Antibody Affinity Maturation in Antigen-Distal Residues

By

Alberto Cisneros III

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMICAL AND PHYSICAL BIOLOGY

May 10th, 2019

Nashville, Tennessee

Approved:

Peggy L. Kendall, M.D. (Chair)

Andrew J. Link, Ph.D

James Chappell, M.D., Ph.D.

Jens Meiler, Ph.D. (Advisor)

James E. Crowe, Jr., M.D. (Advisor)

This work is dedicated to my family.

Thank you for believing in me when I couldn’t.

Acknowledgments

This work would not have been possible without the continued support of my mentors, James E. Crowe, Jr, MD, and Jens Meiler, PhD. Jim, thank you for your unwavering support. When my proteins weren't expressing, and I was pacing the halls in the evenings, hoping for some kind of inspiration, you were there with words of wisdom. You've always been in my corner. Jens, you've always been able to provide tremendous insight and guidance, and have always been willing to share it. I know it hasn't been easy turning a biologist into a full-time Rosetta user; I appreciate all of the extra time you've spent teaching me the basics. You guys are an inspiration, and I can never thank you enough for seeing this through with me. I was also fortunate to have a committee as dedicated to my success as humanly possible. Your continued guidance and understanding helped me through the roughest parts of my graduate career, and I am eternally grateful for your patience.

I would like to thank both the Meiler lab and the Crowe lab for housing me and putting up with my bad jokes. I want to specifically thank a few Crowe lab members and alumni: Gopal Sapparapu, Erica Parrish, Ross Troseth, Rachel Nargi – you've helped me become a better scientist and better person. Gopal, you're a wealth of scientific knowledge and a great friend. Thank you for chatting with me from across the lab bench and walking me through experimental techniques I had never even heard of. Ross, I cannot begin to express my appreciation for your friendship. The afternoon trips to Brown's and rants walking to and from lunch were a welcome distraction from the daily grind. To Erica and Rachel, your understanding of protein expression and purification is the only reason my experiments made it this far. In the Meiler lab, thanks to Amanda, Alyssa, Darwin, and Steven Combs. I didn't even know my way around Linux before I joined the lab, so thanks for putting up with all of

my novice-level questions. Amanda, thank you for walking me through grad school. Who would have thought that two nerds from a state school in Texas grow up to get PhDs? And from Vanderbilt no less! Thanks for being a bright spot in my life.

To my family, thank you for your unconditional love and support. When I was recovering from one of my many back surgeries, you were there to help me walk again. Alex, you're my best friend. I'm sorry you had to pick up everything heavier than 10 lbs. for me for months at a time. Cassandra, thank you for being my confidant and chatting with me on the phone when I was forced to just stare up at the ceiling 24 hours a day. To my partner Brittani, having you near made the last couple of months bearable. Thank you for keeping me focused and encouraging me to give it my all. I love you.

To my mother and father, thank you for believing in me. Thank you for fielding my phone calls at any hour of the day, for keeping me company, for helping me in and out of vehicles, for cooking and cleaning when I couldn't, and for giving me Lucy; she barks a lot but it makes her a good conversationalist. I don't have the words to tell you what you mean to me, so instead I dedicate my PhD and this dissertation to you. We made it.

Table of Contents

CHAPTER

List of Tables

List of Figures

CHAPTER I

Introduction

This thesis is centered around the antigen-recognition molecules produced by human B cells, known as immunoglobulins (Fröland and Natvig, 1972). Immunoglobulins are found in two forms, either as membrane bound B cell receptors or the secreted effector molecules known as antibodies (Hoffman, 2015). Antibodies bind foreign peptides, glycans, and proteins mediating the adaptive immune response by circulating throughout the body and adhering to their specific antigen, thus allowing it to be cleared from circulation (Kirkeby, 2000). In this thesis, I will describe the ternary nature of antibody/antigen interactions and how antigen-distal residues contribute to binding affinity, especially in the context of antibody affinity maturation. This first chapter discusses the mechanisms through which the immune repertoire gains diversity, revolving around the mechanisms that produce antibodies and allow B cells to proliferate. I will introduce HIV as a means through which we can better understand affinity maturation. Finally, I will describe technologies used to explore the relationship between conformational entropy and antibody affinity maturation.

The second chapter of this thesis focuses on techniques for identifying the mechanisms through which non-contact residues contribute to affinity maturation. Prior to my work, dozens of studies had identified the heavy chain/light chain (HC/LC) interface as the primary mediator of the geometry of the paratope (Chailyan, 2011; Masuda, 2006; Hsu, 2014; Dunbar, 2014). Additionally, several studies have attempted to predict the range of motion through computationally expensive means, and many others have sought to identify a single heavy chain – light chain orientation for any given sequence (Bujotzek, 2016; Marze, 2016; Dunbar, 2013). Here, I utilize a novel pipeline that pairs the Rosetta modeling suite with antibody orientation analysis software to interrogate~~s~~ how the composition of the HC/LC interface affects thermodynamic stability and range of HC/LC orientations. In the final chapter, I discuss the results and the thermodynamic implications. I propose several experiments to complement the work presented in this thesis, inform the development of future technologies, and aid engineering endeavors.

Introduction to Antibodies

In humans, B cells begin their development in the bone marrow and complete their maturation in the spleen. This development is delineated by marked changes in surface phenotype, levels of gene expression, and the generation of a unique immunoglobulin molecule comprised of a heavy chain and light chain (HC and LC) (Hardy, 2001). In the bone marrow, hematopoietic stem cells (HSCs) in the bone marrow differentiate into multipotent progenitor cells, then to lymphoid progenitor cells, which receive signals from bone marrow stromal cells and begin B cell development.

Lymphoid progenitor cells become early pro-B cells by beginning the rearrangement of the heavy chain diversity (D) and joining (J) segments, leaving the heavy variable (V) and all light chain gene segments remain in their unrearranged configuration (Allman, 1999). The V(D)J recombination process continues in late pro B cells, which appends a V_H gene segment to the partially rearranged gene to form a fully functional HC gene (Nutt, 1999). A successful V(D)J recombination results in the synthesis of the heavy chain and tested for functionality by binding to the binding immunoglobulin protein (Bip) in the endoplasmic reticulum (ER) (Fritz, 2011). This step of the selection process ensures that the heavy chain folds correctly; nascent heavy chains that fail to bind Bip become targets for degradation, while successful heavy chains associate with a surrogate light chain formed by the VpreB and λ5 proteins, which displace Bip from the C_H1 domain, causing that domain to fold (Taduchi, 2018). At this juncture, they are classified as large pre-B cells and express the rearranged HC alongside the proteins VpreB and λ5, which act as a surrogate light chain, on their surface (Mains, 1983). This trimer, known as the pre-BCR complex (Zhang, 2004), signals the cell to undergo several rounds of proliferation.

The subsequent daughter cells, known as small pre-B cells, then rearrange V and J segments to form light chains. The pre-BCR complex is internalized, and the newly formed light chain replaces the surrogate light chain (Allende 2010). The newly formed B cell receptor, or BCR, is expressed on the surface of the B cell, and the cell is tested for tolerance. Stromal cells and hematopoietic cells express self-antigen on their surface; immature B cells that do not interact with self-antigen are allowed to leave the bone marrow and circulate through the blood, the lymph, and secondary lymphoid organs (SLOs) like lymph nodes and the spleen (Fritz 2014). Immature B cells that bind self-antigen are retained in the bone marrow and undergo receptor editing. Binding to self-antigen signals the B cell to maintain production of the RAG complex (Teigs, 1993). The B cells halt the production of the old light chain, continue VJ recombination to form a new light chain, attempt to form a functional BCR, and are tested against self-antigen again. This process continues until either a new, functional BCR is produced and the B cell avoids interaction with self-antigen, or all light chain VJ rearrangements are exhausted and the B cell undergoes apoptosis (Luring Prak, 2011). Additionally, B cells that bind to soluble self-antigen cease their development and become anergic. Anergic cells no longer continue to express functional BCR and die shortly after. These two tolerance mechanisms are known as “central tolerance” as they occur in the bone marrow (Nemazee, 2017). After leaving the bone marrow, B cells may encounter soluble-self antigen. Upon recognition of self-antigen, auto-reactive B cells become anergic. This mechanism is known and “peripheral tolerance”, as it occurs outside of the bone marrow in the periphery (Pelanda, 2012). When immature B cells leave the bone marrow, they begin to circulate through secondary lymphoid organs, blood, and the lymph. Upon entering the lymph node, B cells are led into primary lymphoid follicles by a gradient of chemokines where they interact with follicular dendritic cells and are stimulated with BAFF, which ensures survival of the B cell and completes the development process (Beyer, 2008). Secondary lymphoid organs also house sites where these mature, naïve B cells are introduced to their specific antigen. Antigen from infected tissue migrates to SLOs and either freely circulates or is presented on follicular dendritic cells (Janeway, 2001). B cells specific for this antigen bind to and internalize it, processing it for presentation to T cells that have T cell receptors specific to the same antigen. Interaction with T cells completes the activation process causing the B cell to proliferate undergo somatic hypermutation, which I will discuss in the next segment.

Antibody Diversity

As one might imagine, recognition of a virtually unlimited number of foreign pathogens and particles requires an incredible amount of antibody diversity. This diversity is determined through four mechanisms. First, during B cell development the immunoglobulin domains are assembled through a process called somatic recombination (Arya, 2018). The variable domains are pieced together from gene segments called the variable (V), diversity (D), and joining (J) segments. The heavy chain is assembled by combining a segment from each of the V, D, and J, genes, whereas the light chain contains only V and J fragments (**Figure I.1**). The heavy chain is encoded by one each of 69 VH, 27 DH, and 6 JH gene segments (Lefranc, 2001).

The second mechanism of diversity occurs during V(D)J recombination as a result of junctional diversity (Alt and Baltimore, 1982). Here, diversity is generated through the removal of nucleotides at the recombination site and subsequent repair to join the segments (Jeske, 1984). These genes have specific sequence motifs adjacent to them called Recombination Signal Sequences, or RSSss. A protein complex containing RAG1 and RAG2 bind specifically to these RSS motifs. The RAG protein complexes bring the the gene segments together and introduce nicks in the dsDNA, cleaving the DNA at the junction which creates hairpin at the end of the gene segments (Jones and Gellert, 2004). Cleavage of the hairpin leaves one side of each gene with a single strand of DNA – an overhang known as Palindromic nucleotides (stemming from the pattern of nucleotides left in the overhang) (Lafaille, 1989). The enzyme terminal deoxynucleotidyl transferase (TdT) processes these overhangs, inserting up to 20 non-templated nucleotides (**Figure I.1**) into the cleaved junctions (Motea and Berdis, 2010). While the exact mechanism through which ligation occurs is still unknown, DNA ligases, protein kinases, and the Artemis nuclease are incorporated into the RAG complex to join the ends of the gene segments together (Malu, 2012). As the name implies, junctional diversity is limited to the junctions formed during recombination. This in turn only generates diversity for the V(D)J junctions, which comprise the CDR3s of both the heavy and the light chains (LeFranc, 2011).

The third mechanism of antibody diversity occurs with heavy chain – light chain pairing, as the light chain is derived from either kappa or lambda genes (Smith, 2016). This light chain diversity is generated through 31 IGKV genes, 5 IGKJ genes, 45 IGLV genes, and 7 IGLJ genes (Lefranc, 2011). While in theory the number of potential light chains (470) when multiplied by the number of potential unique heavy chains (~11,000) gives a total of around 5.2 x 10^6 unique V(D)J HC/LC combinations, the reality is that the combinational diversity may not contribute as much to diversity as expected. Many studies have shown that V_H gene usage can be restricted during infections, and the heterodimeric form that antibodies inhabit provides a means to bind engage a wide variety of antigens, but even if the response were completely random, it is constrained to the immunoglobulin fold by selection mechanisms during B cell development (Wang, 2013).

To compensate for this, additional diversity is garnered through a processes known as somatic hypermutation (SHM), also known as affinity maturation (Maul, 2010), and antibody isotype switching. After B cell activation via antigen recognition and T cell interaction, the enzyme activation-induced cytidine deaminase, or AID, induces point mutations and causes a spike in the rate of mutation by a factor of one million. AID deaminates cytosine nucleotides to uracil, which

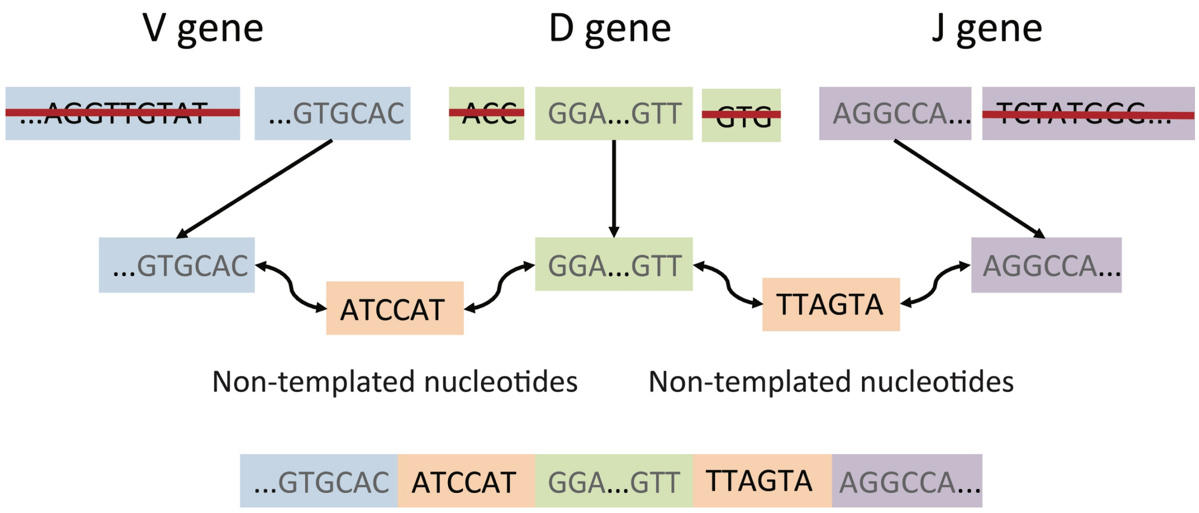


Figure I.1. Junctional diversity is generated during V(D)J recombination. Cartoon representation junctional diversity in naïve B cells.. The heavy chain is built by recombining V (blue), D (green), and J (purple) segments. Cleavage of the RAG-mediated hairpin loops leaves palindromic residues, shown in grey. TdT adds a series of random, non-templated nucleotides (orange) at the junction of each cleaved segment. The segments are ligated together and translated, resulting in an antibody with random amino acids in the VDJ junctions.

causes a uracil-guanine mismatch (Maul 2010). The mechanism through which this results in an error-prone mismatch is still not fully understood (Baretto and Magor, 2011). The DNA is then replicated, and the cells begin to divide. This population of B cells expresses these mutated BCRs with a range of affinity for any given target and has undergone class switching. During proliferation, these genomic rearrangements result in the expression of IgA, IgG, or rarely, IgE isotype antibodies, each of which possesses unique characteristics, function, and structure (Stavnezer, 2004). Activated B cells in follicles change the morphology to secondary follicles containing specialized SLO regions called "germinal centers" (Banerjee, 2016). In germinal centers, B cells that have undergone SHM and class switching continually compete for the antigen presented on FDCs, causing the activation cycle of proliferation and inclusion of mutations to continue (Janeway, 2001). This process can last for weeks, and selects for the BCRs with the highest affinity for their antigen. Some of the surviving B cells migrate to other host SLOs or return to the bone marrow, where they differentiate into plasma cells and secrete high affinity antibody (Chang, 2015), while others become resting, memory B cells that maintain a high affinity BCR (Budeus, 2015).

Antibody Structure

The most common of circulating antibodies is the IgG isotype, which is a homodimer that consists of four polypeptides – two heavy chains, each of which is bound to each other, and identical light chains (**Figure I.2**). The general structure of an IgG molecule can be divided into three segments: The fragment crystallizable (F_c), which contains only the constant regions of the two heavy chains; the fragment antigen-binding (F_{ab}) which contains the variable domain of the heavy and light chain, as well a constant domain for each of the chains; and the fragment variable (F_v), which consists of the variable domains for a heavy chain and it's light chain. This work centers around the Fv region, which consists of two immunoglobulin folds named VH & VL, each of whom is made up of a pair of β sheets (**Figure I.3**). These are built of antiparallel β strands that surround a central hydrophobic core, while VH-VL regions are held together by a series of hydrophobic interactions and sparse hydrogen bonds. The properties inherent to the fold allow for loops to be present at each end (Bork, 1994). Both the heavy and light variable domains contain the complementarity-determining regions – CDRH1, CDRH2, and CDRH3 for the heavy chain and CDRL1, CDRL2, and CDRL3 for the light chain (Schroeder, 2010). Together, these six loops are largely responsible for binding to and recognizing foreign targets or "antigens" (Dondelinger, 2018).

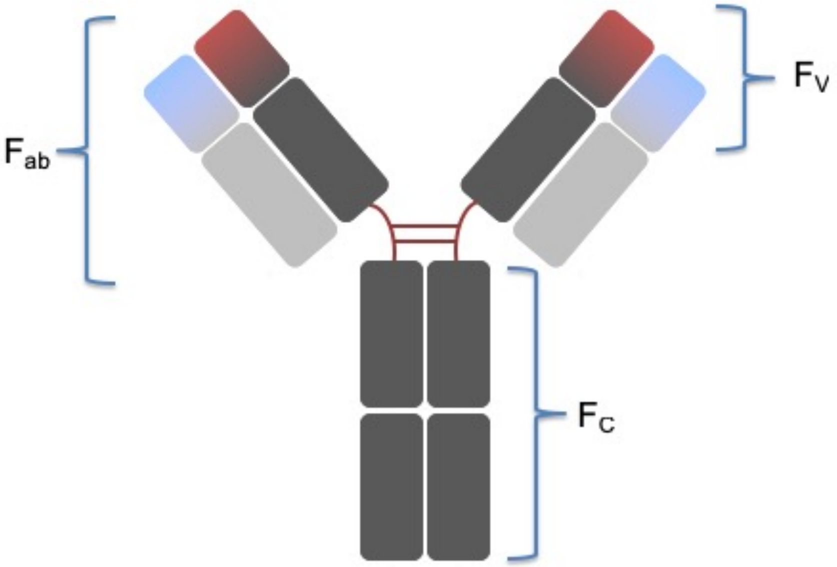


Figure I.2 IgG molecules are homodimers of heterodimers. Cartoon representation of a human IgG molecule. The heavy chain is shown in dark grey, the light chain is shown in light grey. The paratope is shown in blue and red in the F_v region for the light chain and heavy chain, respectively.

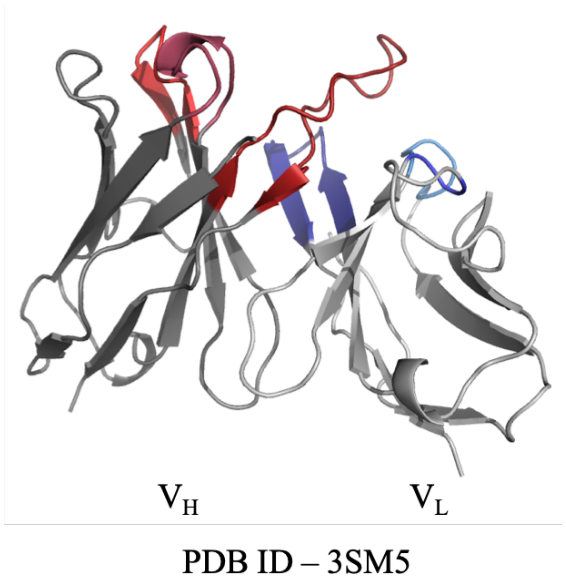


Figure I.3. Structure of the F_v region. Cartoon representation of the CH65 variable region (PDB 3SM5). The heavy chain is shown in dark grey (left), the light chain is shown in light grey (right). CDRH1-3 are shown in red from light red to dark red, and CDRL1-3 is shown in blue from light blue to dark blue, respectively. This representation was reconstructed using PyMol (DeLano, 2002).

Introduction to HIV

The HIV pandemic is a devastating and potentially incurable global health risk. Since the discovery of Human Immunodeficiency Virus as the causative agent of **A**cquired **I**mmuno**d**eficiency **S**yndrome, or **AIDS**, roughly 35 million people have died AIDS-related deaths (UNAIDS,2017). In 2017, it was estimated that 36.9 million infected people were living HIV in 2017, two-thirds of whom do not have access to antiretroviral therapy (UNAIDS, 2017). Nearly 2.0 million people previously uninfected people contracted the virus last year, and 940,000 individuals died AIDS-related deaths (UNAIDS, 2017). The virus is spread through unprotected sexual contact and the bodily fluids associated with it, from mother to child through breastmilk and contact with blood during childbirth, and through penetration of the skin or contact with mucosal membranes by HIV contaminated materials (such as used needles).

HIV is an enveloped retrovirus with a positive strand single-stranded RNA. The virus infects human CD4$^+$ T cells, macrophages, and dendritic cells with its Env glycoprotein. Cell entry centers around the CD4 receptor and a co-receptor, either CCR5 or CXCR4. The envelope glycoprotein forms a trimeric protein called gp160, which contains a transmembrane domain, a series of highly glycosylated, highly variable loops named V1-V5, and the CD4 binding site. In functional virions, gp160 is cleaved into to parts. The first, gp120, is highly glycosylated, contains the variable loops V1-V5, and the CD4 binding site. This protein is responsible for cell adhesion (**Figure I.4)** and initiates a conformational change after binding host CD4 that allows it to interact

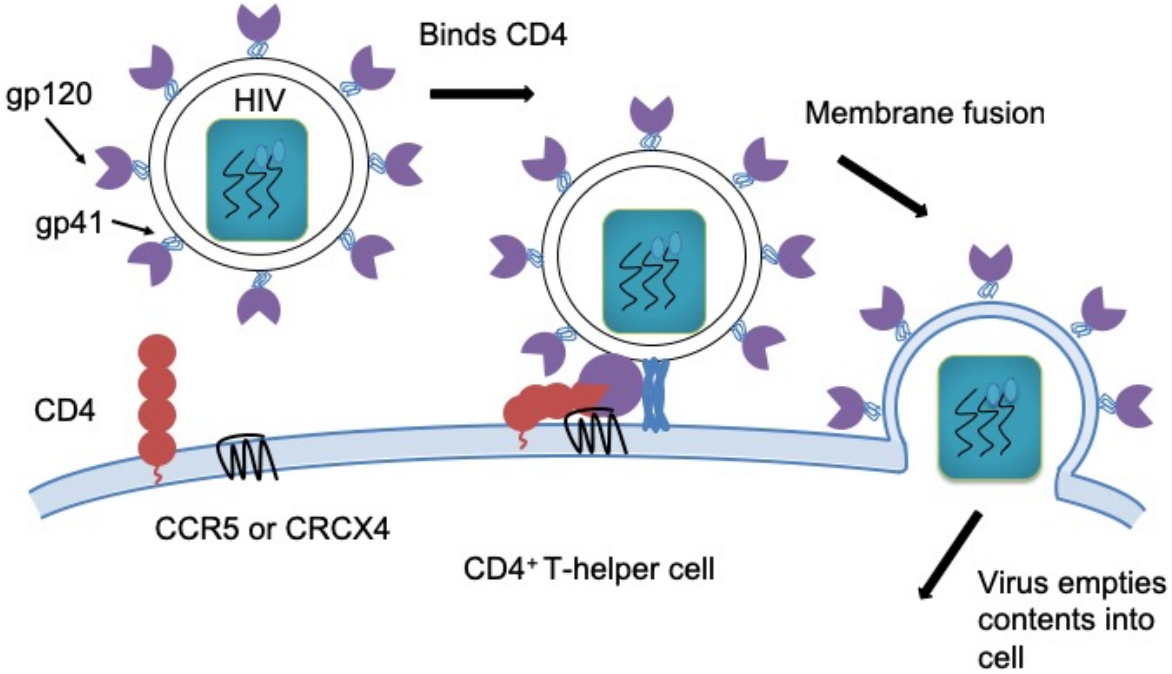


Figure I.4. HIV mechanisms of entry. HIV gp120 (purple) binds host CD4 (red) and undergoes a conformational change that allows it to bind to host-co receptor CCR5 or CRCX4. This change brings gp41 into close proximity with the cell membrane, and initiates gp41machinery involved in membrane fusion. The viral genome and associated proteins enter the cell. Created using Microsoft PowerPoint.

with the host co-receptor (Wilen, 2012). The second product of cleavage, gp41, contains the transmembrane domain and mediates membrane fusion. Like all retroviruses, HIV encodes a reverse transcriptase to create double stranded DNA from its RNA genome This newly synthesized dsDNA genome is eventually integrated into the host genome using integrase, an enzyme that aids the insertion of the viral DNA. HIV maintains a mutation rate of $(4.1 \pm 1.7) \times 10^{-3}$ per base per cell, which is higher than any reported rate for a virus (Cuevas, 2015). Because of this, the initial antibody response to an HIV infection is highly mutated and polyreactive (Liao, 2011), though typically incapable of neutralizing virions. Occasionally, either through the polyreactivity generated by the initial response or random association and continued somatic hypermutation (Mouquet, 2010), antibodies are generated against the HIV CD4 binding site (CD4BS). Broadly neutralizing antibodies to the HIV CD4BS can prevent infection in cells as they target the primary site associated with cell entry. Recent clinical studies have shown that some broadly neutralizing antibodies have the potential to protect against infection or suppress viremia (Scheid, 2016; Bar-on, 2018). These antibodies bind by mimicking host-CD4; the immunoglobulin fold of the heavy chain is strikingly similar to that of host CD4, and uses part of the framework region to interact with the CD4 binding loop (Zhao, 2010) (**Figure I.5**).

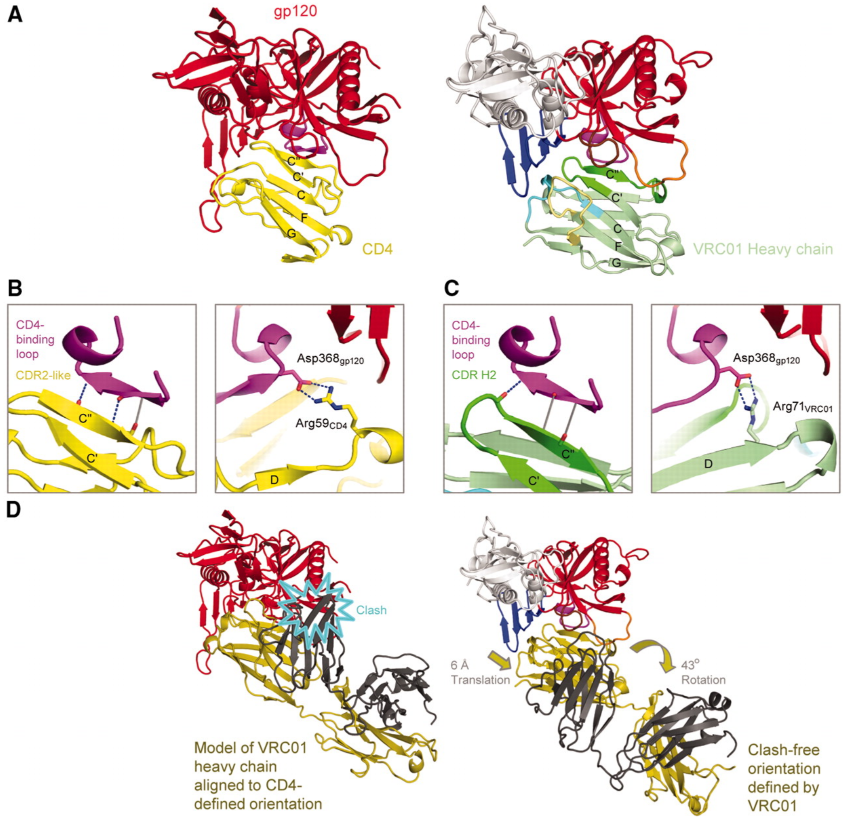


Figure I.5. Structural mimicry of CD4 interaction by antibody VRC01. VRC01 shows how a double-headed antibody can mimic the interactions with HIV-1 gp120 of a single-headed member of the immunoglobulin superfamily such as CD4. **A)** Comparison of HIV-1 gp120 binding to CD4 (N-terminal domain) and VRC01 (heavy chain-variable domain). Polypeptide chains are depicted in ribbon representation for the VRC01 complex (right) and the CD4 complex with the lowest gp120 RMSD (left). The CD4 complex (3JWD) is colored yellow for CD4 and red for gp120, except for the CDR-binding loop (purple). The VRC01 complex is colored as in Fig. 1. Immunoglobulin domains are composed of two β-sheets, and the top sheet of both ligands is labeled with the standard immunoglobulin-strand topology (strands G, F, C, C', C"). **B,C)** Interface details for CD4 **(B)** and VRC01 **(C)**. Close-ups are shown of critical interactions between the CD4-binding loop (purple) and the C" strand as well as between Asp368gp120 and either Arg59CD4 or Arg71VRC01. Hydrogen bonds with good geometry are depicted by blue dotted lines, and those with poor geometry in gray. Atoms from which hydrogen bonds extend are depicted in stick representation and colored blue for nitrogen and red for oxygen. In the left panel of C, the β15-strand of gp120 is depicted to aid comparison with B, though because of the poor hydrogen-bond geometry, it is only a loop. **D)** Comparison of VRC01- and CD4-binding orientations. Polypeptides are shown in ribbon representation, with gp120 colored the same as in **(A)** and VRC01 depicted with heavy chain in dark yellow and light chain in dark gray. When the heavy chain of VRC01 is superimposed onto CD4 in the CD4- gp120 complex, the position assumed by the light chain evinces numerous clashes with gp120 (left). The VRC01-binding orientation (right) avoids clashes by adopting an orientation rotated by 43° and translated by 6-Å. Adapted from (Zhao, 2010).

The Rosetta Software Modeling Suite

The Rosetta software modeling suite is a collection of computational tools designed to create biologically relevant protein models and simulate their interactions with other proteins, peptides, small molecules, and DNA. The Rosetta energy function estimates *in silico* the total free energy of the complex and also the binding free energy of an antibody. Rosetta includes tools to construct comparative models for antibodies and antigens of interest. One of the most commonly used applications of Rosetta is prediction of the structure of a complex between antibody and antigen via docking. This application samples all possible interactions between the two protein partners to identify the biologically relevant interface (Weitzner, 2017). It simultaneously optimizes the conformation of the bound state. The docking algorithm is Monte-Carlo based, and starts with a centroid-mode stage to interrogate the potential docking poses and is followed by an all-atom refinement stage intended to optimize the docked pose and side chain conformations (Gray, 2003), though the flexibility of the docking application allows for either of those steps to be enacted individually. The protocol used for docking is determined by the user, and can be either local or global. In global docking Rosetta randomly orients the two partner proteins; this method is particularly useful when little biological information is known. Alternately, the application can initiate local perturbations, which assumes the pose provided in the input PDB file is close to optimal, and restricts the movements to small perturbations.

One of Rosetta's most notable successes is the design of Top7, a 93 residue protein with a topology previously undiscovered in nature (Kuhlman, 2003). The RosettaDesign algorithm identifies the lowest-energy sequence for any given target structure by iteratively alternating between optimizing the sequence for a static backbone and energetically minimizing the backbone to accommodate the new sequence. Each designable position samples every amino acid (provided that the user does not limit this) using rotamers from the Dunbrack library (Dunbrack and Karplus, 1993). This robust process can be applied in several ways; common applications of this algorithm include the redesign of existing interfaces to alter specificity and deriving the optimal sequence for any given structure. In the following sections I review the Rosetta methods critical for the research I performed in my thesis, centering around docking and design.

Rosetta Energy Function and Relax Protocol

Since its onset, the Rosetta protein modeling suite has been employed to inform a variety of biological studies. Rosetta has been used to successfully design Top7, a 93 residue protein with a topology previously undiscovered in nature (Kuhlman, 2003), redesign protein interfaces for altered specificity (Lewis, 2013), engineer small antibody-mimetic proteins against viral proteins (Fleishman, 2011), determine the structure of proteins from sparse experimental data (Thornburg, 2013; Wang, 2016; Sangha, 2017), create comparative models of antibodies (Weitzner, 2014), and determine the antibody-antigen interface via docking (Weitzner, 2017). These feats are achieved using the Rosetta energy function, which is responsible for scoring the models generated by the aforementioned applications. This energy function is derived from the statistical distribution of geometric parameters in proteins whose structure is known (knowledge-based) and, in some instances, physics-based potentials (Alford, 2017). This is known as a knowledge-based energy function, as it’s score terms are generated by analyzing experimental data. The total energy of a system is calculated as a linear combination of van der Waals interaction, hydrogen bonds, solvation using an implicit water model, electrostatic interactions, among a number of additional terms (Leaver-Fay, 2013). Over the decades the scoring function has been succeeded by a series of incarnations, each derived from and designed to more closely resemble the ever-growing collection of experimental data generated by the scientific community. This is made possible by the continued efforts of the Rosetta community; Rosetta is co-developed by over 50 laboratories around the world, comprised of protein engineers, structural biologists, computational biologists, computer scientists, and experimental biologists.

For many applications, computational modeling using Rosetta centers around modifying a naturally occurring protein backbone structure to accommodate new interactions, create new functional sites, or alter biological activity (Nivón, 2013). Although Rosetta energy functions are constantly being improved, experimental input protein structures almost always have regions that are in a sub-optimal conformation according to the Rosetta energy function. This could be because of inaccuracies in the energy function or in the input structure. Regardless, such frustrations must be removed prior to docking or designing a protein to avoid artefacts in the calculation. For this reason, it is often necessary to energetically minimize an input structure. This energetic minimization, known as "relax", explores the immediate conformational space – iteratively optimizing the sidechain interactions and backbone angles of the protein (**Figure I.6B**) (Tyka, 2011). These small structural changes are made stochastically and evaluated using the all-atom energy function with the aim of identifying the lowest free energy conformation (**Figure I.6A**). This protocol has been shown to both lower the overall energy of a Rosetta model (Bradley, 2005; Conway, 2014) and improve low-resolution crystal structure by refining interactions to more closely mirror nature (Bender, 2016).

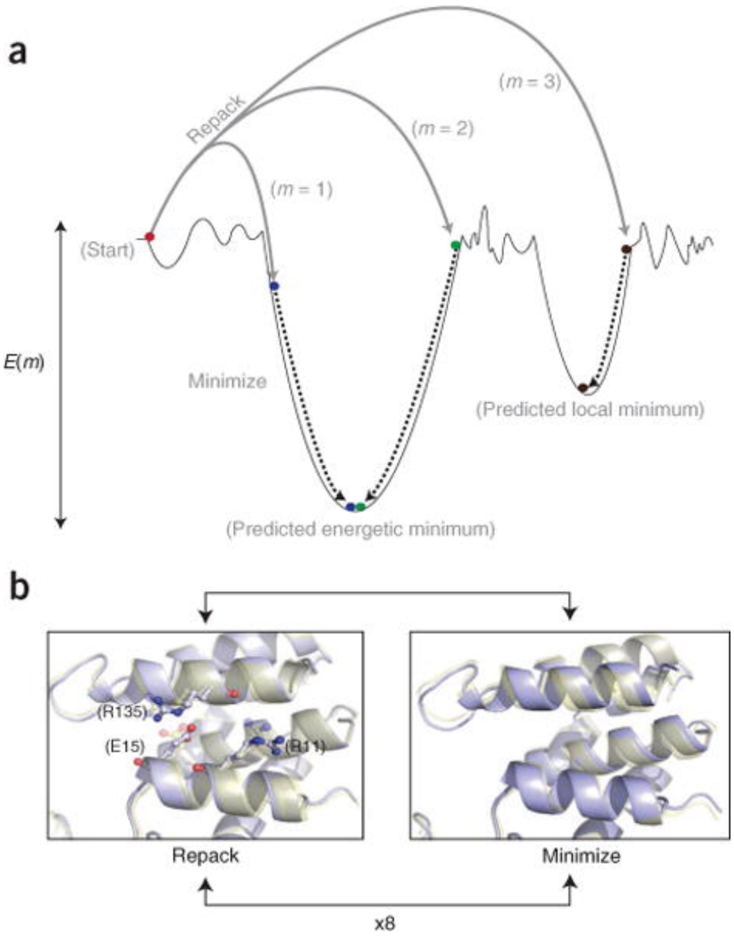


Figure I.6 An overview of Rosetta energetic minimization and all-atom refinement via the relax protocol. (**a**) Simplified energy landscape of a protein structure. The relax protocol combines small backbone perturbations with side-chain repacking. The coupling of Monte Carlo sampling with the Metropolis selection criterion36 allows for sampling of diverse conformations on the energy landscape. The final step is a gradient-based minimization of all torsion angles to move the model into the closest local energy minimum. (**b**) Comparison of structural perturbations introduced by the repack and minimization steps. During repacking, the backbone of the input model is fixed, whereas side-chain conformations from the rotamer library33 are sampled. Comparison of the initial (transparent yellow) and final (light blue) models reveals conservation of the R135 rotamer but changes to the R11 and E15 rotamers. Minimization affects all angles and changes the backbone conformation. Adapted from (Combs, 2013).

Antibody Comparative Modeling including RosettaAntibody Server

High-resolution comparative modeling is crucial for biological engineering applications for which no crystallographic structure is readily available. Comparative modeling in antibodies merges three distinct protocols. The first, called "threading", involves identifying homologous antibody frameworks from existing crystal for a target sequence and aligning a target sequence to that structure. The framework regions are highly conserved across antibodies, which aids the production of biologically relevant models. Rosetta assesses the extent to which the sequence fits that structure. The next step involves grafting non-HCDR3 complementarity determining regions are grafted onto the framework regions. Previous studies have characterized the CDRs found in known antibody structures in an effort to identify patterns in loop structure (Al-Lazikani, 1997; North, 2011). Using a dataset of over 300 non-redundant antibody structures, North et al. found that 85% on non-HCDR3 CDRs can be assigned to one of 72 clusters. Comparative modeling protocols leverage the ordered, canonical nature of the non-HCDR3 CDRs, grafting loops from other known structures onto the models based on the desired length and sequence. Unlike the other CDRs, the HCDR3 does not inhabit "canonical" structures, and is often modelled using a *de novo* approach. These methods often employ a modified kinematic closure (KIC) method, which calculates the all of the conformations for the 6 torsion angles of a peptide chain, and samples N-Cα-C bond angles (Mandell, 2009) (**Figure I.7**). Recently, this modeling method has been paired with HC/LC docking in an effort to modify the HC/LC orientation of the models, thus allowing the HCDR3 to sample conformations across a spectrum of orientations (Sivasubramanian, 2009). In an effort to streamline this process, several labs have generated optimized, standalone algorithms or protocols that can produce comparative models (Adolf-Bryfogle, 2018; Norn, 2017) given when amino acid sequences for the heavy and light chain, the most notable of which is the ROSIE webserver which houses the RosettaAntibody methodology (Sivasubramanian, 2009).

Rosetta antibody docking

While the generation of structurally accurate comparative models in informative, their significance is dependent on establishing how they interact with antigen. High-resolution antibody/antigen co-crystals provide insight into the molecular determinants of binding and

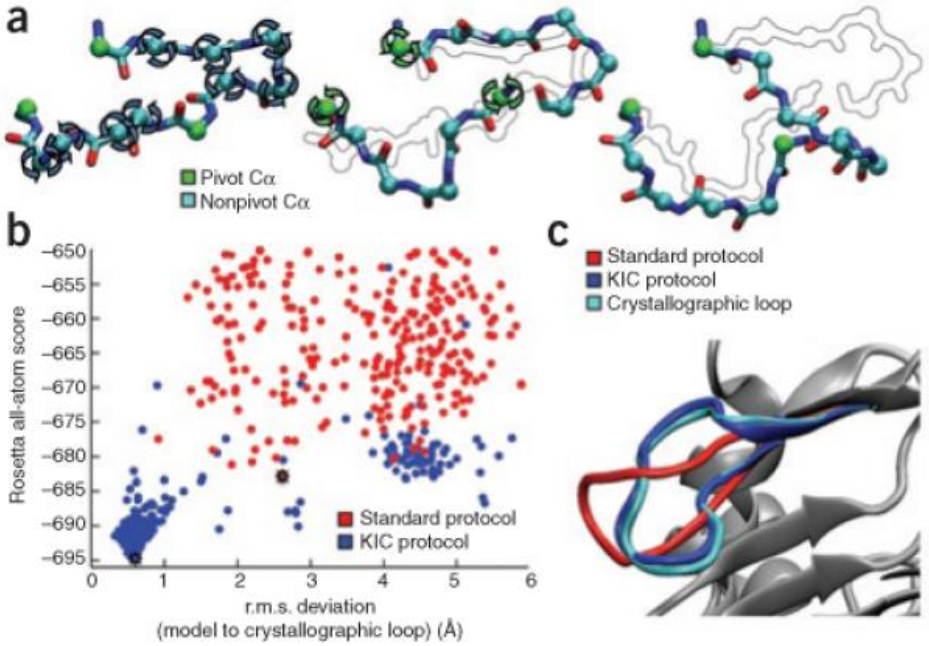


Figure I.7 Loop reconstruction with KIC. (a) In the KIC move, 3 Cα atoms of an N-residue chain are designated as pivots (green spheres); the remaining N – 3 are non-pivot Cα atoms (cyan spheres; left). In a 12-residue loop, 24 torsions are modeled. Non-pivot torsions are sampled from a residue type-specific Ramachandran map, opening the chain (middle). KIC then finds all values for the pivot torsions that close the loop, if any exist, keeping the endpoints fixed (right). The previous state is shown in outline. (b) Performance of the Rosetta KIC protocol and standard protocols on a 12-residue loop (Protein Data Bank (PDB): 1srp). Only KIC densely sampled regions < 1.0 Å r.m.s. deviation from the crystallographic loop. Asterisks mark the lowest-scoring reconstructions from the two methods. The Rosetta all-atom score includes the enthalpy plus the solvation contribution to the entropy but not the configurational entropy. (c) The lowest scoring reconstructions from b are shown. KIC improved reconstruction accuracy to 0.6 Å from 2.6 Å using the standard protocol. Figure adapted from (Mandell, 2009).

neutralization, which in turn informs vaccine design and the development of therapeutic antibodies. Occasionally, experimental methodology fails to produce a viable antibody/antigen co-crystal; the discovery of novel antibody/antigen interactions and the subsequent elucidation of a viable structure of the interaction is limited in throughput, and not all antibody/antigen interactions produce viable, high-resolution structures. In these circumstances, protein-protein docking can be used to predict the biologically relevant interface. This method, previously described in detail, was developed to extensively sample potential interactions between two proteins. Although *de novo* prediction of protein-protein interactions is difficult, antibodies interact with antigens using a limited set of loops and framework regions – the epitope may still need to be determined, but the sample space concerning the paratope is relatively limited. This process is greatly enhanced by the inclusion of known experimental data; the Rosetta docking algorithm has been successfully used to dock an anti-dengue antibody using NMR to limit search space to the predicted epitope (Simonelli, 2013), predict the epitope and binding orientation of *de novo* modeled anti-inflenza proteins against a conserved epitope (Fleishman, 2011), and determine the antibody-antibody, idiotype-anti-idiotype complex by conserving known interface contacts (Vangone, 2014).

Rosetta multi-state and germline polyspecificity

During somatic hypermutation, antibodies gain mutations that either directly increase binding affinity to a target by adding complementary interactions (increasing the enthalpic gain) or pre-configure the paratope for binding (mitigating entropic loss). As described earlier in the introduction, one of the defining features of the human antibody repertoire is its ability to recognize an astonishing number of pathogens with a limited number of unique antigen-naïve BCRs; germline antibodies are polyspecific, and add diversity to our immune system through conformational flexibility. The polyspecificity of germline-encoded antibodies can be recapitulated using a Rosetta design application termed “multi-state design”. Where traditional design optimizes the sequence for a single antibody that binds a specific target, multi-state design can be used to design an antibody that binds to multiple targets simultaneously (Leaver-Fay, 2011). Using this method, Willis et al. showed that in contrast to antibodies that have accumulated mutations during affinity maturation, germline-gene encoded antibodies are inherently better suited for polyspecificity (Willis, 2013). The authors selected three mature antibodies each targeting a different antigen but derived from V_H 5-51, and sought to design the antibodies against the three antigens simultaneously. The germline sequence was favored when binding to multiple antigens was a requirement of the design (**Figure I.8**). Additionally, the positions that were reverted to the germline residue during the simulation showed greater deviation in their phi-psi torsion angles when compared to the mature residues – an indication that the germline sequence is inherently more flexible.

Rosetta design recapitulates antibody maturation

In contrast, the RosettaDesign algorithm provides a means to improve upon antibodies, and given the stochastic nature of both antibody affinity maturation and “design”, sufficiently recapitulate somatic hypermutation. The study performed by Willis et al. concerning germline residues also revealed that allowing Rosetta to design somatically mutated positions often returned the residue of the mature antibody (**Figure I.8**) (Willis, 2013). While somatic mutations incurred *in vivo* might be optimal at their respective positions for a particular interaction, the variable nature of antibodies means that some known antibody/antigen interactions are potentially sub-optimal and therefore prime targets for *in silico* maturation. Willis et al. re-designed the HCDR3 of PG9, an anti-HIV antibody with secondary structure in the HCDR3 (Willis, 2015). The authors used Rosetta to re-design the HCDR3 loop in order to identify variants with an increased affinity to the V1/V2 loops. While the majority of the HCDR3 residues were recovered in sequence (an indication that these residues were already optimal), the authors isolated a variant that demonstrated increased potency and neutralization of HIV by altering a single residue. Sevy et al. employed multi-state design to redesign the HCDR3-mediated, anti-influenza antibody C05 with increased potency and breadth across strains (Sevy, 2019). The resulting variants exhibited improved binding affinity and an increase of breadth with respect to the binding profile of C05. These studies demonstrate how the RosettaDesign algorithm, in conjunction with stochastic mutations and a robust energy function, can both mimic somatic hypermutation in antibodies and continue where it left off.

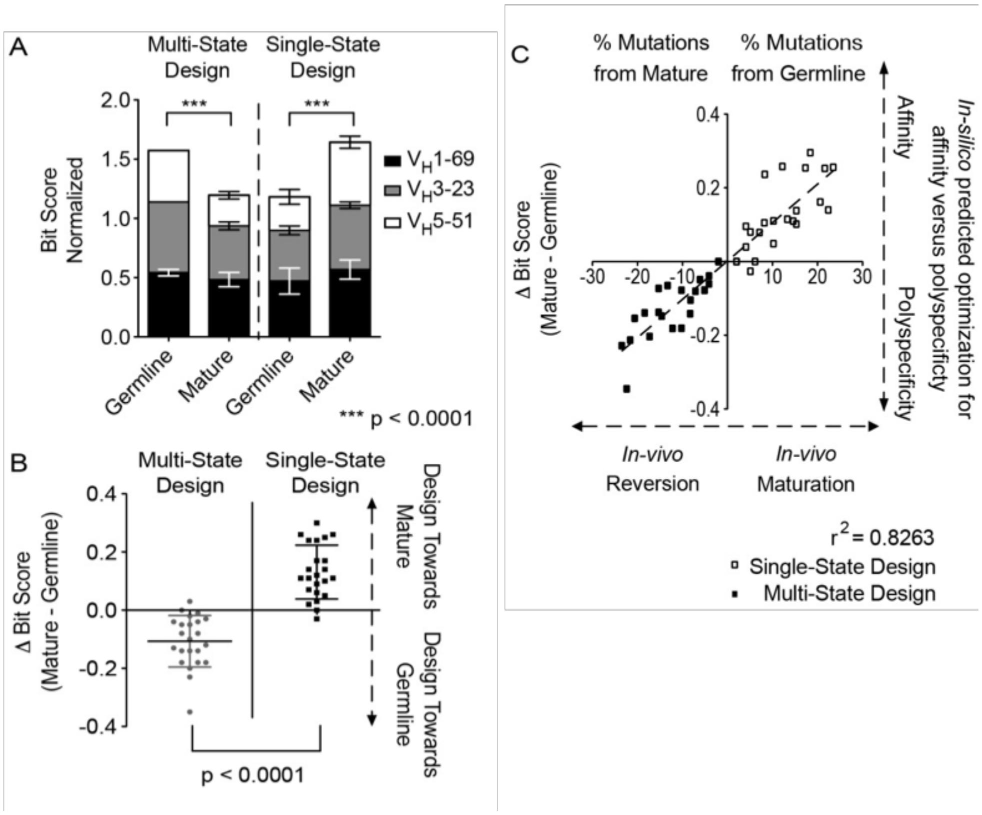


Figure I.8 Multi-state designs toward the germline sequence, single-state to mature sequences. Antibodies encoded by the same inferred germline V_H gene preferred germline sequences when considered in the multi-state design, inferring a more flexible combining site. (A) The bar graph shows the bit-score for each of the three different inferred germline groups and then the sum of the scores in a grouped bar. A perfect design would have a normalized bit-score of 1.0, and summated score of 3.0 for three germline groups. Multi-state design preferred germline sequences for all complexes, while in contrast single-state design preferred mature sequences ($p<0.0001$). (B) The change in bit-score is determined to be the proclivity to either the mature (positive score) or the germline (negative score) sequence. Each complex was assigned a change in bit-score. The change in proclivity between design protocols was significant ($p<0.0001$). (C) Each complex was scored against mature and germline sequences and a difference was calculated (Δbit-score). Positive numbers returned showed a proclivity towards mature sequences, while a negative score suggested a design toward germline. A tight correlation was observed ($r^2=0.8263$) for the *in silico* predicted optimization for specificity versus polyspecificity *(*Δbit-score) and the *in vivo* maturation process (plotted as the mutation percentage away from V_H gene sequence). Adapted from (Willis, 2013).

Computational derivation of HC/LC relative orientation

The relationship between the amino acid composition and the relative orientation between heavy and light chain remains a point of interest among both computational biologists and antibody engineers. Many previous studies have indicated that the interface between the heavy and light chain determines the geometry of the paratope, and the relative orientation between the two domains may act as an additional form of antibody diversity in naïve B cells (Chothia, 1985; Davies and Metzger, 1983; Stanfield, 1993). While each of these studies pointed out the importance of the relative HC/LC orientation , they were unable to identify the determinants of orientation. Early attempts at describing the HC/LC orientation from crystal structures ranged from calculating a single packing angle (Abhinandan, 2010) to deriving four angles that account for various metrics (Marze, 2016). Dunbar et al. developed software known as ABangle, which calculates HC/LC orientation from the six degrees of freedom generated by the association of two proteins (**Figure I.9D**). ABangle uses consensus domains generated from structurally invariant positions across all antibody crystals – these positions were the most conserved in relation to one another, and provide the basis for the rest of the ABangle calculations (**Figure I.9A**). This method provides the greatest description HC/LC orientation as it accounts for all of the degrees of freedom associated with the orientation of two rigid-body objects, and is the basis for the angle calculations in Chapter II.

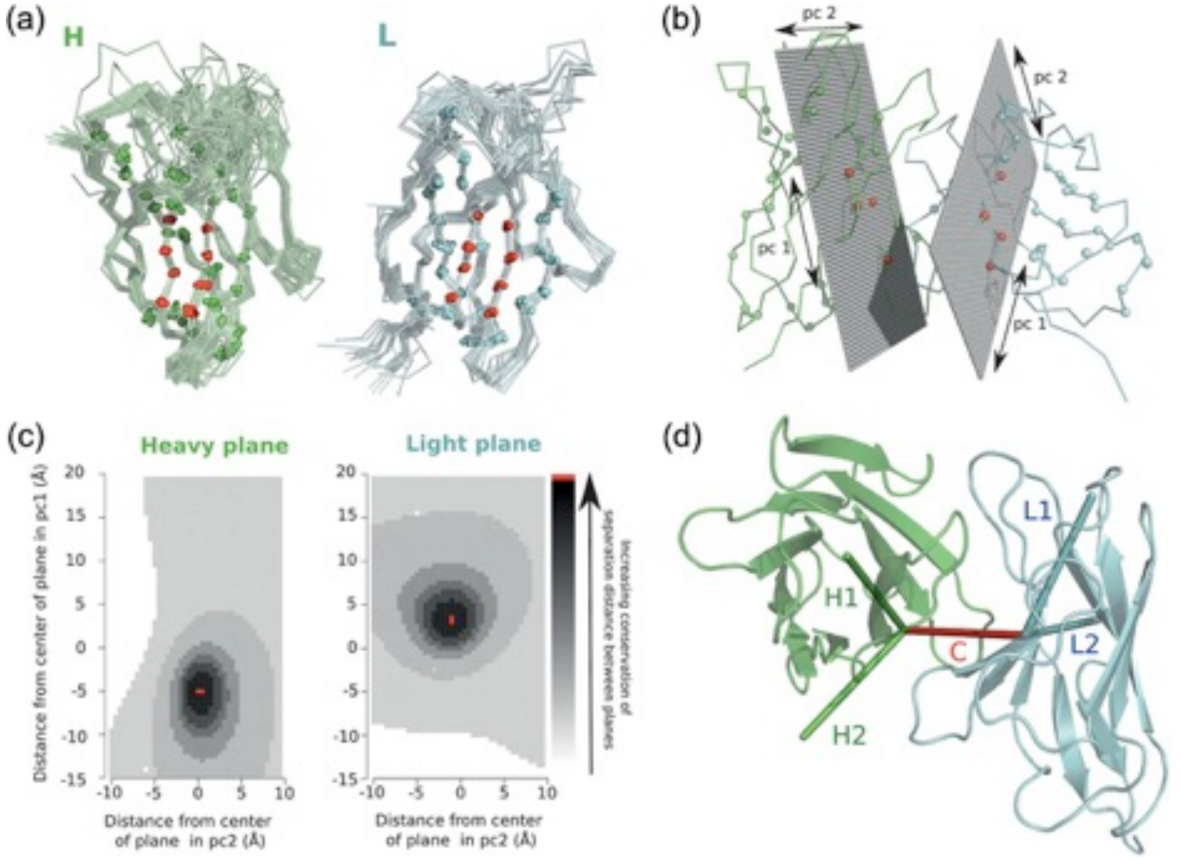

(a)
H
L
(b)
pc 2
pc 2
pc 1
pc 1
(c)
Heavy plane
Light plane
Distance from center of plane in pc1 (Å)
20
15
10
5
0
-5
-10
-15
-10
-5
0
5
10
Distance from center of plane in pc2 (Å)
Increasing conservation of separation distance between planes
(d)
H1
H2
C
L1
L2

Figure I.9. **Construction of consensus domains**. **A)** Superposition of 30 representative VH (green) domains showing the coreset positions (spheres) and the eight positions (red), 240 coordinates sets, used to generate the VH plane. In cyan is the corresponding image for VL. **B**) The average coreset positions (consensus structure) and VH and VL reference planes aligned to the antibody Fv 1B4J_HL. **C**) Calculation of vector C, which runs through the points on the VH and VL reference planes that have the most conserved distance over the 351 Fv structures in the non-redundant set. **D**) Our coordinate system mapped onto 1B4J_HL. H1 and H2 are vectors that are parallel to the principal components used to create the VH reference plane in **(B)**. L1 and L2 are similarly defined for VL. Adapted from (Dunbar, 2013).

Significance

Previous studies have inferred that non-contact residues contribute to overall binding affinity and activity of HIV antibodies, however they were not focused on the unique role of heavy chain/light chain interactions in governing binding affinity or the thermodynamic implications behind such a mechanism (Klein, 2013). In this work I apply a novel pipeline to observe and interpret the changes in orientation that can be attributed to mutations in the HC/LC interface by computational modeling and docking. The models generated for the HC/LC interface reversion exhibited a shift in the range of HC/LC orientations sampled during docking, which is consistent with the concept of mutations in the HC/LC interface as a means of mitigating entropic loss upon binding and increasing the enthalpy for the bound conformation. The results show that highly mutated HIV-specific CD4 binding site antibodies achieve unusual orientation features that are distinguishable from most human antibody heavy chain/light chain orientations in order to bind their epitope and that mutations in the HC/LC interface govern the overall orientation of the CDRs by modulating the range of accessible orientations.

The mechanisms involved in B cell development and subsequently in affinity maturation require a conserved HC/LC interface in order to achieve HC/LC pairing. As previously described in the introduction, B cells undergo somatic hypermutation, incorporating random mutations in their BCRs in order to increase the affinity of their receptors for antigen and compete for survival. Traditionally, mutations in the HC/LC interface have been seen as disruptive to the aforementioned process, as they may prevent efficient HC/LC pairing (Koenig, 2017). The work presented in this thesis challenges this conventional notion of antibody affinity maturation. The findings show that antigen-distal somatic mutations in the HC/LC interface indirectly affect binding affinity through mitigation of entropic loss and pre-configuration for the bound conformation of the antibody, and extends the known molecular determinants of antibody/antigen binding and neutralization to include non-contact residues, thus discerning an additional mechanism through which binding affinity is mediated.

Additionally, while an increasing number of co-crystal structures become available, it is still unknown whether the antibodies in complex are optimal in sequence and structure in terms of affinity for the target. The work defined in this thesis suggests that antibodies devoid of mutations in HC/LC interface can be improved using their bound conformation as a template for *in silico* affinity maturation. The studies defined in chapter II and those proposed in chapter III are of critical importance to antibody engineers that design high affinity interactions and computational biologists looking to create stable proteins.

CHAPTER II

Role of antibody heavy and light chain interface residues in affinity maturation of binding to HIV envelope glycoprotein

Adapted from Cisneros 3rd A, Nargi RS, Parrish EH, Haliburton CM, Meiler J, Crowe Jr. JE. Role of antibody heavy and light chain interface residues in affinity maturation of binding to HIV envelope glycoprotein. *Mol. Syst. Des. Eng*., 2019; Advance Article

Author contributions: I designed and ran all the experiments outlined in this chapter under the mentorship of James Crowe and Jens Meiler. I analyzed all of the data with my mentors and created all of the figures presented in this chapter.

Abstract

The F_V region of an antibody consists of the heavy chain (HC) and light chain (LC) variable domains whose association is maintained by a series of conserved, non-polar interactions. During chronic infections, somatic mutations are induced, often in the HC/LC interface. Sequence variation in these interactions allows the HC and LC domains to inhabit a range of orientations relative to one another. Thus, we hypothesize that these interface mutations are critical to orient and rigidify the HC/LC interface to arrange the paratope for optimal interaction with the antigen, thereby affecting antigen binding affinity allosterically. To test this hypothesis, we measured the HC/LC orientation of a set of broad and potent human HIV neutralizing antibodies. The HC/LC interface of these antibodies contained a large number of mutations and achieved unusual relative

orientations compared to other human antibodies. We expressed and characterized a panel of recombinant HIV CD4 binding site antibodies as the fully matured variant and compared these with variants mutated to the HC/LC interface of the inferred unmutated common ancestor antibody. We found that HC/LC interface reverted antibodies have a reduced affinity, confirming that introduction of somatic mutations in the HC/LC interface was one of the critical steps in affinity maturation. We then used the Rosetta software suite to examine the mechanisms through which these mutations affect binding affinity. We determined to what extent the mutations were critical in altering the relative orientation of HC/LC domains to a conformation that is competent to bind the antigen. We further determined whether the mutations excluded alternative HC/LC conformations that would be incompetent to bind the antigen. These findings suggest that somatic mutations in the HC/LC interface, distant from the antigen/antibody contact region, play a critical role in affinity maturation of HIV antibodies by preconfiguring the bound conformation of the antibody in the orientation required for high affinity recognition of the antigen. Thus, optimization of HC/LC interface could serve as an important tool for maximizing antibody/antigen binding affinity without altering antigen contact residues.

Introduction

The adaptive immune response (occasionally referred to as the “acquired immune response”), is the mechanism through which humans eliminate both bacterial and viral infections (Alberts, 2002) The strength of the adaptive immune response lies in its ability to recognize a vast number of foreign pathogens given a limited number of gene segments and options for gene segment recombination (Koonin, 2015). Prior to B cell activation, the diversity of the antibody repertoire is generated by V(D)J gene segment recombination (Roth 2014); using 69 V_H, 27 D_H, and 6 J_H gene segments, the immune system generates over 11,000 unique VDJ recombination events for the heavy chain alone. Taking into account the 31 IGKV genes, 5 IGKJ genes, 45 IGLV genes, and 7 IGLJ genes that comprise light chains in conjunction with the junctional diversity that stems from recombination, an estimated 10^{11} antibodies can populate an individual's antibody repertoire (Glanville, 2009). These antibodies, while dwarfed in number by the theoretical possible number of epitopes on pathogens that the immune system might encounter, provides compensatory protection through structural flexibility. Germline antibodies that have yet to undergo affinity maturation often are polyspecific and bind multiple targets at low affinity through a flexible binding surface – the paratope (Willis, 2013). However, antibody/antigen complexes are ternary structures (Sherrif, 1987). Since antibodies are formed by the combination of a heavy chain (HC) and a light chain (LC), the complex with the antigen constitutes a three-way interaction (**Figure II.1**). This secondary interface in the F_v region of the antibody allows the variable HCs and variable LCs to take on a wide range of orientations relative to one another and is responsible for determining the geometry of the paratope (Chailyan, 2011; Abhinandan and Martin, 2011).

Additionally, activated B cells undergo affinity maturation – an iterative process involving somatic hyper-mutation (Hwang, 2015), the process by which mutations are made in the rapidly proliferating B cell, diversifying the B cell receptor, and positive selection that ultimately leads to target-specific antibodies (Tiller, 2017). Antibodies that evolve in response to lifelong infections like HIV are often mutated beyond what we see in transient infections (Burton, 2005); upwards of 48% of amino acids in the V_H gene for anti-HIV antibodies like VRC01 are mutated from their germline precursor (Georgiev, 2014). Fera et al. discerned that some anti-HIV CD4 binding-site (CD4BS) antibodies incorporate somatic mutations in the VH-VL interface to alter the geometry of the combining site, accommodating for the insertion of the HIV V5 loop (Fera, 2014) (**Figure II.1**). This change in orientation is thought to mediate breadth of binding and neutralization.

Several recent studies have shown that antigen-distal somatic mutations accumulated in the framework regions (describe framework) of an antibody can drastically affect the breadth of neutralization and binding affinity profile (Georgiev, 2014; Julien, 2017), though the mechanism

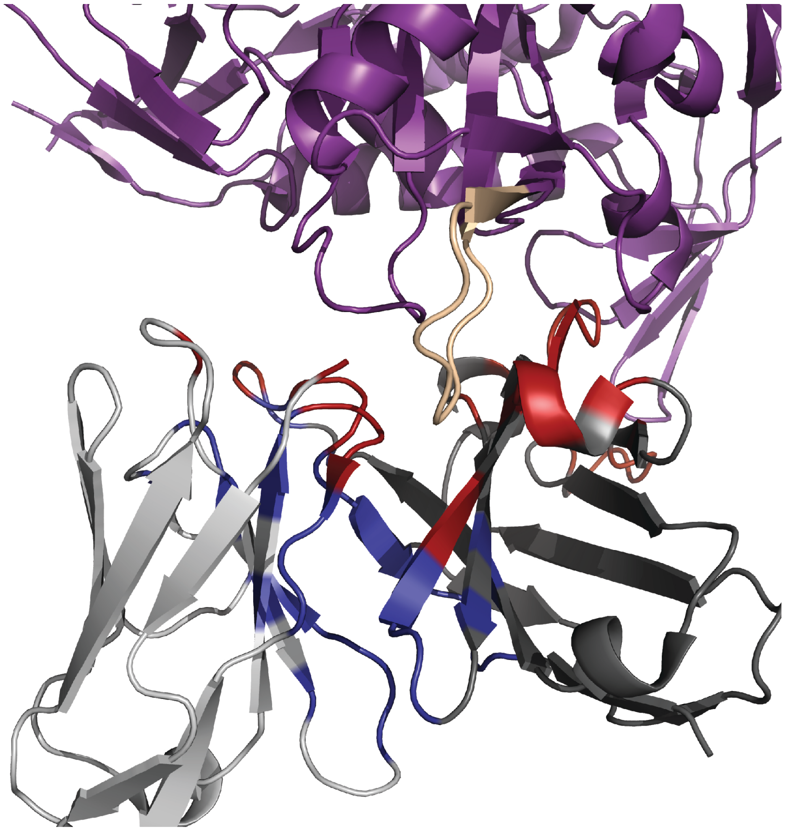


Figure II.1. The ternary nature of antibody/Antigen interfaces. Cartoon representation of antibody-antigen and HC/LC interface using VRC03 and 93TH057 gp120 (PDBID:3SE8). The gp120 is shown in purple, the V5 loop is shown in yellow, the heavy chain is shown in dark grey, and the light chain is shown light grey. The paratope was defined as residues within a Cβ-Cβ cutoff of 8 Å or a pair of non-hydrogen atoms within 5.5 Å across the antibody/antigen interface, and is shown in red. The HC/LC interface was defined using the same interface parameters but excludes residues that are included in the paratope and is shown in blue. (Explain method?)

through which these antigen-distal mutations increase binding affinity remains unknown. Understanding the mechanism through which antigen-distal somatic mutations affect binding affinity offers a new venue through which therapeutic antibodies may be improved and can aid vaccine design. We postulate that a constrained relative orientation between the HC and LC is needed for an antibody to engage its target with maximum affinity, as the relative orientation of this interaction defines the geometry of the antibody paratope. We hypothesize further, that mutations in the interface that enhance affinity of antibody/antigen interaction are introduced in an allosteric manner during antibody maturation. We distinguish two principal mechanisms: **A)** The HC/LC orientation needed to engage the antigen has an increased energy compared to the most likely conformation in the germline antibody. Mutations in the HC/LC interface are needed to 'shift' the HC/LC orientation in a binding competent conformation (**Figure II.2A**). This change largely would confer an enthalpic effect on antigen binding, stabilizing the HC/LC interface in an orientation that allows optimal engagement of the antigen. **B)** In the second scenario, while the germline antibody has its lowest free energy HC/LC orientation at the conformation needed to engage the antigen, a large number of alternative conformations are possible. Mutations in the HC/LC interface are needed to disfavor binding incompetent HC/LC arrangements, *i.e.*, 'tightening' of the conformational ensemble of HC/LC arrangements, thereby reducing the entropic cost of binding by locking the HC/LC orientation in a preconfigured state optimal for binding the antigen (**Figure II.2B**). This concept is consistent with our understanding that recombined germline gene-encoded antibodies are capable of binding to a wide variety of epitopes (Willis, 2013). We appreciate that in reality mixtures of both scenarios are not only possible but also likely. To begin testing this hypothesis, we reverted the HC/LC interface of CD4BS antibodies that contained a large number of somatic mutations. We then characterized the antibodies and their reverted counterparts. We found that mature antibodies bound to gp120 with an increased binding affinity. We employed a structure-based computational approach to predict the amplitude of ‘shifting’, ‘tightening’, and ‘interface stabilization’, as illustrated in **Figure II.2**. The models generated for the HC/LC interface reversion exhibited a different range of HC/LC orientations consistent with the concept of shifting. We also observed weaker and broader minima in the HC/LC interaction, consistent with the concept of tightening. These findings suggest that selection of clones with somatic mutations in the HC/LC interface that preconfigure high-affinity binding sites can act as critical mechanistic component of affinity maturation.

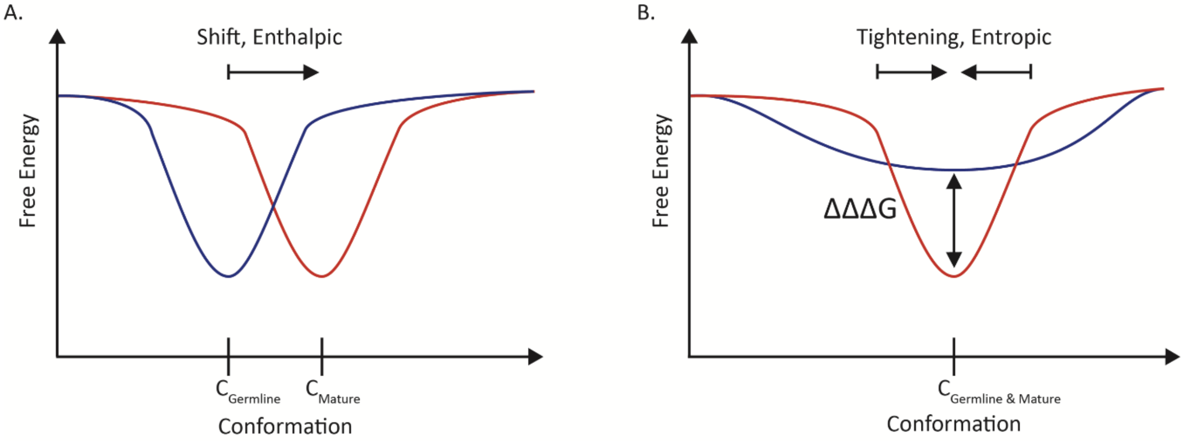
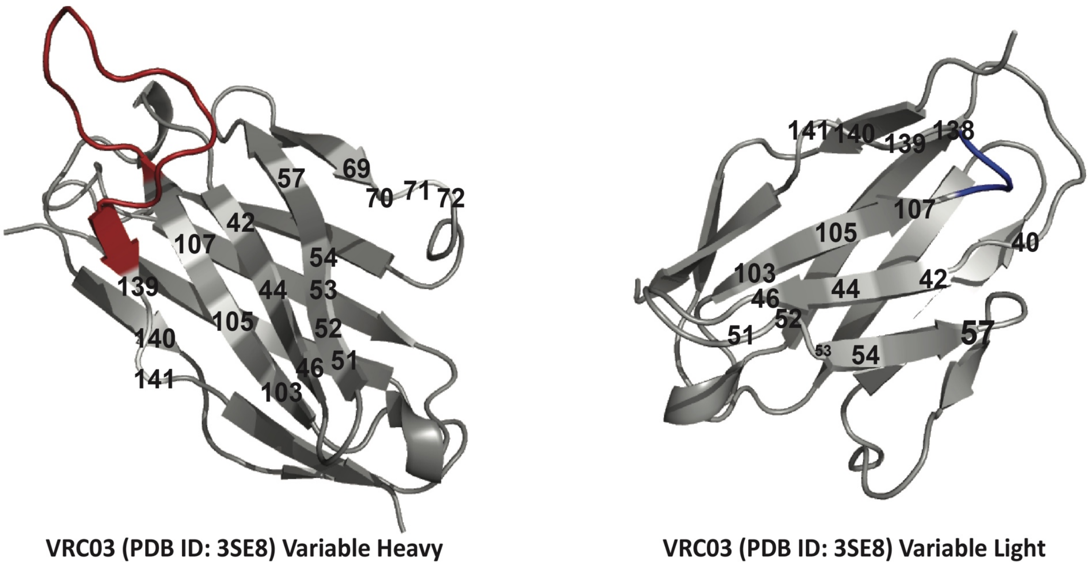


Figure II.2. Energy landscape for the HC/LC complex before and after Somatic Hypermutation. A) The germline antibody has one conformation for which the free energy is minimal at $C_{Germline}$ (shown in blue). After affinity maturation (shown in red), this lowest energy conformation is shifted to create the optimal paratope conformation for the antibody/antigen interaction at C_{Mature}. Somatic mutations that shift this free energy minimum optimize the enthaplic gain for the antibody/antigen interaction. **B)** The germline antibody (shown in blue) has its lowest free energy conformation already at the optimal conformation for antibody/antigen interaction ($C_{Germline}$) but is flexible, alternative low energy conformations exist. After affinity maturation (shown in red), this flexibility is reduced to limit entropic cost of binding and increased stability of the mature conformation for the antibody/antigen interaction (ΔΔΔG).

Results

Definition of the HC/LC interface. In order to define the HC/LC interface, we obtained the coordinates of all available human antibody/antigen co-crystal structures with a resolution better than 3Å in the Protein Data Bank (PDB). The complete set of 466 human antibody/antigen complex structures was downloaded from the PDB in February 2018. After eliminating redundant structures, single-domain antibodies, and point mutants, 301 structures were used for this analysis. We determined the structural parameters of the HC/LC interface for a representative structure using in-house software, using PDB ID: 4M5Z containing the structure of human influenza-specific mAb 5J8 that exhibited ABangle scores near the average for each parameter (**Table S1-S4**). As interface residues, we counted all amino acids that had 1) Cβ atoms of two amino acids i and j within 8 Å across the HC/LC interface or 2) any pair of non-hydrogen atoms within 5.5 Å across the HC/LC interface or 3) a Cβi-Cαi-Cαj angle of less than 75° across the HC/LC interface. Using these criteria, we identified 18 HC and 17 LC interface residues in PDB ID: 4M5Z. Amino acid sequences for the HC and LC variable regions for each antibody were curated from the PDB and numbered with the AHo numbering scheme (Honegger and Plückthun, 2011) using the Antigen receptor Numbering And Receptor Classification (ANARCI) webserver (Dunbar and Deane, 2016). The AHo numbering scheme then was mapped onto the structurally derived interface positions, providing a structurally conserved, sequence-based definition of the HC/LC interface (**Table S2**, **Figure II.3**). We used this definition to create multiple sequence alignments (Crooks, 2004) for antibodies that fail to bind CD4BS (**Figure S2A**) compared to antibodies that bind CD4BS (**Figure S2B**).

Maturation in the HC/LC interface stabilizes variable domain interactions. In order to test the effect of somatic mutations in the HC/LC interface, we first determined the range of orientations of HC and LC that have been described to date in high-resolution structures of ternary antibody/antigen complexes. We used the ABangle software (Dunbar, 2013) to evaluate the relative HC/LC orientation of each of the 301 antibody/antigen co-crystal structures. ABangle determines six features of orientation of heavy and light chain, five angles (designated HL, HC1, LC1, HC2 and LC2) and a distance (dc). The program uses the most structurally conserved residue positions in HC and LC to define domain location and then maps a HC and LC frame plane onto the Fv structure. The tool measures HC/LC orientation essentially by measuring the angles between these two plane segments using a vector with the most conserved length in PDB Fv



VRC03 (PDB ID: 3SE8) Variable Heavy

VRC03 (PDB ID: 3SE8) Variable Light

Figure II.3. Definition of HC/LC interface. Cartoon representations of the heavy chain (left) and light chain (right) domains. The CDR3 of each domain has been colored red (HCDR3) or blue (LCDR3) for reference. Interface positions are colored light grey and have been numbered for clarity.

structures (designated C) as the pivot axis of HC/LC orientation. H1 is the vector running parallel to the first principal component of the HC plane, while H2 runs parallel to the second principal component. L1 and L2 are defined in a similar way on the LC domain. HL is the torsion angle between H1 and L1; HC2 is the bend angle between H2 and C; LC1 is the bend angle between L1 and C; LC2 is the bend angle between L2 and C; dc is the length of C.

The ABangle angle distributions identified antibodies that deviate significantly from the typical HC/LC interface features, i.e., by one to two standard deviations. The distribution of the six ABangle features for human antibody co-crystals is shown in **Figure S1**. We found that the structure of anti-HIV gp120 CD4BS-specific antibodies represented a class of antibodies with unusual features in interface orientation angle HC1 (**Table 1**, **Figure II.4**). The average non-CD4BS antibody had an HC1 angle of 71.3 ± 1.72, while the CD4BS antibodies exhibited an average HC1 angle of 74.0 ± 2.69. These HIV antibodies exhibited a tighter HC/LC interface in

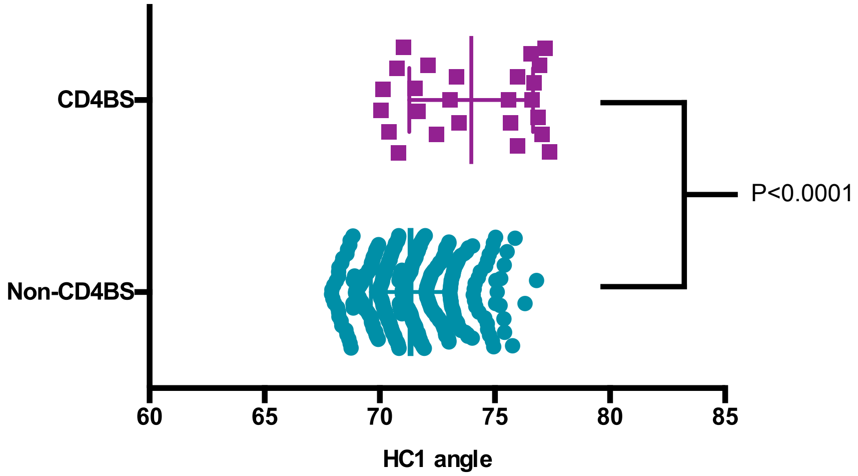


Figure II.4. Comparison of antibody HC1 angle orientations. Histogram showing HC1 angle for CD4BS antibodies (magenta), and non-CD4BS antibodies (teal). CD4BS antibodies possess a larger HC1 angle and smaller dc (distance between domains) than non-CD4BS antibodies.

terms of proximity of the heavy and light immunoglobulin domains. Also, the number of somatic mutations from the inferred germline amino acid sequence in the HC/LC interface with a distance over 5 Å from the antigen ranged from 5 to 12 mutations, with an average of ~9 mutations per antibody (**Table 2**). Unfortunately, the total number of antibody/antigen co-crystal structures is still too small to confidently determine biases in the six ABangle parameters introduced by the germline gene segments. Thus, while CD4BS-specific antibodies deviate statistically significantly from non-CD4BS antibodies, it remains unclear how much of this bias is introduced by the selection of specific germline gene segments.

Construction of Rosetta models of interface-reverted CD4BS antibodies. To determine the effects that these naturally occurring somatic mutations had on the antibody bound conformation, we used Rosetta to construct ensembles of models for the HC/LC interface germline-reverted antibodies. In order to maintain the bound conformation, the protocol was limited to a rigid-body threading, which fixes the backbone coordinates and replaces the side chains in question, followed by a constrained minimization (Nivón,) that allows the structure to adjust to its new sequence while

Table 1. Angle distribution for non-CD4BS and CD4BS antibodies

Angle	Non-CD4BS N=281	CD4BS n=20
HL	-58.7 ± 3.74	-57.6 ± 3.97
HC1	71.3 ± 1.72	74.0 ± 2.69
LC1	120.2 ± 2.30	123.1 ± 2.75
HC2	118.4 ± 2.75	114.8 ± 3.22
LC2	82.9 ± 2.01	83.9 ± 3.56
Dc	16.2 ± 0.27	15.9 ± 0.56

preserving the relative HC/LC orientation. The Rosetta total energy of a system is calculated as a linear combination a series of weighted terms, such as van der Waals interactions, electrostatic interactions, hydrogen bonds, and the Lazaridis-Karplus solvation energy. Using the Rosetta scoring function as a surrogate for free energy, we calculated the HC/LC interface energy **(ΔΔG)** for both the mature and interface-reverted models. Computing $\boldsymbol{\Delta\Delta\Delta G = \Delta\Delta G_{mature} - \Delta\Delta G_{Reverted}}$, we found that that mutations in the HC/LC interface of CD4BS-specific antibodies stabilized the bound conformation (**Table 2**). This finding was true in every case of HIV-specific CD4BS antibody except that of VRC01. This particular mAb differs from the others in that reversion of one somatic mutation restored a canonical glutamine-glutamine interaction in the interface upon reversion.

Table 2. Quantification of differences in orientation between reverted and mature antibodies

Antibody	Tightening[a]	Shift[b]	ΔΔΔG[c]	Mutations[d]
VRC01	0.76 ± 0.12	1.05 ± 0.05	6.28[e] ± 0.34	7
VRC03	1.09 ± 0.18	1.68 ± 0.05	-1.77 ± 0.31	12
VRC-PG04	0.61 ± 0.15	1.33 ± 0.04	-5.79 ± 0.28	9
VRC23	1.64 ± 0.32	1.54 ± 0.06	-3.65 ± 0.45	8
VRC06	0.92 ± 0.13	0.47 ± 0.07	-0.05 ± 0.35	9
12a12	2.48 ± 0.86	0.92 ± 0.12	-3.49 ± 0.25	8
VRC-PG20	0.53 ± 0.04	0.66 ± 0.02	-1.62 ± 0.48	11
VRC07	0.92 ± 0.06	0.91 ± 0.08	-2.50 ± 0.36	5
8ANC131	10.9 ± 2.96	1.53 ± 0.17	-2.23 ± 0.33	12
3BNC117	N/A[f]	N/A	-11.89 ± 0.50	6

[a] The average tightening for each of the ABangle parameters generated by HC/LC docking. Error was calculated for the average of the six ratios for each antibody.

[b] Normalized shift value for each set of distributions generated by HC/LC docking

[c] Change in stability at HC/LC interface. ΔΔΔG = Mature ΔΔG – Reverted ΔΔG.

[d] Number of mutations in the HC/LC interface that do not interact with the antigen

[e] Positive value can be attributed to re-establishing canonical Q-Q interaction in HC/LC interface

[f] Reverted models did not conform to requirements of ABangle software

Mutations in the HC/LC interface shift orientation towards the antigen-bound conformation. Tightly-packed protein cores and interfaces are integral to overall protein stability and the free energy of folding (Kellis, 1988; Geiger-Schuller, 2018). In order to determine how maturation of the HC/LC interface affects the geometry of the paratope, we performed an iterative, small-perturbation docking protocol using the RosettaDock algorithm (Gray, 2003) (**Figure II.5**). Here, we used an all-atom, rigid-body refinement method that incorporated small perturbations in terms of translating and rotating, allowing the structure to explore conformations close to the starting

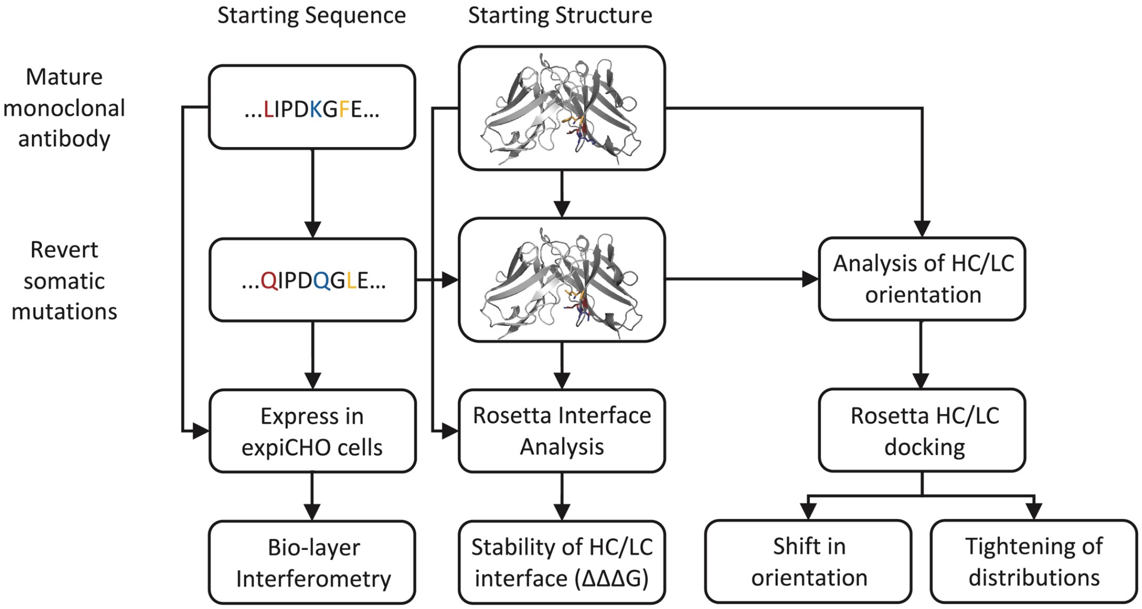


Figure II.5. Flowchart representing methods of HC/LC interface interrogation. The path on the left describes the process through which biophysical characterization of the antibodies takes place; the mature antibodies are expressed in ExpiCHO cells alongside reverted antibodies, and binding to YU2 gp120 is measured through BLI. The center path portrays a simplified overview of the rigid-body modeling process: using a CD4BS antibody structure (cartoon representation of PDB ID: 3se8) as a template, the corresponding reverted amino acid sequence is threaded onto the structure. The reverted model and CD4BS structure are then subjected to constrained minimization, and analysis of the HC/LC interface. The path on the right depicts an overview of the HC/LC docking process. A CD4BS antibody structure and the corresponding reverted model individually undergo small perturbation docking at the HC/LC interface to identify changes in preferred orientation. The docked models are analyzed using ABangle, and the resulting angles are used to identify shifts in orientation and tightening of distribution.

point while preserving the structure of each domain (**Figure II.6**). This procedure was followed by a constrained minimization step that allowed for small adjustments to structure but did not alter the structure enough to affect orientation. These small movements generate models with a wide range of orientations; as the HC/LC interface is reorganized, the structure explores new conformations and the models may converge on a different energetic minimum, generating differences in observed angle distributions.

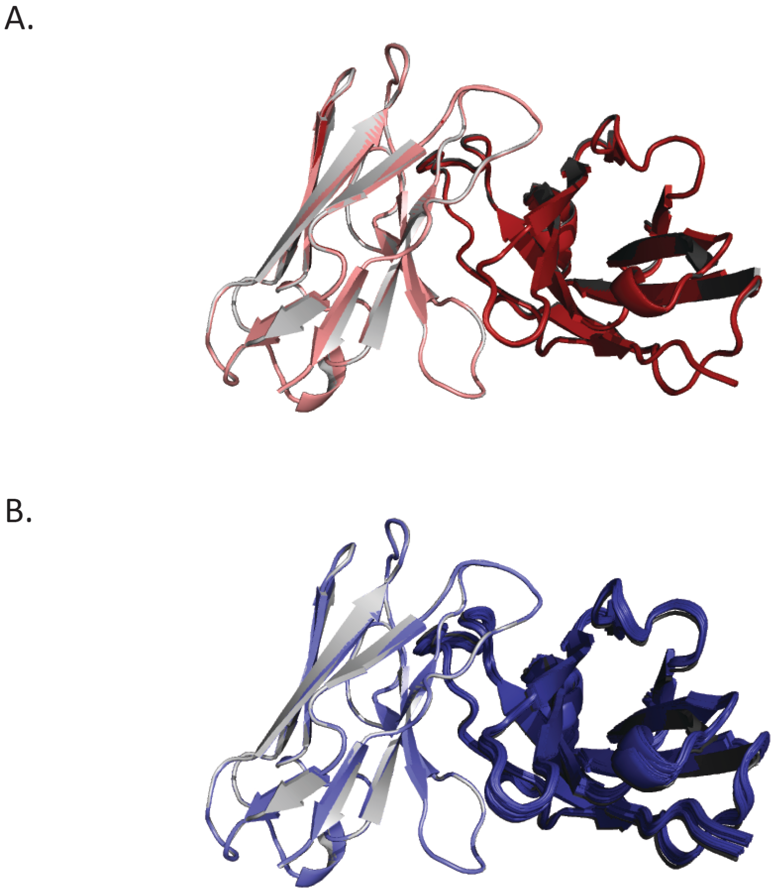


Figure II.6. Affinity maturation in HC/LC interface makes bound conformation more favorable. The top scoring 5% docked models for 8ANC131 were aligned to the light chain (light grey) of the Fv found in the crystal structure (PDBID: 4RWY). **A)** The mature models, shown in dark red (heavy chain) and light red (light chain), maintain the bound conformation after docking, resulting in a "tighter" distribution of angles. **B)** The reverted models, shown in dark blue (heavy chain) and light blue (light chain), vary in HC/LC orientation more than the mature counterpart, resulting in a higher value for the Tightening metric (Table 2).

The changes in orientation were calculated as follows:

(1) ***Normalized shift*** =

$$\frac{1}{6}\sum \frac{\left|\bar{X}_{Reverted(HL,etc\ldots)} - \bar{X}_{mature(HL,etc\ldots)}\right|}{\sigma_{Reverted(HL,etc\ldots)} + \sigma_{mature(HL,etc\ldots)}}$$

where $\bar{X}_{Reverted(HL,etc\ldots)}$ is the mean ABangle value for any angle distribution generated by docking a reverted HC/LC interface, $\bar{X}_{mature(HL,etc\ldots)}$ is the corresponding mean ABangle value for the mature antibody, where $\sigma_{Reverted(HL,etc\ldots)}$ is the standard deviation for any given angle distribution generated by HC/LC docking at a reverted interface, and $\sigma_{mature(HL,etc\ldots)}$ is the standard deviation for the corresponding mature antibody distribution. The Normalized Shift metric provides an estimate of how much the orientation distributions differ between any given mature antibody and its reverted counterpart as a whole. Values greater than one suggest a shift in each category by an average of 1 standard deviation.

(2) ***Tightening*** =

$$\frac{\sigma_{Reverted(HL,etc\ldots)}}{\sigma_{mature(HL,etc\ldots)}}$$

The tightening equation generates a ratio of standard deviations. Values greater than 1 suggest that the mature antibody models embody a tighter angle distribution during HC/LC docking. The standard error (SE) for the shift was calculated using error propagation rules for addition: where shift in HL angle =

$$\left|\bar{X}_{Reverted(HL,etc\ldots)} - \bar{X}_{mature(HL,etc\ldots)}\right|$$

$SE_{(HL)}$ =

$$\sqrt{SE^2_{Reverted(HL)} + SE^2_{mature(HL)}}$$

normalized $\boldsymbol{SE}_{(HL)}$ =

$$\frac{\sqrt{\boldsymbol{SE}^{\mathbf{2}}_{\boldsymbol{Reverted(HL)}} + \boldsymbol{SE}^{\mathbf{2}}_{\boldsymbol{mature(HL)}}}}{\boldsymbol{\sigma}_{\boldsymbol{Reverted(HL)}} + \boldsymbol{\sigma}_{\boldsymbol{mature(HL)}}}$$

and $\boldsymbol{SE}_{(Normalized\ Shift)}$ =

$$\frac{\mathbf{1}}{\mathbf{6}}\sqrt{\boldsymbol{SE}^{\mathbf{2}}_{\boldsymbol{(HL)}} + \boldsymbol{SE}^{\mathbf{2}}_{\boldsymbol{(HC1)}\dots}}$$

A total of 1,000 models was constructed, and the top scoring 5% models were used to evaluate the magnitude of shift in mean angle and the tightening of each distribution (**Table S2**). We calculated the changes in a normalized value for the shift in orientation (Equation 1), and the average tightening between distributions for each angle (**Table 2**). While some variation was observed in the details of how individual mature antibodies differed from their reverted counterparts, we observed two distinct mechanisms through which changes in angle distributions were established: 1) Antibodies whose germline interfaces are not optimized for the orientation needed to interact with the CD4 binding site shift their range of motion by accumulating somatic mutations that lock in the necessary orientation – the optimal orientation is achieved by establishing new electrostatic bonds and van der Waal's interactions that pre-configure the paratope without necessarily restricting the range of conformations accessible to the HC and LC domains, increasing affinity through enthalpic contribution (**Table 2**, mAb VRC-PG04); 2) Antibodies with a range of conformations optimized for the orientation needed to interact with the CD4 binding site accumulate mutations in the HC/LC interface that disfavor suboptimal orientations, tightening the range of accessible conformations – greater affinity for the target is achieved by mitigating the entropic loss upon binding (**Table 2**, mAb 12a12). In some cases, affinity maturation in the HC/LC interface caused both a shift in the optimal orientation and limited the range of favorable conformations (**Table 3**, mAb 8ANC131 and mAb VRC23). 8ANC131 has a noticeably tighter distribution than the other antibodies in the study. This can be attributed to the tight angle distribution produced by docking at the mature HC/LC interface, whose average standard deviation was ~0.13 (**Table S2**).

Binding Studies. We tested our hypotheses using binding assays with recombinant proteins (**Table 3**) to elucidate any change in binding affinity upon reversion to a germline HC/LC interface. We compiled the sequences of the three pairs of antibodies, synthesized cDNAs encoding the antibody variable regions, cloned them into a mammalian expression vector, and expressed each clone as a full-length IgG protein in ExpiCHO cells followed by Protein G column purification. We expressed a panel of three pairs of antibodies (each pair containing the wild-type and the interface-reverted variants). We were able to express both variants for three pairs of antibodies at levels high enough to test accurately in affinity of binding assays. While the antibodies used in this study bind to, and neutralize, a wide variety of HIV strains, their breadth of reactivity for diverse strains converge on select gp120 molecules. We used bio-layer interferometry (Octet RED96, Pall FortéBio) to measure the apparent KD, Ka, and Kd for the interaction of each antibody with recombinant gp120 protein. In every case, the presence of inferred germline gene residues in the HC/LC interface dramatically decreased binding affinity to HIV gp120 (**Table 3**). This apparent loss of binding affinity was caused by a decrease in Ka, while the Kd remained relatively unchanged.

Table 3. Binding affinity KD (nM) for reverted or mature antibodies using bio-layer interferometry

Antibody	K_D (nM)	K_D error (nM)	K_a (1/Ms)	K_d (1/s)
VRC-PG04 reverted	42.8	5.30	3.8×10^{3}	1.4×10^{-4}
VRC-PG04 mature	38.4	0.04	1.2×10^{5}	2.6×10^{-4}
VRC-PG20 reverted	86.4	47.20	1.5×10^{5}	1.0×10^{-2}
VRC-PG20 mature	7.1	0.05	1.5×10^{6}	1.3×10^{-2}
8ANC131 reverted	n.d.	n.d.	n.d.	n.d.
8ANC131 mature	51.1	48.30	3.2×10^{5}	1.7×10^{-2}

Discussion

Antibodies are increasingly used as therapeutic agents, and optimization of antibody structure and function remains a chief concern for biochemical engineers (Chames and Batey, 2009; Leavy, 2010; Beck, 2010). Traditionally, the induction of somatic mutations during antibody maturation was thought to increase binding affinity and specificity to a target by altering the composition of the combining site and creating complementary interactions; pre-configuration of the paratope is not an entirely new concept, but previous analyses have been limited to identifying molecular determinants involved in stabilizing the CDRs (Xu, 2015; Ofek, 2010; Mishra, 2018) or inducing mutations in the HC/LC interface and characterizing the mutants (Chetallier, 1996; Huge, 2003). Many aspects of an antibody-antigen interaction are well-studied upon discovery of a biologically interesting antibody: Da Silva and colleagues demonstrated that the light chain can play in integral part of binding by performing a comprehensive mutagenesis study on the LCDRs of the antibody D5 (Da Silva, 2010); Fera et al. found that the HIV V5 loop alters the orientation of heavy and light domains in CD4BS antibodies by comparing the crystal structures of unbound inferred germline precursors of CH103 antibodies to mature CH103-gp120 complexes; and analysis of the near-pan broadly neutralizing antibody by Huang et al. revealed that N6 adopts a unique heavy chain orientation relative to the binding mode of CD4 and other CD4BS antibodies in order to avoid steric clashes stemming from glycosylation (Huang, 2016). While studies centered around an antibody-antigen interaction often acknowledge unusual bound HC/LC orientations and mention the inclusion of somatic, antigen-distal mutations, the relationship between the two is often unexplored.

Identifying the molecular determinants of binding affinity will facilitate the development of new therapeutic antibodies and may provide an additional target for antibody optimization. In this study, we determined that affinity maturation can take place in the HC/LC interface by introducing mutations that shift the range conformational flexibility towards the optimal configuration by establishing new interactions (enthalpic gain) or by introducing mutations that restrict the conformational space of the V_H and V_L domains, mitigating the entropic loss associated with binding. The concept of antibodies "shifting" the range of energetically favorable in response to the affinity maturation is compatible with the induced fit model of protein-protein interactions (Koshland, 1958). Conformational entropy provides the diversity needed to ensure that a limited antibody repertoire can target and eliminate a virtually limitless number of foreign particles. Upon binding to a target that selects for an orientation that strays from the most energetically favorable *apo* conformation, internalization of the BCR-antigen complex, and successful B cell activation, the B cell gains random mutations in the frameworks that may make the required interaction more favorable.

In principle, the entropic contribution in a binding event involving an antibody and an antigen involves changes in the internal conformational entropy of each participant, the entropy of the solvent, and the entropy involved in association. Some antibodies may encounter an antigen that selects for a conformation near the energetic minimum of the *apo* antibody. Affinity maturation in this scenario centers around "tightening" the range of energetically favorable conformations, minimizing the entropic loss associated binding to the antigen. While some new interactions may be made in the HC/LC interface, this concept is consistent with the conformational selection model of protein-protein interactions (Bosshard, 2001; Chakrabarti, 2016), our understanding that germline-encoded antibodies are inherently more flexible, allowing the paratope to explore a wide range of geometries but become more rigid during affinity maturation (Wong, 2011) , and that mitigating the entropic loss upon binding is an effective means of increasing binding affinity (Lafont, 2007).

This study shows that antibodies that undergo a change in range of favorable HC/LC orientations during affinity maturation experience conformational changes that fit the description for both “shift” and “tightening” (**Table S2**). Although we observed that the HC1 angle for CD4BS antibodies is significantly different from the average HC1 angle for non-CD4BS antibodies, we see additional changes in many aspects of orientation between individual antibodies and their germline-reverted counterparts. For instance, docking of the HC/LC interface in VRC-PG04 and its reverted counterpart revealed changes in the HC1 and HC2 angles, as well as a narrowing in the distribution of the HL, HC1, and HC2 angles.

We measured the effects of mutations in the HC/LC interface on antibody/antigen binding affinity, Fv region thermodynamic stability, and relative orientation of the heavy and light chain. Through a combination of *in vitro* BLI kinetic assays and computational experiments, we found that antibody/antigen binding affinity can be increased by inducing mutations in the HC/LC interface that preconfigure the combining site of the antibody. The antibodies used in this study inhabit an unusual orientation; these antibodies neutralize HIV by inserting the HC domain into the CD4-binding site on HIV gp120, causing the heavy chain to pull away from the light chain slightly in order to accommodate the gp120 V5 loop (**Figure II.1**).

There are several limitations with this study. First, we assumed that the germline-reverted heavy and light chain interfaces were compatible with otherwise heavily mutated antibodies. By focusing on reversions in the HC/LC interface of the antibodies used in this study, we created a chimeric antibody that may disrupt heavy and light interactions in an unexpected manner. Second, we assume that Rosetta can find the optimal *apo* conformation during a docking simulation given the bound conformation. Our computational approach only approximates antibody flexibility through the range of orientations that the heavy and light chains can access through small perturbation docking. While a direct comparison of the flexibility for each antibody is preferred, a comprehensive sampling with techniques like nuclear magnetic resonance (Soares, 2004) or molecular dynamics analysis (Margreitter, 2016) through conformational space for a molecule as large as the antibody F_V region is both difficult and computationally expensive. To compensate for this limitation, we adopted a method that truncates the computational time and can parallelize experiments for multiple antibodies. Third, the Rosetta software can only approximate the thermodynamic stability of a structure. The software uses an implicit water model, which may not accurately account for hydrogen bonding networks formed in the interface (Alford, 2017). Water coordination in an interface can contribute to the affinity of the interaction (Marino, 2016), though in the case of the HC/LC interface, the disruption of intermolecular hydrophobic interactions may lead to destabilization; the presence of water in the HC/LC interface may change the optimal HC/LC orientation (Herold, 2017). Finally, the study was limited to the collection of relative K_D of mature and germline-reverted antibodies. While a thermodynamic approach is preferred, the reverted antibodies did not express at a high enough concentration for an accurate determination of enthalpy and entropy through a resource demanding technique like isothermal titration calorimetry.

Our findings suggest that during chronic infections such as HIV, as B cells in germinal centers are exposed to repetitive rounds of somatic mutation of the antigen receptor and positive selection, they incorporate mutations in the heavy and light chain interface that may indirectly improve binding affinity. Though we did not see a universal trend in rigidification, or “tightening”, upon affinity maturation, a recent study by Jeliazkov et al. suggests that affinity maturation of the HCDR3 loop does not always result in rigidification (Jeliazkov, 2018), which is consistent with our results. Additionally, our findings suggest that the typically well-conserved HC/LC interface can tolerate mutations and could serve as a hotspot for engineering antibodies with maximal affinity.

Methods

Selection of antibody/antigen complexes. Every published, human antibody/antigen (protein) complex was collected from the Protein Data Bank. The HC/LC orientation of each complex was assessed using the downloadable version of the ABangle software (Dunbar, 2013) and angle distributions were plotted using the Prism version 7 software (GraphPad). The complexes that comprised the tails of each distribution were examined for mutations in the HC/LC interface, and a subset of complexes that bound the same antigen was selected based on two criteria: 1) the average angle of these complexes must be at least one standard deviation than the average angle across all antibodies in at least one of the six ABangle metrics used to determine orientation, and 2) every antibody in the subset chosen must have at least one amino acid mutation from germline in the HC/LC interface. Residues appropriate for consideration as HC/LC mutants were selected using the Rosetta InterfaceAnalyzer module (Lewis, 2011). These interfaces were inspected manually using PyMOL (version 1.8.4.0) to ensure that the residues in question do not make contact the antigen. Mutations in the HC/LC interface were detected by submitting the nucleotide sequence, found at GenBank ® (Benson, 2012) for the heavy and light chains to the international ImMunoGeneTics information system (Lefranc, 2011) for alignment.

Rigid-body analysis of HC/LC interfaces in gp120-CD4BS co-crystal structures. The PDB files of each co-crystal were altered to isolate the F_V region of the antibody, removing all other components of the antibody/antigen interaction, and Rosetta was used to revert mutated residues in the HC/LC interface. This procedure was accomplished through rigid-body modeling, which prevents perturbation of the backbone and preserves the HC/LC orientation. Minimization was limited to repacking of the amino acids and 100 models were made for each antibody and its HC/LC germline-reverted form using this protocol. Models were ranked according to the total score assigned by Rosetta in Rosetta Energy Units (REU), and the top five models for each antibody were used to calculate the change that these mutations have on the stability of the interface (ΔΔΔG).

Docking and measuring change in HC/LC orientation. We used rigid-body docking to explore the effects of mutations in the HC/LC interface on relative orientation on the heavy and light chains. Briefly, using the RosettaDock algorithm (Chaudhury, 2011), iterative rounds of docking and repacking were performed on the template PDB files. 1,500 models were generated using the protocol, and the top scoring (total REU) 50 models were used for further analysis. ABangle software was used to evaluate the orientation of each model, and the resulting angle distributions were compared using the previously described equations.

Antibodies and gp120 expression. cDNAs encoding antibody heavy and light chains were cloned into IgG expression vectors for mammalian cells (McLean, 2003). The DNA of the two vectors was mixed together at a 1:1 molar ratio and transfected into ExpiCHO cells (Thermo-Fischer) with a 1:1 ratio of DNA to ExpiFectamine CHO (Thermo-Fischer). Antibodies and their variants were cultured at a 1-liter volume, and the supernatant was collected at day 14. Antibodies were purified from supernatant on MabSelect SuRe columns (GE Healthcare). cDNA encoding the HIV envelope protein YU2 gp120 was cloned into the pCDNA3.4 plasmid vector and transfected into 293-F HEK cells, as previously described (Willis, 2015).

Biolayer interferometry (BLI). The binding affinity for the panel of mature and interface reverted CD4BS IgG to monomeric YU2 gp120 core was determined by BLI using an Octet RED96 instrument (Pall FortéBio, CA, USA). The antibodies were diluted in PBS with 0.05% Tween 20 at pH 7.4, then captured using anti-human IgG Fc capture (AHC) tips. Testing using a series of YU2 gp120 protein concentrations (60, 30 or 15 µg/mL) was used to calculate the equilibrium dissociation constant (K_D). An unliganded sensor (devoid of CD4BS IgG) was used as a reference sensor in order to correct for non-specific gp120 binding. The binding traces were processed using the FortéBio Data Analysis Software v9.0, and the processed binding curves were fitted using the "Heterogeneous Ligand 2:1 interaction" model.

CHAPTER III

Conclusions and future directions

Concluding remarks

Antibody affinity maturation is an integral part of the adaptive immune system. During a chronic infection or persistent disease like HIV, somatic mutations are accumulated not only in the interface between antibody and viral protein but also throughout the interface between the antibody's heavy and light chain (HC/LC interface). We find that these mutations can radically change the relative orientation of the variable domains, preconfiguring the paratope for its bound conformation. Additionally, these mutations indirectly affect binding affinity; for example, CD4BS antibodies with a germline-reverted HC/LC interface have a dampened binding affinity to HIV YU2 gp120 when compared to their mature antibody counterparts. I hypothesized that these mutations induce alternate orientations through a combination of shifting the energetic minimum and limiting the amount of conformational entropy by stabilizing the HC/LC interface in a way that favours the bound conformation.

This study began by making an effort to determine the conformational diversity of bound antibodies found in all the human, antibody-antigen co-crystals represented in the Protein Data Bank. Prior to this study, we knew that the difference in HC/LC orientation between the apo and bound conformations of an antibody differed by ~5 degrees when comparing the HL value as long as the antibodies in question bind proteins (Dunbar, 2013). This suggests that most (if not all) protein-binding, antibody/antigen interactions include some re-orientation of the heavy chain with respect to the light chain. The study presented in this thesis focuses on antibodies against the HIV gp120 CD4 binding site. The reason for this is fourfold: **1)** These antibodies are heavily mutated; reverting the mutations in the interface back to their germline residue is less likely to cause dissociation of the HC/LC interface than re-designing a germline interface as both mature and reverted interfaces are known to have existed in nature; **2)** these antibodies have been well-studied it was known that a blanket reversion of all antigen-distal somatic mutations disrupts binding and neutralization (Klein, 2013); **3)** in order to bind the CD4 binding site, these antibodies inhabit a unique orientation that resolves clashes with V5 loop (Fera, 2014); **4)** high-resolution co-crystal structures depicting the antibodies in complex with gp120 were readily available at the Protein Data Bank. These conditions were optimal for the study, but it is unclear whether mutations in the HC/LC interface affect all antibodies in the same way. Given the diversity of the results in Chapter II, it is likely that the "potency" associated with mutations that increase affinity allosterically is conditional. Even amongst the CD4 antibodies studied in this thesis, we see a mixture of shift and tightening during maturation. This may be caused by subtle differences between these antibodies in CD4 buried surface area (**Figure III.1**) and types of interactions with gp120 (**Figure III.2**). This raises one important future questions to be studied: Are these observations, i.e. the shift and tightening in orientation that was observed in Chapter II, translatable to other antibody/antigen interactions?

As a ternary complex, antibody/antigen interactions can be divided into two categories concerning the mode of binding. The first, which consists of antibodies that only interact with antigen using the heavy chain, can take a variety of forms. For example, the interaction between influenza hemagglutinin (HA1) head domain and the anti-influenza antibody C05 is mediated entirely by the HCDR3 (Ekiert, 2012), while the HA stem domain targeting antibody CR6261 interacts with the antigen using HCDR1, HCDR2, and HCDR3. These antibodies differ in the

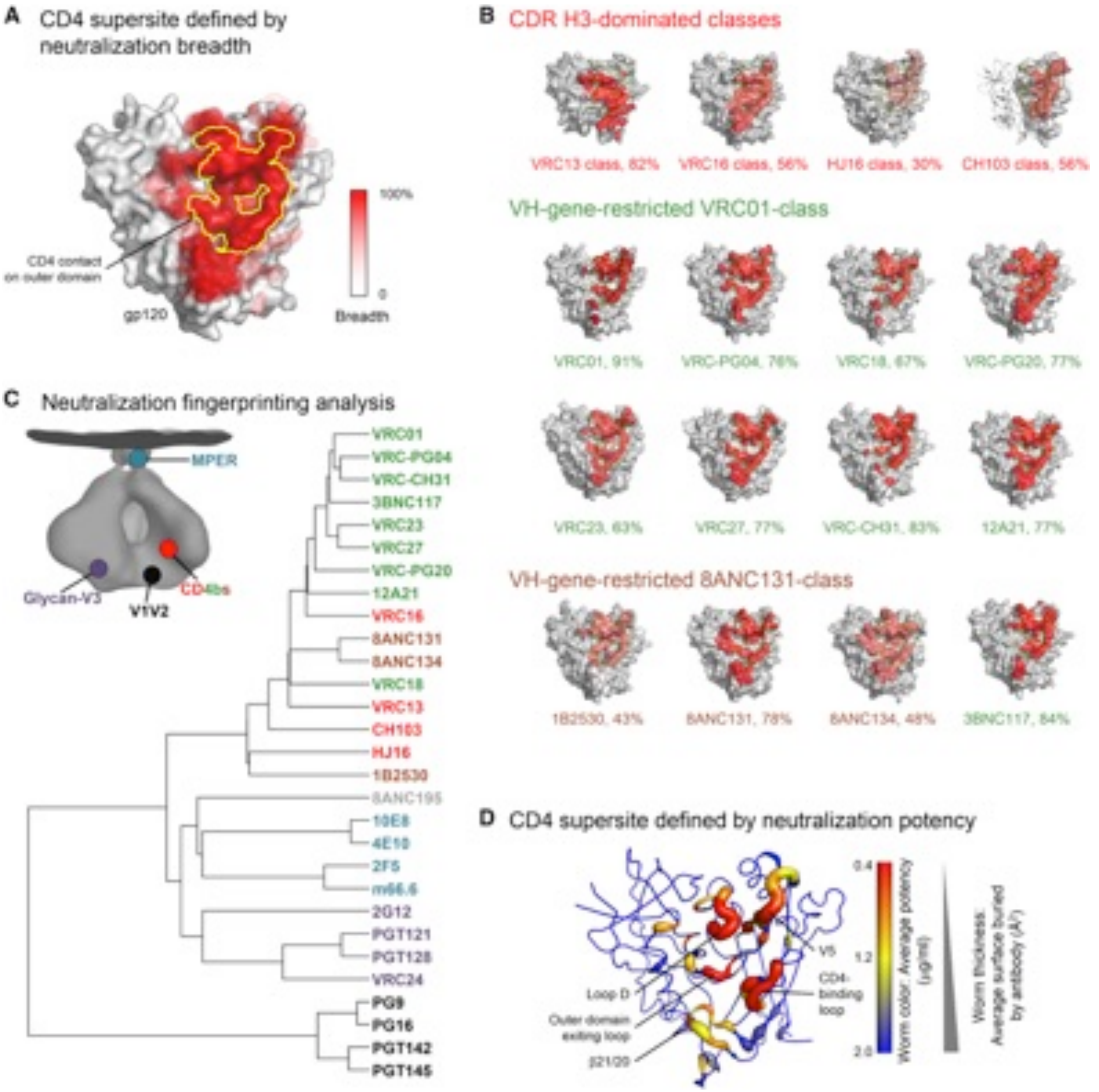


Figure III.1 The CD4 supersite. **A)** Antibodies from 14 donors define an immunological supersite of HIV-vulnerability. A composite of how the breadth-coded epitope surfaces shown in B are mapped to the gp120 surface. The yellow outline defines the outer-domain contact of the CD4 receptor. **B)** Epitopes of CD4bs antibodies colored by breadth. **C)** Dendrogram constructed from similarities in neutralization fingerprint based on serologic analysis with a 178-virus panel; insert shows the HIV-1 viral spike, with membrane at top, with major epitopes labeled; epitope colors correspond to antibody colors in the dendrogram. **D)** Potency of CD4-binding site antibodies mapped to the supersite. The worm representation of HIV-1 gp120 is colored by averaged antibody potency, with thickness representing average buried binding surface area of corresponding residues; notably, in addition to the outer domain contact on gp120 for CD4, neighboring regions in the inner domain and on strands β20/21 contribute to the supersite. Adapted from (Zhao, 2015).

engagement of the light chain to the heavy chain; CR6261 moves the light chain out of the way to prevent clashes with glycosylation at the base of the helix (Ekiert, 2012), while the C05 light chain is 12 Å from the antigen in its bound conformation . The second mode of binding uses both heavy and light chains to bind to the antigen. The paratope of CD4BS antibodies like 8ANC131 includes both heavy and light chains, but the degree to which light chains are incorporated into the paratope is highly variable, even among antibodies that target the same epitope due to small differences in binding orientation (Huang, 2016).

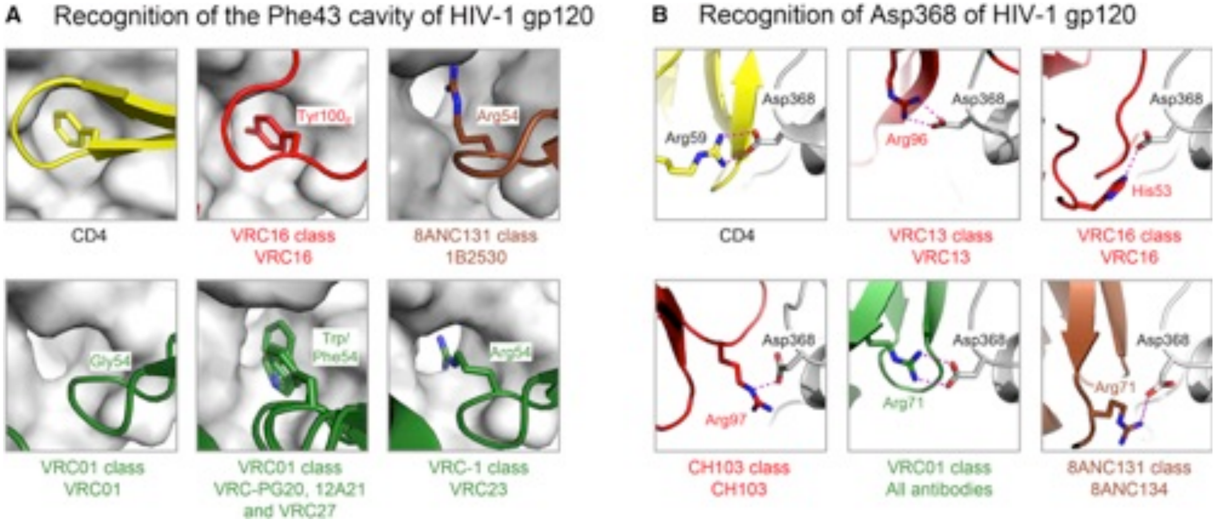


Figure III.2 Paratopes of Effective CD4bs Antibodies Are Extremely Diverse **A)** Antibodies from 14 donors define an immunological supersite of HIV-vulnerability. A composite of how the breadth-coded epitope surfaces shown in B are mapped to the gp120 surface. The yellow outline defines the outer-domain contact of the CD4 receptor. **B)** Epitopes of CD4bs antibodies colored by breadth. **C)** Dendrogram constructed from similarities in neutralization fingerprint based on serologic analysis with a 178-virus panel; insert shows the HIV-1 viral spike, with membrane at top, with major epitopes labeled; epitope colors correspond to antibody colors in the dendrogram. **D)** Potency of CD4-binding site antibodies mapped to the supersite. The worm representation of HIV-1 gp120 is colored by averaged antibody potency, with thickness representing average buried binding surface area of corresponding residues; notably, in addition to the outer domain contact on gp120 for CD4, neighboring regions in the inner domain and on strands β20/21 contribute to the supersite

These differences in binding modes likely alter how mutation in the HC/LC interface might affect binding affinity. Antibodies like C05 do not need to shift the HC/LC orientation, as binding is mediated entirely by HCDR3; antigen distal mutations in the HC/LC interface are likely geared toward mitigating entropic loss. In contrast, CR6261 must rotate its light chain in order to engage its target optimally (Ekiert, 2012), and mutations in the HC/LC might contribute more to shifting the optimal orientation in order to interact with the epitope (increasing enthalpic gain). Antibodies that utilize both heavy and light chains, however, may require a combination of shifting and tightening like the panel of CD4BS antibodies in Chapter II.

Additionally, the conformational selection model of protein-protein interactions is not limited to antibody/antigen interactions; flexible, conformationally diverse interactions may benefit from mutations that mitigate the entropic loss upon binding. A recent study by Li et al. suggests that an alignment of the conformational entropy of the partner proteins is one of the most important determinants of protein-protein interaction (Li, 2019). The authors use molecular dynamics to interrogate the interactions between HIV gp120 and host CD4 in an effort to determine the effects of CD4 binding on the conformational entropy and molecular motions of gp120. They found that the association of gp120 and CD4 greatly reduces conformational fluctuations of gp120 while simultaneously increasing the stability of the bound conformation. Additionally, the gp120/CD4 interaction greatly restricts the movement of the V1/V2 loops, preventing gp120 from returning to its closed, unbound state. In this case, the authors conclude that the development of small molecules to lock in the “open” conformation of gp120 would aid in viral neutralization, as the interaction between CD4 and gp120 is completely dominated by conformational selection. One future direction of this research could be to take this temporal aspect of dynamics into account as the study presented in this thesis exclusively modelled the structural dynamics excluding the time scale.

An additional study that can be performed with existing technology involves creating and characterizing antibodies with a “synthetic interface”, wherein the HC/LC is re-designed using the mutated positions as mutational hotspots. Using the mature antibody in the bound conformation, Rosetta can search for the “optimal” sequence of mutations in those positions. Should Rosetta design a better (or different) HC/LC interface than the mature antibody, the *in silico* maturation results will result in models with angle distributions that resemble the mature antibodies’ ranges of orientation, as the stabilizing effects preconfigure the antibody for the bound conformation. In order to determine how stabilizing mutations might affect the binding affinity and thermodynamic profile of an antibody with an HC/LC orientation closer to the means of distributions for all human antibody co-crystal structures, it is imperative that *in silico* maturation is enacted on a variety of non-HIV interactions that reflect the “average” bound antibody structure in the PDB. After expressing and testing these Rosetta-generated mutants, we should have a clearer picture of the ternary nature of antibody/antigen interactions.

References

Abhinandan KR, Martin AC. Analysis and prediction of VH/VL packing in antibodies. *Protein Eng Des Sel*. 2010 Sep;23(9):689-97

Adolf-Bryfogle J, Kalyuzhniy O, Kubitz M, et al. RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *PLoS Comput Biol*. 2018;14(4):e1006112. Published 2018 Apr 27. doi:10.1371/journal.pcbi.1006112

Alberts B, Johnson A, Lewis J, et al. The Adaptive Immune System. In: Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21070/

Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol.* 1997 Nov 7;273(4):927-48

Alford RF, Leaver-Fay A, Jeliazkov JR, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*. 2017;13(6):3031-3048. doi:10.1021/acs.jctc.7b00125.

Allende M. L., G. Tuymetova, B. G. Lee, E. Bonifacino, Y. P. Wu, R. L. Proia. 2010. S1P1 receptor directs the release of immature B cells from bone marrow into blood. *J. Exp. Med*. 207: 1113–1124.

Allman D, J. Li J, Hardy RR. 1999. Commitment to the B lymphoid lineage occurs before DH-JH recombination. *J. Exp. Med*. 189: 735–740.

Alt FW, Baltimore D. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. *Proc Natl Acad Sci U S A*. 1982;79(13):4118-22.

Arya R, Bassing CH. V(D)J Recombination Exploits DNA Damage Responses to Promote Immunity. *Trends Genet*. 2017;33(7):479-489.

Banerjee A, Sindhava V, Vuyyuru R, et al. YY1 Is Required for Germinal Center B Cell Development. *PLoS One*. 2016;11(5):e0155311. Published 2016 May 11. doi:10.1371/journal.pone.0155311

Barreto, VM, Magor,BG. Activation-induced cytidine deaminase structure and functions: A species comparative view. *Developmental & Comparative Immunology*. 2011; 35(9): 991-1007, https://doi.org/10.1016/j.dci.2011.02.005.

Bar-On Y, Gruell H, Schoofs T, Pai JA1 Nogueira L, Butler AL, Millard K, Lehmann C, Suárez I, Oliveira TY, Karagounis T, Cohen YZ, Wyen C, Scholten S, Handl L, Belblidia S, Dizon JP, Vehreschild JJ, Witmer-Pack M, Shimeliovich I, Jain K, Fiddike K, Seaton KE, Yates NL, Horowitz J, Gulick RM, Pfeifer N, Tomaras GD, Seaman MS, Fätkenheuer G, Caskey M, Klein F, Nussenzweig MC. Safety and anti-viral activity of combination HIV-1 broadly neutralizing antibodies in viremic individuals. *Nat Med*. 2018 24(11): 1701–1707.

Beck A, Wurch T, Bailly C, Corvaia N. Strategies and challenges for the next generation of therapeutic antibodies. *Nat Rev Immunol*. 2010;10(5):345-52. doi: 10.1038/nri2747.

Bender BJ, Cisneros A, Duran AM, et al. Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. *Biochemistry*. 2016;55(34):4748-63.

Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res*. 2012;41(Database issue):D36-42.

Bosshard HR. Molecular recognition by induced fit: How fit is the concept? *News Phys Sci*. 2001;16:171–173.

Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science*. 2005 16;309(5742):1868-71.

Budeus B, Schweigle de Reynoso S, Przekopowitz M, Hoffmann D, Seifert M, Küppers R. Complexity of the human memory B-cell compartment is determined by the versatility of clonal diversification in germinal centers. *Proc Natl Acad Sci U S A*. 2015;112(38):E5281-9.

Bujotzek A, Lipsmeier F, Harris SF, Benz J, Kuglstatter A, Georges G. VH-VL orientation prediction for antibody humanization candidate selection: A case study. *MAbs*. 2015;8(2):288-305.

Burton DR, Stanfield RL, Wilson IA. Antibody vs. HIV in a clash of evolutionary titans. *Proc Natl Acad Sci U S A*. 2005;102(42):14943-8.

Campana D, Janossy G, Bofill M, Trejdosiewicz LK, Ma D, Hoffbrand AV, Mason DY, Lebacq AM, Forster HK. Human B cell development. I. Phenotypic differences of B lymphocytes in the bone marrow and peripheral lymphoid tissue. *J Immunol*. 1985;134(3):1524-30

Chailyan A, Marcatili P, Tramontano A. The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *The Febs Journal*. 2011;278(16):2858-2866. doi:10.1111/j.1742-4658.2011.08207.x.

Chakrabarti KS, Agafonov RV, Pontiggia F, et al. Conformational Selection in a Protein-Protein Interaction revealed by Dynamic Pathway Analysis. *Cell reports*. 2016;14(1):32-42. doi:10.1016/j.celrep.2015.12.010.

Chames P, Baty D. Bispecific antibodies for cancer therapy: the light at the end of the tunnel? *mAbs*. 2009;1(6):539-547.

Chaudhury S, Berrondo M, Weitzner BD, Muthu P, Bergman H, Gray JJ. Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. Uversky VN, ed. *PLoS ONE*. 2011;6(8):e22477. doi:10.1371/journal.pone.0022477.

Chetallier J, Van Regenmortel MH, Vernet T, Altschuh D. Functional Mapping of conserved residues located at the VL and VH domain interface of a Fab. *J Mol Biol*. 1996;264(1):1-6.

Chang X, Li B, Rao A. RNA-binding protein hnRNPLL regulates mRNA splicing and stability during B-cell to plasma-cell differentiation. *Proc Natl Acad Sci U S A*. 2015;112(15):E1888-97.

Chothia C, Novotný J, Bruccoleri R, Karplus M. Domain association in immunoglobulin molecules. The packing of variable domains. *J Mol Biol*. 1985 5;186(3):651-63.

Cisneros 3rd A, Nargi RS, Parrish EH, Haliburton CM, Meiler J, Crowe Jr. JE. Role of antibody heavy and light chain interface residues in affinity maturation of binding to HIV envelope glycoprotein. *Mol. Syst. Des. Eng***.**, 2019; Advance Article

Combs SA, Deluca SL, Deluca SH, et al. Small-molecule ligand docking into comparative models with Rosetta. *Nat Protoc*. 2013;8(7):1277-98.

Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci*. 2013;23(1):47-55.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. *Genome Research*. 2004;14(6):1188-1190. doi:10.1101/gr.849004.

Da Silva GF, Harrison JS, Lai JR. Contribution of light chain residues to high affinity binding in an HIV-1 antibody explored by combinatorial scanning mutagenesis. *Biochemistry*. 2010;49(26):5464-72.

Davies DR, Metzger H. Structural basis of antibody function. *Annu Rev Immunol*. 1983;1:87-117.

DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter On Protein Crystallography*. 2002; 40, 82-92.

Dondelinger M, Filée P, Sauvage E, et al. Understanding the Significance and Implications of Antibody Numbering and Antigen-Binding Surface/Residue Definition. *Front Immunol*. 2018;9:2278. Published 2018 Oct 16. doi:10.3389/fimmu.2018.02278

Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*. 2016;32(2):298-300. doi:10.1093/bioinformatics/btv552.

Dunbar J, Fuchs A, Shi J, Deane CM. ABangle: characterizing the VH-VL orientation in antibodies. *Protein Eng Des Sel*. 2013;26(10):611-620. https://doi.org/10.1093/protein/gzt020

Dunbar J, Knapp B, Fuchs A, Shi J, Deane CM. Examining variable domain orientations in antigen receptors gives insight into TCR-like antibody design. *PLoS Comput Biol*. 2014;10(9):e1003852. Published 2014 Sep 18. doi:10.1371/journal.pcbi.1003852

Ekiert DC, Bhabha G, Elsliger MA, et al. Antibody recognition of a highly conserved influenza virus epitope. *Science*. 2009;324(5924):246-51.

Ekiert DC, Kashyap AK, Steel J, et al. Cross-neutralization of influenza A viruses mediated by a single antibody loop. *Nature*. 2012;489(7417):526-32.

Feige MJ, Hendershot LM, Buchner J. How antibodies fold. *Trends Biochem Sci*. 2009;35(4):189-98.

Fera D, Schmidt AG, Haynes BF, et al. Affinity maturation in an HIV broadly neutralizing B cell lineage through reorientation of variable domains. *Proc Natl Acad Sci U S A*. 2014;111(28):10275-80.

Fleishman SJ, Whitehead TA, Ekiert DC, et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*. 2011;332(6031):816-21.

Fritz JM, Weaver TE. The BiP cochaperone ERdj4 is required for B cell development and function. *PLoS One*. 2014;9(9):e107473. Published 2014 Sep 15. doi:10.1371/journal.pone.0107473

Fröland SS, Natvig JB. Lymphocytes with membrane-bound immunoglobulin (B-lymphocytes) in new-born babies. *Clin Exp Immunol*. 1972;11(4):495-505.

Geiger-Schuller K, Sforza K, Yuhas M, Parmeggiani F, Baker D, Barrick D. Extreme stability in de novo-designed repeat arrays is determined by unusually stable short range interactions. *PNAS*. 2018;(115(29): 7539-7544. DOI: 10.1073/pnas.1800283115

Georgiev IS, Rudicell RS, Saunders KO, et al. Antibodies VRC01 and 10E8 neutralize HIV-1 with high breadth and potency even with immunoglobulin-framework regions substantially reverted to germline. *Journal of immunology* (Baltimore, Md : 1950). 2014;192(3):1100-1106. doi:10.4049/jimmunol.1302515.

Glanville J, Zhai W, Berka J, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *PNAS*. 2009;106(48):20216-20221. doi:10.1073/pnas.0909775106.

Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D, Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Mol. Biol*. 2003; 331(1): 281-299. https://doi.org/10.1016/S0022-2836(03)00670-3.

Hardy R. R, Hayakawa K. B Cell Development Pathways. *Ann Rev of Immunol*. 2001;19(1): 595-621.

Herold EM, John C, Weber B, et al. Determinants of the assembly and function of antibody variable domains. *Scientific Reports*. 2017;7:12276. doi:10.1038/s41598-017-12519-9.

Hoffman W, Lakkis FG, Chalasani G. B Cells, Antibodies, and More. *Clin J Am Soc Nephrol*. 2015;11(1):137-54.

Honegger A, Plückthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol*. 2001;309(3):657-70.

Hsu H-J, Lee KH, Jian J-W, et al. Antibody Variable Domain Interface and Framework Sequence Requirements for Stability and Function by High-Throughput Experiments. *Structure*. 2014;22(1):22-34. doi:10.1016/j.str.2013.10.006

Huang J, Kang BH, Ishida E, et al. Identification of a CD4-Binding-Site Antibody to HIV that Evolved Near-Pan Neutralization Breadth. *Immunity*. 2016;45(5):1108-1121.

Hugo N, Weidenhaupt M, Beukes M, Xu B, Janson JC, Vernet T, Altschuh D. VL position 34 is a key determinant for the engineering of stable antibodies with fast dissociation rates. *Protein Eng*. 2003; 16(5):381-6.

Hwang JK, Alt FW, Yeap L-S. Related Mechanisms of Antibody Somatic Hypermutation and Class Switch Recombination. *Microbiol spectrum*. 2015;3(1):10.1128/microbiolspec.MDNA3-0037-2014. doi:10.1128/microbiolspec.MDNA3-0037-2014.

Peripheral tolerance. *Immunology*. 1994;83(Suppl 1):44–47.

Janeway CA Jr, Travers P, Walport M, et al. The generation of diversity in immunoglobulins. In :Immunobiology: The Immune System in Health and Disease. 5th edition. New York: Garland Science; 2001.

Jeliazkov JR, Sljoka A, Kuroda D, et al. Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification. *Front Immunol*. 2018;9:413. Published 2018 Mar 2. doi:10.3389/fimmu.2018.00413

Jeske DJ, Jarvis J, Milstein C, Capra JD. Junctional diversity is essential to antibody activity. *J Immunol*. 1984;133(3):1090-2.

Jones, JM, Gellert, M, The taming of a transposon: V(D)J recombination and the immune system. *Immunol Rev*, 2004;200: 233-248. doi:10.1111/j.0105-2896.2004.00168.x

Julian MC, Li L, Garde S, Wilen R, Tessier PM. Efficient affinity maturation of antibody variable domains requires co-selection of compensatory mutations to maintain thermodynamic stability. *Scientific Reports*. 2017;7:45259. doi:10.1038/srep45259.

Kellis JT Jr, Nyberg K, Sali D, Fersht AR. Contribution of hydrophobic interactions to protein stability. *Nature*. 1988;333(6175):784-6.

Klein F, Diskin R, Scheid JF, et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell*. 2013;153(1):126-38.

Koenig P, Lee CV, Walters BT, et al. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proc Natl Acad Sci U S A*. 2017;114(4):E486-E495.

Koonin EV, Krupovic M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat Rev Genet*. 2014;16(3):184-92.

Koshland DE. Application of a Theory of Enzyme Specificity to Protein Synthesis. *PNAS*. 1958;44(2):98-104.

Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A*. 2000;97(19):10383-8.

Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. *Science*. 2003 21;302(5649):1364-8.

Lafont V, Armstrong AA, Ohtaka H, Kiso Y, Mario Amzel L, Freire E. Compensating enthalpic and entropic changes hinder binding affinity optimization. *Chem Biol Drug Des*. 2007;69(6):413-22.

Leaver-Fay A, Jacak R, Stranges PB, Kuhlman B. A generic program for multistate protein design. *PLoS One*. 2011;6(7):e20937.

Leavy O. Therapeutic antibodies: past, present and future. *Nat Rev Immunol*. 2010; 10(5): 297. doi: 10.1038/nri2763

Lefranc MP. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc*. 2011;2011(6). pii: pdb.top115. doi: 10.1101/pdb.top115. PMID: 21632786 LIGM:382

Lewis SM, Kuhlman BA. Anchored Design of Protein-Protein Interfaces. Uversky VN, ed. *PLoS ONE*. 2011;6(6):e20872. doi:10.1371/journal.pone.0020872.

Li Y, Deng L, Yang LQ, Sang P, Liu SQ. Effects of CD4 Binding on Conformational Dynamics, Molecular Motions, and Thermodynamics of HIV-1 gp120. *Int J Mol Sci*. 2019;20(2):260. Published 2019 Jan 10. doi:10.3390/ijms20020260

Liao HX, Chen X, Munshaw S, et al. Initial antibodies binding to HIV-1 gp41 in acutely infected subjects are polyreactive and highly mutated. *J Exp Med*. 2011;208(11):2237-49.

Luning Prak ET, Monestier M, Eisenberg RA. B cell receptor editing in tolerance and autoimmunity. *Ann N Y Acad Sci*. 2011;1217:96-121.

Malu S, De Ioannes P, Kozlov M, et al. Artemis C-terminal region facilitates V(D)J recombination through its interactions with DNA Ligase IV and DNA-PKcs. *J Exp Med*. 2012;209(5):955-63.

Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods*. 2009;6(8):551-2.

Margreitter C, Mayrhofer P, Kunert R, Oostenbrink C. Antibody humanization by molecular dynamics simulations—in-silico guided selection of critical backmutations. *J Mol Recog*. 2016;29(6):266-275. doi:10.1002/jmr.2527.

Marino SF, Olal D, Daumke O. A complex water network contributes to high-affinity binding in an antibody–antigen interface. *Data in Brief*. 2016;6:394-397. doi:10.1016/j.dib.2015.12.023.

Marze NA, Lyskov S, Gray JJ. Improved prediction of antibody VL-VH orientation. *PEDS*. 2016;29(10):409-418.

Masuda K, Sakamoto K, Kojima M, Aburatani T, Ueda T, Ueda H. The role of interface framework residues in determining antibody V(H)/V(L) interaction strength and antigen-binding affinity. *FEBS J*. 2006; 273(10):2184-94.

Maul RW, Gearhart PJ. AID and somatic hypermutation. *Adv Immunol*. 2010;105:159-91.

Maul RW, Gearhart PJ. Controlling somatic hypermutation in immunoglobulin variable and switch regions. *Immunol Res*. 2010;47(1-3):113-22.

McLean GR, Nakouzi A, Casadevall A, Green NS. Human and murine immunological cassettes. Mol *Immunol*. 2000;37(14):837-45.

Mishra AK, Mariuzza RA. Insights into the Structural Basis of Antibody Affinity Maturation from Next-Generation Sequencing. *Front Immunol*. 2018;9:117. Published 2018 Feb 1. doi:10.3389/fimmu.2018.00117

Motea, E. A., & Berdis, A. J. Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochimica et biophysica acta*. 2009; 1804(5), 1151-66.

Mouquet H, Scheid JF, Zoller MJ, et al. Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation. *Nature*. 2010;467(7315):591-5.

Nemazee D. Mechanisms of central tolerance for B cells. *Nat Rev Immunol*. 2017;17(5):281-294.

Nivón LG, Moretti R, Baker D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. Zhang Y, ed. *PLoS ONE*. 2013;8(4):e59004. doi:10.1371/journal.pone.0059004.

North B, Lehmann A, Dunbrack RL. A new clustering of antibody CDR loop conformations. *J Mol Biol*. 2010;406(2):228-56.

Norn CH, Lapidoth G, Fleishman SJ. High-accuracy modeling of antibody structures by a search for minimum-energy recombination of backbone fragments. *Proteins*. 2016;85(1):30-38.

Nutt SL,Heavey B, Rolink AG, Busslinger M. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature*. 1999; 401: 556–562.

Ofek G, McKee K, Yang Y, et al. Relationship between Antibody 2F5 Neutralization of HIV-1 and Hydrophobicity of Its Heavy Chain Third Complementarity-Determining Region. *Journal of Virology*. 2010;84(6):2955-2962. doi:10.1128/JVI.02257-09.

Pelanda R, Torres RM. Central B-cell tolerance: where selection begins. *Cold Spring Harb Perspect Biol*. 2012;4(4):a007146. Published . doi:10.1101/cshperspect.a007146

Peng HP, Lee KH, Jian JW, Yang AS. Origins of specificity and affinity in antibody-protein interactions. *Proc Natl Acad Sci U S A*. 2014;111(26):E2656-65.

Poosarla VG, Li T, Goh BC, Schulten K, Wood TK, Maranas CD. Computational de novo design of antibodies binding to a peptide with high affinity. *Biotechnol Bioeng*. 2017;114(6):1331-1342.

Rao SP, Riggs JM, Friedman DF, Scully MS, LeBien TW, Silberstein LE. Biased VH gene usage in early lineage human B cells: evidence for preferential Ig gene rearrangement in the absence of selection. *J Immunol*. 1999; 163(5):2732-40.

Roth DB. V(D)J Recombination: Mechanism, Errors, and Fidelity. *Microbiol Spectr*. 2014;2(6):10.1128/microbiolspec.MDNA3-0041-2014.

Sakaguchi N, Melchers F. Lambda 5, a new light-chain-related locus selectively expressed in pre-B lymphocytes. *Nature*. 1986; 324:579–82. doi: 10.1038/324579a0

Sangha AK, Dong J, Williamson L, et al. Role of Non-local Interactions between CDR Loops in Binding Affinity of MR78 Antibody to Marburg Virus Glycoprotein. *Structure*. 2017;25(12):1820-1828.e2.

Scheid JF, Horwitz JA, Bar-On Y, et al. HIV-1 antibody 3BNC117 suppresses viral rebound in humans during treatment interruption. *Nature*. 2016;535(7613):556-60.

Schroeder HW, Cavacini L. Structure and function of immunoglobulins. *J Allergy Clin Immunol*. 2010;125(2 Suppl 2):S41-52.

Sevy AM, Wu NC, Gilchuk IM, et al. Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses. *Proc Natl Acad Sci U S A*. 2019;116(5):1597-1602.

Sivasubramanian A, Sircar A, Chaudhury S, Gray JJ. Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins*. 2009;74(2):497-514.

Sheriff S, Silverton EW, Padlan EA, et al. Three-dimensional structure of an antibody-antigen complex. *Proc Natl Acad Sci U S A*. 1987;84(22):8075-9.

Simonelli L, Pedotti M, Beltramello M, et al. Rational engineering of a human anti-dengue antibody through experimentally validated computational docking. *PLoS One*. 2013;8(2):e55561.

Smith K, Shah H, Muther JJ, Duke AL, Haley K, James JA. Antigen nature and complexity influence human antibody light chain usage and specificity. *Vaccine*. 2016;34(25):2813-20.

Soares MR, Bisch PM, Campos de Carvalho AC, Valente AP, Almieda FCL. Correlation between conformation and antibody binding: NMR structure of cross-reactive peptides from T. cruzi, human and L. braziliensis. 2004;560(1-3):134-140. https://doi.org/10.1016/S0014-5793(04)00088-2

Stanfield RL, Takimoto-Kamimura M, Rini JM, Profy AT, Wilson IA. Major antigen-induced domain rearrangements in an antibody. *Structure*. 1993 Oct 15;1(2):83-93.

Stavnezer J, Amemiya CT. Evolution of isotype switching. 2004; 16(4):257-75.

Thornburg NJ, Nannemann DP, Blum DL, et al. Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. *J Clin Invest*. 2013;123(10):4405-9.

Tiegs SL, Russell DM, Nemazee D. Receptor editing in self-reactive bone marrow B cells. *J Exp Med*. 1993;177(4):1009-20.

Tiller KE, Chowdhury R, Li T, et al. Facile Affinity Maturation of Antibody Variable Domains Using Natural Diversity Mutagenesis. Frontiers in Immunology. 2017;8:986. doi:10.3389/fimmu.2017.00986.

Tyka MD, Keedy DA, André I, et al. Alternate states of proteins revealed by detailed energy landscape mapping. *J Mol Biol*. 2010;405(2):607-18.

Vangone A, Abdel-Azeim S, Caputo I, et al. Structural basis for the recognition in an idiotype-anti-idiotype antibody complex related to celiac disease. *PLoS One*. 2014;9(7):e102839. Published 2014 Jul 30. doi:10.1371/journal.pone.0102839

Victora GD, Mesin L. Clonal and cellular dynamics in germinal centers. *Curr Opin Immunol*. 2014;28:90-6.

Wang F, Ekiert DC, Ahmad I, et al. Reshaping antibody diversity. *Cell*. 2013;153(6):1379-93.

Wang RY, Song Y, Barad BA, Cheng Y, Fraser JS, DiMaio F. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *Elife*. 2016;5:e17219. Published 2016 Sep 26. doi:10.7554/eLife.17219

Weitzner BD, Jeliazkov JR, Lyskov S, et al. Modeling and docking of antibody structures with Rosetta. *Nat Protoc*. 2017;12(2):401-416.

Willis JR, Briney BS, DeLuca SL, Crowe JE, Meiler J. Human Germline Antibody Gene Segments Encode Polyspecific Antibodies. Peters B, ed. *PLoS Comp Bio*. 2013;9(4):e1003045. doi:10.1371/journal.pcbi.1003045.

Willis JR, Sapparapu G, Murrell S, et al. Redesigned HIV antibodies exhibit enhanced neutralizing potency and breadth. *The Journal of Clinical Investigation*. 2015;125(6):2523-2531. doi:10.1172/JCI80693.

Wong SE, Sellers BD, Jacobson MP. Effects of somatic mutations on CDR loop flexibility during affinity maturation. *Proteins*. 2011;79(3):821-9. doi: 10.1002/prot.22920.

Xu H, Schmidt AG, O'Donnell T, et al. Key mutations stabilize antigen-binding conformation during affinity maturation of a broadly neutralizing influenza antibody lineage. *Proteins*. 2015;83(4):771-780. doi:10.1002/prot.24745.

Zhang M, Srivastava G, Lu L. The pre-B cell receptor and its function during B cell development. *Cell Mol Immunol*. 2004;1: 89–94.

Zhou T, Georgiev I, Wu X, et al. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science*. 2010;329(5993):811-7.

Zhou T, Lynch RM, Chen L, et al. Structural Repertoire of HIV-1-Neutralizing Antibodies Targeting the CD4 Supersite in 14 Donors. *Cell*. 2015;161(6):1280-92.

APPENDIX

Protocol capture for Chapter II

Introduction

The following is a protocol capture that demonstrates how to determine the stability at the HC/LC interface using an antibody/antigen co-crystal and generate an ensemble of HC/LC docked models. In this example, we will be making models of the antibody VRC-PG04.

The version of Rosetta used for the entirety of this study is: Rosetta_2015.12.57698, released on May 5th, 2015.

All input materials for this protocol capture can be downloaded from https://github.com/ac1546/HC_LC_docking.

The ABangle software can be found at http://www.stats.ox.ac.uk/~dunbar/abangle/.

Preparing input structures

The PDB structure 3se9 was downloaded from the Protein Data Bank (https:www.rcsb.org/) and processed manually in PyMol. The gp120 component (Chain G), waters, and salt ions were removed. Next, we remove the constant region of the Fab in order to lessen the time needed to generate models. Here, we removed residues 113-216 of the heavy chain and residues 108-214 of the light chain. The molecule was saved as 3se9_Fv_clean.pdb to denote the type of fragment the pdb contains and whether or not this contains atoms that Rosetta cannot process.

Defining the HC/LC interface

Here, we use the InterfaceAnalyzer application to define an HC/LC interface using the following command:

```
/path_to_rosetta/rosetta/main/source/bin/InterfaceAnalyzer.linuxgccrelease -s
3se9_Fv_clean.pdb -tracer_data_print true -pack_input true -pack_separated true -
score:weights talaris2013.wts
```

Near the end of the output is a PyMol selection defining the residues that comprise the HC/LC interface:

```
select 3se9_Fv_clean_interface,
/3se9_Fv_clean//H/1+3+4+6+35+37+39+43+44+45+46+47+48+49+50+57+58+59+89+
91+92+93+94+99+100+100+100+100+100+100+100+101+102+103+104+105+106+1
08+ +
/3se9_Fv_clean//L/31+32+33+34+35+36+38+41+42+43+44+45+46+47+48+49+50+51
+52+53+55+56+57+58+85+87+89+90+91+96+97+98+99+100+101+
```

In order to determine which of the HC/LC interface residues do not interact with the antigen, this selection is modified so that it works with the unmodified 3se9 structure:

```
    select 3se9_Fv_clean_interface, /3se9
//H/1+3+4+6+35+37+39+43+44+45+46+47+48+49+50+57+58+59+89+91+92+93+94+
99+100+100+100+100+100+100+100+101+102+103+104+105+106+108+ + /3se9
//L/31+32+33+34+35+36+38+41+42+43+44+45+46+47+48+49+50+51+52+53+55+56
+57+58+85+87+89+90+91+96+97+98+99+100+101+
```

To ensure that only antigen-distal residues are considered, the paratope residues are defined using the following commands:

```
    select paratope, byres(chain H+L within 5.5 of chain G)
    color red, paratope
```

Next, we identified mutations in the HC/LC interface. To do this, we downloaded the nucleotide sequences for VRC-PG04 from GenBank (accession numbers **JN159466.1** – light chain, and **JN159464.1** – heavy chain). The nucleotide sequences are then entered into IMGT V-Quest (http://www.imgt.org/IMGT_vquest/vquest), and the mutations from germline in the HC/LC interface were identified manually using the resulting alignments. Additionally, we only selected for mutations whose side chains face the interface. The resulting mutations were formatted into a residue file or "resfile", which tells Rosetta which residue to place at any given position in a model.

The mutations in the HC/LC interface of VRC-PG04 were reverted to their inferred germline residue using 3se9_germline.resfile:

```
NATAA
EX 1 EX 2
start
91 H PIKAA Y #F
32 L PIKAA Y #H
34 L PIKAA A #T
38 L PIKAA Q #K
43 L PIKAA A #P
44 L PIKAA R #K
49 L PIKAA Y #F
46 L PIKAA G #A
53 L PIKAA S #K
```

Construction of Rosetta models of interface-reverted CD4BS antibodies

To determine the effects that these naturally occurring somatic mutations had on the antibody bound conformation, we used Rosetta to construct ensembles of models for the HC/LC interface germline-reverted antibodies. Since we want to understand how antigen-distal mutations contribute to the bound conformation, the protocol was limited to a rigid-body threading. The following command was used to generate models for the mature antibody:

```
        /path_to_rosetta/rosetta/main/source/bin/relax.default.linuxgccrelease -flip_HNQ
-no_optH false -relax:constrain_relax_to_start_coords -score:weights talaris2013.wts -
relax:ramp_constraints false -s 3se8_Fv_clean.pdb -nstruct 100 -scorefile 3se9.fasc -
out:suffix "_mature”
```

Interface reverted models were generated using the following command:

```
        /path_to_rosetta/rosetta/main/source/bin/relax.default.linuxgccrelease -flip_HNQ
-no_optH false -relax:constrain_relax_to_start_coords -score:weights talaris2013.wts -
relax:ramp_constraints false -s 3se9_Fv_clean.pdb -nstruct 100 -out:suffix "_revert" -
relax:respect_resfile 1 -packing:resfile 3se9_germline.resfile -scorefile
3se9_reverted.fasc
```

At this point, both the mature and interface reverted models have been generated, but we still want to evaluate the effect that the mutations have on the interface. To ensure that we calculate across the same interface that we defined earlier, we’re going to use the Interface Analyzer application again, but in the form of a mover. For the sake of this example, it isn’t necessary, but it makes batch processing much easier. This method takes a “flags” or “options” file, an xml file, and the input pdb, all of which are available in the “inputs” folder on the github page. The interface energy, ΔΔG, was calculated for each model using the following commands:

```
/dors/meilerlab/apps/rosetta/rosetta_2015.12.57698/main/source/bin/rosetta_scripts.default.linuxgccrelease @iface_analyzer.flags -s *mature*pdb -parser:protocol iface_analyzer_VH_VL.xml -out:file:score_only -scorefile iface_3se9_mature.fasc
```

```
/dors/meilerlab/apps/rosetta/rosetta_2015.12.57698/main/source/bin/rosetta_scripts.default.linuxgccrelease @iface_analyzer.flags -s *revert*pdb -parser:protocol iface_analyzer_VH_VL.xml -out:file:score_only -scorefile iface_3se9_reverted.fasc
```

Next, the top 10 scoring models for each treatment were identified and their metrics collected using these commands:

```
cat iface_3se9_mature.fasc | sort –nk 2 | head -10 > top10_mature.fasc
cat iface_3se9_reverted.fasc| sort –nk 2 | head -10 > top10_reverted.fasc
```

The sixth column in these "top10" scorefiles represents the value for interface energy. The change in average interface energy, ΔΔΔG, is equal to Mature ΔΔG – Reverted ΔΔG. Negative values indicate a more favourable interface in the bound conformation.

HC/LC docking

Next, we determined how the mutations in the HC/LC interface affect HC/LC orientation by performing small perturbation docking. The docking step also takes an options file, and xml, and the starting model. In order to provide a direct comparison between mature and reverted interfaces, we restricted the docking protocol so that it does not alter the structural integrity of the domains; the minimization step employed an atom coordinate constraint, ensuring that the relax protocol itself would not skew angle measurements. We use this to generate 1000 models for each category, and analyse the top %5 for each, ranking by ΔΔG. Small perturbation docking was enacted using the following commands for the mature and reverted models:

Mature

```
/path_to_rosetta/rosetta/main/source/bin/rosetta_scripts.default.linuxgccrelease
@docking.flags -s 3se9_Fv_clean.pdb -parser:protocol small_pert.xml -out:file:scorefile
3se9_dock_mature.fasc -nstruct 1000 -out:suffix “_mature”
```

Reverted

```
/path_to_rosetta/rosetta/main
/source/bin/rosetta_scripts.default.linuxgccrelease @docking.flags -s
3se9_Fv_clean.pdb -resfile 3se9_germline.resfile -parser:protocol small_pert_revert.xml
-out:file:scorefile 3se9_dock_revert.fasc -nstruct 1000 -out:suffix _dock_revert
```

The models are then ranked by ΔΔG, and the top %5 are used to evaluate change in orientation.

```
cat 3se9_dock_mature.fasc | sort –nk 10 | head -50 > 3se9Top50Mature.fasc
cat 3se9_dock_revert.fasc | sort –nk 10 | head -50 > 3se9Top50Revert.fasc
```

In order to determine how somatic mutations in the HC/LC interface affect heavy and light chain relative orientation, we used ABangle to calculate the relative HC/LC orientation for each of the top-scoring models. This software calculates six parameters by mapping two reference planes onto the Fv domains, drawing a distance vector between them, and measuring five angles – a torsion angle and four bend angles, between the two planes while using the distance vector as a pivot axis. Additionally, ABangle can take in a list of PDB files to evaluate in the form of .dat files. Generating the .dat file is accomplished through the following commands:

```
cat 3se9Top50*fasc | grep dock | awk '{print($NF".pdb"}' > 3se9Top50.dat
```

ABangle was used to calculate relative orientation through the following command:

```
ABangle –i 3se9top50.dat –usernumbered
```

The resulting angles are found in /path_to_ABangle/ABangleData/UserAngles.dat. The average values, standard deviations, and standard error were calculated for each type of model (mature and revertant) across the six ABangle parameters. The resulting values were used in the following equations to calculate the shift in average angle and tightening of each distribution.

(3) ***Normalized shift*** =

$$\frac{1}{6}\sum \frac{\left|\bar{X}_{Reverted(HL,etc\ldots)} - \bar{X}_{mature(HL,etc\ldots)}\right|}{\sigma_{Reverted(HL,etc\ldots)} + \sigma_{mature(HL,etc\ldots)}}$$

where $\bar{\boldsymbol{X}}_{\boldsymbol{Reverted}(\boldsymbol{HL,etc}\dots)}$ is the mean ABangle value for any angle distribution generated by docking a reverted HC/LC interface, $\bar{\boldsymbol{X}}_{\boldsymbol{mature}(\boldsymbol{HL,etc}\dots)}$ is the corresponding mean ABangle value for the mature antibody, where $\boldsymbol{\sigma}_{\boldsymbol{Reverted}(\boldsymbol{HL,etc}\dots)}$ is the standard deviation for any given angle distribution generated by HC/LC docking at a reverted interface, and $\boldsymbol{\sigma}_{\boldsymbol{mature}(\boldsymbol{HL,etc}\dots)}$is the standard deviation for the corresponding mature antibody distribution. The Normalized Shift metric provides an estimate of how much the orientation distributions differ between any given mature antibody and its reverted counterpart as a whole. Values greater than one suggest a shift in each category by an average of 1 standard deviation.

(4) ***Tightening*** =

$$\frac{\boldsymbol{\sigma}_{\boldsymbol{Reverted}(\boldsymbol{HL,etc}\dots)}}{\boldsymbol{\sigma}_{\boldsymbol{mature}(\boldsymbol{HL,etc}\dots)}}$$

The tightening equation generates a ratio of standard deviations. Values greater than 1 suggest that the mature antibody models embody a tighter angle distribution during HC/LC docking.

The standard error (SE) for the shift was calculated using error propagation rules for addition where:

shift in HL angle = $\left|\bar{\boldsymbol{X}}_{\boldsymbol{Reverted}(\boldsymbol{HL,etc}\dots)} - \bar{\boldsymbol{X}}_{\boldsymbol{mature}(\boldsymbol{HL,etc}\dots)}\right|$

SE*(HL)* **=**

$$\sqrt{\boldsymbol{SE}^2_{\boldsymbol{Reverted}(\boldsymbol{HL})} + \boldsymbol{SE}^2_{\boldsymbol{mature}(\boldsymbol{HL})}}$$

normalized ***SE****(HL)* =

$$\frac{\sqrt{\boldsymbol{SE}^2_{\boldsymbol{Reverted}(\boldsymbol{HL})} + \boldsymbol{SE}^2_{\boldsymbol{mature}(\boldsymbol{HL})}}}{\boldsymbol{\sigma}_{\boldsymbol{Reverted}(\boldsymbol{HL})} + \boldsymbol{\sigma}_{\boldsymbol{mature}(\boldsymbol{HL})}}$$

and ***SE****(Normalized Shift)* =

$$\frac{\mathbf{1}}{\mathbf{6}}\sqrt{\boldsymbol{SE}^2_{(\boldsymbol{HL})} + \boldsymbol{SE}^2_{(\boldsymbol{HC1})\dots}}$$

Supplementary Data

Supplementary tables can be downloaded at:

https://github.com/ac1546/Dissertation

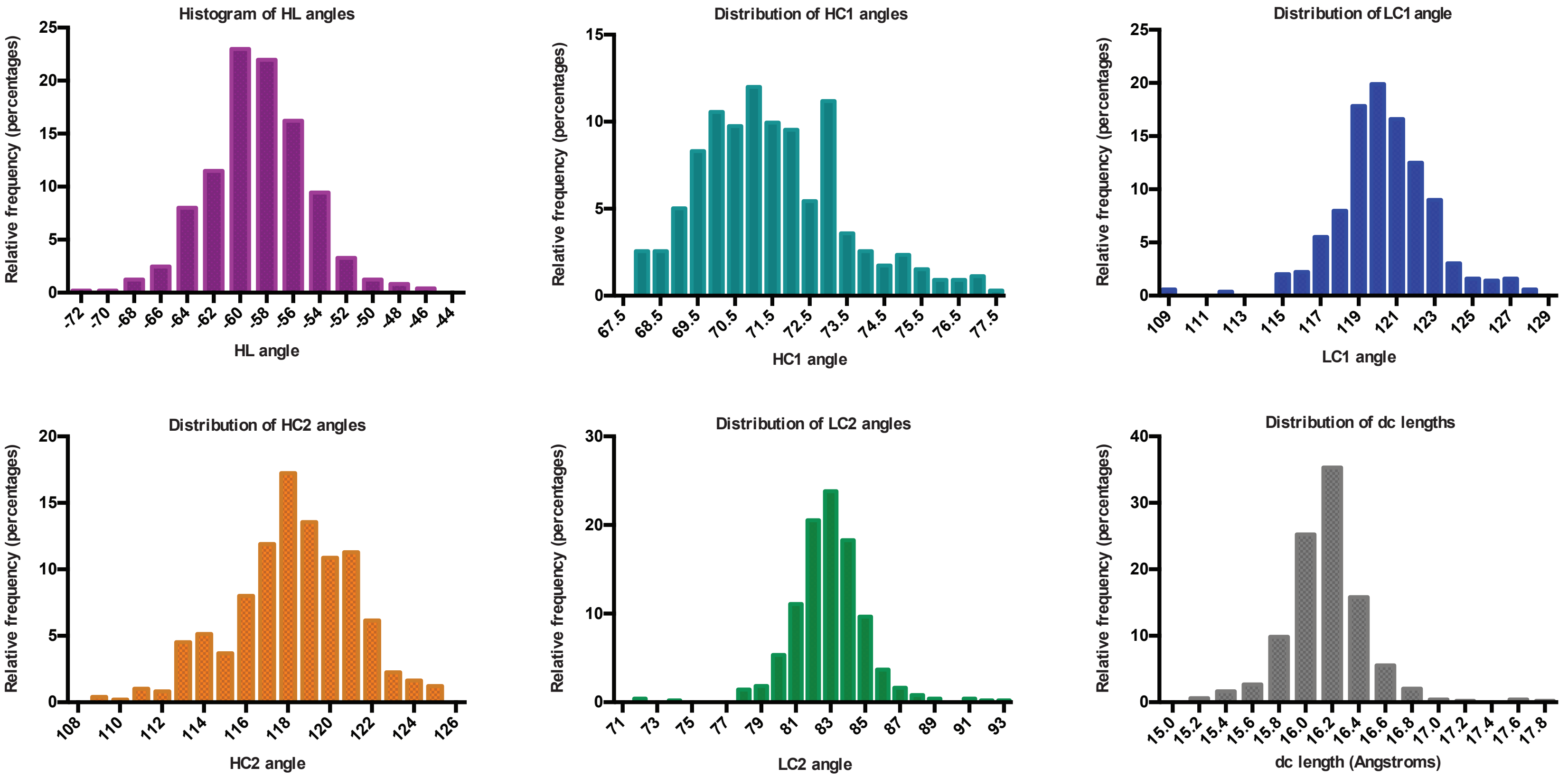


Figure S1. Distribution of human antibody/antigen co-crystals. Human antibody/antigen co-crystals were analyzed with ABangle. The resulting angle distributions for the HL (purple), HC1 (teal), LC1 (blue), HC2 (orange), LC2 (green), and dc (grey) values were produced using Prism. These metrics have a mean value of -58.6 (HL), 71.33 (HC1), 120.32 (LC1), 118.236 (HC2), 82.9 (LC2), and 16.15 (dc).

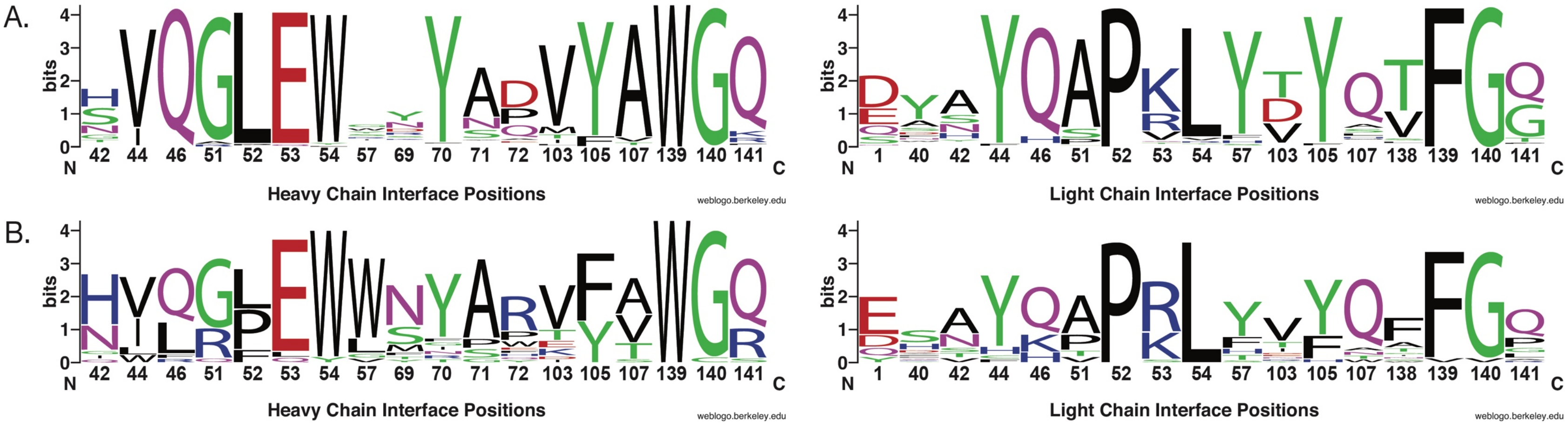


Figure S2. Sequence variation at the HC/LC interface. (A) WebLogos11 showing frequency of amino acids in the HC/LC interface for antibodies that fail to bind to the CD4 binding site (CD4BS) epitope. (B) WebLogos showing the frequency of amino acids in the HC/LC interface for CD4BS antibodies.

APPENDIX II

Publications and contributions

Bender, B. J., Cisneros, A., Duran, A. M., Finn, J. A., Fu, D., Lokits, A. D., Mueller, B. K., Sangha, A. K., Sauer, M. F., Sevy, A. M., Sliwoski, G., Sheehan, J. H., DiMaio, F., Meiler, J., … Moretti, R. (2016). Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. *Biochemistry*, *55*(34), 4748-63.

Abstract:

Previously, we published an article providing an overview of the Rosetta suite of biomacromolecular modeling software and a series of step-by-step tutorials [Kaufmann, K. W., et al. (2010) *Biochemistry 49*, 2987–2998]. The overwhelming positive response to this publication we received motivates us to here share the next iteration of these tutorials that feature *de novo* folding, comparative modeling, loop construction, protein docking, small molecule docking, and protein design. This updated and expanded set of tutorials is needed, as since 2010 Rosetta has been fully redesigned into an object-oriented protein modeling program Rosetta3. Notable improvements include a substantially improved energy function, an XML-like language termed "RosettaScripts" for flexibly specifying modeling task, new analysis tools, the addition of the TopologyBroker to control conformational sampling, and support for multiple templates in comparative modeling. Rosetta's ability to model systems with symmetric proteins, membrane proteins, noncanonical amino acids, and RNA has also been greatly expanded and improved.

My contribution to this publication as co-author centered around describing caveats that a user encounters while employing Rosetta as a modeling tool. I delineate the stochastic, abbreviated nature of the score function and note that exhaustive sampling may be required in order to approach the local energy minimum (much less global energy minimum) for a structure or complex through any given application.

BIOCHEMISTRY
including biophysical chemistry & molecular biology

Current Topic

pubs.acs.org/biochemistry

Protocols for Molecular Modeling with Rosetta3 and RosettaScripts

Brian J. Bender,[†,‡] Alberto Cisneros, III,[‡,§] Amanda M. Duran,[‡,‖] Jessica A. Finn,[‡,⊥] Darwin Fu,[‡,‖] Alyssa D. Lokits,[‡,#] Benjamin K. Mueller,[‡,‖] Amandeep K. Sangha,[‡,‖] Marion F. Sauer,[‡,§] Alexander M. Sevy,[‡,§] Gregory Sliwoski,[‡,‖] Jonathan H. Sheehan,[‡] Frank DiMaio,[@] Jens Meiler,[†,‡,§,‖,⊥,#] and Rocco Moretti[*,‡,‖]

[†]Department of Pharmacology, Vanderbilt University, Nashville, Tennessee 37232-6600, United States

[‡]Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37240-7917, United States

[§]Chemical and Physical Biology Program, Vanderbilt University, Nashville, Tennessee 37232-0301, United States

[‖]Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States

[⊥]Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville, Tennessee 37232-2561, United States

[#]Neuroscience Program, Vanderbilt University, Nashville, Tennessee 37235, United States

[@]Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States

Ⓢ *Supporting Information*

ABSTRACT: Previously, we published an article providing an overview of the Rosetta suite of biomacromolecular modeling software and a series of step-by-step tutorials [Kaufmann, K. W., et al. (2010) *Biochemistry 49*, 2987−2998]. The overwhelming positive response to this publication we received motivates us to here share the next iteration of these tutorials that feature *de novo* folding, comparative modeling, loop construction, protein docking, small molecule docking, and protein design. This updated and expanded set of tutorials is needed, as since 2010 Rosetta has been fully redesigned into an object-oriented protein modeling program Rosetta3. Notable improvements include a substantially improved energy function, an XML-like language termed "RosettaScripts" for flexibly specifying modeling task, new analysis tools, the addition of the TopologyBroker to control conformational sampling, and support for multiple templates in comparative modeling. Rosetta's ability to model systems with symmetric proteins, membrane proteins, noncanonical amino acids, and RNA has also been greatly expanded and improved.

Obtaining atomic-detail accurate models for all proteins, natural and engineered, in all relevant functional states, alone and in complex with all relevant interaction partners by crystallography or nuclear magnetic resonance (NMR) is impaired by the vast number of possible protein sequences and interactions. In some cases, it is complicated by experimental obstacles and is often time and cost intensive. Additional difficulties arise when the dynamic properties of proteins and their interactions with other molecules are to be studied from crystallographic snapshots. Here, computational modeling of the structure and dynamics of proteins and interactions can complement experimental techniques. Such computational models add atomic detail not present in low-resolution or limited experimental data, model states that are not tractable for experimental structure determination, simulate conformational flexibility and plasticity of states, and prioritize states for crystallization or study with other experimental techniques.

At the same time, prediction and design of protein structure *in silico* is a formidable task: the need to model thousands of atoms instantiates the sampling challenge of testing a large number of possible arrangements or conformations. The need to complete these calculations in a finite time creates the scoring challenge of developing an energy function that is rapid but still accurately identifies biologically relevant, low-free energy states.

The Rosetta software suite represents a compilation of computational tools aimed at obtaining physically relevant structural models of proteins and their interactions with other proteins, small molecules, RNA, and DNA. Rosetta has contributed to the advancement of structural biology by tackling challenges in *de novo* protein design,[1−3] comparative modeling,[4,5] protein design,[6−11] protein−protein docking,[12−15] and protein−small molecule docking.[16−18] Additionally, Rosetta can be applied to RNA/DNA structure prediction,[19,20] the incorporation of noncanonical amino acids,[21,22] and other difficult structural challenges such as membrane protein structure prediction[23] and modeling of symmetric proteins.[24,25]

Rosetta developers follow the hypothesis that a single, unified energy function should be able to accomplish all of these complex tasks; furthermore, the continuous optimization of this

Received: May 5, 2016
Revised: July 29, 2016
Published: August 4, 2016

DOI: 10.1021/acs.biochem.6b00444

Table 1. Publically Accessible Web Servers Running Rosetta[a]

server	address	protocols offered
ROSIE	rosie.rosettacommons.org	many, including small molecule docking, protein design, RNA design, etc.[46]
Robetta	robetta.bakerlab.org	structure prediction[51]
Rosetta.design	rosettadesign.med.unc.edu	protein design[104]
FlexPepDock	flexpepdock.furmanlab.cs.huji.ac.il	flexible peptide docking[105]
RosettaBackrub	kortemmelab.ucsf.edu/backrub	backbone remodeling and design[106]
FunHunt	funhunt.furmanlab.cs.huji.ac.il	classification of protein−protein complex interactions[107]
CS-Rosetta	csrosetta.bmrb.wisc.edu	structure prediction based on chemical shift data
RosettaDiagrams	rosettadiagrams.org	setup protocols through visual diagrams

[a] All web servers listed are free for noncommercial use.

energy function to improve one structural problem will ultimately improve performance for other modeling tasks. Important components of the energy function are statistically derived, i.e., using protein models derived from high-resolution crystallographic data in the Protein Data Bank (PDB) as a knowledge base.[1,6,16,23,26−35] For speed, the energy function is pairwise decomposable and employs a distance cutoff. For many sampling tasks, Rosetta employs a Monte Carlo search steered by the Metropolis criterion (MCM).[27] Rosetta is continually developed and rigorously tested by a consortium of international academic laboratories known as the RosettaCommons (www.rosettacommons.org). Herein, we present a global review of generalized Rosetta protocols and applications, as well as descriptions of novel functionalities recently introduced.[26,36,37]

Detailed tutorials and examples are included as Supporting Information. The tutorials herein supersede our previous tutorials put forward in "Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You".[38]

■ MAKING ROSETTA ACCESSIBLE

Rosetta is extremely powerful for many applications in structural biology, but for many years, it was limited by the fact that users needed an extensive background in C++ and the Unix environment to be able to construct new protocols. An ongoing effort by many groups has been taken to eliminate these boundaries, allowing greater flexibility and ease of use for the novice and intermediate user. These updates include customizable protocols using XML or Python. The updates using XML (RosettaScripts)[36] or Python (PyRosetta)[39] allow users to customize protocols without learning C++, by combining prewritten Rosetta objects and defining their behavior without having to write and recompile new C++ code. In addition, the Rosetta community now offers multiple web interfaces for application-specific tasks.

Other tools have been added, not to run Rosetta but to improve users' experience, such as graphical user interfaces (GUIs) to visualize Rosetta operations and to generate input files,[40] and PyMOL integration for real-time molecular visualization.[41] These tools offer users intuitive control over structural modeling without sacrificing flexibility and power.

RosettaScripts. RosettaScripts is an XML-like language for specifying modeling protocols through the Rosetta framework.[36] It allows users to define a set of Rosetta objects and execute them in a defined order to develop full protocols. Rosetta objects in RosettaScripts fall under four main categories: Movers, which are objects that modify a structure in some way; Filters, which evaluate properties of a structure; TaskOperations, which control the degrees of freedom of Rosetta's side-chain placement routines; and ScoreFunctions, which evaluate the energy of a structure. By combining these four elements, users are able to leverage many different sampling and scoring algorithms, with fine control over sampling degrees of freedom and protocol flow. All objects defined under these categories are customizable, which is a distinct advantage of RosettaScripts over conventional command line applications. For example, a user can define multiple score functions to be used in different sections of a protocol and then combine several protocols into a single XML protocol (i.e., protein−protein docking and design). This flexibility has made a number of scientific advances possible, such as *de novo* design of an influenza binder,[10] protein−protein docking based on hybrid structural methods,[42] and HIV vaccine design.[43]

PyRosetta. Because of the popularity of Python as a programming language in the computational biology community, a Python-based implementation of Rosetta was developed, termed PyRosetta.[39] PyRosetta consists of Python bindings for the major functions and objects of Rosetta, allowing all of these objects to be run from a Python environment. One advantage is the ability to combine Rosetta protocols with other popular structural biology software, such as PyMOL[44] and BioPython.[45] PyRosetta includes access to the same set of Rosetta objects for sampling and scoring that are described above for RosettaScripts, as well as many others. Unlike RosettaScripts, PyRosetta can be run in either script mode or interactive mode. Interactive mode allows the user to inspect their objects in real time while prototyping a new protocol.[39] Notably, PyRosetta is available for Windows in addition to Linux and Mac OSX, expanding the availability of Rosetta to researchers who use a Windows environment.

Web Interfaces. We are aware of eight Web servers that have been created to allow nonexperts to make use of Rosetta's functionality (Table 1). These Web servers allow Rosetta to be used with almost no learning curve, making the boundary to entry even lower than that of the scripting protocols mentioned above. In particular, ROSIE [the Rosetta Online Server that Includes Everyone (http://rosie.rosettacommons.org)][46] has been set up to easily provide a web interface for new Rosetta protocols.

Other Tools. Since the publication of RosettaScripts and PyRosetta, new tools have been developed to make running a Rosetta protocol even more intuitive. An interface to PyMOL was developed by Baugh et al., which allows users to visualize their molecules being manipulated by Rosetta as the protocol is being run.[41] While the viewer was originally developed for use with PyRosetta, it has since been extended for RosettaScripts. This visualization tool is especially useful for new users with experience in structural biology but new to computation.

Table 2. Standard Rosetta Score Function Terms

score term	definition
low-resolution scoring terms	
env	hydrophobicity term for each amino acid
vdw	steric repulsion between two residues
pair	probability of two residues interacting
rg	radius of gyration
cbeta	solvation term based on a number of surrounding residues
hs_pair, ss_pair, and sheet	secondary structure terms
high-resolution scoring terms (talaris2014)	
fa_atr, fa_rep, and fa_intra_rep	decomposed 6−12 Lennard-Jones potential
fa_sol	EEF1 solvation term
pro_close	proline ring closure energy
omega	omega backbone dihedral potential
dslf_fa13	updated disulfide geometry potential
rama	potential of ϕ and ψ angles for each amino acid
p_aa_pp	probability of an amino acid given a set of ϕ and ψ angles
fa_dun	rotamer likelihood
hbond_sr_bb, hbond_lr_bb, hbond_bb_sc, and hbond_sc	combined covalent−electrostatic hydrogen bond potentials for α-helices, β-sheets, side-chain backbone, and side-chain−side-chain interactions, respectively
yhh_planarity	tyrosine hydroxyl out-of-plane penalty
fa_elec	Coulombic electrostatic potential between two residues with a distance-dependent dielectric (deprecates fa_pair)

In addition, several GUIs for Rosetta have been developed to eliminate the need to run Rosetta exclusively through the Unix command line.[40] The PyRosetta Toolkit was developed to serve as a GUI for running PyRosetta, with menus to guide the user through the relevant Rosetta options that are needed for a protocol.[47] InteractiveRosetta is a GUI for running Rosetta protocols with an integrated molecular visualization window and user-friendly controls for implementing common Rosetta protocols.[40] Through these GUIs, users can generate input files for Rosetta protocols using a "point and click" interface while also running protocols seamlessly in the same window.

■ SAMPLING AND SCORING IN ROSETTA

Rosetta Sampling. While the approaches used by different protocols vary, in general Rosetta utilizes a Monte Carlo Metropolis sampling algorithm to quickly and efficiently determine the quality of structural trajectories. Rosetta further differentiates between sampling backbone and side-chain conformations within two separate refinement tasks. In addition, backbone sampling can be performed on a global or local scale. Large-scale backbone sampling utilizes 3-mer and 9-mer fragments derived from the Protein Data Bank (PDB), while local refinements of the backbone optimize ϕ and ψ angles without disturbing the global fold. Side-chain sampling also utilizes information derived from the PDB to create a "rotamer" library of observed conformations to reduce the conformational search space. For a more detailed discussion of Rosetta sampling, see ref 27.

Rosetta Scoring. The Rosetta score, or energy function, is a linear, weighted sum of terms combining knowledge- and physics-based potentials gathered from protein structural features within the PDB. The score function is used during Rosetta modeling to evaluate Monte Carlo sampling and for scoring the final output pose. With the implementation of Rosetta3, the score function is treated as a separate entity such that it can be repeatedly called and rapidly processed in a manner independent of the protocol at hand.[26] Additionally, score terms are grouped into a hierarchy based on potentials related to one entity (i.e., χ-angle probability), two interacting entities (i.e., hydrogen bonding potential), and terms that require the analysis of the entire model (i.e., radius of gyration).

Low- versus High-Resolution Scoring. In low-resolution scoring, or "centoid" mode, the side chain of each residue is removed and represented instead as a super atom ("centroid"), at a position that roughly approximates the center of mass of that side chain, averaged across likely side-chain states (or at the Cα atom for glycine). This greatly reduces the degrees of freedom that must be sampled during low-resolution backbone movement while preserving chemical and structural features of a given residue. Typical low-resolution sampling involves replacement of the backbone conformation with peptide fragments three and nine amino acids in length that are derived from the PDB. Peptide fragments are generated from the primary sequence of the protein. Centroid-mode scoring and sampling are used during the initial stages of protein modeling where exhaustive searches of conformational space are performed such as *de novo* protein folding, loop building, and rigid-body protein−protein docking.[1,12,27,28] Common score terms used in centroid mode are listed in Table 2.

High-resolution scoring, or "full-atom" mode, allows for full representation of all atoms of each side chain. In full-atom mode, conformational sampling relies on evaluating side-chain rotamers (derived from the PDB) during a Metropolis Monte Carlo simulated annealing protocol to find the global minimum.[29] Full-atom scoring was originally developed for protein design but has seen several improvements throughout Rosetta's history to the current *talaris2014* score function.[6,30−32,37] We have provided an additional example tutorial for the user on the basics of Rosetta scoring; see the scoring_and_prep folder in the Supporting Information.

Score Function Optimization. The score function is a linear weighted sum of energy terms; therefore, the weights can be parametrized to generate meaningful scores for predicted models. These are often fit against benchmark sets of modeling challenges to guide prediction of native structures. An algorithm "optE" was developed to streamline this weighting term optimization.[32] This algorithm excels at setting reference weights for amino acids. Using the approximation that a native, evolved sequence is close to the optimal sequence for a structure,[30] optE attempts to find reference weights that minimize the divergence from native sequence profiles. Via optimization of the *talaris2014* score function for sequence recovery (~40%), performance in novel design tasks is also improved.

Like previous iterations of the full-atom score function, *talaris2014* sums separate physics- and knowledge-based potentials. It was found that combining physics- and knowledge-based information in a given score term led to improved Lennard-Jones and hydrogen bonding score terms.[31] The combined covalent−electrostatic hydrogen bonding terms were further updated with improved geometry and parametrization for sp^2-hybridized hydrogen bond acceptors.[37] Scoring potentials of knowledge-based score terms were smoothed with the use of bicubic-spline interpolation.[32] An updated rotamer library was included with an adaptive kernel formulation, which allows for smoother potentials of Ramachandran-based score terms.[33] Ideal atomic coordinates for amino acids, the geometry of disulfide bonds, and the hydroxyl sampling of serine and threonine residues were also expanded and improved. The free energies of solvation (LK_DGFREE) were updated to improve the EEF1 solvation energy potential of buried residues. Lastly, a new term that describes the Coulombic electrostatic potential between two residues with a distance-dependent dielectric (fa_elec) was introduced and replaces the previous statistics-based potential (fa_pair).[32] Further refinements were made to reduce the influence of the hydrogen bonding terms. This resulted in improved sequence recovery, rotamer recovery, and model discrimination.[37] As of writing, these updates culminated in the *talaris2014* score function, which is the default for current versions of Rosetta. All *talaris2014* score terms are listed in Table 2.

Continual optimization of the Rosetta score function means that the default score function varies with Rosetta version: *score12* for versions prior to Rosetta 3.5, *talaris2013* for weekly releases until 2016.10, and *talaris2014* for Rosetta 3.6 and weekly releases since 2016.11. Further score function refinement is ongoing, and it is likely that future Rosetta releases will have a different default score function. Additionally, while Rosetta strives to have a single all-atom score function to encompass all modeling tasks, several application-specific scoring potentials have been developed to include new score terms and optimized score term weights. These include, but are not limited to, modified score functions for small molecule docking,[16] protein−protein docking,[12,34] and membrane protein modeling,[23,35] as well as specialized score functions for low-resolution sampling stages.

Limitations and Caveats. Ongoing improvements made by the Rosetta community have led to increasingly accurate modeling protocols; however, there are still several hurdles that must be overcome for Rosetta to accurately produce nativelike models. First, Rosetta sampling is stochastic in nature. Therefore, not every modeling trajectory will sample a regional minimum on the score function. Second, the score function is heuristic and abbreviated for speed. It fails to fully recapitulate the fundamental forces. Therefore, minima of the energy function are not guaranteed to describe biologically relevant states. Third, even with its rapid score function, Rosetta is unable to exhaustively sample all possible structural space due to computational time restraints. Fourth, many Rosetta protocols are optimized for local resampling and require a starting model, which may not exist for some systems.

Evaluating Interfaces. Some biological applications of Rosetta focus on improving, creating, or otherwise altering a well-defined protein−protein, protein−small molecule, or protein−DNA interface. These protocols typically inhabit a much smaller search space and in some cases rely solely on rigid-body optimization to generate a desired interaction.[48] In these instances, a series of specific interactions is evaluated, and the widely used "score vs RMSD" plot (see Figure 5h for an example of a score vs RMSD plot) is repurposed to look at small changes at the interface; here, plotting the "interface score" against the "interface RMSD" prevents small, meaningful changes from being lost in the larger fluctuations when scoring the entire model or computing the RMSD over all atoms. Additionally, analytical tools like the Interface Analyzer provide a series of useful calculations that include binding energy,[49] shape complementarity, the number of buried, unsatisfied hydrogen bonds, and the solvent accessible area buried at the interface. These metrics can be used in conjunction with RosettaHoles[50] to generate a packing statistic score for the interface.

■ *DE NOVO* STRUCTURE PREDICTION

De novo protein structure prediction is one of the greatest remaining challenges in computational structural biology. This process models the tertiary structure of a protein from its primary amino acid sequence. Importantly, *de novo* modeling differs from template-based or comparative protein modeling in that structural predictions are not based upon a known homologous structure. To address the challenge of predicting a protein's structure *de novo*, Rosetta uses short peptide "fragments" to assemble a complete protein structure.

The Rosetta *de novo* protein folding algorithm continues to follow the steps described in our previous review.[38] Briefly, short peptide fragments of known protein structures are obtained from the PDB and are inserted into an extended-chain protein following a Monte Carlo strategy.[1] In that sense, Rosetta *de novo* protein folding is not truly *de novo*; it combines a very large number of small templates. The hypothesis is that while not every protein fold is yet represented in the PDB, the conformation of small peptide fragments is exhaustively sampled. These peptide fragments are used to alter the backbone conformation of the extended-chain protein, folding it toward a low-energy tertiary structure. The process is repeated to create an ensemble of models. Finally, these low-resolution models can be filtered on the basis of pass/fail criteria provided by the user. These models can be clustered, and an energy minimization step applied to refine an all-atom model with the high-resolution energy function.

Generating Peptide Fragments. *De novo* protein folding relies on the assembly of short peptide fragments, usually generated as a preprocessing step. First, the primary protein sequence is used to generate secondary structure predictions. Next, the sequence, secondary structure predictions, and NMR data (if available) are used to pick candidate three- and nine-amino acid fragments from the PDB. Finally, these candidate fragments are scored, and the best N fragments are written to a fragment library file. The ROBETTA Web server (http://robetta.bakerlab.org) is available for noncommercial use and allows users to generate fragment libraries using a simple interface.[51] Additionally, Gront et al. have developed the

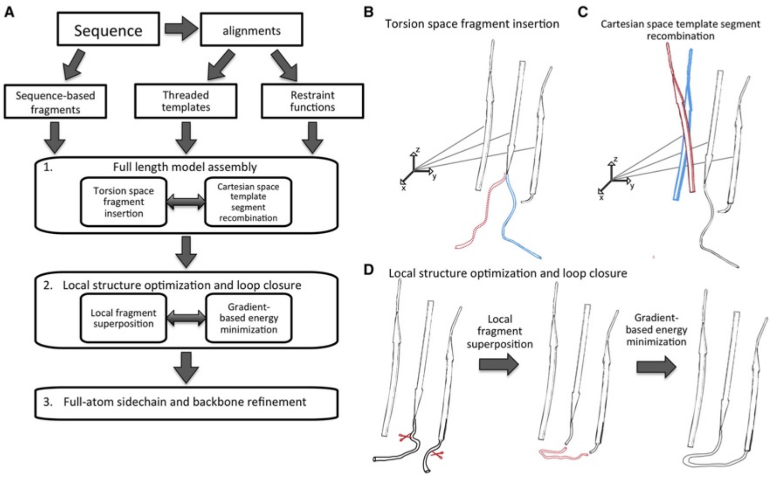

A
Sequence
alignments
Sequence-based fragments
Threaded templates
Restraint functions
1. Full length model assembly
Torsion space fragment insertion
Cartesian space template segment recombination
2. Local structure optimization and loop closure
Local fragment superposition
Gradient-based energy minimization
3. Full-atom sidechain and backbone refinement
B
Torsion space fragment insertion
C
Cartesian space template segment recombination
z
y
x
D
Local structure optimization and loop closure
Local fragment superposition
Gradient-based energy minimization

Figure 1. Multitemplate comparative modeling with Rosetta. (A) General workflow of the RosettaCM protocol. (B) Fragment insertion (blue, before insertion; red, after insertion). (C) Recombination of template segments. (D) Fragment insertion and minimization for loop closure. Reprinted with permission from ref 5. Copyright 2013 Elsevier.

FragmentPicker that provides users with total control over the fragment picking protocol.[52]

TopologyBroker. The TopologyBroker,[53] a tool that allows for more complex simulations, is an improvement added to Rosetta since our last review. The conformational space searched during a Rosetta *de novo* modeling simulation is vast, and successful searches often integrate prior knowledge with sampling. In *de novo* protein folding, this prior knowledge may be in the form of β-strand pairing constraints or the formation of a rigid chunk of the target fold based on a structurally homologous domain. Previously, protocol developers were restricted to a sequential sampling approach in which Rosetta could readily violate one set of these constraints while sampling to satisfy the other. The TopologyBroker was developed to create a consensus sampling approach that satisfies all of the requested constraints without requiring additional code development for each unique system; instead, the Broker provides an Application Program Interface (API) that allows for plug-and-play applications to generate complex sampling strategies.

Benchmarking *De Novo*. The *de novo* modeling capabilities of the object-oriented Rosetta software suite ("Rosetta3") were assessed in the CASP8 (Critical Assessment of protein Structure Prediction) experiment.[3] For 13 targets in the assessment, no homologous templates were identified and Rosetta's *de novo* modeling protocol was used to predict the structure of these targets. Following the observation that Rosetta *de novo* structural predictions are sometimes improved by using nonstandard fragment sizes, a range of fragment lengths were used when modeling the CASP8 targets. Longer fragment lengths were found to improve modeling of α-helical proteins, while shorter fragment lengths mainly improved modeling of β-strand proteins.

Limitations of *De Novo*. Because *de novo* structure prediction is such a powerful tool and yet such a complex challenge, it is critically important to understand the limitations of the algorithm. Rosetta performs well at folding small, globular, soluble proteins as well as small, simple membrane proteins containing 80−100 residues. However, large and complex proteins present additional difficulties that are not easily overcome by *de novo* techniques alone. Instead, users must incorporate other biochemical information to obtain nativelike models. Ongoing work shows that the incorporation of residue−residue co-evolution information can significantly improve the prediction accuracy during *de novo* modeling trails.[2] Other techniques such as homology modeling and using experimental constraints are discussed below.

Furthermore, because *de novo* structural prediction will sample many potential protein folds, it is necessary to generate large numbers of models (>10000) to adequately sample the conformational space. Extensive computational resources are needed to generate this number of models, and the use of distributed computational methods (such as computational clusters) is recommended. An example tutorial for the *de novo* prediction of a protein structure with Rosetta is included in the Supporting Information. This tutorial, protein_folding, provides an outline for a basic *de novo* protocol. Structural prediction of a soluble protein is described, both with and without the application of experimentally derived restraints. Also, a brief review of model analysis is covered. Instructions on how to run a membrane protein *de novo* protocol are included in a subfolder of the protein_folding tutorial.

■ COMPARATIVE MODELING

Comparative modeling differs from *de novo* methods in that it utilizes a known protein structure as the starting scaffold or template for structural prediction. If the template structure is a homologous protein, one speaks often of "homology modeling". Comparative modeling is a useful strategy for predicting protein structure and function when experimental methods fail or would be too resource intensive to employ. It increases the probability of obtaining realistic conformational predictions, especially when the target, or desired protein, is greater than 150 amino acids in length and/or adopts a complex tertiary fold. However, it requires that a related, often homologous, structure has been determined experimentally; this is termed the template. Ideally, the sequence identity between the target and the template is >30%, although proteins with lower sequence identity may still be used for comparative modeling when their tertiary fold is conserved.

The latter case will be examined within the tutorial provided with the Supporting Information. This tutorial, rosetta_cm, outlines the basic steps necessary for comparative modeling in Rosetta. The tutorial focuses on the use of RosettaRelax and RosettaMembrane, as well as information for implementing basic restraints.

Over the past several years, comparative modeling in Rosetta has incorporated many improvements, specifically the use of multiple templates and a specific low-resolution scoring functions.[5] Previously published protocols of comparative modeling with Rosetta suggested using multiple templates to obtain diversity and flexibility.[4] However, models were built on individual templates. The new RosettaCM protocol allows for integration of multiple templates with *de novo* fragments into a single structural model of the protein.[5] Hence, this multitemplate, multistage protocol samples a broader structural landscape and can select well-scoring subtemplates for different regions of the protein to be modeled.

A highly detailed description of RosettaCM design, sampling, and scoring has previously been published.[5] Users are encouraged to refer to this work for a comprehensive assessment of RosettaCM applications, considerations, and caveats. Herein, we will briefly describe features of RosettaCM as they apply to the protocol presented.

Starting Templates. Before utilizing RosettaCM, starting templates must be identified through remote homologue detection methods such as PSIBLAST.[54] When homologues are not found using sequence-based methods, three-dimensional (3D) fold recognition software may be used to obtain suitable templates. As with other modeling software, RosettaCM performance improves with higher sequence similarity and identity.

Three Stages of Multitemplate Comparative Modeling. Multitemplate RosettaCM is a three-stage process in which the best scoring model from each stage is utilized as the input for the following step (Figure 1). The output of stage 1 is a full-length, assembled model that is generally correct in topology. However, segment boundaries where templates are mended can be suboptimal in geometry and energetically frustrated. To resolve these energetic frustrations and to explore the conformational space around this starting model, stage 2 of RosettaCM iteratively improves local environments through a series of fragment insertions, side-chain rotamer sampling, and gradient-based energy minimization of the entire structure using a RosettaCM-specific low-resolution energy function. The best model from this cycle is then moved to stage 3 for a final round of all-atom refinement that improves side-chain geometries, backbone conformations, and packing density before converging on a final output model.

Modeling Loops. In previous Rosetta comparative modeling protocols, a user-defined, "loop" closure step was required to remove chain breaks, reconcile long unstructured coils, or rebuild regions of low sequence similarity (all of which are defined as "loops" within the Rosetta framework). Two different algorithms are available: Cyclic Coordinate Descent (CCD) and Kinematic Loop Closure (KIC). Briefly, CCD quickly closes roughly 99% of loops utilizing a robotics-inspired iterative approach to manipulate dihedral angles of three residue backbone atoms between user-specified C-terminal and N-terminal anchor points. The second loop building algorithm, KIC, explicitly determines all possible combinations of torsion angles within the defined segment using polynomial resultants.[55] While being slower than CCD, KIC determines more accurate loop structures, provided the anchor points are optimally set. Both algorithms within Rosetta can be used in conjunction with fragments derived from the PDB to build regions of missing electron density, poor homology, or backbone gaps.

Unlike the single-template loop building application, comparative modeling with multiple templates closes chain breaks and rebuilds loops internally during stage 2. *De novo* fragment insertions are encouraged in regions of weak backbone geometry, while template-based fragment insertions anneal chain breaks and low-electron density regions. Additional smoothing occurs with the RosettaCM-specific scoring function. This internal step removes the need for additional loop closures by the user. However, it is encouraged for the user to critically examine all output models to validate structural accuracy.

■ PROTEIN–PROTEIN DOCKING

Determining the optimal binding orientation and interface of two or more protein binding partners has many biological and pharmaceutical applications, yet determining the structure of protein–protein complexes by biochemical techniques is slow and laborious. RosettaDock is a useful tool for computationally predicting protein–protein interactions by employing an algorithm that simulates a biophysical encounter of two or more binding partners and optimizes the conformation of the bound state. The RosettaDock algorithm includes a multiscale, Monte Carlo-based docking algorithm that begins with a centroid-mode stage to identify docking poses, followed by an all-atom refinement stage to optimize rigid-body position and side-chain conformations.[12]

Global versus Local Docking. The initial pose for docking is determined by either global docking or local perturbation. Global docking randomly orients one of the two binding partners in relation to the other to determine an initial binding interface. This is useful when there is no biological or structural evidence to suggest a starting pose. Local perturbation allows the user to define a general starting pose for the binding partners when prior experimental knowledge exists; this initial placement greatly decreases the conformational search space and improves the sampling density close to the starting pose, although this may bias models toward the starting conformation. The tutorial included in the protein–protein_docking folder illustrates one application of Rosetta protein–protein local docking by using two known binding positions of the CR6261 antibody to influenza antigen hemagglutinin (HA) subtypes H1 and H5.

Low-Resolution versus High-Resolution Docking. The full RosettaDock algorithm begins with low-resolution docking. The first step involves rigid-body movements of the binding partners that rotate and translate in relation to one another.[13] The score function is used to achieve a threshold acceptance rate of rigid-body moves.[12] A high-resolution docking mode follows in which the lowest-energy structures and/or largest clusters assessed from the centroid-mode stage are selected for high-resolution refinement. Centroid pseudoatoms are replaced with all-atom side chains in their initial unbound conformations followed by additional fine-grained rigid-body docking.

Improvements to RosettaDock. The addition of RosettaScripts and PyRosetta to Rosetta now gives users the flexibility to modularize the centroid mode and all-atom mode of RosettaDock to suit case-specific applications. This was done by splitting RosettaDock into three major classes: DockingProtocol, DockingLowRes, and DockingHighRes.[13] The increase in flexibility showed an only marginal increase in successful predictions; however, it is particularly adept at predicting antibody−antigen complexes.[13] The modularization of RosettaDock has allowed users to also incorporate additional features within their docking protocols, including additional parameters for nonprotein moieties and protonation states,[49,56,57] flexible peptide-chain docking using FlexPepDock,[14] and *de novo* peptide docking.[15]

FlexPepDock. The FlexPepDock *de novo* docking algorithm is similar to the RosettaDock algorithm in that it begins with sampling rigid-body moves from the initial protein−peptide complex. Although not included in the tutorial, this step also includes iterative peptide fragment insertions and random moves of the peptide backbone using decreasing simulated temperature weights. Next, the low-resolution model is improved using an all-atom refinement stage by peptide side-chain placement optimization using a Monte Carlo search of "small" and "shear" moves described by Rohl et al.[1] Each round of refinement also includes a decreasing repulsive van der Waals weight term and an increasing attractive van der Waals term to allow a greater degree of perturbation within the binding pocket without causing the peptide and protein to separate during energy minimization. The FlexPepDock *de novo* benchmark demonstrated that the protocol produces near-native models with 86% accuracy (Figure 2).[15]

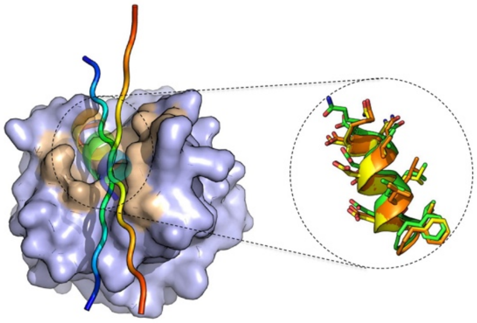


Figure 2. Protein−peptide interface prediction using FlexPepDock *ab initio*. Structure prediction of the Che-Z-derived peptide bound to CheY (PDB entry 2FMF) from two opposite starting orientations converges onto the same final conformation resembling the structure of the native peptide. The left panel is a general view of the CheY receptor (gray; interface residues colored light brown), the two initial, extended peptide conformations (rainbow cartoons), and the final helical peptide conformation (rainbow, transparent cartoon). The right panel is a detailed atomic view of the top FlexPepDock *ab initio* predictions from two simulations (yellow and orange) and the native peptide conformation (green). Reprinted from ref 15.

PROTEIN−SMALL MOLECULE DOCKING

Protein−small molecule docking aims to capture the binding interactions between a protein and a small molecule. This includes recapitulating the binding pose and quantifying the interaction strength. RosettaLigand, Rosetta's protein−small molecule docking protocol, is designed to consider both protein and small molecule flexibility.[16,17] It uses a two-phase docking approach similar to Rosetta's protein−protein docking: a low-resolution phase of rapid sampling based on shape complementarity followed by a high-resolution phase of Monte Carlo minimization of side-chain rotamers and small molecule conformers. The models undergo a final gradient minimization of the protein and molecule torsion degrees of freedom before they are output along with an interface score as a proxy for binding free energy. A small molecule docking tutorial (ligand_docking) included in the Supporting Information demonstrates this optimized protocol.

Improvements to RosettaLigand. In contrast to the previously published RosettaLigand protocol,[38,58] this tutorial replaces the independent translation/rotation low-resolution sampling steps with the new Transform algorithm.[18] The Transform algorithm couples translational, rotational, and conformational sampling into a single Monte Carlo process. In a benchmark case, the Transform algorithm demonstrated a 10−15% improvement in docking success rate and an effective 30-fold speed increase over the classical methods.[18] The improved search time permits the use of RosettaLigand for screening medium-sized small molecule libraries, protocols for which are found in the Supporting Information. For screening work with much larger libraries, Rosetta's Docking Approach using Ray-Casting (DARC) is a GPU-accelerated method demonstrated to be successful for protein−protein interface small molecules.[59] It should also be noted that screening applications use a simplified scoring function because of the computational complexity of fully flexible protein high-resolution refinement.

Customizable Small Molecule Docking Protocols. The RosettaLigand protocol can now be customized through the RosettaScripts XML interface, allowing for greater flexibility of use.[58] Additional features now include docking with explicit interface water molecules, which demonstrated 56% recovery for failed docking cases across a CSAR (Community Structure−Activity Resource) benchmark of 341 diverse structures.[60] Design of interfaces can now be incorporated into a single step for the docking and design of protein−small molecule binding pockets.[61] These RosettaScripts-based protocols have also been used to predict absolute binding energies for HIV-1 protease−inhibitor complexes with an *R* value of 0.71.[62]

Research questions often focus on small molecules binding to a target protein without an experimentally determined structure. Such cases require first building models of the receptor using *de novo* Rosetta, RosettaCM, or similar protein modeling protocols. When docking small molecules into protein models, Kaufmann and Meiler observed a nativelike binding pose among the top 10 scoring comparative models for 21 of 30 test cases.[63] Furthermore, docking results were significantly better in cases utilizing protein templates containing a small molecule of similar chemotype compared to templates with dissimilar small molecules or proteins in the apo state. A full Rosetta protocol linking comparative modeling and small molecule docking is available in ref 4. Combs et al.

utilized the previously discussed independent translation/rotation low-resolution sampling but can be easily modified to the new Transform sampling.

Small Molecule Docking in Membrane Proteins. Because of their biological importance and the challenges of experimentally determining their structures, membrane proteins are particularly attractive targets for the comparative model docking strategy. While the comparative modeling portion may be handled in a membrane environment, to date, Rosetta handles small molecule docking in a soluble environment. Nguyen et al. demonstrated the applicability of the soluble simplification for G protein-coupled receptors (GPCRs).[64] RosettaLigand sampled near-native poses when docking small molecules into comparative models of GPCRs, but selecting correct small molecule poses by Rosetta score alone remains challenging (Figure 3). The use of templates with high sequence identity, knowledge-based binding pocket filters, and experimental contacts are recommended methods for improving accuracy. Additional algorithm development and benchmarking are being pursued to fully integrate RosettaLigand with the RosettaMembrane framework.[23]

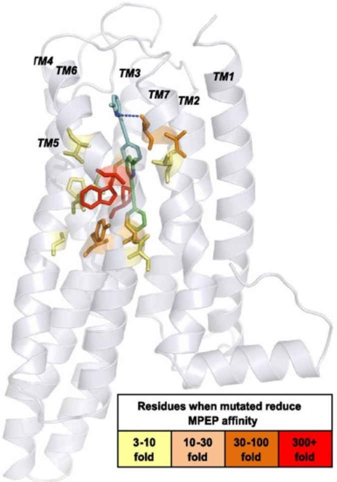


Figure 3. Application of RosettaLigand docking of negative allosteric modulator MPEP into a comparative model of the mGlu5 transmembrane domain. The predicted lowest-energy MPEP docking position (cyan) is close to residues demonstrating a change in MPEP modulations upon mutation (yellow to red). Reprinted with permission from ref 108. Copyright 2013 American Society for Pharmacology and Experimental Therapeutics.

■ INCORPORATING EXPERIMENTAL DATA

While Rosetta can sample near-native structures in a variety of situations, knowledge of limited experimental information can guide sampling and discriminate conformations inconsistent with experimental data, allowing more accurate determination of structures with less sampling. The incorporation of experimental data most commonly takes the form of modifications to the energy function. Addition of experiment-based scoring terms can make the energy landscape less rugged, allowing Rosetta sampling to more rapidly converge on relevant conformations.

For the incorporation of such information, Rosetta has a flexible restraint system (termed "constraints" in Rosetta parlance). Rosetta constraints have a two-part organization: specification of structural measurements such as distances or angles and a function that converts the measurement into an energetic penalty. A wide variety of measurements and functional transformations are currently available within Rosetta, and these can be freely mixed and matched according to the particular use case. There are also built-in tools for incorporating experimental data, allowing users to select only the best of a set of potentially inaccurate restraints. The flexibility of these restraints allows them to be applied in a diversity of situations, from incorporation of nuclear Overhauser enhancement (NOE) distances from NMR spectroscopy[65] to the use of mass spectrometry cross-linking information[66] to the use of custom potentials derived from probability distributions matching EPR/DEER measurements.[67,68]

Although the constraint system provides flexibility when incorporating experimental data for most Rosetta protocols, other experimental data types may reflect more complex structural parameters and require specialized scoring terms. Residual dipolar couplings,[69] pseudocontact shifts,[70] and small-angle X-ray scattering[71] have all been incorporated into Rosetta using specialized score terms, as have several techniques for working with electron microscopy (EM)- and X-ray-based electron density.[37,72,73] An example tutorial for using X-ray crystallography data and electron density maps with Rosetta, structure_refinement, is provided in the Supporting Information.

Improvements in image data analysis and electron detectors have led to advances in electron microscopy, producing electron density maps at resolutions as high as 3 Å for complex molecular machines. However, model building into these near-atomic resolution electron density maps is still difficult and error prone. DiMaio et al. have developed methods in Rosetta that incorporate medium- to high-resolution (3−5 Å) cryo-EM maps for density-guided structure determination and structure refinement.[72−74]

Protein Structure Prediction with Cryo-EM Restraints. This method takes advantage of near-atomic-resolution cryo-EM density maps for protein structure prediction. Using this method, highly accurate models of proteins up to 660 amino acids in length can be determined without homologous structures. This method includes density-traced backbone conformation and side-chain density agreement for sequence assignment during structure prediction. Structure determination starts with obtaining nine-residue fragments centered on each amino acid in the sequence using the Fragment Picker as mentioned previously in the *de novo* folding section. These fragments are then docked into the electron density map using a translational and rotational search to identify possible fragment placement. To further refine these placements, side-chain information is used to identify fragments with physically realistic side-chain conformations consistent with the experimental data. Finally, the largest mutually consistent subset of fragment placements is selected. A subset of placements is scored with a low-resolution score function that evaluates their pairwise consistency. Monte Carlo-simulated annealing finds a subset of fragment placements optimizing this score function. This assignment will not necessarily assign a position for each residue. This process is conducted iteratively until 70% of the sequence has been assigned a backbone conformation. For consecutive iterations, the portion of the density map already covered in the previous step is excluded from fragment placements. Finally, Rosetta loop modeling and an all-atom refinement step, both guided by the cryo-EM density map, fill in any missing regions in the model.

Cryo-EM-restrained protein structure prediction[72] yielded models within 2.0−3.1 Å all-atom RMSD compared to experimentally determined structures. Structure determination of proteins rich in β-sheets is challenging for this method because of the conformational variability of the structure. Medium-resolution (4.8 Å) density maps provide another challenge during partial structure building for this method.

Density-Guided Iterative Local Refinement. This structure refinement protocol[74] includes techniques from X-ray crystallographic refinement, *de novo* structure prediction, segment rebuilding, and all-atom refinement from comparative modeling in Rosetta to predict models of proteins at atomic-level accuracy starting from a low-resolution model (with the correct topology). Like comparative modeling techniques, backbone fragments are inserted onto a template structure via superposition and minimization to close the peptide bonds. In density-guided structure rebuilding, before the peptide bonds are closed the fragments are optimized to fit the density after superposition. The backbone fragments that do not fit into this density are replaced by backbone fragments derived from the PDB. Peptide bond, backbone, and side-chain geometries are maintained during this step by coordinate constraints at the fragment end points and Ramachandran and rotameric constraints, respectively. This density-guided rebuilding step is followed by alternative refinement of model coordinates and atomic *B* factors until a good correlation is obtained between the model and the density map. Finally, the quality of the refined model is evaluated using all-atom energies as well as agreement with the experimental data, using the Fourier Shell Correlation between the model and map.

With homologues as starting points, the structure of the 20S proteasome, periplasmic domains PrgH and PrgK of the needle complex, and a peptide fiber assembly were refined using this method.[74] The accuracy of the refined models was tested against the quality (sequence identity) of the starting model, the number of images used for the reconstruction of the map, and the resolution of the density map. Density-guided iterative local rebuilding generated >75% accurate models for maps up to 4.4 Å resolution and less accurate models for maps with a resolution lower than 5 Å. This suggests that to successfully refine a model, the helix pitch, individual β-strands, and some of the aromatic side chains should be partially visible in the density map.

Among many applications, the density-guided iterative local rebuilding technique for structure determination in Rosetta has been used to determine structures of the peroxisomal Pex1/Pex6 ATPase complex with a unique double-ring,[75] type VI secretion system contractile sheath in *Vibrio cholerae*,[76] and SIRV2 virion that infects the *Sulfolobus islandicus* hyperthermophilic acidophile.[77]

Refinement with Phenix and Rosetta. Phenix[78] is state-of-the art X-ray refinement software used to determine crystal structures of biomolecules. The Rosetta structure modeling methodology has been combined with the Phenix refinement method to improve structure determination at low and high resolutions. Phenix benefits from the detailed all-atom force field and more effective conformational search and minimization procedures that exist in Rosetta. The Phenix.Rosetta refinement approach[73] utilizes Phenix for bulk solvent correction to calculate electron density maps and refine atomic *B* factors while the Rosetta force-field, minimization, and sampling techniques are used to optimize the model geometry. This method includes alternative real and reciprocal space refinement to improve model structure. Rosetta force-field constrains the refinement to physically plausible conformations, and density maps restrain the Rosetta side-chain and backbone sampling during refinement.

The Phenix.Rosetta refinement method was tested against conventional refinement in Phenix, CNS,[79] and REFMAC.[80] On 26 models with density map resolution ranging from 3.0 to 4.5 Å, Phenix.Rosetta refinement generated models with superior geometry in terms of free *R* factor, MolProbity score, and RMSD compared to that of the published structures.[73]

Phenix.Rosetta refinement has been successfully adapted to determine structures of the flavin binding center of the NqrC subunit of sodium-translocating NADH:quinone oxidoreductase,[81] the full-length protein and regulatory domain of *Pseudomonas aeruginosa* OxyR,[82] the apo-TrmBL2 structure to understand nonspecific binding of DNA by TrmBL2,[83] and the $\alpha\beta$ T cell antigen receptor (TCR)−CD1a complex.[84]

Phenix.mr_rosetta is another model rebuilding technique that integrates structure modeling tools from Rosetta with crystallographic structure determination tools in Phenix.[85] This technique can be used to determine challenging structures for which simple molecular replacement procedures usually fail, when starting models are based on a remote homologue with <30% sequence identity.[86,87] The phenix.mr_rosetta algorithms allow users to identify suitable templates and refine them with Rosetta before performing molecular replacement and then rebuilding the models with Rosetta and Phenix autobuilding tools. Electron density map-guided energy optimization, combinatorial side-chain packing, and torsional space minimization are used to improve molecular replacement models before applying crystallographic model building techniques. Phenix.mr_rosetta allows rapid structure determination without experimental phase information given the availability of homologues structures with >20% sequence identity, diffraction data sets of better than 3.2 Å resolution, and four or fewer copies in the asymmetric unit cell.

■ PROTEIN DESIGN

Inverse Folding Problem. Protein design is a unique protocol in that instead of finding the optimal conformation of a particular sequence, it aims to determine an optimal sequence for a given conformation. For this reason, it is often termed the "inverse protein folding problem".[38] Generally, there are two main design strategies: design for stability and design for function. The stability protocol considers the entire protein for design, and the score terms of interest are generally focused on improved packing. The design for function protocol is usually a localized design, centered on a specific region, domain, pocket, etc., of a protein with a focused energy function that governs precise interactions, such as electrostatics or hydrogen bonding.

Protein design involves iterative optimization of sequence and structure. During the fixed backbone side-chain optimization step, sequence space is sampled simultaneously with side-chain conformational space using Monte Carlo-simulated annealing by exchanging all possible amino acids at user-specified designable positions while evaluating the predicted energy.[6] This is followed by flexible backbone minimization to optimize the model. The first successful use of *de novo* RosettaDesign produced a sequence for a fold not seen in the PDB.[6] The experimentally determined structure had an RMSD of 1.1 Å from the computationally design model. An example tutorial for protein design, protein_design, is provided in the Supporting Information.

Design for Stability. Protein stability can be affected by a single-point mutation. Kellogg et al. evaluated several protocols with varying levels of flexibility and sampling and determined one method in particular to be useful for single-point mutations.[7] This method was made into the application ddg_monomer. When ddg_monomer was tested on a set of 1210 single-point mutants from the ProTherm database, the correlation of predicted ddGs to experimental ddGs was 0.69 while the stability classification accuracy was 0.72.

While ddg_monomer is a tool for predicting how a single-point mutation affects the stability of a protein, RosettaVIP (void in packing) is a design strategy that has been developed to identify single-point mutations that could improve the stability of a protein.[9] When Borgo et al. fully designed proteins, they found that the hydrophobic cores of the designed models were poorly packed when compared to their respective native proteins. RosettaVIP was able to identify packing deficiencies and sample a much smaller sequence space to fill the void in packing, resulting in a more stable design.

Design for Functionality. In addition to stabilizing monomeric proteins, RosettaDesign can be used to design interfaces between proteins. Fleishman et al. established a dock design protocol that optimizes the sequence of a protein to bind a surface patch of a target protein during design. Docking was used to optimize the positioning of the interacting proteins at the interface. Experimentally determined structures had an interface very similar to those of the designed models.[10]

Other types of interfaces of interest for design applications are protein−small molecule interfaces. Tinberg et al.[11] provided a great example of using RosettaDesign to design for affinity as well as stability (Figure 4). First, RosettaMatch[88] was used to find a stable scaffold for design for binding a particular small molecule. Next, RosettaDesign was used to maximize the binding affinity between the protein and small molecule. Finally, a second round of design was used to minimize destabilization due to mutagenesis in the first round. To ensure these mutations were meaningful, design was guided by a multiple-sequence alignment. The resulting most energetically favorable model was the highest-affinity binder in experimental studies and had a cocrystal structure that agreed with the computational model.

Most design algorithms in Rosetta are performed while considering a single fixed backbone structure. Recently, efforts to consider several structures during the design process have been undertaken to tackle more difficult design problems. A generalized multistate design protocol was introduced in 2011[8] to help in cases in which design should occur to satisfy multiple conformations or to design specificity toward one state and negative design against other states. Willis et al.[89] showed that RosettaMultistateDesign was capable of predicting residues that were important for polyspecificity when designing the heavy-chain variable region of an antibody. Sevy et al. introduced a new approach to multistate design that accelerates the process of multistate design by reducing the sequence search space,[90] allowing more complex backbone movements to be incorporated into a design protocol.

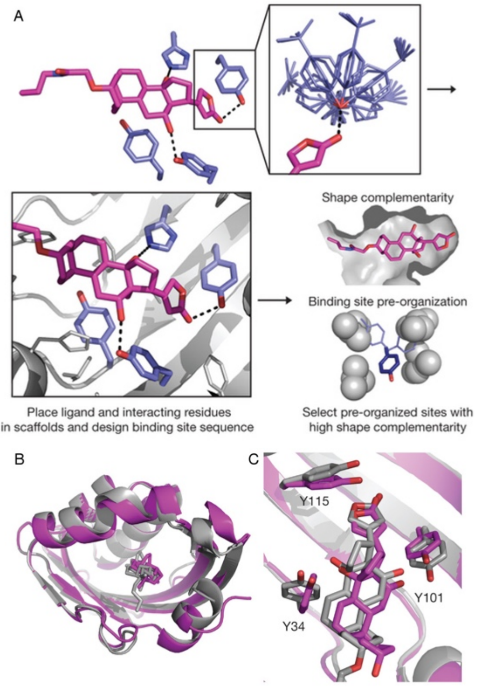

A
Shape complementarity
Binding site pre-organization
Place ligand and interacting residues
in scaffolds and design binding site sequence
Select pre-organized sites with
high shape complementarity
B
C
Y115
Y101
Y34

Figure 4. Design of protein−ligand interactions for high affinity and selectivity. (A) The design approach involved specifying binding interactions between the protein and ligand followed by design of the binding site. Finally, only designs in which shape complementarity was better than what is seen in native complexes were selected for experimental characterization. (B) Design crystal structure (purple) and computational model (gray) of the protein−ligand complex resulting from design for high affinity and selectivity. The RMSD was 0.54 Å, while the bound form (C) had an RMSD of 0.99 Å. Reprinted with permission from ref 11. Copyright 2013 Macmillan Publishers Ltd.

■ ADDITIONAL ROSETTA METHODS

Symmetry. Previously, Rosetta2 was limited in its ability to model large symmetric complexes.[24] In 2011, DiMaio et al. introduced a new mode in Rosetta to model symmetric proteins.[25] This allowed protocols to sample and score large, symmetric complexes much more quickly and with less memory usage as this approach samples only symmetric degrees of freedom, greatly reducing the search space. The underlying assumption, however, is that the interactions between all subunits are symmetric. The current implementation of RosettaSymmetry can create complex symmetric assemblies through the use of a symmetry definition file for a symmetric or nearly symmetric structure from the PDB. In the case of *de novo* folding, a symmetry definition file must be generated from scratch.

Membrane. RosettaMembrane has been the method used to model helical transmembrane proteins for several years. RosettaMembrane consists of both low-resolution[35] and high-resolution[91] scoring functions that were developed to describe how the protein interacts with the membrane environment. Recently, RosettaMP, a new framework for modeling membrane proteins in Rosetta, was developed to facilitate communication between model sampling and scoring.[23] Work is ongoing to adapt existing protocols to be compatible with RosettaMP.

Noncanonical Amino Acids and Noncanonical Backbones. Rosetta was initially developed to predict the three-dimensional structure of proteins using the 20 canonical amino acids. However, the expansion to include noncanonical amino acids (NCAAs) and noncanonical backbones (NCBs) is important, as they allow for the flexibility to create more precise interactions between proteins,[92] metal ions,[93] or antigens.[94] While the expansion to include more diverse structures is critical, the addition is nontrivial.

The addition of NCAAs requires the modification of both the scoring function and how the space is explored. These hurdles, however, are not easy to clear, as Rosetta is built on a foundation of knowledge-based components within its scoring function. Most of these knowledge-based score terms come from published protein structures, and few NCAAs have a statistically relevant representation in the PDB. Therefore, developers need to rework key components of the Rosetta scoring function.[21] All score terms were then reweighted to account for the changes in the score terms. Along with the new score terms, the authors created rotamer libraries for 114 NCAAs, as well as a tool, MakeRotLib, for creating rotamers for user-supplied NCAAs.

An effort was also undertaken to add noncanonical backbones to Rosetta, and in the initial attempt, five new backbones were added.[22] The first hurdle in the addition of an NCB is defining what a "residue" is. In Rosetta,3[26] the "residue" became the central object; therefore, with NCBs, a repeating subunit must be defined. Additionally, new backbone sampling movers must be created, or the backbone must be fixed, as the NCB will have flexibilities different from those of a linear chain of three singly bonded atoms. The final key point in the addition of NCBs is the creation of new rotamer libraries for the side chain. Even if the side-chain atoms are identical to those of a canonical side chain, the chemical change in the backbone will cause different flexibilities, due to sterics or electrostatics. A peptoid (a backbone structure identical to the canonical backbone, with the only change being the side-chain branches from the nitrogen instead of the α-carbon) rotamer generator has been created[95] for users to create rotamers for their own side chains. However, care must be taken when creating rotamers for a blended backbone system.[96]

The main considerations for a user attempting to use NCAAs and/or NCBs in Rosetta are understanding the chemical properties of their side chain and/or backbone and properly representing this knowledge in Rosetta. The correct score terms need to be used, as the standard knowledge-based score terms will not apply. An appropriate rotamer library and/or mover must be added to allow for proper sampling of the protein landscape. Finally, the user must understand that because work on NCAAs and NCBs is still limited, novel score terms or sampling methods may be required.

RNA. Structural predictions of RNA molecules require confronting the same challenges as protein modeling: sampling the conformational space of the heteropolymer and accurately scoring different conformations. Rosetta applies the same strategies developed for modeling proteins to address these challenges in nucleic acids.[19] An assembly of fragments that have been observed in known RNA structures is used to produce nativelike tertiary models. This procedure also captures sequence-dependent local conformational biases. A centroid-based "low-resolution" scoring function is used for selection of initial models. It is knowledge-based (statistically derived from frequencies observed in known crystal structures) and includes scoring terms for base pairing, base stacking, and compactness and terms for maintaining coplanarity and disfavoring steric clashes. In current protocols, models are subsequently refined using a full-atom physics-based energy potential and a Stepwise Ansatz.[20] This protocol is termed FARFAR, Fragment Assembly of RNA with Full-Atom Refinement, and is available using the distributed Rosetta suite and the ROSIE Web server.[46]

The Stepwise Ansatz (Figure 5) has been benchmarked on loops and hairpins up to 10 nucleotides in length. For larger structures, additional information is needed to restrict the conformational sampling to a tractable amount. This

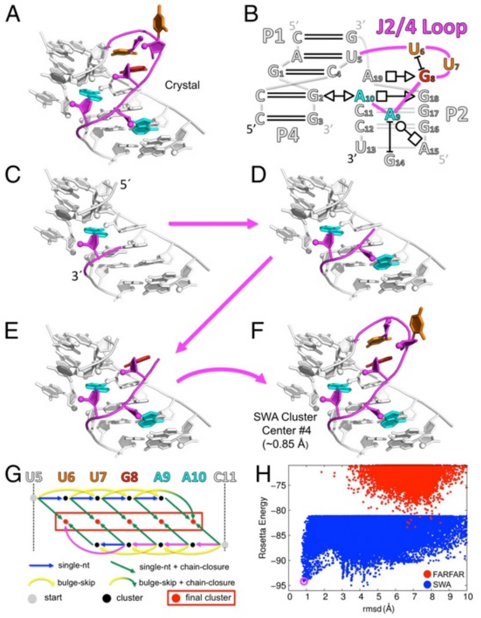


Figure 5. Stepwise assembly (SWA) structure modeling method for RNA. Illustration of the J2/4 loop from the three-way junction of a TPP-sensing riboswitch (PDB entry 3DV2). (A) Crystallographic conformation of the five-nucleotide loop (colored). (B) Schematic of the three-way junction. (C−F) The loop is built in a stepwise manner, starting from the 3′ end. (G) A directed acyclic graph recursively covers all possible build-up paths. The steps shown in panels C−F are colored magenta. Gray vertices correspond to the starting point with none of the loop nucleotides built. Black vertices are partially built subregions. Red vertices correspond to the ending points with the loop completely built. (H) Energy vs RMSD from the crystal for models generated by SWA (blue points) and by the prior method (FARFAR, red points). The SWA fourth lowest-energy cluster center (purple circle) is within atomic accuracy of the crystallographic model (0.85 Å RMSD). Reprinted with permission from ref 109. Copyright 2011 National Academy of Sciences.

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

information can come from both predictions and experimental data. Secondary structure predictions can be made using algorithms that take into account structures of homologous sequence. Chemical mapping experiments provide useful reactivity data that help assign base pairing status to each nucleotide. Multidimensional chemical mapping, such as "M2, mutate-and-map"[97] and Multiplexed -OH Cleavage Analysis by paired-end sequencing (MOHCA-seq),[98] can provide specific pairwise proximity information. Additional efficiency is gained by the preassembly of helical structures as input to the fragment assembly step. These methods have performed well in the recent blind prediction experiments called "RNA-Puzzles".[99]

The same techniques used for structure prediction can also be applied to structure refinement, to improve the quality of RNA crystallographic models in the presence of X-ray data. This procedure has been implemented in the "Enumerative Real-space Refinement ASsisted by Electron density under Rosetta" ERRASER-Phenix pipeline[100] and was demonstrated to improve the geometrical parameters and model quality of 24 RNA-containing structures in the PDB, including small pseudoknots and large ribosomal subunits.[101]

NMR structure determination of proteins or nucleic acids typically relies on a large number of NOE measurements to derive distance constraints for structure calculations. Using a relatively small number of measurements of only ^{1}H chemical shift values, CS-Rosetta-RNA was demonstrated to provide sufficient information to determine the structures of 23 noncanonical RNA motifs at high resolution.[102] This functionality is also available on the ROSIE Web server.[46]

RNA Design. RNA can be designed using Rosetta's RNA Redesign algorithm. It performs fixed backbone design on 3D RNA structures to produce sequences that best stabilize a given 3D conformation.[103] The success rate for a benchmark set of 15 RNA crystal structures was 45% sequence recovery overall and 65% sequence recovery for noncanonical sequences (not Watson–Crick or G-U). Finally, the algorithm was able to predict a sequence that would increase the thermostability of domain IV of the signal recognition particle.

■ CONCLUSIONS

The Rosetta software suite represents a compilation of computational tools aimed at obtaining physically relevant structural models of proteins, RNA, and small molecule interactions. Herein, we presented a general outline of updated Rosetta applications, protocols, frameworks, and functionalities with the aim of improving user success. All protocols are generalizable and can be applied to an extended list of biological queries that other structure-determining methods may not be able address.

Improvements to the variety of Rosetta interfaces (RosettaScript, PyRosetta, and many web interfaces) allow the user a high degree of flexibility and personalization for each specific structural problem, as well as providing a previously unavailable entry point for novice users.

The current, default Rosetta score function (*talaris2014*) has been optimized and improved with new score terms as well as reweighted knowledge- and physics-based potentials. Rosetta also incorporates a new release of the Dunbrack rotamer library.[33]

De novo structure prediction has greatly improved with the implementation of the TopologyBroker, which was developed to create consensus sampling that satisfies all user-requested constraints without requiring additional code development for each unique system. Recent progress in comparative modeling applications has broadened the possible conformational search space by incorporating multiple starting templates. Protocols for protein–protein docking now include flexibility to modularize the coarse-grained and high-resolution modes of RosettaDock, giving the user more freedom to incorporate additional features in the docking process while narrowing the computational search space. Improvements in protein–small molecule docking utilize an improved *Transform* algorithm that increases both the speed and quality of this tool in obtaining more nativelike conformations. Likewise, the flexibility in incorporating experimentally derived constraints for most protocols has also greatly improved. To tackle the challenge of the inverse folding problem, new implementations of multistate design permit users to optimize sequences while considering several structures simultaneously.

Continuous developments in Rosetta have enhanced its utility by adding functionality to model proteins embedded in the membrane, expansion into nontraditional protein modeling by adding noncanonical amino acids, noncanonical backbones, and nucleic acids, and adding the ability to model ever-larger proteins by the addition of symmetry.

Installation and Licensing. The Rosetta licenses are available at http://www.rosettacommons.org/software free of charge for academic and governmental laboratories. Rosetta is compatible with most Unix-based operating systems and is distributed as source code. A user manual describing compilation, installation, and usage for the current release can be found at http://www.rosettacommons.org/docs/latest/. Demos and tutorials for additional Rosetta protocols can be found at http://www.rosettacommons.org/demos/latest/. Interested developers can join the RosettaCommons organization to contribute to the Rosetta software package.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.biochem.6b00444.

> Step-by-step tutorials for six of the protocols discussed in the paper (*de novo* folding, comparative modeling, protein–protein docking, protein–ligand docking, working with electron density, and protein design) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*Department of Chemistry, Vanderbilt University, 7330 Stevenson Center, Station B 351822, Nashville, TN 37235. E-mail: rocco.moretti@vanderbilt.edu. Telephone: +1 (615) 936-6594.

Author Contributions

B.J.B., A.C., A.M.D., J.A.F., D.F., A.D.L., B.K.M., A.K.S., M.F.S., and A.M.S. contributed equally to this work.

Funding

This work was supported by a grant from the National Institutes of Health (NIH) (R01GM073151) and the RosettaCommons. Work in the Meiler laboratory is supported by the NIH (R01 GM080403, R01 GM099842, R01 DK097376, R01 HL122010, R01 GM073151, and U19 AI117905) and the National Science Foundation (CHE 1305874).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge Ray Y.-R. Wang and Sam DeLuca for assistance in preparing the tutorials. The table of contents graphic incorporates a CC-BY-SA 4.0 licensed photo by Hans Hillewaert.

■ REFERENCES

(1) Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004) Protein structure prediction using Rosetta. *Methods Enzymol. 383*, 66–93.

(2) Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D. E., Kamisetty, H., Grishin, N. V., and Baker, D. (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife 4*, e09248.

(3) Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B. H., Das, R., Grishin, N. V., and Baker, D. (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Struct., Funct., Genet. 77* (Suppl. 9), 89–99.

(4) Combs, S. A., Deluca, S. L., Deluca, S. H., Lemmon, G. H., Nannemann, D. P., Nguyen, E. D., Willis, J. R., Sheehan, J. H., and Meiler, J. (2013) Small-molecule ligand docking into comparative models with Rosetta. *Nat. Protoc. 8*, 1277–1298.

(5) Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013) High-resolution comparative modeling with RosettaCM. *Structure 21*, 1735–1742.

(6) Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science 302*, 1364–1368.

(7) Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Struct., Funct., Genet. 79*, 830–838.

(8) Leaver-Fay, A., Jacak, R., Stranges, P. B., and Kuhlman, B. (2011) A generic program for multistate protein design. *PLoS One 6*, e20937.

(9) Borgo, B., and Havranek, J. J. (2012) Automated selection of stabilizing mutations in designed and natural proteins. *Proc. Natl. Acad. Sci. U. S. A. 109*, 1494–1499.

(10) Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M., Wilson, I. A., and Baker, D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science 332*, 816–821.

(11) Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., and Baker, D. (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature 501*, 212–216.

(12) Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol. 331*, 281–299.

(13) Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H., and Gray, J. J. (2011) Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One 6*, e22477.

(14) Raveh, B., London, N., and Schueler-Furman, O. (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Struct., Funct., Genet. 78*, 2029–2040.

(15) Raveh, B., London, N., Zimmerman, L., and Schueler-Furman, O. (2011) Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. *PLoS One 6*, e18934.

(16) Meiler, J., and Baker, D. (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins: Struct., Funct., Genet. 65*, 538–548.

(17) Davis, I. W., and Baker, D. (2009) RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol. 385*, 381–392.

(18) DeLuca, S., Khar, K., and Meiler, J. (2015) Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. *PLoS One 10*, e0132508.

(19) Das, R., and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U. S. A. 104*, 14664–14669.

(20) Das, R. (2013) Atomic-accuracy prediction of protein loop structures through an RNA-inspired Ansatz. *PLoS One 8*, e74830.

(21) Renfrew, P. D., Choi, E. J., Bonneau, R., and Kuhlman, B. (2012) Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. *PLoS One 7*, e32637.

(22) Drew, K., Renfrew, P. D., Craven, T. W., Butterfoss, G. L., Chou, F. C., Lyskov, S., Bullock, B. N., Watkins, A., Labonte, J. W., Pacella, M., Kilambi, K. P., Leaver-Fay, A., Kuhlman, B., Gray, J. J., Bradley, P., Kirshenbaum, K., Arora, P. S., Das, R., and Bonneau, R. (2013) Adding diverse noncanonical backbones to rosetta: enabling peptidomimetic design. *PLoS One 8*, e67051.

(23) Alford, R. F., Koehler Leman, J., Weitzner, B. D., Duran, A. M., Tilley, D. C., Elazar, A., and Gray, J. J. (2015) An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput. Biol. 11*, e1004398.

(24) André, I., Bradley, P., Wang, C., and Baker, D. (2007) Prediction of the structure of symmetrical protein assemblies. *Proc. Natl. Acad. Sci. U. S. A. 104*, 17656–17661.

(25) DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D., and André, I. (2011) Modeling symmetric macromolecular structures in Rosetta3. *PLoS One 6*, e20450.

(26) Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol. 487*, 545–574.

(27) Rohl, C. A., Strauss, C. E., Chivian, D., and Baker, D. (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins: Struct., Funct., Genet. 55*, 656–677.

(28) Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol. 268*, 209–225.

(29) Dunbrack, R. L., and Karplus, M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol. 230*, 543–574.

(30) Kuhlman, B., and Baker, D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A. 97*, 10383–10388.

(31) Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J., and Baker, D. (2011) Structure-guided forcefield optimization. *Proteins: Struct., Funct., Genet. 79*, 1898–1909.

(32) Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., Thompson, J., Davis, I. W., Pache, R. A., Lyskov, S., Gray, J. J., Kortemme, T., Richardson, J. S., Havranek, J. J., Snoeyink, J., Baker, D., and Kuhlman, B. (2013) Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol. 523*, 109–143.

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

(34) Bazzoli, A., Kelow, S. P., and Karanicolas, J. (2015) Enhancements to the Rosetta Energy Function Enable Improved Identification of Small Molecules that Inhibit Protein-Protein Interactions. *PLoS One 10*, e0140359.

(35) Yarov-Yarovoy, V., Schonbrun, J., and Baker, D. (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins: Struct., Funct., Genet. 62*, 1010−1025.

(36) Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E. M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., and Baker, D. (2011) RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One 6*, e20161.

(37) O'Meara, M. J., Leaver-Fay, A., Tyka, M., Stein, A., Houlihan, K., DiMaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., and Kuhlman, B. (2015) A Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput. 11*, 609−622.

(38) Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., and Meiler, J. (2010) Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry 49*, 2987−2998.

(39) Chaudhury, S., Lyskov, S., and Gray, J. J. (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics 26*, 689−691.

(40) Schenkelberg, C. D., and Bystroff, C. (2015) InteractiveROSETTA: a graphical user interface for the PyRosetta protein modeling suite. *Bioinformatics 31*, 4023−4025.

(41) Baugh, E. H., Lyskov, S., Weitzner, B. D., and Gray, J. J. (2011) Real-time PyMOL visualization for Rosetta and PyRosetta. *PLoS One 6*, e21931.

(42) Thornburg, N. J., Nannemann, D. P., Blum, D. L., Belser, J. A., Tumpey, T. M., Deshpande, S., Fritz, G. A., Sapparapu, G., Krause, J. C., Lee, J. H., Ward, A. B., Lee, D. E., Li, S., Winarski, K. L., Spiller, B. W., Meiler, J., and Crowe, J. E., Jr. (2013) Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. *J. Clin. Invest. 123*, 4405−4409.

(43) Jardine, J., Julien, J. P., Menis, S., Ota, T., Kalyuzhniy, O., McGuire, A., Sok, D., Huang, P. S., MacPherson, S., Jones, M., Nieusma, T., Mathison, J., Baker, D., Ward, A. B., Burton, D. R., Stamatatos, L., Nemazee, D., Wilson, I. A., and Schief, W. R. (2013) Rational HIV immunogen design to target specific germline B cell receptors. *Science 340*, 711−716.

(44) *PyMOL Molecular Graphics System* (2015) Schrödinger, LLC, Portland, OR.

(45) Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics 25*, 1422−1423.

(46) Lyskov, S., Chou, F. C., Conchúir, S., Der, B. S., Drew, K., Kuroda, D., Xu, J., Weitzner, B. D., Renfrew, P. D., Sripakdeevong, P., Borgo, B., Havranek, J. J., Kuhlman, B., Kortemme, T., Bonneau, R., Gray, J. J., and Das, R. (2013) Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). *PLoS One 8*, e63906.

(47) Adolf-Bryfogle, J., and Dunbrack, R. L., Jr. (2013) The PyRosetta Toolkit: a graphical user interface for the Rosetta software suite. *PLoS One 8*, e66856.

(48) Lewis, S. M., Wu, X., Pustilnik, A., Sereno, A., Huang, F., Rick, H. L., Guntas, G., Leaver-Fay, A., Smith, E. M., Ho, C., Hansen-Estruch, C., Chamberlain, A. K., Truhlar, S. M., Conner, E. M., Atwell, S., Kuhlman, B., and Demarest, S. J. (2014) Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface. *Nat. Biotechnol. 32*, 191−198.

(49) Kilambi, K. P., Pacella, M. S., Xu, J., Labonte, J. W., Porter, J. R., Muthu, P., Drew, K., Kuroda, D., Schueler-Furman, O., Bonneau, R., and Gray, J. J. (2013) Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20−27. *Proteins: Struct., Funct., Genet. 81*, 2201−2209.

(50) Sheffler, W., and Baker, D. (2009) RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci. 18*, 229−239.

(51) Kim, D. E., Chivian, D., and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res. 32*, W526−531.

(52) Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E., and Baker, D. (2011) Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS One 6*, e23294.

(53) Porter, J. R., Weitzner, B. D., and Lange, O. F. (2015) A Framework to Simplify Combined Sampling Strategies in Rosetta. *PLoS One 10*, e0138220.

(54) Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389−3402.

(55) Mandell, D. J., Coutsias, E. A., and Kortemme, T. (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods 6*, 551−552.

(56) Guilhot-Gaudeffroy, A., Froidevaux, C., Azé, J., and Bernauer, J. (2014) Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities. *PLoS One 9*, e108928.

(57) Kilambi, K. P., Reddy, K., and Gray, J. J. (2014) Protein-protein docking with dynamic residue protonation states. *PLoS Comput. Biol. 10*, e1004018.

(58) Lemmon, G., and Meiler, J. (2012) Rosetta Ligand docking with flexible XML protocols. *Methods Mol. Biol. 819*, 143−155.

(59) Gowthaman, R., Miller, S. A., Rogers, S., Khowsathit, J., Lan, L., Bai, N., Johnson, D. K., Liu, C., Xu, L., Anbanandam, A., Aubé, J., Roy, A., and Karanicolas, J. (2016) DARC: Mapping Surface Topography by Ray-Casting for Effective Virtual Screening at Protein Interaction Sites. *J. Med. Chem. 59*, 4152−4170.

(60) Lemmon, G., and Meiler, J. (2013) Towards ligand docking including explicit interface water molecules. *PLoS One 8*, e67536.

(61) Allison, B., Combs, S., DeLuca, S., Lemmon, G., Mizoue, L., and Meiler, J. (2014) Computational design of protein-small molecule interfaces. *J. Struct. Biol. 185*, 193−202.

(62) Lemmon, G., Kaufmann, K., and Meiler, J. (2012) Prediction of HIV-1 protease/inhibitor affinity using RosettaLigand. *Chem. Biol. Drug Des. 79*, 888−896.

(63) Kaufmann, K. W., and Meiler, J. (2012) Using RosettaLigand for small molecule docking into comparative models. *PLoS One 7*, e50769.

(64) Nguyen, E. D., Norn, C., Frimurer, T. M., and Meiler, J. (2013) Assessment and challenges of ligand docking into comparative models of G-protein coupled receptors. *PLoS One 8*, e67302.

(65) Zhang, Z., Porter, J., Tripsianes, K., and Lange, O. F. (2014) Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta. *J. Biomol. NMR 59*, 135−145.

(66) Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G., Aebersold, R., and Malmström, L. (2013) Cross-link guided molecular modeling with ROSETTA. *PLoS One 8*, e73411.

(67) Alexander, N. S., Stein, R. A., Koteiche, H. A., Kaufmann, K. W., McHaourab, H. S., and Meiler, J. (2013) RosettaEPR: rotamer library for spin label structure and dynamics. *PLoS One 8*, e72851.

(68) Hirst, S. J., Alexander, N., McHaourab, H. S., and Meiler, J. (2011) RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. *J. Struct. Biol. 173*, 506−514.

(69) Sgourakis, N. G., Lange, O. F., DiMaio, F., André, I., Fitzkee, N. C., Rossi, P., Montelione, G. T., Bax, A., and Baker, D. (2011) Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J. Am. Chem. Soc. 133*, 6288−6298.

(70) Schmitz, C., Vernon, R., Otting, G., Baker, D., and Huber, T. (2012) Protein structure determination from pseudocontact shifts using ROSETTA. *J. Mol. Biol. 416*, 668−677.

(71) Rossi, P., Shi, L., Liu, G., Barbieri, C. M., Lee, H. W., Grant, T. D., Luft, J. R., Xiao, R., Acton, T. B., Snell, E. H., Montelione, G. T., Baker, D., Lange, O. F., and Sgourakis, N. G. (2015) A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta. *Proteins: Struct., Funct., Genet. 83*, 309−317.

(72) Wang, R. Y., Kudryashev, M., Li, X., Egelman, E. H., Basler, M., Cheng, Y., Baker, D., and DiMaio, F. (2015) De novo protein

structure determination from near-atomic-resolution cryo-EM maps. *Nat. Methods 12*, 335−338.

(73) DiMaio, F., Echols, N., Headd, J. J., Terwilliger, T. C., Adams, P. D., and Baker, D. (2013) Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods 10*, 1102−1104.

(74) DiMaio, F., Song, Y., Li, X., Brunner, M. J., Xu, C., Conticello, V., Egelman, E., Marlovits, T. C., Cheng, Y., and Baker, D. (2015) Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods 12*, 361−365.

(75) Blok, N. B., Tan, D., Wang, R. Y., Penczek, P. A., Baker, D., DiMaio, F., Rapoport, T. A., and Walz, T. (2015) Unique double-ring structure of the peroxisomal Pex1/Pex6 ATPase complex revealed by cryo-electron microscopy. *Proc. Natl. Acad. Sci. U. S. A. 112*, E4017−4025.

(76) Kudryashev, M., Wang, R. Y., Brackmann, M., Scherer, S., Maier, T., Baker, D., DiMaio, F., Stahlberg, H., Egelman, E. H., and Basler, M. (2015) Structure of the type VI secretion system contractile sheath. *Cell 160*, 952−962.

(77) DiMaio, F., Yu, X., Rensen, E., Krupovic, M., Prangishvili, D., and Egelman, E. H. (2015) Virology. A virus that infects a hyperthermophile encapsidates A-form DNA. *Science 348*, 914−917.

(78) Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 66*, 213−221.

(79) Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 54*, 905−921.

(80) Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 67*, 355−367.

(81) Borshchevskiy, V., Round, E., Bertsova, Y., Polovinkin, V., Gushchin, I., Ishchenko, A., Kovalev, K., Mishin, A., Kachalova, G., Popov, A., Bogachev, A., and Gordeliy, V. (2015) Structural and functional investigation of flavin binding center of the NqrC subunit of sodium-translocating NADH:quinone oxidoreductase from Vibrio harveyi. *PLoS One 10*, e0118548.

(82) Jo, I., Chung, I. Y., Bae, H. W., Kim, J. S., Song, S., Cho, Y. H., and Ha, N. C. (2015) Structural details of the OxyR peroxide-sensing mechanism. *Proc. Natl. Acad. Sci. U. S. A. 112*, 6443−6448.

(83) Ahmad, M. U., Waege, I., Hausner, W., Thomm, M., Boos, W., Diederichs, K., and Welte, W. (2015) Structural Insights into Nonspecific Binding of DNA by TrmBL2, an Archaeal Chromatin Protein. *J. Mol. Biol. 427*, 3216−3229.

(84) Birkinshaw, R. W., Pellicci, D. G., Cheng, T. Y., Keller, A. N., Sandoval-Romero, M., Gras, S., de Jong, A., Uldrich, A. P., Moody, D. B., Godfrey, D. I., and Rossjohn, J. (2015) $\alpha\beta$ T cell antigen receptor recognition of CD1a presenting self lipid ligands. *Nat. Immunol. 16*, 258−266.

(85) Terwilliger, T. C., Dimaio, F., Read, R. J., Baker, D., Bunkoczi, G., Adams, P. D., Grosse-Kunstleve, R. W., Afonine, P. V., and Echols, N. (2012) phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. *J. Struct. Funct. Genomics 13*, 81−90.

(86) DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwai, H., Pokkuluri, P. R., and Baker, D. (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature 473*, 540−543.

(87) DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W., and Baker, D. (2009) Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol. 392*, 181−190.

(88) Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Röthlisberger, D., and Baker, D. (2006) New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci. 15*, 2785−2794.

(89) Willis, J. R., Briney, B. S., DeLuca, S. L., Crowe, J. E., and Meiler, J. (2013) Human germline antibody gene segments encode polyspecific antibodies. *PLoS Comput. Biol. 9*, e1003045.

(90) Sevy, A. M., Jacobs, T. M., Crowe, J. E., and Meiler, J. (2015) Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. *PLoS Comput. Biol. 11*, e1004300.

(91) Barth, P., Schonbrun, J., and Baker, D. (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc. Natl. Acad. Sci. U. S. A. 104*, 15682−15687.

(92) Sievers, S. A., Karanicolas, J., Chang, H. W., Zhao, A., Jiang, L., Zirafi, O., Stevens, J. T., Münch, J., Baker, D., and Eisenberg, D. (2011) Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature 475*, 96−100.

(93) Mills, J. H., Khare, S. D., Bolduc, J. M., Forouhar, F., Mulligan, V. K., Lew, S., Seetharaman, J., Tong, L., Stoddard, B. L., and Baker, D. (2013) Computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy. *J. Am. Chem. Soc. 135*, 13393−13399.

(94) Xu, J., Tack, D., Hughes, R. A., Ellington, A. D., and Gray, J. J. (2014) Structure-based non-canonical amino acid design to covalently crosslink an antibody-antigen complex. *J. Struct. Biol. 185*, 215−222.

(95) Renfrew, P. D., Craven, T. W., Butterfoss, G. L., Kirshenbaum, K., and Bonneau, R. (2014) A rotamer library to enable modeling and design of peptoid foldamers. *J. Am. Chem. Soc. 136*, 8772−8782.

(96) Butterfoss, G. L., Drew, K., Renfrew, P. D., Kirshenbaum, K., and Bonneau, R. (2014) Conformational preferences of peptide-peptoid hybrid oligomers. *Biopolymers 102*, 369−378.

(97) Cordero, P., Kladwang, W., VanLang, C. C., and Das, R. (2014) The mutate-and-map protocol for inferring base pairs in structured RNA. *Methods Mol. Biol. 1086*, 53−77.

(98) Cheng, C., Chou, F.-C., Kladwang, W., Tian, S., Cordero, P., and Das, R. (2014) MOHCA-seq: RNA 3D models from single multiplexed proximity-mapping experiments. *bioRxiv*, 004556.

(99) Miao, Z., Adamiak, R. W., Blanchet, M. F., Boniecki, M., Bujnicki, J. M., Chen, S. J., Cheng, C., Chojnowski, G., Chou, F. C., Cordero, P., Cruz, J. A., Ferré-D'Amaré, A. R., Das, R., Ding, F., Dokholyan, N. V., Dunin-Horkawicz, S., Kladwang, W., Krokhotin, A., Lach, G., Magnus, M., Major, F., Mann, T. H., Masquida, B., Matelska, D., Meyer, M., Peselis, A., Popenda, M., Purzycka, K. J., Serganov, A., Stasiewicz, J., Szachniuk, M., Tandon, A., Tian, S., Wang, J., Xiao, Y., Xu, X., Zhang, J., Zhao, P., Zok, T., and Westhof, E. (2015) RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA 21*, 1066−1084.

(100) Chou, F. C., Echols, N., Terwilliger, T. C., and Das, R. (2016) RNA Structure Refinement Using the ERRASER-Phenix Pipeline. *Methods Mol. Biol. 1320*, 269−282.

(101) Chou, F. C., Sripakdeevong, P., Dibrov, S. M., Hermann, T., and Das, R. (2012) Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat. Methods 10*, 74−76.

(102) Sripakdeevong, P., Cevec, M., Chang, A. T., Erat, M. C., Ziegeler, M., Zhao, Q., Fox, G. E., Gao, X., Kennedy, S. D., Kierzek, R., Nikonowicz, E. P., Schwalbe, H., Sigel, R. K., Turner, D. H., and Das, R. (2014) Structure determination of noncanonical RNA motifs guided by ^{1}H NMR chemical shifts. *Nat. Methods 11*, 413−416.

(103) Das, R., Karanicolas, J., and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods 7*, 291−294.

(104) Liu, Y., and Kuhlman, B. (2006) RosettaDesign server for protein design. *Nucleic Acids Res. 34*, W235−238.

(105) London, N., Raveh, B., Cohen, E., Fathi, G., and Schueler-Furman, O. (2011) Rosetta FlexPepDock web server–high resolution modeling of peptide-protein interactions. *Nucleic Acids Res. 39*, W249–253.

(106) Lauck, F., Smith, C. A., Friedland, G. F., Humphris, E. L., and Kortemme, T. (2010) RosettaBackrub–a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Res. 38*, W569–575.

(107) London, N., and Schueler-Furman, O. (2008) FunHunt: model selection based on energy landscape characteristics. *Biochem. Soc. Trans. 36*, 1418–1421.

(108) Gregory, K. J., Nguyen, E. D., Reiff, S. D., Squire, E. F., Stauffer, S. R., Lindsley, C. W., Meiler, J., and Conn, P. J. (2013) Probing the metabotropic glutamate receptor 5 (mGlu(5)) positive allosteric modulator (PAM) binding pocket: discovery of point mutations that engender a ″molecular switch″ in PAM pharmacology. *Mol. Pharmacol. 83*, 991–1006.

(109) Sripakdeevong, P., Kladwang, W., and Das, R. (2011) An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proc. Natl. Acad. Sci. U. S. A. 108*, 20573–20578.

Bangaru, S., Nieusma, T., Kose, N., Thornburg, N. J., Finn, J. A., Kaplan, B. S., King, H. G., Singh, V., Lampley, R. M., Sapparapu, G., Cisneros, A., Edwards, K. M., Slaughter, J. C., Edupuganti, S., Lai, L., Richt, J. A., Webby, R. J., Ward, A. B., … Crowe, J. E. (2016). Recognition of influenza H3N2 variant virus by human neutralizing antibodies. *JCI insight*, *1*(10), e86673.

Abstract:

Since 2011, over 300 human cases of infection, especially in exposed children, with the influenza A H3N2 variant (H3N2v) virus that circulates in swine in the US have been reported. The structural and genetic basis for the lack of protection against H3N2v induced by vaccines containing seasonal H3N2 antigens is poorly understood. We isolated 17 human monoclonal antibodies (mAbs) that neutralized H3N2v virus from subjects experimentally immunized with an H3N2v candidate vaccine. Six mAbs exhibited very potent neutralizing activity ($IC_{50} < 200$ ng/ml) against the H3N2v virus but not against current human H3N2 circulating strains. Fine epitope mapping and structural characterization of antigen-antibody complexes revealed that H3N2v specificity was attributable to amino acid polymorphisms in the 150-loop and the 190-helix antigenic sites on the hemagglutinin protein. H3N2v-specific antibodies also neutralized human H3N2 influenza strains naturally circulating between 1995 and 2005. These results reveal a high level of antigenic relatedness between the swine H3N2v virus and previously circulating human strains, consistent with the fact that early human H3 seasonal strains entered the porcine population in the 1990s and reentered the human population, where they had not been circulating, as H3N2v about a decade later. The data also explain the increased susceptibility to H3N2v viruses in young children, who lack prior exposure to human seasonal strains from the 1990s.

Acquiring co-crystal structures for each of the biologically-interesting HA-antibody interactions depicted in the study would be a time-consuming and expensive task. My contribution to this work centered around the homology modeling of various H3 hemagglutinin head domains using Rosetta. These models were used to visualize the epitopes for interpreting *in-vitro* binding data and inferring the mechanism through which antibodies that target HA head can neutralize the virus without inhibiting cell entry. After careful study we determined that antibodies like H3v-47 neutralize the virus by preventing budding.

JCI insight

RESEARCH ARTICLE

Recognition of influenza H3N2 variant virus by human neutralizing antibodies

Sandhya Bangaru,[1] Travis Nieusma,[2] Nurgun Kose,[3] Natalie J. Thornburg,[3,4] Jessica A. Finn,[1] Bryan S. Kaplan,[5] Hannah G. King,[3] Vidisha Singh,[3] Rebecca M. Lampley,[3] Gopal Sapparapu,[3,4] Alberto Cisneros III,[6] Kathryn M. Edwards,[4] James C. Slaughter,[3,7] Srilatha Edupuganti,[8,9] Lilin Lai,[8,9] Juergen A. Richt,[10] Richard J. Webby,[5] Andrew B. Ward,[2] and James E. Crowe Jr.[1,3,4,]

[1]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, Tennessee, USA. [2]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA. [3]Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, Tennessee, USA. [4]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, USA. [5]Infectious Diseases, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [6]Chemical and Physical Biology Program, Vanderbilt University University, Nashville, Tennessee, USA. [7]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA. [8]The Hope Clinic of Emory Vaccine Center, Emory University School of Medicine, Atlanta, Georgia, USA. [9]Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, Georgia, USA. [10]College of Veterinary Medicine, Kansas State University, Manhattan, Kansas, USA.

Since 2011, over 300 human cases of infection, especially in exposed children, with the influenza A H3N2 variant (H3N2v) virus that circulates in swine in the US have been reported. The structural and genetic basis for the lack of protection against H3N2v induced by vaccines containing seasonal H3N2 antigens is poorly understood. We isolated 17 human monoclonal antibodies (mAbs) that neutralized H3N2v virus from subjects experimentally immunized with an H3N2v candidate vaccine. Six mAbs exhibited very potent neutralizing activity (IC_{50} < 200 ng/ml) against the H3N2v virus but not against current human H3N2 circulating strains. Fine epitope mapping and structural characterization of antigen-antibody complexes revealed that H3N2v specificity was attributable to amino acid polymorphisms in the 150-loop and the 190-helix antigenic sites on the hemagglutinin protein. H3N2v-specific antibodies also neutralized human H3N2 influenza strains naturally circulating between 1995 and 2005. These results reveal a high level of antigenic relatedness between the swine H3N2v virus and previously circulating human strains, consistent with the fact that early human H3 seasonal strains entered the porcine population in the 1990s and reentered the human population, where they had not been circulating, as H3N2v about a decade later. The data also explain the increased susceptibility to H3N2v viruses in young children, who lack prior exposure to human seasonal strains from the 1990s.

Introduction

Annual outbreaks of influenza A viruses (IAVs) in humans are a major global health problem, causing more than 250,000 deaths every year (1). In addition to yearly epidemics, novel influenza viruses originating from other animals periodically cross the species barrier to humans and cause pandemics with high morbidity and mortality rates. IAVs are enveloped viruses that contain the antigenic hemagglutinin (HA) and neuraminidase (NA) surface glycoproteins. HA encodes the receptor-binding site (RBS) and fusion peptide essential for attachment and entry into the host cell and is the primary target for potent neutralizing antibodies (2). The globular head domain that contains the sialic acid–binding (SA-binding) pocket is the major antigenic portion of the HA and tolerates high sequence variability. As a consequence, influenza viruses undergo constant antigenic drift that allows escape from antibody-mediated immunity. There are currently 18 known subtypes of IAVs that fall into 2 broad groups based on the HA sequences and phylogeny (3). Of these, only H1 and H3 subtypes currently circulate in humans. Preferential binding of particular HA molecules to different types of SA receptors on host cells is the major determinant of host specificity (4). The HA of avian IAVs has high affinity for α 2,3–linked SA, whereas human influenza viruses have high affinity for α 2,6–linked SA (4–7).

The IAV genome is segmented, and the virus is capable of superinfecting cells with a heterologous IAV

Conflict of interest: The authors have declared that no conflict of interest exists.

Submitted: February 23, 2016
Accepted: June 2, 2016
Published: July 7, 2016

Reference information:
JCI Insight. 2016;1(10):e86673.
doi:10.1172/jci.insight.86673.

in a single animal. These features allow for reassortment of the influenza genome in intermediate hosts, such as swine or poultry, enabling emergence of strains that are capable of crossing the species barrier to humans (8). In particular, swine may act as a mixing vehicle for IAVs, because their upper respiratory tract epithelial cells possess both α 2,3- and α 2,6–linked SA receptors, which allow infection with both avian and human IAV (6). Although swine influenza viruses do not generally infect humans, sporadic cases of human infections with swine H1N1 and H3N2 have been documented since 1958 (9). Reassorted swine influenza viruses that are capable of infecting humans can cause severe disease and pose a pandemic threat due to lack of preexisting immunity to the virus. The H1N1 influenza pandemic in 2009–2010 was associated with a virus of swine origins and is an example of a swine virus that was able to transmit easily in the human population and cause disease (10).

Influenza viruses that circulate in pigs are designated "variant" viruses when they cause human infections. Swine-origin IAV H3N2v viruses containing the matrix gene from the 2009 H1N1 pandemic virus were first detected in humans in July 2011. Since then, there have been at least 345 reported cases of human infections with H3N2v viruses, with a high prevalence in children (11–13). A recent study showed that all children <5 years old and >80% up to 14 years old lack protective serum antibody titers against H3N2v (14). Most cases of H3N2v-associated disease have been associated with exposure to swine, with very limited human-human transmission (12). H3N2v is antigenically distinct from the currently circulating H3N2 seasonal strains, and it has been determined that vaccination with 2010–2011 annual trivalent inactivated virus does not induce neutralizing antibodies against the variant H3N2 virus (14). Lack of preexisting immunity to the variant virus, especially in children, may be a major concern if a highly transmissible H3N2v outbreak occurs (14–16).

Here, we describe the characterization of human mAbs to H3N2v HA isolated from individuals vaccinated with an experimental monovalent inactivated H3N2v vaccine candidate. We used these mAbs to define the molecular basis of strain specificity or cross-reactivity for human neutralizing antibodies recognizing the HA of H3 seasonal or emerging H3 variant viruses. The results indicate that polymorphisms in the 150 helix and the 190 loop, located near the RBS on HA, play a major role in escape of H3N2v virus from immunity induced by seasonal H3N2 vaccines. Furthermore, our results reveal that the HA protein of H3N2v strains is antigenically similar to the human H3N2 IAV strains that circulated during the late 1990s, during which several H3N2 spillover events have been suggested to occur from humans into US swine (17–19).

Results

Isolation of H3N2v-reactive human mAbs from vaccinated donors. Healthy adult donors received 2 doses of subvirion H3N2v vaccine (15 μg of HA/dose) 21 days apart in an open–label trial, the results of which were reported previously (20). Peripheral blood samples were obtained from volunteers after informed consent on the day of vaccination (day 0) and 21 days after the second dose of vaccine (day 42). Cryopreserved peripheral blood mononuclear cells (PBMCs) were immortalized by EBV transformation, and we collected supernatants from the resulting lymphoblastoid cell lines. Supernatants were screened by ELISA for binding to recombinant HA protein from the H3N2v strain A/Minnesota/11/2010 (designated here as the MNv strain) or one of two representative H3N2 seasonal strains, A/Victoria/361/2011 (designated here as the Victoria strain) or A/Wisconsin/67/2005. The frequency of H3N2 seasonal or variant-reactive B cell lines was reported previously in the description of the vaccine trial results, and those studies indicated a significant B cell response to the H3N2v HA among the vaccinated individuals (20). The majority of the B cells that secreted H3 HA-reactive antibodies on day 42 recognized the variant HA specifically, with limited cross-reactivity between variant virus-reactive antibodies secreted by B cells and those secreted by seasonal virus-reactive B cells (Supplemental Figure 1; supplemental material available online with this article; doi:10.1172/jci.insight.86673DS1). Transformed B cell lines with supernatants that showed reactivity against the MNv HA in ELISA were selected for fusion to generate human hybridoma cell lines secreting mAbs. We used PBMCs collected on study day 42 from a total of 12 subjects to isolate 36 cloned hybridomas secreting mAbs. The IgG subclass and light chain type of the 17 neutralizing antibodies are presented along with the donor number for the sample from which they were isolated (Figure 1).

Binding and neutralization profile of H3N2v mAbs. The neutralization potential of the clones in this antibody panel against MNv virus was determined by microneutralization assay using the MNv virus. Seventeen of thirty-six mAbs exhibited neutralizing activity against the variant virus when tested in concentrations as high as 10 μg/ml (Figure 1). The IC_{50} values are shown as a heat map, with increased color

Monoclonal antibody	Subject number	IgG subclass	Light chain	EC50 (ng/mL) for indicated strain				H3N2v IC50 (ng/mL)	H3N2v HAI (μg/mL)
				H3N2v Minnesota 2010	H3N2 Perth 2009	H3N2 Victoria 2011	H3N2 Texas 2012		
H3v-126	27	1	λ	3	>	>	>	3	0.16
H3v-71	37	1	λ	10	>	>	>	7	0.31
H3v-47	10	1	κ	4	18	26	38	8	>5.0
H3v-104	41	1	κ	14	>	>	>	9	0.16
H3v-98	37	1	κ	19	>	>	>	14	0.32
H3v-81	37	1	λ	29	>	>	>	59	1.25
H3v-141	54	1	κ	25	>	>	>	137	1.25
H3v-45	9	3	λ	14	>	>	>	257	1.25
H3v-95	37	3	λ	9	159	1	239	402	2.5
H3v-62	37	1	κ	4	>	1,625	>	438	>5.0
H3v-84	37	1	κ	2	88	5	135	482	>5.0
H3v-11	1	1	κ	32	>	144	>	564	5.0
H3v-21	1	1	κ	11	>	352	>	697	>5.0
H3v-86	37	1	κ	27	357	13	368	956	>5.0
H3v-7	1	1	κ	34	>	>	>	1,226	>5.0
H3v-9	2	1	κ	38	>	>	>	1,301	>5.0
H3v-79	37	1	λ	86	353	>	219	1,602	>5.0

Figure 1. Characterization of 17 neutralizing monoclonal antibodies (mAbs). The antibodies are arranged in the order of neutralization potency (column 9) with the most potent antibodies at the top. Seventeen mAbs isolated by human B cell hybridoma generation exhibited neutralization potential (shown as half-maximal inhibitory concentration [IC_{50}]) at <5 μg/ml against the H3N2v virus by microneutralization assay. Nine antibodies exhibited hemagglutinin inhibition (HAI) activity, indicating that they disrupt receptor-binding function of the virus. The mAbs were tested for binding against HA from H3N2v or 3 seasonal strains (shown as half-maximal effective concentration [EC_{50}]). The > symbol indicates that binding was not detected at the maximum concentration tested (2 μg/ml). The experiments for determining EC_{50} (n = 4), IC_{50} (n = 3), and HAI (n = 3) were conducted twice independently.

intensity corresponding to an increase in neutralizing potency. Six mAbs (found at the top of Figure 1 and designated as H3v-126, H3v-71, H3v-47, H3v-104, H3v-98, and H3v-81) showed very potent neutralization against the virus, with IC_{50} values of less than 100 ng/ml. We also determined the ability of 4 mAbs (H3v-98, H3v-104, H3v-71, and H3v-45) to neutralize 4 H3N2 strains representing each antigenic cluster circulating in swine, A/Swine/Texas/4199-2/98 (cluster I), A/Swine/Colorado/23619/99 (cluster II), A/Swine/Oklahoma/18089/99 (cluster III), and A/Ohio/13/2012 (cluster IV) (Supplemental Table 1). H3v-98 and H3v-104 showed potent inhibiting activity against 3 of 4 strains. H3v-71 and H3v-45 exhibited activity against 2 strains at low concentrations.

The 17 H3N2v-neutralizing mAbs were tested for binding to MNv and 3 seasonal strains A/Perth/16/2009, Victoria, and A/Texas/50/2012. All of the antibodies had half-maximal effective concentrations (EC_{50}) for binding below 100 ng/ml to MNv HA (Figure 1). Increasing intensity of the orange cell fill color in the EC_{50} column corresponds to increasing binding for the indicated HA. The antibodies displayed a differential binding pattern — 9 mAbs bound specifically to the MNv HA, while the other 8 bound to both the variant HA and a combination of seasonal HAs (Figure 1). Interestingly, all but one (H3v-47) of the potently neutralizing mAbs exhibited a variant-specific binding phenotype without detectable cross-reactivity for the HA of H3 seasonal strains (Figure 1) or other HA subtypes (Supplemental Table 2).

We determined the nucleotide sequence of the antibody heavy chain variable gene regions (Table 1). The 17 neutralizing antibodies had unique HCDR3 sequences, indicating that the mAbs represented independent clones. We also performed next-generation sequence analysis of antibody gene repertoires for a subset of donors (donors 10, 27, 37, and 41). Interestingly, we were able to identify on the day of immunization (day 0) a member of the clonal lineage for 2 of the 6 mAbs obtained from these donors, suggesting that the clone was induced by prior infection (Supplemental Table 3). The frequency of those clones was greatly expanded on day 7, the expected day of peak plasmablast circulation in peripheral blood (Supplemental Table 3). A clonal lineage of one of the clones, H3v-104, is shown in Supplemental Figure 2, with the amino acid sequence alignment of 296 clonal variants shown in Supplemental Table 4.

Potently neutralizing mAbs block the RBS. Due to the HA specificity demonstrated by the mAbs, we anticipated

Table 1. Antibody variable genes encoding 17 H3v-reactive human mAbs

mAb	Heavy chain			HCDR3 length	HCDR3 amino acid sequence
	V_H	D	J_H		
H3v-126	2-70D*04	6-6*01	4*02	16	ARTDGGSISSAAYFES
H3v-71	1-69*01	3-10*02	5*02	16	AREGLGSVIIGPWFDP
H3v-47	1-69*05	2-2*01	3*02	21	ARGASKVEPAAPAYSDAFDM
H3v-104	1-69*13	5-12*01	4*02	16	ARDYYRGEFSGYDFES
H3v-98	3-23*01	1-1*01	4*02	13	AKSSFTKGSPFDY
H3v-81	3-49*04	1-7*01	6*02	19	SREAANWNYPYHYSNGMDV
H3v-141	3-15*01	3-10*01	4*02	15	TTDNSFYYGSGYFDH
H3v-45	3-9*01	1-26*01	5*02	17	AKDGASGGTYYEAGFDP
H3v-95	3-21*01	5-18*01	3*02	15	ARDLSVYSYGGAFDI
H3v-62	3-30*04	2-15*01	5*02	14	ARRFCTGGSCYLDP
H3v-84	3-48*02	3-16*01	4*02	14	ARDGAVVFGVPFDT
H3v-11	1-18*01	2-21*01	4*02	17	ARRSRAWGLSKQGPLDY
H3v-21	1-2*02	5-18*01	4*02	14	ARGYNLGYLVLFDY
H3v-86	6-1*01	3-22*01	4*02	15	ARGIQHWMMVVAFDH
H3v-7	3-33*04	2-15*01	4*02	15	AKDRDGGVARAPLDY
H3v-9	4-34*01	3-3*01	4*02	20	ARGRPSDESWSGYLDNGFDF
H3v-79	1-69*02	6-19*01	6*02	22	AVRAFSTAVAGKGPWHYYGMDV

that the majority of the potently neutralizing mAbs mediated neutralization by binding to the less conserved head domain of HA. We performed hemagglutination inhibition (HAI) assays in order to identify mAbs that interfered with the receptor-binding function of HA. A total of 9 of the 17 neutralizing mAbs exhibited HAI activity against the MNv virus, suggesting that these mAbs function by blocking virus binding to the SA receptor (Figure 1). The most potent mAbs that displayed variant-specific binding phenotype also exhibited HAI activity, suggesting that the variant-specific polymorphisms around the RBS in the head domain region on MNv HA play a major role in determining the unique antigenic profile of the variant virus.

In vivo efficacy of H3v-126, H3v-104, and H3v-71. We tested representative mAbs as prophylaxis in a mouse challenge study. Groups of 6- to 8-week-old female DBA/2J mice at 5 animals per group were injected with 100 μg of individual mAbs by the intraperitoneal route on the day prior to virus challenge. Controls ($n = 10$) were injected with PBS. The modestly increased body weight for all 3 groups of mAb-treated animals when compared with the PBS-treated animals was not statistically different (Figure 2).

Competition-binding studies. In order to determine if the neutralizing mAbs bound to common or diverse epitopes on HA, we performed competition-binding assays using biolayer interferometry with all of the neutralizing antibodies. Fifteen of the seventeen neutralizing mAbs were classified into competition-binding groups based on their ability to block other mAbs from binding to the HA; we were unable to detect good binding signal for H3v-81 or H3v-7 with biolayer interferometry. The 15 neutralizing mAbs tested fell into 3 major competition-binding groups, with some overlap between the

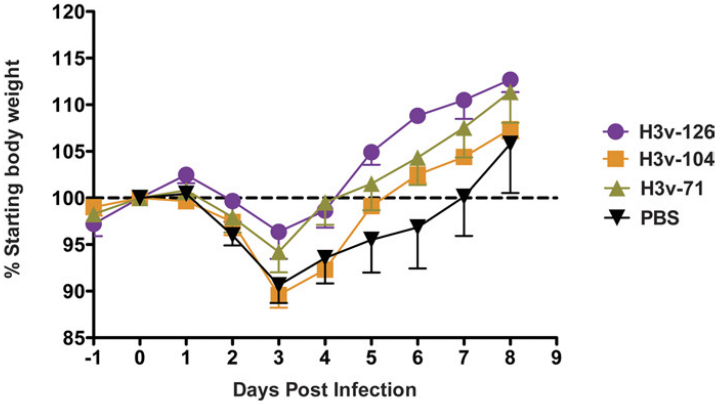
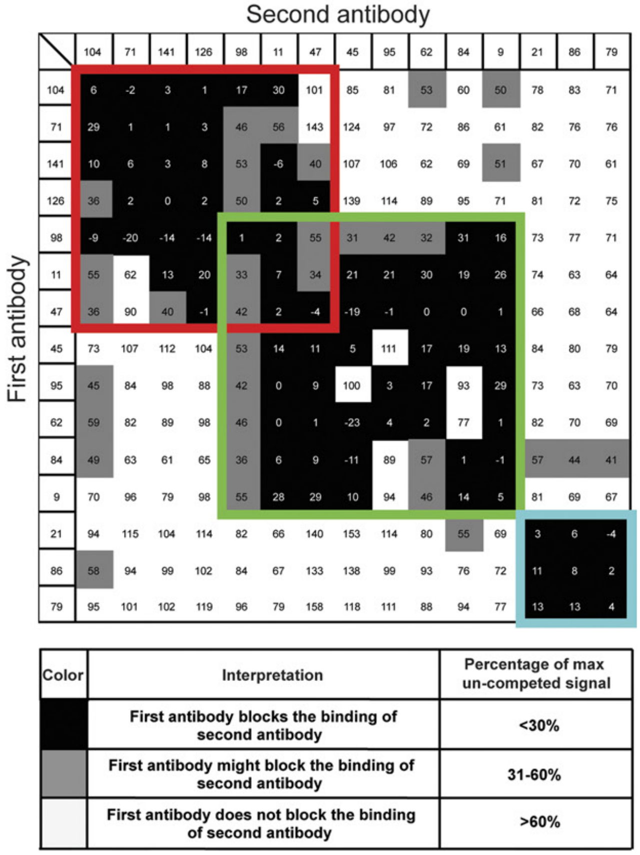


Figure 2. Prophylactic efficacy of H3v-104, H3v-126, and H3v-71 in mice. Groups of mice ($n = 5$) were treated with 100 μg of individual mAbs 24 hours before challenge with mouse-adapted A/Minnesota/11/2010 X203 virus. Controls ($n = 10$) were injected with PBS. The weight loss of mice was measured daily for 14 days after inoculation (day 0).


Second antibody
First antibody

	104	71	141	126	98	11	47	45	95	62	84	9	21	86	79
104	6	-2	3	1	17	30	101	85	81	53	60	50	78	83	71
71	29	1	1	3	46	56	143	124	97	72	86	61	82	76	76
141	10	6	3	8	53	-6	40	107	106	62	69	51	67	70	61
126	36	2	0	2	50	2	5	139	114	89	95	71	81	72	75
98	-9	-20	-14	-14	1	2	55	31	42	32	31	16	73	77	71
11	55	62	13	20	33	7	34	21	21	30	19	26	74	63	64
47	36	90	40	-1	42	2	-4	-19	-1	0	0	1	66	68	64
45	73	107	112	104	53	14	11	5	111	17	19	13	84	80	79
95	45	84	98	88	42	0	9	100	3	17	93	29	73	63	70
62	59	82	89	98	46	0	1	-23	4	2	77	1	82	70	69
84	49	63	61	65	36	6	9	-11	89	57	1	-1	57	44	41
9	70	96	79	98	55	28	29	10	94	46	14	5	81	69	67
21	94	115	104	114	82	66	140	153	114	80	55	69	3	6	-4
86	58	94	99	102	84	67	133	138	99	93	76	72	11	8	2
79	95	101	102	119	96	79	158	118	111	88	94	77	13	13	4

Color	Interpretation	Percentage of max un-competed signal
■	**First antibody blocks the binding of second antibody**	**<30%**
▨	**First antibody might block the binding of second antibody**	**31-60%**
□	**First antibody does not block the binding of second antibody**	**>60%**

Figure 3. Competition binding of neutralizing antibodies to H3N2v A/Minnesota/11/2010 hemagglutinin (HA) protein. Biolayer interferometry was used to perform competition-binding assays. The HA was loaded onto Ni-NTA tips, and binding to 2 successive antibodies was tested. The binding signal for each antibody was obtained from a single association step of the mAb onto HA. If binding of the first antibody blocked the binding of the second antibody by reducing its binding signal by more than 70%, it was defined as a competitor, indicated in black. The values in the table indicate the percentage of the maximum uncompeted binding signal. The red box indicates group 1, the binding group comprising the potently neutralizing mAbs. The green box and the blue box represent group 2 (partially overlaps with group 1) and group 3, respectively. The experiment was conducted twice independently.

groups designated groups 1 and 2 and a distinct group 3 (Figure 3). Notably, 4 of the most potent neutralizing clones (H3v-98, H3v-104, H3v-126, and H3v-71) that also exhibited HAI activity and variant specificity fell into the same competition-binding group, group 1.

Mutagenesis experiments and electron microscopy revealed regions on the H3N2v HA important for its immune escape from antibodies induced by seasonal H3N2 viruses. The 6 most potently neutralizing mAbs that displayed a variant-specific phenotype (i.e., mAbs that bound specifically to the MNv HA and not to the seasonal HAs) were chosen for fine epitope mapping. We sought to determine the residues important for immune escape of the H3N2v virus from antibodies induced by seasonal vaccines. We considered that the HA of MNv has 52 polymorphisms as compared with the Victoria seasonal strain. We initially performed a mutagenesis screen by introducing variant-specific polymorphisms into a cDNA encoding the Victoria HA protein to identify mutations that would enhance binding to the variant-specific mAbs. 17 variant-specific polymorphisms were introduced into Victoria HA as single or double mutations. Three double-mutant HA molecules, Victoria I202V/T203I, A163E/L164Q, and R142G/N144V, enhanced binding to mAbs H3v-98 and H3v-104 in comparison to wild-type Victoria HA (Supplemental Figure 3).

Based on the results of the initial mutagenesis screen, we targeted the region around the RBS and introduced residues from seasonal strains into the H3N2v HA to identify residues that disrupted binding to wild-type MNv. The EC_{50} values for MNv HA mutants that disrupt binding to each antibody are shown with representative binding curves (Figure 4). Mutation of Y155T/N156K/L157S disrupted binding to both mAb H3v-71 and mAb H3v-81, whereas mutant MNv N158G/Y159S/K160T did not bind to mAb H3v-71 and MNv K189N/T192I/N193R did not bind to H3v-81. A single-mutation Y137S disrupted binding to H3v-141. In addition, a triple alanine mutant, MNv L194A/Y195A/V196A, did not bind to mAbs H3v-98, H3v-104, H3v-71, H3v-81, and H3v-104.

We also performed electron microscopy (EM) of MNv HA in complex with Fab portions of the 3 most potent variant-specific antibodies, H3v-104, H3v-126, and H3v-71 (Figure 5, A–C), along with the Fab of a stem-binding antibody CR9114 (used as reference in each complex). The EM reconstructions revealed that all 3 antibodies bound to the RBS on HA, with overlapping footprints (Figure 5D). This finding was consistent with the competition-binding data and demonstrated that H3v-126, H3v-71, and H3v-104 all fall under the same competition-binding group (group 1). The H3v-104 mAb displayed a comparatively broader footprint on HA, extending below the RBS, supporting our previous observation that the Victoria R142G/N144V mutant enhanced binding to mAb H3v-104 but not to H3v-126 or H3v-71. H3v-71 bound

A

Graph Legend	H3N2v MN mutants	EC_{50} (μg/mL)					
		H3v-98	H3v-104	H3v-71	H3v-81	H3v-141	H3v-126
▲	MN wild-type	0.13	0.05	0.09	0.29	0.07	0.04
●	MN Y137S	0.13	0.06	0.09	0.34	>	0.04
●	MN Y155T/N156K/L157S	0.19	0.06	>	>	0.30	0.04
★	MN N158G/Y159S/K160T	0.30	0.08	>	0.32	0.45	0.04
▼	MN K189N/T192I/N193R	0.18	0.06	0.12	>	0.61	0.04
◆	MN L194A/Y195A/V196A	2.89	9.00	>	>	>	0.06

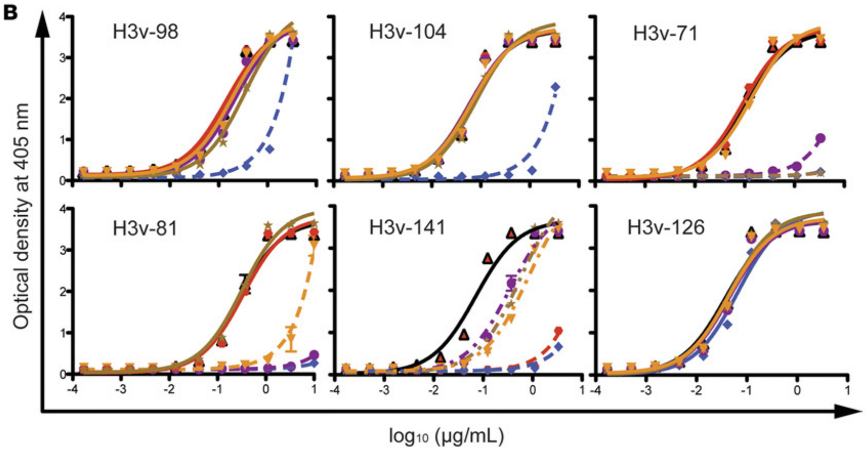


Figure 4. Binding of H3 variant-specific antibodies to mutated Minnesota hemagglutinin (HA) proteins. Mutagenesis of MNv HA was performed to determine antigenic residues important for recognition by variant-specific mAbs. Mutants of H3N2v A/Minnesota/11/2010 (MN) HA were generated by site-directed mutagenesis, and the half-maximal effective concentration (EC_{50}) values were determined by performing ELISA with serial dilutions of each antibody against the mutant HAs. The table (**A**) shows EC_{50} values and the graph (**B**) shows binding curves. The mutants that disrupted binding completely or decreased the EC_{50} by greater than 5-fold are represented as dashed or dotted lines, respectively, and are indicated by red EC_{50} values in the table. The > symbol indicates that binding was not detected at the maximum concentration tested (10 μg/ml). The experiments for determining the EC_{50} (n = 4) values were performed twice independently.

similarly to H3v-126 but appeared to interact more with the upper rim of the RBS (Figure 5D). This finding could explain why mutations in the 190 helix and 150 loop affected binding of H3v-71 but not of H3v-126. Collectively, the results from mutagenesis and EM structural studies indicated that variant-specific HA residues residing in 2 major structural features surrounding the RBS, the 150 loop and the 190 helix, principally account for antigenic distinction of the variant virus from current seasonal strains.

The HA of H3N2v virus is antigenically related to that of older human seasonal H3N2 IAVs. The majority of the severe cases of influenza infection caused by H3N2v viruses to date in the US have occurred in children, suggesting partial immunity in adults. We sought to identify the nature of preexisting immunity to the variant virus in the adult population that might provide partial protection against severe disease. Soluble HA proteins belonging to 12 different seasonal H3N2 strains that circulated in humans from 1968 to 2013 were cloned, expressed, and assayed for binding to 6 potently neutralizing mAbs with variant-specific binding and neutralizing phenotypes. Interestingly, 3 mAbs, H3v-98, H3v-104, and H3v-71, showed strong binding to H3N2 viruses that circulated between 1997 and 2004 but not to the HA of H3 subtype strains that circulated before or after that period (Supplemental Figure 4). To determine if these mAbs had the potential to neutralize these older seasonal strains, we performed microneutralization assays to test activity against 12 viruses that circulated between 1968 and 2013. Indeed, H3v-98 and H3v-104 neutralized all of the strains tested between 1995 and 2005, whereas H3v-71 neutralized all of the viruses between 1997

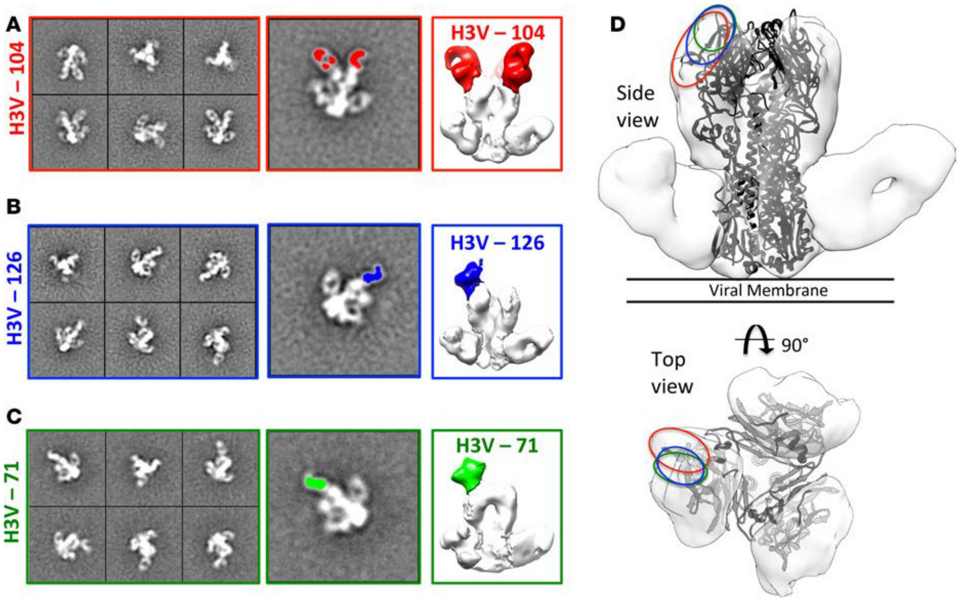


Figure 5. Negative stain EM images of hemagglutinin-Fab complexes. In each case the stem-binding antibody CR9114 was added to the complex in order to improve 3D reconstructions. (**A**) Reference-free 2D class averages of complex containing Fab 104 (left), single class average with Fab colored in red (middle), and 3D reconstruction (right). (**B**) Reference-free 2D class averages of complex containing Fab 126 (left), single class average with Fab colored in blue (middle), and 3D reconstruction (right). (**C**) Reference-free 2D class averages of complex containing Fab 71 (left), single class average with Fab colored in green (middle), and 3D reconstruction (right). (**D**) Side and top views of HA-Fab 126-CR9114 with Fab 126 removed and crystal structure of H3V (4FNK) fitted. Binding sites of the 3 antibodies described in **A**–**C** is highlighted using colors corresponding to Fabs.

and 2002 (Figure 6). We used a representative panel of live-virus H3 seasonal strains for the neutralization assays that matched as closely as possible the strains that we had used to make recombinant HA molecules for the seasonal strain HA-binding assays. Although we did not detect binding for mAbs H3v-98 or H3v-104 to the HA from A/New York/55/2004 or A/Hiroshima/52/2005, we did observe potent neutralization against the A/Wisconsin/67/2005 strain that was available for neutralization testing. In summary, prior infection with seasonal H3N2 strains that circulated during 2004 and 2005 might induce antibodies that cross-react with H3N2v viruses but likely does not guarantee the presence of protective antibodies against H3N2v because of antigenic heterogeneity.

Discussion

We report here the isolation of human mAbs against H3N2v virus and the use of them to determine the molecular basis for the antigenic distinction between H3 seasonal strains and the H3N2v viruses. Seventeen antibodies neutralized MNv H3N2v at concentrations of less than 10 μg/ml. Four of the seventeen neutralizing mAbs exhibited ultrapotent neutralizing activity (IC_{50} < 10 ng/ml). Three of these four ultrapotent neutralizing H3N2v mAbs also displayed potent inhibiting activity against swine H3N2 strains belonging to different antigenic clusters. Collectively, these results suggest that some of these H3N2v mAbs could be used in humans as therapeutics against many H3N2 strains circulating in swine in the case of zoonotic transmission event.

The H3 HA-binding breadth of IgG secreted by memory B cells from individuals vaccinated with monovalent inactivated H3N2v revealed that there is limited cross-reactivity between antibodies secreted by variant-reactive B cells and H3 seasonal virus-reactive B cells. About half of the H3v-reactive clones that were isolated displayed specific binding to the MNv HA, with no detectable cross-reactivity for HAs from 3 recent seasonal H3N2 strains. The potently neutralizing variant-specific antibodies clustered into the same competition-binding group and used the same virus neutralization mechanism by blocking the RBS on HA.

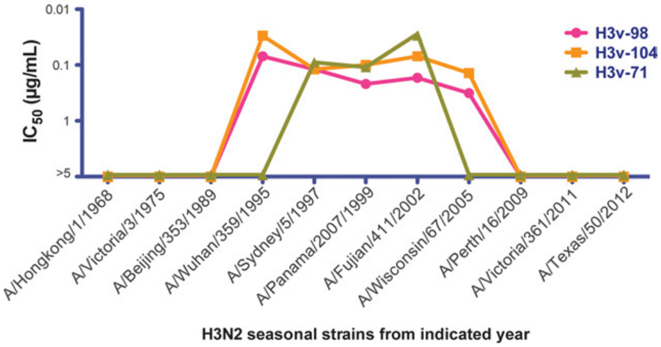


Figure 6. Neutralization of H3N2 seasonal strains. The monoclonal antibodies were tested for neutralization activity by microneutralization assay against 11 seasonal strains isolated between 1968 and 2012. The half-maximal inhibitory concentration (IC_{50}) values for H3v-98 (pink), H3v-104 (orange), or H3v-71 (green) are indicated on the *y* axis, and the H3N2 strains used for neutralization are indicated on the *x* axis. The IC_{50} values are represented as baseline if neutralization was not detected at any concentration less than 5 μg/ml. The microneutralization assay for determining the IC_{50} (*n* = 3) values was performed twice independently.

This finding suggested that common antibodies induced by seasonal trivalent inactivated influenza virus (TIV) vaccination that recognize immunodominant epitopes adjacent to the RBS on HA fail to contribute to cross-protection against variant virus. A recent study showed that the molecular basis for antigenic drift in human H3N2 seasonal strains from 1968 to 2006 was attributed to 7 single amino acid substitutions at positions 145, 155, 156, 158, 159, 189, and 193 in HA (21). Six of these residues are located in the antigenic site B (upper rim of RBS). Fine-epitope mapping of 5 potently neutralizing H3N2v-specific mAbs revealed a similar pattern in context to H3N2v antigenicity. The residues important for variant specificity of these mAbs were located primarily in the 150-loop and 190-helix antigenic elements near the RBS.

Historically, swine IAVs have caused sporadic human infections, but these outbreaks had been mostly self-limiting. The 2009 H1N1 pandemic was an example of a swine IAV variant, containing the same subtype of HA and NA as that of the circulating human strain, which caused a human pandemic. A recent rise in swine-origin H3N2v human infections has raised concerns of another swine-origin influenza virus pandemic (11, 13). Surveillance of IAVs in live-animal markets in Minnesota that included human, swine, and environmental samples provided evidence of interspecies transmission of IAVs from swine to human, signifying a potential risk to persons attending live markets (22). Moreover, H3N2 was identified as the predominant subtype circulating in the swine population; a majority of these swine H3N2 isolates had the same genomic constellation as the H3N2v viruses isolated previously from humans (22). Vaccination with seasonal TIV does not provide protection against H3N2v viruses, indicating these swine-origin viruses have the potential to spread within unprotected human populations (14).

The HA from the swine-origin 2009 pandemic H1N1 strain is antigenically similar to human H1N1 viruses that circulated between 1918 and 1943 (23). It is hypothesized that IAV classical swine H1N1 was introduced into the domestic swine population sometime during the 1918 human pandemic and remained relatively static in swine for greater than 80 years until the end of the 20th century (24, 25). In light of the 2009 pandemic, swine have been thought to act as a potential frozen reservoir of human IAVs, which can result in pandemics through reintroduction of the human-origin virus into a then susceptible human population at a remote time decades later (23). The reactivity of the human mAbs that we isolated from H3N2v vaccines revealed that the same phenomenon may be occurring now with H3 viruses, as human H3 seasonal strains from the past may be harbored in an antigenically static manner in farm swine in the US.

During late 1990s, there were several spillovers of H3N2 virus from humans to pigs resulting in the introduction of H3N2 viruses into the US swine population (17–19). The H3N2v viruses that were isolated from humans in 2011 were phylogenetically closer to seasonal strains that circulated during late 1990s than to current H3 human strains (19, 26). We hypothesized that the swine-origin H3N2v viruses might be antigenically related to the seasonal H3N2 viruses from late 1990s, and thus the H3N2v-specific antibodies that showed no reactivity against the current seasonal strains might recognize the older seasonal viruses. Indeed, we found that 3 potently neutralizing variant-specific mAbs (H3v-98, H3v-104, and H3v-71) had the ability to neutralize seasonal strains that circulated between 1997 and 2002. Additionally, H3v-98 and H3v-104 also neutralized strains from 1995 and 2005. These results indicate that the H3N2v viruses that caused human infections from 2011 to 2013 are antigenically related to the human ancestral H3N2 strains that circulated from 1995 to 2005. Taken together, the data suggest that children who were born after 2005 do not have any preexisting cross-reactive immunity against H3N2v virus because they were not exposed to the earlier H3 seasonal strains. Therefore, these children should be the priority target population for vaccination to prevent an H3N2v outbreak.

The phenomenon of antigenic similarity between swine IAVs and ancestral human strains might be explained by several mechanisms. First, swine may act as a relatively static reservoir of human IAVs. Second, differences in the location of the immunodominant antibody epitopes on HA for swine and humans might allow preservation of antigenic epitopes targeted by humans. We compared the HA1 sequences from H3N2v Minnesota virus and the seasonal Victoria 2011 virus with that of the Sydney 1997 strain (a possible ancestor to the H3N2v Minnesota virus). The HA1 subunit of H3N2v Minnesota and H3N2 Victoria strains has 37 or 32 polymorphisms, respectively, when compared with the HA1 sequence of H3N2 Sydney 1997, suggesting that these strains have drifted at a relatively similar rate in swine compared with the drift in humans. Thirteen of these positions that are situated primarily in the upper or lower rim of the RBS (antigenic sites A and B) showed variation in both Minnesota and Victoria HA. In contrast, there were comparatively more polymorphisms in the H3N2v Minnesota HA sequence in antigenic sites C and E, whereas polymorphisms in the 220-loop (antigenic site D) were seen predominantly in the Victoria 2011 strain. These host-based differences in the observed immunodominance pattern provide a possible explanation of why H3N2 viruses in the swine population retain antigenic features of their human ancestral strains. This recurring phenomenon, in which swine populations act as a static/divergent antigenic reservoir of previously circulating human IAVs should thus be taken into serious consideration for pandemic preparedness.

Methods

Influenza viruses. The seed stock of H3N2v strain MNv was obtained from Terrence Tumpey (US CDC). The working stocks used for microneutralization assay and HA inhibition assays were made from the supernatant of virus-infected MDCK cell culture monolayers in plain Dulbecco's Modified Eagle Medium (Gibco DMEM, Invitrogen, 11965) with 2 μg/ml of TPCK-trypsin. The seasonal H3N2 strains A/Fujian/411/2002 (FR-1146), A/Perth/16/2009 (FR-370), A/Wisconsin/67/2005 (FR-397), and A/Texas/50/2012 (FR-1210) were provided by the Influenza Reagent Resource (http://www.influenzareagentresource.org/) of the US CDC. Two H3N2 seasonal strains, Victoria (NR-44022) and A/Sydney/5/1997 (NR-12278), were obtained from BEI Resources.

Recombinant soluble HA proteins. Sequences encoding the HA genes of interest were optimized for expression, and cDNAs were synthesized (Genscript) as soluble trimeric constructs by replacing the transmembrane and cytoplasmic domain sequences with cDNAs encoding the GCN4 trimerization domain and a His-tag at the C-terminus. Synthesized genes were subcloned into the pcDNA3.1(+) mammalian expression vector (Invitrogen). HA protein was expressed by transient transfection of 293F cells with polyethylenimine transfection reagent and was grown in expression medium (Freestyle 293 Expression Medium; Invitrogen, 12338). The supernatants were harvested after 7 days, filter-sterilized with a 0.4-μm filter, and purified with HisTrap TALON FF crude columns (GE Healthcare Life Sciences).

PBMC isolation and hybridoma generation. A cohort of 25 donors was vaccinated twice, 21 days apart, with 15 μg of HA/0.5-ml dose of reassortant MNv NYMC X-203, as part of NIH-sponsored clinical research trials of this experimental vaccine (DMID protocol 12-0011). The details of the clinical trial were described previously (20). PBMCs from these donors were isolated from day 0 (day of first vaccination) and protocol day 42 (3 weeks after the second dose of vaccine) by density gradient separation on Ficoll and cryopreserved. The human lymphocytes from day 42 were thawed and immortalized by transformation with EBV substrain B95.8 in the presence of CpG10103, cyclosporin A, and a Chk2 inhibitor (27). The cells were plated in a 384-well plate, and 8 days later the supernatants from these transformed B cells were used to screen for the presence of antibodies that bound to soluble H3N2 MNv HA using capture ELISA. The positive wells containing B cells secreting anti-H3N2v antibodies were expanded onto irradiated human PBMC feeder layers for 4 days and then fused with HMMA2.5 myeloma cells using a Cytopulse PA4000 electrofusion device. After fusion, human hybridomas were selected in medium with HAT solution containing ouabain, and several rounds of limiting dilution passages were performed in 384-well culture plates to isolate cell lines of the hybridomas with the highest level of secretion of IgG (27).

mAb production and purification. The hybridoma cell lines with the highest level for IgG expression for each clone were selected as single cells using flow cytometric sorting to obtain clones secreting mAbs. Once hybridoma clones were obtained following sorting and growth in a 384-well plate, we expanded them first into wells of a 48-well plate and then further into a 75-cm^2 flask to 70% confluency in hybridoma growth medium (ClonaCell-HY medium E from STEMCELL Technologies, 03805). The cells then were washed and expanded equally to four 225-cm^2 flasks for antibody expression in serum-free medium (GIBCO Hybridoma-SFM, Invitrogen, 12045084). The supernatant was harvested after 3 weeks, filtered with a 0.4-μm filter, and the monoclonal IgGs were purified by affinity chromatography using protein G columns (GE Life Sciences, Protein G HP Columns).

EC_{50} binding analysis. The EC_{50} concentration for each antibody was determined as described previously (28). Briefly, we performed ELISA using plates coated with the HA of interest at 2 μg/ml overnight at 4°C and then blocked with 5% nonfat dry milk, 2% goat serum, and 0.1% Tween-20 in PBS for 1 hour. Three-fold dilutions of the mAb starting from 10 μg/ml were added to the wells and incubated for 1 hour, followed by incubation for 1 hour at 1:4,000 dilution of anti-human IgG alkaline phosphatase conjugate (Meridian Life Science, W99008A). The plates were washed 3 times between each step with PBS containing 0.1% Tween-20. Phosphatase substrate solution (1 mg/ml p-nitrophenol phosphate in 1 M Tris aminomethane) was added to the plates and incubated for 1 hour, and the optical density values were measured at 405-nm wavelength on a BioTek plate reader. Each dilution was done in triplicate, and the EC_{50} values were calculated in Prism software (GraphPad) using nonlinear regression analysis.

HAI and neutralization assays to determine IC_{50} values. Neutralization potential of all of the mAbs was determined by microneutralization assay and HAI assay. For microneutralization, 50 μl of 2-fold serial dilutions of each antibody starting at 20 μg/ml was incubated with 50 μl of 100 $TCID_{50}$ of the virus in viral growth medium (VGM) for 1 hour at room temperature. VGM consists of plain DMEM with 2 μg/ml of TPCK-trypsin and 50 μg/ml gentamicin. The MDCK cell monolayer cultures were washed 2 times with 100 μl PBS containing 0.1% Tween-20, and the virus-antibody mixture then was added to cells and incubated for 32 hours at 37°C. The cells were washed again and fixed with 100 μl of 80% methanol/20% PBS. The presence of influenza nucleoprotein in the fixed cells was determined by ELISA using a 1:8,000 dilution of mouse anti-NP antibody (BEI Resources, NR 4282) as the primary antibody and a 1:4,000 dilution of goat anti-mouse alkaline phosphate conjugate as the secondary antibody (ThermoFisher Scientific, 31320). Each dilution was tested in duplicate and the half-maximal inhibitory concentration (IC_{50}) was determined by nonlinear regression analysis of $\log_{10}$(inhibitor) vs. response function, using Prism software (GraphPad). An IC_{50} value of 2 μg/ml was used as the threshold to determine the presence of functional neutralization. For performing the HAI assay, we used turkey red blood cells (Rockland) that were diluted to 0.5% in Alsever's solution (Sigma, A3551). 25 μl of 4 hemagglutination units of virus were incubated with 25 μl of 2-fold dilutions of the mAb, starting at 10 μg/ml in PBS for 1 hour at 37°C. 50 l of the virus-antibody mixture was incubated with turkey red blood cells for 1 hour at room temperature. The HAI titer was defined as the highest dilution of antibody that inhibited hemagglutination of red blood cells. Each dilution was performed in duplicate.

In vivo efficacy of H3N2v mAbs. On the day prior to infection, 6- to 8-week-old female DBA/2J mice (Jackson Laboratories, 000671) (n = 5) were injected by i.p. route with 1, 10, or 100 μg of antibody H3v-71, H3v-104, or H3v-126 in 50 μl PBS. Controls were treated with PBS (n = 10). On day 0, mice were inoculated intranasally with 10^7 PFU of A/Minnesota/10/11 X-203 virus under isoflurane anesthesia and then weighed daily. All treated mice survived for the duration of the study. One animal treated with PBS alone died.

Competition-binding groups. Biolayer interferometry using an Octet Red instrument (ForteBio) was used to confirm mAb-HA binding and to perform competition-binding assays. The HA was loaded onto ForteBio Ni-NTA tips at a concentration of 25 μg/ml, and binding to 2 successively applied mAbs at 100 μg/ml was tested. All of the dilutions were made in 1X kinetic buffer (ForteBio, 18-5032). The actual binding signal for each mAb was obtained after 300 seconds of a single association step of the mAb on to HA. If binding of the first antibody blocked the binding of the second antibody by reducing its actual binding signal by more than 70%, it was defined as a competitor. If binding of the first antibody did not block the binding of the second antibody by reducing its actual binding signal by less than 30%, it was defined as a noncompetitor. A signal reduction between 30% and 70% was defined as partial blocking.

Site-directed mutagenesis of genes encoding HA proteins. Primers for site-directed mutagenesis were designed using the Agilent QuikChange Primer Design program (Agilent Technologies). The QuikChange Lightning Multi-site Mutagenesis kit (Agilent, 210515-5) was used to introduce multiple mutations into cDNAs encoding the HA genes of H3N2 MNv or H3N2 Victoria, according to the manufacturer's instructions. These mutant HAs were tested for antibody binding in ELISAs, as above, to determine EC_{50} values for binding and to identify amino acids that comprise the epitope.

EM. SEC-purified MNv complexed with Fab H3v-104 and Fab CR9114 was diluted to 14.7 μg/ml, applied to freshly glow-discharged 400-mesh carbon-coated copper grids, and negatively stained with 2% uranyl formate. Similar complexes containing MNv and Fab CR9114 were then prepared separately with both Fab H3v-126 and Fab H3v-71 at 7.2 μg/ml and 5.8 μg/ml, respectively. The complexes containing Fab H3v-104 or Fab H3v-126 were imaged at ×92,000 magnification on an FEI Talos at 200 keV, resulting in a pixel size of 1.57 Å/pixel (calibrated using catalase crystal diffraction), with a dose of 23.96 e/Å^2. The complex containing Fab H3v-71 was imaged at ×52,000 magnification on an FEI Tecnai T12 at 120 kV TEM, resulting in a pixel size of 2.05 Å/pixel, with a dose of 25.41 e/Å^2. All data were collected using Leginon Multi-Scale Imaging (MSI-raster 3.1) software (29). The Talos is equipped with an FEI Ceta 4 k × 4 k CMOS and the T12 TEM is equipped with a Teitz F416 4 k × 4 k CMOS.

EM data processing. For the complex containing Fab H3v-104 DoGpicker was used to automatically select particles from 450 raw micrographs that were then binned by 2, resulting in a 3.14 Å/pixel size, and placed into 128 × 128 pixel boxes (30). Particles were aligned with Iterative MRA-MSA and ISAC, resulting in a final stack of 3,618 raw particles (31, 32). Class averages from ISAC were used to create a common lines initial model in EMAN2 (33). Model refinement was conducted in EMAN, resulting in a 21.3 Å resolution reconstruction based on a 0.5 FSC cutoff value (34). The same processing pipeline was used to prepare a reconstruction of the complex containing Fab H3v-126 from 306 raw micrographs and a final stack of 3,202 particles. Because of variable occupancy of the Fab several rounds of MRA-MSA were conducted to isolate particles containing a single-bound Fab, H3v-126. A model of HA bound with Fab CR9114 only was used as an initial model for refinement, resulting in a resolution of 19.8 Å. The complex containing Fab H3v-71 also had variable Fab occupancy. Five rounds of MRA-MSA and one round of ISAC were conducted to isolate particles with only 1 Fab H3v-71 bound, resulting in a final stack of 3,931 particles in 192 pixel boxes. A model of HA bound with Fab CR9114 only was used as an initial model for refinement, resulting in a resolution of 27.2 Å.

Antibody heavy and light chain variable gene sequence analysis. Antibody heavy and light chain genes for each of the neutralizing mAbs were cloned from the hybridoma lines after single-cell flow cytometric sorting to biologically clone the cell lines. RNA was extracted from these hybridoma clones using the RNeasy mini kit (Qiagen, 74106), followed by RT-PCR amplification of antibody gene cDNAs. PCR products encoding antibody heavy or light chain genes were cloned individually into the pGEM-T vector, and Sanger nucleotide sequence analysis was used to determine the antibody cDNA sequences. Analysis of variable gene sequences was performed using the international ImMunoGeneTics (IMGT) information system (http://imgt.org/). We also performed next-generation sequence analysis of antibody gene repertoires from 4 selected donors and defined in detail clonal lineage for 1 clone. First, antibody heavy chain variable gene sequences were obtained using RT-PCR and Illumina MiSeq 2 × 300 paired-end amplicon sequence analysis from PBMCs collected on day 0, 7, or 42 after immunization. For donor 41 and the H3v-104 clone, we identified those sequences sharing use of the V_H1-69 and J_H4 gene segments that encode H3v-104; the complete repertoire sequence database of donor 41 is available at the NCBI Sequence Read Archive at accession SRP075907. Then, CD-HIT was used to identify the HCDR3 sequences of 296 sequences that clustered within 85% identity with the HCDR3 sequence of H3v-104. We built a Phylip lineage using the alakazam and shazam packages of the Change-O software suite and visualized the network in Cytoscape.

Statistics. The IC_{50} values were calculated after log transformation of antibody concentrations using a 3-parameter nonlinear fit analysis of antibody $\log_{10}$ concentration vs. response with R^2 values > 0.85. The EC_{50} values were calculated after log transformation of antibody concentrations using sigmoidal dose-response nonlinear fit analysis with R^2 values of > 0.85. Change in mouse weights over days 1–8 were compared by antibody using repeated-measures analysis of covariance, controlling for weight at day 0. All statistics were analyzed using Prism software version 5 (GraphPad). A *P* value less than 0.05 was considered significant.

Study approval. PBMCs were collected at the Emory University Vaccine Treatment and Evaluation Unit after informed consent from otherwise healthy subjects with prior history of experimental H3N2v subunit vaccination, as described in the Methods. The protocol and consent form were approved prior to study by the Emory University Institutional Review Board Committee, Atlanta, Georgia, USA. The animal protocol covering the H3N2v influenza virus challenge infections of passively immunized mice was reviewed and approved by the Institutional Animal Care and Use Committee at St. Jude Children's Research Hospital.

Author contributions

SB, TN, JAF, NJT, BSK, AC, JAR, RJW, ABW, and JEC designed the research study and experiments. SB, TN, JAF, NK, NJT, BSK, HGK, VS, RML, GS, AC, JAR, and ABW conducted laboratory experiments and acquired and analyzed data. BSK and RJW conducted animal studies. JCS gave statistical review. SE, LL, and KME conducted the parental vaccine study and acquired PBMCs. SB and JEC wrote the manuscript. All authors read and commented on the manuscript.

Acknowledgments

The authors thank F. Smith-House (Vanderbilt) and H. Liu (Kansas State) for technical support, Mark J. Mulligan (Emory) for clinical support during human sample acquisition, and Jill Janssen of the Vanderbilt Clinical Trials Center for regulatory support. The original clinical trial of the H3N2v vaccine (DMID protocol 12-0011) was supported by NIAID, NIH and performed in the Emory Vaccine Center. SB was supported by the Vanderbilt International Scholar Program. This work was supported by NIH grants R01 AI106002 and U19 AI117905; NIAID contracts HHSN272200900047C and HHSN272201400007C (to JEC), HHSN27220800007C (KME), and HHSN272200800005C and HHSN266200700006C (to BSK, RJW, and JAR); and the Vanderbilt NIH Clinical and Translational Science Award UL1 RR024975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Address correspondence to: James E. Crowe Jr., Vanderbilt Vaccine Center, Vanderbilt University Medical Center, 11475 MRB IV, 2213 Garland Avenue, Nashville, Tennessee 37232-0417, USA. Phone: 615.343.8064; E-mail address: james.crowe@vanderbilt.edu.

1. World Health Organization. Influenza (Seasonal). WHO Web site. http://www.who.int/mediacentre/factsheets/fs211/en/. Accessed June 13, 2016.
2. Dreyfus C, et al. Structure of a classical broadly neutralizing stem antibody in complex with a pandemic H2 influenza virus hemagglutinin. *J Virol*. 2013;87(12):7149–7154.
3. Tong S, et al. New world bats harbor diverse influenza A viruses. *PLoS Pathog*. 2013;9(10):e1003657.
4. Suzuki Y, et al. Sialic acid species as a determinant of the host range of influenza A viruses. *J Virol*. 2000;74(24):11825–11831.
5. Rogers GN, Paulson JC. Receptor determinants of human and animal influenza virus isolates: Differences in receptor specificity of the H3 hemagglutinin based on species of origin. *Virology*. 1983;127(2):361–373.
6. Ito T, et al. Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *J Virol*. 1998;72(9):7367–7373.
7. Connor RJ, et al. Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. *Virology*. 1994;205(1):17–23.
8. Brown IH. The pig as an intermediate host for influenza A viruses between birds and humans. *Int Congr Ser*. 2001;1219:173–178.
9. Myers KP, et al. Cases of swine influenza in humans: a review of the literature. *Clin Infect Dis*. 2007;44(8):1084–1088.
10. Smith GJD, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*. 2009;459(7250):1122–1125.
11. Centers for Disease Control Prevention. Update: Influenza A (H3N2)v transmission guidelines — five states, 2011. *MMWR Morb Mortal Wkly Rep*. 2012;60(51–52):1741–1744.
12. Finelli L, Swerdlow DL. The emergence of influenza A (H3N2)v virus: what we learned from the first wave. *Clin Infect Dis*. 2013;57(suppl 1):S1–S3.
13. Lindstrom S, et al. Human infections with novel reassortant influenza A(H3N2)v viruses, United States, 2011. *Emerging Infect Dis*. 2012;18(5):834–837.
14. Skowronski DM, et al. Cross-reactive and vaccine-induced antibody to an emerging swine-origin variant of influenza A virus subtype H3N2 (H3N2v). *J Infect Dis*. 2012;206(12):1852–1861.
15. Centers for Disease Control Prevention. Antibodies cross-reactive to influenza A (H3N2) variant virus impact of 2010-11 seasonal influenza vaccine on cross-reactive antibodies-United States. *MMWR Morb Mortal Wkly Rep*. 2012;61(14):237–241.
16. Waalen K, et al. Age-dependent prevalence of antibodies cross-reactive to the influenza A (H3N2) variant virus in sera collected in Norway in 2011. *Euro Surveill*. 2012;17(19): 20170.
17. Zhou NN, et al. Emergence of H3N2 reassortant influenza A viruses in North American pigs. *Vet Microbiol*. 2000;74(1–2):47–58.
18. Zhou NN, et al. Genetic reassortment of avian, swine, and human influenza A viruses in American pigs. *J Virol*. 1999;73(10):8851–8856.
19. Webby RJ, et al. Evolution of swine H3N2 influenza viruses in the United States. *J Virol*. 2000;74(18):8243–8251.
20. Keitel WA, et al. Safety and immunogenicity of a subvirion monovalent unadjuvanted inactivated influenza A/H3N2 variant (H3N2v) vaccine in healthy persons ≥ 18 years old. *J Infect Dis*. 2015;212(4):552–561.
21. Koel BF, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*. 2013;342(6161):976–979.
22. Choi MJ, et al. Live animal markets in Minnesota: A potential source for emergence of novel influenza A viruses and interspecies Transmission. *Clin Infect Dis*. 2015;61(9):1355–1362.

23. Manicassamy B, et al. Protection of mice against lethal challenge with 2009 H1N1 influenza A virus by 1918-like and classical swine H1N1 based vaccines. *PLoS Pathog*. 2010;6(1):e1000745.
24. Nelson MI, et al. Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog*. 2008;4(2):e1000012.
25. Vincent AL, Ma W, Lager KM, Janke BH, Richt JA. Swine influenza viruses a North American perspective. *Adv Virus Res*. 2008;72:127–154.
26. Lina B, et al. S-OtrH3N2 viruses: use of sequence data for description of the molecular characteristics of the viruses and their relatedness to previously circulating H3N2 human viruses. *Euro Surveill*. 2011;16(50):20039.
27. Smith SA, et al. Persistence of circulating memory B cell clones with potential for dengue virus disease enhancement for decades following infection. *J Virol*. 2012;86(5):2665–2675.
28. Thornburg NJ, et al. Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. *J Clin Invest*. 2013;123(10):4405–4409.
29. Suloway C, et al. Automated molecular microscopy: The new Leginon system. *J Struct Biol*. 2005;151(1):41–60.
30. Voss NR, et al. DoG Picker and TiltPicker: Software tools to facilitate particle selection in single particle electron microscopy. *J Struct Biol*. 2009;166(2):205–213.
31. Hohn M, et al. SPARX, a new environment for Cryo-EM image processing. *J Struct Biol*. 2007;157(1):47–55.
32. van Heel M, et al. A new generation of the IMAGIC image processing system. *J Struct Biol*. 1996;116(1):17–24.
33. Tang G, et al. EMAN2: An extensible image processing suite for electron microscopy. *J Struct Biol*. 2007;157(1):38–46.
34. Ludtke SJ, et al. EMAN: Semiautomated software for high-resolution single-particle reconstructions. *J Struct Biol*. 1999;128(1):82–97.

Finn, J. A., Koehler Leman, J., Willis, J. R., Cisneros, A., Crowe, J. E., & Meiler, J. (2016). Improving Loop Modeling of the Antibody Complementarity-Determining Region 3 Using Knowledge-Based Restraints. *PloS one*, *11*(5), e0154811.

doi:10.1371/journal.pone.0154811

Abstract:

Structural restrictions are present even in the most sequence diverse portions of antibodies, the complementary determining region (CDR) loops. Previous studies identified robust rules that define canonical structures for five of the six CDR loops, however the heavy chain CDR 3 (HCDR3) defies standard classification attempts. The HCDR3 loop can be subdivided into two domains referred to as the "torso" and the "head" domains and two major families of canonical torso structures have been identified; the more prevalent "bulged" and less frequent "non-bulged" torsos. In the present study, we found that Rosetta loop modeling of 28 benchmark bulged HCDR3 loops is improved with knowledge-based structural restraints developed from available antibody crystal structures in the PDB. These restraints restrict the sampling space Rosetta searches in the torso domain, limiting the φ and ψ angles of these residues to conformations that have been experimentally observed. The application of these restraints in Rosetta result in more native-like structure sampling and improved score-based differentiation of native-like HCDR3 models, significantly improving our ability to model antibody HCDR3 loops.

My contribution to this study involved building and preparing the benchmark set and the design of the experiment. The initial, PDB-derived structures were processed, in some cases repacked, before they were used as templates for the *de novo* modeling of HCD3 loops.

RESEARCH ARTICLE

Improving Loop Modeling of the Antibody Complementarity-Determining Region 3 Using Knowledge-Based Restraints

Jessica A. Finn[1], Julia Koehler Leman[7], Jordan R. Willis[5], Alberto Cisneros, III[5], James E. Crowe, Jr[1,3,5,6], Jens Meiler[2,4,5]*

1 Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville, Tennessee, United States of America, **2** Department of Chemistry, Vanderbilt University, Nashville, Tennessee, United States of America, **3** Department of Pediatrics, Vanderbilt University, Nashville, Tennessee, United States of America, **4** Department of Pharmacology, Vanderbilt University, Nashville, Tennessee, United States of America, **5** Chemical and Physical Biology Program, Vanderbilt University, Nashville, Tennessee, United States of America, **6** Vanderbilt Vaccine Center, Vanderbilt University, Nashville, Tennessee, United States of America, **7** Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland, United States of America

* jens@meilerlab.org

OPEN ACCESS

Citation: Finn JA, Koehler Leman J, Willis JR, Cisneros A, III, Crowe JE, Jr, Meiler J (2016) Improving Loop Modeling of the Antibody Complementarity-Determining Region 3 Using Knowledge-Based Restraints. PLoS ONE 11(5): e0154811. doi:10.1371/journal.pone.0154811

Editor: Mitchell Ho, National Cancer Institute, NIH, UNITED STATES

Received: February 4, 2016

Accepted: April 19, 2016

Published: May 16, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the National Institute of Allergy and Infectious Diseases grants R56 AI110750 and U19 AI117905, and the National Institute of Allergy and Infectious Diseases Contract HHSN272201400024C. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Structural restrictions are present even in the most sequence diverse portions of antibodies, the complementary determining region (CDR) loops. Previous studies identified robust rules that define canonical structures for five of the six CDR loops, however the heavy chain CDR 3 (HCDR3) defies standard classification attempts. The HCDR3 loop can be subdivided into two domains referred to as the “torso” and the “head” domains and two major families of canonical torso structures have been identified; the more prevalent “bulged” and less frequent “non-bulged” torsos. In the present study, we found that Rosetta loop modeling of 28 benchmark bulged HCDR3 loops is improved with knowledge-based structural restraints developed from available antibody crystal structures in the PDB. These restraints restrict the sampling space Rosetta searches in the torso domain, limiting the φ and ψ angles of these residues to conformations that have been experimentally observed. The application of these restraints in Rosetta result in more native-like structure sampling and improved score-based differentiation of native-like HCDR3 models, significantly improving our ability to model antibody HCDR3 loops.

Introduction

The field of antibody-mediated immunity has long benefited from structural studies of protein-protein interactions, in most cases through the determination of co-crystal structures of antibodies in complex with their antigens. Such studies often reveal the molecular mechanism of pathogen neutralization [1–4]. However, the size and complexity of the antibody repertoire coupled with the substantial resources needed for experimental structure determination prohibit such studies on a comprehensive scale. B cell development leads to the generation of a large population of unique antibody proteins, and it is theorized that this diverse antibody repertoire may contain 10^{11} or more different protein sequences [5,6]. Recent studies determined that the circulating antibody repertoire contains at least 10^6 unique sequences, a number still far too large for comprehensive experimental structural studies [7,8].

Analysis of antibody structures determined by X-ray crystallography revealed conservation of structural features even in the regions of the antibody with the most sequence diversity, the six complementarity determining region (CDR) loops, which are responsible for antigen binding. Three of these loops are contributed by the heavy chain component of the fragment variable (Fv) domain of the antibody (HCDRs), and three are contributed by the light chain Fv domain (LCDRs). Two studies have identified robust rules that define canonical structures for five of the six CDR loops [9,10]. However, the HCDR3 defies classification attempts. The HCDR3 is encoded by the junction of three gene segments (V, D and J genes) connected by random nucleotide additions or deletions that are not encoded in the antibody germline gene segments, but rather introduced by the host enzyme terminal deoxynucleotidyl transferase during antibody gene recombination. The HCDR3 is therefore significantly more diverse in sequence length and composition than the other CDR loops, which are encoded by either a single gene segment (heavy and light chain CDRs 1 and 2) or by a simplified junction (LCDR3) [11–13]. As a result a large and diverse conformational space is observed for HCDR3s. Accordingly, HCDR3 is often especially important for antigen recognition and binding as has been revealed in previous structural studies [14].

The Rosetta software suite for macromolecular modeling can *de novo* predict the structure of a protein or portions thereof. The tertiary structure of a protein is determined from its primary sequence by pairing effective sampling techniques with knowledge-based energy functions. These energy functions for the most part assume that optimal geometries within proteins can be derived from a statistical analysis of the available structural information stored in the Protein Data Bank [15,16]. Similar approaches are used during comparative modeling, when structurally divergent regions (typically loops) of otherwise homologous proteins must be predicted [17]. Rosetta is capable of predicting antibody structures with low root mean square deviation (RMSD) to experimental structures outside the HCDR3; however accurately modeling the HCDR3 loop remains a challenge [18–21].

In an effort to classify canonical structures of the HCDR3 loop, prior work has subdivided it into two domains: the less diverse "torso" and the more variable "head" (Fig 1) [9,10]. Two major families of canonical torso structures have been identified, and are referred to as "bulged" and "non-bulged" torsos [10]. In this study, the geometries of the bulged torso domain have been used to develop restraints that restrict the sampling space of the HCDR3 torso and result in more native-like models when *de novo* modeling the entire HCDR3 loop.

Previous studies have used restraints to model the bulged HCDR3 torso, following rules previously described by Shirai et al. wherein a pseudodihedral angle restraint was calculated from the Cα atoms of residues T5, T6, T7 and the following initial residue of Framework 4 to define the bulged or non-bulged torso [18,21–24]. Weitzner et al. [18] utilized RosettaAntibody implemented within the Rosetta 3 framework to predict the structures of 11 previously unpublished antibody structures for the second antibody modeling assessment (AMA-II) [21]. The longest HCDR3 loop in AMA-II contained 16 residues, and was predicted by the RosettaAntibody team with an RMSD of 3.70Å to the native HCDR3 loop [18, 21]. Shirai et al. also competed in AMA-II, and used their torso restraint rules to filter results generated by a pipeline that includes both Spanner and OSCAR for loop structure prediction; in comparison to the RosettaAntibody team described above, their best model for the longest HCDR3 loop had an RMSD of 3.29Å to the native HCDR3 loop [21, 23].

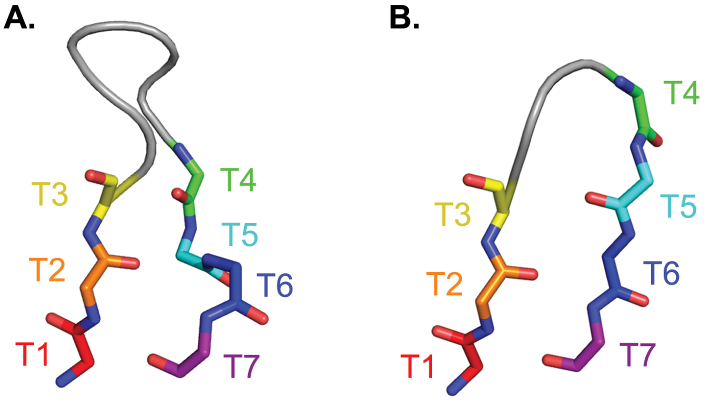


Fig 1. Defining the HCDR3 torso. The torso is defined as the first three and last four residues of the HCDR3 loop, numbered from T1 to T7. Main chain atoms are shown for bulged (panel A; PDBID 1UYW) and non-bulged (panel B; PDBID 2J88) torsos. In many (but not all) bulged torsos, a side-chain interaction between T2 and T6 causes the C-terminal side of the torso to bulge outward; the lack of such an interaction in non-bulged torsos leaves the beta-strand structure intact.

doi:10.1371/journal.pone.0154811.g001

In this study, a novel set of restraints was tested on 28 previously crystallized human antibodies with HCDR3 loops of increasing length and structural complexity. We expect that these restraints will improve modeling of antibodies for which no structural information is available, providing a means by which comprehensive structural studies of antibodies may be accomplished.

Results

Measuring bulged and non-bulged torso dihedral angles

An annotated list of antibodies was used to cull experimentally derived structures from the Protein Data Bank (PDB), expanding upon the list published by North et al. [10]. Following the IMGT conventions for defining the HCDR3, where the first HCDR3 residue occurs immediately following the V-gene residue Cys104 and the last HCDR3 residue occurs immediately preceding the J-gene residue Trp118, the torso is defined as the first three and the last four residues of the HCDR3 [10,25]. Accordingly, torso domain regions were pulled from these structures as two short peptide fragments (T1-T3 and T4-T7) and clustered using Rosetta at a threshold of 2 Å to separate bulged and non-bulged torsos. Previous studies identified a sequence motif (Arg or Lys at T2 and Asp at T6) that contributes to bulged torso formation in some but not all cases; these key residues were conserved in our bulged cluster, with 80% of bulged structures presenting Arg or Lys at T2, 73% presenting Asp at T6, and 65% retaining the complete T2/T6 sequence motif (S1 Fig) [9,10]. We found that germline-encoded regions of the antibody sequence often contribute these critical residues, as the end of the V gene segment contributes the first two to three torso residues while the J gene segment contributes the last four torso residues. The T2/T6 sequence motif that is often found in bulged torsos is present in 73% of naïve V and 92% of J germline gene allele segments (S1 Fig).

The φ and ψ angles of the seven torso residues of each antibody structure were measured, with key differences between bulged and non-bulged torsos identified in the ψ angles of residues T4 and T6 (Table 1). However, upon further study of previously defined torso clusters we

Table 1. Bulged and non-bulged dihedral angle measurements.

Torso Residue	Bulged		Non-bulged	
	φ	ψ	φ	ψ
T1	-145 ± 9	148 ± 12	-146 ± 12	145 ± 16
T2	-101 ± 22	142 ± 13	-109 ± 20	136 ± 26
T3	-107 ± 32	137 ± 33	-119 ± 44	138 ± 51
T4	-121 ± 49	161 ± 48	-82 ± 49	3 ± 59
T5	-95 ± 35	98 ± 26	-126 ± 43	136 ± 53
T6	-87 ± 18	-30 ± 26	-118 ± 34	129 ± 24
T7	-126 ± 14	134 ± 10	-125 ± 19	136 ± 11

The average and standard deviation of φ and ψ angles were calculated from existing human and mouse antibody crystal structures available in the PDB. Torso structures were clustered as bulged (n = 218) and non-bulged (n = 38) using a cluster radius of 2 Å.

doi:10.1371/journal.pone.0154811.t001

observed that the ψ angle of T4 is able to form two distinct conformations in both bulged and non-bulged torso clusters, and the T4 ψ angle does not distinguish between bulged and non-bulged torso clusters; the differences we observed when comparing all bulged antibodies to all non-bulged antibodies were due to the limited sample size of structures available in the PDB for these sub-conformations (S2 Fig) [10]. This is in contrast to for example T5, where a larger standard deviation is observed but still a statistically significant preference for a smaller ψ angle in a bulged torso exists. Average φ and ψ angles were calculated as follows:

$$atan2\left(\frac{\sum \sin\alpha}{n}, \frac{\sum \cos\alpha}{n}\right) \tag{1}$$

An approximate standard deviation was found using the following equations. For the vector v:

$$\vec{v} = \left(\frac{\sin\alpha}{n}, \frac{\cos\alpha}{n}\right) \tag{2}$$

Approximate standard deviation is calculated using:

$$\sqrt{2 \times [1 - \vec{v}]} \tag{3}$$

It is worth noting that straightforward average and standard deviation calculations are insufficient when handling circular values such as dihedral angles.

Derivation of restraints for bulged torso conformation

It has been observed that Rosetta rarely samples the bulged torso conformation when modeling HCDR3 loops [14]. Due to this limitation, coupled with the greater amount of experimentally derived structural data available for bulged torsos than non-bulged torsos and the fact that bulged torsos are more prevalent in the human antibody repertoire, we chose to focus on developing restraints to improve modeling of HCDR3 loops with bulged torsos. Rosetta uses a defined format to read in experimentally derived restraints. We used our measurements to generate dihedral angle restraints following a circular harmonic scoring function. Since the ψ angle measurement of T4 varies by 180 degrees between known bulged torso clusters, this measurement was omitted from our calculated restraints (S2 Fig).

Table 2. Experimentally derived antibodies used to benchmark bulged torso restraints.

PDB ID	HCDR3 Length	Resolution (Å)	Source
1WT5	11	2.10	*Humanized*
2G75	11	2.28	*Human*
4G5Z	11	1.83	*Human*
3QRG	12	1.70	*Human*
4G6K	12	1.90	*Humanized*
4LLU	12	2.16	*Human*
1FVC	13	2.20	*Humanized*
3HI5	13	2.50	*Human*
4HFW	13	2.60	*Human*
4FQH	14	2.05	*Human*
4NM4	14	2.65	*Human*
8FAB	14	1.80	*Human*
3G6A	15	2.10	*Human*
3TNM	15	1.85	*Human*
3W9D	15	2.32	*Human*
1AQK	16	1.84	*Human*
1DQL	16	2.60	*Human*
1OM3	16	2.20	*Human*
1U6A	17	2.81	*Human*
3AAZ	17	2.20	*Humanized*
4M5Y	17	1.55	*Human*
3INU	18	2.50	*Human*
3QEH	18	2.59	*Human*
4F58	18	2.49	*Human*
1HZH	20	2.70	*Human*
4LKC	22	2.20	*Human*
1RHH	24	1.90	*Human*
4FNL	26	2.30	*Human*

28 high-resolution antibody structures solved by X-ray crystallography were used to benchmark the bulged torso restraints. Each of these antibody structures was solved in the absence of antigen (*i.e.*, apo structures) and all residues in the HCDR3 loops were resolved.

doi:10.1371/journal.pone.0154811.t002

Modeling HCDR3 loops using bulged torso restraints

Following the protocol capture outlined in Supplemental Information, these restraints were used to model and score the HCDR3 loops from 28 benchmark antibodies whose structures had been previously determined by X-ray crystallography (Table 2). These 28 benchmark structures represent HCDR3 lengths from 11 to 26 residues, with a mean length of 16 residues, spanning a range regularly observed in human antibody repertoires that also have a mean HCDR3 length of 16 amino acids [26]. Each of the benchmark antibodies was crystallized in the absence of an antigen (*i.e.*, apo) in order to avoid attempts to model conformations achieved by induced fit with a binding partner.

Restraints function as a penalty during Rosetta's scoring protocol, *i.e.*, a positive energy value is added when a dihedral angle leaves the allowed range. In this case, models formed with native-like bulged torso dihedral angles would have no (or only a very small) penalty from the restraint term, whereas models that deviated from the bulged torso dihedral angles would be penalized with a positive energy score. When restraints were applied during modeling, we

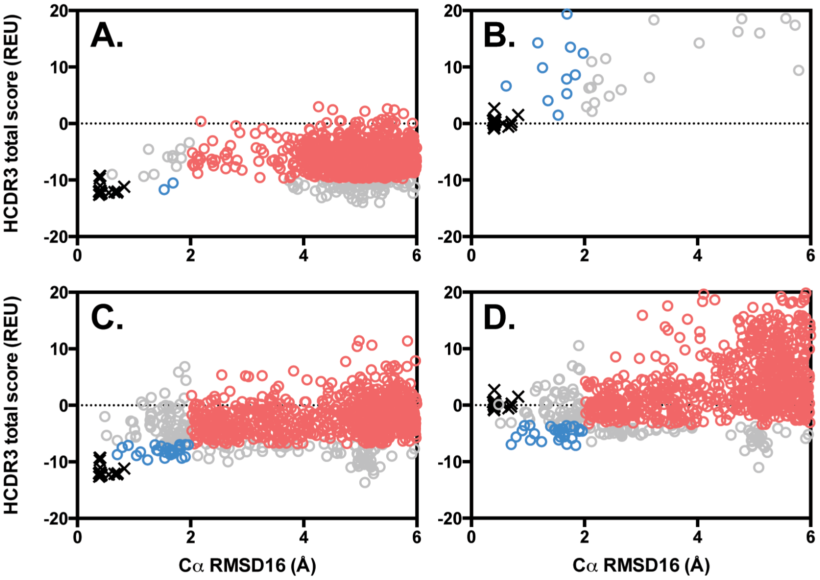


Fig 2. Bulged torso restraints improve native-like HCDR3 sampling and recovery. Using Rosetta LoopModel, 1,000 models of the benchmark antibody 4G5Z (circles) were generated with or without bulged restraints and these models were then scored with or without bulged restraints (panel A, modeled and scored without restraints; panel B, modeled without but scored with restraints; panel C, modeled with but scored without restraints; panel D, modeled and scored with restraints). The native crystal structure 4G5Z was also minimized using Rosetta FastRelax, generating 20 structures (black x's). The total HCDR3 score (in Rosetta Energy Units, or REU) is shown versus the Cα root mean square deviation of the HCDR3 loop, normalized to that of a protein loop containing 16 residues (RMSD16, in Å) to the native crystal structure. Models with scores ranked in the top 10% and RMSD16 ≤ 2 Å have been colored blue, while models with scores ranked below the top 10% and RMSD16 > 2 Å have been colored red. Improved native-like HCDR3 sampling is observed as a greater density of low RMSD16 models (blue circles) in comparison to Panel A, while improved model recovery is defined as a greater correlation between RMSD16 and score (colored vs. gray circles) in comparison to Panel A, as seen in panels C and D.

doi:10.1371/journal.pone.0154811.g002

observed a higher density of low-scoring, low-RMSD models (Fig 2C, blue circles, n = 26) than when modeling without restraints (Fig 2A, blue circles, n = 2). These low-scoring, low-RMSD models are defined as scoring in the top 10% of models, with Cα RMSD16 to the native structure of ≤ 2 Å (represented as blue circles, whereas models scoring below the top 10% of models with Cα RMSD16 > 2 Å are represented as red circles in Fig 2 and S2 Fig. When restraints were applied during scoring but not during modeling (Fig 2B) we found that the resulting models incur substantial restraint penalties due to non-native-like sampling of the torso domain, however the correlation between score and RMSD16 is improved. During application of this protocol wherein a native structure is unavailable, the ability to identify native-like models by score alone is extremely valuable. When applying restraints during both modeling and scoring, Rosetta generates a model population where an increased number of native-like structures correlate with low scores (Fig 2D, blue circles, n = 30; also see S1 File) as compared to experiments modeled and scored without restraints (Fig 2A, blue circles, n = 2; also see S1 File). Finally, we found that the application of these restraints results in more models whose backbone structures agree with bulged torso measurements defined in the literature (n = 719 with restraints, n = 33 without restraints; see S3 Fig) [14,22].

The results of modeling the 28 benchmark HCDR3 loops with or without bulged torso restraints can be found in Figs 3–5. We observed changes in both conformational sampling and in model discretion by score when restraints were applied. To analyze improvements in conformational sampling, models were ranked by RMSD16 to their native structure (Fig 3) and to study changes in scoring discretion the models were ranked by HCDR3 score (Fig 4). Finally, models were clustered using a package called Calibur and the best cluster by average HCDR3 score was analyzed (Fig 5).

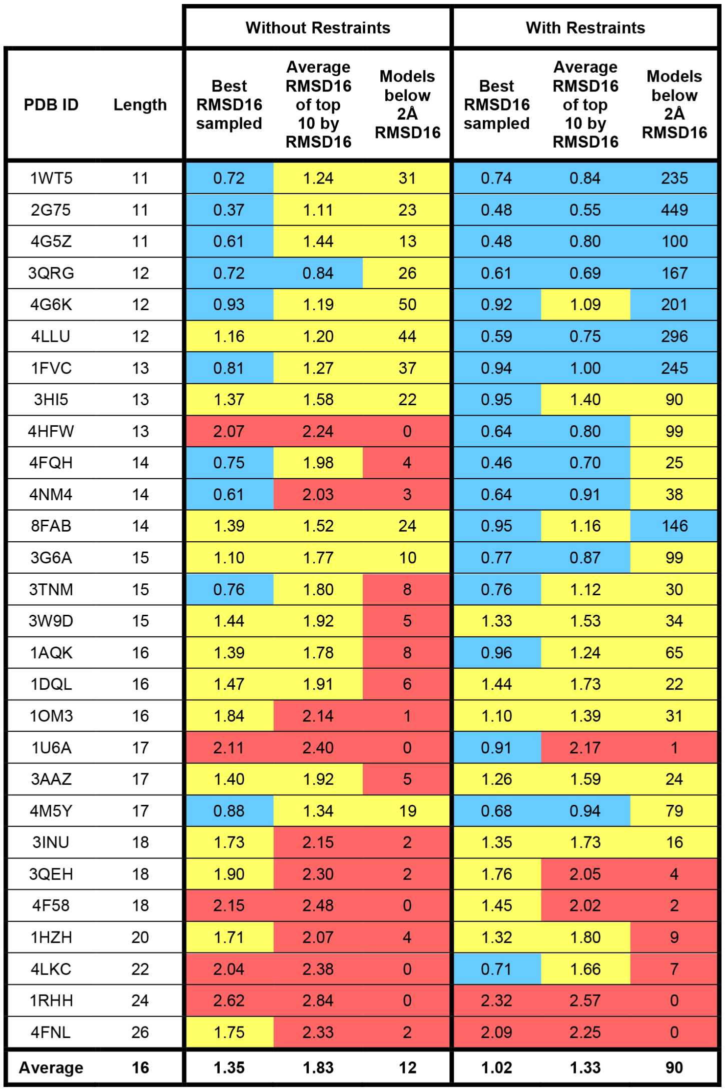
Bulged HCDR3 restraints improve native-like conformational sampling

Modeling with bulged torso restraints improved native-like conformational sampling (the number of models with RMSD16 below 2 Å) in 26 out of 28 benchmark cases (Fig 3); in the remaining case of benchmark antibody 1RHH with an HCDR3 loop of 24 residues, no models below 2 Å were observed when modeling with or without restraints, and in the case of 4FNL with an HCDR3 loop of 26 residues, 2 models below 2 Å were observed when modeling without restraints, compared to no models sampled below 2 A when modeling with restraints. On average, 90 models below 2 Å were generated with restraints, compared to only 12 models below 2 Å without restraints. The best RMSD sampled using bulged torso restraints was below 1 Å in 18 out of 28 cases with restraints, compared to 10 out of 28 cases without restraints. The average difference in the best RMSD sampled was 0.33 Å lower when restraints were applied during modeling. Furthermore, the average RMSD16 of the most native-like 10% of models (when ranked by RMSD16) is below 1 Å in 11 out of 28 cases when restraints are applied, compared to just 1 of 28 cases without restraints, revealing improved depth of high-resolution native-like sampling.

State-of-the-art computational methods to construct loop regions in proteins work reliably until about eight residues, and provide good results from some loops up to twelve residues [18–21]. Beyond this limit, the conformational space often becomes too large to be sampled exhaustively. Many HCDR3 loops are longer and specialized methods are needed to limit the conformational space. Our analyses describe better sampling of native-like structures during modeling of these diverse HCDR3 loops when our torso restraints are used, with qualitative changes in performance observed at 14 and 18 amino acids.

Bulged HCDR3 restraints improve scoring discretion

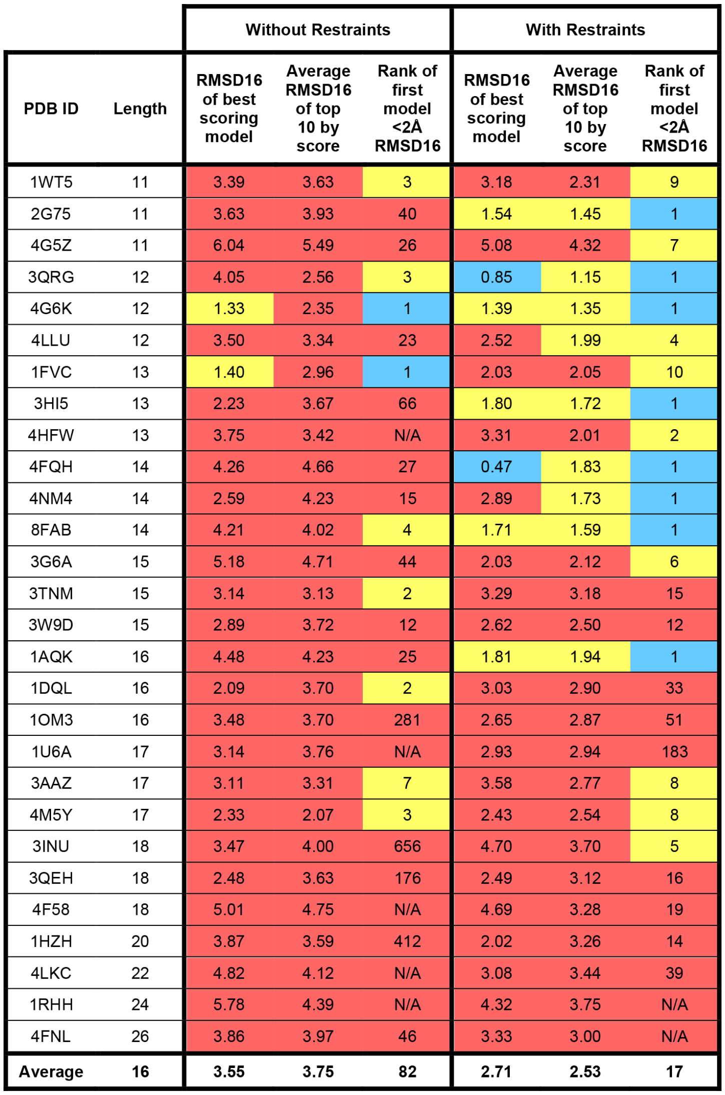
The ability to identify native-like HCDR3 loops by score when *de novo* modeling using Rosetta is of critical importance. Unfortunately, we found the predictive ability of Rosetta's scoring function in the absence of restraints to be lacking; when ranking models by HCDR3 score, only 2 of 28 benchmark cases resulted in a top-scoring model with RMSD16 $<$ 2 Å (Fig 4). However when restraints were applied, ranking models by score resulted in 7 of 28 cases with an RMSD16 below 2 Å and two of those with RMSD16 below 1 Å (antibody 3QRG, 12 amino acids long and 4FQH, 14 amino acids long). On average, the RMSD16 of the best scoring model improved by 0.84 Å when restraints were used during modeling and scoring. Because restraints improve sampling, there was also a marked improvement in the average RMSD16 of the top 10 models ranked by score; when restraints are applied, the average is below 2 Å in 9 out of 28 cases, but no results below 2 Å were found when restraints were not used. On average, there is an improvement of 1.22 Å in the average RMSD16 of the top 10 models ranked by score. The average rank of the first model below 2 Å is 17 when restraints are applied and in 8 of 28 cases the first-ranking model is below 2 Å, compared to only 2 out of 28 cases resulting in a first-ranking model below 2 Å and an average rank of 82 when restraints are not used. Altogether these analyses reveal that the bulged torso restraints improve scoring discretion of native-like structures, but that further improvement to the scoring of HCDR3 loops is needed [27].



		Without Restraints			With Restraints		
PDB ID	Length	Best RMSD16 sampled	Average RMSD16 of top 10 by RMSD16	Models below 2Å RMSD16	Best RMSD16 sampled	Average RMSD16 of top 10 by RMSD16	Models below 2Å RMSD16
1WT5	11	0.72	1.24	31	0.74	0.84	235
2G75	11	0.37	1.11	23	0.48	0.55	449
4G5Z	11	0.61	1.44	13	0.48	0.80	100
3QRG	12	0.72	0.84	26	0.61	0.69	167
4G6K	12	0.93	1.19	50	0.92	1.09	201
4LLU	12	1.16	1.20	44	0.59	0.75	296
1FVC	13	0.81	1.27	37	0.94	1.00	245
3HI5	13	1.37	1.58	22	0.95	1.40	90
4HFW	13	2.07	2.24	0	0.64	0.80	99
4FQH	14	0.75	1.98	4	0.46	0.70	25
4NM4	14	0.61	2.03	3	0.64	0.91	38
8FAB	14	1.39	1.52	24	0.95	1.16	146
3G6A	15	1.10	1.77	10	0.77	0.87	99
3TNM	15	0.76	1.80	8	0.76	1.12	30
3W9D	15	1.44	1.92	5	1.33	1.53	34
1AQK	16	1.39	1.78	8	0.96	1.24	65
1DQL	16	1.47	1.91	6	1.44	1.73	22
1OM3	16	1.84	2.14	1	1.10	1.39	31
1U6A	17	2.11	2.40	0	0.91	2.17	1
3AAZ	17	1.40	1.92	5	1.26	1.59	24
4M5Y	17	0.88	1.34	19	0.68	0.94	79
3INU	18	1.73	2.15	2	1.35	1.73	16
3QEH	18	1.90	2.30	2	1.76	2.05	4
4F58	18	2.15	2.48	0	1.45	2.02	2
1HZH	20	1.71	2.07	4	1.32	1.80	9
4LKC	22	2.04	2.38	0	0.71	1.66	7
1RHH	24	2.62	2.84	0	2.32	2.57	0
4FNL	26	1.75	2.33	2	2.09	2.25	0
Average	**16**	**1.35**	**1.83**	**12**	**1.02**	**1.33**	**90**

Fig 3. Torso restraints improve sampling of bulged HCDR3 loops. For each benchmark antibody structure, 1,000 models were generated with or without bulged torso restraints. The number of models below 2 Å RMSD16 to the native structure, the best RMSD16 sampled, and the average RMSD16 of the best 10 models ranked by RMSD16 are provided. For RMSD16-containing cells, blue shading represents RMSD16 ≤ 1 Å; yellow shading represents RMSD16 between 1 and 2 Å; red represents RMSD16 > 2 Å. For cells containing the number of models below 2 Å, blue shading represents ≥ 100 models; yellow shading represents ≥ 10 models; red shading represents fewer than 10 models.

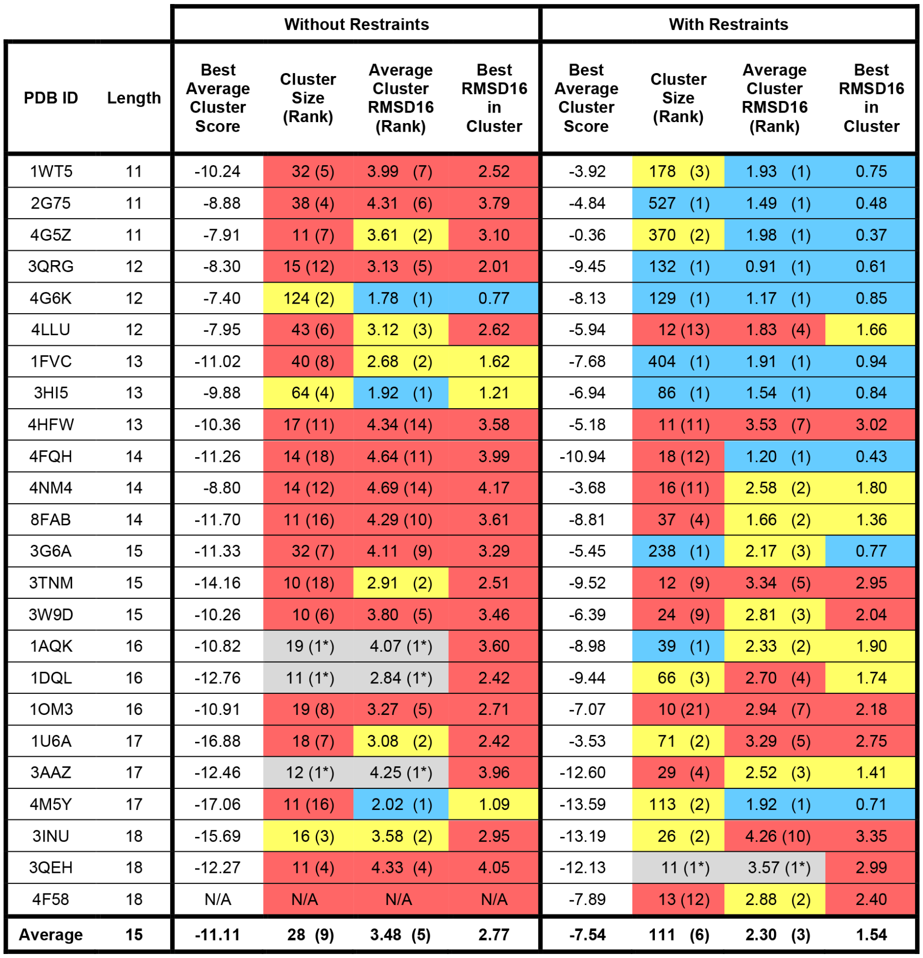
doi:10.1371/journal.pone.0154811.g003



		Without Restraints			With Restraints		
PDB ID	Length	RMSD16 of best scoring model	Average RMSD16 of top 10 by score	Rank of first model <2Å RMSD16	RMSD16 of best scoring model	Average RMSD16 of top 10 by score	Rank of first model <2Å RMSD16
1WT5	11	3.39	3.63	3	3.18	2.31	9
2G75	11	3.63	3.93	40	1.54	1.45	1
4G5Z	11	6.04	5.49	26	5.08	4.32	7
3QRG	12	4.05	2.56	3	0.85	1.15	1
4G6K	12	1.33	2.35	1	1.39	1.35	1
4LLU	12	3.50	3.34	23	2.52	1.99	4
1FVC	13	1.40	2.96	1	2.03	2.05	10
3HI5	13	2.23	3.67	66	1.80	1.72	1
4HFW	13	3.75	3.42	N/A	3.31	2.01	2
4FQH	14	4.26	4.66	27	0.47	1.83	1
4NM4	14	2.59	4.23	15	2.89	1.73	1
8FAB	14	4.21	4.02	4	1.71	1.59	1
3G6A	15	5.18	4.71	44	2.03	2.12	6
3TNM	15	3.14	3.13	2	3.29	3.18	15
3W9D	15	2.89	3.72	12	2.62	2.50	12
1AQK	16	4.48	4.23	25	1.81	1.94	1
1DQL	16	2.09	3.70	2	3.03	2.90	33
1OM3	16	3.48	3.70	281	2.65	2.87	51
1U6A	17	3.14	3.76	N/A	2.93	2.94	183
3AAZ	17	3.11	3.31	7	3.58	2.77	8
4M5Y	17	2.33	2.07	3	2.43	2.54	8
3INU	18	3.47	4.00	656	4.70	3.70	5
3QEH	18	2.48	3.63	176	2.49	3.12	16
4F58	18	5.01	4.75	N/A	4.69	3.28	19
1HZH	20	3.87	3.59	412	2.02	3.26	14
4LKC	22	4.82	4.12	N/A	3.08	3.44	39
1RHH	24	5.78	4.39	N/A	4.32	3.75	N/A
4FNL	26	3.86	3.97	46	3.33	3.00	N/A
Average	**16**	**3.55**	**3.75**	**82**	**2.71**	**2.53**	**17**

Fig 4. Torso restraints improve recovery of native-like bulged HCDR3 loops. For each benchmark antibody structure, 1,000 models were generated with or without bulged torso restraints. The number of models below 2 Å RMSD16 to the native structure, best RMSD16 sampled, average RMSD16 of the best 10 models ranked by RMSD16, RMSD16 of the best model ranked by Rosetta score, average RMSD16 of the top 10 models ranked by Rosetta score, and the rank of the first model below 2 Å when sorted by Rosetta score are provided. For RMSD16-containing cells, blue shading represents RMSD16 ≤ 1 Å; yellow shading represents RMSD16 between 1 and 2 Å; red represents RMSD16 > 2 Å. For rank-containing cells, blue shading represents rank 1; yellow shading represents ranks 2 to 10; red shading represents ranks > 10.

doi:10.1371/journal.pone.0154811.g004



		Without Restraints				With Restraints			
PDB ID	Length	Best Average Cluster Score	Cluster Size (Rank)	Average Cluster RMSD16 (Rank)	Best RMSD16 in Cluster	Best Average Cluster Score	Cluster Size (Rank)	Average Cluster RMSD16 (Rank)	Best RMSD16 in Cluster
1WT5	11	-10.24	32 (5)	3.99 (7)	2.52	-3.92	178 (3)	1.93 (1)	0.75
2G75	11	-8.88	38 (4)	4.31 (6)	3.79	-4.84	527 (1)	1.49 (1)	0.48
4G5Z	11	-7.91	11 (7)	3.61 (2)	3.10	-0.36	370 (2)	1.98 (1)	0.37
3QRG	12	-8.30	15 (12)	3.13 (5)	2.01	-9.45	132 (1)	0.91 (1)	0.61
4G6K	12	-7.40	124 (2)	1.78 (1)	0.77	-8.13	129 (1)	1.17 (1)	0.85
4LLU	12	-7.95	43 (6)	3.12 (3)	2.62	-5.94	12 (13)	1.83 (4)	1.66
1FVC	13	-11.02	40 (8)	2.68 (2)	1.62	-7.68	404 (1)	1.91 (1)	0.94
3HI5	13	-9.88	64 (4)	1.92 (1)	1.21	-6.94	86 (1)	1.54 (1)	0.84
4HFW	13	-10.36	17 (11)	4.34 (14)	3.58	-5.18	11 (11)	3.53 (7)	3.02
4FQH	14	-11.26	14 (18)	4.64 (11)	3.99	-10.94	18 (12)	1.20 (1)	0.43
4NM4	14	-8.80	14 (12)	4.69 (14)	4.17	-3.68	16 (11)	2.58 (2)	1.80
8FAB	14	-11.70	11 (16)	4.29 (10)	3.61	-8.81	37 (4)	1.66 (2)	1.36
3G6A	15	-11.33	32 (7)	4.11 (9)	3.29	-5.45	238 (1)	2.17 (3)	0.77
3TNM	15	-14.16	10 (18)	2.91 (2)	2.51	-9.52	12 (9)	3.34 (5)	2.95
3W9D	15	-10.26	10 (6)	3.80 (5)	3.46	-6.39	24 (9)	2.81 (3)	2.04
1AQK	16	-10.82	19 (1*)	4.07 (1*)	3.60	-8.98	39 (1)	2.33 (2)	1.90
1DQL	16	-12.76	11 (1*)	2.84 (1*)	2.42	-9.44	66 (3)	2.70 (4)	1.74
1OM3	16	-10.91	19 (8)	3.27 (5)	2.71	-7.07	10 (21)	2.94 (7)	2.18
1U6A	17	-16.88	18 (7)	3.08 (2)	2.42	-3.53	71 (2)	3.29 (5)	2.75
3AAZ	17	-12.46	12 (1*)	4.25 (1*)	3.96	-12.60	29 (4)	2.52 (3)	1.41
4M5Y	17	-17.06	11 (16)	2.02 (1)	1.09	-13.59	113 (2)	1.92 (1)	0.71
3INU	18	-15.69	16 (3)	3.58 (2)	2.95	-13.19	26 (2)	4.26 (10)	3.35
3QEH	18	-12.27	11 (4)	4.33 (4)	4.05	-12.13	11 (1*)	3.57 (1*)	2.99
4F58	18	N/A	N/A	N/A	N/A	-7.89	13 (12)	2.88 (2)	2.40
Average	**15**	**-11.11**	**28 (9)**	**3.48 (5)**	**2.77**	**-7.54**	**111 (6)**	**2.30 (3)**	**1.54**

Fig 5. Cluster analysis of bulged HCDR3 loop modeling. Calibur was used to cluster the 1,000 models generated with or without bulged torso restraints for each antibody, using a threshold of 2.0. Clusters containing less than 1% of the total models were omitted from analysis; models generated for benchmark antibodies 4F58, 1HZH, 4LKC, 1RHH and 4FNL did not produce any large clusters upon analysis (N/A). Average Rosetta score was calculated for each cluster, and the cluster with the lowest average score was selected as the "correct" cluster. The size of this correct cluster (and it's rank among cluster sizes), its average RMSD16 to the native structure (and rank among average RMSD16 measurements) are provided. Cells containing rank data are shaded blue if the value represents the top rank, yellow for ranks 2–3, and red for ranks >3; if only one cluster (1*) was found, the cell is shaded gray. For RMSD16-containing cells, blue shading represents RMSD16 ≤ 1 Å; yellow shading represents RMSD16 between 1 and 2 Å; red represents RMSD16 > 2 Å. Values were omitted from column averages if ≤1 cluster was found.

doi:10.1371/journal.pone.0154811.g005

Clustering bulged HCDR3 loop models

Using the clustering package Calibur [28], we analyzed the HCDR3 models generated with and without bulged restraints (Fig 5). Only clusters containing >1% of models (10 or more) were considered. For models made based on structures with 20 or more amino acids in the HCDR3 loop, no sufficiently large clusters were found. For the other benchmark structures, clusters were sorted by average cluster HCDR3 score, with the lowest average HCDR3 score being chosen as the "correct" cluster. This approach to selecting the "correct" conformation is common when *de novo* modeling HCDR3 loops, as the native structure of the loop is not known outside of benchmark studies. When restraints were used during modeling, the rank of the cluster size (how large a cluster is compared to other clusters) improved in 18 out of 24 cases over experiments where restraints were not used. When restraints were applied during modeling, the average RMSD16 of the correct cluster improved in 21 out of 24 cases. The average RMSD16 for the best cluster by score was top-ranking in 9 out of 24 cases when restraints were applied during modeling, compared to just 3 out of 24 cases when restraints were not used, which reveals the predictive power of our scoring metrics when restraints are applied.

Discussion

There is a growing body of work surrounding canonical structures of antibody CDR loops, first described by Chothia and colleagues and updated as recently at 2011 by the Dunbrack group [9,10]. These groups have shown that that five of the six CDR loops take on canonical structures, and that the remaining HCDR3 forms only a few canonical classes of structure in its torso domain. Our work builds upon this background, and has led to the development of knowledge-based structural restraints from available crystal structures of HCDR3 loops with bulged torsos. We have shown that these restraints can be used to restrict the sampling space Rosetta searches during *de novo* loop modeling, limiting the torso domain to the φ and ψ angles of these residues that have been experimentally observed. These torso restraints improve native-like structure sampling and score-based differentiation of native-like HCDR3 models. We have also shown that such structural restraints improve Rosetta's ability to model longer HCDR3 loops than previously possible, extending the range of the technique to cover more biologically relevant HCDR3 loop lengths.

While this study focuses on benchmarking new knowledge-based restraints against antibodies whose structures have been experimentally determined, the true value of these restraints is in their ability to improve *de novo* antibody modeling. Such antibody structural predictions are a more rapid approach than experimental structural techniques, and can improve our understanding of host-pathogen interactions, provide insight into mechanisms of viral infection, and may lead to new monoclonal antibody therapeutics or vaccine candidates. Combined with our prior understanding of canonical CDR loops, which had made it possible to homology model much of the functional surface of the antibody (the "paratope") using Rosetta, we can now predict the remaining HCDR3 which is critical in many antibody-antigen interactions. The central dogma of structural biology, that structure dictates function, lets us expect that improved accuracy in modeling HCDR3 will lead to improved accuracy in modeling antibody/antigen interactions which in turn leads to improved prediction of antibody function. We recognize that further experiments would be needed to prove this. Finally, upcoming advances in antibody sequencing, including the ability to sequence endogenously paired heavy and light chains, will provide the last critical insight in antibody modeling; we must now come to understand restrictions at the heavy and light chain interface that alter the paratope, and incorporate such restrictions into our structural predictions.

Although we have applied this approach to improving human antibody modeling, we recognize that this approach to structural restraint development is applicable to many other protein families in which structurally diverse surface loops with key functional importance are supported upon more structurally restricted framework regions [27]. Obvious examples include proteins with the PDZ domain and peptidase C1 domain protein families, which were found to use bulged HCDR3-like loops to recognize and bind their substrates [14]. Finally, we have shown that knowledge-based structural restraints can be calculated easily and applied to improve modeling of novel loops not previously solved by experimental techniques, provided enough experimentally derived structural data is available for framework regions of functional loops in other protein families, and that canonical classes of those regions can be defined.

Materials and Methods

Calculating bulged and non-bulged torso dihedral angles

A collection of antibody heavy chain variable domains was manually curated from the PDB, building upon a published list [10] (S2 File). The torso residues of these structures were extracted from the PDB files and were clustered using Rosetta Cluster with a cluster radius of 2 Å to separate bulged and non-bulged antibody torsos. φ and ψ dihedral angles of the seven torso residues were found using Biopython [29], with average and approximate standard deviation calculated using Eqs 1 and 2 (Table 1).

Generating HCDR3 loop models

The complete protocol for generating the HCDR3 loop models using Rosetta is described in S3 File and example file input and output is provided in S4 File. In brief, structure files for each benchmark antibody were downloaded from the PDB and were cleaned such that only a single variable domain remained. Input files for loop modeling were generated with the assistance of a suite of python scripts, and fragments were selected using the fragment picker. Centroid loop modeling was accomplished using cyclic coordinate descent (CCD), followed by a kinematic closure (KIC) full-atom refinement [30–32].

HCDR3 torso sequence analysis

Sequences of the seven torso residues were taken from each of the PDB files of the bulged antibody torso cluster found above and used to generate a WebLogo using the default webserver settings [33] (S1 Fig). A second WebLogo was generated using the sequences of the torso residues taken from the IMGT human V_H and J_H gene segments [34] (S1 Fig).

Supporting Information

S1 Fig. Bulged torso structures share similar sequences, which are germline-encoded. Previous studies identified a sequence motif in bulged torso structures, which are formed primarily via a side-chain interaction between either Arg or Lys (R/K) at T2 and Asp (D) at T6. A consensus sequence from bulged torsos culled from the PDB shows the prevalence of these residues at these positions (panel A). These residues are germline-encoded, as observed in a consensus sequence of the V_H and J_H gene segments that contribute to the torso domain (panel B).
(TIF)

S2 Fig. Average φ and ψ angles observed for each torso residue in known bulged and non-bulged clusters. North et al. [10] defined seven canonical torso conformations from experimentally-determined antibody structures. Two of these clusters are considered bulged (H3-anchor-1 and H3-anchor-3; blue) and two are considered non-bulged (H3-anchor-2 and H3-anchor-5; red). φ and ψ angles are well defined for both bulged and non-bulged HCDR3 torso residues. Bulged and non-bulged torsos are differentiated by their ψ angle at T6. The ψ

angle at T4 is bimodal for both bulged and non-bulged HCDR3 torsos, with ~180 degrees separating the two clusters within each definition.
(TIF)

S3 Fig. Bulged torso restraints improve sampling of HCDR3 torso angles. Using Rosetta LoopModel, 1,000 models of the benchmark antibody 4G5Z were generated without (red) or with (blue) bulged restraints. The τ_{101} angle and α_{101} dihedral angle defined by Weitzner et al. [14] were calculated for each model. Gray regions of the plot denote ± 3σ of the mean angles calculated for bulged HCDR3 torsos by Weitzner et al. [14]. Improved recovery of bulged torsos was observed as a greater density of points in the center gray region when restraints were applied (n = 719), versus when no restraints were applied (n = 33).
(TIF)

S1 File. Bulged torso restraints improve native-like HCDR3 sampling and recovery. As in Fig 2, 1,000 models of each benchmark antibody were generated and scored with or without bulged restraints using Rosetta LoopModel (comparable to Fig 2A and 2D). Models with scores ranked in the top 10% and RMSD16 $\leq$ 2 Å have been colored blue, while models with scores ranked below the top 10% and RMSD16 $>$ 2 Å have been colored red. The native crystal structure was also minimized using Rosetta FastRelax, generating 20 structures (black x's). The total HCDR3 score vs. the HCDR3 Cα RMSD16 to the native crystal structure is shown.
(PDF)

S2 File. HCDR3 definitions file. This file contains two comma separated value tables. The first table represents the non-bulged antibody structures used to calculate dihedral angle values, and lists the PDB file, chain ID, HCDR3 start residue and HCDR3 end residue when each chain in the PDB file has been renumbered sequentially starting from 1. The second file lists these values for the bulged antibody structures used to calculate the dihedral angle values.
(TXT)

S3 File. Rosetta protocol. A complete protocol has been provided, including Rosetta version number, for individuals who wish to utilize our methodology.
(PDF)

S4 File. Rosetta protocol capture. This archive contains example input and output files needed to run the Rosetta protocol described in S3 File.
(ZIP)

Author Contributions

Conceived and designed the experiments: JAF JRW JEC JM. Performed the experiments: JAF JRW AC. Analyzed the data: JAF JKL JRW AC JEC JM. Contributed reagents/materials/analysis tools: JAF JKL AC. Wrote the paper: JAF JEC JM. Independently verified the protocol: JKL.

References

1. Hashiguchi T, Fusco ML, Bornholdt ZA, Lee JE, Flyak AI, Matsuoka R, et al. Structural basis for Marburg virus neutralization by a cross-reactive human antibody. Cell. 2015; 160: 904–912. doi: 10.1016/j.cell.2015.01.041 PMID: 25723165
2. Hong M, Lee PS, Hoffman RMB, Zhu X, Krause JC, Laursen NS, et al. Antibody recognition of the pandemic H1N1 Influenza virus hemagglutinin receptor binding site. J Virol. American Society for Microbiology; 2013; 87: 12471–12480. doi: 10.1128/JVI.01388-13 PMID: 24027321
3. Whittle JRR, Zhang R, Khurana S, King LR, Manischewitz J, Golding H, et al. Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. Proc Natl

Acad Sci USA. National Acad Sciences; 2011; 108: 14216–14221. doi: 10.1073/pnas.1111497108 PMID: 21825125

4. Li Y, O'Dell S, Walker LM, Wu X, Guenaga J, Feng Y, et al. Mechanism of neutralization by the broadly neutralizing HIV-1 monoclonal antibody VRC01. J Virol. American Society for Microbiology; 2011; 85: 8954–8967. doi: 10.1128/JVI.00754-11 PMID: 21715490
5. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci USA. National Acad Sciences; 2009; 106: 20216–20221. doi: 10.1073/pnas.0909775106 PMID: 19875695
6. Trepel F. Number and distribution of lymphocytes in man. A critical analysis. Klin Wochenschr. 1974; 52: 511–515. PMID: 4853392
7. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. Reindl M, editor. PLoS ONE. Public Library of Science; 2011; 6: e22365. doi: 10.1371/journal.pone.0022365 PMID: 21829618
8. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med. NIH Public Access; 2009; 1: 12ra23. PMID: 20161664
9. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. J Mol Biol. 1998; 275: 269–294. doi: 10.1006/jmbi.1997.1442 PMID: 9466909
10. North B, Lehmann A, Dunbrack RL. A new clustering of antibody CDR loop conformations. J Mol Biol. 2011; 406: 228–256. doi: 10.1016/j.jmb.2010.10.030 PMID: 21035459
11. Finn JA, Crowe JE. Impact of new sequencing technologies on s^tudies of the human B cell repertoire. Curr Opin Immunol. 2013; 25: 613–618. doi: 10.1016/j.coi.2013.09.010 PMID: 24161653
12. Fanning LJ, Connor AM, Wu GE. Development of the immunoglobulin repertoire. Clin Immunol Immunopathol. 1996; 79: 1–14. PMID: 8612345
13. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983; 302: 575–581. PMID: 6300689
14. Weitzner BD, Dunbrack RL, Gray JJ. The origin of CDR H3 structural diversity. Structure. 2015; 23: 302–311. doi: 10.1016/j.str.2014.11.010 PMID: 25579815
15. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, Meiler J. Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry. American Chemical Society; 2010; 49: 2987–2998. doi: 10.1021/bi902153g PMID: 20235548
16. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 1997; 268: 209–225. doi: 10.1006/jmbi.1997.0959 PMID: 9149153
17. Rohl CA, Strauss CEM, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins. Wiley Subscription Services, Inc., A Wiley Company; 2004; 55: 656–677. doi: 10.1002/prot.10629 PMID: 15103629
18. Weitzner BD, Kuroda D, Marze N, Xu J, Gray JJ. Blind prediction performance of RosettaAntibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization. Proteins. 2014; 82: 1611–1623. doi: 10.1002/prot.24534 PMID: 24519881
19. Sircar A, Kim ET, Gray JJ. RosettaAntibody: antibody variable region homology modeling server. Nucleic Acids Res. Oxford University Press; 2009; 37: W474–9. doi: 10.1093/nar/gkp387 PMID: 19458157
20. Almagro JC, Beavers MP, Hernandez-Guzman F, Maier J, Shaulsky J, Butenhof K, et al. Antibody modeling assessment. Proteins. Wiley Subscription Services, Inc., A Wiley Company; 2011; 79: 3050–3066. doi: 10.1002/prot.23130 PMID: 21935986
21. Almagro JC, Teplyakov A, Luo J, Sweet RW, Kodangattil S, Hernandez-Guzman F, et al. Second antibody modeling assessment (AMA-II). Proteins. 2014; 82: 1553–1562. doi: 10.1002/prot.24567 PMID: 24668560
22. Shirai H, Kidera A, Nakamura H. Structural classification of CDR-H3 in antibodies. FEBS Lett. 1996; 399: 1–8. PMID: 8980108
23. Shirai H, Ikeda K, Yamashita K, Tsuchiya Y, Sarmiento J, Liang S, et al. High-resolution modeling of antibody structures by a combination of bioinformatics, expert knowledge, and molecular simulations. Proteins. 2014; 82: 1624–1635. doi: 10.1002/prot.24591 PMID: 24756852
24. Whitelegg NR, Rees AR. WAM: an improved algorithm for modelling antibodies on the WEB. Protein Eng. 2000; 13: 819–824. PMID: 11239080

25. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol. 2003; 27: 55–77. PMID: 12477501

26. Briney BS, Willis JR, Hicar MD, Thomas JW, Crowe JE. Frequency and genetic characterization of V (DD)J recombinants in the human peripheral blood antibody repertoire. Immunology. 2012; 137: 56–64. doi: 10.1111/j.1365-2567.2012.03605.x PMID: 22612413

27. Das R, Baker D. Macromolecular modeling with rosetta. Annu Rev Biochem. Annual Reviews; 2008; 77: 363–382. doi: 10.1146/annurev.biochem.77.062906.171838 PMID: 18410248

28. Li SC, Ng YK. Calibur: a tool for clustering large numbers of protein decoys. BMC Bioinformatics. BioMed Central Ltd; 2010; 11: 25. doi: 10.1186/1471-2105-11-25 PMID: 20070892

29. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. Oxford University Press; 2009; 25: 1422–1423. doi: 10.1093/bioinformatics/btp163 PMID: 19304878

30. Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. J Mol Biol. 2007; 373: 503–519. doi: 10.1016/j.jmb.2007.07.050 PMID: 17825317

31. Canutescu AA, Dunbrack RL. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci. Cold Spring Harbor Laboratory Press; 2003; 12: 963–972. doi: 10.1110/ps.0242703 PMID: 12717019

32. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods. Nature Publishing Group; 2009; 6: 551–552. doi: 10.1038/nmeth0809-551 PMID: 19644455

33. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. Cold Spring Harbor Lab; 2004; 14: 1188–1190. doi: 10.1101/gr.849004 PMID: 15173120

34. Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. Oxford University Press; 2005; 33: D256–61. doi: 10.1093/nar/gki010 PMID: 15608191