THE DEVELOPMENT OF COMPUTATIONAL METHODS
FOR DESIGNING ANTIBODIES AND OTHER PROTEINS

A Dissertation in
Chemical Engineering
by
Robert J. Pantazes

© 2014 Robert J. Pantazes

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2014
The dissertation of Robert J. Pantazes was reviewed and approved* by the following:

Costas D. Maranas  
Donald B. Broughton Professor of Chemical Engineering  
Dissertation Adviser  
Chair of Committee

Andrew Zydney  
Walter L. Robb Chair and Professor of Chemical Engineering  
Head of the Department of Chemical Engineering

Themis Matsoukas  
Professor of Chemical Engineering

Howard Salis  
Assistant Professor of Chemical Engineering

Arthur Lesk  
Professor of Biochemistry and Molecular Biology

*Signatures are on file in the Graduate School.
Abstract

Proteins are polymers of amino acids that have essential and diverse roles in organisms, including: structure (e.g. actin), catalysis of chemical reactions (e.g. cytochrome p450), and signaling (e.g. insulin-like growth factor I). Given the wide range of functions that proteins fulfill in nature, there is much interest in utilizing them for human needs in many areas, such as biofuels production, materials, and medicine. However, nature rarely provides a protein that is perfect for a specific human application, necessitating the use of engineering methods to improve or create desired properties. Computations are an essential tool for the de novo design of proteins. This dissertation focuses on the use of antibodies as a model protein system to develop de novo protein design methods.

Due to their many useful experimental and medicinal applications, antibody structures and their natural mechanisms of generation have been extensively studied. They are an excellent system for learning de novo protein design principles, as their structures have many modular features and their functions are limited to binding, not catalysis. Antigen binding by antibodies is primarily driven by the complementarity determining regions (CDRs). Models of the possible backbone conformations of the CDRs (i.e. canonical structures) were generated. These models were used in the Optimal Complementarity Determining Regions (OptCDR) method to allow the de novo design of antibody CDRs to bind any specified antigen epitope. Next, a database of Modular Antibody Parts (MAPs) analogous to the human germline genes used to make antibodies was created and shown to be able to predict antibody structures with a high degree of accuracy. Analysis of calculations involving OptCDR, MAPs, and other work outside the scope of this dissertation suggested that the computational protein engineering methods currently in
use needed to be improved. This led to the development of the Iterative Protein Redesign & Optimization Suite of Programs for the (re)design of proteins. Numerous collaborations have been established to experimentally validate the computational predictions and the research is progressing towards the \textit{de novo} design of fully human antibodies.
# Table of Contents

List of Figures ........................................................................................................... vii  
List of Tables ............................................................................................................. viii 
List of Abbreviations ................................................................................................. ix  
Acknowledgements ................................................................................................... xi  

**Chapter 1. Introduction** ......................................................................................... 1  
Modeling and Predicting Protein Structure .................................................................. 3  
Designing Stabilized Proteins ....................................................................................... 5  
Engineering Proteins for Molecular Interactions ....................................................... 5  
Designing New Proteins ............................................................................................... 8  
Using Antibodies for De Novo Protein Design ......................................................... 9  
References .................................................................................................................... 14  

**Chapter 2. Optimal Complementarity Determining Regions** ................................. 29  
Canonical Structure Identification .............................................................................. 30  
Step 1: Selection of CDR Canonical Structures ....................................................... 31  
Step 2: Initialization of CDR Amino Acid Sequences ............................................. 35  
Step 3: Structure and Sequence Refinement with IPRO ......................................... 36  
Step 4: Library Generation ......................................................................................... 37  
Energy Function Testing ............................................................................................. 38  
Results ......................................................................................................................... 39  
Hepatitis C Capsid Peptide .......................................................................................... 39  
Fluorescein .................................................................................................................. 41  
VEGF ........................................................................................................................... 43  
Discussion .................................................................................................................... 45  
References .................................................................................................................... 57  

**Chapter 3. The Database of Modular Antibody Parts** ........................................... 61  
Construction and content ......................................................................................... 62  
Germline Gene Assignment ....................................................................................... 63  
V*, CDR3, and J* Structure Delimitation .................................................................. 65  
Database Generation and Use .................................................................................. 67  
Database Evaluation .................................................................................................. 70  
Broadly-Neutralizing Antibodies .............................................................................. 71  
Discussion .................................................................................................................... 72  
References .................................................................................................................... 87  

**Chapter 4. The IPRO Suite of Programs** ............................................................... 91  
Implementation ........................................................................................................... 93  
IPRO Iterations ........................................................................................................... 96  
Backbone Perturbation ............................................................................................... 97  
Optimal Rotamer Selection ....................................................................................... 97  
Target Molecule Redocking ..................................................................................... 100  
Iteration Evaluation .................................................................................................. 101  
Ensemble Structure Refinements ............................................................................. 101  
OptGraft Binding Pocket Selection MILP ................................................................. 103  
IPRO ............................................................................................................................ 105  
OptGraft ...................................................................................................................... 106
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>OptCDR</td>
<td>107</td>
</tr>
<tr>
<td>MAPs</td>
<td>108</td>
</tr>
<tr>
<td>OptZyme</td>
<td>109</td>
</tr>
<tr>
<td>Mutator</td>
<td>109</td>
</tr>
<tr>
<td>Discussion</td>
<td>110</td>
</tr>
<tr>
<td>References</td>
<td>120</td>
</tr>
<tr>
<td>Chapter 5. Summary and Future Work</td>
<td>123</td>
</tr>
<tr>
<td>Summary</td>
<td>124</td>
</tr>
<tr>
<td>OptCDR</td>
<td>124</td>
</tr>
<tr>
<td>MAPs</td>
<td>125</td>
</tr>
<tr>
<td>The IPRO Suite of Programs</td>
<td>126</td>
</tr>
<tr>
<td>OptMAVEn</td>
<td>127</td>
</tr>
<tr>
<td>MAPs Structure Selection MILP</td>
<td>128</td>
</tr>
<tr>
<td>Antigen Positioning</td>
<td>131</td>
</tr>
<tr>
<td>IPRO Affinity Maturation</td>
<td>132</td>
</tr>
<tr>
<td>Experimental Validation</td>
<td>133</td>
</tr>
<tr>
<td>Anti-FLAG (DYKD) Antibodies</td>
<td>133</td>
</tr>
<tr>
<td>Gammabodies</td>
<td>135</td>
</tr>
<tr>
<td>Anti-CD20 Antibodies</td>
<td>138</td>
</tr>
<tr>
<td>Discussion</td>
<td>139</td>
</tr>
<tr>
<td>References</td>
<td>157</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1.1: Requirements for Successful Computational Protein Engineering ...........12
Figure 1.2: Antibody Structure Overview ..............................................................13
Figure 2.1: The Four Clusters of H2 CDRs ..........................................................48
Figure 2.2: The Score Distributions of Native and Decoy Antibody – Antigen Complexes ..........................................................49
Figure 2.3: CDR – Hepatitis C Capsid Peptide Complexes ..................................50
Figure 2.4: CDR – Fluorescein Complexes .........................................................51
Figure 2.5: CDR – VEGF Complexes .................................................................52
Figure 3.1: MAPs Antibody Structure Prediction Workflow ..................................75
Figure 3.2: Details of Antibody Sequence and Structure ......................................76
Figure 3.3: Observed Amino Acid Change Numbers in Human Antibodies ...........78
Figure 3.4: Observed Amino Acid Change Frequencies in Human Antibodies .......79
Figure 3.5: Clustering of V, D, and J Region Structures ........................................80
Figure 3.6: The Experimental and MAPs Predicted Structures of Antibodies CH65 and 4E10 ..........................................................81
Figure 4.1: The Format of the IPRO Suite of Programs .........................................112
Figure 4.2: Starting an IPRO Suite of Programs Experiment .................................113
Figure 4.3: Example of IPRO Terminology .........................................................114
Figure 4.4: The STANDARDS.py File .................................................................115
Figure 4.5: The Contents of an IPRO Experiment’s Folder ....................................116
Figure 4.6: The Contents of an Experiment_Details.txt File .................................117
Figure 4.7: IPRO Iteration Steps .........................................................................118
Figure 5.1: Average Epitope Positions ...............................................................142
Figure 5.2: Libraries of OptCDR-Designed Anti-FLAG Antibodies ......................144
Figure 5.3: Preliminary Binding Data for Anti-FLAG Antibodies .........................145
Figure 5.4: Aβ Superimposed on a V-KAPPA Domain ......................................147
Figure 5.5: Hydrophobic Regions of the V-KAPPA Domain ..............................148
List of Tables

Table 2.1: Sequence-Based Rules for Identifying CDRs.................................53
Table 2.2: Computational and Experimental Binding Data for the Antibodies........54
Table 2.3: A Library of Hepatitis C Capsid Binding Antibody CDRs ..................55
Table 2.4: Two Libraries of Fluorescein-Binding Antibody CDRs .....................56
Table 3.1: Human Germline V, D, and J Genes and Alleles ...........................82
Table 3.2: Alternative Delimitation of Structure ...........................................83
Table 3.3: Number of Model Structures in the MAPs Database .......................84
Table 3.4: Amino Acid Changes in the Ten Best Predicted Prototypes for the Broadly-Neutralizing anti-Influenza Antibody CH65 ..........................85
Table 3.5 Predicted Changes in Interaction Energy for Amino Acid Changes in the Broadly-Neutralizing anti-HIV Antibody 4E10 .................................86
Table 4.1: The OptCDR Canonical Structures .............................................119
Table 5.1: Calculated Energies for MAPs - Antigen Complexes ......................149
Table 5.2: The EEh Anti-FLAG Library CDR Sequences ..................................152
Table 5.3: The EEf Anti-FLAG Library CDR Sequences ..................................154
Table 5.4: The M2h Anti-FLAG Library CDR Sequences ..................................155
List of Abbreviations

Aβ: amyloid-β

C-KAPPA: κ constant domain

CASP: Critical Assessment for protein Structure Prediction

CDR: Complementarity Determining Region

D: Diversity gene

DG: Design Group

DM: Design Molecule

DP: Design Position

FR: Framework region

Gammabody: Grafted Amyloid Motif Antibody

GAMS: General Algebraic Modeling System

IgG: γ immunoglobulin

IGF-1: Insulin-like growth factor-I

IGF-2: Insulin-like growth factor-II

IMGT®: the international ImMunoGeneTics information system

IPRO: Iterative Protein Redesign & Optimization

J: Joining gene

MAPs: Modular Antibody Parts

MILP: Mixed Integer Linear Programming

OptCDR: Optimal Complementarity Determining Regions

PDB: Protein Data Bank

PIGS: Prediction of ImmunoGlobulin Structure
RMSD: root-mean-square deviation

Rotamer: rotational isomer

scFv: single chain variable fragment

TM: Target Molecule

U: a setting in the IPRO Suite of Programs that the user may specify

V: Variable gene

V-KAPPA: κ light variable domain

V-LAMBDA: λ light variable domain

VDW: Van der Waals

VEGF: Vascular endothelial growth factor

VH: Heavy variable domain

VL: Light variable domain

WAM: Web Antibody Modeling
Acknowledgements

I would like to begin by thanking my adviser, Costas Maranas, who has been incredibly helpful to me throughout my time at Penn State. His wisdom, guidance, and patience as I worked through matters both professional and personal have been invaluable. This dissertation would not be possible without him and I am very grateful. I would also like to thank my other dissertation committee members for their insights and recommendations throughout my graduate school experience. In addition, I would like to thank all of the Maranas Lab members past and present who have provided support, feedback, advise, recommendations, and laughs in the last several years.

I would like to thank my family and friends for their love and support, as well as their patience every time I said I was “almost” done with graduate school when there was still a long way to go. I would especially like to thank my parents, Lynn and Jeff, who provided me with a wonderful and loving home to grow up in and always encouraged the love of science and engineering that made this dissertation possible. Most importantly, I want to thank my wife, Kelley, who is my life and by far the best person I know. I can’t wait to keep spending the rest of my life with her.
Chapter 1. Introduction
Proteins are polymers of amino acids that have essential and diverse roles in organisms, including: structure (e.g. actin), catalysis of chemical reactions (e.g. cytochrome p450), and signaling (e.g. insulin-like growth factor I). Since they have such broad, naturally occurring utility, there is much interest in utilizing proteins to meet many of humanity’s needs in areas such as biofuel production [1, 2] and medicine [3]. However, it is highly unlikely that a naturally occurring protein optimally suited for a particular human objective is known. The sparsity of optimal, natural proteins has led to significant research in protein engineering methods [4]. Although purely experimental design efforts relying on combinatorial library construction and screening have been widely successful, the lessons learned do not easily generalize to inform the redesign of other systems. Additionally, they can be time-consuming and expensive and it is very unlikely that purely experimental techniques will ever be able to \textit{de novo} design a novel protein well suited for a particular task. In contrast, lessons learned in computationally designing one protein are generally applicable to other systems and computational protein engineering methods have the potential to allow for rational, \textit{de novo} design of new proteins.

Proteins have been previously computationally designed to bind new ligands [5], proteins [6], and nucleic acids [7], to improve protein stability [8, 9], as well as to introduce novel enzymatic activity [10, 11], demonstrating that the fundamental rudiments of molecular recognition and interactions can be adequately captured via computational design. Despite these successes, predictably changing or even improving a protein’s function in response to a performance target remains a formidable challenge. Successful \textit{de novo} computational protein design requires accurate \textit{structure} prediction, protein \textit{stability} at the desired operating conditions, and correct modeling of the protein’s \textit{interactions} with
other molecules (e.g. substrates, ligands, cofactors, etc.). Figure 1.1 illustrates how these three key areas can be brought together to facilitate de novo protein design.

**Modeling and Predicting Protein Structure**

Reliable protein structure prediction is paramount in protein design, as protein geometry and flexibility along with proper presentation of charges and molecular groups on the surface determine function (or lack-thereof). The central dogma behind protein structure prediction is that the native structure reaches a conformation that achieves (near) global minimum energy. The bi-annual Critical Assessment for protein Structure Prediction (CASP) benchmarks the current state of the art in protein structure prediction, with CASP9 completed in the summer of 2010. Using a feature space representation Kim *et al.* [12] sought to understand why identification of the native state is so challenging and discussed how the magnitude of the sampling problem dictates whether the problem can be solved with extra computational resources or if improved algorithms must be developed beforehand.

The development of improved protein structure prediction algorithms has been the focus of a number of publications. McAllister and Floudas [13] developed improved bounding methods for the structure search problem. In contrast to trimming the search space, Hahn and coworkers [14] sought to search more rapidly using a cluster expansion technique, albeit at the cost of introducing a controllable error. A popular concept that reduces the conformational search space is the use of rotamers (short for rotational isomers) of the statistically preferred conformations of amino acid side chains dependent upon the protein backbone geometry. Berkholz *et al.* [15] discussed how the backbone geometry varies as a function of the backbone dihedral angles. Havranek and Baker [16] considered
how to identify acceptable backbone changes that would allow rotamers to assume optimized orientations. Krivov and coworkers [17] developed SCWRL4 to more accurately and quickly predict side-chain conformations in proteins, while Shandler et al. [18] used foldamers to explore generation of better rotamer libraries. Blum and coworkers [19] developed a new method for de novo protein structure prediction that combines conformational space annealing and genetic algorithms that achieved significant improved over a standard Rosetta implementation.

While it is customary in protein design to assume a single, well-defined backbone geometry, this does not always hold true. Xue et al. [20] developed the meta-analysis tool PONDR-FIT to develop predictions for disordered regions of proteins. An alternative to using a single backbone structure is to model an ensemble of low-energy protein structures. Allen and coworkers [21] developed a multi-state design algorithm for modeling protein properties (e.g., stability, activity, and solubility) that are dependent upon backbone conformation variability. McAllister and Floudas [22] combined the αBB deterministic global optimization approach with conformational space annealing to predict lower energy protein structures (i.e., unique and ensembles thereof) and compared results with other methods. Allen and Mayo [23] developed MSD-FASTER and Subramani et al. [24] created ICON to generate and screen ensembles of low energy protein structures.

There have been several publications in the last two years where authors have customized and deployed computational protein structure prediction systems to specific protein classes. Correia et al. [25] successfully designed protein scaffolds to present target epitopes recognized by antibodies. Luo and coworkers [26] used computations to model
the allosteric changes of eight single-point mutations of αIIbβ3 to the integrin headpiece and observed conformational changes propagating from the headpiece to the legs of the integrin. In more general applications, Rosetta has been used to predict the structures of oligomers with near atomic-level accuracy [27], which should be helpful in conjunction with NMR data to resolve structures and to model the allosteric changes of ligand-free proteins from their bound states [28].

Designing Stabilized Proteins

After an appropriate structure for a protein has been modeled, care must be taken to ensure that it will be stable at the desired pH and temperature. Although literature attention to this topic has waned, it remains a critical factor in protein engineering. Belien et al. [29] used the pKD software to improve the low-pH stability of the B. subtilis endo-β-1,4-xylanase by making mutations that affected the local pKₐ of key residues. Heinzelman and coworkers [30] used SCHEMA to recombine several parent cellulases to design a library of thermostabilized proteins. Tian and coworkers [31] used computations to identify glycine to proline mutations to thermostabilize proteins by exploiting the fact that glycine has the highest conformational entropy of any amino acid whereas proline has the lowest. Joo et al. [32] used a more general computational approach to identify thermo-unstable residues and correcting mutations. Finally, Gribenkon et al. [33] and Gao et al. [34] used computations to identify thermostabilizing mutations while imposing active site geometry criteria to safeguard the activity of the redesigned proteins.

Engineering Proteins for Molecular Interactions

Computational protein design for a given function relies on optimizing a complex choreography of interactions with other molecules. An important class of protein
interaction partners is other proteins. Tuncbag and coworkers [35] developed a computational method to identify “hot-spot” residues that are the most important in mediating protein-protein interactions. In a study aimed at redesigning the interactions between a high-affinity pair of proteins, (i.e., acetylcholinesterase and fasciculin), Sharabi et al. [36] found that changes in the interaction energy, rather than total energy, correlated well with the experimental changes in binding energy. Guntas and coworkers [37] used a joint computational and experimental approach to redesign the ubiquitin-ligase E6AP to act on the unnatural partner Ubc12 in an effort to demonstrate efficiency advantages computations can offer. Yosef et al. [38] used ORBIT to switch the specificity of Calmodulin between its two main-target interaction partners, demonstrating the plasticity of interactions in signaling networks.

On the opposite side of the size-scale for protein interaction partners are metal ions. Their small size allows for more computationally complex descriptions of molecular interactions. Hayik et al. [39] used a mixed QM/MM protocol to predict metal ion binding energies. The C. Maranas Lab contributed in this area by developing the OptGraft method [40] to identify the location in a target protein that can best accommodate a metal ion binding site along with beneficial mutations in the surrounding residues. Wang et al. [41] used a similar approach specific to zinc ions.

A class of protein interaction partners with increasing attention in the literature is nucleic acids. Ashworth and Baker [42] used computations to assess the degree of optimization in known protein-DNA interactions and identified the contribution of individual residues. Several groups used nucleic acid binding proteins as targets for specificity alterations: Liu et al. [43] increased the specificity of a nucleoside kinase for 3’-deoxythymidine,
Lopes and coworkers [44] modified asparaginyl-tRNA synthetase to favor binding of aspartyl-adenylate, and Murphy et al. [45] used loop remodeling to alter the specificity of a human guaninine deamylase for ammelide over guanine.

Many other studies aimed to engineer proteins for optimizing their interactions with a variety of target molecules. Yang and coworkers [46] used free-energy perturbation calculations on the free and transition-state butyrylcholinesterase to identify high-activity mutants for the hydrolysis of cocaine. Berrondo et al. [47] analyzed the structural and regulatory consequences of mutations in the N-terminus arm of AraC, which is a gene expression regulatory protein that promotes expression when bound to arabinose and suppresses it otherwise. Chaudhury and Gray [48] used computational docking techniques to identify residues in an HIV protease that were important for activity and found they were residues that tended to confer drug-resistance. Grigoryan et al. [49] were successful in designing orthogonal interaction partners for specific members of the B-ZIP family of proteins in spite of their high sequence and structural homology. Chica et al. [50] destabilized the fluorophore ground state and stabilized the excited state to design improved red-fluorescent proteins. The C. Maranas Lab created the Iterative Protein Redesign & Optimization (IPRO) method [51, 52] and used it to change the cofactor specificity of C. boidini xylose reductase from NADPH to NADH [53]. In addition, the OptZyme method was recently developed to improve enzyme catalytic properties [54].

Two recent papers build upon the pioneering de novo computational design of an enzyme that catalyzes the Kemp elimination reaction [55]. Khersonsky and coworkers [56] computationally generated and experimentally screened proposed beneficial mutations for this enzyme while Kiss et al. [57] used computations to rank-order and evaluate
active and inactive *in silico* designed enzymes for the Kemp elimination reaction, finding that molecular dynamics was most useful in explaining the experimental findings.

**Designing New Proteins**

By bringing to bear structure elucidation, stability safeguards, and molecular interaction descriptions, a number of efforts achieved *de novo* design of novel proteins. Masica *et al.* [58] who used computations to *de novo* design peptides that can influence calcite binding. Fry and coworkers [59] designed a heterotetrameric protein that can selectively bind a chromophore whereas Koder *et al.* [60] designed an O$_2$ binding protein with properties similar to natural globin proteins with the key improvement of being able to bind O$_2$ better than CO. Finally, Siegel *et al.* [61] computationally *de novo* designed an enzyme to catalyze the Diels-Alder reaction, for which no naturally occurring enzyme was known beforehand.

Although there have been successes, the dream of efficiently, predictably and reliably computationally designing improved proteins remains beyond reach. Baker [62] eloquently reviewed in detail many of the unresolved challenges facing computational enzyme design. Biological systems are significantly more complicated than the idealized abstractions imposed by the assumptions used in computational protein engineering. It is increasingly realized that proteins rarely have unique functions; instead they participate in multiple interactions and processes in ways that may confound the ability of computations to assess their fitness. New steps are needed in integrating computational design methods, experimental screening protocols and structure identification techniques to achieve new milestones in protein design.
Using Antibodies for De Novo Protein Design

One particularly intriguing target to address the challenges raised by Baker is antibodies. Antibodies are proteins in vertebrate immune systems that are able to bind a diverse set of molecules, ranging from proteins and peptides to haptens (i.e. small molecules), with high specificity and affinity. Figure 1.2 illustrates a γ immunoglobulin (IgG), which are composed of pairs of heavy and light chains folded into the well-known “Y” shape. Two identical heavy chains form the stem and each branch is constructed from a light chain and the end of one of the heavy chains. The entire stem and approximately half of each branch are collectively referred to as the constant region while the ends of each branch are the variable regions. Each variable region is composed of a highly conserved framework region (FR) and six complementarity determining regions (CDRs), also known as hypervariable regions. The six CDRs, three in the light variable domain (VL) (L1, L2, L3) and three in the heavy variable domain (VH) (H1, H2, H3), are known to be responsible for the majority of antibody binding interactions. Humanization [63] is a common experimental technique where the CDRs from a non-human antibody are attached to the framework of a human antibody, thereby retaining the binding properties of the non-human antibody while decreasing or eliminating its immunogenicity. It is well established that for all CDRs, except H3, there are discrete sets of conformations that their secondary structures assume, known as canonical structures [64, 65]. Notably, there is a subset of antibodies found in camelids that lack light chains and have shorter than normal heavy chains [66].

The human immune system creates IgG through V-(D)-J recombination, a process where random variable (V), diversity (D), and joining (J) germline genes are combined to create
an antibody variable domain [67]. Only VH have D genes, while VL domains are created from two separate gene loci, \( \kappa \) (V-KAPPA) and \( \lambda \) (V-LAMBDA). All gene junctions occur within the third CDR of the variable domains and nucleotide insertions and deletions may occur when the genes are recombined. It is the presence of the D gene and the junctions with the V and J genes that gives rise to the significant structural variability observed in the H3 CDR. Junctional diversity introduced during V-(D)-J recombination gives a theoretical limit of up to \( 2 \times 10^{12} \) unique antibodies that the human immune system may produce (a number that is not reached due to antibodies that are out-of-frame, not expressed, etc.) [68]. By combinatorially shuffling only a few hundreds of genes, the immune system can produce billions of unique antibodies. Once an initial antibody has been generated and binds a foreign molecule, it undergoes somatic hypermutation to increase its affinity with its antigen, further increasing the diversity of possible antibodies.

Ever since antibody-producing mouse B cells were first immortalized by fusion with cancer cells to create hybridomas [69], antibodies have been extensively used in a variety of contexts. Experimentally, their ability to recognize molecules with high specificity and affinity has been leveraged in popular assays such as ELISA [70] and ELISPOT [71]. Medicinally, they are a profitable drug class, with the five top-selling antibodies (Rituxan [72], Remicade [73], Herceptin [74], Humira [75], and Avastin [76]) having revenues in excess of $11 billion in 2006 [77].

Therefore, it is not surprising that over the years there have been extensive efforts geared towards designing antibodies and libraries thereof. A number of experimental techniques [78-85] have been developed and successfully applied to design antibodies to bind
desired antigens or to improve the binding characteristics of an existing antibody. For example, Chen et al. [79] used monovalent phage display to increase the affinity of an antibody against vascular endothelial growth factor (VEGF) by 100 fold. While libraries have been computationally designed for other protein targets [86-88], computational antibody design methods have so far focused on altering existing antibodies to improve their properties [89-92]. Two notable examples are the use of a variety of computational methods to predict beneficial mutations for an antibody – antigen complex [91] and the rational redesign of an anti-VLA1 antibody through modification of a single CDR [90]. Other computational efforts have focused on predicting antibody structures from only their amino acid sequences [93-95]. Immunoinformatics tools have been developed to identify the genes used to create antibodies from nucleotide sequences [96-101], amino acid sequences [101-106], and three-dimensional structures [103, 107].

Although existing experimental and computational antibody design methods have made important contributions, especially in developing life-saving medications, no efforts have yet been focused on the de novo computational design of antibodies. With their many modular and well understood structural features and functions that are limited to binding, not catalysis, antibodies make an excellent system for developing de novo protein design methods.
I. Structure Modeling and Prediction

Protein primary structure

Modeled protein tertiary structure

Figure 1.1: Requirements for Successful Computational Protein Engineering. *De novo* computational protein engineering requires successfully meeting three design challenges: I) proper modeling of protein structure, II) ensuring the protein is stable at the desired operating conditions, and III) obtaining the proper interactions with other molecules. Simultaneously achieving each of these design targets is required for IV) successful *de novo* computational protein engineering. The *C. boidini* xylose reductase in III and Diels-Alder enzyme in IV are highlights and are intended as examples, not all-inclusive portrayals, of these categories.
Figure 1.2: Antibody Structure Overview. This is an illustration of an IgG structure [PDB: 1HZH] [108]. The heavy chain constant domains and the light chain constant domains are shown in dark gray, the variable domains are shown in yellow, and the CDRs are shown in orange. Because the constant domains are conserved among all antibodies, design efforts only need to focus on the variable domains. The most important part of the variable domains for antibody function (i.e. antigen binding) are the CDRs.
References


29. Belien T, Joye IJ, Delcour JA, Courtin CM: **Computational design-based molecular engineering of the glycosyl hydrolase family 11 B. subtilis XynA endoxylanase improves its acid stability.** *Protein engineering, design & selection: PEDS* 2009, **22**(10):587-596.


45. Murphy PM, Bolduc JM, Gallaher JL, Stoddard BL, Baker D: Alteration of enzyme specificity by computational loop remodeling and design. *Proceedings*


59. Fry HC, Lehmann A, Saven JG, DeGrado WF, Therien MJ: Computational design and elaboration of a de novo heterotetrameric alpha-helical protein
that selectively binds an emissive abiological (porphinato)zinc chromophore.


96. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP: IMGT/HighV-QUEST: the IMGT(R) web portal for immunoglobulin (IG) or antibody and T cell
receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome research* 2012, 8(1):26.


Chapter 2. Optimal Complementarity Determining Regions
The CDRs from one antibody can be attached to the framework of another, often without significant losses in antigen binding affinity [1]. This suggests that the CDRs are the smallest portion of an antibody’s structure for the de novo design of binding. Therefore the Optimal Complementarity Determining Regions (OptCDR) method for the design of novel antibody CDRs to bind any specified antigenic epitope with high affinity and specificity was developed.

OptCDR is a four-step workflow to design libraries of CDRs to bind a specified antigen epitope. The first step is the selection of the combination of CDR canonical structures (only backbones) that are most likely to have favorable binding with the antigen. During the second step, the amino acid sequences of the selected structures are initialized one at a time using a rotamer library, energy functions, and a mixed-integer linear programming (MILP) optimization formulation. This is followed by several thousand iterations of a modified version of the previously developed Iterative Protein Redesign and Optimization (IPRO) [2, 3] procedure, which simultaneously refines the backbone structures and amino acid sequences of the CDRs. The fourth step of OptCDR is library generation by accumulation of the most-promising mutations for the CDRs.

**Canonical Structure Identification**

The first step of OptCDR requires the use of a library of canonical structures that spans the full range of structural diversity exhibited by CDRs in known antibodies. This library was constructed by performing a clustering analysis on the CDRs identified from 883 antibody structures downloaded from the Protein Data Bank (PDB) [4]. Sequence-based rules, given in Table 2.1, were employed to identify the CDRs in each antibody. The rules were based on those from the Web Antibody Modeling (WAM) [5] method and
then expanded to include amino acids with similar chemical properties. When these sequence-based rules were not sufficient (approximately 25% of cases), the framework of the antibody variable region was aligned to a consensus framework and structural analysis was used to identify the CDRs. Through utilization of sequence and structural analyses, at least one CDR was identified in 833 of the 883 structures and all six in 700 of them.

A clustering procedure inspired by the work of Martin and Thronton [6] was used to group similar CDRs into canonical clusters. At the end of the clustering, a canonical structure was selected from each cluster as the structure with the smallest deviations in the sines and cosines of its phi, psi, and omega dihedral angles from the averages of the cluster. The clustering was carried out such that all members of a cluster had a backbone atom (N, Cα, C) root-mean-square deviation (RMSD) of no more than 1.5 Å from the canonical structure. A number of modifications were made \textit{a posteriori} to resolve structural inconsistencies, such as improper attachment to the variable region framework or significant clashing (i.e. at least one pair of atoms less than the sum of the Van der Waals (VDW) radii apart) with other canonical structures. Clusters containing only one or two members for all CDRs except H3 were discarded. Ultimately, this process resulted in 13 L1, 1 L2, 9 L3, 8 H1, 4 H2, and 124 H3 CDR canonical structures. The four clusters of H2 CDRs are shown in Figure 2.1 as an example of how similar structures are grouped together and dissimilar structures are disaggregated.

\textbf{Step 1: Selection of CDR Canonical Structures}

Canonical structures for CDRs describe only their backbone conformation with no information about the side chains. Due to the absence of residue type information we
developed a scoring system to distinguish between structures that have the potential to exhibit favorable binding with the antigen using only backbone atom – antigen distance information. The hypothesis is that the antigen should be placed within reach of the antibody CDRs (i.e. less than eight angstroms) but avoid detrimental clashes (i.e. closer than the sum of two atoms VDW radii). Specifically, if the distance between a backbone atom in the structure and the closest atom in the antigen is less than the sum of their VDW radii, then the contact is penalized with a negative score. Alternatively, if the two atoms are more than eight angstroms apart then the contact is considered to be unlikely to contribute to binding and a zero score is assigned. Only if the distance is between the sum of the two atoms VDW radii and eight angstroms is a positive score assigned. A penalty score of -5 for steric clashes and a reward score of +1 for potential to contribute favorably to binding were used. The score for a given canonical structure is computed as the sum of the scores of the atoms in its backbone. It is worth noting that all steric clashes between the backbone atoms of the CDRs and the antigens are equally penalized. The reason for this uniform treatment of clashes is that if an antigen is clashing with the backbone atoms of a CDR then most likely there would be significant clashes with the side chain.

The problem of selecting the highest scoring combination of non-clashing canonical structures can be mathematically represented using a MILP representation. This requires the definition of the index set \( I = \{i \mid L1, L2, L3, H1, H2, H3\} \) denoting the six CDRs and sets \( C^i = \{c \mid 1, \ldots, C^i\} \) encoding the number of canonical structures for a given CDR (i.e., 13, 1, 9, 8, 4, and 124 respectively). Set \( I_{c^{\text{clash}}} \) contains all pairwise canonical structure combinations \((i_1,c_1)\) and \((i_2,c_2)\) that share at least one pair of atoms that are closer then the
sum of the two atoms VDW radii (i.e. steric clash). The importance of excluding sterically clashing canonical structure pairs is bolstered by the fact that out of 521 antibody structures with resolution no worse than 2.5 Å, there were only two such clashing pairs (PDB 1LGV, 1OCW) and both such clashes were small (less than 0.2 Å). Parameter $S_{i,c}$ encodes the score contribution of the $c^{th}$ structure of the $i^{th}$ CDR. Binary variable $y_{i,c}$ encodes which structure $c$ is selected for the $i^{th}$ CDR. Specifically, $y_{i,c}$ is equal to one if structure $c$ has been selected for CDR $i$ and zero otherwise. The resulting MILP optimization problem is posed as:

$$\text{Maximize } \sum_{c=1}^{6} \sum_{i=1}^{C} y_{i,c} S_{i,c}$$

Subject to:

$$\sum_{c=1}^{C} y_{i,c} = 1, \quad i \in I$$

$$y_{i1,c1} + y_{i2,c2} \leq 1, \quad (i1,c1,i2,c2) \in IC_{\text{clash}}$$

Equality constraint (2) ensures that exactly one structure is selected for each CDR while inequalities (3) preclude the simultaneous presence of two canonical structures that sterically clash. The optimization formulation described collectively by Equations (1-3) is solved to global optimality using CPLEX [7] accessed in the General Algebraic Modeling System (GAMS) [8]. The solution of the above MILP formulation yields the highest scoring combination of canonical structures for a given position of the antigen.

The scoring system was tested on 254 native antibody-antigen complexes with resolutions no worse than 2.5 Å in which all six CDRs had been identified and 25,400 decoy complexes. The decoy complexes were generated by randomly changing the CDR
canonical structures of the native complexes 100 times each. As shown in Figure 2.2, the
distribution of scores for the native complexes was notably better than those for the decoy
complexes. Using a cutoff score of 52, the scoring system achieves a sensitivity of 85%
and a specificity of 80%, where sensitivity is the percentage of native complexes above
the cutoff and specificity is the percentage of decoy complexes below the cutoff.
Notably, a slightly better separation between native and decoy complexes could be
achieved with a cutoff of ten angstroms, but it was not adopted because it created an
abnormal bias for longer canonical structures. The highest scoring combination of CDR
canonical structures was also identified for each one of the 254 antibody–antigen
complexes. On average, the native complexes scored within 23% of the highest scoring
combination of CDR canonical structures for their antigen positions and in ten cases the
specific CDRs of an antibody scored better than the best combination of canonical
structures. This is consistent with the expectation that native antibodies are generally very
good but rarely optimal at recognizing a particular antigen.
Proper position of the antigen in the antibody binding pocket is addressed by iteratively
solving the MILP for different antigen locations. For each type of antigen (hapten,
peptide, protein), average Cartesian coordinates and standard deviations for the center of
mass of the portion of the epitope were calculated. Several thousand antigen positions
and orientations are randomly generated with normally distributed departures for the
three translational and uniformly distributed departures for the three rotational degrees of
freedom. When it is desired to target a specific epitope, the rotational degrees of freedom
can be limited to ensure that the epitope is the only portion of the antigen interacting with
the CDRs. Afterwards, several antigen positions with the highest scoring combinations of
canonical structures are retained. It is well established that antibodies typically form differently shaped binding pockets depending on the type of antigen: flat surfaces for proteins, grooves for peptides, and cavities for haptens [9-11]. Interestingly, the first step of OptCDR recapitulates these types of binding pockets for the different classes of antigens without the inclusion of any specific restraints to bias the selection towards these types of pockets.

Step 2: Initialization of CDR Amino Acid Sequences

The second step of OptCDR is the initialization of the amino acid sequences of the selected canonical structures one at a time in an order that is indicative of their typical proximity to antigens (e.g. H3, L3, H2, H1, L1, L2) and can be specified by the user of OptCDR. This is done using energy functions (VDW and electrostatics from CHARMM [12] and Lazaridis-Karplus implicit solvation [13]), a rotamer library, and a previously developed MILP rotamer optimization formulation [3] that selects the lowest energy combination of rotamers. Three additional sets of constraints are implemented in the MILP to ensure that the amino acids selected are consistent with the chosen canonical structures.

The first two sets of constraints limit the total percentage usage of each amino acid and all charged amino acids in a given CDR to below one standard deviation greater than their average percentage usage in the downloaded database of natural antibodies. The third set of constraints limit the specific amino acids permitted at each position in a canonical structure. If a canonical structure was observed at least 30 times, then only the amino acids that were encountered before at a given position are permitted. If a canonical structure was observed fewer than 30 times, then only the amino acids of the same “type”
are permitted. The chemical types used are charged (D,E,K,R,H), aliphatic (G,L,I,A,V,M), aromatic (F,W,Y), and polar (S,T,C,N,Q). Proline is treated as an aliphatic residue, but no rotamers of it exist in the rotamer library so it is never a selected amino acid.

After the amino acid sequences of all six CDRs have been initialized, a rigid-body docking step is carried out to refine the position of the antigen in the antibody binding pocket. This is followed by sequential solution of the MILP optimization problem for each CDR in the same order as before and then another rigid-body docking procedure.

**Step 3: Structure and Sequence Refinement with IPRO**

In the third step of OptCDR, several thousand iterations of a modified version of IPRO [2, 3] are carried out. In each iteration, a CDR is randomly selected and the phi and psi dihedral angles of each residue in the CDR are perturbed in CHARMM [12]. Next, the amino acid sequence of the CDR is re-determined using the rotamer selection MILP. This is followed by an energy minimization in CHARMM, during which weak harmonic constraints (0.05 kcal/mol) are used on all CDRs except H3 to ensure that their canonical structures are not changed. During this energy minimization similar constraints may also be used on the antigen to prevent or allow antigen conformational rearrangements. The Metropolis criterion, as in simulated annealing, is used to determine whether or not to retain the results of the iteration. If they are retained, a rigid-body docking procedure is carried out followed by another CHARMM energy minimization. If the post-docking interaction energy is lower than the pre-docking energy, then the post-docking results are retained. Otherwise, the pre-docking results are kept and the retained structures serve as the starting points for the next iteration. Over the course of many iterations, IPRO
progressively identifies sets of backbone perturbations and amino acid sequence mutations that improve the interaction energy between the CDRs and the antigen. The results from applying IPRO to specific, experimentally examined systems are discussed later in this chapter. IPRO has previously been used to design several protein-substrate systems that were experimentally verified to have the desired properties [14, 15]. The concept of iterating between structure perturbation and sequence/rotamer redesign was pioneered by the D. Baker Lab at the University of Washington [16] in the design of Top7, a computationally designed protein with a novel fold. The backbone perturbation followed by optimal rotamer selection allows IPRO to simultaneously mutate every amino acid in a CDR. This can uncover multiple non-additive mutations that would be unlikely to be sampled using directed evolution experiments. The use of the Metropolis criterion in IPRO safeguards against being trapped in local energy minima, which allows for the identification of multiple solutions with very different sequences.

**Step 4: Library Generation**

In the final step of OptCDR, the rotamer selection MILP is again employed to identify multiple low-energy amino acid sequences for each CDR. Starting with the final structures from step 3, the MILP formulation is solved multiple times for each CDR to identify not only the best but also the second, third, etc. best solutions by accumulating integer cuts that exclude previously visited solutions. This identifies multiple amino acid sequences for each CDR and a library of targeted size can be constructed from the lowest energy combinations of CDRs. By using these four steps, OptCDR can expediently generate multiple novel and diverse libraries of antibody CDRs to bind any specified antigen.
Energy Function Testing

As a validation of the energy functions used in OptCDR two tests were carried out. First, we used IPRO to examine the 95 mutants to an anti-VLA1 antibody [PDB: 1MHP] that had been computationally predicted to improve affinity with a 12% experimental success rate [17]. Following the experimental verification step the authors a posteriori refined their methods so that 26% of their predicted favorable mutants were correct. Using the rotamer selection and energy minimization using IPRO we calculated the change in interaction energy compared with wild-type for all 95 mutants. We observed an overall accuracy of 78% at predicting the experimental effect of the mutations and 50% of our predicted favorable mutants were in agreement with the experiment. These results demonstrate the efficacy of the energy functions used within OptCDR to distinguish between mutants that improve/decrease affinity.

As a follow up, we tested the ability of OptCDR to pinpoint the correct amino acid sequence for a given CDR structure. We chose 38 high-resolution antibody-antigen complexes and attempted to find even lower energy amino acid sequences for their CDRs. For each antibody, we used IPRO’s energy minimization and energy calculation functions to refine the initial crystal structures and determine the wild-type CDR-antigen interaction energies. Subsequently, the rotamer selection MILP was used to generate the five lowest-energy, non-wild-type amino acid sequences for each CDR in each complex. The corresponding rotamers were patched into the CDRs’ backbones and IPRO’s energy minimization and energy calculation functions were used to determine the interaction energy of the predicted low-energy CDRs. For 67% of the CDRs, no sequence with a lower energy than the native sequence could be found. Only 24% of the non-native CDR
amino acid sequences had lower interaction energies than the native sequences, and the
differences between the improved non-native and native sequences were minor (less than
9%). These results indicate that OptCDR recognizes the wild-type sequences known to
lead to effective binding, for a given antigen and CDR combination, as the best or near-
best binders.

Results
Three systems were selected to test OptCDR’s efficacy: a peptide from the capsid of
hepatitis C [PDB: 1N64] [18], fluorescein [PDB: 1FLR] [19], and VEGF [PDB: 1CZ8] [20]. Several computational metrics are used to draw comparisons, including interaction
energy defined as the minimized energy of the antigen - CDRs complex minus the energy
of the CDRs and the energy of the antigen individually. It is approximated within
CHARMM [12] using the VDWs, electrostatics, bonds, angles, dihedral angles, improper
dihedral angles, and generalized Born with molecular volume integration implicit
solvation energy functions. Contacts are defined as the number of CDR atoms within
three angstroms of the antigen and polar contacts are determined using PyMOL [21]. All
computations were carried out on 3.0 GHz Intel Xeon processors with 4 GB of RAM.
Each complete antibody library design was generated on its own processor and all were
completed in less then 12 days of computations, with an average of about 9 days.

Hepatitis C Capsid Peptide
Hepatitis C is a virus that infects approximately 3.2 million people in the United States
(http://www.cdc.gov/hepatitis/C/cFAQ.htm#statistics). The CDRs of antibody 19D9D6
[PDB: 1N64] [18] that bind a peptide (residues 13-40) from the capsid of the hepatitis C
virus with a $K_d$ of $1.3 \pm 0.1$ nM are shown in Figure 2.3A. This system was selected as a
general test of OptCDR to generate promising design alternatives. We first examined the extent of improvements that can be achieved in the computationally accessible metrics of binding quality by only mutating the original antibody structure without altering the CDR canonical structures. The results obtained using IPRO show improvements in all three binding metrics (see Table 2.2). Interestingly, the mutations identified are confined to the CDRs with the fewest antigen contacts. This trend of predicting mutations in the periphery of the antibody binding pocket is consistent with a previous computational study that was experimentally validated [22]. In this case, since 19D9D6 is already a high-affinity antibody, the results indicate that the dominant interactions in the center of the binding pocket are already effective and binding could only be improved further through repacking of the edges of the antibody-antigen interface.

We next used OptCDR to design three sets of antibody CDRs to bind the peptide instead of relying on only adding point mutations to 19D9D6. We first assumed a conservative posture by imposing harmonic constraints that insured that the antigen conformation did not change significantly upon binding. The three generated designs exhibit highly diverse antigen locations/orientations, canonical structure selections, and amino acid sequences but all share the groove that is typically observed for peptide-binding antibodies. Significant improvements in all computational binding metrics are observed (see Table 2.2) over the case of using only mutations. Table 2.3 depicts the predicted lowest-energy amino acid sequences for the CDRs of the structure shown in Figure 2.3B. By combining the predicted mutations, a library of CDRs can be generated (maximum library size = $1.1 \times 10^{14}$). By prioritizing mutations based on their binding scores, libraries of any smaller size can be culled from the original.
Finally, we removed the harmonic constraints on the antigen, allowing for its conformation to radically change in response to the interaction energy minimization step, and used OptCDR to generate three additional sets of antibody CDRs to bind the peptide. As seen in Table 2.2, allowing conformational changes to the peptide in response to energy minimization led to additional improvements in interaction energy and polar contacts, though no effect on the number of contacts. It should be noted that the conformational changes were not forced or pre-specified, but came about as a response to interactions with the CDRs during the IPRO step (i.e., Step 3) of OptCDR. In this case, the conformational changes were all between 4.2 and 5.2 Å RMSD from the initial peptide conformation. These results demonstrate how antigen conformational changes upon antibody binding may be an important contributor in informing antibody-antigen interactions. OptCDR allows for a user-specified presence, absence, and modulation of the strength of any imposed harmonic constraints.

**Fluorescein**

We next turned our attention to the hapten fluorescein for a comparison with experimental results. Anti-fluorescein antibodies have been used as the test system by several experimental directed evolution efforts [23-25] with different display technologies to improve the binding of a scFv derived from the anti-fluorescein antibody 4-4-20. This is a particularly interesting system because 4-4-20’s affinity for fluorescein is near the affinity ceiling of the tertiary immune response (0.7 ± 0.3 nM) [23]. Boder et al. [23] used yeast surface display to identify a scFv with 14 mutations exhibiting a ~6500 fold improvement in $K_d$, while Fukuda et al. and Jeremutus et al. [24, 25] used
mRNA and ribosome display, respectively, to identify mutants with ~30 fold improvement in $K_d$.

First, we used IPRO to evaluate these experimentally identified mutants: the best mutants identified by Boder et al. (2000) and Fukuda et al. (2006) and the consensus mutants identified by Fukuda et al. (2006) and Jermutus et al. (2001). Since some of the mutations were in FRs, the entire variable domains were used to create the mutants, not just the CDRs. SwissParam (http://swissparam.ch/) was used to create the topology and parameter files needed in CHARMM for fluorescein. Once the mutant antibodies were modeled, we extracted their CDRs so the calculated interaction energies could be directly compared with OptCDR results. Table 2.2 shows the computational and experimental improvements over the wild-type scFv of the four mutants. Note that the rank order of the mutants in terms of improvement over wild-type using $K_d$ and interaction energy, as quantified in OptCDR, match. Furthermore, the H3 CDR in the Jermutus et al. (2001) mutant shows the highest RMSD from the wild-type structure, which matched the experimental observation of increased flexibility of this CDR due to the removal of a salt bridge.

Next, IPRO was used to computationally affinity mature the CDRs of antibody 4-4-20, leading to numerous mutations (40 total) in all CDRs except H2. The computational binding results are detailed in Table 2.2. While the improvement in interaction energy over the wild-type antibody is not quite as large as that of the best experimental mutant, it does surpass the calculated energies for all other mutants. We also used OptCDR to design two sets of CDRs to bind fluorescein. Their computational binding metrics are given in Table 2.2, amino acid sequences in Table 2.4, and structures in Figure 2.4.
Interestingly, the interaction energies of the OptCDR designs do not reach the same levels as those of the computationally and experimentally affinity matured versions of 4-4-20, but they do surpass the wild-type antibody. Both designs share a number of features that appear favorable to binding. First, in both cases fluorescein is positioned within a deep cavity between the L3 and H2 CDRs on one side and the H3 CDR on the other. Both designs have long H3 CDRs folded mostly over the top of the fluorescein molecules to trap them in place, and it is this position of the H3 CDRs that lead to the notable increase in contacts between the experimental and OptCDR designs (Table 2.2). For both designs, the edges of the cavity have polar residues with each fluorescein oxygen involved in at least one polar contact. Finally, the sides of both cavities are composed of aliphatic and aromatic residues stabilizing the core hydrophobic portion of fluorescein. We hypothesize that the H3 CDRs of the unbound designs are sufficiently flexible to allow fluorescein access to the binding pocket.

**VEGF**

Finally, OptCDR designs for binding VEGF are contrasted against an affinity-matured antecedent of the antibody medication bevacizumab [26] to examine OptCDR’s epitope targeting abilities. VEGF has been shown [27] to promote tumor proliferation and growth. A number of anti-VEGF antibody-based drugs, including bevacizumab [26], with low nanomolar affinity are available. The resolved structure [PDB: 1CZ8] of an affinity-matured bevacizumab antecedent shows VEGF situated within a pocket primarily formed by the heavy chain CDRs [20]. In all OptCDR results, a harmonic constraint was used to prevent the structure of VEGF from changing significantly during calculations.
IPRO-based computational affinity maturation of 1CZ8 led to only minimal improvements in the interaction energy, as well as the number of contacts and polar contacts, which is a testament to the thoroughness of the experimental affinity maturation that 1CZ8 has already undergone. When we used OptCDR to predict novel CDRs to bind the same epitope of VEGF targeted by 1CZ8, the best design, shown in Figure 2.5B, has binding metrics that are comparable with the existing antibody (slightly greater interaction energy and a few more contacts and polar contacts). Thus the range of binding metrics for OptCDR (see Table 2.2) only reaches the levels of 1CZ8, although the predicted structures are all notably different and most of which exhibit the planar binding-pocket expected for protein-binding antibodies.

Next, we explored another set of OptCDR designs by targeting an epitope on the opposite side of VEGF from the portion bound by 1CZ8 and our other designs. To the best of our knowledge, this portion of VEGF is not recognized by cellular VEGF receptors or any designed antibodies. One of the predicted designs is shown in Figure 2.5C and the computational binding metrics are given in Table 2.2. Even though the obtained designs are very different as they target a completely different epitope of VEGF, the computational binding metrics achieved are quite similar in value. The results obtained demonstrate the efficacy of OptCDR to generate designs that bind VEGF with equivalent computational binding characteristics as the wild-type antibody with novel CDRs or targeting a different epitope of the VEGF molecule, alluding to the built-in redundancy of molecular recognition.

We decided to further explore VEGF-binding designs by focusing on only three out of the six CDRs, as in nanobodies. Nanobodies are single domain proteins derived from the
variable domain of heavy chains from a special subset of antibodies in camelids that lack light chains and thus have only three CDRs instead of six. OptCDR was used to generate nanobody CDR designs by considering only the H1, H2, and H3 CDRs. Despite the reduction in the number of structural degrees of freedom, OptCDR identified designs based solely on the heavy chain CDRs (see Table 2.2) that had similar computational binding metrics to the six CDR designs. This surprising finding is consistent with the experimental observation that nanobodies can have binding affinities that are equivalent to antibodies despite their smaller size. We believe that OptCDR achieved this through the selection of longer than typical canonical structures, especially for the H3 CDR, as is typical in experimental nanobodies. In addition, the absence of the light chain allows the antigen to assume positions and orientations that are normally prohibited due to steric clashes. A representative design plotted in Figure 2.5D illustrates how the selection of longer structures for the H3 domain counteracts the loss of the light chain.

Discussion

While there are many experimental techniques to design and redesign antibodies to bind target antigens, to date computational methods have focused only on redesigning existing antibodies to have improved binding affinities. To the best of our knowledge, OptCDR is the first general computational method for the de novo design of antibodies to bind any specified epitope of an antigen. Its four-step procedure works by first selecting appropriate CDR canonical structures to bind the antigen, then filling in their amino acid sequences, followed by a simultaneous structural and sequence refinement, and finally a library generation step through accumulation of the best mutations to each CDR. By using these four steps, OptCDR can generate multiple novel and diverse libraries of
antibodies to bind any specified antigen. We recognize that the FRs of antibody variable domains can contribute to antibody affinity [1]. In this first effort, however, we chose to exclusively focus on the CDRs as they are the most important factors in antibody recognition. This fact is manifested by the largely conserved affinity upon humanization of antibodies by grafting the CDRs onto human antibody frameworks. Nevertheless, we believe that the OptCDR workflow is open to future extensions that will target the design of the entire antibody variable domains.

OptCDR can be thought of as a computational analogue to the human immune system (or directed evolution experiments). At the start of an infection, B cells produce initial antibodies that bind the pathogen. As the infection progresses, the generated antibodies undergo an evolutionary process where their pathogen-binding affinities are improved by sequential identification of favorable mutations. The first two steps of OptCDR, the identification of appropriate combinations of canonical structures and the filling in of their amino acid sequences, can be viewed as the identification of an initial antibody to bind the antigen. The third step of OptCDR is the sequential identification of perturbation/mutation combinations that lead to improved antigen binding. The use of a methodology similar to the immune system gives OptCDR the same flexibility as the immune system: the ability to generate antibodies to bind a wide range of possible epitopes of antigens. A strength and limitation of this paradigm is that it selects antibodies on the basis of having adequate but not necessarily optimal binding affinities. This leads to an initial pool of structurally diverse antibody designs whose affinity for the targeted antibody can be ratcheted up further through the accumulation of additional mutations. The diversity of possible antibody design space in response to an antigen
challenge suggests that OptCDR identified solutions are unlikely to converge to the amino acid choices of existing antibodies.

We have tested OptCDR on three antigens that span the range of antigen sizes bound by antibodies: the hapten fluorescein, a peptide from the capsid of hepatitis C, and the protein VEGF. All three antigens are bound by their wild-type antibodies with low nanomolar affinities, ($K_d = 0.7, 1.3, \text{ and } 0.11 \text{ nM, respectively}$), making them challenging test cases. Nonetheless, for all three cases OptCDR arrived at multiple novel sets of CDRs that form the appropriate types of antibody binding pockets for the given antigen, have interaction energies comparable to or better than the native antibodies, and share many features that appear to be conducive to high-affinity binding. By taking advantage of the highly conserved structural features of antibodies, OptCDR achieves the *de novo* design for a specific class of protein molecules.
Figure 2.1: The Four Clusters of H2 CDRs. The members of the clusters are shown in blue, and the canonical structure of each cluster is shown in orange. All members of each cluster have an RMSD of no more than 1.5 angstroms with the canonical structure of that cluster.
Figure 2.2: The Score Distributions of Native and Decoy Antibody-Antigen Complexes. 254 native and 25,400 decoy antibody-antigen complexes were examined using the scoring function and good separation between them was observed. Using a cutoff of 52, a sensitivity of 85% and a specificity of 80% were achieved, where sensitivity is the percentage of native complexes above the cutoff and specificity is the percentage of decoy complexes below it.
Figure 2.3: CDR – Hepatitis C Capsid Peptide Complexes. The capsid peptides are shown as red spheres, the CDRs are shown as orange ribbons, and CDR residues within 4 angstroms of the peptide are explicitly shown. All images are from the same perspective. A) is the natural antibody-peptide complex [PDB: 1N64], B) is an OptCDR design with no peptide conformational change, and C) is an OptCDR design with peptide conformational change.
Figure 2.4: CDR – Fluorescein Complexes. Fluorescein is shown as cyan spheres with oxygens and hydrogens shown in red and white, respectively. The CDRs are shown as orange ribbons and all CDR residues within four angstroms of fluorescein are explicitly shown. All the complexes are shown from the same perspective relative to the antibody binding pocket. A) is the structure of antibody 4-4-20 and B) and C) are the two OptCDR fluorescein-binding designs.
Figure 2.5: CDR – VEGF Complexes. VEGF is shown as green spheres, the CDRs are shown as orange ribbons, and CDR residues that are within four angstroms of VEGF are explicitly shown. All images are from the same perspective relative to the antibody binding pocket. A) is the structure of PDB 1CZ8. B) is the best OptCDR design generated to bind the same portion of VEGF as bevacizumab using all 6 CDRs while C) is the best design to bind an epitope on the opposite side of VEGF. D) is the best nanobody OptCDR design.
Table 2.1: Sequence-Based Rules for Identifying CDRs. XXX corresponds to any permitted amino acid. “/” separate amino acids that are acceptable at a given position while “-“ distinguish between different positions. For the end of the H2 CDR, rule (A) was applied first and if that failed then rule (B) was used.

<table>
<thead>
<tr>
<th>CDR Loop</th>
<th>Starting Position</th>
<th>Preceded by:</th>
<th>Length:</th>
<th>Followed by:</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>1 Residue after first CYS</td>
<td>CYS</td>
<td>10-17 Residues</td>
<td>TRP - TYR/LEU/PHE/VAL - GLN/LEU/PHE</td>
</tr>
<tr>
<td>L2</td>
<td>14-19 Residues After L1</td>
<td>ILE/VAL/MET - TYR/LYS/PHE/HIS/GLY</td>
<td>7 Residues</td>
<td>GLY/GLU</td>
</tr>
<tr>
<td>L3</td>
<td>1 Residue after first CYS after L2</td>
<td>CYS</td>
<td>7-13 Residues</td>
<td>PHE - GLY - XXX - GLY</td>
</tr>
<tr>
<td>H1</td>
<td>4 Residues after first CYS</td>
<td>CYS - XXX - XXX - XXX</td>
<td>10-12 Residues</td>
<td>TRP - ILE/VAL/MET/ALA/PHE - ARG/LYS - GLN/ASN/LYS/GLU/HIS</td>
</tr>
<tr>
<td>H2</td>
<td>Approximately 15 Residues after H1</td>
<td>TRP/TYR/LEU - ILE/LEU/VAL/MET - GLY/ALA/SER</td>
<td>9-15 Residues</td>
<td>A) TYR - XXX - XXX - XXX - LYS - GLN/LYS/ARG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B) TYR/PHE - XXX - XXX - XXX - XXX - LYS - GLN/LYS/ARG</td>
</tr>
<tr>
<td>H3</td>
<td>3 Residues after CYS Approximately 40 Residues after H2</td>
<td>CYS - XXX - XXX</td>
<td>3 - 22 Residues</td>
<td>TRP - GLY - XXX - GLY</td>
</tr>
</tbody>
</table>
Table 2.2: Computational and Experimental Binding Data for the Antibodies. The various computational binding metrics are defined in the text at the start of the Results Section and all experimental $K_d$ values were taken from the appropriate publications. VEGF – epitope 1 is the epitope bound by bevacizumab while epitope 2 is on the opposite side of VEGF. The nanobody designs have only three CDRs, all other antibodies have six.

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Antigen</th>
<th>Interaction Energy (kcal/mol)</th>
<th>Contacts</th>
<th>Polar Contacts</th>
<th>Experimental $K_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>19D9D6</td>
<td>Hepatitis C Capsid Peptide</td>
<td>-62.6</td>
<td>74</td>
<td>8</td>
<td>1.3 nM</td>
</tr>
<tr>
<td>IPRO Affinity Maturation of 19D9D6</td>
<td>Hepatitis C Capsid Peptide</td>
<td>-78.1 to -80.1</td>
<td>81 to 87</td>
<td>14 to 15</td>
<td>NA</td>
</tr>
<tr>
<td>OptCDR Designs</td>
<td>Hepatitis C Capsid Peptide</td>
<td>-88.2 to -104.8</td>
<td>88 to 115</td>
<td>18 to 23</td>
<td>NA</td>
</tr>
<tr>
<td>OptCDR Design with Antigen Rearrangement</td>
<td>Hepatitis C Capsid Peptide</td>
<td>-123.6 to -175.8</td>
<td>88 to 112</td>
<td>23 to 31</td>
<td>NA</td>
</tr>
<tr>
<td>4-4-20</td>
<td>Fluorescein</td>
<td>-49.5</td>
<td>22</td>
<td>4</td>
<td>0.7 nM</td>
</tr>
<tr>
<td>Boder <em>et al.</em> Best Design</td>
<td>Fluorescein</td>
<td>-78.7</td>
<td>23</td>
<td>4</td>
<td>48 fM</td>
</tr>
<tr>
<td>Fukuda <em>et al.</em> Consensus Design</td>
<td>Fluorescein</td>
<td>-67.8</td>
<td>22</td>
<td>5</td>
<td>~1 nM (wt 32 nM)</td>
</tr>
<tr>
<td>Fukuda <em>et al.</em> Best Design</td>
<td>Fluorescein</td>
<td>-70.4</td>
<td>24</td>
<td>3</td>
<td>0.88 nM (wt 32 nM)</td>
</tr>
<tr>
<td>Jermutus <em>et al.</em> Consensus Design</td>
<td>Fluorescein</td>
<td>-74.9</td>
<td>24</td>
<td>3</td>
<td>~37 pM</td>
</tr>
<tr>
<td>IPRO Affinity Maturation of 4-4-20</td>
<td>Fluorescein</td>
<td>-77.8</td>
<td>17</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>OptCDR Designs</td>
<td>Fluorescein</td>
<td>-55.2 to -57.3</td>
<td>71 to 79</td>
<td>8 to 9</td>
<td>NA</td>
</tr>
<tr>
<td>PDB 1CZ8</td>
<td>VEGF - epitope 1</td>
<td>-110.8</td>
<td>86</td>
<td>21</td>
<td>0.11 nM</td>
</tr>
<tr>
<td>IPRO Affinity Maturation of 1CZ8</td>
<td>VEGF - epitope 1</td>
<td>-111.0 to -116.0</td>
<td>89 to 99</td>
<td>21 to 22</td>
<td>NA</td>
</tr>
<tr>
<td>OptCDR Design</td>
<td>VEGF - epitope 1</td>
<td>-82.0 to -109.0</td>
<td>73 to 110</td>
<td>11 to 25</td>
<td>NA</td>
</tr>
<tr>
<td>OptCDR Design</td>
<td>VEGF - epitope 2</td>
<td>-88.6 to -98.4</td>
<td>57 to 85</td>
<td>15 to 20</td>
<td>NA</td>
</tr>
<tr>
<td>OptCDR Nanobodies</td>
<td>VEGF - epitope 1</td>
<td>-82.2 to -92.6</td>
<td>79 to 86</td>
<td>20 to 22</td>
<td>NA</td>
</tr>
</tbody>
</table>
Table 2.3: A Library of Hepatitis C Capsid Binding Antibody CDRs. The CDR sequences and predicted mutations to them to form a library of up to $1.1*10^{14}$ antibodies that can all bind the peptide from the capsid of hepatitis C.

<table>
<thead>
<tr>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>QGTTQKMVAASGMD</td>
<td>GGTSRRRTNSKS</td>
<td>TQYSTWGGKYQNSQF</td>
<td>GFSIRRSSNVGVTLMDFKT</td>
<td>EIGSRGNNTMFDFKFT</td>
<td>ARSGKTEAQGTRGGWMDEDDNDKA EQTISA KT NQKR QY E NM M AD M N</td>
</tr>
<tr>
<td>SQNM YD AN F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.4: Two Libraries of Fluorescein-Binding Antibody CDRs. The amino acid sequences of the two OptCDR designed fluorescein-binding antibodies and the predicted mutations to the CDRs to form libraries of antibodies.

<table>
<thead>
<tr>
<th>Design</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ETQSSGGSNTDKSWVE</td>
<td>GGTKSRT</td>
<td>TQWKEQGSNGA</td>
<td>EVTGTQFGVN</td>
<td>SVSGDAKTG</td>
<td>KKNMAGDDEGRGDSSTNRAQ</td>
</tr>
<tr>
<td></td>
<td>S N Q M</td>
<td>S S</td>
<td>N S A</td>
<td>A N F</td>
<td>I AGGNN</td>
<td>D A ST NDD QLAANKQEE</td>
</tr>
<tr>
<td></td>
<td>T N A</td>
<td>N A</td>
<td>A N M</td>
<td>Y Y</td>
<td>D K</td>
<td>S D DD E N DQ SN MR L</td>
</tr>
<tr>
<td>2</td>
<td>ESSQLVHSGKTFMS</td>
<td>EGTVSKG</td>
<td>QNFSSYGGKFQ</td>
<td>GFSITRTEGVS</td>
<td>GVSSTGKTE</td>
<td>AGGQYESASKKE</td>
</tr>
<tr>
<td></td>
<td>AT N LDKN</td>
<td>N T</td>
<td>N Y T</td>
<td>L Y W</td>
<td>I AS N</td>
<td>S N A TR</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>A</td>
<td>N</td>
<td>F</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


8. GAMS Home Page [http://www.gams.com]


matured therapeutic antibody using structure-based computational design.


Chapter 3. The Database of Modular Antibody Parts
The OptCDR method addresses the *de novo* design of antigen binding through the CDRs. However, FR residues can contribute to the binding capacity of an antibody and OptCDR did not constrain the CDR sequences to be fully human. To address these issues, a computational analogue to the human immune response is envisioned. The immune response beings with the generation of antibodies by V-(D)-J recombination of human germline genes. Therefore, a database of Modular Antibody Parts (MAPs) that are analogous to the human germline genes was created for predicting antibody tertiary structures. Figure 3.1 illustrates the MAPs workflow, which allows for predicting the structure of any mutated (usually affinity matured) antibody. First, the prototype sequences for VH and VL are predicted from germline genes. Next, model structures of the prototype sequences are created by identifying and assembling the closest MAPs structures. As detailed below, the MAPs database has structures for V* (V region FR1 to FR3), CDR3, and J* (J region FR4). Finally, the antibody structure is predicted by incorporating the amino acid changes of the antibody compared to the prototype.

**Construction and content**

The structurally conserved features of antibodies allow for standardized numbering schemes describing each position in an antibody, including the IMGT® unique numbering [1-4] from IMGT®, the international ImMunoGeneTics information system® [5, 6]. IMGT® is a well-curated source of antibody sequences, genes, structures, and standards used extensively in this chapter. Figure 3.2 shows the IMGT® unique numbering [1-4], the approximate start and end points of the V, D, and J regions (which depend on the length of the germline regions and trimming during V-(D)-
J rearrangement), the conserved start and end points of the FRs and CDRs, and the location of FRs and CDRs within an antibody variable domain.

**Germline Gene Assignment**

The first step in Figure 3.1 is the assignment of germline V, D, and J genes to an antibody to identify a prototype sequence. From the IMGT/GENE-DB [7] we downloaded all human germline V, D, and J genes and retained all complete (i.e., no missing nucleotides) and unique genes and alleles, spanning a total diversity of approximately $10^8$ possible antibodies (see Table 3.1). The adopted hypothesis for assigning germline genes to an antibody variable domain is that the gene combination with the fewest amino acid changes from the corresponding portion of the antibody’s sequence is the one most likely to have been used in V-(D)-J recombination.

When compared to the germline genes, the junctions (in IMGT® positions 104-118) which result from the V-(D)-J rearrangement usually show nucleotide deletions at the 3’ V region and 5’ J region, and for a VH on both ends of the D region, due to exonuclease trimming. The junctions also show random nucleotide insertions by the terminal deoxynucleotidyl transferase [8]. For the VH, this makes the assignment of the D gene only possible at the nucleotide level [9, 10]. Therefore, the determination of the number of amino acid changes between a set of V, D, and J regions and a VH or VL is carried out as follows:

1. The total number of amino acid changes in the V and J regions outside of CDR3 is counted. Note that amino acid insertions and deletions are rare (observed rates of 0.00-0.09% per region) outside of CDR3 and are penalized as amino acid changes.
2. The CDR3 nucleotides of the 3’ V and 5’ J regions are fixed in place at the start and end, respectively, of CDR3.

3. In CDR3, the V and J combination is first evaluated for the presence of nucleotide gaps or overlaps (i.e., insertions or deletions). If there is a gap between the end of the V region and the start of the J region, the gap is filled so that the minimum number of amino acid changes is introduced. Conversely, if the regions overlap, deletions that cause the fewest amino acid changes are chosen. Then, the total number of amino acid changes in CDR3 is counted. Note that nucleotide insertions and deletions are penalized based on the number of amino acids they affect (e.g., 1-3 insertions affect 1 amino acid, 4-6 affects 2, etc.). This procedure is sufficient for VL and for the few VH that lack D genes.

4. For VH, the D region is positioned at every possible position from the beginning of CDR3 to the end. For each possible position, the number of amino acids changes is assessed as in Step 3. The V, D, and J gene combination that results in the fewest amino acid changes is selected as the prototype sequence for the target antibody. In the event that two or more sets of genes have the same number of amino acid changes, the minimum number of nucleotide changes is used as a tie-breaker (e.g., mutating Ile (att) to Val (gtt) or Thr (act) requires a minimum of one nucleotide change, but mutating Val to Thr requires at least two). This procedure is very fast for VL (< 0.5 s) and reasonably quick for VH (5 – 60 s) due to the added complexity of considering the D regions.

To test the gene assignment protocol, we downloaded 7,652 VH, 2,247 V-KAPPA and 1,605 V-LAMBDA unique, human antibodies from IMGT/LIGM-DB [11] and assigned
germline genes to them. The observed rates and number of amino acid changes in each of
the FRs and CDRs along with amino acid insertions and deletions in CDR3 are shown in
Figures 3.3 and 3.4. The observed frequencies and locations of amino acid changes match
expected trends. The average number of amino acid changes (i.e., VH 17.6, V-KAPPA
and V-LAMBDA both 9.0) compare quite well with the expected 10-15 amino acid
changes per variable domain [12]. In addition, the average rate of amino acid changes in
the “hypervariable” CDRs is much higher than in the FRs. This is especially pronounced
in the VH CDR3, which is well-known to often have the most antigen contacting
residues. Interestingly, FR3 also accumulates a much higher number of amino acid
changes in the VH compared with the VL. The confirmation of expected experimental
trends in the average frequency and number of amino acid changes alludes to the efficacy
of the gene assignment protocol.

**V*, **CDR3**, and **J* Structure Delimitation**

With this established gene assignment protocol, we next turn our attention to determining
if antibodies with the same germline genes assume the same structures. Reliable
prediction of antibody tertiary structures, modeled from their prototype amino acid
sequences, hinges upon the hypothesis that antibody regions that share the same
prototype sequence assume similar structures. We downloaded 1,168 human, humanized,
chimeric, and mouse antibody structures from the IMGT/3Dstructure-DB [13, 14] as well
as all complete mouse IGH, IGK, and IGL V, D, and J genes from the IMGT/GENE-DB
[7]. Prototype sequences were identified for all antibodies, with human genes used for
human antibodies and mouse genes used for mouse, chimeric, and humanized antibodies.
Note that chimeric antibodies have mouse variable domains and humanized antibodies
have mouse CDRs attached to human FRs. Mouse genes were used for the humanized antibodies so that CDR3 were modeled using the appropriate genes.

With the prototype sequences determined for the antibody structures, a clustering procedure similar to one used for just the CDRs in OptCDR was used to determine if identical regions give rise to similar structures. At the end of the clustering procedure, the structure with the smallest average backbone atom (N, Ca, C) RMSD with all other structures in a cluster was selected as the model structure. The clustering was carried out so that all members of a cluster have a backbone atom RMSD of no more than 2.0 Å with the model structure. This distance cutoff was also used to assess previous antibody structure prediction methods [15-17].

An initial analysis was conducted to determine if using models of the V, D, and J genes was an effective approach to generate the MAPs database. Figure 3.5 gives representative results of the clustering process. For the V regions and the light J regions the procedure led to almost entirely unique inferences for structure, but for the VH D and J regions antibodies with identical regions often had different structures. This implies that assignment of the germline regions may be sufficient to predict the structures of VL but is insufficient for VH. However, we observed that for both VH and VL the clustering procedure appeared to work well outside of CDR3 and poorly within CDR3. This includes the 3’ V regions where “fraying” of the clusters was observed in the last few residues. This suggests that a modified description, listed in Table 3.2, utilizing CDR3 as a structural component instead of D regions may improve the description of antibody structures. We therefore selected an alternative delimitation of structures for both VH and VL: (i) V* (FR1-FR3, IMGT® positions 1-104), containing all the sequence before
CDR3 (i.e., most of the V region), (ii) CDR3 (IMGT® positions 105-117), and (iii) J* (FR4, IMGT® positions 118-128), encompassing everything after CDR3 (i.e., most of the J region). This alternative concatenation of the genetic information provides a more succinct description of structural diversity.

When the clustering procedure is applied to the antibodies based on the prototype V*, CDR3, and J* sequences the clustering results show one structure per one prototype sequence. After discounting obvious explanations for structural differences (e.g. unusual antibody structures like light chain dimers or triabodies, mutations known to alter crystal packing, structural perturbations caused by the linkers in single-chain variable fragments, antigen binding causing conformational changes, etc.), only 9 VH V*, 18 VH CDR3, and 6 V-KAPPA CDR3 structures out of the 1168 antibodies used were found to differ from other structures with the same prototype sequence. All other regions had at most a single outlier. Most of the differing CDR3 structures were those that were naturally diverse, owing to the expected high diversity of their junctions, suggesting that larger structural changes are to be expected.

**Database Generation and Use**

The observation that each prototype V*, CDR3, and J* sequence has a single structure indicates that a database of these structures can be used to model antibody variable domains. The MAPs database was constructed to contain as much structural diversity as possible. For each V*, CDR3, and J* prototype sequence, the antibody from the 1168 structures that required the fewest amino acid changes was selected as the model. In the event that two or more antibodies shared the same minimum number of amino acid changes, the structure with the smallest average backbone atom RMSD with the other
possible models was chosen. The model structures had their sequences mutated to the germline sequence using IPRO’s optimal rotamer selection procedure [18, 19] and any structural inconsistencies were then corrected with a CHARMM energy minimization [20]. All CHARMM energy minimizations in this chapter were molecular mechanics minimizations and were carried out using the “all27_prot_na” topology and parameter files, the angl, bond, dihe, elec, impr, urey, and vdw energy terms, and no solvation. The created models were stored in the MAPs database as PDB files using the IMGT® unique numbering for the V domain [1-4].

Each unique human J* prototype in Table 3.1 was modeled using this procedure. All J* models required two or fewer amino acid changes. Human V* prototypes shown in Table 3.1 were also similarly modeled using a cutoff on the maximum number of amino acid changes allowed. No V* prototype was selected if it required more than one standard deviation greater than the average number of amino acid changes expected (cutoffs of 16 amino acid changes for heavy and 14 for both κ and λ light, as determined from the data in Figure 3.3). This cutoff was imposed to ensure that the V* models were accurate structural representations of their sequences.

Each CDR3 prototype was modeled using a cutoff on the maximum number of allowable amino acid changes. An analysis of the RMSDs between all pairs of VH CDR3 with the same number of amino acids shows that on average one extra amino acid change gives rise to a 1/3 Å increase in backbone RMSD ($R^2 = 0.97$). Therefore a cutoff of six amino acid changes was used, as this number of changes would likely cause an average change of 2.0 Å in backbone RMSD, which is the similarity cutoff used during the clustering procedure and in previous antibody structure prediction methods [15-17]. Prototype
CDR3s whose model required six or fewer amino acid changes were changed to the predicted sequence in the same manner as the V* and J* models, while those requiring seven or more changes were included using the experimentally determined model structure. The modeling of the CDR3 ensured that the maximum amount of structural diversity was included in the MAPs database for this essential binding feature, at the expense of introducing some mouse and non-prototype sequences.

The size statistics of the MAPs database are presented in Table 3.3. The MAPs database is composed of 929 “parts” that can be assembled to create $2.3 \times 10^{10}$ unique antibodies. This is in fact more antibodies than can be assembled by the human immune system through rearrangement of the V, D, and J genes. However, the complex mechanisms and junctional diversity of V-(D)-J recombination significantly increases the number of antibodies that the immune system can potentially generate. The MAPs database contains all currently observed structural diversity of antibodies with CDR that can encompass a wide range of possible positions.

The MAPs database can be used to model antibody structures as shown in Figure 3.1. For a target affinity matured antibody with an unknown structure, a prototype sequence is first computationally identified. Next, the V*, CDR3, and J* structures in the MAPs database that have the closest sequence to the prototype are identified. The models are mutated to the prototype sequence using the optimal rotamer selection MILP from IPRO [18, 19] and relaxed using a CHARMM energy minimization [20] step. Finally, mutating the prototype antibody followed by another CHARMM energy minimization generates the predicted structure of the target antibody.
Database Evaluation

The efficacy of the MAPs database for predicting antibody tertiary structures from their amino acid sequences was assessed. A cross-validation set of 260 antibodies from the 1168 downloaded was selected. These 260 antibodies were not used in creating any of the model structures of the database, contained both VH and VL, and had experimental resolutions no worse than 2.5 Å. The structures were predicted using the workflow from Figure 3.1 with a mean RMSD of 1.900 ± 0.325 Å accounting for all atoms in all residues in both variable domains. The mean experimental resolution of the structures was 2.074 ± 0.274 Å. As the V* and J* model structures are based on human genes, it is to be expected that the predicted human antibodies (56 out of 260) had a slightly better mean RMSD than the predicted mouse, chimeric, and humanized antibodies (i.e. 1.771 ± 0.184 Å versus 1.933 ± 0.345 Å). In contrast, there is no significant difference between the mean RMSDs of the 207 bound antibody complexes (i.e. 1.876 ± 0.320 Å) and the 53 unbound complexes (i.e. 1.992 ± 0.330 Å).

We relied on three popular online servers for predicting antibody structures (i.e., WAM [17], Prediction of ImmunGlobulin Structure (PIGS) [15], and RosettaAntibody [16]) to benchmark the effectiveness of the introduced method. WAM’s published results show that 16 out of 19 (i.e., 84%) predicted antibodies had VH CDR3 backbone RMSD values no worse than 2.0 Å. Our results meet the same RMSD criterion for a much larger set (i.e., 249 out of 260) and percentage (i.e., 96%) of predicted structures. PIGS provided the backbone RMSD of all amino acids in four antibodies (i.e., 1.08, 1.11, 1.16, and 1.42 Å). The corresponding mean and median using the MAPs database are comparable (1.134 ± 0.365 Å and 1.022 Å, respectively). RosettaAntibody published RMSD results for the
backbone atoms of all amino acids in the CDRs of 54 antibodies that have a median RMSD of 1.4 Å with 80% of them less than 2.0 Å. MAPs based structure prediction yields a median RMSD of 1.256 Å with 209 out of 260 (80%) better than 2.0 Å. These results demonstrate that using the MAPs database to predict antibody structures is at least as accurate as existing methods and in most cases better.

**Broadly-Neutralizing Antibodies**

We also briefly explored the efficacy of using the MAPs database to support antibody engineering and design using two broadly neutralizing antibodies: the anti-influenza antibody CH65 [PDB: 3SM5] [21] and the anti-HIV antibody 4E10 [PDB: 2FX7] [22].

The experimental and MAPs predicted structures for CH65 and 4E10 (see Figure 3.6) have an all-atom RMSD of 2.046 and 2.099 Å, respectively. Due to its publication date, the structure of CH65 was not one of the 1168 considered when creating the MAPs database. The overall quality of its MAPS based predicted structure is comparable to that of 4E10 with some structure prediction discrepancies only within the VH CDR3. This CDR is 20 amino acids long and has at least a ten amino acid difference from any VH CDR3 structure in the MAPs database. Nevertheless, the backbone atom RMSD of the VH CDR3 from the experimental structure is only 2.160 Å.

The first step in the proposed workflow for antibody structure prediction in Figure 3.1 is the prediction of the best prototype sequences for any given antibody. Table 3.4 lists the amino acid changes in the ten best prototype sequences for antibody CH65, containing 21 positions in VH and 12 positions in V-LAMBDA. The VH amino acid changes are predominantly allotted within FR3 and CDR3 while the V-LAMBDA amino acid changes are distributed throughout the sequence. A more uniform distribution of amino
acid changes in VH could be obtained, at the expense of introducing more total amino acid changes, by limiting how often each gene is used in the predicted results. For example, this may be desirable in the construction of a combinatorial library to bind a particular antigen epitope.

The identification of the prototype sequence and the generation of prototype and affinity matured models simplifies the assessment of interesting amino acid changes relevant to a particular antibody. The anti-HIV antibody 4E10 contains 28 VH and 14 V-KAPPA amino acid changes from the prototype sequence. Table 3.5 reports the changes in the computed interaction energy between the antibody and antigen for each one of the 42 amino acid changes accumulated during affinity maturation. The mutations were created using the optimal rotamer selection MILP from IPRO and a CHARMM energy minimization. The computed interaction energies should be considered for the qualitative trends they imply rather than their absolute values. Only two of the observed amino acid changes are computationally predicted to be detrimental to binding while 37 are predicted to be weakly and three strongly beneficial. This suggests that the amino acid changes needed to convert the prototype structure to the affinity matured antibody generally improve binding. This lends credibility to the computational identification of prototypes because affinity maturation, as expected, led to binding improving whereas random mutations would have been detrimental to binding.

**Discussion**

This chapter introduced a database of Modular Antibody Parts that can be used in the *de novo* design of antibodies in an analogous fashion to V-(D)-J recombination. Using the structural diversity encompassed within 1168 experimental antibody structures we
compiled the MAPs database that contains 929 parts that can be combined to create $2.3 \times 10^{10}$ unique antibodies. The prediction of 260 antibody structures not used in making any of the MAPs database models revealed that this database can be used to reliably predict antibody tertiary structures. In contrast to previous antibody structure prediction methods [15-17], MAPs allows for antibody structure prediction without the need for de novo folding calculations every time. The all-atom, modular nature of the MAPs database allows for the pre-calculation of pairwise structural component interaction energies. The computational savings do not come at the expense of accuracy of prediction as the RMSD of the predicted structures is at least as accurate as earlier methods.

Antibody structures are known to be affected to some extent by the relative orientation of the VH and VL, the specific canonical structures used, and the FRs to which they attach. However, despite ignoring all these factors, the procedure described in Figure 3.1 operating on the MAPs database provides very high fidelity of antibody structure prediction, alluding to a high degree of modularity of available conformations.

The broadly-neutralizing anti-HIV antibody 4E10 was used to demonstrate the ability of the MAPs database to model the nature of amino acid changes upon affinity maturation. Almost all of the accumulated amino acid changes were predicted to be beneficial to binding, thus providing indirect evidence regarding the validity of the prototypes. The proposed workflow is generally agnostic to the method used to identify prototypes for an antibody. If a user has higher confidence in a prototype sequence different from the one identified by the method described here, it is possible to directly import it in the calculations.
Currently, almost all antibodies are designed entirely using experimental methods. Once a promising antibody is identified additional affinity improvements are sought after using random mutagenesis and directed evolution protocols. Knowing where to target mutations and what type of mutations to explore can greatly improve the efficiency of experimental methods. In the germline repertoire of V, D, and J genes, evolution has retained many similar genes with amino acid changes at key positions that are likely to influence binding to a range of substrates. As described for the broadly neutralizing anti-influenza antibody CH65, it is possible to rapidly identify multiple potential prototypes for any given antibody. By contrasting the mutations between the target antibody and the prototypes both promising positions and amino acid combinations likely to confer improved binding affinity can be quickly compiled providing cues for combinatorial library design.
**Figure 3.1: MAPs Antibody Structure Prediction Workflow.** The MAPs database can be used to predict an antibody structure in three steps. First, a prototype sequence is predicted. In this work, that is accomplished by assigning germline genes that minimize the number of amino acid changes between the prototype and target antibody. However, prototype sequences from alternative methods that a user has confidence in can be used. In the second step, a prototype model is created by assembling and mutating the most similar models in the MAPs database. Finally, a predicted model of the antibody is created by mutating the prototype structure. The variable domain depicted for steps two and three is the VH of PDB 3NCJ [23] with four of the mutations highlighted.

1) **Prototype Sequence Prediction**

Assign V, D, and J germline genes with the fewest amino acid changes from the mature antibody sequence

Antibody: `evq...lKi...tNy...carQLY...tvss`
Prototype: `evq...lRi...tSy...carGYS...tvss`

2) **Prototype Structure Prediction**

Identify MAPs structures most similar in sequence to the predicted prototype sequence, assemble them, and mutate to the prototype sequence

3) **Mature Structure Prediction**

Mutate the predicted prototype structure to match the affinity mature antibody
**Figure 3.2: Details of Antibody Sequence and Structure.** The IMGT® unique numbering scheme is shown in a vertical orientation. This includes the start and end points of the FRs and CDRs, the corresponding structures in the VH of the broadly neutralizing anti-HIV antibody 4E10, and the approximate start and end points of the V, D, and J regions. The listed starting and ending positions of the V, D, and J regions are only approximate due to the small variations in the lengths of these genes and/or owing to trimming by exonuclease. The starting and ending positions of the FRs and CDRs are shown with a larger font size. VL lack D genes but generally have longer V genes to compensate.
Figure 3.3: Observed Amino Acid Change Numbers in Human Antibodies. Here we show the numbers of amino acid changes in each of the seven regions of antibody structure for VH, V-KAPPA, and V-LAMBDA. Panels G and H show the number of nucleotide insertions and deletions in CDR3, as this is the only region where those events are common.
Figure 3.4: Observed Amino Acid Change Frequencies in Human Antibodies. Here we show the percent of amino acids that are changed in each of the seven regions of antibody structure for VH, V-KAPPA, and V-LAMBDA. Panels G and H show the percentage of nucleotide insertions and deletions in CDR3, as this is the only region where those events are common.
Figure 3.5: Clustering of V, D, and J Region Structures. These panels show representative clusters of structures with the same V, D, and J regions, respectively. All structures in the same cluster have the same color. For the VH V region and VL, the clustering worked well. However, for the VH D and J regions the clustering procedure did not work (i.e., antibodies with the same germline regions have different structures). Interestingly, all of the deviations occur within CDR3. This observation led to the development of an alternative delimitation of structure: V* (everything before CDR3, FR1-FR3, IMGT positions 1-104), CDR3, and J* (everything after CDR3, FR4, IMGT positions 118-128).
Figure 3.6: The Experimental and MAPs Predicted Structures of Antibodies CH65 and 4E10. The experimental structures are shown in light gray and the MAPs predicted structures are shown in black. CH65 has an all-atom RMSD of 2.046 Å and 4E10 has an all-atom RMSD of 2.099 Å.
Table 3.1: Human Germline V, D, and J Genes and Alleles. Germline genes and alleles downloaded from the IMGT/GENE-DB [7]. Only the genes and alleles that had no missing nucleotides were retained for use in assigning germline genes to an antibody.

<table>
<thead>
<tr>
<th></th>
<th>VH</th>
<th>V-KAPPA</th>
<th>V-LAMBDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Retained</td>
<td>Total</td>
</tr>
<tr>
<td>V</td>
<td>324</td>
<td>236</td>
<td>97</td>
</tr>
<tr>
<td>D</td>
<td>44</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>13</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>Antibodies</td>
<td>1.85 $10^5$</td>
<td>9.61 $10^4$</td>
<td>873</td>
</tr>
<tr>
<td>Total Antibodies</td>
<td>$3.21 \times 10^8$</td>
<td>Retained Antibodies</td>
<td>$1.29 \times 10^8$</td>
</tr>
</tbody>
</table>
Table 3.2: Alternative Delimitation of Structure. Positions are according to the IMGT® unique numbering. Positions 104 (2nd-CYS) and 118 (J-PHE or J-TRP) belong to FR3 and FR4, respectively, and are the anchors of CDR3.

<table>
<thead>
<tr>
<th>V* (FR1-FR3)</th>
<th>IMGT Start</th>
<th>IMGT End</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDR3</td>
<td>105</td>
<td>117</td>
</tr>
<tr>
<td>J* (FR4)</td>
<td>118</td>
<td>128</td>
</tr>
</tbody>
</table>
Table 3.3: Number of Model Structures in the MAPs Database. The number of models for each region of structure for each type of antibody variable domain is tabulated.

<table>
<thead>
<tr>
<th></th>
<th>VH</th>
<th>V-KAPPA</th>
<th>V-LAMBDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>V*</td>
<td>141</td>
<td>67</td>
<td>38</td>
</tr>
<tr>
<td>CDR3</td>
<td>428</td>
<td>199</td>
<td>39</td>
</tr>
<tr>
<td>J*</td>
<td>5</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>$3.0 \times 10^5$</td>
<td>$6.7 \times 10^3$</td>
<td>$1.0 \times 10^4$</td>
</tr>
<tr>
<td>Possible Antibodies</td>
<td></td>
<td></td>
<td>$2.3 \times 10^{10}$</td>
</tr>
</tbody>
</table>
Table 3.4: Amino Acid Changes in the Ten Best Predicted Prototypes for the Broadly-Neutralizing anti-Influenza Antibody CH65. The amino acid changes of the top ten prototype (P) sequences (21 VH, 12 V-LAMBDA positions) are listed. In VH all positions outside of CDR3 have the same amino acid listed because the same V gene was selected. Within CDR3 different D gene selections give rise to different amino acid changes. In V-LAMBDA different V and J gene selections give rise to different amino acid changes throughout the entire domain. Sequence numbering is according to the IMGT® unique numbering [1-4].

<table>
<thead>
<tr>
<th>Position</th>
<th>Antibody</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
</tr>
<tr>
<td>36</td>
<td>D</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>38</td>
<td>H</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>39</td>
<td>I</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>40</td>
<td>N</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>57</td>
<td>H</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>64</td>
<td>D</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>83</td>
<td>A</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>91</td>
<td>V</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>92</td>
<td>N</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>93</td>
<td>G</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>95</td>
<td>K</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>107</td>
<td>G</td>
<td>V</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>108</td>
<td>G</td>
<td>Q</td>
<td>S</td>
<td>I</td>
<td>W</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>Y</td>
</tr>
<tr>
<td>109</td>
<td>L</td>
<td>L</td>
<td>I</td>
<td>V</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>S</td>
</tr>
<tr>
<td>110</td>
<td>E</td>
<td>E</td>
<td>A</td>
<td>G</td>
<td>Q</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>S</td>
</tr>
<tr>
<td>111</td>
<td>P</td>
<td>R</td>
<td>A</td>
<td>A</td>
<td>L</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>111A</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>T</td>
<td>W</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>W</td>
</tr>
<tr>
<td>111B</td>
<td>S</td>
<td>S</td>
<td>P</td>
<td>T</td>
<td>L</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>Y</td>
</tr>
<tr>
<td>111C</td>
<td>V</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>H</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>112C</td>
<td>D</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position</th>
<th>Antibody</th>
<th>2</th>
<th>17</th>
<th>27</th>
<th>36</th>
<th>42</th>
<th>54</th>
<th>55</th>
<th>56</th>
<th>116</th>
<th>117</th>
<th>123</th>
<th>126</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>S</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>17</td>
<td>Q</td>
<td>Q</td>
<td>K</td>
<td>K</td>
<td>Q</td>
<td>K</td>
<td>K</td>
<td>Q</td>
<td>K</td>
<td>K</td>
<td>K</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>27</td>
<td>D</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>36</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>42</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>54</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>I</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>I</td>
<td>V</td>
<td>I</td>
<td>V</td>
<td>I</td>
<td>V</td>
</tr>
<tr>
<td>55</td>
<td>C</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>56</td>
<td>Y</td>
<td>D</td>
<td>Y</td>
<td>D</td>
<td>D</td>
<td>Y</td>
<td>D</td>
<td>D</td>
<td>Y</td>
<td>D</td>
<td>D</td>
<td>Y</td>
<td>D</td>
</tr>
<tr>
<td>116</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>W</td>
<td>A</td>
<td>W</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>117</td>
<td>I</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>123</td>
<td>K</td>
<td>K</td>
<td>K</td>
<td>K</td>
<td>Q</td>
<td>K</td>
<td>K</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
</tr>
<tr>
<td>126</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
</tbody>
</table>
Table 3.5: Predicted Changes in Interaction Energy for Amino Acid Changes in the Broadly-Neutralizing anti-HIV Antibody 4E10. Note that 37 of 42 amino acid changes are predicted to be weakly beneficial, 3 are predicted to be strongly beneficial, and 2 are predicted to be weakly detrimental to antigen binding.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Mutation</th>
<th>ΔΔG (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VH</td>
<td>K 14 R</td>
<td>-2.54</td>
</tr>
<tr>
<td>VH</td>
<td>K 20 T</td>
<td>-1.975</td>
</tr>
<tr>
<td>VH</td>
<td>T 29 S</td>
<td>-3.071</td>
</tr>
<tr>
<td>VH</td>
<td>S 36 T</td>
<td>-2.343</td>
</tr>
<tr>
<td>VH</td>
<td>I 39 L</td>
<td>-3.495</td>
</tr>
<tr>
<td>VH</td>
<td>Q 48 R</td>
<td>-2.846</td>
</tr>
<tr>
<td>VH</td>
<td>I 56 V</td>
<td>-1.889</td>
</tr>
<tr>
<td>VH</td>
<td>I 59 L</td>
<td>-3.449</td>
</tr>
<tr>
<td>VH</td>
<td>G 63 T</td>
<td>-2.573</td>
</tr>
<tr>
<td>VH</td>
<td>A 65 T</td>
<td>-1.801</td>
</tr>
<tr>
<td>VH</td>
<td>Q 69 P</td>
<td>-1.691</td>
</tr>
<tr>
<td>VH</td>
<td>K 70 R</td>
<td>-2.226</td>
</tr>
<tr>
<td>VH</td>
<td>V 76 I</td>
<td>-1.678</td>
</tr>
<tr>
<td>VH</td>
<td>K 82 R</td>
<td>-2.143</td>
</tr>
<tr>
<td>VH</td>
<td>M 89 L</td>
<td>-2.158</td>
</tr>
<tr>
<td>VH</td>
<td>S 92 N</td>
<td>-2.549</td>
</tr>
<tr>
<td>VH</td>
<td>S 96 P</td>
<td>-2.451</td>
</tr>
<tr>
<td>VH</td>
<td>I 109 T</td>
<td>-0.923</td>
</tr>
<tr>
<td>VH</td>
<td>I 111 G</td>
<td>-1.026</td>
</tr>
<tr>
<td>VH</td>
<td>F 111a W</td>
<td>-2.648</td>
</tr>
<tr>
<td>VH</td>
<td>V 111c W</td>
<td>-3.922</td>
</tr>
<tr>
<td>VH</td>
<td>V 112d L</td>
<td>-2.517</td>
</tr>
<tr>
<td>VH</td>
<td>I 112c G</td>
<td>-3.328</td>
</tr>
<tr>
<td>VH</td>
<td>I 112b K</td>
<td>-45.027</td>
</tr>
<tr>
<td>VH</td>
<td>D 113 G</td>
<td>-45.886</td>
</tr>
<tr>
<td>VH</td>
<td>D 116 A</td>
<td>5.989</td>
</tr>
<tr>
<td>VH</td>
<td>V 117 H</td>
<td>-0.445</td>
</tr>
<tr>
<td>VH</td>
<td>M 123 L</td>
<td>-1.908</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>L 11 Q</td>
<td>-2.11</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>S 30 G</td>
<td>-3.713</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>S 36 N</td>
<td>0.412</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>S 37 N</td>
<td>-2.128</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>Y 38 K</td>
<td>-55.314</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>K 45 R</td>
<td>-2.614</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>A 68 P</td>
<td>-2.604</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>T 69 S</td>
<td>-2.681</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>I 71 V</td>
<td>-2.177</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>P 72 A</td>
<td>-2.41</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>S 109 Q</td>
<td>-2.969</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>P 115 L</td>
<td>-4.398</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>W 116 S</td>
<td>-5.325</td>
</tr>
<tr>
<td>V-KAPPA</td>
<td>I 126 V</td>
<td>-2.531</td>
</tr>
</tbody>
</table>
References


2. Lefranc MP: IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harbor protocols* 2011, 2011(6):633-642.

3. Lefranc MP: IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harbor protocols* 2011, 2011(6):643-651.


13. Ehrenmann F, Kaas Q, Lefranc MP: **IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or**
antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic acids research* 2010, **38**(Database issue):D301-307.


Chapter 4. The IPRO Suite of Programs
Calculations using the methods presented in Chapters 2 and 3, as well as work outside the scope of this dissertation, suggested that significant improvements were needed in the computational protein engineering tools used by the C. Maranas Lab. In the last decade, the C. Maranas Lab has developed a number of computational protein design methods. The Iterative Protein Redesign & Optimization (IPRO) [1, 2] method uses alternating protein backbone perturbations and amino acid sequence mutations to improve the binding of proteins to novel ligands. IPRO was used to change the cofactor specificity of *Candida boidinii* xylose reductase from its native cofactor of nicotinamide adenine dinucleotide phosphate to nicotinamide adenine dinucleotide. OptGraft [3] identifies the residues in a protein that can best support a novel binding site and then finds mutations in surrounding residues to ensure the proper geometry of the binding residues. It was used to introduce a calcium-binding site from thioredoxin into the first domain of CD2. OptZyme [4] uses ground state substrates and transition state analogues to improve the catalytic properties of enzymes. The OptCDR method designs novel CDRs to bind any specified antigen epitope and was described in Chapter 2.

However, while these various protein design methods all share common features, the developed codes were typically incompatible with one another and difficult to learn to use. In this chapter the development of an integrated IPRO Suite of Programs to rectify these issues is described. The IPRO Suite shares a common, modular core of code that allows for full-integration of current and future protein engineering methods with one another. In addition, there is now a single user interface that is substantially easier to learn how to use than the previous programs. The current implementation of the IPRO Suite of Programs being described here contains IPRO, OptGraft, OptCDR, OptZyme,
the MAPs database for antibody structure prediction from Chapter 3, and a Mutator program for predicting the properties of specific protein mutants.

**Implementation**

The IPRO Suite of Programs is a collection of programs, files, and folder for use in the UNIX operating system. The programs are primarily written in Python, as well as C++ and GAMS [5]. The IPRO Suite of Programs currently only supports the use of the CHARMM [6] force field and the PDB [7] file format. There is a long-term goal of supporting additional force fields and molecular structure formats.

Figure 4.1 shows the contents of the IPRO Suite folder, which are seven folders and the Python user-interface program, Start_Experiment.py. The “databases” folder is a repository for the MAPs database and the canonical structures used in OptCDR. “Experiments” is a folder where individual IPRO Suite experiments are conducted. The “input_files” folder is a location to store various input files (e.g. topology and parameter files for CHARMM). “Modules” stores the Python and C++ functions that are needed for running the programs of the IPRO Suite, which are stored in the “programs” folder. “References” contains constant information related to the IPRO Suite of Programs (e.g. the positions of the CDRs in OptCDR) and “structures” is where files of molecules are stored.

All user inputs to the IPRO Suite of Programs are handled by the Start_Experiment.py program. This program asks the user for all of the information needed to run a particular type of IPRO Suite experiment, validates that the information is acceptable as it is collected, and starts the calculations of the experiment. Figure 4.2 illustrates how the Start_Experiment.py program begins: by identifying who is running the experiment, the
type of IPRO Suite of Programs experiment they want to run, and finding a unique name for the experiment. A folder with that name will be created in the “experiments” folder of the IPRO Suite and all calculations for the experiment will be carried out there.

Once the basic information of the experiment has been provided, the user is next asked to identify the Design Molecules (DMs) and Target Molecules (TMs) for the experiment. DMs are the proteins the user is designing to have modified interactions with the TMs. To understand this terminology, consider the hypothetical situation of mutating the VH domain of the antibody in PDB: 3KR3 [8] to have improved binding to insulin-like growth factor-II (IGF-2) while preventing binding to insulin-like growth factor-I (IGF-1) [PDB: 1IMX] [9], as illustrated in Figure 4.3. In this case, the VH and VL of the antibody are DMs and IGF-1 and IGF-2 are TMs.

After the molecules being used in the experiment have been specified, information about how to use the force field is asked for (i.e. input files and energy terms). Also, certain calculations necessitate the use of pairwise additive energy functions and the generalized Born with molecular volume integration method utilized for implicit solvation in CHARMM [10] is not pairwise additive. Therefore, if the user elects to use implicit solvation, they must also provide input files for how to use Lazaridis-Karplus implicit solvation [11], a method that is pairwise additive.

Next, the “Start_Experiment.py” program will automatically confirm that the provided DM and TM structures are complete (i.e. no missing residues) and compatible with the force field and implicit solvation input files. Once this validation is completed, the user will be prompted to provide many other pieces of information specific to the type of IPRO Suite of Programs experiment being run. Two of the most important types of
information for most kinds of IPRO Suite experiments are Design Positions (DPs) and Design Groups (DGs). A DP is a residue in a DM that is permitted to mutate to modify the interactions with the TMs. Note that a DM is permitted to not have any DPs, as in the example in Figure 4.3 where the heavy chain of the antibody is being mutated and the light chain is not. A DG is a subset of TMs that the DMs are interacting with at the same time and with the same objective for modifying their interactions. In the example in Figure 4.3, IGF-2 is in one DG, where the objective is to improve binding between the DMs and TMs, and IGF-1 is in a second DG, where the objective is to suppress binding.

The IPRO Suite of Programs has four ways in which it can modify the interactions between DMs and TMs in a DG: binding may be continuously improved, it may be maintained above or below its initial value, or it may be constantly suppressed. The remainder of the information specific to each type of IPRO Suite of Programs experiment will be described when discussing the individual programs. Henceforth, $U$ will be used to indicate any setting that the user is permitted to specify themselves and default values will always be noted. The default values are stored in the STANDARDS.py file in the modules folder of the IPRO Suite. Figure 4.4 shows the start of the file. Upon installation of the IPRO Suite of Programs, certain portions of this file must be edited (e.g. the InstallFolder needs to be the global system path of the IPRO Suite folder) and the default settings may be changed to match a user’s expected defaults. These specifications, and others about how to use the IPRO Suite of Programs, are stored in a README.txt file in the modules folder.

Once a user has provided all of the necessary information for how to run an IPRO Suite of Programs experiment, the experiment’s folder is set up. The folder corresponding to
the example in Figure 4.2 is shown in Figure 4.5. It contains three folders, two text files, and one Python program. The “structures” and “input_files” folders respectively contain the necessary molecule structures and input files for the experiment. The “results” folder is where the different structures generated by the experiment are stored. All structures created by the IPRO Suite of Programs use the residue numbering of their initial, experimentally determined structure files. The “Experiment_Details.txt” file stores all of the information the user provided about how to run the IPRO Suite of Programs experiment. It is intended to be both human and machine readable, and the start of this file is shown in Figure 4.6. The “NAME_Summary.txt” file, where NAME is the name of the experiment (i.e. Demonstration), keeps a record of the calculations carried out during the experiment, including when they begin, how long they take, and what was done in each iteration of calculation. Finally, the “IPRO.py” program is what actually carries out the calculations of this IPRO experiment.

**IPRO Iterations**

The core of the IPRO Suite of Programs is IPRO, a method to redesign proteins to have improved properties (e.g. binding, catalysis, etc.). The original implementation of IPRO suggested stabilizing mutations to a combinatorial library of recombined proteins, with the limiting case being the redesign of a single protein [2]. However, since then IPRO has only been used to redesign single proteins and the functionality to redesign a combinatorial library has been discarded.

The six steps of an iteration of IPRO are described in Figure 4.7. First, a randomly selected set of contiguous residues is selected in a DM and the backbone phi and psi dihedral angles are randomly perturbed. Then residues in and around the perturbed region
are repacked and mutated using a rotamer library [12] and MILP optimization. The third step is a local, rigid-body redocking of the TMs. This is followed by a global energy minimization and energy calculation. The calculated energy between the DMs and TMs is used in the Metropolis criterion to determine whether or not to retain the new DMs as the starting point for the next iteration. By utilizing these six steps over the course of thousands of iterations, IPRO can successfully identify combinations of backbone and amino acid changes that give improved properties for the system.

**Backbone Perturbation**

In the first step of IPRO, a random backbone perturbation to a DM is carried out. This begins by randomly selecting a DP. Next, the total size (one to five residues) of the perturbed region and the position of the DP in this contiguous stretch of residues are randomly chosen. For each of the perturbed residues, the phi and psi dihedral angles are randomly perturbed using a Gaussian distribution centered at zero degrees with a standard deviation of 1.5°. No modification greater in magnitude than five degrees is permitted. Five residues on each side of the perturbed region are free to move during the perturbation to prevent the dihedral angle changes from causing long-range structural effects. These 11-15 residues are mutated to glycine and an energy minimization with strong restraints on the perturbed dihedral angles is carried out to make the perturbed DM structure. Residues outside of the 11-15 residues are not permitted to move during the perturbation. Each DG has the same random perturbation applied to it.

**Optimal Rotamer Selection**

The second step of IPRO is the use of a rotamer library and MILP optimization to repack the amino acid side chains in and around the perturbed region. The 11-15 residues that
were mutated to glycine during the backbone perturbation are always repacked. The user has the option of also repacking residues that are spatially close to the perturbed region. The default definition of close is having at least one heavy atom (i.e. not hydrogen) within 4.5 Å of a heavy atom in a perturbed residue [13]. Only DPs within the perturbed region are permitted to mutate. By default, all DPs can mutate to any amino acid, but the user may specify the permitted kinds of amino acid mutations for each DP.

The rotamer library does not contain proline rotamers. Therefore a proline rotamer was created to include the option of mutating to proline during IPRO Suite of Programs experiments. 1000 proline residues were collected from a database of antibodies [14] and the proline with the smallest average heavy-atom RMSD with the other residues was selected to serve as the rotamer. This was residue 14 in molecule chain B of PDB: 1CFQ [15]. This rotamer is only used when it is sterically permitted, as determined by an analysis of literature data [16], or when a residue is already a proline and is not mutating. The rotamer – non-rotamer (i.e. the TMs and all parts of the DMs that are not being replaced by rotamers) and rotamer – rotamer energies are calculated using the pairwise additive, non-bonded energy terms from the force field (i.e. VDW, electrostatics, and implicit solvation). The rotamer library has statistical terms that the user may choose to include in the rotamer – non-rotamer energies, but by default they are excluded. When selecting the rotamers, the rotamer – non-rotamer energies are calculated first. Next, the total number of rotamers is reduced below a threshold $U$ (default is 1400) to make sure the rotamer selection MILP can solve in a reasonable amount of time and not exceed the memory limits of the processor (four gigabytes for the C. Maranas Lab). This is done by retaining the $U$ lowest energy rotamers (default is three) of each permitted
kind of amino acid at each position, as well as all rotamers that are within an energy
cutoff \( U \) (default is 40 kcal/mol) of the lowest energy rotamer at each position. In the
event that the total number of rotamers exceeds the maximum rotamer number threshold
after applying these limitations, the minimum number of rotamers of each kind of amino
acid and the energy cutoff are reduced until the total number of rotamers is below the
threshold. Once the number of rotamers has been reduced to an acceptable number, the
rotamer – rotamer energies are calculated.

Once all of the rotamer – non-rotamer and rotamer – rotamer energies have been
calculated, MILP optimization can be used to select the minimum energy arrangement of
rotamers. This requires the definition of the following sets: the residues receiving
rotamers, \( i,j = 1,\ldots,N \), and the rotamers at position \( i, r,s = 1,\ldots,R_i \). The energy of rotamer
\( r \) at position \( i \) with the non-rotamer portions of the DG is stored in parameter \( EC_{i,r} \) and
the energy between rotamer \( r \) at position \( i \) and rotamer \( s \) at position \( j \) is stored in
parameter \( ER_{i,r}^{j,s} \). Binary variable \( y_{i,r} \) is equal to one if rotamer \( r \) is selected at position \( i \)
and zero otherwise. Similarly, binary variable \( w_{i,r}^{j,s} \) is equal to one if rotamer \( r \) at position
\( i \) and rotamer \( s \) at position \( j \) are selected simultaneously and zero otherwise. Using these
sets, parameters, and variables, the optimal combination of rotamers can be selected by
minimizing:

\[
\sum_{i=1}^{N} \sum_{r=1}^{R_i} y_{i,r} EC_{i,r} + \sum_{i=1}^{N-1} \sum_{r=1}^{R_i} \sum_{j=i+1}^{N} \sum_{s=1}^{R_j} w_{i,r}^{j,s} ER_{i,r}^{j,s}
\]  

(1)

Equation 1 is the objective function of the MILP, and minimizes the net energy of the
selected rotamers with each other and with the non-rotamer portions of the DG. This
equation is subject to:

\[
\sum_{r=1}^{R_i} y_{i,r} = 1 \quad \forall i \mid 1, \ldots, N
\]  

(2)
Equation 2 ensures that exactly one rotamer is selected at each position.

\[ w_{l,r}^{i,s} = y_{l,r} y_{j,s} \quad \forall i | 1, ..., N - 1, \forall r | 1, ..., R_i, \forall j | i + 1, ..., N, \forall s | 1, ..., R_j \] (3)

Equation 3 makes sure that \( w_{l,r}^{i,s} \) only assumes a value of one if both \( y_{l,r} \) and \( y_{j,s} \) have values of one. However, Equation 3 is non-linear, which would substantially increase the time it takes for the optimization formulation to solve. It can be linearized using:

\[ w_{l,r}^{i,s} = \sum_{s=1}^{R_j} y_{j,s} \quad \forall i | 1, ..., N - 1, \forall r | 1, ..., R_i, \forall j | i + 1, ..., N \] (4)

\[ w_{l,r}^{i,s} = \sum_{r=1}^{R_i} y_{l,r} \quad \forall i | 1, ..., N - 1, \forall j | i + 1, ..., N, \forall s | 1, ..., R_j \] (5)

Equations 1, 2, 4, and 5 form the optimal rotamer selection MILP and can be solved to optimality using the CPLEX solver [17]. The IPRO Suite of Programs now supports calling the CPLEX solver directly from Python, as well as the previously supported method, GAMS [5]. The optimal combinations of rotamers are selected sequentially for each DG, with all DGs after the first being constrained to have a sequence that matches that of the first. Therefore it is important for the user to select the “most important” DG as the first in an experiment.

**Target Molecule Redocking**

The third step of an IPRO iteration is a local, rigid-body redocking of the TMs. Because docking may take a long time for some systems, this step is only carried out every \( U \) iterations (default is three). Docking uses the same pairwise additive energy functions used during the rotamer selection step. During docking, a TM is randomly perturbed along and around the X, Y, and Z axes. The perturbations are generated using a Gaussian distribution centered at zero, with user-defined standard deviations (defaults of 0.2 Å and 2.0°). After the TM is perturbed, the net interaction energies of all of the TMs with each other and with the DMs is evaluated and the movement of the TM is kept using simulated
annealing if the net interaction energy has improved. Each TM is sequentially randomly
perturbed during an iteration of docking, and $U$ iterations are carried out (default is 500).
A constant cooling schedule is used for simulated annealing and the user may set the
temperatures at the start and end of docking (the defaults retain 25% of positions within
10 kcal/mol of the best at the start of docking and 10% of positions within 10 kcal/mol of
the best at the end of docking).

**Iteration Evaluation**

The fourth and fifth steps of an IPRO iteration are a global energy minimization and
interaction energy calculation, where interaction energy is the energy of the minimized
DG minus the energy of only the DMs from the minimized DG minus the energy of only
the TMs from the minimized DG. Finally, the calculated interaction energy is used in the
Metropolis criterion to determine whether or not to retain the results of the IPRO
iteration. The user may set the temperature used to make the decision (default is to retain
25% of designs within 10 kcal/mol of the best design). All DGs must pass the selection
criteria for the results of the iteration to be retained.

**Ensemble Structure Refinements**

A favorable result of an IPRO design iteration is a specific backbone structure and amino
acid sequence that has been evaluated at a local energy minimum and found to have
improved properties relative to the experiment’s objectives. This energy minimum is
unlikely to be the global energy minimum and it is possible that the calculated properties
observed for it are an artifact that is not representative of the true properties of the
system. To address this issue, the IPRO Suite of Programs now includes optional
ensemble structure refinements. These refinements are carried out by generating an
ensemble of $U$ (default is ten) structures, each of which is generated from $U$ “refinement” iterations (default is 25) of IPRO. During refinement iterations, the objective for every DG is to minimize the complex energy and the normal objectives of DGs are not considered. In addition, no mutations are permitted during refinement iterations. In essence, refinement iterations are searching for improved local energy minima for a specific DM amino acid sequence. After all refinement iterations have been completed, the average properties of the refinement ensemble are evaluated to determine whether or not a particular design is actually the best identified so far. This ensemble approach to evaluating structures has given a high correlation ($R^2 = 0.960$) between calculated interaction energies and experimentally measured binding affinities in previous work [4]. If structure refinements are being done, there is the option of using binding energy instead of interaction energy to evaluate DGs. This is done by generating an additional DG during structure refinements that contains only the DMs. Analysis during the development of the IPRO Suite of Programs revealed that allowing binding energy calculations without structure refinements resulted in highly inconsistent results between different repetitions of the same experiment.

**Parallelization**

Due to certain system incompatibilities of the C. Maranas Lab, it was impossible to write the IPRO Suite of Programs code so that it can be parallelized. Therefore the IPRO Suite of Programs was written in such a way that processors running separately from one another can work on the same IPRO Suite experiment. An independent processor carries out each iteration of IPRO. The results of an iteration are shared between processors by writing the information to files, something that must be done anyway so that the user may
have access to it. If a processor finds a result that is the best so far in an experiment, that information is always shared between processors. The user has the option of saying whether or not results retained by the Metropolis criterion but are not the best observed so far should be shared between processors or if each processor should pursue an independent trajectory until a best result is found. If a processor completes an iteration and determines that an ensemble structure refinement is necessary, a system flag is created that other processors will identify, causing them to delete their ongoing iterations and participate in the ensemble structure refinement.

**OptGraft Binding Pocket Selection MILP**

OptGraft inserts a binding site from one protein into a novel host and then uses IPRO to suggest mutations that improve the geometry of the binding pocket. The selection of the host residues for the binding pocket in the novel protein is done using a MILP optimization formulation that has been modified from its original implementation [3] to solve much faster. The MILP requires the definition of sets defining the residues in the novel host protein, \( i,j = 1,\ldots,N \), and the residues of the binding pocket in the original protein, \( k,l = 1,\ldots,K \). The MILP makes selections based on Ca – Ca and Cβ – Cβ distances. The atom – atom distances in the novel host protein are stored in parameters \( RA_{i,j} \) and \( RB_{i,j} \), respectively. The distances between the atoms in the binding site residues are stored in parameters \( ra_{k,l} \) and \( rb_{k,l} \). Binary variable \( Y_{i,k} \) is equal to one if residue \( i \) in the novel host protein is selected as the equivalent to residue \( k \) in the binding site and zero otherwise. Similarly, binary variable \( W_{i,l}^{j,k} \) is equal to one if residue \( i \) is selected as the equivalent to residue \( k \) and residue \( j \) is simultaneously selected as the equivalent of residue \( l \) and is zero otherwise. The MILP can then be written to minimize:
\[ \sum_{k=1}^{K-1} \sum_{i=1}^{N} \sum_{l=i+1}^{K} \sum_{j=1}^{N} \left[ (RA_{i,j} - ra_{k,l})^2 + (RB_{i,j} - rb_{k,l})^2 \right] W_{i,k}^{j,l} \] (6)

Equation 6 minimizes the sum of the squared differences in distances between the \( \alpha \) and \( \beta \) carbons in the selected binding site and in the original binding site, thus maintaining the original geometry. This equation is subject to:

\[ \sum_{i=1}^{N} Y_{i,k} = 1 \quad \forall k \in \{1, ..., K\} \] (7)

This equation makes sure that every residue in the binding site is assigned to exactly one residue in the novel host protein.

\[ \sum_{k=1}^{K} Y_{i,k} \leq 1 \quad \forall i \in \{1, ..., N\} \] (8)

Equation 8 makes certain that no novel host residue is assigned as more than one residue from the binding site.

\[ W_{i,k}^{j,l} = Y_{i,k} Y_{j,l} \quad \forall i \in \{1, ..., N\}, \forall j \in \{1, ..., N\}, \forall k \in \{1, ..., K-1\}, \forall l \in \{k+1, ..., K\} \] (9)

This establishes the relationship between \( W_{i,k}^{j,l} \), \( Y_{i,k} \), and \( Y_{j,l} \), which is nonlinear and must be linearized. The modification in this formulation is by linearizing Equation 9 in a more efficient manner. This starts by summing over the \( i \) positions in the novel host protein on both the left and right hand sides of Equation 9.

\[ \sum_{i=1}^{N} W_{i,k}^{j,l} = \sum_{i=1}^{N} Y_{i,k} Y_{j,l} \quad \forall j \in \{1, ..., N\}, \forall k \in \{1, ..., K-1\}, \forall l \in \{k+1, ..., K\} \] (10)

\( Y_{j,l} \) is independent of the index \( i \) being summed over in Equation 10 and can be moved outside of the summation. Combining this with Equation 7 gives:

\[ \sum_{i=1}^{N} W_{i,k}^{j,l} = Y_{j,l} \quad \forall j \in \{1, ..., N\}, \forall k \in \{1, ..., K-1\}, \forall l \in \{k+1, ..., K\} \] (11)

\[ \sum_{j=1}^{N} W_{i,k}^{j,l} = Y_{i,k} \quad \forall i \in \{1, ..., N\}, \forall k \in \{1, ..., K-1\}, \forall l \in \{k+1, ..., K\} \] (12)

Equation 12 results from the same analysis as Equation 11 when a summation over \( j \) is used instead. Together, Equations 11 and 12 linearize Equation 9. Equations 6-8, 11, and
12 form the binding site selection MILP and can be solved to optimality by calling the CPLEX [17] solver directly from Python or by using GAMS [5].

**OptCDR Canonical Structure Database**

The set of antibody structures examined to make the MAPs [14] database had their CDRs collected and clustered to generate a new database of canonical structures that follow IMGT® unique numbering [18-21]. The clustering was carried out as described in Chapter 2. It was conducted such that at the end of clustering all structures within a cluster had a backbone atom (N, Cα, C) RMSD of no more than 1.5 Å from the model structure of the cluster, where the model is the structure with the smallest average backbone atom RMSD with all other members of the cluster. All clusters for the H3 CDR were retained and clusters with at least three members were retained for CDRs H1, H2, L1, L2, and L3. Several *a posteriori* modifications were made to reduce clashing among the canonical structure models. The old and new canonical structure databases are listed in Table 4.1. Of note is the significant increase in the number of L3 canonical structures. This is consistent with the expectation of structural diversity in this CDR due to the junction of a Variable and Joining gene introduced during V-(D)-J recombination. It is unclear why the original OptCDR analysis did not identify more L3 canonical structures.

**IPRO**

IPRO redesigns proteins to have modified binding interactions with one or more substrates. To run an IPRO experiment, the user must provide all of the information needed to run the experiment: the user’s name, the name of the experiment, the DMs and TMs, how to use the force field, how to use implicit solvation, the DPs, the kinds of amino acids mutations permitted for each DP, the DGs, how to run docking, how to
repack rotamers, the temperature for the Metropolis criterion, whether or not to share
results retained by the Metropolis criterion between processors, the number of design
iterations to run, and whether or not and how to do ensemble structure refinements. In
addition, the IPRO Suite of Programs permits four kinds of restraints on atom positions
that are rigorously enforced throughout all functions: atoms may be permanently fixed in
place, atoms may be kept near their positions in their experimentally determined
structures, the distance between two atoms in the same DG may be restrained, and the
dihedral angle between four atoms in the same DG may be restrained.

Once the information has been provided and Start_Experiment.py has created the folder
to run the experiment in, the calculations are begun. They start with an energy
minimization of the initial DGs, the replacement of the side chains of all DM amino acids
with the most similar rotamer from the rotamer library, and a second energy minimization
of the DGs. This provides an initial starting point for the experiment. If ensemble
structure refinements are being used, $2U$ refinement iterations are carried out to further
refine this initial structure. Once the initial structure is completed, the $U$ design iterations
are run, being interrupted as necessary by ensemble structure refinements whenever a
best result is found. Over the course of thousands of iterations, IPRO can identify
combinations of structure perturbations and sequence mutations that provide
improvements in the interactions between the DMs and TMs.

**OptGraft**

OptGraft inserts a binding site from one protein into a novel host and then uses IPRO to
suggest mutations that improve the geometry of the binding pocket. To run an OptGraft
experiment, a user must provide the same information as for an IPRO experiment, with
the added necessity of identifying the binding site in one molecule and the particular DM it is being added to. In addition, residues in the novel host DM that are close to the new binding site, defined in the same way as used during rotamer repacking, are automatically selected as DPs. The user may exclude some or all of these DPs and select others if they so desire. Also, strong distance restraints on the $\alpha$ and $\beta$ carbon distances of the binding site residues are used throughout the OptGraft experiment to make sure the generated design has the appropriate geometry. An OptGraft experiment begins by inserting the novel binding site into the designated DM and then proceeds in the same manner as an IPRO experiment.

**OptCDR**

The workflow of OptCDR is described in Chapter 2. To run an OptCDR experiment, the user must provide most of the same information as for an IPRO experiment. However, there is no specification of DMs or DPs as the antibody is designed from scratch and every residue in the CDRs may mutate. In addition, the user must specify whether to design a VH, a VL, or both, $U$ random initial placements of the antigen during the first step of OptCDR, and $U$ unique antibody CDR designs to generate. OptCDR also now has the ability to include FR residues during calculations if the user so desires. These FR residues are left in their experimental configuration throughout OptCDR.

An OptCDR experiment begins by generating $U$ random initial placements for the antigen and identifying the best combination of canonical structures for each. The best positions are retained for independent library designs, as specified by the user. For each combination of canonical structures, the amino acids of the canonical structures are initialized from glycine one CDR at a time (due to memory limitations for the MILP). As
described in Chapter 2, restraints are imposed on what amino acids are allowed at each position in a CDR based on experimental observations of the canonical structure. In addition, the maximum percent usage of each amino acid and all charged amino acids is limited, based on the maximum percent usage observed in each CDR, using integer cuts in the rotamer selection MILP. Each CDR is initialized, the docking function of IPRO is run, and then the amino acids of each CDR are reassigned to create an initial antibody. Once this structure is generated, IPRO is used to computationally affinity mature the designed CDRs. During an iteration of IPRO in OptCDR, a CDR is randomly selected and the entire backbone is randomly perturbed, instead of only perturbing a portion of it. Also, the sequence restraints already mentioned for initializing the CDR sequences are enforced in IPRO. Finally, the rotamer selection MILP is iteratively solved, using integer cuts to eliminate previous solutions, to generate a library of mutants to the affinity matured antibody design.

**MAPs**

The MAPs database is integrated into the IPRO Suite of Programs to predict antibody structures from their sequences, as described in the original publication [14]. Predicting an antibody’s structure is the IPRO Suite of Programs experiment that is most simple to run. A user must provide their name, the type of experiment (i.e. MAPs), the name of a folder to store the experiment’s results in, and a heavy variable domain sequence and / or a light variable domain sequence. The IPRO Suite automatically assigns germline genes to the antibody sequences, collects the most similar MAPs pieces and mutates them to the germline gene sequence, mutates this antibody structure to the provided sequences, and runs an energy minimization to generate the final predicted antibody structure.
**OptZyme**

OptZyme [4] uses ground state substrates and transition state analogues in IPRO to predict mutations that give better catalytic properties. Running an OptZyme experiment is identical to running an IPRO experiment, except the user must specify which TMs are the ground state substrates and which are the transition state analogues. If multiple catalytic reactions are being simultaneously considered (e.g. suppressing catalysis or binding of a competitor), the user must specify which ground state substrates and which transition state analogues are paired. This information is used to automatically generate DGs with the appropriate objectives.

**Mutator**

Often, it is desirable to computationally analyze specific mutants of a protein for properties, rather than try to design for those properties. To facilitate this, a Mutator program was created that uses the ensemble structure refinement function of the IPRO Suite of Programs. Running a Mutator experiment is identical to running an IPRO experiment, except instead of identifying DPs the user must identify the amino acid changes of the mutants. Multiple mutants to the same system can be analyzed in a single experiment. The residues that are being mutated in any mutant are treated as pseudo-DPs so that IPRO iterations can select regions of structure to perturb. A Mutator experiment begins by doing a structure refinement of the wildtype system of molecules. This refined complex is used as the starting point for each mutant. The rotamer selection MILP is used to generate an initial version of the mutant and then an ensemble structure refinement is used to evaluate its properties.
Discussion

The IPRO Suite of Programs has been created as a single tool and source of code to integrate the many C. Maranas Lab computational protein engineering methods. All of the programs described here are run through the same user interface, the Start_Experiment.py program. This program asks the user for all of the information that is needed to run an experiment and validates that the provided information is acceptable. The user must only answer yes and no questions and provide specifications for exactly what they want to do in an experiment and the program takes care of setting up the experiment and identifying any problems.

The IPRO Suite of Programs has also been modified to provide enhanced documentation. Every experiment is run in a uniquely named folder and the detailed instructions for running the experiment are stored in an “Experiment_Details.txt” file. In addition, a summary of every experiment is created as the experiment’s calculations are carried out and all result structures are automatically saved with the time and date they were generated. Therefore it is always possible to review this automatically generated data to determine exactly what was done, even years later.

All programs in the IPRO Suite now use the same core of code so that there are no discrepancies between different methods. The code is also written in a modular, object-oriented fashion so that it is easy to add additional features to the IPRO Suite of Programs without causing compatibility issues. In this way, we intend to add new features to the IPRO Suite of Programs (e.g. supporting new force fields) as well as adding new computational protein engineering methods to this common Suite of Programs as they are developed.
The various programs in the IPRO Suite have been successfully used for a variety of protein engineering applications in the last decade. Now they have been integrated together, improved for faster calculations (e.g. OptGraft), and modified to give improved confidence in the results (e.g. ensemble structure refinements). There is a new, single user interface that is intended to be easy to learn to use for someone without extensive programming or protein engineering experience, opening up the IPRO to novice users.
Figure 4.1: The Format of the IPRO Suite of Programs. This is the contents of the IPRO Suite of Programs folder. It contains seven directories: databases contain the MAPs database and the OptCDR canonical structures, experiments is where individual experiments are carried out, input_files contains files that are needed as inputs (e.g. topology and parameter files for CHARMM), modules contains the Python and C++ files needed to run various programs, programs contains the Python files that run particular types of IPRO Suite experiments, references has files of constant information (e.g. how an antibody should be positioned), and structures stores the files of molecules. In addition, there is one program: Start_Experiment.py is the user interface for beginning any IPRO Suite experiment.
<table>
<thead>
<tr>
<th>Welcome to the Iterative Protein Redesign &amp; Optimization (IPRO) Suite of Programs from the Costas Maranas Laboratory in the Chemical Engineering Department of the Pennsylvania State University.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please make sure you have read the provided documentation. That is where the methodologies, capabilities, and terminology of the IPRO Suite are defined and explained. It is assumed you have a working knowledge of this information.</td>
</tr>
<tr>
<td>To begin, please provide some basic information about your experiment.</td>
</tr>
<tr>
<td>It appears you are rjp5003. Is this correct?</td>
</tr>
<tr>
<td>yes</td>
</tr>
<tr>
<td>What type of IPRO Suite Experiment is this?</td>
</tr>
<tr>
<td>IPRO</td>
</tr>
<tr>
<td>Are you sure this is correct?</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>What would you like to name this IPRO Experiment?</td>
</tr>
<tr>
<td>Demonstration</td>
</tr>
<tr>
<td>Are you sure this is correct?</td>
</tr>
<tr>
<td>Y</td>
</tr>
</tbody>
</table>

**Figure 4.2: Starting an IPRO Suite of Programs Experiment.** All IPRO Suite of Programs experiments are created and run by the “Start_Experiment.py” program in the main IPRO Suite folder. This program starts by identifying who is running the experiment and confirming that information is correct, asking what type of IPRO Suite program is being run, and asking for a unique name of the experiment. Overall, the “Start_Experiment.py” program asks the user for the information needed to run a particular type of IPRO Suite of Programs experiment, validates that the information is acceptable and will work, and starts the calculations of the experiment.
Figure 4.3: Example of IPRO Terminology. Here we have an antibody – IGF-2 complex [PDB: 3KR3] that is hypothetically being designed to have improved binding to IGF-2 while preventing binding to IGF-1. The VH and VL of the antibody are Design Molecules and IGF-1 and IGF-2 are Target Molecules in this example. Binding to IGF-1 and IGF-2 should be evaluated separately, so there are two Design Groups. Each Design Group contains all of the Design Molecules and has a separate binding objective.
Figure 4.4: The STANDARDS.py File. Located in the IPRO Suite modules folder, STANDARDS.py is the location of the default settings of the IPRO Suite of Programs. Here the beginning of that file is shown. The Start_Experiment.py program accesses this information to provide the user with easy answers to its many questions. When the IPRO Suite is installed, users should modify the contents of this file to match their expected default settings. For example, the defaultUser is “rjp5003”, which is the Pennsylvania State University user ID of Robert J. Pantazes and should be changed.
Figure 4.5: The Contents of an IPRO Experiment’s Folder. These are the contents of the folder for the “Demonstration” experiment created in Figure 4.2. It contains three folders, two text files, and one Python program. The “input_files” and “structures” folders respectively store the files and structures needed for running the experiment, while the “results” folder stores the structures generated by the experiment. Since this is an IPRO experiment, those are the each of the best structures identified by the experiment, stored by the iteration they were generated on. The Python program “IPRO.py” is what will actually run the IPRO experiment. “Experiment_Details.txt” stores all of the information the user provided about how the experiment should be run and “Demonstration_Summary.txt” contains a summary of the calculations carried out during the experiment.
<table>
<thead>
<tr>
<th>Basic Experiment Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>User: rjp5003</td>
</tr>
<tr>
<td>Type: IPRO</td>
</tr>
<tr>
<td>Name: Demonstration</td>
</tr>
<tr>
<td>File Format: PDB</td>
</tr>
<tr>
<td>Force Field: CHARMM</td>
</tr>
<tr>
<td>Folder: /gpfs/work/r/rjp5003/IPRO_Suite/experiments/Demonstration/</td>
</tr>
</tbody>
</table>

**How to run IPRO**
- IPRO Iterations: 1000
- IPRO Annealing Temperature: 3640.000
- Annealing Sharing: yes
- Energy Calculation: Interaction

**How to use CHARMM**
- CHARMM Topology File: top_all27_prot_na.rtf
- CHARMM Parameter File: par_all27_prot_na.prm
- CHARMM Energy Term: angl
- CHARMM Energy Term: bond
- CHARMM Energy Term: dihe
- CHARMM Energy Term: elec
- CHARMM Energy Term: impr
- CHARMM Energy Term: urey
- CHARMM Energy Term: vdw
- CHARMM Iterations: 5000

**How to use Implicit Solvation**
- Use Solvation: no

**Figure 4.6: The Contents of an Experiment_Details.txt File.** This is the start of the contents of the “Experiment_Details.txt” file corresponding to the experiment begun in Figure 4.2 and shown in Figure 4.5. The “Experiment_Details.txt” file contains all of the information the user provides for how to run a particular IPRO Suite of Programs experiment. It is intended to be both a human-readable record of how the experiment was conducted as well as instructions to the IPRO Suite of Programs on how to carry out the experiment.
Figure 4.7: IPRO Iteration Steps. An iteration of IPRO is composed of 6 steps. First, a set of contiguous residues is randomly selected and their phi and psi dihedral angles are randomly perturbed. Next, a rotamer library, energy functions, and MILP optimization is used to repack residues in and around the perturbed region. This is followed by a local, rigid-body TM redocking. Finally, an energy minimization and energy calculation are carried out. The calculated energy is used in the Metropolis criterion to determine whether or not to retain the results of the iteration or discard them.
Table 4.1: The OptCDR Canonical Structures. Here are the number of amino acids and number of canonical structures for each CDR in the original OptCDR publication and updated to use IMGT® unique numbering. The definition of each CDR is different in the two cases and approximately 300 additional antibodies were analyzed for the new dataset. However, it is interesting to note that the number of L3 CDRs significantly increased. This is consistent with the expected structural diversity introduced by the gene junction between the Variable and Joining genes in this CDR during V-(D)-J recombination.

| CDR | Chapter 2 | | Chapter 4 | | |
|-----|-----------|-----------|-----------|-----------|
|     | Amino Acids | Canonical Structures | Amino Acids | Canonical Structures |
| H1  | 10-12      | 8          | 8-10       | 9          |
| H2  | 9-12       | 4          | 7-10       | 5          |
| H3  | 3-22       | 124        | 5-24       | 135        |
| L1  | 10-17      | 13         | 5-12       | 13         |
| L2  | 7          | 1          | 3          | 1          |
| L3  | 8-11       | 9          | 7-13       | 16         |
References


5. GAMS Home Page [http://www.gams.com]


Crystallographic analysis of anti-p24 (HIV-1) monoclonal antibody cross-

16. Lovell SC, Davis IW, Arendall WB, 3rd, de Bakker PI, Word JM, Prisant MG, 
Richardson JS, Richardson DC: Structure validation by Calpha geometry: 

2007.

18. Ehrenmann F, Giudicelli V, Duroux P, Lefranc MP: IMGT/Collier de Perles: 
IMGT standardized representation of domains (IG, TR, and IgSF variable 
and constant domains, MH and MhSF groove domains). Cold Spring Harbor 

19. Lefranc MP: IMGT unique numbering for the variable (V), constant (C), and 
groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harbor 

20. Lefranc MP: IMGT Collier de Perles for the variable (V), constant (C), and 
groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harbor 

M, Da Piedade I, Rouard M et al: IMGT unique numbering for 
immunoglobulin and T cell receptor constant domains and Ig superfamily C-
Chapter 5. Summary and Future Work
Summary

There are numerous examples of successful computational redesigns of proteins, but the de novo design of proteins remains a formidable and difficult-to-achieve task. It was the objective of this dissertation to use antibodies as a model protein system to learn about de novo protein design principles. Antibodies were selected because they have many modular structural features that aid in modeling their atomic coordinates and because their functions are primarily limited to binding, not catalysis. In short, antibodies are one of the easiest, albeit still difficult, family of proteins to design and therefore are an excellent choice on which to practice de novo design.

OptCDR

In Chapter 2, the Optimal Complementarity Determining Regions method to design antibody CDRs specific to any specified antigen epitope was developed. OptCDR is a four-step workflow. First, the combination of canonical structures that is most likely to be able to favorably bind the antigen is selected. This is done using a database of canonical structures generated from an analysis of over 800 antibody structures, a geometric criterion to describe which structures best fit around the antigen, and a MILP optimization formulation. Once the initial canonical structures are selected, their amino acid sequences are then iteratively identified using a rotamer library, pairwise additive energy functions, and a rotamer selection MILP. Care is taken to ensure that the CDR amino acid compositions and sequences are consistent with naturally observed trends. In the third step of OptCDR, iterations of IPRO are used to find CDR backbone structure changes and amino acid mutations that provide better binding to the target antigen. Finally, OptCDR is completed by generating an antibody library of the most promising
CDR mutants for antigen binding. By utilizing these four steps, OptCDR can design libraries of antibody CDRs specific to any specified epitope.

Three test cases were selected for OptCDR: the hapten fluorescein, a peptide from the capsid of hepatitis C, and the protein VEGF. In all three cases, OptCDR was able to generate antibody CDRs that appear computationally promising. The designs had the expected types of geometries based on the size of the antigen, had many polar contacts, and had calculated interaction energies that were comparable to or better than high-affinity, naturally occurring antibodies for the antigens. These promising computational features suggest that OptCDR is capable of the de novo design of the portion of an antibody’s structure that is most important to antigen binding.

**MAPs**

The next development beyond OptCDR is the de novo design of entire antibody variable domains. It is our belief that developing a computational analogue to the immune response is the most promising pathway to variable domain design. In the immune system, a naïve antibody is created through V-(D)-J recombination, where randomly selected V, D, and J germline genes are recombined to make an antibody variable domain. Therefore, in Chapter 3 a database of Modular Antibody Parts analogous to the human germline genes was created.

An analysis of over 1100 antibody structures revealed that it is impossible to predict the backbone structure of the amino acids corresponding to a D gene, primarily because of the nucleotide insertions and deletions that occur between genes during V-(D)-J recombination. Therefore an alternative definition of antibody structure was developed: V* (most of the V gene), CDR3, and J* (most of the J gene). This alternative
concatenation of antibody structures provided a one-to-one correlation between predicted germline gene sequences and affinity matured antibody structures. Using the V*, CDR3, and J* definition of antibody structure, the MAPs database of 929 antibody “parts” based on the human germline genes was generated. A test of 260 high-resolution antibody structures not used to make any of the MAPs models revealed that the MAPs database can be used to predict antibody tertiary structures with an average all-atom RMSD of 1.900 Å.

The IPRO Suite of Programs

Experiences with OptCDR, MAPs, and other protein design projects outside the scope of this dissertation strongly suggested that the computational methods being used needed significant improvement. The various programs required expert-level knowledge to use and, despite utilizing similar methods, were often incompatible with one another. In Chapter 4, an integrated Iterative Protein Redesign & Optimization Suite of Programs was developed. Created from a common core of object-oriented code and with a new, simpler user interface, the IPRO Suite of Programs is a single tool to address a wide variety of protein redesign problems.

The core of the IPRO Suite of Programs is IPRO, a six-step workflow to find backbone structure perturbations and amino acid mutations that lead to improved protein properties. In the first step of IPRO, the dihedral angles of a randomly selected portion of the protein’s backbone are perturbed. The amino acids in and around this perturbed region are then repacked using a rotamer library, pairwise additive energy functions, and a MILP optimization formulation. This is followed by a rigid-body, local redocking of the target molecule(s), a global energy minimization, and an interaction energy calculation.
Finally, the Metropolis criterion is used to determine whether or not to retain the results of an iteration. Over the course of thousands of iterations, IPRO can identify structure and sequence changes that lead to improved properties.

To help improve confidence in the predicted designs, an ensemble structure refinement was added to IPRO, which is started whenever an improved design is identified by an iteration. Independent copies of the complex of molecules are generated and run through tens of iterations of IPRO. During these iterations no mutations are allowed and structures are retained or discarded based on whether or not the global energy of the system of molecules has been improved. The ensemble average interaction energy can then be used to provide a better estimate of the true properties of the mutated protein.

The IPRO Suite of Programs includes six unique protein engineering methods. IPRO redesigns an existing protein to have modified interactions with a substrate. OptCDR, as presented in Chapter 2, designs unique antibody CDRs against any specified antigen epitope. The MAPs database can be used to predict antibody tertiary structures. OptGraft inserts binding sites from one protein into another. OptZyme uses ground state structures and transition state analogues to improve the catalytic properties of enzymes. Finally, Mutator makes and evaluates specific mutants of a protein.

**OptMAVEn**

OptCDR, the MAPs database, and the IPRO Suite of Programs have lain the foundation for the development of an Optimal Method of Antibody Variable region Engineering (OptMAVEn) for the *de novo* design of entire, fully human antibody variable domains against any specified epitope. OptMAVEn is envisioned as a two-step workflow analogous to the human immune response. In the first step, the combination of MAPs
structures that is best able to bind the specified epitope is selected. This is equivalent to a B cell producing a naïve antibody through V-(D)-J recombination and the antibody binding to a pathogen. In the immune response, this is followed by affinity maturation where beneficial mutations are sequentially identified. The second step of OptMAVEN is the use of IPRO to progressively identify amino acid sequence changes that yield improved antigen binding, with additional constraints to ensure the antibody becomes or remains fully human.

**MAPs Structure Selection MILP**

In the first step of OptMAVEN, the combination of MAPs structures that best binds a specified antigen’s epitope is identified using an MILP. For a specific antigen position, the interaction energies between the antigen and the MAPs structures can be calculated using pairwise energy functions, as in the rotamer selection step of an IPRO iteration (see Chapter 4). The one modification is that the VDW radii of the atoms are reduced by 50% to avoid penalizing minor steric clashes that could be easily fixed by an energy minimization.

The problem of selecting the optimal combination of MAPs structures can then be mathematically represented using a MILP formulation. This requires the definition of the index sets $V = \{ v \mid VH, V$-KAPPA, V-LAMBDA $\}$, denoting the three types of antibody variable domains, and $R = \{ r \mid V^*,$ CDR3, $J^* \}$, denoting the three types of MAPs antibody structures. Also required are the sets $S_{v,r} = \{ s \mid 1, \ldots, S_{v,r} \}$, encoding the specific MAPs structures for each variable domain and region and set $IC^{\text{clash}}$, containing all pairwise MAPs structure combinations $(v1, r1, s1)$ and $(v2, r2, s2)$ that sterically clash with one another. Parameter $IE_{v,r,s}$ encodes the calculated energy between structure $S_{v,r}$
and the antigen. Switching parameters $H$ and $L$ have values of one if a VH and VL, respectively, are being designed and zero if they are not. This allows the user to possibly design only a VH or VL, as would be appropriate for a nanobody. Binary variable $y_{v,r,s}$ encodes whether or not structure $S^{v,r}$ has been selected (i.e. a value of one) or not (i.e. a value of zero).

The MILP formulation can then be written to minimize the energy between the selected parts and the antigen:

$$\sum_{v=1}^{V} \sum_{r=1}^{R} \sum_{s=1}^{S} y_{v,r,s} E_{v,r,s}$$  

This equation is subject to:

$$\sum_{s=1}^{S} y_{vH,v*,s} = H$$  

$$\sum_{s=1}^{S} y_{v-KAPPA,v*,s} + \sum_{s=1}^{S} y_{v-LAMBDAY,v*,s} = L$$

Equations 2 and 3 control whether or not a VH or VL, respectively, is designed. They ensure that exactly one V* structure is selected if a VH or VL should be designed or none if not.

$$\sum_{r=CDR3}^{J*} \sum_{s=1}^{S} y_{v,r,s} = \sum_{s=1}^{S} y_{v,v*,s} \quad \forall v \in V$$

Equation 4 requires that the number of CDR3 and J* structures selected for each type of variable domain is equal to the number of V* structures selected (i.e. zero or one).

Writing the formulation in this way ensures that when a VL is designed, all structures used will be either V-KAPPA or V-LAMBDAY.

$$y_{v1,r1,s1} + y_{v2,r2,s2} \leq 1 \quad \forall (v1, r1, s1, v2, r2, s2) \in I_C^{clash}$$

This final inequality ensures that no structures that clash with one another and would form an unstable antibody are simultaneously selected. Equations 1-5 form the MAPs.
structure selection MILP, which can be solved to optimality using CPLEX [1] called directly from Python or from within GAMS [2].

A non-redundant database of 125 antibody – antigen complexes was created to evaluate the quality of the structures selected by the MAPs structure selection MILP. These PDB files were selected because they each had a VH and a VL, had resolutions no worse than 2.5 Å, and had antigens that were composed entirely of 150 or fewer amino acids. The use of amino acids is necessitated by the lack of CHARMM [3] parameters for non-protein molecules. Table 5.1 lists these antibodies and four calculated interaction energies: the energy of the parts with the antigen calculated with pairwise energy functions, the energy from pairwise energy functions of the assembled antibody after a CHARMM energy minimization, the CHARMM-calculated interaction energy of the selected antibody after the energy minimization, and the CHARMM-calculated interaction energy of the native antibody – antigen complex.

Only two antibodies, PDB 2AI6 and 3IU3, had parts selected for them with positive interaction energies. After an energy minimization, all designed antibodies had negative interaction energies when calculated by both pairwise energy functions and CHARMM. This suggests that for favorable antigen positions (e.g. native positions), it is possible for the MAPs structure selection MILP to design initial antibodies that successfully bind the antigen. Of particular note is the fact that for 72 / 125 cases (i.e. 57.6%), the designed antibodies had better CHARMM-calculated interaction energies than the native antibodies, by an average of 93.4 ± 65.7 kcal/mol. In the other 53 cases, the native complex was better by an average of 92.4 ± 89.3 kcal/mol. These interaction energies
changes are well within the range of improvements that would be expected from a typical IPRO affinity maturation.

It is not surprising that OptMAVEn can identify initial antibodies in some cases that are better than the native antibodies for two reasons. First, at any given moment the immune system will not have all possible naïve antibodies available, so sub-optimal antibodies can be expected to be chosen in many cases. In contrast, the MAPs structure selection MILP will always be able to pick an optimal antibody. Second, the objective of an immune response is to eliminate a problem as rapidly as possible, not to design optimally binding antibodies. Although strong binding is important, other factors such as stability, concentration, and when an antibody is discovered play a role. Since OptMAVEn is trying to design optimal binders, it is to be expected that the designed antibodies would differ from the naturally occurring ones. Nonetheless, it is promising that the initial antibodies selected by OptMAVEn appear, in very general terms, to bind about as well as an average affinity matured antibody.

**Antigen Positioning**

Having established that the MAPs structure selection MILP makes good choices when antigens are in good positions, it must be determined how to find good initial antigen positions. 750 antibody - antigen complexes were collected, with each complex having a resolution no worse than 2.5 Å and containing both a VL and a VH. The epitopes of the antigens were defined as any residue with at least one heavy atom within 4.5 Å of a CDR heavy atom [4] and the centers of mass of these epitopes were calculated. Figure 5.1 shows the distributions of the average center of mass of the epitopes along the X, Y, and Z axes. Using these distributions as a guideline for where to expect antigens to be
positioned, a grid search will be conducted to find good initial antigen positions. Seven coordinates along the X axis (-10 Å to 5 Å with steps of 2.5 Å), seven coordinates along the Y axis (-5 Å to 10 Å with steps of 2.5 Å), and eleven coordinates along the Z axis (3.75 Å to 16.25 Å with steps of 1.25 Å) will be considered, along with six rotations around the Z axis at each position, for a total of 3234 initial positions of the antigen. At each position, the best combination of MAPs parts to bind the antigen will be selected and the interaction energy of the parts determined. The lowest energy positions will be retained for 1000 iterations each of simulated annealing to find good initial antigen position and antibody structure combinations.

**IPRO Affinity Maturation**

In the second step of OptMAVeN, IPRO will be used to affinity mature the antibody to have improved binding to the antigen. IPRO will be implemented as described in Chapter 4, with several minor modifications. First, for each FR residue in the VH and V-KAPPA or V-LAMBDA domains, only amino acids that have been observed at that position in at least one percent of known human antibodies will be permitted. These permissions are based on the sequences of 7,652 VH, 2,247 V-KAPPA, and 1,605 V-LAMBDA unique, human antibody sequences downloaded from the IMGT/LIGM-DB [5]. In addition, one important consideration for designing antibodies for medical purposes is to reduce their immunogenicity so that they do not cause an immune reaction. Therefore, it is desirable to be able to design fully human antibodies to eliminate immunogenicity concerns. A major trigger in these immune responses is the type II major histocompatibility complex proteins (MHC II) displaying peptides of nine amino acids (9-mers) [6]. Although there are computational deimmunization programs available [6, 7], the extensive calculations
they require are incompatible with the rotamer selection MILP of IPRO. Therefore a conservative approach that only permits human 9-mers that have been previously observed is utilized. Analysis of the 11,504 human antibody sequences identified 330,590 unique human 9-mers. These 9-mers will be used to create integer cuts for the rotamer selection MILP that will ensure that the “humanness” of all 9-mers involved in mutations is either maintained at its initial level or increased. To keep the problem tractable, the maximum size of a perturbation during an IPRO iteration will be reduced from five to three (only residues within the perturbation are ever permitted to mutate). IPRO will be run treating every residue in the VH and VL as DPs, with a 3-fold higher preference for perturbing and mutating CDRs over FRs.

Experimental Validation

Although calculations show that the computational methods presented in this dissertation are promising, experimental confirmation of their efficacy is necessary. Towards this end, several experimental collaborations have been established. These experiments will be used to evaluate both successes and failures of the computational methods so that refinements and improvements to the methods may be identified.

Anti-FLAG (DYKD) Antibodies

OptCDR, as described in Chapters 2 and 4, was used to generate libraries of antibody CDRs against the FLAG antigen [8] (i.e. DYKD), with the objective of generating 90 total antibodies in three libraries. The first library, EEh, contains 12 unique sets of VH OptCDR-designed CDRs with several mutants of each unique set for a total size of 40 antibodies. The designed CDRs are attached to the VH of PDB: 3NN8, which is an anti-HIS tag antibody based on 3D5 [9], with the heavy FR and VL unmodified. The second
library, EEf, contains three unique sets of all six CDRs, with several mutants of each for a total size of ten antibodies, attached to the VH and VL of 3NN8. The final library, M2h, contains 12 unique sets of VH CDRs, with several mutants of each for a total library size of 40. The CDRs are attached to the VH of the anti-FLAG antibody M2 [PDB: 2G60] [10] and the heavy FR and VL are unmodified. The antigen was a 12-mer peptide from PDB: 1R77 [11] (GGDYKDDDDKGG) containing FLAG. These libraries are illustrated in Figure 5.2.

For each of the three libraries, OptCDR was used to generate 30 unique CDR designs, which were each affinity matured. The libraries were reduced to their final number of CDR designs (12, 3, and 12, respectively) based on their calculated interaction energies and observations of the complex structures. The primary determining factor in eliminating designs was the observed interactions between the CDRs and the specified DYKD epitope. Although OptCDR succeeded in always having interactions with the target epitope, many of the designs also had interactions with the neighboring pairs of residues (i.e. GG and DD). These designs were removed from consideration. For each of the selected structures, the top ten CDR mutant sequences were identified. Several of these were rationally selected for each antibody to generate the final libraries described in Figure 5.2. The CDR sequences of the final libraries are listed in Tables 5.2, 5.3, and 5.4, respectively.

The J. Maynard Lab at the University of Texas, Austin, conducted a preliminary analysis of the binding properties of the antibodies. They were expressed as single-chain variable fragments (scFv) on the surface of Escherichia coli and binding with a protein containing the FLAG peptide was monitored by ELISA [12]. Each library was screened with
approximately four times coverage. Figure 5.3 shows the binding results for each library, all of which show at least some positive binding to FLAG. Therefore OptCDR has successfully designed multiple, unique antibodies to bind a specified antigen epitope. Interestingly, the M2h library, which was based on an anti-FLAG antibody, shows the lowest percent of positive binding. This is possibly explained by known experimental concerns with the scFv expression of M2. Further analysis is ongoing to determine the binding affinities and structures of the most promising designs.

**Gammabodies**

Alzheimer’s disease is a devastating illness associated with misfolding and aggregation of amyloid-β (Aβ), particularly Aβ residues 1-42. The P. Tessier Lab at Rensselaer Polytechnic Institute has been working on designing special Aβ binding antibodies known as Grafted Amyloid Motif Antibodies (gammabodies) [13, 14]. A gammabody is a single antibody variable domain with an aggregation-prone region from Aβ42 inserted into a CDR, typically CDR3. Because Aβ42 has been shown to stack in-register with homotypic interactions causing aggregation [15-18], the hypothesis was that those homotypic interactions could mediate binding with the gammabodies. This was tested for two aggregation-prone regions, Aβ15-24 (QKLVFFAED) and Aβ33-42 (GLMVGGVVIA), and Aβ binding was experimentally observed [13]. The use of charged residues before and after the inserted aggregation-prone regions has also proven effective for reducing gammabody - gammabody aggregation [14].

The P. Tessier Lab is now interested in designing antibodies with Aβ aggregation-prone regions grafted into two CDRs. Specifically, residues 18-21 (VFFA) and 34-39 (LMVGGV). The MAPs database was scanned for pairs of CDRs with the proper
geometry to support the homotypic interactions that mediate Aβ binding. An exhaustive search for the pair of CDR structures that best match the geometry of Aβ was conducted. For a given pair of CDR structures, the analysis started with Aβ18-21 corresponding to the first four residues of the first CDR and Aβ34-39 corresponding to the first seven residues of the second CDR. Aβ was superimposed on the CDR structures with a minimized RMSD for the N, Cα, C, and Cβ of all eleven Aβ residues. The RMSDs of Aβ18-21 and Aβ34-39 were then calculated separately and the results were ranked based on the sum of these two independent RMSDs. This procedure ensures that Aβ is positioned as well as possible relative to the CDR residues but allows for an evaluation of how well Aβ aligns to each CDR independently. Once the residues at the start of each CDR had been compared with Aβ18-21 and Aβ34-39, the Aβ34-39 window in the second CDR was advanced by one residue and the new combination was evaluated. After all possibilities for the position of Aβ34-39 had been evaluated in the second CDR, the window of Aβ18-21 in the first CDR was advanced by one residue and the Aβ34-39 window was reset to the beginning of the second CDR. In this way, all possible positions of the Aβ18-21 and Aβ34-39 residues in all possible combinations of CDRs were evaluated.

It was determined that the combination of CDR structures that best match the geometry of Aβ18-21 and Aβ34-39 are both in a V-KAPPA: a CDR1 sequence of QSVLYSSNNKNY, with LYSS corresponding to Aβ18-21, and a CDR3 sequence of QQYNNWPPRIT, with WPPRIT corresponding to Aβ34-39. Although no experimentally determined structure exists for that antibody, it was determined that it would be the result of the V-(D)-J recombination of IGKV4-1*01 and IGKJ5*01. Two V
gene nucleotide mutations would be needed (tat -> aat to generate residue 108N and agt -> aat to generate residue 109N), along with an insertion of nag during recombination. These changes are well within the range of expected changes during V-(D)-J recombination and somatic hypermutation, based on Figure 3.3. The superposition of a MAPs predicted structure of this antibody with Aβ is shown in Figure 5.4.

There are obvious aggregation concerns for a single antibody variable domain with two grafted amyloid motifs in it. Inserting charged residues before or after the grafted sequence is a previously established method of preventing aggregation in gammabodies [14]. Coincidentally, there is a polar NN motif next to the grafting site for both motifs (after the motif in CDR1 and before it in CDR3). Therefore mutating these NN motifs to DD was explored. Model structures with the grafted motifs and charged mutations in CDR2 [19] were generated without charged mutations in either CDR1 or CDR3, with charged mutations only in CDR1, with charged mutations only in CDR3, and with charged mutations in both CDRs. Molecular dynamics simulations were performed using NAMD [20] with parameters specified by the CHARMM [3] force field. Each trajectory was slowly heated by running for 80 fs and then rescaling the atomic velocities to match a 1K temperature increase. Each system was equilibrated for 0.25 ns at 300K using the NVT ensemble followed by data collection for 5 ns using the NVE ensemble. In all cases, the CDR structures stayed relatively constant (i.e. conformed to a canonical structure). However, it was observed that inserting charged residues into CDR1 changed the canonical structure of that CDR, causing it to move towards the β hairpin turn in FR3 (residues 84-87). Inserting charged residues into CDR3 did not change its canonical structure. Experimental testing is now underway for the V-KAPPA variable domain made
from recombination of IGKV4-1*01 and IGKJ5-1 as described here, with grafted amyloid motifs in CDRs 1 and 3 and with charged mutations in CDR3.

A preliminary analysis was conducted to identify additional aggregation prone regions on the V-KAPPA domain. This domain is evolved to be positioned “next to” a VH and “on top” of a constant κ domain (C-KAPPA) in an IgG and it is therefore expected that there will be hydrophobic patches in those areas. Amino acids were separated into hydrophilic (DEGKNQRST), neutral (ACHP), and hydrophobic (FILMVWY) groups based on hydrophobicity data [4]. Figure 5.5 shows the V-KAPPA domain with the hydrophilic residues in green, the neutral residues in yellow, and the hydrophobic residues in red. As expected, hydrophobic patches were identified in the portions of the V-KAPPA that have contacts with VH and C-KAPPA in IgG. This data will be applied to suggest hydrophilic mutations if initial experiments fail to produce soluble gammabodies.

**Anti-CD20 Antibodies**

CD20 is a peptide antigen that is a therapeutically relevant target in B-cell lymphomas and leukemias. Several antibodies, including Rituximab [PDB: 2OSL] [21], are known to bind with ~8 nM affinities [22]. In collaboration with the T. Wood Lab at the Pennsylvania State University, four antibody libraries of approximately the same size (i.e., ~5 \(10^8\)) against CD20 will be designed and constructed to fairly assess the quantitative benefit of using computations in antibody library design. The first two antibody libraries will mimic traditional practices in antibody design [23] by first performing random mutagenesis of the entire variable domain, and secondly targeted saturation mutagenesis of six rationally selected positions. In the third library, the value of computations in guiding point mutations (i.e., affinity maturation) will be assessed by
using IPRO, as described in Chapter 4, to optimally redesign 12 positions, including the six positions of the second library. The fourth library will examine the improvements in affinity that can be gained by computationally designing the entire variable domains using OptMAVEEn. If possible, structural resolution for a select list of designs to elucidate the structural changes associated with affinity improvements will be explored. The experimental generation of the first two libraries has been initiated, but no results are yet available.

**Discussion**

*De novo* protein design is a very challenging problem. This dissertation has made significant contributions towards designing antibody CDRs (OptCDR), modeling antibody variable domains (MAPs), and redesigning proteins in general (IPRO Suite of Programs). Ongoing experimental validations and continued computational method development will allow for further refinement of the computational methods described here, along with hopefully providing a novel and powerful tool for the design of antibody medications (OptMAVEEn). The major advance presented here is the development of general methods for the *de novo* design an entire family of proteins. The *de novo* design methods developed in this dissertation have focused on modeling portions of antibody structures and selecting optimal combinations of structures to bind a particular antigen. However, it is important to note that the success of these and other protein design methods remain limited by the accuracy of force fields’ descriptions of protein structures and functions.

Extending these methods of *de novo* protein design beyond antibodies would be very challenging, but is feasible. First, all known protein structures from the PDB would need
to be analyzed and separated into secondary structure elements (i.e. α helixes, β sheets, disordered loops, etc.). This is approximately two orders of magnitude more structures than were analyzed in this dissertation, suggesting that calculations can be accomplished in a reasonable time period. Next, the structures would need to be clustered together to generate a database of parts that spans the diversity of what has been observed in known protein structures. These clusters can be used to determine rules of what kinds of amino acid sequences are permitted for different parts and which are not, facilitating later design stages. The most challenging portion of this proposed protein design workflow is determining how to assemble the parts. An advantage of antibodies is that their modular features allow the databases discussed in this dissertation to be generated such that the parts have a fixed global position relative to one another. This is not true for proteins in general. Therefore, a user would have to define a functional site for their protein (i.e. catalysis, binding, etc.) and then a genetic algorithm would have to search for arrangements of parts that fit that geometry. Developing such an algorithm is a formidable challenge, but should be possible. Then, once an initial protein’s structure is generated, IPRO or a related method can be used to suggest mutations that improve the protein’s stability and functionality.

This proposed general de novo protein engineering workflow is probably viable but also definitely a challenging manner of extending the de novo antibody design methods developed in this dissertation. The challenges inherent in it mean that it may never be realized; yet doing so would be an important break-through in protein engineering that would have numerous experimental and real-world applications. Even if a general de novo protein design method is not created based on this research, important contributions
to antibody design methods have been discussed here. The ability to design antibodies to
bind any specified epitope is a useful objective that is extremely difficult to meet
experimentally. Doing so while eliminating immunogenicity, as in OptMAVEn, may be
an essential tool for developing valuable medications.
**Figure 5.1: Average Epitope Positions.** The average positions of 750 antigen epitopes were calculated and plotted. It was found that in the X and Y directions, these positions form bell-curve distributions with no significant difference in the positions of antigens based on their size. In contrast, in the Z direction, larger antigens were generally placed further from the antibody. The coordinates along the X, Y, and Z directions in this figure will form the grid points for the initial antigen placement search.
Figure 5.2: Libraries of OptCDR-Designed Anti-FLAG Antibodies. These are the three libraries of 90 total OptCDR-designed anti-FLAG antibodies. The first library in panel A, EEh, contains 12 sets of VH CDRs, and 40 total antibodies, attached to the anti-HIS tag antibody 3D5. The second library in panel B, EEf, is made up of ten total antibodies that are mutants of three unique designs of all six CDRs attached to the FR of 3D5. The final library in panel C, M2h, is 40 total mutants of 12 VH CDR designs attached to the anti-FLAG antibody M2.
A) EEh

Activity Ratio:
FLAG Binding/Blank

73.2% Positive Binders

B) EEf

Activity Ratio

40.3% Positive Binders

C) M2h

Activity Ratio

7.8% Positive Binders
Figure 5.3: Preliminary Binding Data for Anti-FLAG Antibodies. These plots show preliminary binding data, as determined by phage ELISA, for the three anti-FLAG OptCDR designed antibody libraries. All three libraries show at least some positive binding, indicating that OptCDR successfully designed multiple, unique antibodies against the specified antigen epitope.
Figure 5.4: Aβ Superimposed on a V-KAPPA Domain. This is the Aβ – V-KAPPA complex that has the most similar geometry between a pair of CDRs and the structure of the Aβ aggregation prone regions. The V-KAPPA domain is shown in dark gray with Aβ in light gray. On the left, in red, are the Aβ18-21 residues and the corresponding residues in CDR1 and on the right, in orange, are Aβ34-39 and the corresponding CDR3 residues. It is clear that while the geometry between the structures is not identical, there is a high degree of similarity.
Figure 5.5: Hydrophobic Regions of the V-KAPPA Domain. In this figure, hydrophilic residues are shown in green, neutral residues are shown in yellow, and hydrophobic residues are shown in red. A visual examination of the structure revealed three hydrophobic surfaces: the CDRs (only CDR3 is visible in this figure), the “side” of the antibody that normally interacts with VH, and the “bottom” of the antibody that normally interacts with C-KAPPA.
Table 5.1: Calculated Energies for MAPs – Antigen Complexes. These are the 125 non-redundant antibody-antigen complexes used to evaluate the quality of antibody designs selected by the MAPs structure selection MILP. All energies are in kcal/mol and four interaction energies are reported for each antigen: the energy of the MAPs selected structures using pairwise energy functions, the energy of the CHARMM minimized structure using pairwise energy functions, the energy calculated by CHARMM, and the energy calculated by CHARMM for the native antibody. 57.6% of the MAPs selected structures have better CHARMM-calculated interaction energies than the native antibody-antigen complexes.

<table>
<thead>
<tr>
<th>Name</th>
<th>MILP</th>
<th>Min.</th>
<th>CHARMM</th>
<th>Native</th>
<th>Name</th>
<th>MILP</th>
<th>Min.</th>
<th>CHARMM</th>
<th>Native</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ACY</td>
<td>-290.5</td>
<td>-274</td>
<td>-325.8</td>
<td>-206.3</td>
<td>1BJ1</td>
<td>-35.8</td>
<td>-85.5</td>
<td>-146.6</td>
<td>-180.2</td>
</tr>
<tr>
<td>1CE1</td>
<td>-218.5</td>
<td>-178.4</td>
<td>-220.5</td>
<td>-428.8</td>
<td>1CFS</td>
<td>-139.7</td>
<td>-195.2</td>
<td>-259.8</td>
<td>-272.5</td>
</tr>
<tr>
<td>1CFT</td>
<td>-274.8</td>
<td>-295.8</td>
<td>-329.9</td>
<td>-139.6</td>
<td>1CU4</td>
<td>-214.5</td>
<td>-305.1</td>
<td>-386.2</td>
<td>-500</td>
</tr>
<tr>
<td>1DZB</td>
<td>-275.2</td>
<td>-329.5</td>
<td>-454.3</td>
<td>-408.4</td>
<td>1EGJ</td>
<td>-206.3</td>
<td>-191.3</td>
<td>-304.2</td>
<td></td>
</tr>
<tr>
<td>1F90</td>
<td>-267.6</td>
<td>-260.4</td>
<td>-285.6</td>
<td>-192.7</td>
<td>1EJO</td>
<td>-299.2</td>
<td>-267.6</td>
<td>-336.6</td>
<td>-285.6</td>
</tr>
<tr>
<td>1FPT</td>
<td>-225.8</td>
<td>-250.4</td>
<td>-334.7</td>
<td>-473.6</td>
<td>1GGI</td>
<td>-306.6</td>
<td>-262.7</td>
<td>-268</td>
<td>-321.4</td>
</tr>
<tr>
<td>1HH6</td>
<td>-161.3</td>
<td>-103.8</td>
<td>-175.5</td>
<td>-231.9</td>
<td>1HIN</td>
<td>-152.6</td>
<td>-187.9</td>
<td>-237.6</td>
<td>-195.2</td>
</tr>
<tr>
<td>1I8I</td>
<td>-352.5</td>
<td>-346.6</td>
<td>-403</td>
<td>-474</td>
<td>1H9R</td>
<td>-214.9</td>
<td>-221.5</td>
<td>-324.8</td>
<td>-237.9</td>
</tr>
<tr>
<td>1JHL</td>
<td>-260.1</td>
<td>-260.2</td>
<td>-332.1</td>
<td>-277.3</td>
<td>1JP5</td>
<td>-131.5</td>
<td>-137.1</td>
<td>-192.5</td>
<td>-214</td>
</tr>
<tr>
<td>1JRH</td>
<td>-253.8</td>
<td>-219.3</td>
<td>-304.5</td>
<td>-464</td>
<td>1KB5</td>
<td>-358.7</td>
<td>-265.7</td>
<td>-306.9</td>
<td>-265.8</td>
</tr>
<tr>
<td>1KC5</td>
<td>-140.3</td>
<td>-154.9</td>
<td>-167.4</td>
<td>-188.4</td>
<td>1KCR</td>
<td>-180</td>
<td>-183.1</td>
<td>-272</td>
<td>-97.1</td>
</tr>
<tr>
<td>1KIQ</td>
<td>-135.4</td>
<td>-173.5</td>
<td>-251.2</td>
<td>-171.2</td>
<td>1KTR</td>
<td>-190.7</td>
<td>-173.1</td>
<td>-216.3</td>
<td>-6.6</td>
</tr>
<tr>
<td>1MLC</td>
<td>-149.1</td>
<td>-117.4</td>
<td>-207.5</td>
<td>-288.2</td>
<td>1MVU</td>
<td>-293</td>
<td>-313.8</td>
<td>-356.5</td>
<td>-115.9</td>
</tr>
<tr>
<td>1N64</td>
<td>-113.1</td>
<td>-124.1</td>
<td>-192.1</td>
<td>-197.2</td>
<td>1NAK</td>
<td>-170.6</td>
<td>-205.5</td>
<td>-268.8</td>
<td>-307.5</td>
</tr>
<tr>
<td>Code</td>
<td>Value1</td>
<td>Value2</td>
<td>Value3</td>
<td>Value4</td>
<td>Value5</td>
<td>Value6</td>
<td>Value7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1NSN</td>
<td>-384.2</td>
<td>-372.6</td>
<td>-492.6</td>
<td>-209.8</td>
<td>1OAZ</td>
<td>-169.1</td>
<td>-203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1OBE</td>
<td>-2.2</td>
<td>-125.2</td>
<td>-201</td>
<td>-373</td>
<td>1ORS</td>
<td>-246.1</td>
<td>-242.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1P4B</td>
<td>-91.5</td>
<td>-63.6</td>
<td>-114.2</td>
<td>-126.1</td>
<td>1PZ5</td>
<td>-84.5</td>
<td>-85.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1QKZ</td>
<td>-123.5</td>
<td>-134.5</td>
<td>-191.7</td>
<td>-142.8</td>
<td>1NZQ</td>
<td>-149.5</td>
<td>-214.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1RJL</td>
<td>-168.4</td>
<td>-148.3</td>
<td>-201.3</td>
<td>-210.5</td>
<td>1SM3</td>
<td>-227.2</td>
<td>-238.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1TET</td>
<td>-137.9</td>
<td>-92.9</td>
<td>-158.6</td>
<td>-221.6</td>
<td>1TQB</td>
<td>-311.1</td>
<td>-234.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1TZG</td>
<td>-111</td>
<td>-111.3</td>
<td>-192.1</td>
<td>-220.1</td>
<td>1TZH</td>
<td>-121.7</td>
<td>-150.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1TZI</td>
<td>-133.8</td>
<td>-138.2</td>
<td>-159.7</td>
<td>-36.8</td>
<td>1U8J</td>
<td>-239.5</td>
<td>-252.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1UWX</td>
<td>-269.4</td>
<td>-263</td>
<td>-304.7</td>
<td>-127</td>
<td>1V7M</td>
<td>-131.8</td>
<td>-182.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1W72</td>
<td>-204.2</td>
<td>-217.8</td>
<td>-273.8</td>
<td>-238.4</td>
<td>1XGY</td>
<td>-298.4</td>
<td>-258</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1XIW</td>
<td>-213.2</td>
<td>-253.5</td>
<td>-321.4</td>
<td>-478.1</td>
<td>1ZA3</td>
<td>-348.4</td>
<td>-395.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1ZTX</td>
<td>-329.6</td>
<td>-309.9</td>
<td>-400</td>
<td>-315.1</td>
<td>2A6I</td>
<td>173.6</td>
<td>-139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2B1H</td>
<td>-219.5</td>
<td>-294.2</td>
<td>-363.7</td>
<td>-600.1</td>
<td>2BDN</td>
<td>-189.8</td>
<td>-435.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2CK0</td>
<td>-75.9</td>
<td>-73.1</td>
<td>-124.1</td>
<td>-109.2</td>
<td>2DQJ</td>
<td>-178.5</td>
<td>-274.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2EH8</td>
<td>-218</td>
<td>-164.9</td>
<td>-210.2</td>
<td>-276.5</td>
<td>2FJH</td>
<td>-267.4</td>
<td>-158.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2G5B</td>
<td>-224.4</td>
<td>-205.6</td>
<td>-256.9</td>
<td>-173</td>
<td>2HIP</td>
<td>-110.8</td>
<td>-107.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2HFG</td>
<td>-210.5</td>
<td>-172.9</td>
<td>-217.2</td>
<td>-207.6</td>
<td>2HH0</td>
<td>-282.9</td>
<td>-306</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2HKF</td>
<td>-210.2</td>
<td>-247.6</td>
<td>-282.8</td>
<td>-477.7</td>
<td>2HRP</td>
<td>-215.2</td>
<td>-221.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2HVK</td>
<td>-307</td>
<td>-365.3</td>
<td>-437</td>
<td>-505.2</td>
<td>2IFF</td>
<td>-276.4</td>
<td>-209.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2IGF</td>
<td>-261.7</td>
<td>-296.2</td>
<td>-355.3</td>
<td>-327</td>
<td>2J4W</td>
<td>-258.1</td>
<td>-292.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2JEL</td>
<td>-283.1</td>
<td>-255.9</td>
<td>-354.1</td>
<td>-232.1</td>
<td>2OQJ</td>
<td>-116</td>
<td>-128.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2OR9</td>
<td>-77.9</td>
<td>-65.1</td>
<td>-102.2</td>
<td>-266.8</td>
<td>2OSL</td>
<td>-31.9</td>
<td>-77.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2QHR</td>
<td>-251</td>
<td>-254.4</td>
<td>-329.8</td>
<td>-471.3</td>
<td>2QR0</td>
<td>-254.4</td>
<td>-259.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2QSC</td>
<td>-261.3</td>
<td>-277</td>
<td>-334</td>
<td>-198</td>
<td>2R0W</td>
<td>-224.9</td>
<td>-196.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2R29</td>
<td>-118.5</td>
<td>-122.8</td>
<td>-205.7</td>
<td>-327.6</td>
<td>2VWE</td>
<td>-70.1</td>
<td>-108.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2VXQ</td>
<td>-286</td>
<td>-341.4</td>
<td>-441.4</td>
<td>-387.7</td>
<td>2VXS</td>
<td>-78.3</td>
<td>-95.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2ZPK</td>
<td>-110.8</td>
<td>-114.5</td>
<td>-155.9</td>
<td>-126.7</td>
<td>2UQ</td>
<td>-81.1</td>
<td>-81.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code</td>
<td>3AB0</td>
<td>-327</td>
<td>-350.9</td>
<td>-443.4</td>
<td>-300.5</td>
<td>3BAE</td>
<td>-230.5</td>
<td>-232.2</td>
<td>-307.4</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>3BDY</td>
<td>-197.8</td>
<td>-156.8</td>
<td>-211</td>
<td>-170.3</td>
<td>3BK</td>
<td>-133.5</td>
<td>-53.6</td>
<td>-35.4</td>
<td>-131.3</td>
</tr>
<tr>
<td>3CVH</td>
<td>-167.2</td>
<td>-153.1</td>
<td>-153.1</td>
<td>-189.5</td>
<td>3CXD</td>
<td>-212.9</td>
<td>-204.6</td>
<td>-265.3</td>
<td>-294.6</td>
</tr>
<tr>
<td>3D85</td>
<td>-98.8</td>
<td>-111.9</td>
<td>-196.3</td>
<td>-180.7</td>
<td>3DVG</td>
<td>-128.9</td>
<td>-189.9</td>
<td>-222.2</td>
<td>-194.5</td>
</tr>
<tr>
<td>3E8U</td>
<td>-171.5</td>
<td>-172.5</td>
<td>-231.4</td>
<td>-199.7</td>
<td>3EO1</td>
<td>-242.3</td>
<td>-170</td>
<td>-237</td>
<td>-121.6</td>
</tr>
<tr>
<td>3ETB</td>
<td>-311</td>
<td>-338.5</td>
<td>-410.5</td>
<td>-592.7</td>
<td>3EYU</td>
<td>-274.2</td>
<td>-239.2</td>
<td>-269</td>
<td>-240.7</td>
</tr>
<tr>
<td>3F58</td>
<td>-279.4</td>
<td>-284.1</td>
<td>-316.2</td>
<td>-286.3</td>
<td>3FFD</td>
<td>-182.4</td>
<td>-236</td>
<td>-324.1</td>
<td>-511.9</td>
</tr>
<tr>
<td>3FN0</td>
<td>-90.1</td>
<td>-118.2</td>
<td>-182</td>
<td>-176</td>
<td>3G5V</td>
<td>-170.2</td>
<td>-201.9</td>
<td>-279.9</td>
<td>-261.7</td>
</tr>
<tr>
<td>3G6D</td>
<td>-540.7</td>
<td>-609.2</td>
<td>-686.5</td>
<td>-728.8</td>
<td>3GGW</td>
<td>-152.4</td>
<td>-160.3</td>
<td>-221.8</td>
<td>-88.1</td>
</tr>
<tr>
<td>3GHB</td>
<td>-75.1</td>
<td>-60.1</td>
<td>-80.1</td>
<td>-351.7</td>
<td>3GHE</td>
<td>-196</td>
<td>-217</td>
<td>-304.8</td>
<td>-386.8</td>
</tr>
<tr>
<td>3GJG</td>
<td>-58.4</td>
<td>-74.2</td>
<td>-114.5</td>
<td>-84.7</td>
<td>3HR5</td>
<td>-374.3</td>
<td>-342.1</td>
<td>-415</td>
<td>-340</td>
</tr>
<tr>
<td>3IU3</td>
<td>7.1</td>
<td>-174.4</td>
<td>-306.5</td>
<td>-801</td>
<td>3KS0</td>
<td>-157.3</td>
<td>-185.7</td>
<td>-248.4</td>
<td>-148</td>
</tr>
<tr>
<td>3L5W</td>
<td>-437.4</td>
<td>-459.5</td>
<td>-534.7</td>
<td>-522</td>
<td>3L5Y</td>
<td>-358.9</td>
<td>-406.6</td>
<td>-476</td>
<td>-570.5</td>
</tr>
<tr>
<td>3LQA</td>
<td>-162.4</td>
<td>-175.9</td>
<td>-224.4</td>
<td>-81.8</td>
<td>3MLR</td>
<td>-231.9</td>
<td>-303.2</td>
<td>-390.1</td>
<td>-345.4</td>
</tr>
<tr>
<td>3MLS</td>
<td>-271.1</td>
<td>-306.5</td>
<td>-401.9</td>
<td>-406.2</td>
<td>3MLW</td>
<td>-434.8</td>
<td>-498.1</td>
<td>-596.7</td>
<td>-490</td>
</tr>
<tr>
<td>3MLX</td>
<td>-115.3</td>
<td>-109.9</td>
<td>-143.4</td>
<td>-174.1</td>
<td>3MLY</td>
<td>-183.2</td>
<td>-192.7</td>
<td>-230.5</td>
<td>-246.5</td>
</tr>
<tr>
<td>3NFP</td>
<td>-222.6</td>
<td>-287</td>
<td>-414.6</td>
<td>-351.7</td>
<td>3NH7</td>
<td>-403.4</td>
<td>-432.3</td>
<td>-568.9</td>
<td>-782.7</td>
</tr>
<tr>
<td>3O0R</td>
<td>-173.5</td>
<td>-180.1</td>
<td>-233.9</td>
<td>-120.5</td>
<td>3O2D</td>
<td>-83.8</td>
<td>-102.7</td>
<td>-152.5</td>
<td>-67.1</td>
</tr>
<tr>
<td>3P30</td>
<td>-175.9</td>
<td>-193.3</td>
<td>-287.7</td>
<td>-304.5</td>
<td>3PJS</td>
<td>-292.6</td>
<td>-302.6</td>
<td>-371.6</td>
<td>-121</td>
</tr>
<tr>
<td>3QG6</td>
<td>-109.1</td>
<td>-103.6</td>
<td>-182.7</td>
<td>-130.2</td>
<td>3QWO</td>
<td>-417.2</td>
<td>-418.4</td>
<td>-476.3</td>
<td>-338.9</td>
</tr>
<tr>
<td>3RKD</td>
<td>-210.4</td>
<td>-231.6</td>
<td>-325.2</td>
<td>-347.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.2: The EEh Anti-FLAG Library CDR Sequences. The H1, H2, and H3 CDR sequences of the antibodies in the EEh library. The designs are given two numbers (e.g. 1.1), where the first number is the number of the unique CDR design from the OptCDR calculations and the second is the number of the mutant of that combination of CDRs. A mutant number of 0 refers to the original predicted sequence.
<table>
<thead>
<tr>
<th>Antibody</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEh1.0</td>
<td>GYSITRSSAG</td>
<td>NGASQDSK</td>
<td>AMFDQVGRGHW</td>
</tr>
<tr>
<td>EEh1.1</td>
<td>GYSITRSSAG</td>
<td>NGASQDSK</td>
<td>AGFKNVMRGHY</td>
</tr>
<tr>
<td>EEh1.2</td>
<td>GYSITRSSAG</td>
<td>IGADRSST</td>
<td>AMFDQVGRGHW</td>
</tr>
<tr>
<td>EEh1.4</td>
<td>GDSLTSYNAG</td>
<td>NGASQDSK</td>
<td>AMFDQVGRGHW</td>
</tr>
<tr>
<td>EEh1.5</td>
<td>GYSITRSSAG</td>
<td>NSADKYGS</td>
<td>AMFDQVGRGHW</td>
</tr>
<tr>
<td>EEh2.0</td>
<td>GDSVTSYNAG</td>
<td>TASGESNK</td>
<td>SRKNMAGW</td>
</tr>
<tr>
<td>EEh2.2</td>
<td>GDSVTSYNAG</td>
<td>VGSRSENQ</td>
<td>SRKNMAGW</td>
</tr>
<tr>
<td>EEh2.3</td>
<td>GDSVTSYNAG</td>
<td>TASGESNK</td>
<td>SKTMRAFG</td>
</tr>
<tr>
<td>EEh2.4</td>
<td>GDSVTSYNAG</td>
<td>TASGESNK</td>
<td>VKNSRGAF</td>
</tr>
<tr>
<td>EEh3.0</td>
<td>GRSYTAQG</td>
<td>SRDNSEGGAT</td>
<td>SAGDF</td>
</tr>
<tr>
<td>EEh3.1</td>
<td>GRSYTAQG</td>
<td>SRDNSEGGAT</td>
<td>GMAFD</td>
</tr>
<tr>
<td>EEh3.3</td>
<td>GATYKSQG</td>
<td>SRDNSEGGAT</td>
<td>SAGDF</td>
</tr>
<tr>
<td>EEh4.0</td>
<td>GGSLTRSNMA</td>
<td>ANGKTMGES</td>
<td>AGWDNYYEGRF</td>
</tr>
<tr>
<td>EEh5.0</td>
<td>GESAVNG</td>
<td>SGDKGNSHAT</td>
<td>AGYRNWKGFEF</td>
</tr>
<tr>
<td>EEh5.2</td>
<td>GDSVSTIG</td>
<td>SGDKGNSHAT</td>
<td>AGYRNWKGFEF</td>
</tr>
<tr>
<td>EEh5.5</td>
<td>GESAVNG</td>
<td>SGSKGNSDAK</td>
<td>AGYRNWKGFEF</td>
</tr>
<tr>
<td>EEh5.6</td>
<td>DFNGQATG</td>
<td>SGDKGNSHAT</td>
<td>AGYRNWKGFEF</td>
</tr>
<tr>
<td>EEh6.0</td>
<td>GDSLTSAMG</td>
<td>GSGKTVAES</td>
<td>AKSGF</td>
</tr>
<tr>
<td>EEh6.1</td>
<td>GFSVSTKMG</td>
<td>GSGKTVAES</td>
<td>AKSGF</td>
</tr>
<tr>
<td>EEh6.2</td>
<td>GDSLSTFNAG</td>
<td>GSGKTVAES</td>
<td>AKSGF</td>
</tr>
<tr>
<td>EEh6.3</td>
<td>GDSLTSAMG</td>
<td>GSFGNKVDQ</td>
<td>AKSGF</td>
</tr>
<tr>
<td>EEh6.5</td>
<td>GDSLTSAMG</td>
<td>GSGKTVAES</td>
<td>SRGMY</td>
</tr>
<tr>
<td>EEh7.0</td>
<td>GDSLTSYSAG</td>
<td>MRTSGEA</td>
<td>VAFKSMIDGGW</td>
</tr>
<tr>
<td>EEh7.2</td>
<td>GDSLTSYSAG</td>
<td>VRSSEGNA</td>
<td>VAFKSMIDGGW</td>
</tr>
<tr>
<td>EEh7.3</td>
<td>GDSLSTSTAMG</td>
<td>MRTSGEA</td>
<td>VAFKSMIDGGW</td>
</tr>
<tr>
<td>EEh7.4</td>
<td>GDSLTSYSAG</td>
<td>MRTSGEA</td>
<td>DGWKQMYPEGAF</td>
</tr>
<tr>
<td>EEh8.0</td>
<td>GVSYENSG</td>
<td>SGSTQTHKDAT</td>
<td>SMKRGAV</td>
</tr>
<tr>
<td>EEh8.1</td>
<td>GESMSGQA</td>
<td>SGSTQTHKDAT</td>
<td>SMKRGAV</td>
</tr>
<tr>
<td>EEh8.3</td>
<td>GVSYENSG</td>
<td>SGSTQTHKDAT</td>
<td>SMKRGAV</td>
</tr>
<tr>
<td>EEh8.4</td>
<td>GVSYENSG</td>
<td>SGSTQTHKDAT</td>
<td>ASKRGLV</td>
</tr>
<tr>
<td>EEh8.5</td>
<td>GETYNASG</td>
<td>SGSTQTHKDAT</td>
<td>SMKRGAV</td>
</tr>
<tr>
<td>EEh9.0</td>
<td>GYTFSSGD</td>
<td>SKDQGENGAT</td>
<td>ARFKGSDMY</td>
</tr>
<tr>
<td>EEh9.1</td>
<td>GYTFSSGD</td>
<td>SKDQGENGAT</td>
<td>VTRKESGAF</td>
</tr>
<tr>
<td>EEh11.0</td>
<td>GDSVTSFNAG</td>
<td>MSSNGDKQ</td>
<td>DGFYQAMKGRY</td>
</tr>
<tr>
<td>EEh13.0</td>
<td>GDSVTSYSAG</td>
<td>GAVMST</td>
<td>AKSTGRYDF</td>
</tr>
<tr>
<td>EEh13.4</td>
<td>GDSVTSYSAG</td>
<td>LGAMSV</td>
<td>AKSTGRYDF</td>
</tr>
<tr>
<td>EEh13.6</td>
<td>GDSLSSFNAG</td>
<td>GAVMST</td>
<td>AKSTGRYDF</td>
</tr>
<tr>
<td>EEh14.0</td>
<td>GDSLSSYNAG</td>
<td>VGAMST</td>
<td>VRNEWSGAF</td>
</tr>
<tr>
<td>EEh14.3</td>
<td>GDSLSSYNAG</td>
<td>MAGVST</td>
<td>VRNEWSGAF</td>
</tr>
<tr>
<td>EEh14.4</td>
<td>GDSLSSYNAG</td>
<td>GALMNI</td>
<td>VRNEWSGAF</td>
</tr>
</tbody>
</table>
**Table 5.3: The EEf Anti-FLAG Library CDR Sequences.** The CDR sequences of the antibodies in the EEf library. Each antibody is named using two numbers (e.g. 5.1). The first number refers to the number of the unique CDR combination generated by OptCDR (i.e. 5 is the fifth design out of 30) and the second number specifies the mutant of the CDR design. A mutant number of 0 refers to the original OptCDR design with no mutations.

<table>
<thead>
<tr>
<th>Antibody</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEf5.0</td>
<td>GDIGSSYT</td>
<td>SSGGQEART</td>
<td>AHSRAMQLFDSSGGKE</td>
<td>QGSRAN</td>
<td>AGD</td>
<td>SGFENSQARSM</td>
</tr>
<tr>
<td>EEf5.1</td>
<td>GYIGSSHT</td>
<td>SSGGQEART</td>
<td>AHSRAMQLFDSSGGKE</td>
<td>QGSRAN</td>
<td>AGD</td>
<td>SGFENSQARSM</td>
</tr>
<tr>
<td>EEf15.0</td>
<td>GFSIKGAN</td>
<td>VRGDAST</td>
<td>ADRKMYSFYSGGGEA</td>
<td>NARSGS</td>
<td>DGN</td>
<td>SSWRNQGAKSI</td>
</tr>
<tr>
<td>EEf15.2</td>
<td>GFSIKGAN</td>
<td>VRGDAST</td>
<td>ADRKMYSFYSGGGEA</td>
<td>GATKSQ</td>
<td>DGN</td>
<td>SSWRNQGAKSI</td>
</tr>
<tr>
<td>EEf15.3</td>
<td>GFSIKGAN</td>
<td>VNDSGEA</td>
<td>ADRKMYSFYSGGGEA</td>
<td>NARSGS</td>
<td>DGN</td>
<td>SSWRNQGAKSI</td>
</tr>
<tr>
<td>EEf15.4</td>
<td>GFSIKGAN</td>
<td>VRGDAST</td>
<td>ADRKMYSFYSGGGEA</td>
<td>NARSGS</td>
<td>DGN</td>
<td>SAFDQTNKYVG</td>
</tr>
<tr>
<td>EEf15.5</td>
<td>GFSIKGAN</td>
<td>VRGDAST</td>
<td>ADRKMYSFYSGGGEA</td>
<td>NARSGS</td>
<td>DGN</td>
<td>SSWRNQGAKSI</td>
</tr>
<tr>
<td>EEf24.0</td>
<td>AFSGMTSG</td>
<td>INTSSEGK</td>
<td>ALDAIMKSGGRW</td>
<td>SRNVGANDS</td>
<td>RGN</td>
<td>GSYDLSMKQA</td>
</tr>
<tr>
<td>EEf24.2</td>
<td>AFSGMTSG</td>
<td>INTSSEGK</td>
<td>ALKSMARNGGDW</td>
<td>SRNVGANDS</td>
<td>RGN</td>
<td>GSYDLSMKQA</td>
</tr>
<tr>
<td>EEf24.5</td>
<td>GWSMATQG</td>
<td>INTSSEGK</td>
<td>ALDAIMKSGGRW</td>
<td>SRNVGANDS</td>
<td>RGN</td>
<td>GSYDLSMKQA</td>
</tr>
</tbody>
</table>
Table 5.4: The M2h Anti-FLAG Library CDR Sequences. The H1, H2, and H3 CDR sequences of the antibodies in the M2h library. The designs are given two numbers (e.g. 1.1), where the first number is the number of the unique CDR design from the OptCDR calculations and the second is the number of the mutant of that combination of CDRs. A mutant number of 0 refers to the original predicted sequence.
<table>
<thead>
<tr>
<th>Antibody</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2h1.0</td>
<td>GVSGKSTA</td>
<td>SGSNGDSHAK</td>
<td>KMFENYADGGY</td>
</tr>
<tr>
<td>M2h1.1</td>
<td>GVSGKSTA</td>
<td>SGSNGDSHAK</td>
<td>ALFEQGVGDGMW</td>
</tr>
<tr>
<td>M2h1.3</td>
<td>GVSGKSTA</td>
<td>SGDQSJKSGEM</td>
<td>KMFENYADGGY</td>
</tr>
<tr>
<td>M2h1.4</td>
<td>GYSAKQSG</td>
<td>SGSNGDSHAK</td>
<td>KMFENYADGGY</td>
</tr>
<tr>
<td>M2h1.5</td>
<td>GVSGKSTA</td>
<td>SGSNSEGDTK</td>
<td>KMFENYADGGY</td>
</tr>
<tr>
<td>M2h4.0</td>
<td>GDISTSNGW</td>
<td>SGSRGNDSAK</td>
<td>AYYESMLRGGW</td>
</tr>
<tr>
<td>M2h4.1</td>
<td>GDVSSGNT</td>
<td>SGSRGNDSAK</td>
<td>AYYESMLRGGW</td>
</tr>
<tr>
<td>M2h4.2</td>
<td>GDISTSNGW</td>
<td>FRSGQKNKGET</td>
<td>AYYESMLRGGW</td>
</tr>
<tr>
<td>M2h4.3</td>
<td>GSIVTSGD</td>
<td>SGSRGNDSAK</td>
<td>AYYESMLRGGW</td>
</tr>
<tr>
<td>M2h4.4</td>
<td>GDISTSNGW</td>
<td>SGSQGKSHAM</td>
<td>AYYESMLRGGW</td>
</tr>
<tr>
<td>M2h4.5</td>
<td>GDISTSNGW</td>
<td>FKSGQNSGAT</td>
<td>AYYESMLRGGW</td>
</tr>
<tr>
<td>M2h5.0</td>
<td>GGSLSSTKVA</td>
<td>LAGMN</td>
<td>GARMLYDFAGEY</td>
</tr>
<tr>
<td>M2h5.5</td>
<td>GGSLSSTKVA</td>
<td>LGMAQS</td>
<td>GARMLYDFAGEY</td>
</tr>
<tr>
<td>M2h6.0</td>
<td>GENATISG</td>
<td>SGSRGNKDAM</td>
<td>AVWHSMFDGGY</td>
</tr>
<tr>
<td>M2h6.1</td>
<td>GENATISG</td>
<td>SGDRSNSGAK</td>
<td>AVWHSMFDGGY</td>
</tr>
<tr>
<td>M2h6.5</td>
<td>GENATISG</td>
<td>SGSRGNKDAK</td>
<td>GMWHSVLEGAY</td>
</tr>
<tr>
<td>M2h7.0</td>
<td>GRIGSSNY</td>
<td>VTKEGSI</td>
<td>SAGDY</td>
</tr>
<tr>
<td>M2h7.1</td>
<td>GRIGSSNY</td>
<td>MSSDGKT</td>
<td>SAGDY</td>
</tr>
<tr>
<td>M2h7.3</td>
<td>GDISNSGY</td>
<td>VTKEGSI</td>
<td>SAGDY</td>
</tr>
<tr>
<td>M2h8.0</td>
<td>GRTGSMQA</td>
<td>MGSDAKT</td>
<td>SAGEY</td>
</tr>
<tr>
<td>M2h8.3</td>
<td>GRTGSMQA</td>
<td>MGSDAKT</td>
<td>AVGEF</td>
</tr>
<tr>
<td>M2h10.0</td>
<td>GASVKSTG</td>
<td>IGAVQS</td>
<td>GKDASSGYVFILDVAGAFME</td>
</tr>
<tr>
<td>M2h10.1</td>
<td>GASVKSTG</td>
<td>IGAVQS</td>
<td>GDRVSNAWLFIGEVA VGAMYAK</td>
</tr>
<tr>
<td>M2h11.0</td>
<td>GFIISTSGD</td>
<td>SKSNGDGHA T</td>
<td>ARDESKYYGAGF</td>
</tr>
<tr>
<td>M2h11.2</td>
<td>GFIISTSGD</td>
<td>SKSNGDGHAT</td>
<td>ARDESQWKAGF</td>
</tr>
<tr>
<td>M2h11.5</td>
<td>GYSGSIND</td>
<td>SKSNGDGHAT</td>
<td>ARDESKYYGAGF</td>
</tr>
<tr>
<td>M2h12.0</td>
<td>GYIIGTNKD</td>
<td>SGSNGYSDTK</td>
<td>TKSELRGFAV</td>
</tr>
<tr>
<td>M2h12.1</td>
<td>GYIIGTNKD</td>
<td>SGSNGYSDTK</td>
<td>TKSELRGFAV</td>
</tr>
<tr>
<td>M2h12.3</td>
<td>GYIIGTNKD</td>
<td>SGRQSNGDAK</td>
<td>TKSELRGFAV</td>
</tr>
<tr>
<td>M2h12.4</td>
<td>GYIISSTGD</td>
<td>SGSNGYSDTK</td>
<td>TKSELRGFAV</td>
</tr>
<tr>
<td>M2h12.5</td>
<td>GYIIGTNKD</td>
<td>SGRQSNGDAK</td>
<td>TKSELRGFAV</td>
</tr>
<tr>
<td>M2h13.0</td>
<td>GVSNSRAQG</td>
<td>LAMGST</td>
<td>GKDEVNSIAF</td>
</tr>
<tr>
<td>M2h13.3</td>
<td>GVSGLKSTA</td>
<td>LAMGST</td>
<td>GKDEVNSIAF</td>
</tr>
<tr>
<td>M2h13.4</td>
<td>GVSNSRAQG</td>
<td>LAMGST</td>
<td>SVDENRLAF</td>
</tr>
<tr>
<td>M2h14.0</td>
<td>GATTVRSRG</td>
<td>TNAEGVGKS</td>
<td>SMGDF</td>
</tr>
<tr>
<td>M2h14.4</td>
<td>GASLSRNG</td>
<td>TNAEGVGKS</td>
<td>SMGDF</td>
</tr>
<tr>
<td>M2h15.0</td>
<td>GDSVVSSTGMA</td>
<td>TQGKSVGRS</td>
<td>KMYRSWGDAGY</td>
</tr>
<tr>
<td>M2h15.1</td>
<td>GGSVSTNRMA</td>
<td>TQGKSVGRS</td>
<td>KMYRSWGDAGY</td>
</tr>
<tr>
<td>M2h15.3</td>
<td>GDSVVSSTGMA</td>
<td>TQGKSVGRS</td>
<td>AVWDSIGRGKY</td>
</tr>
<tr>
<td>M2h15.4</td>
<td>GDSVVSSTGMA</td>
<td>TQGKSVGRS</td>
<td>EMFDQAYKGGY</td>
</tr>
</tbody>
</table>
References


2. **GAMS Home Page** [http://www.gams.com]


14. Perchiacca JM, Ladiwala AR, Bhattacharya M, Tessier PM: Aggregation-resistant domain antibodies engineered with charged mutations near the


VITA
Robert J. Pantazes

Education

B.S. Chemical Engineering – The Pennsylvania State University 2008
Ph.D. Chemical Engineering – The Pennsylvania State University 2014

Awards and Honors

Personal Excellence Award for Outstanding Presentation at the Graduate Research Symposium, 2011
Leighton Riess Graduate Fellowship, 2011
Personal Excellence Award for Outstanding Candidacy Exam, 2009
Top Up Scholarship, 2008

Publications


