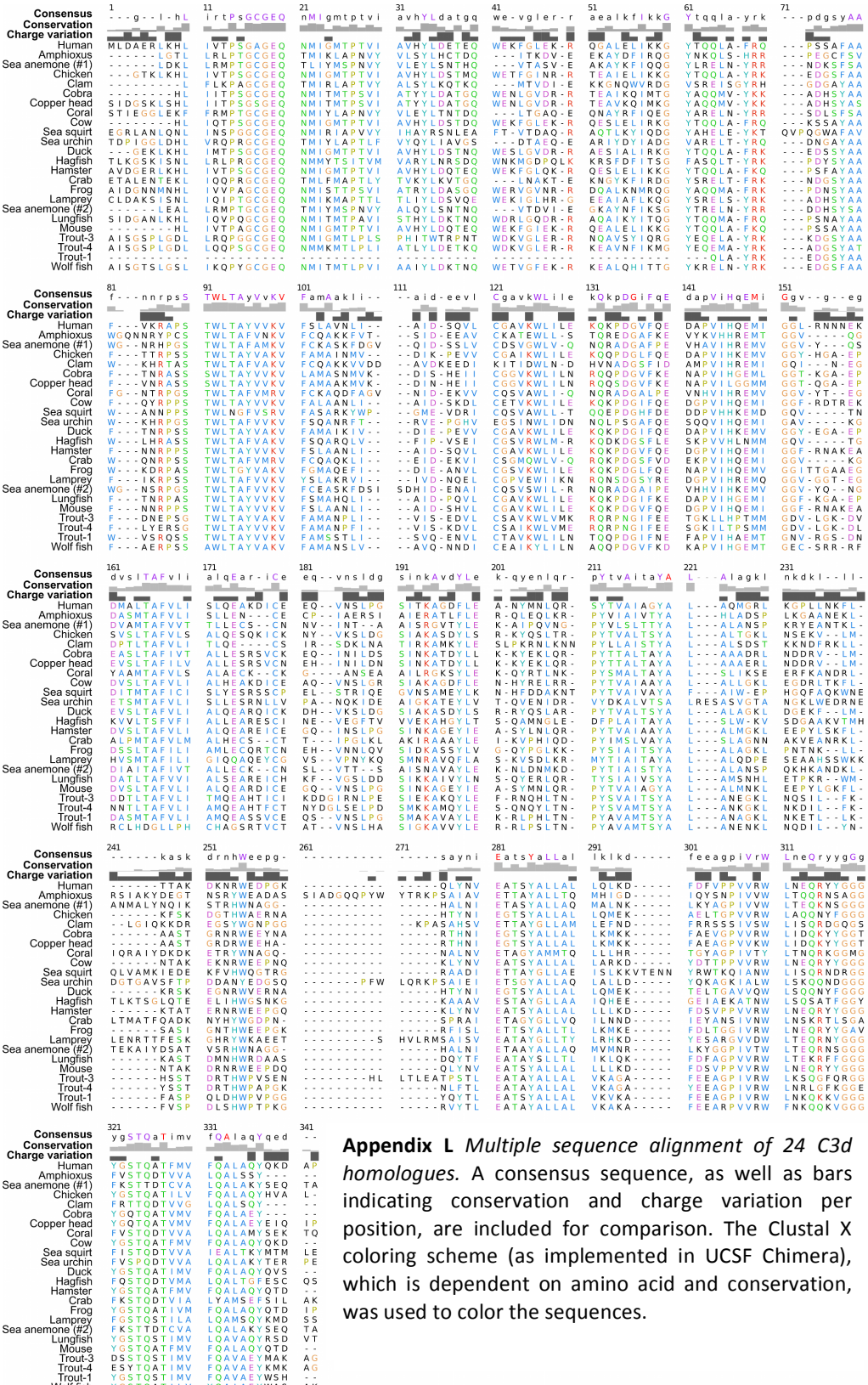
Appendix I *Spatial distributions of electrostatic potential of the non-MG C3b domains (PDB Code 2I07).* The amino acid ranges for the non-MG C3b domains are: alpha-Nt (730-745), C3d (or TED, 970-1264), CUB (912-969 & 1265-1330), LNK (578-642), and C345C (1475-1641). The net charge of each domain is marked in the figure. Electrostatic potentials were calculated using ionic strengths corresponding to 0 mM counterion concentration. The color code for isopotential contours is blue for positive and red for negative electrostatic potential. Isopotential contours are plotted at ±3 kBT/e. Ribbon representations of each domain are also provided to illustrate the orientations used for the isopotential contours.



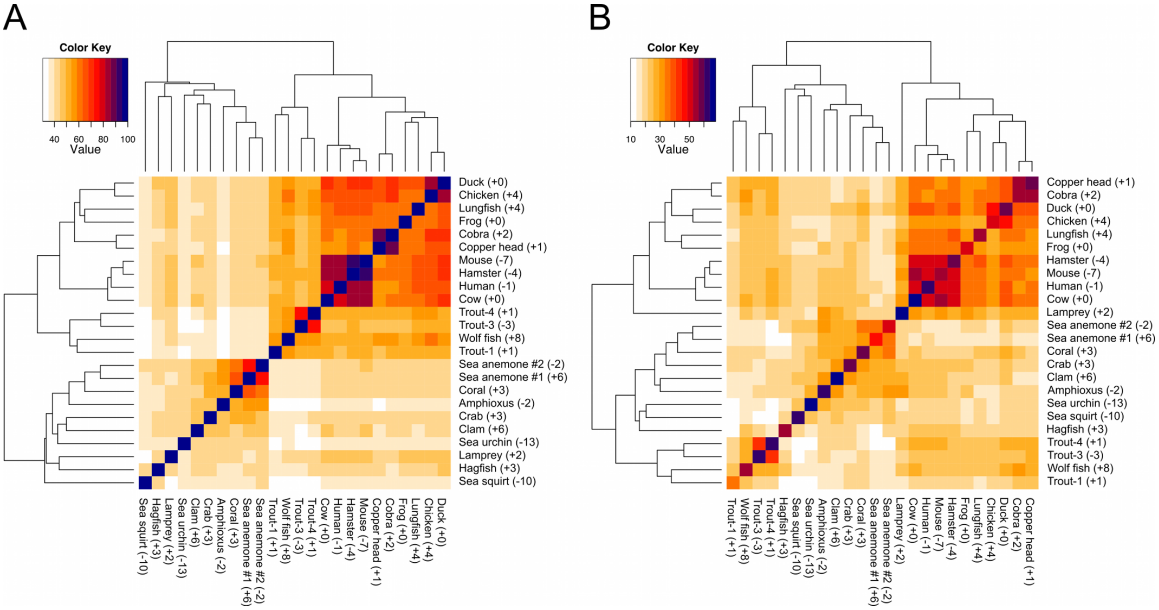
Appendix J *Complement regulators and receptors bind the acidic "hot-spot" of complement C3d*. Cumulative electrostatic similarity distribution for 24 homologues projected onto the surface of human C3d (same as Figure 2B) [blue – green – red; low to high similarity] with host ligands superimposed. Ribbon representations are used for the host ligands: FH 4 – white (PDB: 2WII); FH 19/20 – gray (PDB: 2XQW); CR2 – black (PDB: 3OED). Two rotations of C3d are provided to show the two electrostatic "hot-spots": (A) CR2-face and (B) thioester-face.



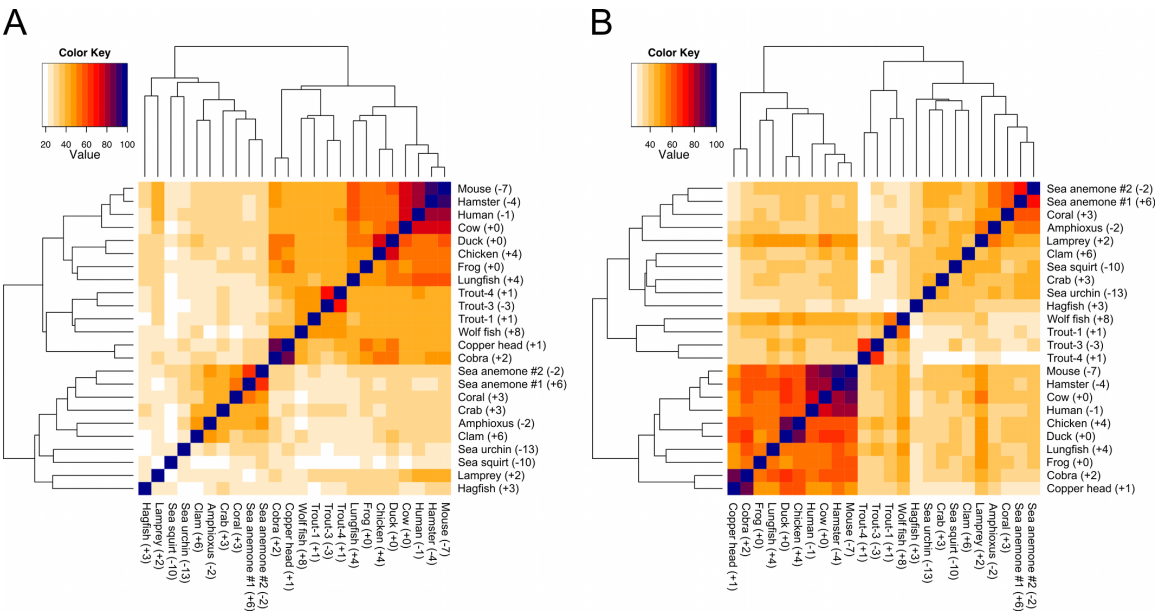
Appendix K *Pathogenic inhibitors of the complement system target the conserved electrostatic "hot-spots" of complement C3d.* Cumulative electrostatic similarity distribution for 24 homologues projected onto the surface of human C3d (same as Figure 2B) [blue – green – red; low to high similarity] with S. aureus virulence factors superimposed. Ribbon representations are used for the S. aureus virulence factors: Ehp – white (PDB: 2NOJ); Efb-C – gray (PDB: 2GOX); Sbi – black (PDB: 2WY7). Two rotations of C3d are provided to show the two electrostatic "hot-spots": (A) CR2-face and (B) thioester-face.



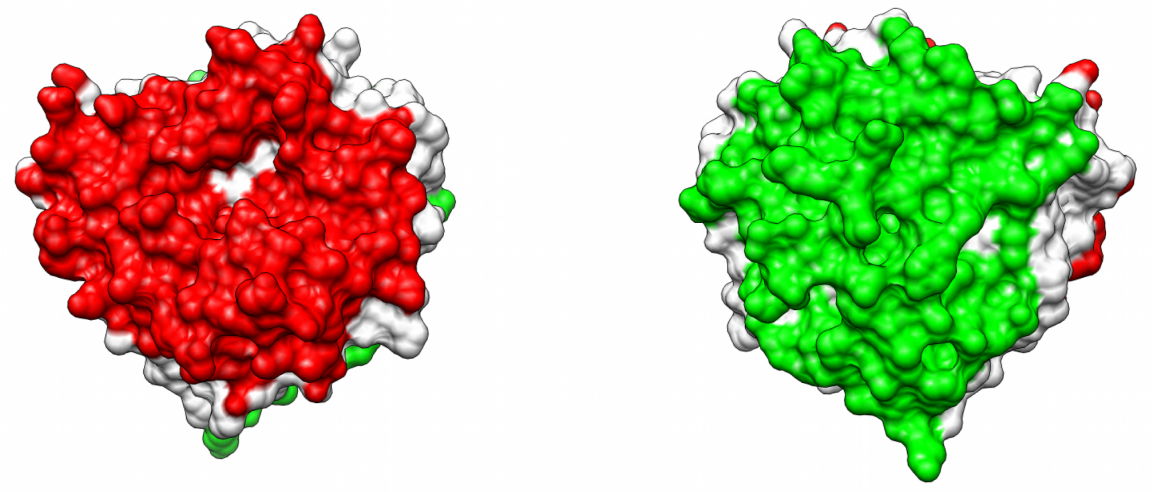
Appendix L *Multiple sequence alignment of 24 C3d homologues.* A consensus sequence, as well as bars indicating conservation and charge variation per position, are included for comparison. The Clustal X coloring scheme (as implemented in UCSF Chimera), which is dependent on amino acid and conservation, was used to color the sequences.



Appendix M *Sequence and charge clustering of C3d homologues using whole sequences.* Dendrograms with distance matrix heatmaps illustrate sequence clustering based on: (A) percent identity and (B) number of positions with the same charge. Net charge of each sequence is provided in parentheses.



Appendix N *Sequence similarity clustering for the two sides of C3d.* Dendrograms with distance matrix heatmaps illustrate clustering of the 24 C3d homologues based on percent identities within the two functional regions as defined by Appendix O: (A) CR2-face and (B) thioester-face. Net charge of each homologue is provided in parentheses.



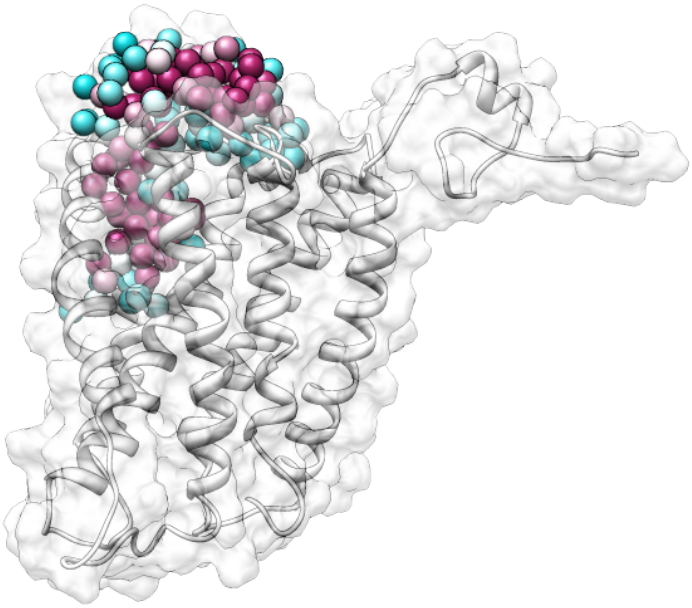
Appendix O *Illustration of the two functional regions used for sequence analysis of complement C3d*. Residues were assigned to the two regions according to their x-coordinates. The CR2-face (colored in red) includes residues that contain at least one atom with an x-coordinate ≤ (mean(x) – 5 Å), while thioester face includes residues that contain at least one atom with an x-coordinate > (mean(x) + 5 Å).

Appendix P *List of complement C3d homologues with UniProt accession ID, human C3d percent identity, and net charge.*

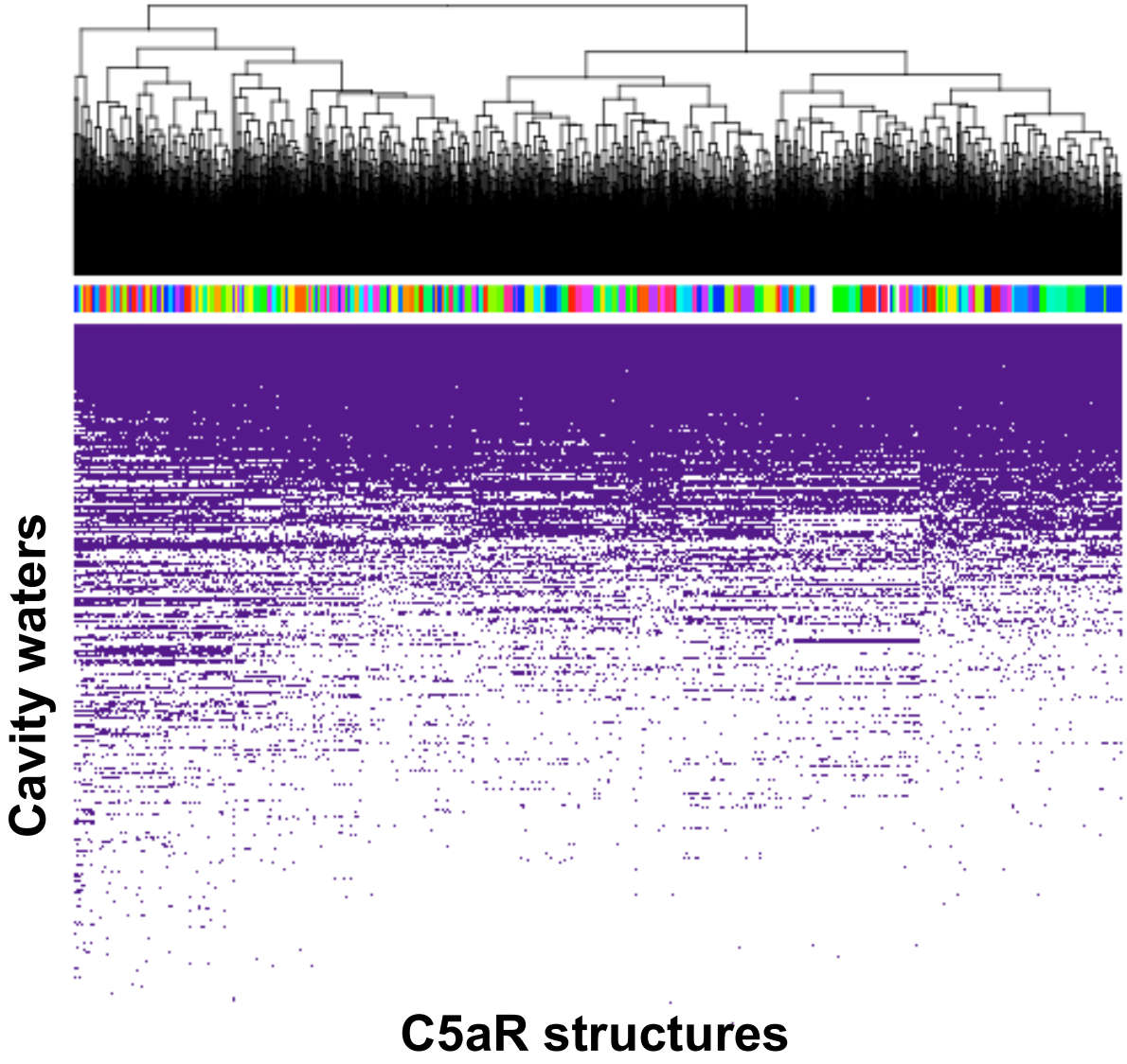
Name	Scientific name	UniProt ID	Percent identity	Net charge
Human	*Homo sapiens*	P01024	-	-1
Hamster	*Mesocricetus auratus*	Q9Z115	85	-4
Mouse	*Mus musculus*	P01027	85	-7
Cow	*Bos taurus*	Q693V9	80	+0
Chicken	*Gallus gallus*	Q2MV09	65	+4
Duck	*Anas platyrhynchos*	B5AG23	65	+0
Cobra	*Naja naja*	Q01833	58	+2
Copper head	*Austrelaps superbus*	A0RZG2	57	+1
Trout-1	*Oncorhynchus mykiss*	P98093	49	+1
Trout-3	*Oncorhynchus mykiss*	Q98977	48	-3
Trout-4	*Oncorhynchus mykiss*	Q9DDV9	51	+1
Wolf fish	*Anarhichas minor*	Q98TS6	52	+8
Lungfish	*Protopterus aethiopicus*	Q9W6G1	64	+4
Frog	*Xenopus laevis*	Q91588	60	+0
Lamprey	*Lampetra japonica*	Q5SGB5	47	+2
Hagfish	*Eptatretus burgeri*	P98094	40	+3
Sea squirt	*Halocynthia roretzi*	O97019	37	-10
Amphioxus	*Branchiostoma belcheri*	Q969A4	37	-2
Sea anemone (#2)	*Haliplanella lineata*	D4Q9Z9	43	-2
Sea anemone (#1)	*Nematostella vectensis*	B9X079	41	+6
Coral	*Acropora millepora*	A0T397	42	+3
Sea urchin	*Strongylocentrotus purpuratus*	O44344	39	-13
Crab	*Tachypleus tridentatus*	B6ZH52	42	+6
Clam	*Venerupis decussatus*	C0JPJ2	42	+6

Appendix Q *Comparison of predicted interactions between C3d homologues and human CR2.* Homology models of the hypothetical complexes formed between 24 homologues of C3d and human CR2 were generated with Modeller, using the complex of human C3d-CR2 (PDB Code: 3OED) as a template. The solvation free energy of association ($\Delta\Delta G_{solv}$) for each complex was calculated using PDB2PQR/APBS according to the thermodynamic cycle of Figure 2-1B. The species are ranked according to $\Delta\Delta G_{solv}$.

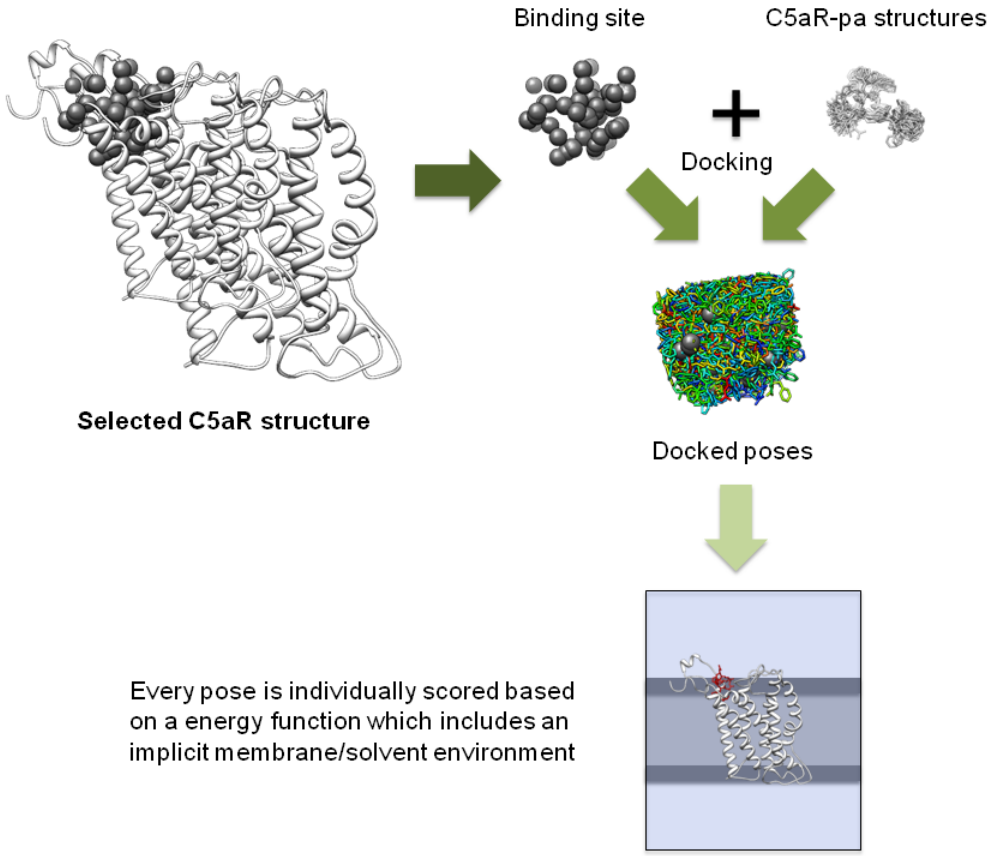
Name	**$\Delta\Delta G_{solv}$ (kJ/mol)**
Sea urchin	-280.4
Mouse	-256.0
Sea squirt	-204.7
Hamster	-191.3
Human	-167.6
Amphioxus	-147.3
Cow	-134.0
Trout-3	-98.7
Sea anemone (#2)	-96.9
Coral	-88.3
Frog	-87.0
Copper head	-84.0
Trout-1	-80.2
Duck	-60.8
Hagfish	-57.2
Lamprey	-56.0
Cobra	-50.8
Trout-4	-43.0
Chicken	-30.2
Lungfish	-25.6
Clam	-4.3
Sea anemone (#1)	7.2
Crab	12.0
Wolf fish	52.0



Appendix R *Water representation of C5aR cavity shape.* The cavity waters were selected according to the procedure described in section 4.2.2. Water molecules are presented using only the oxygen atom, and the molecule color is indicative of the occupancy of the molecule over the course of the trajectory (maroon is high occupancy and cyan is low occupancy).



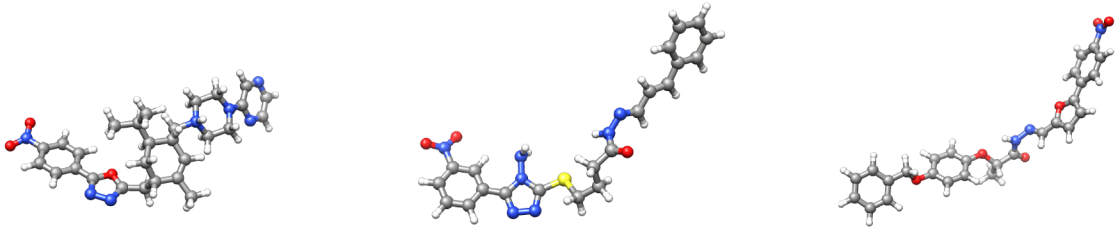
Appendix S *Clustering of C5aR structures based on cavity shape similarity.* The cavity shapes were characterized based on binary water finger prints as described in section 4.2.2. In the heatmap above, each vertical column represents the binary water fingerprint of a single C5aR cavity and each row represents a specific cavity water; here purple indicates that a water is present in a C5aR cavity structure while white indicates that it is absent. The columns have been rearranged according to the dendrogram (top), which was produced using hierarchical clustering as described in section 4.2.2. The horizontal color bar below the dendrogram is provided to visualize the clusters produced during tree cutting.



Appendix T *Schematic of docking procedure used for generating the C5aR:PMX-53 complex.* This procedure corresponds to the methods described in section 4.2.2.

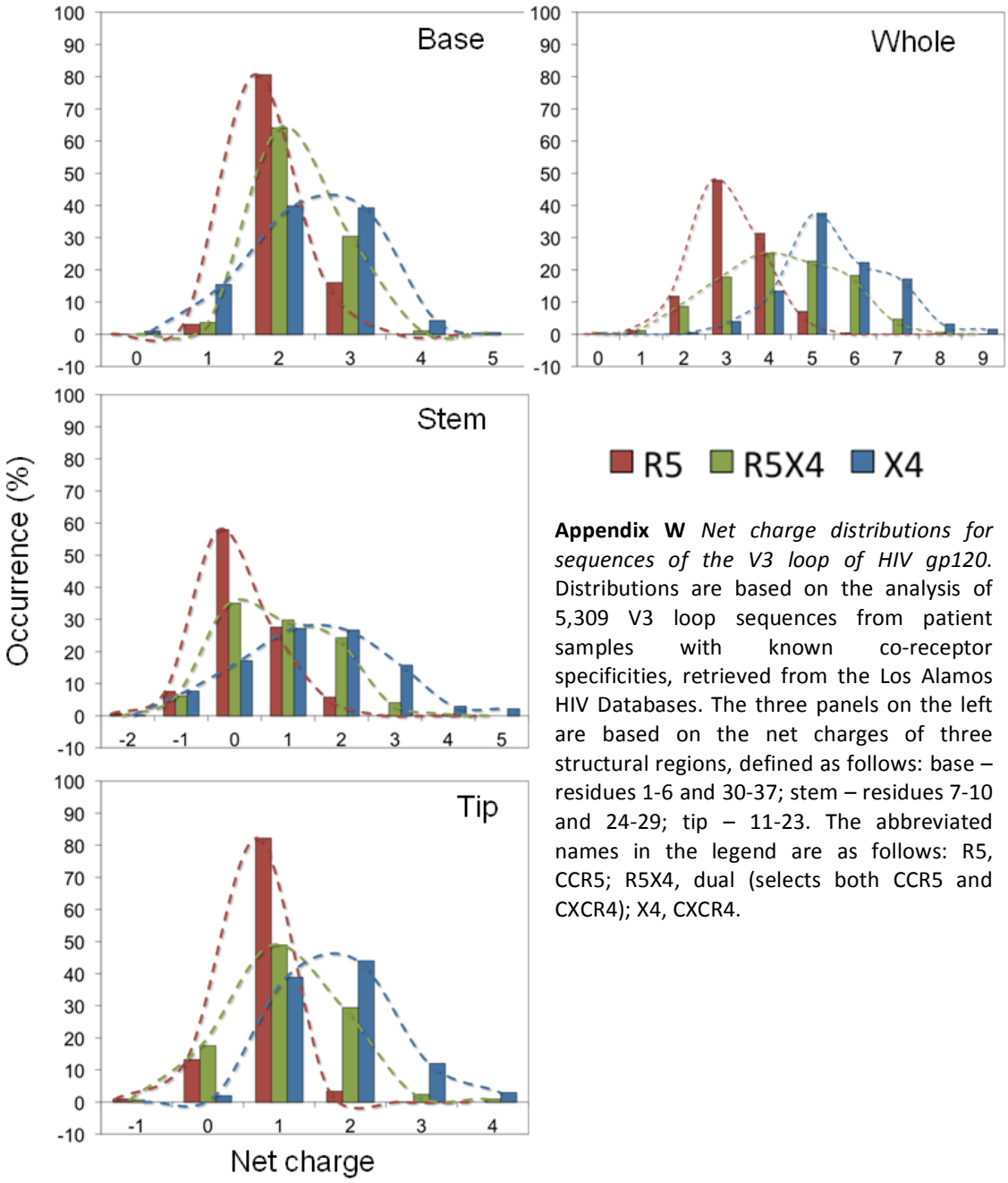
Appendix U *List of top 20 ZincPharmer compounds predicted for C5aR based on a Amber Score ranking.* Initial screening was performed by using ZincPharmer to identify compounds similar to PMX-53 based on pose 22 (overall lowest energy conformer) of the C5aR:PMX-53 model. Approximately 300 compounds were identified and score using the Amber Score function of DOCK6. Since there were only 20 compounds with a favorable (negative) Amber Score, no additional classification or clustering was performed.

#	ZINC ID	Amber Score
1	ZINC67890466	-135.6
2	ZINC58989853	-62.8
3	ZINC20311483	-60.2
4	ZINC09261828	-58.2
5	ZINC58989860	-56.5
6	ZINC28711445	-47.5
7	ZINC64228391	-44.9
8	ZINC20974239	-43.5
9	ZINC67890464	-35.3
10	ZINC03217449	-34.2
11	ZINC67879092	-29.4
12	ZINC03024412	-29.1
13	ZINC67879068	-26.8
14	ZINC08441234	-26.7
15	ZINC19855996	-19.5
16	ZINC59500310	-12.3
17	ZINC13510865	-10.7
18	ZINC00725906	-8.6
19	ZINC40917971	-7.9
20	ZINC67879081	-6.2



ZINC67890466 ZINC58989853 ZINC20311483

Appendix V *Molecular representations of top 3 predicted C5aR ligands based on pharmacophore search of pose 22.* Compounds are presented as ball and stick models with atoms being colored by element type.



Appendix W *Net charge distributions for sequences of the V3 loop of HIV gp120.* Distributions are based on the analysis of 5,309 V3 loop sequences from patient samples with known co-receptor specificities, retrieved from the Los Alamos HIV Databases. The three panels on the left are based on the net charges of three structural regions, defined as follows: base – residues 1-6 and 30-37; stem – residues 7-10 and 24-29; tip – 11-23. The abbreviated names in the legend are as follows: R5, CCR5; R5X4, dual (selects both CCR5 and CXCR4); X4, CXCR4.

Appendix X *CCR5 specific V3 loop sequence composition per position.*

Pos.	Amino acids	%

	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*M*	*N*	*L*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	*Y*	-	
1	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**
2	1	0	0	1	0	0	0	13	0	1	0	0	0	0	0	1	82	2	0	0	0	**82**
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	**100**
4	0	0	0	0	1	0	0	0	0	0	0	1	97	0	0	0	0	0	0	1	0	**97**
5	0	0	0	0	1	10	1	0	0	1	72	0	0	0	0	13	0	0	0	1	0	**72**
6	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	**99**
7	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	**99**
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	**99**
9	0	0	0	0	0	0	0	1	1	0	0	0	0	0	95	2	0	1	0	0	0	**95**
10	0	0	0	1	0	0	0	0	71	0	0	0	0	6	14	0	8	0	0	0	0	**71**
11	0	0	0	1	0	19	0	0	0	0	0	0	0	0	1	79	0	0	0	0	0	**79**
12	0	0	0	0	0	0	0	82	0	4	0	0	0	0	0	0	2	12	0	0	0	**82**
13	0	0	0	0	0	1	48	0	0	0	8	0	10	0	20	3	7	0	1	0	0	**48**
14	0	0	0	0	1	0	0	77	0	13	0	8	0	0	0	0	0	1	0	0	0	**77**
15	3	0	0	0	0	96	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	**96**
16	0	0	0	0	0	0	0	0	0	0	0	1	96	0	0	0	0	0	1	0	0	**96**
17	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**99**
18	1	0	0	0	0	2	0	0	8	0	0	0	0	29	56	4	0	0	0	0	0	**56**
19	77	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	13	7	0	0	0	**77**
20	0	0	0	0	79	0	0	3	0	2	0	7	0	0	0	0	0	1	5	3	0	**79**
21	0	0	1	0	7	0	1	0	0	0	0	0	0	0	0	1	0	0	0	91	0	**91**
22	62	0	0	0	0	0	0	0	1	1	0	0	0	0	4	0	31	0	0	0	0	**62**
23	2	0	0	1	0	2	0	0	0	0	1	0	0	0	0	0	93	0	0	0	0	**93**
24	0	0	5	6	0	81	0	0	2	0	3	0	0	0	0	0	1	0	0	0	0	**81**
25	5	0	38	29	0	4	0	11	1	0	1	0	0	7	0	1	1	2	0	0	0	**38**
26	0	0	0	0	0	0	0	93	0	0	0	0	0	0	0	0	0	6	0	0	0	**93**
27	0	0	0	0	0	12	0	81	0	0	0	0	0	0	0	0	3	2	0	0	0	**81**
28	0	0	10	0	0	87	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	**87**
29	0	0	67	0	0	0	0	12	1	0	18	0	0	0	1	0	0	0	0	0	0	**67**
30	0	0	0	0	0	0	0	86	0	0	0	0	0	0	13	0	0	0	0	0	0	**86**
31	1	0	0	0	0	0	0	0	3	0	0	0	0	9	87	0	0	0	0	0	0	**87**
32	12	0	0	1	0	0	1	0	12	0	0	0	0	70	2	1	0	0	0	0	0	**70**
33	87	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	4	0	**87**
34	0	12	0	0	1	0	66	0	0	0	0	0	0	1	1	1	0	0	0	17	0	**66**
35	0	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	**87**
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	**99**
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	**99**

Appendix Y *CCR5/CXCR4 specific V3 loop sequence composition per position.*

Pos.	Amino acids																				%
	A	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***K***	***M***	***N***	***L***	***P***	***Q***	***R***	***S***	***T***	***V***	***W***	***Y***	-	
1	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**
2	1	0	0	0	0	0	0	12	0	2	0	0	0	0	0	5	80	0	0	0	0	**80**
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	**100**
4	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	**100**
5	0	0	1	0	0	24	0	0	0	0	52	0	0	0	0	9	0	0	0	14	0	**52**
6	0	0	1	1	0	0	0	0	1	0	92	0	0	0	1	2	2	0	0	1	0	**92**
7	0	0	0	0	0	0	0	1	5	0	88	0	0	0	0	2	0	0	0	4	0	**88**
8	0	0	0	1	0	0	0	4	2	0	0	0	0	0	2	0	90	0	0	0	0	**90**
9	0	0	0	0	0	0	0	4	1	0	0	0	0	1	92	0	0	0	0	0	0	**92**
10	0	0	0	0	0	0	0	0	65	0	0	0	0	12	15	0	6	0	0	0	0	**65**
11	0	0	1	0	0	23	1	0	0	0	0	0	0	0	22	50	0	0	0	0	0	**50**
12	0	0	0	0	2	0	0	48	0	1	0	1	0	0	1	0	11	34	1	0	0	**48**
13	0	0	0	0	0	1	37	1	0	0	2	0	3	0	32	9	9	0	0	3	0	**37**
14	1	0	0	0	0	0	1	63	0	10	0	16	0	0	0	0	4	4	0	0	0	**63**
15	1	0	0	0	0	96	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	**96**
16	1	0	0	0	0	2	1	2	0	0	0	2	88	3	0	0	0	0	0	0	0	**88**
17	0	0	0	0	0	95	0	0	0	0	0	0	1	0	2	0	0	1	0	0	0	**95**
18	0	0	0	0	0	3	1	0	4	0	0	0	0	41	47	2	0	0	0	1	0	**47**
19	53	0	0	0	1	3	0	0	0	0	0	0	1	0	2	2	20	15	0	2	0	**53**
20	2	0	0	0	54	1	0	4	0	0	0	13	0	2	0	0	0	5	12	6	0	**54**
21	1	0	0	0	12	0	6	0	0	0	0	0	0	0	1	1	1	1	1	75	0	**75**
22	33	0	0	0	2	0	1	0	2	0	0	0	0	0	3	0	51	6	0	2	0	**51**
23	4	0	0	0	1	1	2	0	1	0	5	0	0	2	2	2	78	0	0	3	0	**78**
24	3	0	12	6	0	52	0	0	10	0	3	1	0	0	6	0	5	0	1	0	0	**52**
25	8	0	18	13	0	4	0	15	5	0	1	0	0	12	11	1	7	5	0	0	0	**18**
26	0	0	0	0	0	3	0	88	1	0	1	0	0	1	1	0	1	3	0	0	0	**88**
27	0	0	1	0	0	23	0	59	2	0	1	1	0	0	0	0	5	7	0	0	0	**59**
28	0	0	21	0	0	69	0	4	1	0	3	0	0	1	0	0	0	1	0	0	0	**69**
29	0	0	60	0	0	3	0	23	0	0	9	0	1	0	1	0	0	2	0	0	0	**60**
30	0	0	0	0	0	3	0	67	3	0	2	0	0	1	22	0	1	1	0	0	0	**67**
31	1	0	1	0	0	0	1	2	3	0	1	0	0	20	70	0	0	0	0	0	0	**70**
32	24	0	0	0	0	0	1	3	22	0	0	0	0	42	7	0	0	1	0	0	0	**42**
33	69	1	0	0	0	0	17	0	0	0	0	0	1	1	5	0	0	0	0	6	0	**69**
34	1	24	0	0	0	0	50	0	0	0	0	0	1	2	0	0	0	0	0	20	1	**50**
35	3	70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	**70**
36	0	2	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	95	**95**
37	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	**96**

Appendix Z *CXCR4 specific V3 loop sequence composition per position.*

Pos.	Amino acids																					%
	A	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***K***	***M***	***N***	***L***	***P***	***Q***	***R***	***S***	***T***	***V***	***W***	***Y***	**-**	
1	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**
2	4	0	0	0	0	0	0	4	2	1	0	0	0	0	0	9	79	0	0	0	0	**79**
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	**100**
4	0	0	0	0	0	0	1	0	0	0	0	0	99	0	0	0	0	0	0	0	0	**99**
5	0	0	3	0	9	9	2	0	0	0	39	0	0	0	0	6	0	0	0	32	0	**39**
6	0	0	1	3	0	0	0	1	6	0	75	0	0	0	1	0	8	0	0	5	0	**75**
7	1	0	0	0	0	0	3	1	13	0	63	0	2	0	0	1	12	0	0	3	0	**63**
8	2	0	0	1	0	0	0	18	8	2	0	1	0	0	2	0	64	1	0	0	0	**64**
9	0	0	0	1	0	0	0	7	4	0	1	0	0	0	84	3	0	0	0	0	0	**84**
10	0	0	0	0	0	0	0	1	40	1	1	0	0	9	28	1	16	0	0	1	0	**40**
11	0	0	0	0	0	8	3	1	2	2	1	0	0	2	34	47	0	0	0	0	0	**47**
12	0	0	0	0	3	0	0	34	1	2	0	4	0	0	4	0	28	22	0	0	0	**34**
13	0	0	0	0	0	9	22	2	0	0	1	1	5	0	31	11	12	0	0	6	0	**31**
14	0	0	0	0	1	2	2	71	1	6	0	9	1	0	0	1	3	2	0	0	0	**71**
15	0	0	0	0	0	90	0	5	0	0	0	0	2	2	0	0	0	0	0	0	0	**90**
16	0	0	0	0	0	7	1	8	0	0	0	12	63	4	2	0	2	1	0	0	0	**63**
17	0	0	0	0	0	89	1	0	0	0	0	1	3	1	3	2	0	0	0	0	0	**89**
18	0	0	0	0	0	6	1	0	3	0	0	0	3	19	66	0	0	2	0	0	0	**66**
19	51	0	0	0	2	9	0	0	0	0	0	0	0	3	5	0	1	26	0	1	0	**51**
20	3	0	0	0	39	0	1	0	0	1	0	6	1	6	2	8	0	6	8	18	0	**39**
21	4	0	0	0	5	0	2	0	0	0	0	2	0	0	3	5	6	0	0	74	0	**74**
22	14	0	0	0	9	0	0	14	12	1	0	0	0	0	5	1	35	5	0	4	0	**35**
23	10	0	0	1	0	3	1	0	1	4	1	1	0	0	1	0	69	2	0	7	0	**69**
24	6	0	7	6	0	28	0	0	25	0	1	0	0	2	16	0	7	1	0	0	0	**28**
25	2	0	12	10	0	4	0	32	5	1	4	2	0	14	4	2	5	3	0	0	0	**32**
26	0	0	1	0	0	17	0	70	5	0	0	0	0	0	1	0	2	4	0	0	0	**70**
27	0	0	3	3	0	29	0	26	5	1	4	0	0	2	4	0	13	9	0	0	0	**29**
28	0	0	26	0	0	49	0	14	1	0	3	0	0	0	1	0	1	3	0	2	0	**49**
29	0	0	39	0	0	5	0	25	3	0	4	0	1	0	6	1	9	2	0	2	0	**39**
30	0	0	1	0	0	7	0	46	2	1	3	1	0	7	30	0	1	0	0	0	0	**46**
31	7	0	5	0	0	2	0	3	9	2	2	0	2	20	48	0	0	0	0	0	0	**48**
32	31	0	0	0	0	0	1	7	17	0	0	0	0	17	20	0	0	1	0	5	0	**31**
33	48	7	0	0	2	0	20	0	0	0	0	0	1	5	8	0	0	0	0	8	0	**48**
34	5	31	0	0	0	0	22	0	1	0	0	0	0	7	0	0	0	0	0	26	7	**31**
35	8	49	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	2	38	**49**
36	0	5	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	1	88	**88**
37	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	**92**

Appendix AA *UCSF Chimera Introduction*

Additional Help
All of the commands and functions discussed in this brief tutorial are explained in much more detail in the UCSF Chimera User Guide:
http://www.cgl.ucsf.edu/chimera/1.4/docs/UsersGuide/index.html

For complete list of all command line functions and options please see the following link:
http://www.cgl.ucsf.edu/chimera/1.4/docs/UsersGuide/framecommand.html

Getting Started
All of the examples herein will be based on structures of the SUMO1:SENP2 (PDB Code: 1tgz) and SUMO2:SENP2 (PDB Code: 2io0) complexes. Both files can be imported directly into Chimera by using the drop-down menu and going to File > Fetch by ID… (A similar notation will be used throughout when referring to the drop down menus.).

After opening the PDB files, I recommend removing all solvent or heteratoms which can be achieved using the following command:

del ~protein #Which deletes everything that's not (~) the protein.

Superposition
The first step for this comparison is to superimpose the two models (A "model" for Chimera refers to a file, e.g. one PDB file, and not to whether or not the structure was experimentally determined.). The easiest method is MatchMaker, which can be found under Tools > Structure Comparison > MatchMaker. The MatchMaker function first performs a sequence alignment to determine what portions of the two models are similar, prior to fitting the atomic coordinates. MatchMaker can superimpose two structures even if the there are not the same number of atoms in each structure, where other methods require that the same number of atoms be given.

One measure of the fit or similarity between two structures is the root mean squared deviation, or RMSD, which can be calculated for all or a subset of atoms. In Chimera the command line function rmsd will perform this calculation when given two atom specifications (The format for atom specification is **#***model***:***residues***.***chain***@***atom*. See below for examples, also see Quick Reference Guide). The rmsd function requires that the same number of atoms be given by each specification, so you will need to examine the sequence alignment from MatchMaker to determine which residues are conserved.

rmsd #0:366-589 #1:366-589
rmsd #0:366-589@ca,n,c,o #1:366-589@ca,n,c,o
rmsd #0:366-589@ca #1:366-589@ca

If it is desired that only a specific subset of the structure be used for the superposition, which cannot be achieved using MatchMaker, the match function provides such functionality. The

match function performs a least squares fit based on the specified atom coordinates, and automatically returns an RMSD.

match #0:366-589.A@ca #1:366-589.A@ca
match #0:366-589.A@ca,n,c,o #1:366-589.A@ca,n,c,o
match #0:366-589.A #1:366-589.A

Representations

All representations (Atom/Bond, Ribbon, Surface) can be found under the Actions drop-down menu, with each having its own options. If representations are applied with no atoms selected, the representation will be applied to all models. Below you will find a few examples of the commands to achieve these different representations using the command line. Multiple representation types can be used simultaneously, and atom specifiers can be used to apply to representations to only portions of the model.

Atom/Bond:

show	(All atoms and bonds will be shown)
rep bs	(Change atom/bond represenation to ball and stick, also can use stick and wire, or alternatively you can use sphere for a van der Waals representation.)
show #0@ca	(C-alpha trace. Here this should show only atoms in model #0.)
show #0@n,ca,c,o	(Backbone)
~show	(Hide atom/bonds)

Ribbon:

ribbon	(Show all ribbons)
ribrep edged	(Change representation. Other options are flat or smooth, which is default.)
~ribbon	(Hide all ribbons)

Surface:

surface #0	(Surface model #0.)
surfrep dot	(Surfaces can be represented in dot, mesh, or solid, where solid is the default.)
surftransparency 70	(Change surface transparency. Takes a percent as an input, where 0 is opaque and 100 is fully transparent.)

Color

The color of specific models or representations can be changed by using the Actions > Color menu. Selections can be made using the command line or the mouse prior to applying a color, if no selection is made it is applied to all models.

color gray #0:.A	(Color chain A of model 0 gray.)
color yellow hydrophobic	
color red negative	

color blue positive
rainbow

Analyzing Complexes

It is often desired to analyze which interactions or contacts contribute to the stability of proteins or protein complexes. One approach to achieve such an analysis is to use zone selections, in which one can select residues (zr) or atoms (zr) according to its proximity to other atoms or residues.

show #0:.A zr < 5 & #0:.B	(Show residues with at least one atom within 5 Å of chain A of model #0 and is within chain B of model #0. The second portion of this selection, "& #0:.B", is used to exclude other models and chain A.)
sel #0:.A zr < 5 & #0:.B	(Alternatively the same arguments can be used to select the residues, which can subsequently be written to file at Actions > Write List …)

For charge-charge interactions it is typically desired to find specific atoms, those that are possibly charged, with in the vicinity of a specific charged group. Such a comparison can be made using similar syntax to above, as follows:

sel #0:63.B@CZ za < 5 & :glu@cd & #0:.A

Hydrogen bonds can be determined by: Tools > Structure Analysis > FindHBonds

Delta SASA of binding for a complex can be determined as follows (requires newer version of Chimera, 1.4.1 or higher):

measure buriedArea #0:.A #0:.B
measure buriedArea #1:.A #1:.B

Saving Images

When saving images it is typically better to have a white background to save ink etc., and this can be set through the Action > Color drop down menu, or with the following command: set bg_color white

Chimera is capable of producing very high-resolution images for publication. This can be achieved by File > Save Image…, and following the dialog boxes. The copy command can be used as shown in the following example: copy file /Users/ckieslich/test.png png

! Extra note: It is also possible to save edited structures as PDBs and to save Sessions where models and representations are save for later use (File > …)

Appendix BB *Dock 6 Outline*

I) Prepare Structures

1. For the receptor: Use DockPrep function in Chimera
    - Deletes solvent
    - Mutates MSE to MET
    - Adds H
        - Possible Methods
            - Steric only
            - Consider H-bond network
        - Histidine Protonation
            - Residue name based
            - Individually chosen
            - Unspecified (determined by method)
    - Adds Charge
    - (Newer Version: Can replace incomplete side chains using a library of rotomers. If this option is not used, it is necessary to mutate residues with incomplete side chains to Ala or Gly)
    - After completing above steps, need to save two files: 1) mol2 file and 2) PDB without hydrogens (needed to create molecular surface later)
2. For the ligand:
    - Protein ligand: Repeat above steps but no PDB file needed
    - Small molecule ligand: 2 options (Repeat above (No PDB) or submit to ZINC)

II) Generate Spheres

1. Generate molecular surface
    - Based on PDB file of receptor
    - DMS (included with DOCK)
        - Uses a ball with the radius of water (like APBS)
2. Generate spheres surrounding receptor
    - Spheres generated for entire surface
        - Spheres are produced with sizes on a given range (default: 1.4 - 4Å)
        - Produced spheres are tangent to the molecular surface
    - Filtered, keeping only the largest sphere associated with each surface atom
    - Filtered set of spheres are then clustered using a single linkage algorithm
3. Select a subset of spheres
    - Can choose the largest cluster (Typically represents binding site)
    - Can select spheres within a given radius of a desired location
    - Add spheres manually, if a known binding site is poorly represented

III) Grid Generation

1. Defines area used for calculation of docking score
2. Must answer a series of questions.

IV) Docking

1. Rigid Ligand
    - Typically used to explore the matching and minimization algorithms
    - Also useful if searching a conformationally expanded database
    - In conceptual terms, the algorithm matches the centers of the ligand heavy atom to the centers of the receptor site spheres. The algorithm follows the steps below:
        - Generate node
        - Label as match if atom and sphere edges are equivalent
        - Extend match by adding more nodes
        - Exhaustively generate set of non-degenerate matches
        - Use matches to create transformation matrices to move the entire molecule
    - *node* = pairing of one heavy atom and one sphere center
    - *edge length* = Euclidean distance between atom or sphere centers
    - Once an orientation has been generated, the interaction between the ligand and the receptor can be energetically optimized, in this case using a simplex minimizer (Nelder, et al. *Computer Journal* 1965). During minimization, the ligand is allowed to be flexible, but the receptor remains rigid. The final score in the output file is the best pose generated from the orienting and minimization procedure.
2. Flexible Ligand
    - Allows ligand to structurally rearrange in response to the receptor
    - All bonds within molecular rings are treated as rigid.
    - To treat such phenomena as sugar puckering and chair-boat hexane conformations, the user needs to supply each ring conformation as a separate input molecule.
    - If the molecule does not have a ring, the largest rigid segment is specified as the anchor.
    - Additional bonds may be specified as rigid by the user.
    - Next is the identification of flexible bonds
    - The location of each flexible bond is used to partition the molecule into rigid segments, where a segment is the largest local set of atoms that contains only non-flexible bonds.

V) Amber Scoring

When a user calls for Amber score, the program performs minimization, MD simulation, and more minimization on the individual ligand, the individual receptor, and the ligand-receptor complex, and calculates the score as follows:

$$\mathbf{E_{binding} = E_{complex} - (E_{receptor} + E_{ligand})}$$

where E is obtained from:
$E = E_{MM} + (E_p\text{-sol} + E_{np}\text{-sol})$
$E_{MM} = E_{bad} + E_{vdW} + E_{es}$
E_p-sol: Electrostatic part of solvation energy using GB
E_{np}-sol: Non-polar part of solvation energy using SA

8. CIRICULUM VITAE

EDUCATION

University of California, Riverside; Riverside, California
The Marlan & Rosemary Bourns College of Engineering
PhD. Bioengineering, August 2012
Advisor: Dimitrios Morikis
Title: Design and Applications of a Computational Framework for Protein and Drug Design

Saint Louis University; St. Louis, Missouri
Parks College of Engineering, Aviation and Technology
B.S. Biomedical Engineering, May 2007

RESEARCH EXPERIENCE

University of California, Riverside; Riverside, California
Graduate Research Assistant, Fall 2007 – Present

- Dr. Dimitrios Morikis, Department of Bioengineering
- Development of computational framework for the analysis of the role of electrostatics in protein function, with applications in protein and drug design.
- Computational design of antagonists of complement immune system activation as potential therapeutics for autoimmune and cardiovascular disease.

University of Warsaw; Warsaw, Poland
NSF/DAAD CESRI Fellow, Summer 2010

- Dr. Jan Antosiewicz, Department of Biophysics
- Development of tools for the computational modeling of the pH dependent properties of proteins, including association and structural stability.

Osaka University; Osaka, Japan
NSF EAPSI Fellow, Summer 2009

- Dr. Susumu Date, Cybermedia Center
- Application of parallel and grid computing strategies for the distribution of docking simulations in virtual drug screening.

Saint Louis University; St. Louis, Missouri
Undergraduate Researcher, Summer 2006 – Spring 2007

- Dr. David Barnett, Department of Biomedical Engineering
- Finite element modeling of anode location for EMG monitoring of iliosacral screw placement

TEACHING EXPERIENCE

University of California, Riverside
Graduate Teaching Assistant, Fall 2007 – Spring 2010

- Biomolecular Engineering: Conducted hands-on tutorials of computational tools for protein modeling and graded modeling projects. [Spring 2010]
- Biophysics and Biothermodynamics: Responsible for leading discussion sections to supplement lectures and graded weekly assignments. [Fall 2007,2008, & 2009]
- Bioinstrumentation Lab: Led labs varying from EKG circuit construction/analysis to mechanical testing of chicken bone, and graded lab reports. [Spring 2008]

Saint Louis University
Undergraduate Teaching Assistant, Fall 2006

- Biocomputation: Was responsible for leading lab/discussion a section in which students were able to gain hands-on programming experience to supplement lectures.

VOLUNTEER EXPERIENCE

Science Oplympiad Coach; November 2010 – April 2012
Matthew Gage Middle School; Riverside, California

- Led two teams of 7^{th} and 8^{th} grade students in the design, construction, and analysis of trebuchets capable of throwing objects such as a tennis ball.
- Led two teams of 7^{th} and 8^{th} grade students in the design, construction, and analysis of 2-liter soda bottle compressed-air driven rockets.

PUBLICATIONS

1. **Kieslich, C.A.** and Morikis, D. (2012) The two sides of complement C3d: Evolution of electrostatics in a link between innate and adaptive immunity. *PLoS Comp. Bio.* Submitted
2. **Kieslich, C.A.**, Shin, D., López de Victoria, A., González-Rivera, G., Morikis, D. (2012) A predictive model for HIV-1 co-receptor selectivity and disease progression. *AIDS Res. Hum. Retro.* Submitted.
3. **Kieslich, C.A.***, Tamamis, P*, Woodruff, T.M., Archontis, G., Morikis, D. (2012) Quasi-dynamic pharmacophore for C5aR antagonists. *J. Phys. Chem. B* Submitted.
4. Bellows-Peterson, M.L., Fung, H.K., and Floudas, C.A., **Kieslich, C.A.**, Zhang, L, and Morikis, D, Wareham, K.J., and Monk, P.N., Hawksworth, O.A., and Woodruff, T.M. (2012) De novo peptide design with C3a receptor agonist and antagonist activities: Theoretical predictions and experimental validation. *J. Med. Chem.,* 55(9): 4159-4168
5. **Kieslich, C.A.***, Tamamis, P*, Gorham, R.D., López de Victoria, A., Sausman, N.U., Archontis, G., Morikis, D. (2011) Exploring protein-protein and protein-ligand interactions in the immune system using molecular dynamics and continuum electrostatics. *Curr. Phys. Chem. In Press.*
6. El-Assaad, A.M., **Kieslich, C.A.**, Gorham, R.D., and Morikis, D. (2011) Electrostatic exploration of the C3d-FH4 interaction using a computational alanine scan. *Mol. Immunol.*, 48(15/16): 1844-50.

7. López de Victoria, A., **Kieslich, C.A.**, Rizos, A.K., Krambovitis, E., and Morikis, D. (2011) Clustering of HIV-1 subtypes based on gp120 V3 loop electrostatic properties. *BMC Biophys.,* 5(3):1-16.
8. Gorham, R.D., **Kieslich, C.A.**, and Morikis, D. (2011) Complement Inhibition by Staphylococcus aureus: Electrostatics of C3d–EfbC and C3d–Ehp Association. *Cell. Mol. Bioeng.,* 48(1): 32-43.
9. Gorham, R.D., **Kieslich, C.A.**, Nichols, A., Sausman, N.U., Foronda, M., and Morikis, D. (2011) Calculation of free energy of protein association using Poisson-Boltzmann electrostatics: validation with experimental kinetic data. *Biopolymers,* 95(11): 746-54.
10. Hakkoymaz, H., **Kieslich, C.A.**, Gunopulos, D., and Morikis, D. (2011) Molecular similarity determination using multi-resolution analysis. *Mol. Inform.,* 30(8): 733-46.
11. **Kieslich, C.A.**, Goodman, G., Vazquez, H., López de Victoria, A., and Morikis, D. (2011) The effect of electrostatics on Factor H function and related pathologies, *J. Mol. Graph Mod.,* 29(8): 1047-55.
12. Gorham, R., **Kieslich, C. A.**, and Morikis, D. (2011) Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Ann. Biomed Eng,* 39(4): 1252-63.
13. **Kieslich, C.A.**, Gorham, R.D., and Morikis, D. (2011) Is the rigid-body assumption reasonable? Insights into the effects of dynamics on the electrostatic analysis of barnase-barstar. *J. Non-Cryst. Solids,* 357(2): 707-716.
14. **Kieslich, C.A.**, Yang, J., Gunopulos, D., and Morikis, D. (2011) Automated computational protocol for alanine scans and clustering of electrostatic potentials: application to C3d.CR2 association, *Biotech. Prog.*, 27(2): 316-325.
15. Chae, K., Gonong, B.J., Kim, S.C., **Kieslich, C.A.**, Morikis, D., Balasubramanian, S., and Lord, E.M. (2010) A multifaceted study of stigma/style cysteine-rich adhesion (SCA)-like Arabidopsis lipid transfer proteins (LTPs) suggests diversified roles for these LTPs in plant growth and reproduction, *J. Exp. Botany*, 61(15): 4277-4290.
16. Cheung, A., Yang, J., **Kieslich, C.A.**, and Morikis, D., (2010) Solvation effects in calculated electrostatic association free energies for the C3d-CR2 complex, and comparison to experimental data, *Biopolymers*, 93(6): 509-519.
17. Pyaram, K., **Kieslich, C. A.**, Yadav, V. N., Morikis, D., and Sahu, A. (2010) Influence of electrostatics on the complement regulatory functions of Kaposica, the complement inhibitor of Kaposi.s sarcoma-associated herpesvirus, *J.Immunol.,* 184(4): 1956-1967.
18. Chae, K., **Kieslich, C. A.**, Morikis, D., Kim, S.C., and Lord E. M. (2009) Arabidopsis lipid transfer protein 5 functions in pollen tube growth in the pistil transmitting tract, *Plant Cell*, 21: 3902-3914.

POSTERS

1. **Kieslich C.A.** and Morikis, D. (2012) The role of electrostatics in the function of homologous thioester containing proteins: Insights into the evolution of the complement C3d:CR2 interaction. Biophysics Society Annual Meeting, San Diego, CA.
2. **Kieslich C.A.**, López de Victoria, A., Shin, D., González-Rivera, G., and Morikis, D. (2012) Charge discriminates coreceptor selectivity for HIV-1. Biophysics Society Annual Meeting, San Diego, CA.

3. **Kieslich, C.A.**, Gorham, R.D., and Morikis, D. (2011) Molecular similarity-based computational framework for protein and drug design. NBCR Summer Institute, University of California, San Diego, CA.
4. **Kieslich C.A.**, Liao, J., and Morikis, D. (2010) Development of a high-throughput computational protocol, AESOP, and its application to the electrostatic analysis of the SUMO-1:SENP2 complex. Biophysics Society Annual Meeting, San Francisco, CA.
5. **Kieslich, C.A.**, Liao, J., and Morikis, D. (2008) Exploration of SUMO-SENP2 interactions using a high throughput computational protocol. Biomedical Engineering Society Annual Fall Meeting, St Louis, MO.
6. **Kieslich, C.A.** and Barnett, D. W. (2007) A finite element model of stimulus evoked EMG monitoring during iliosacral screw insertion. Biomedical Engineering Society Annual Fall Meeting, Hollywood, CA.

PLATFORM PRESENTATIONS

1. **Kieslich, C.A.** and Morikis, D. (2012) Computational design of complement system inhibitors targeting C5aR. Linking Tobacco Control Research and Practice for a Healthier California, Young Investigator Session, Sacramento, CA.
2. **Kieslich, C.A.**, Tamamis, P., Archontis, G., and Morikis, D. (2012) Development of a quasi-dynamic pharmacophore model for cyclic peptidic antagonists of complement receptor C5aR. ACS 243rd National Meeting, San Diego, CA.
3. **Kieslich, C.A.**, Gorham, R.D., and Morikis, D. (2011) Molecular similarity-based computational framework for protein and drug design. Best Student Poster Presentation NBCR Summer Institute, University of California, San Diego, CA.
4. **Kieslich, C.A.**, Gorham, R.D., Gasik, Z., Antosiewicz, J., and Morikis, D. (2011) AESOP: Analysis of Electrostatic Similarities of Proteins. 12th Annual Systemwide Bioengineering Symposium, Bioengineering Institute of California, Santa Barbara, CA.
5. **Kieslich, C.A.**, Gorham, R.D., Nichols, A., Sausman, N.U., Foronda, M., and Morikis, D. (2011) Electrostatic clustering and free energy calculations for evaluation of protein-protein complexes. ACS 241st National Meeting, Anaheim, CA.
6. **Kieslich C.A.**, Liao, J., and Morikis, D. (2008) Exploration of SUMO-SENP2 interactions using a high throughput computational protocol. 9th Annual Systemwide Bioengineering Symposium, Bioengineering Institute of California, Riverside, CA.

AWARDS/HONORS

- Co-Chair Bioengineering Graduate Student Association; 2009 – 2011
- Best Participant Poster at 2011 NBCR Summer Institute, University of California, San Diego
- 2011 NBCR Summer Institute Scholarship, University of California, San Diego
- UCOP Tobacco Related Disease Research Program Dissertation Fellow (2010 – 2012)
- NSF/DAAD Central Europe Summer Research Institute Research Fellow, Poland 2010
- TechHorizons (UCR) Student Poster Competition 2009, Second Place
- NSF East Asia and the Pacific Summer Institute Research Fellow, Japan 2009
- Outstanding Teaching Assistant, Department of Bioengineering, University of California, Riverside (2007-2008)