Antibody Loop Modeling Methods and Applications

Colleen Murrett

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2015

ABSTRACT

Antibody Loop Modeling Methods and Applications

Colleen Murrett

This thesis describes improvements to our protein loop structure prediction algorithm and use of this algorithm to inform a computational investigation of anti-HIV antibodies. First, in Section I, we outline improvements to the Protein Local Optimization Program ("Plop") that allow us to reliably restore long loops containing secondary structure elements, and predict nativelike conformations for loops whose surroundings deviate from the native crystal structure context. Shifting to focus exclusively on antibody hypervariable loop prediction, we also benchmark our results in the community-wide Second Antibody Modeling Assessment. Plop can now be reliably deployed as a tool for understanding important biological systems. In Section II, we start from a system of interest – broadly neutralizing antibodies against HIV-1 – with the long-term goal of computationally identifying more potent VRC01-class anti-HIV antibodies. We show proof of concept results for predicting relative binding affinity upon mutation using free energy perturbation (FEP) simulations for the VRC01 antibody binding to glycoprotein gp120. Using the protocols developed in Section I, we provide case studies for using Plop to understand key FEP results and guide future experiments.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am very much grateful for the support and guidance of my advisor, Prof. Richard Friesner, over the course of my graduate career. His persistence and tenacity in tackling challenging yet relevant scientific problems has greatly benefitted me as a graduate student and will stay with me in my future endeavors.

I am also grateful for the support of my committee members, Prof. Barry Honig and Prof. Ann McDermott, and my thesis committee members, Prof. Ruben Gonzalez and Prof. Larry Shapiro.

My work on Plop owes much to the guidance and contributions of Dr. Edward Miller and Steven Jerome. Dr. Kai Zhu and Dr. Ben Sellers helped me get started with antibody cases, and I am grateful for the opportunity to build off their outstanding work. I also want to acknowledge Dr. Joseph Bylund, and Dr. Dahlia Goldfeld for their assistance and input with respect to Plop.

Working on anti-HIV antibody modeling project has been a highlight of my time at Columbia, and I thank everyone who I've had the privilege to learn from and collaborate with on this work, including Profs. Barry Honig and Larry Shapiro, Dr. Tatyana Gindin, Dr. Chaim Schramm, Dr. Robert Abel, Dr. Lingle Wang, and Dr. Anthony Clark.

The Friesner Group has grown, shrunk, and grown over the past five years and I am grateful to have worked alongside Prof. Kateri DuBay, Dr. Shulu Feng, Dr. Michelle Lynn Hall, Dr. Thomas Hughes, Prof. Jianing Li, Dr. Peilin Liao, Dilek Okus, Dr. Katarina Roos, Prof. Severin Schneebeli, Andrew Weisman, Dr. Jing Zhang, and all those previously mentioned. I am particularly appreciative of the support and camaraderie of Andrew, Peilin, and Steve in the West Wing. At the end of the day, Maureen Carothers, Betty Cusack, and Calman Lobel make all of this work possible. Thank you!

At some point in the graduate school, people I once considered mere classmates became my professional scientific network. I want to acknowledge several people who have been instrumental in my own pathway from student to professional: Jieling Zhu, Dr. Tracy Y. Wang, Dr. Neena Chakrabarti, Dr. Lindsay Leone, Dr. Glen Hocky, Dr. Michelle Lynn Hall (again), Margaret Elliott, Dr. Holly Wolcott, Dr. Donald Chang, Dr. Stephen Thomas, Ahmet-Hamdi Cavusoglu, Brendan Roach, Dr. Danielle Sedbrook, Joseph Ulichny, Nathan Daly, and Colin Kinz-Thompson.

And I am extraordinarily grateful for the support of my friends and family. Special thanks to Laura; Anthony, Zhibai, and the community at St. Ignatius Loyola; Cole, Devon, and everyone at Limelight; Eryn, Paul & Jane, Chris, Mark. Even more special thanks to Mom, Dad†, and Steve.

This thesis is dedicated to

Sister Kathryn O'Brien, Mr. Sean McGuan, Mrs. Karen Adams, and Dr. Sydney Peterson,

who taught me creativity, compassion, critical thinking, mathematics, chemistry, and the sheer joy of learning: my early foundation for working on the frontiers of science.

I. Loop Prediction Methods Development

Chapter 1 : Introduction

At its simplest level, a protein is described by its primary structure: the specific amino acids, in order, that make up the protein chain, strung together like beads. A protein's role in the cell, though, is determined largely by how this chain of residues organizes in space – its secondary structure, the chain of beads crumpled into a ball. Its ability to interact with other molecules and perform various biological tasks relies on a particular arrangement of residues, and atoms within those residues. Therefore, a detailed, atomistic understanding of a protein's structure can elucidate how it goes about its job in the cell – or why it fails to complete its tasks.

Protein structures are determined primarily using x-ray crystallography, with nuclear magnetic resonance (NMR) spectroscopy catching up as an alternative approach. These experimental methods have produced reams of valuable atomic-level structural data – over 100,000 structures and counting deposited in the Protein Data Bank (PDB).

However, both x-ray crystallography and NMR are complicated, time- and reagent-consuming methods. Even for systems that readily crystallize or fall under NMR's size limitations, experimental structure determination is not a practical option for, as an example, characterizing tens or hundreds of similar systems to determine the structural effect of a few residue mutations. Further, some

systems, such as trans-membrane proteins, are quite challenging to crystallize and structurally characterize through standard methods.

Purely computational methods for protein structure determination have long been proposed as a way of supporting crystallography and closing the gap between the number of sequenced proteins and the comparatively much smaller set of protein structures. Further, computational approaches can be used to guide rational development and design of novel proteins that have not yet been synthesized, much less crystallized. Ideally, a computational structure prediction tool takes a protein sequence and returns spatial coordinates for each atom more quickly and cheaply than experimental methods, while retaining the same level of accuracy – that is, returning the same structure.

Computational options have been proposed along the spectrum from all-atom molecular dynamics (MD) simulations, which take considerable computational resources to slowly fold a protein into the presumed native conformation, to informatics-based approaches that quickly match bits of the protein's sequence to a database of known sequence-structure sets.

Homology modeling falls near the latter end of this spectrum. This technique uses sequence homologs of known structure, typically found in the PDB, to construct a new structural model under the assumption that these similar sequences will fold in similar conformations. For many proteins, considering the number and diversity of available structures, this is a well-founded assumption. Protein secondary structure largely consists of regular motifs, such as alpha helices and beta strands, which are identifiable from sequence data and amenable to this sort of pattern matching. Further, building a model – even for a large protein – from homologs is a computationally trivial task, achievable in a matter of minutes.

However, not all protein regions are as easily identifiable as an alpha helix – and in many proteins, the diverse, unsearchable regions are the most critical to the protein's unique function (or malfunction). Relying on homology modeling in such instances is unhelpful at best. Typically, the regions with low sequence homology and thus uncertain structure are loops. Protein loops, sometimes defined as regions with an absence of secondary structure elements, are nonetheless critical to the protein's overall conformation and often essential to the protein's activity – for example, whether it forms the necessary shape to bind with a drug molecule. And not only are loops less likely to share sequence identity with known structures – they are also prone to forming different conformations from similar sequences. This flexibility and structural diversity further limits homology modeling's utility in loop regions.

An independent method for loop structure prediction is therefore an important supplement to homology modeling. Because loops are relatively small and localized – by definition, they are constrained by the core of the protein – it is feasible to shift up the spectrum and deploy more computationally involved tools to model these regions.

The Protein Local Optimization Program (PLOP) is one such tool. Broadly, it builds up possible loops *ab initio* and ranks them according to energy, calculated via an physics-based all-atom force field. This ensures that a diverse set up loop conformations are sampled, whether or not they are linked to this sequence in the PDB, and implicit solvation (along with clustering possible loops) keeps the computational requirements manageable.

The following chapters will focus on PLOP's methodology and testing. In Chapter 2, the section continues with a review of PLOP, originally published in *Current Opinion in Structural Biology*, placing it into context within the broader goals of protein structure prediction and outlining recent improvements.

Next, Chapter 3 details PLOP's hierarchical sampling protocols, including methods for accurately predicting loops containing secondary structure elements, as published in *Journal of Computational and Theoretical Chemistry*. These results show that PLOP can consistently predict long and complex loops in the crystal structure context – i.e., where the loop is deleted and rebuilt without any changes to the surrounding protein residues' positioning or conformation. We then begin to evaluate PLOP's performance in more complicated environments that approach the challenges of refining loops in true homology models.

The challenges presented by loops in non-native environments come to the forefront in Chapter 4, where I introduce a new scoring term that penalizes residue pairs joined by highly unusual dihedrals. This drastically improves our predictive capabilities for loops in homology models. Here, I also begin to focus on antibody systems.

Chapter 5 reports the results from a completely blinded community-wide antibody structure prediction challenge, the Second Antibody Modeling Assessment (AMA-II). Originally published in *Proteins: Structure, Function, and Bioinformatics*, this collaborative effort with researchers at Schrodinger, Inc. tests our loop prediction methodology against the state of the art, compares loop predictions in the crystal structure context to loop predictions in homology models, and underscores the utility of *ab initio* prediction for modeling the most variable antibody loops.

Ultimately, the algorithmic improvements discussed in this section bring PLOP to prime time. Instead of selecting test cases to develop generalized protocols, we can identify biologically interesting proteins and use the now-developed protocols to build reliable homology models and analyze systems in meaningful ways. This will be demonstrated in Section II.

Chapter 2 : Computational Methods for High Resolution Prediction and Refinement of Protein Structures

In this chapter, we review progress towards reliable *ab initio* protein loop prediction and refinement to sub-Ångström accuracy[i]. Methodological improvements to our loop prediction protocol are discussed in two main areas: implicit solvation and conformational sampling.

2.1 Introduction

Prediction of protein structure to atomic resolution has been a long-standing goal of computational biophysics. For a protein with a very different sequence from that of any protein with known structure, this task is daunting and requires a large component of *ab initio* simulation. However, in the vast majority of cases, there is significant homology between the target sequence and one or more sequences where experimental structures are available, and highly successful homology modeling approaches, based on employing one or more known structures as templates, are used routinely. The Critical Assessment of protein Structure Prediction (CASP) competitions [1-2], which have been held biannually since 1994, are primarily focused on homology modeling, and document the very substantial progress that has been made, even on cases with low but detectable sequence identity.

While homology models have been very useful in a wide range of applications[3-12], in many cases these models do not yet predict atomic details at high resolution. This limits their utility in a number

[i] Reproduced with permission from *Current Opinion in Structural Biology* 2013 *23* (2), 177–184. Copyright 2013 Elsevier Ltd.

of important applications, such as structure based drug design or QM/MM computation of enzyme mechanisms. A key weakness in typical homology modeling methods is their reliance on knowledge-based scoring functions [13] and lack of rigorous treatment of the physical chemistry of protein and solvent interactions. Generally, if a target and template have extremely high sequence identity, one may obtain a high-resolution model based on the target. However, even for these best-scenario cases, differing residues can cause local regions of the model to contain nontrivial structural deviations, which can have an important impact on molecular recognition or chemical reactivity.

In this review, we focus on the question of what is necessary and sufficient to convert standard homology models into reliable high-resolution structures, a process that we refer to as structural refinement. Conceptually, the simplest approach to refinement is to run an all-atom molecular dynamics simulation using explicit solvent models [14]. Unfortunately, the timescale to rearrange all of the atoms in the protein from the homology model starting point to the native structure is quite long compared to the length of molecular dynamics trajectories accessible with current technology (typically ~1 microsecond, ~1 millisecond with a great expenditure of computational resources)[15].

The alternative to molecular dynamics is conformational search algorithms[16-20]. These methods do not attempt to reproduce the exact dynamical trajectory of the system, but rather search the phase space of possible structures by making relatively large displacements of torsion angles, followed by minimization. The necessary complement to this sort of sampling is implicit, or continuum, treatment of aqueous solvation. If an explicit representation of solvent is employed, conformational search methods will require extensive, and expensive, rearrangement of water molecules to accompany every proposed conformational change, thus negating the primary advantage of the

algorithm – the ability to make large moves that sample phase space efficiently, as opposed to the very small displacements that are possible in molecular dynamics.

The challenge for conformational search methods is thus twofold: (1) developing an implicit solvent model capable of the requisite level of accuracy, and (2) designing sampling algorithms that will converge the phase space search for the diverse range of protein structures.

Below, we will focus on the development of conformational search and continuum solvent models over the past 5 years, and the progress that has been made in the ability to refine protein structures.

2.2 Side Chain and Loop Prediction: Optimization and Testing of Models and Algorithms for Protein Structural Refinement

There are two widely used continuum solvent models that offer the promise of providing sufficient accuracy to carry out protein structural refinement: the Poisson-Boltzmann (PB) [21-23] and Generalized Born (GB) equations [24-26]. In our laboratory, we have focused on the use of the GB equation, extensively parameterized to reproduce experimental crystallographic protein structure data. Our approach is to repredict individual side chains, and then increasingly long loop regions, in the context of the native protein environment. We use the OPLS protein force field [27-29] for the molecular mechanics component of the model. The solvent model has been significantly modified in order to achieve robust agreement with the experimental data [30].

Table 2.1 presents results for charged and polar single side chain prediction using two versions (one recent, one of older vintage) of our continuum solvation model. Single side chain prediction

involves few degrees of freedom, and thus can be performed via exhaustive conformational sampling by standard algorithms [31-32]. Only side chains from high-resolution crystal structures, with significant electron density displayed for all atoms of the side chain, are used in the test set. A root mean square deviation (RMSD) from the experimental coordinates of less than 1.5Å is considered successful. The success probabilities shown represent a very substantial improvement over the use of standard GB models such as that described in refs. [25, 33-35].

		OPLS2005	OPLS2005
Residue	**# of cases**	**VSGB2.0**	**VSGB1.0**
Arg	144	84.0%	82.6%
Asn	252	91.7%	88.5%
Asp	293	94.9%	92.5%
Cys	92	100.0%	100.0%
Gln	159	83.2%	77.6%
Glu	151	86.2%	84.9%
His	83	95.2%	91.6%
Lys	121	90.1%	88.4%
Thr	316	94.3%	92.6%
Tyr	404	99.1%	98.6%
Ser	221	88.0%	86.1%
All	**2236**	**91.6%**	**89.6%**

Table 2.1 Prediction results for polar and charged protein residue side chains.

The tabulated percentages reflect the fraction of side chains where the predicted side chain geometry was within 1.5 Å root-mean-square deviation of the experimentally observed geometry**.**

Many modifications of the "standard" GB model (e.g., from ref. [25, 33]) were required to achieve these results; however, one major alteration stands out as crucial: we allow for effective polarization of protein groups by charged residues, via increasing the value of the internal dielectric constant in this type of interaction. This "variable dielectric model" [30, 36] eliminates the well-known problem of dramatic overprediction of salt bridge formation by GB and other continuum based models. Overall, it leads not only to a large improvement in the accuracy of side chain prediction, but a substantial reduction in energy errors when an incorrect side chain conformation is predicted. If the correct conformation is very close in energy to the incorrect one, the impact on overall structural prediction is going to be much less important.

We next take the model derived from optimizing results for side chain prediction and perform loop predictions for a data set of short to medium length loops (6-12 residues in length) in the native protein environment. Further optimization of the model is carried out, and again, one crucial modification emerges: the standard surface area model for hydrophobicity performs poorly in the congested protein environment.

We therefore replace this model with a hydrophobic scoring function derived from protein-ligand docking calculations, which has been optimized to reproduce binding affinities of ligands in protein receptors [9, 30, 37]. This hydrophobic model better estimates the benefit of placing hydrophobic protein side chains in the hydrophobic core of the protein, as opposed to allowing water molecules to occupy such highly unfavorable regions.

Finally, the resulting model is tested, without any further adjustment, in its ability to predict long loops (14-20 residues in length) in the native protein environment. We have described our

hierarchical loop prediction algorithm in detail in prior publications [37-39]. These energy model improvements, along with improvements in sampling to better explore loop candidates and phase space described in ref. [40], have made it possible, in the context of native proteins, to reliably predict loops up to 18 residues. In all loop predictions, the final conformation reported is of the lowest energy loop.

Table 2.2 presents results for these tests, which encompass a test set of 115 loops in total. Up to 18 residues, the RMSDs of the predictions to experiment are sub-Angstrom for the most recent version of the continuum model (VSBG 2.0). This is a remarkable advance over previous results in the literature, (including our own results with VSBGB 1.0, which was not as well optimized), and nears the limit imposed by experimental uncertainty. Results of this quality are likely sufficient for use in structure based drug design and other applications. The problem now becomes predicting loops to this level of accuracy in more complicated homology model environments.

		OPLS2005/VSGB2.0				**OPLS2005/VSGB1.0**			
Loop Length	Number of Cases	Median Backbone RMSD (Å)	Average Backbone RMSD (Å)	Average Side Chain RMSD (Å)	% of Cases with RMSD < 2Å	Median Backbone RMSD (Å)	Average Backbone RMSD (Å)	Average Side Chain RMSD (Å)	% of Cases with RMSD < 2Å
14	36	0.38	0.51	1.67	100.0	0.67	1.19	2.51	91.7
15	30	0.54	0.63	1.85	100.0	0.75	1.55	3.07	73.3
16	14	0.43	0.70	1.85	100.0	0.80	1.43	3.20	78.6
17	9	0.57	0.62	1.84	100.0	1.92	2.30	4.25	66.7
18	16	0.60	0.80	1.78	100.0	3.45	4.18	5.59	37.5
19	7	1.60	1.41	3.46	100.0	1.31	2.65	3.87	57.1
20	3	1.68	1.59	2.88	100.0	1.12	1.43	2.71	66.7
All	**115**	**0.52**	**0.69**	**1.91**	**100.0**	**1.04**	**1.89**	**3.37**	**73.0**

Table 2.2 Prediction results for 14 to 20 residue length loops.

The backbone RMSD's of the predictions are computed by first superimposing the predicted protein structure and the experimentally observed protein structure while excluding the predicted loop. The backbone RMSD values include only the C and N atoms tracing the protein backbone. The tabulated side chain RMSD values were computed including all heavy atoms of the side chains**.**

2.3 Prediction of Larger and More Complex Protein Regions

Having established the validity of the energy model in the above tests, we next explore prediction capabilities for larger and more complex protein regions. In these cases, as in homology model systems, the sampling challenges are increased in various dimensions, and the number of alternative structures that must be rejected in favor of the correct structure in many cases becomes exponentially larger.

We use the same energy model and sampling algorithms as are discussed above, with one important addition. We have noticed that many incorrect loop predictions display a pattern of backbone torsion angle pairs that is *never* observed in the Protein Data Bank (PDB). Structures containing such angle combinations cannot be correct. We have constructed a table linking regions of torsion angle space to their observed frequency in protein crystal structures, and from it built an empirical scoring function that penalizes loops built with torsion angle pairs found in the extremely sparsely populated regions. These penalties are helpful towards both sampling and scoring, by eliminating candidate loops with unacceptably high strain energy. In the work reported below, this new addition to the scoring function was used to improve results in a selected subset of challenging cases, such as those where side chains surrounding a loop-helix-loop or loop-hairpin-loop are allowed to vary.

2.4 Prediction of Loop-Helix-Loop and Loop-Hairpin-Loop regions

Many long loops contain embedded regions of secondary structure: small α-helices, 3_{10} helices, or β-hairpins. Where there are embedded helices, a modification of our usual algorithm is employed. We use sequence-based secondary structure prediction methods [40-42] to locate possible helices, and then run simulations in which these potential helices are seeded into the simulation. The seeding is achieved by using a separate, helical-dihedral library during loop buildup over the residues sequence-based secondary structure prediction assigns as helical [39]. The final energy is compared between the simulations with and without the seeded helices, and the lowest energy result is selected. For hairpin-containing loops, no special sampling is required.

For both loop-helix-loop and loop-hairpin-loops, the CPU time required for a single PLOP run is on the order of an hour on a 1 GHz AMD Opteron 265 processor. The time for a prediction is directly a function of the loop size, rather than the secondary-structure size; the rate-limiting step is the minimization of all atoms in each candidate loop. A full loop prediction requires multiple runs of PLOP, up to 400, however up to 40 can be run in parallel at once, as described in Zhu, *et al.*[36]

Table 2.3 summarizes the results obtained for a test set of 33 high-resolution loop-helix-loops, while Table 2.4 does similarly for a test set of 40 high-resolution loop-hairpin-loops, both distributed as explained therein. Consistent sub-Angstrom RMSDs are obtained in all cases of prediction in the native environment for these more challenging and diverse structures, demonstrating the reliability of both the energy model and sampling algorithm.

Helix Length	Number of Cases	Without Helix Seeding				With Helix Seeding Permitted			
		RMSD (Å)		ΔE (kcal/mol)		RMSD (Å)		ΔE (kcal/mol)	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
4	12	0.53	1.29	3.89	12.39	0.50	0.62	-1.89	-6.96
5	7	0.91	1.09	-7.94	-6.19	0.47	0.99	-7.94	-7.14
6	4	1.00	0.95	2.06	11.11	0.65	0.74	-5.80	-5.78
7	5	0.55	1.94	0.51	5.19	0.93	0.81	-2.04	-2.76
8	5	0.81	0.98	4.33	6.66	0.44	0.48	-0.86	-2.48

Table 2.3 Results of 33 loop-helix-loop predictions performed with or without helix seeding permitted.

For the predictions where helix seeding was permitted, we selected the lowest energy loop predicted across all simulations. The source of a possible helix may have come from sequence based secondary-structure prediction or from observations of helices formed in non-seeded predictions. The lowest energy loop may also have come from a non-seeded if no lower energy loops were found after seeding. With helix seeding included in our sampling methodology, sub-Ångstrom mean and median RMSDs are found for all helix lengths considered.

Hairpin Length	Number of Cases	Dipeptide Dihedral Library			
		RMSD (Å)		ΔE (kcal/mol)	
		Median	Mean	Median	Mean
6	11	0.41	1.07	-5.61	-5.05
7	2	1.13	1.13	-21.38	-21.38
8	15	0.63	0.89	-6.87	-7.53
9	7	0.51	0.89	-5.00	-5.74
10	2	0.42	0.42	-7.32	-7.32
11	1	0.53	0.53	-10.55	-10.55
12	1	0.30	0.30	-3.06	-3.06
13	1	0.44	0.44	-0.04	-0.04

Table 2.4 Results of 40 loop-hairpin-loop predictions.

The ΔE value compares the energy of the lowest energy loop against the crystal structure loop coordinates, minimized using our energy function. The RMSD reported is of the lowest energy loop prediction**.**

Figure 2.1a illustrates an example loop-helix-loop prediction. The prediction target here is a 16-residue loop containing a 6-residue helix. The prediction performed in the absence of external knowledge of the helix results in a 1.47 Å RMSD but a ΔE of 50.45 kcal/mol indicative of a large sampling error. By exploiting information provided by PSIPRED[40], a popular secondary-structure based prediction program, we are able to seed a helix and improve the sampling to reach a 0.31 Å RMSD and a ΔE of -1.37 kcal/mol.

Figure 2.1b illustrates our success in predicting loop-hairpin-loops without any special sampling. Here a 16 residue loop-hairpin-loop from PDB: 2ZWA containing an 11-residue hairpin is predicted with a 0.53 Å RMSD.

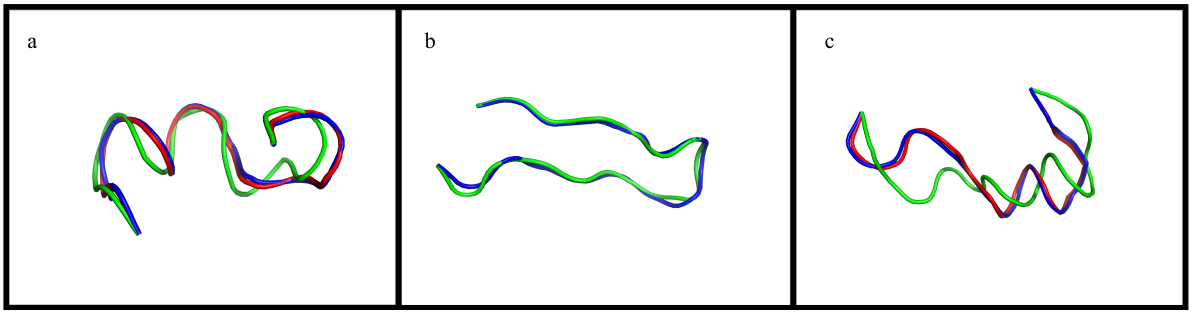

Figure 2.1 Example loop-helix-loop, loop-hairpin-loop, and perturbed native predictions.

a. The target loop-helix-loop is a 16-residue protein segment from PDB 2RJ2. The native structure is shown in blue. The prediction performed without helical seeding is shown in green and yields a 1.47 Å RMSD and a ΔE of 50.45 kcal/mol. Helical seeding performed using information obtained from PSIPRED[41], a popular sequence-based secondary structure prediction program, is shown in red. The seeded helix prediction results in a superior 0.31 Å RMSD prediction with a native-like ΔE of -1.37 kcal/mol. b. The target loop-hairpin-loop is a 16-residue loop containing an 11-residue hairpin from PDB 2ZWA. The native structure is shown in blue while the prediction is in green. The hairpin is predicted with a 0.53 Å RMSD and ΔE of -10.55 kcal/mol. c. The target is a 16-residue loop from PDB 1L5W containing a 5-residue helix. The native structure is shown in blue while the perturbed native is shown in green. The perturbed native loop-helix-loop has an RMSD of 3.00 Å. The residues within 7.5Å of this perturbed loop conformation were minimized with the loop held fixed. This placed the surrounding environment in a non-native minimum. However, for simplicity, these surrounding residues are not illustrated here. The resultant loop-helix-loop repredicted from this perturbed environment is shown in red and has a 0.54 Å RMSD a ΔE of -15.03 kcal/mol.

We then select one structure from each of the helix lengths in Table 2.3, a total of five structures, and one structure from each of the hairpin lengths in Table 2.4, a total of seven structures, and perform a more demanding test on each of them: prediction of the structure in an environment where the backbone is in the native conformation, but the surrounding side chains (up to 7.5 Å distant) are arranged around a non-native structure, selected from Loop-Helix-Loop or Loop-Hairpin-Loop predictions with RMSDs greater than or equal to 3 Å from the native structure.

Figure 2.1c illustrates an example perturbed-native loop prediction. Reprediction of the native loop under such conditions is far more demanding, since the local environment no longer serves as a "guide" to the correct structure. Nevertheless, the results are of more or less the same quality as the original predictions in the native environment. For loop-helix-loop predictions performed in a perturbed native environment, the mean (RMSD, ΔE) was (0.65 Å, -9.52 kcal/mol). These are comparable to the predictions of these same loops in the native environment where the mean (RMSD, ΔE) was (0.47 Å, -1.49 kcal/mol). For loop-hairpin-results, the results are similar with predictions in the perturbed environment being restored to a mean (RMSD, ΔE) of (0.79 Å, -4.73 kcal/mol) compared to (0.38 Å, -4.51 kcal/mol) in the native. These results are encouraging with regard to transferability of the algorithm to homology modeling.

2.5 Prediction of the ECL2 loop in GPCRs

G-Protein coupled receptors, or GPCRs, constitute one of the most important classes of pharmaceutical targets in the human genome [43-47]. Recent experimental breakthroughs have resulted in a substantial number of GPCR structures being available in the PDB. However, this is still a small fraction of the total number of GPCRs in the genome, and high-resolution homology modeling of the additional structures, with particular emphasis on accurate prediction of the second extracellular loop (ECL2), would be extremely valuable in drug discovery efforts [44-45, 48].

As a first step, we have predicted the structure of the ECL2 loop in a number of GPCRs in the PDB in the context of the native structure [49]. For the four GPCRs included in recent publications, the ECL2 loop is extremely long (26-32 residues) and has a highly complex structure, in some cases with

an embedded helix or strand. The result of these efforts is shown in Figure 2.2. Considering the challenge presented by this loop, the fully *ab initio* predictions are very reasonable. We have also performed one prediction of the ECL2 of a homology model of β2AR, obtaining a similar quality result: a very encouraging step in overall methodologies [50].

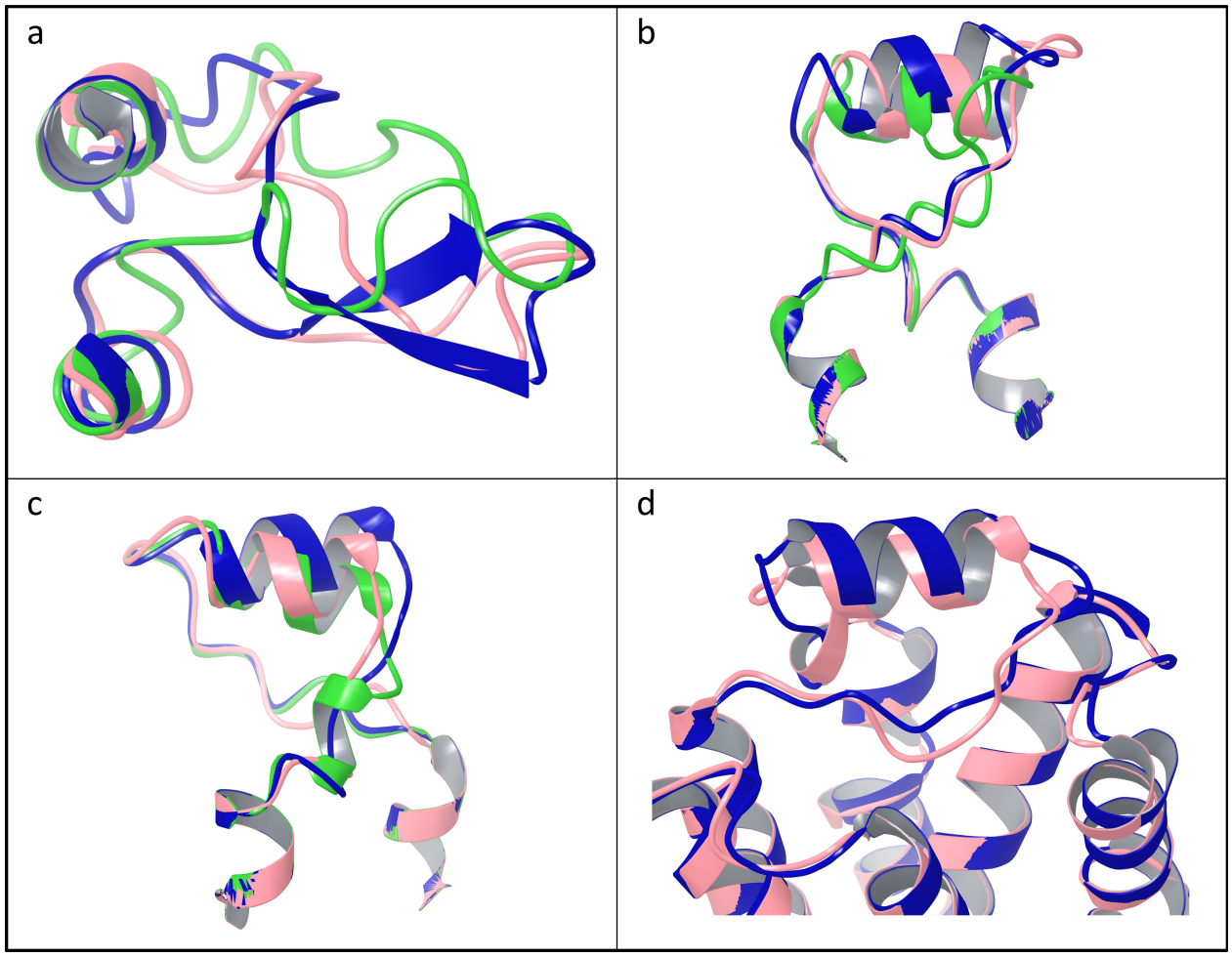

Figure 2.2 Visualization of predicted ECL2s vs. their native counterpart.

a. The ECL2 of bRh. The blue loop is the native, the pink loop is the final ECL2 prediction while the rest of the protein is held fixed in its crystallographic position. The green loop is the final ECL2 prediction while the nearby loops and side chains are in non-native positions. Note that for these loop predictions, a simulated membrane was included (see ref. 48). b. The ECL2 of β1AR. The blue loop is the native, the pink loop is the final ECL2 prediction while the rest of the protein is held fixed in its crystallographic position. The green loop is the final ECL2 prediction while the nearby

loops and side chains are in non-native positions. c. The ECL2 of β2AR. The blue loop is the native, the pink loop is the final ECL2 prediction while the rest of the protein is held fixed in its crystallographic position. The green loop is the final ECL2 prediction while the nearby loops and side chains are in non-native positions. d. Again, the ECL2 of β2AR. The blue loop is the native, the pink loop is the final ECL2 prediction in the context of a homology model of β2AR.

In order for these calculations to succeed, we had to incorporate protein membrane interactions via new explicit membrane calculations. Briefly, molecular dynamic simulations were run with explicit membrane molecules surrounding the receptor. Similarity between the MD structure and the native target loop regions justified the use of loop predictions performed on the MD structure. During these loop predictions, up to three key torsional bonds of the rotating lipid heads were sampled together with all surrounding side chains within 7.5 Å of the loop. These calculations are explained in greater detail in Goldfeld *et al.* [50,51].

2.6 Conclusion

We have shown that the use of a continuum solvation model and a molecular mechanics force field, along with an efficient conformational sampling algorithm, is capable of yielding accurate and reliable predictions for relatively large protein regions (up to as many as 30 residues). The refinement of homology models differs from these problems in that the entire backbone of the protein will present some sort of deviation from the native coordinates - large in some cases, small in others, but overall representing delocalization of the structural errors, as opposed to the model problems we have studied where the error is localized to one well defined protein region. Based on the results above, the primary difficulty at this point is to devise a sampling algorithm that can handle delocalized errors of this type. There are many possibilities that build on the ability to efficiently refine quite large local regions, but extensive computational experiments will be required to identify effective methods, and to optimize the methodology for what will likely be a greatly expanded need for computational resources. The continued rapid reduction in the cost/performance of computing provides the means to meet this aspect of the challenge, however, and we are optimistic that practical solutions, from other groups as well as our own, will begin to appear in the next several years.

Chapter 3 : Prediction of Long Loops with Embedded Secondary Structure using the Protein Local Optimization Program

In this chapter, we detail improvements to the Protein Local Optimization Program (PLOP)[ii]. This work builds on previous development efforts for sampling increasingly long loops and extends PLOP's capabilities to accurately model loops containing small secondary structure elements such as alpha helices or hairpins. The results presented here demonstrate PLOP's effectiveness at restoring native loop conformations in the crystal structure context, a prerequisite for accurately loop prediction in a homology model or other non-native environment.

3.1 Introduction

Continual advances in loop prediction have yielded accurate modeling from twelve-residue loops[51] up to loops as long as twenty residues[30, 37]. These methods have managed to achieve near-atomic accuracy performing loop prediction in the presence of the crystal structure environment – a necessary, but not sufficient condition for realistic homology modeling.

Historically, loop prediction was first approached analytically by Go and Scheraga[52] in 1970. Demonstrated was the ability to predict, by solving a set of equations, the conformation of peptide fragments containing up to six rotatable torsions. This analytical method was updated 21 years later by Palmer and Scheraga[53]. Here, the authors relax constraints on the original formulation by permitting each residue in the loop to adopt independent bond lengths or bond angles. However, the analytical method still remained limited to six torsion angles - three residues assuming the

[ii] Reproduced with permission from *Journal of Chemical Theory and Computation* 2013 *9* (2), 1846–1864. Copyright 2013 American Chemical Society.

backbone ω torsion remained fixed. To accommodate larger loops, Palmer and Scheraga extend the method by permitting additional torsions, beyond the six that can be analytically determined, so long as they are independently set prior to the calculations. Thus, their method requires that the algorithm be repeated numerous times over a conformational search of these additional independent torsions. Hence, for larger loops combinatorics must be considered.

Moult and James in 1986 proposed one of the first combinatorial searches through a discrete set of torsions[54]. Here, the authors described the use of a systematic search through torsion angles obtained from a Ramachandran plot. For loops as small as five residues, their method yields about 10^{10} conformations, already an intractable number. To cope with the combinatorial explosion the authors, employ the use of rules and filters to restrict and prune the number of conformations to a manageable subset before performing more expensive scoring. Loops are scored with using a simple pairwise electrostatic energy function and a surface area based hydrophobic term.

Later methods vary in both the sampling rules and scoring function. Bruccoleri and Karplus in 1987 released CONGEN, from which our algorithm draws some similarity[19, 51]. There the authors use the CHARMM energy function[55] to score loops. In 1992, Bassolino-Kilmas and Bruccoleri advance CONGEN to permit directed loop buildup which takes into account information from partially built structures[56]. In 2003, DePristo et al.[57] and de Bakker et al.[58] use the AMBER forcefield[59] and Generalized Born solvation model[25, 33] for scoring loops. Loop buildup is performed using, among other modifications, a fine-grained torsion library that is residue-specific. Like CONGEN, our work draws similarities to this last method[51]. We note that this historical review is not exhaustive but is intended to highlight the origins of loop prediction as it relates to this work.

In general, the use of combinational exploration of torsion space for loop buildup has within it two sub-problems, sampling problems where coping with the combinatorics of loop buildup requires the development of clever pruning strategies, and energy problems where the minimization, scoring and ranking of the resultant loops must be computationally affordable yet accurate enough to identify the best conformation among those produced.

Throughout the literature, the functional definition of a loop has been a local segment of the protein that is free of secondary structure other than, perhaps, three-residue 3_{10} helices, but lies between large, likely well-conserved, secondary structure elements[37, 60]. Indeed, initial homology models are often constructed on the assumption that secondary structure elements are conserved between the template and the target[61]. However, this loop definition has not always been strictly followed. Notable cases of loops containing secondary structure are the ECL2 loops of human β2-adrenergic receptor[62] and turkey β1-adrenergic receptor[63], both G-protein coupled receptors (GPCRs). These loops are actually loop-helix-loops (LHLs) containing an eight-residue α-helix. Spinach Rubisco is another example. The active site is composed of a highly conserved α/β barrel. Lying between each α/β pair are loops, of which loop 5 contains a five-residue α-helix and two residues that form part of βF, a β-strand external to the active site, and loop 8 which contains a four-residue α-helix[64].

Recent attempts have been made to model the GPCR LHLs and have been met with significant success reaching an accuracy as high as a 1.59 Å RMSD[49, 65]. As the method we provide here exists along a continuum of protein structure prediction methods, one that shares significant applicability to secondary structure-free loops, we retain the loose definition of the word 'loops', and here refer to loops as a region of the protein that may contain secondary structure but is flanked by even larger

secondary structure elements. Presented in greater detail below is a precise definition, which was strictly enforced, to select a set of test of cases.

Throughout the literature, predictions performed on loops containing secondary structure are scant. Zhu, Xie and Honig presented a refinement protocol that addresses loop-helix-loops and loop-hairpin-loops, referred to more generally as protein segments in the chapter, using a knowledge-based potential[66]. What is explored is the refinement of these segments, rather than the prediction of the segments *de novo*. Consequently, the success of their refinement is dependent on the difficulty of the initial structure. For hairpins and loop-helix-loops, close to 70% of their refinements yield predictions with an RMSD of 2.0 Å or better. In these cases, the secondary structure elements are kept fixed with their native torsions and moved as a rigid body. However, as our method discussed in this chapter is independent of the conformation of the input loop (although it is dependent on the conformation of the surrounding environment) results cannot be directly compared.

Alternatively, Rohl *et al.*, described *de novo* loop construction using the Rosetta algorithm[16]. Included in their test set are predictions of ten loops, referred to as structurally variable regions, of 13 to 34 residues in length. These predictions were done in the crystal structure environment and do include loops containing secondary structure. Although some of the members of their test set include, for example, loop-helix-loops, only ten cases were done in the context of the native protein – too few to permit comparisons between our method without relying on anecdotal information. Instead, the authors concentrate on the more ambitious task of loop prediction in an unrefined homology model. Finally, we note in a previous study, our attempt to address the challenges of helix packing[67]. In Li *et al.,* we explored placement of a helix in a loop-helix-loop but treated the helix as a rigid body. Although the method relies on prior knowledge of the presence of a helix, for large helices, this is

not unreasonable, as is stated above, because significant segments of secondary structure tend to be conserved across homologous structures. Indeed, the smallest helix considered in this study was eight-residues.

To the best of our knowledge, no studies have been performed that systematically address the challenges of *de novo* prediction of loops containing secondary structure, particularly for cases when *a priori* knowledge about the presence of small secondary structure is noisy at best. As loop prediction matures to accurate prediction of larger and larger loops, it becomes awkward to exclude cases of secondary structure-embedded loops. In this work, we propose a method to predict long loops containing possibly multiple helices or a hairpin. Our initial test set is composed of loops containing between 8 and 17 residues. The secondary structure length explored ranges from 3 to 13 residues, although in principle, prediction of loops containing larger secondary structure segments remains tractable.

For loop-helix-loops, we constructed a separate dihedral library taken from a non-redundant set of high-resolution Protein Data Bank[68] structures containing α-helices. The user is required to specify which residues this helical dihedral library is to be applied to, termed the helical bounds. Results with exact helical bounds taken from the crystal structure were used as an initial validation. More relevant to actual structure prediction and refinement, we then concentrated on accurate loop prediction using helical bounds supplied by either sequence-based secondary structure prediction algorithms or previous loop predictions performed without the use of our helical dihedral library. That is, in many cases, nascent helices were predicted without supplying any expectation of a helix. This suggested a propensity for this loop to include a helix and allow us to repredict the loop using our helical dihedral library. Throughout all sampling methods explored, what remains crucial is that purely from

our energy model, we are able to pick out the loop with the lowest, or near lowest RMSD relative to the native structure. Finally, for loops containing either helices or hairpins, we explored loop reprediction in a perturbed local environment, similar to an environment encountered in full homology models, although without deviations of the backbone from the native structure, and established success in restoring the native loop conformation. The results are generally satisfactory with loop-helix-loop predictions from imprecise helical bounds routinely reaching sub-Ångström RMSD and hairpin predictions reaching similar atomic accuracy.

3.2 Materials and Methods

Selection of Test Cases

All PDB structures that were available as of August 30, 2010 were searched. A global criteria was used to select structures that satisfy the following properties:

1. A sequence identity between any two proteins must be ≤ 50%
2. Only crystal structures were selected
3. The resolution of the crystal structure must be < 2.0Å
4. Structures reporting only Cα coordinates were excluded
5. A minimum R_{work} of 0.25 was enforced.
6. The pH of the crystal structure was restricted to lie between 6.0 and 8.0.
7. The exclusion of proteins due to sequence identity was performed using the PISCES web server[69] (http://dunbrack.fccc.edu/PISCES.php). Loops were selected using a local criterion that satisfies the following:
8. The average temperature factor of atoms within the loop must be ≤ 35.

9. The real-space R-factor[70] of any residues in a selected target loop must not be greater than 0.200.
10. All residues within the loop or interacting with any residues within the loop must be free of alternate conformations.
11. To reduce effects due to loop-ligand interactions, the minimum distance between any loop atom and any atom as part of a neutral ligand must be > 4 Å. For charged ligands, this cutoff is increased to 6.5 Å.

The real-space R-factor was found by reference to the Uppsala Electron Density Server[71] (http://eds.bmc.uu.se/eds/). The above criteria are similar to what was used to create test sets in our past publications[36-37].

Identification of Secondary Structure-Containing Loops

In our most recent publications, loops were defined as being a segment of the protein absent of secondary structure[37, 66]. To identify loops containing secondary structure, an alternative definition was proposed. For loops containing secondary structure, the loop must be bounded by a span of secondary-structure larger than the greatest contiguous span of secondary structure within the loop. For example, if a loop contained, at most, a six-residue α-helix, then flanking the loop must be residues that are a part of a secondary structure element of at least seven residues in length. Furthermore, the first and last residue of a loop must also not display secondary structure. Assignment of secondary structure on a per residue basis was done using the DSSP program[72].

A loop was defined as a loop-helix-loop only if there were no other types of secondary structure present other than turns and helices (including 3_{10} and αα-helices), i.e. any loop containing both β-

bridges and helical residues was discarded from this study. A total of 35 loop-helix-loop regions were identified which were either 16 or 17 residues in length in all. This loop length was chosen to select cases that were considered sufficiently difficult to demonstrate the efficacy of our approach. In our previous publication, loops free of secondary-structure were successfully predicted up 17 residues in length[37].

For loops containing β-hairpins, it became necessary to distinguish between a β-hairpin and a segment that is part of a larger β-sheet. To make such a distinction, the following criteria were used:

1. The loop must contain the secondary structure pattern strand-turn-strand.
2. However, the turn residues need not be immediately adjacent to a strand residue.
3. The loop must be free of helices.
4. The strand residues comprising part of the pattern in criterion 1 must be forming backbone hydrogen bonds only to other residues within the loop.
5. The hydrogen-bonding pattern must be anti-parallel.

For hairpins, requiring loops be either 16 or 17 residues in length yielded too few test cases. Thus, a loop was accepted so long as it was not greater than 17 residues. A total of 41 cases satisfying the above hairpin criteria were identified.

Single-Loop Prediction

Single loop prediction is performed through individual runs of the Protein Local Optimization Program (PLOP). Briefly, PLOP operates through four stages: buildup, closure, clustering, and scoring. Full details can be found in Jacobson *et al.*[51], however, the salient features will be presented here and the modifications of the PLOP protocol utilized in this work will be described.

Loop buildup is begun with a backbone dihedral angle library constructed from rotamers frequently observed in crystal structures. Initially, the library contained a set of dihedrals on a single amino acid basis[51]. As larger loops were explored, efficient exploration of conformational space dictated the use of a dipeptide dihedral library[30, 37]. In this approach, a library is constructed from each of the 400 (20 x 20) possible dipeptide pairs and used in a sequence specific manner during buildup. For example, a loop containing an arginine–alanine dipeptide would explore sampling from a different rotamer library than an arginine–valine dipeptide. This implicitly treats the individual amino acid torsions as coupled.

In helices, the backbone torsions are highly coupled to form the necessary hydrogen-bonding network. It was therefore natural to extend the use of a dipeptide dihedral library to exploit coupled backbone torsions across the four residues, or greater, of an α-helix. As such, for residues considered to be helical, a separate n-residue α-helical library was used for loop buildup, where n is four or larger. The aspects of this α-helical library are discussed in greater detail below. In β-hairpins, non-local torsional coupling is present and so to enforce torsional coupling during loop buildup would heavily constrain both the coupled, hydrogen-bonding residues, as well as the intervening turn residues. Although such an approach may still be fruitful, we found that for β-hairpins, our previous dipeptide torsional library was effective and so we did not explore further the use of an alternative β-hairpin library.

Loop buildup is performed simultaneously from both ends of the loop up to the C_α atom on the closure residue. In our prior publications, the closure residue was simply picked as the midpoint of the loop[36-37, 51]. For the loop-helix-loops described in this work, the closure residue, shared by both halves of a loop, cannot be permitted to bisect a helix. As is described further below, the helical

library is based on the construction of entire helices, and not helical fragments. If the closure residue of the loop were a part of a helix, the helix would be split between both halves of the loop. Thus for this work, we were forced to alter the designation of the closure residue. The closure residue is initially set with the equation

$$C_{a,closure} = N_{term,LHL} + \left(Length_{LHL} - 1 \pm Length_{Helix} \right) / 2$$

where + is used when the C-terminus loop is the longer loop and − for when the N-terminus loop is longer or if both flanking loops are of equal length. $N_{term,LHL}$ refers to the residue number of the N-terminus of the loop-helix-loop. Should the closure lie adjacent to the helix, the closure residue is shifted one residue further away from the helix. This is to afford extra flexibility to the residues that precede loop closure.

Clarifying by example, consider the LHL predicted in PDB 1BKR (Figure 1). Predicted was the 17-residue loop-helix-loop from G75 – D91 containing a 4-residue alpha helix from P82 to I85. When predicting this loop without the helical library the closure residue is at the midpoint of the LHL, residue 83, highlighted in white in Figure 3.1. This residue intersects the helix and so cannot serve as the closure residue when employing the helical dihedral library from segments 82-85. Application of the above equation places the closure residue adjacent to the helix at residue D81, but for further flexibility, the closure residue is assigned to be residue L80 on the N-terminus loop, two residues away from the start of the helix. As in our previous work, the Cartesian positions of the two closure C_α atoms are averaged and the remaining atoms of the loop backbone are generated using standard geometry algorithms to close the loop.

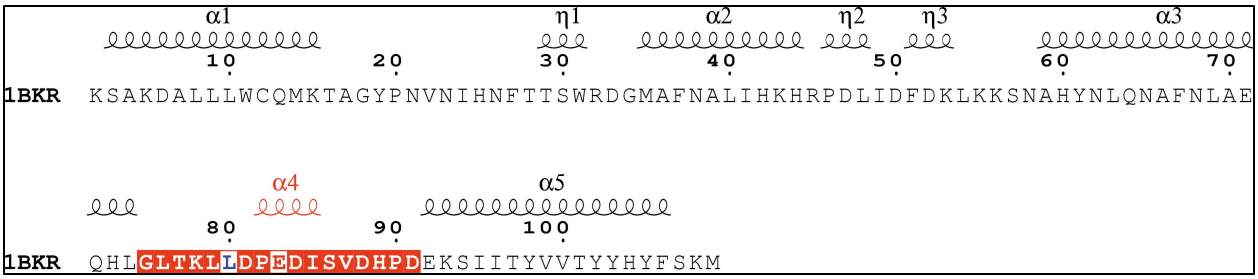

Figure 3.1 Loop-helix-loop predicted in PDB 1BKR.

The target loop-helix-loop residues are highlighted red from residues 75–32. The helix of interest, labeled α4, spans residues 82–85. . Loop prediction without the helical library would assign the closure residue to be residue 83, highlighted in white. The LHL method places the closure residue at position 80. This figure was generated using ESPript [28].

During loop buildup, nascent loops undergo preliminary screening through the use of a parameter termed the overlap factor (*ofac*). The *ofac* is defined as the ratio of the distance between two atom centers to the sum of their atomic radii. A lower *ofac* cutoff allows for a higher overlap between the van der Waals radii. If during loop buildup, a backbone atom is placed with a smaller *ofac* than permitted by the threshold, then that candidate loop is discarded.

Three additional screens are used to reject unreasonable loops early in their construction:

1. For the current residue(s) being predicted, there must exist at least one acceptable side-chain conformation, based on sampling a 30° side-chain rotamer library.
2. The loop must not travel further than 6.32 Å away from every C_α atom in the protein. This is an empirically determined value and is meant to reject loops that fail to form contacts with the rest of the protein.
3. The distance between the latest residue predicted and the closure residue must be less than a threshold beyond which closure is not considered possible. For example, a statistical analysis of a set of >500 proteins found that the maximum C_α - C_α distance that can be spanned by four residues is 13.97 Å.

Full details of these screening methods are given in Jacobson *et al.*[51]

An additional screening method is also employed to enforce broad sampling of conformational space. During loop buildup via single dihedrals, all pairs of states must obey the relationship $\Delta\phi^2 + \Delta\psi^2 > R_{eff}^2$ $\Delta\phi^2 + \Delta\psi^2 > R_{eff}^2$, where R_{eff} is the "effective resolution" of (ϕ, ψ) space. The effective resolution is adaptively set during loop buildup. The total number of loop candidates is constrained to lie between a minimum of 512 loops up to a maximum of 10^6 loops. This constrains the number of loop candidates to a tractable size. We achieve this by initially setting the effective resolution to a coarse value of 300º and then gradually improve the resolution to finer values down to a minimum of 5º (the resolution limit of the dihedral library). For loop buildup using the dipeptide dihedral library, the effective resolution relationship becomes:

$$\Delta\phi_1^2 + \Delta\psi_1^2 + \Delta\omega^2 + \Delta\phi_2^2 + \Delta\psi_2^2 > R_{eff}^2 \; \Delta\phi_1^2 + \Delta\psi_1^2 + \Delta\omega^2 + \Delta\phi_2^2 + \Delta\psi_2^2 > R_{eff}^2.$$

Loop buildup using the helical dihedral library did not utilize any effective resolution relationship. Principally, this was because the size of the helical dihedral library is significantly smaller than the single peptide or dipeptide dihedral library. Due to a "lever effect", a small change in the dihedrals at one end of a helix can significantly alter the coordinates of the opposite end of the helix. This effect becomes more dramatic for larger helices. To exclude what few candidate loops are produced during buildup because of a resolution cutoff would be to ignore this lever effect. Greater detail about the construction and composition of the helix dihedral library is presented below.

To prevent expensive optimization of similar loop candidates, the k-means clustering algorithm[73-74] is employed and only one representative loop per cluster is passed onto side chain sampling and

optimization. The number of clusters is set to be four times the number of residues in a loop, excluding residues initially flagged as helical during input to loop prediction, up to a preset maximum of 50 clusters. The number of clusters determines the number of representative loops passed onto side chain sampling/loop optimization and is empirically set to balance the conformational space that must be accurately scored against computational expense. Since the entire helix is constructed as a whole from the helical library, it would seem awkward to count the helical residues the same as the non-helical ones and so helical residues are excluded when determining the number of clusters to optimize. For the loops described in this chapter, this often had little consequence. For a 17-residue loop with a four-residue helix the maximum number of clusters, set at 50, is reached. The most common helical size was four residues (see Figure 3.2, below). For a 16-residue loop with a four-residue helix, the number of clusters is 48. Only for the few cases, such as PDB 2JA2, where a 16-residue loop contains an eight-residue helix, were the number of clusters, set to 32, significantly different from the maximum value of 50. These cases are the exception, and as is described later, the results from these cases, despite the reduced number of clusters, were excellent.

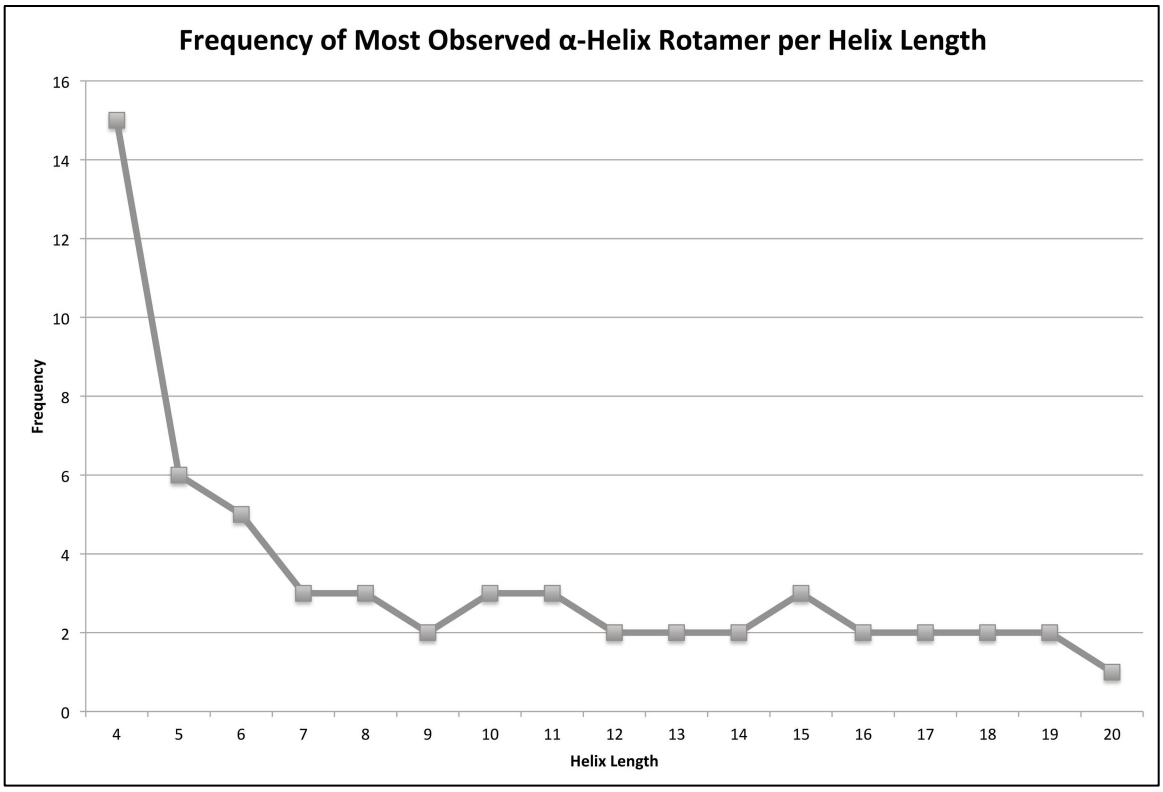

Figure 3.2 Plot of the frequency observed of an α-helix rotamer per helix length.

After a six-residue α-helix, rotamers were only observed no more frequently than three times.

Side chain sampling is performed using a 10º-resolution rotamer library constructed by Xiang and Honig[75]. The algorithm for side-chain optimization works by initially placing side-chains in a random rotamer state onto the backbone. Self-consistent optimization is then performed where all side-chains but one are held fixed while the free side chain is minimized. With the exception of loop prediction in a perturbed native environment, the default of one round of side-chain randomization per entire loop minimization was found sufficient. When considering perturbed native environments, where the surrounding side chains are included in refinement, additional rounds of side-chain randomization/self-consistent optimization is performed separately to compare to predictions done without this extra sampling. The lowest energy side-chain rotamers are selected across any additional rounds of side-chain randomization. After self-consistent side chain rotamers

are selected, the complete loop, with both side chains and backbone atoms, is then energy minimized. Full details about side-chain optimization are described in our past publications[51, 76].

Scoring is done using an augmented form of the Optimized Potential for Liquid Simulations (OPLS) all-atom force field[28-29, 76]. For solvation, an implicit model was used based on the Surface Generalized Born model as described initially in Ghosh *et al.*[31] A variable dielectric approach is used to treat polarization from protein side chains[34]. Additional corrections were added to the energy model to better account for π-π interactions, self-contact interactions, and hydrophobic interactions. The force field, solvation model, and all correction terms are discussed in greater detail in Li *et al.*[30] The protonation state of all titratable residues was set using the Independent Cluster Decomposition Algorithm of Li *et al.*[77]

Since we evaluate our loop prediction method against published crystal structures, crystal-packing effects were taken into consideration. The crystallographic asymmetric unit, as well as all atoms from other surrounding unit cells that are within 30 Å, are included in the simulation. The coordinates of all copies of the asymmetric units are updated for steric clash checking and energy calculation throughout the course of the loop prediction.

Construction of the helical dihedral library

As a natural extension to the dipeptide dihedral library, we constructed a helical dihedral library to exploit the coupled torsions present in an α-helix. An initial set of PDB structures was obtained from the precompiled culled PDB lists from the PISCES web server[69]. The parameters used to cull the structures were a percentage identity cutoff of 30%, a resolution cutoff of 2.0 Å or better, and an R-factor cutoff of 0.25. The PDB list was obtained on October 16, 2007. The list contained 3900 PDB structures. Using an internal PLOP implementation of the DSSP algorithm[72], α-helices were

identified with lengths ranging from four to twenty residues. The ϕ,ψ angles for the helical residues were extracted. We ignored values for the ω dihedral and instead used 180° during loop buildup. Deviations from the *trans* conformation are permitted during loop minimization. The dihedral angles were rounded and binned to a 10° resolution. The frequency of each binned helical rotamer was counted per helix length. In structures containing homomultimeric proteins, the helix was only counted once. We did not include helical fragments from larger helices as part of the set of dihedrals for smaller helices. That is, the torsions in a 6-residue α-helix are kept separate from the torsions in a 4-residue α-helix. This adherence to the use of only complete helices was rigidly followed throughout loop prediction. Specifically, loop buildup from both ends of the loop was done such that the helix was not divided between both loop halves. When predicting a subsection of a loop, as is done during hierarchical loop prediction, in any instance where a subsection of the helix was predicted, the dipeptide dihedral library from Zhao *et al.*[37] was used instead.

Initially, we sought to include all rotamers observed with a frequency above a set cutoff. However, this approach was problematic. Despite the large number of PDB structures, for large helices, many rotamer sets do not appear more than once. For example, in a 9-residue helix containing 18 dihedral angles (ϕ,ψ), a single 10° difference in any ϕ,ψ angle would place that rotamer in a new bin. For helices of this length, a helical rotamer was not observed with a frequency greater than twice (Figure 3.2). Beyond a six-residue α-helix, rotamers were observed no more frequently than three times. We therefore felt that there was no suitable frequency cutoff to use. Ultimately, we arbitrarily decided to set the library to contain $2\times Length_{Helix}$ rotamers and populated the library with the most frequent rotamers that conformed closest to ideal helical dihedral angles of $(\phi,\psi)=(-60^o,-40^o)$. Any non-

ideality in a helix was left to be predicted during loop minimization and the multiple stages of loop refinement described in the following section.

Hierarchical Loop Prediction

Hierarchical Loop Prediction was first described by Jacobson *et al.*[51] in 2004 and then expanded by Zhu *et al.*[36] in 2006. In short, multiple runs of PLOP are performed where increasing constraints are applied to subsequent rounds of loop predictions. The lowest energy loops from each PLOP run are passed onto subsequent, constrained rounds of refinement. The lowest energy loop across all PLOP runs and all constraints is considered the final structure.

Hierarchical loop prediction is begun with an initial set of candidate loops that are predicted by running PLOP at discrete values of the overlap factor (*ofac*). In this work, we permitted the *ofac* to vary from 0.3 to 0.7 in increments of 0.05. The best 15 loops, in terms of energy, are passed onto a *Ref* stage. A *Ref* stage constrains the C_α atoms of any new prediction to lie within a set radius of the C_α coordinates of the previous stage. In this case, the *Ref1* stage used a 4 Å radius. The best 20 loops from this stage are passed onto a *Fix-n* stage. In a *Fix-n* stage, we repredict a subset of the original target loop but use the output from a previous stage as the scaffold, holding a total of *n* terminal residues fixed. For example, in a *Fix3* stage, we hold three terminal residues fixed, and repredict the interior loop residues that remain. There are a total of four possible ways to fix three terminal residues:

1. Fix three N-terminal residues
2. Fix three C-terminal residues
3. Fix two N-terminal residues and one C-terminal residue
4. Fix one N-terminal residue and two C-terminal residues

All four possibilities are explored when selecting the lowest energy loop from the *Fix3* stage. In general, there are $n + 1$ possible combinations for a given *Fix-n* stage. We ran a total of eight *Fix* stages from *Fix1* to *Fix8*. The *Fix1* stage passed the top 10 loops onto *Fix2*. Each subsequent *Fix* stage passed one less loop onto a subsequent stage so that the *Fix8* stage passed only the top three predictions. Finally, a second *Ref* stage is run, *Ref2*, where a 6 Å C_α constraint is used. In total, taking into account all permutations in the *Fix* stages as well as the *Init* stage and *Ref* stages, there is a minimum of 334 PLOP runs per hierarchical loop prediction. The minimum number of PLOP runs can be exceeded by adaptively varying the *ofac* during hierarchical loop prediction, described in greater detail below.

To accommodate our helical dihedral library, we modified hierarchical loop prediction method in two ways:

1. The generation of our helical library was based on complete helices. To be precise, the helical library for four-residue helices is taken only from the coordinates of helices that are exactly four residues. We do not include in our four residue helical library segments of, for example, an eight-residue helix spanning four residues in length. As such, we do not construct our loops using a separate set of "partial" secondary structural elements. As a result of this, *Fix* stages that would constrain part of a helix, instead revert to using our general dihedral library for the individual PLOP run.
2. The use of a helical library also resulted in a large number of individual PLOP runs that failed to produce any candidate helices. This can happen under normal circumstances, say, during a late *Fix* stage where the majority of the loop is kept constrained and only a small subset of the loop is resampled. Loop construction in these late *Fix* stages requires the

residue buildup to occur without violating our *ofac* criterion despite being in an environment made all the more crowded by the unconstrained segments of the loop. This problem becomes compounded when working with a helical library. Since loop buildup with a helical library appends the helix onto a nascent loop in a single step, a slight displacement of the preceding residue leads to a large displacement of the terminal end of the helix – a sort of lever effect. If this crude displacement of the terminal residue of a helix places the loop in a steric clash with the surrounding environment, the loop candidate could be rejected due to the *ofac* criterion. In these cases, the outcome of a loop prediction becomes all the more sensitive to the *ofac* parameter. To further decouple the effect the *ofac* has on a successful loop prediction, any individual PLOP run beyond the *Init* stage that fails to succeed past loop buildup is automatically rerun with a lower *ofac* down to the lowest *ofac* sampled during the *Init* stage. In a PLOP run, the rate-limiting factor is during side chain optimization/minimization, rather than during loop buildup. Restarting a PLOP job after a failed buildup stage is on an order of magnitude of one minute. Since this procedural augmentation can apply to loop-helix-loops as much as it can to other loops, this improved sampling adjustment was applied to all cases studied in this work, regardless of the dihedral library used.

Calculation of RMSD

The success of loop prediction was gauged by using the backbone RMSD calculated against the native, crystal structure conformation of the loop. RMSD was calculated by superimposing the protein backbone, excluding the loop, and using the N, C_α, and C coordinates of the loop to compute the deviations. Unless otherwise stated, we report the RMSD for the lowest energy predicted loop.

Calculation of the relative energy

Similar to RMSD, at the conclusion of complete hierarchical loop prediction, we report the relative energy of our predicted structure against the energy of the minimized native. This relative energy is defined as $\Delta E = E_{prediction} - E_{native}$. A final structure that has a poor RMSD but a calculated energy that is erroneously superior to the native would thus have a negative ΔE and would indicate a failure of our energy model. Minimization of the target for comparison against predictions is necessary to permit a fair comparison between structures but is particularly important when comparing to crystal structures as the PDB structures obtained have, in all the structures examined in this chapter, no explicit hydrogen atoms. The minimization of the native was performed similarly to minimization/optimization of candidate loop structures as described above in the Single-Loop Prediction subsection of the methods. For the native, the target loop is first minimized followed by side chain sampling using the protocol described above in the Single-Loop Prediction section. For predictions done in a perturbed native environment, ΔE reports are still against the energy of the minimized native. For these cases, all additional surrounding residues that are included in the prediction are also minimized in the native to permit an accurate comparison. In instances when we used additional rounds of side chain sampling, the native loop, during minimization, was also permitted identical number of additional side chain sampling.

Sequence based secondary structure prediction

Loop prediction using the helical dihedral library requires the user to provide a range of loop residues, known as the helical bounds, over which to apply this library. To serve as an initial test of our method without the complication of uncertainty in the existence and size of a helix, we predicted loop-helix-loops from previously published crystal structures. In these experiments, the helical bounds were known *a priori*. After we had observed success using exact helical bounds, we

tested the robustness of this method in a more realistic setting where the helical bounds were supplied by popular sequence-based secondary structure prediction software. Specifically, we ran local copies of the secondary structure prediction packages SSPro4[41, 78] and PSIPRED[40]. The output of either of these programs is a secondary structure assignment across each of the residues contained in the protein chain of interest. We examined the secondary structure assignments only for the residues that spanned our particular loops. Often times, these assignments labeled more than one set of intra-loop residues as helical. In particular, the loops discussed in this chapter are sometimes bounded by larger helices and these secondary structure assignment algorithms had occasionally assigned the terminal residues of the loop to be a part of that larger flanking helix. In other cases, three, two or even a single intra-loop residue was assigned as helical. As the loop-helix-loop prediction method described in this chapter is intended for α-helices (helices of four residues or larger), assigning less than four residues as helical is not useful for our purposes. Thus, for simplicity, the largest intra-loop helical segment predicted by SSPro4 or PSIPRED, spanning at least four residues, was used as the inputted helical bounds. When both PSIPRED and SSPro4 offered useable helical bounds, we performed loop prediction with both bounds separately and compared the results.

Loop prediction in an inexact environment

Unless otherwise noted, all loop predictions in this work were done by deleting the loop residues but leaving all surrounding side chains intact, thereby preserving the crystal structure environment. In an actual homology modeling experiment, the surrounding side chains are unlikely to be placed *a priori* in their correct native conformation. To test the effectiveness of our method in refining loops in an inexact environment, we followed the approach of Sellers *et al.*[38] to perturb the surrounding side chains to a reasonable but non-native conformation. To do this, we ran multiple rounds of PLOP to

predict the loop of interest in the crystal structure and selected a loop with a backbone RMSD of no better than 3 Å. A list of surrounding residues is obtained by noting all residues that are within 7.5 Å of any candidate predicted loop, not just the one loop with a 3 Å RMSD. The union of the side chains from the surrounding residue list as well as the loop side chains is minimized with the 3 Å backbone RMSD loop held in place. At this point, the surrounding side chains are "biased" towards the 3 Å RMSD loop. This structure then provides the surrounding environment for subsequent tests of our loop prediction methods.

Dipeptide Rotamer Frequency Score

For a number of challenging cases, we experimented with the use of a new addition to our energy model that penalizes loop conformations that are constructed with seldom-observed dipeptide dihedrals. The dipeptide rotamer frequency-based scoring term employed a greatly expanded dipeptide rotamer library (garnered from ~7500 high-quality PDB structures) that incorporated the frequency of each rotamer in this subset of the PDB. This information was used to penalize loop dipeptides whose combination of (ϕ,ψ) angles fall in an extremely unpopulated region of the five-dimensional dipeptide analogue to the well-known Ramachandran plot. The set of five angles for each dipeptide in the predicted loop, using a "sliding window" scheme, is compared against the new library to find the nearest dipeptide rotamer. Two criteria determine whether a penalty will be applied to the dipeptide:

1. If the Euclidean distance between the loop dipeptide and the nearest rotamer in the library is greater than a certain, empirically determined cutoff.

2. If the total population of rotamers within a set radius of the loop dipeptide is below a certain threshold.

The form of this penalty term, its implementation, and its successes in improving loop prediction in crystal structure and homology model environments will be discussed in detail in an upcoming publication. This term was used in two situations:

1. For all of the predictions in inexact environments. This is a substantially more challenging sampling and scoring problem, and the information contained in the dipeptide score can be expected to improve results systematically.
2. For a small subset of the predictions in the native environment where difficulties in the standard prediction approach were encountered.

To date, we have not found any cases where this term worsens results. However, more extensive tests are underway and will be presented in a subsequent publication

3.3 Results and Discussion

Description of Test Cases

Application of the discriminating criteria used to select suitable LHL test cases yielded a set of 35 loop-helix-loops of 16 or 17 residues in length. These loops exhibited a distribution of helix size as shown in Figure 3.3. The distribution indicates a diversity of helix sizes within a 16- or 17-residue loop. Although the helical library described in this work is only for α-helices, loops were included that contained 3^{10} helices, either separate from an α-helix already present in the loop, or as the sole secondary structure of the loop. It is these former cases where a loop contains both a 3^{10} helix with an α-helix that led to the non-zero frequency for helices of length three (Figure 3.3).

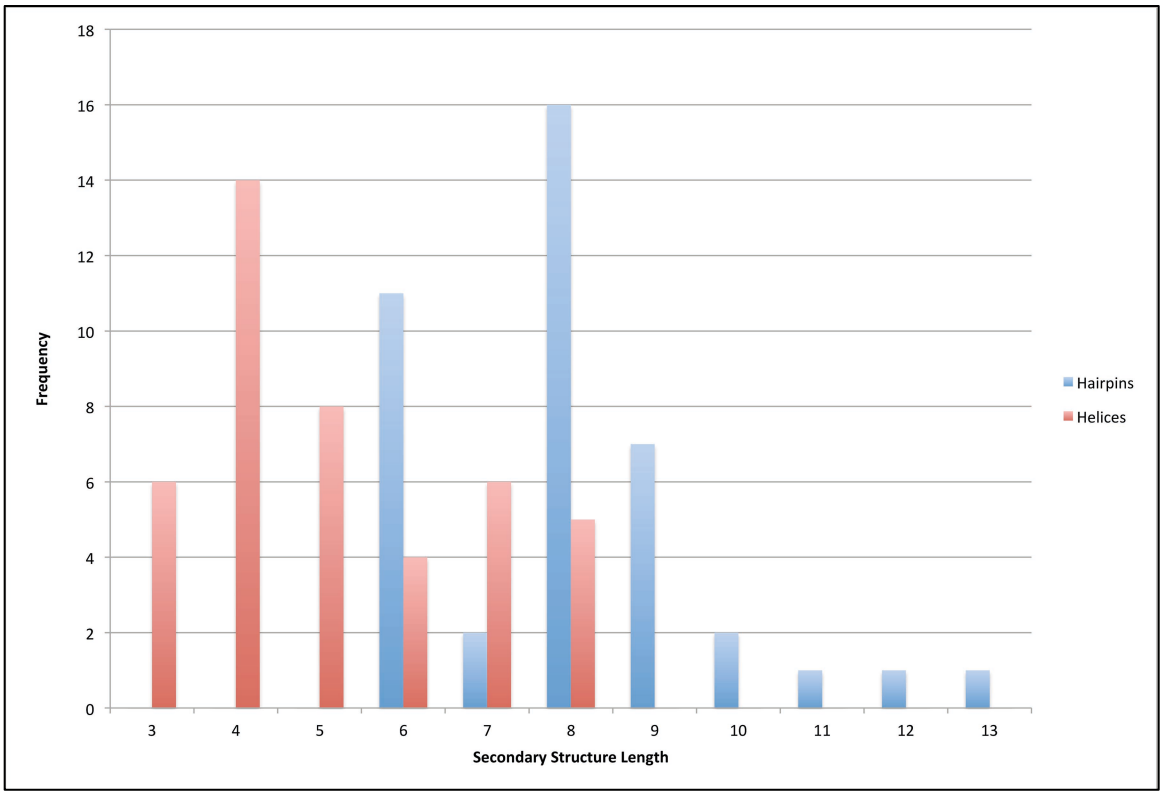

Figure 3.3 Distribution of secondary-structural elements within the test set of loops.

Helices of length 3 were from 3_{10}-helices found in loops already containing an α-helix. Hairpin length includes the terminal hydrogen bonded residues as well as all residues in between.

PDB 1W27 contains a noteworthy example of a multi-helical loop. The 17-residue loop, contains a 4-residue 3^{10} helix and 5-residue α-helix separated by a single residue, D302 (Figure 3.4). Evidently, residue D302 permits flexibility in the backbone to transition from one helical type to another. We explored the use of our α-helical library in three approaches: 1) Loop prediction given the α-helix as the helical bounds; 2) Loop prediction given the 3^{10}-helix as the helical bounds; 3) Loop prediction where the 3^{10} and α-helix bounds are combined to yield a 10-residue "α-helix." The results of these approaches are described in greater detail below.

PDB 2VPN was another case of a multi-helical loop. The 16-residue loop of interest is composed of a 4-residue α-helix and a 7-residue α-helix separated by a single residue, E102 (Figure 3.5). Residue E102 is kinked, according to DSSP, failing to form the periodic hydrogen bond expected of an α-helix. As in the 1W27 case, we tried three approaches to predicting this loop.

For β-hairpins, a set of 41 cases was collected satisfying the criteria described in the methods section. The size of the hairpin region ranged from 6 to 13 residues within loops up to 17 residues in length. Hairpin size is defined to be the number of residues from the start of the first β-strand to the end of the second β-strand, including all non-β residues in between. Hairpins occurred most frequently as either six or eight residues in length (Figure 3.3). However, since the formation of the coordinated hydrogen bonds is what is most challenging in loop-hairpin-loop prediction, we feel it is useful to describe the distribution of hydrogen bonds across our set of β-hairpins. Hairpins contained from four to eight hydrogen bonded residues with the number of coil/turn residues contained within the hairpin ranging from two to seven residues (Figure 3.6). Thus, this test set of β-hairpin containing loops required the successful prediction of at least one specific hydrogen bond spanning at most seven residues.

Predictions performed in the crystal structure environment

A total of 35 loop-helix-loop (LHL) cases and 41 beta-hairpin cases were predicted in the crystal structure environment. In the crystal structure environment, the loop of interest is deleted and rebuilt while the surrounding residues remain fixed. In this work, we compare the predictions done using a helical dihedral library versus predictions performed using the standard PLOP dihedral library[37].

Loop-Helix-Loops predicted using the dipeptide dihedral library versus the helical dihedral library with exact helical bounds

As a first test of the helical dihedral library, we performed loop prediction on the set of 35 LHL cases either with the previous dipeptide dihedral library[37] or with the helical library described in this work. Experiments such as these were primarily meant to ensure that in the absence of uncertainty in the size and location of the helix, our helical library method could succeed. A prediction performed where the helix is postulated from secondary structure prediction software is our primary methodological algorithm to be used in realistic prediction situations, and is discussed later. Table 3.1 provides a summary of the results as a function of helix length. Compared to the dipeptide dihedral library, the helical dihedral library consistently displays improved accuracy, with mean and median RMSD always below 1 Å. No strong correlation is noted between the size of the internal helix and the results from either dihedral library. This suggests, consistent with past results[30, 36-37], that the difficulty in loop prediction lies with the size of the loop, rather than the secondary structure contained in the loop, at least for helices up to eight residues in length.

Helix Length	Number of Cases	Dipeptide Dihedral Library				Helical Dihedral Library with Exact Helical Bounds			
		RMSD (Å)		ΔE (kcal/mol)		RMSD (Å)		ΔE (kcal/mol)	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
4	12	0.53	1.29	3.89	12.39	0.55	0.99	-0.02	-3.18
5	7	0.91	1.09	-7.94	-6.19	0.51	0.80	-3.74	-3.41
6	4	1.00	0.95	2.06	11.11	0.62	0.77	2.47	2.75
7	5	0.55	1.94	0.51	5.19	0.79	0.91	2.52	1.59
8	5	0.81	0.98	4.33	6.66	0.36	0.41	-4.22	-2.18

Table 3.1 Comparison of Loop-Helix-Loop predictions with the dipeptide dihedral library versus the helical dihedral library.

The two noteworthy multi-helical loops found in PDB 1W27 and 2VPN are excluded in this table. The ΔE value compares the energy of the lowest energy loop against the crystal structure loop coordinates, minimized using our energy function. The RMSD reported is of the lowest energy loop prediction and corresponds with the ΔE.

For LHLs containing a four-residue helix, both dihedral libraries appear to perform similarly. As might be expected, the helical library shows the greatest advantage for predictions containing an eight-residue helix with superior median and mean RMSD values by around 0.5 Å. It is likely that the coordinated hydrogen bonds that need to be formed are easily generated when explicit helical dihedrals spanning the precise residues are deliberately introduced during sampling. This seems particularly relevant for the LHL in PDB 2YR5. This is a 16-residue loop containing a 7-residue α-helix (Figure 3.7).

The dipeptide dihedral library produces a 7.26 Å RMSD loop with a ΔE of -0.9 kcal/mol relative to the minimized crystal structure, while the helical dihedral library leads to a 1.11 Å RMSD loop with a ΔE of -18.34 kcal/mol. The dipeptide dihedral library clearly fails to form the native helix, forming instead a loop that protrudes out in solution. The prediction with the helical library is dramatically superior but forms a larger nine-residue α-helix. Evidently, the shorter seven-residue α-helix "seeds" the larger helix. Considering the large negative ΔE energy relative to the native, these additional two helical residues may be the result of an energy error incorrectly favoring formation of additional helical residues. While slightly detrimental to the accuracy of this particular loop prediction, as is discussed in greater detail below, the use of a shorter helix to "seed" a larger one is later exploited to find the lowest energy loop.

Two PDB structures, 1W27 and 2VPN, each contain a multi-helical loop-helix-loop that still satisfied the criteria stated above for selecting loops (Figure 3.4, Figure 3.5). These cases provided an opportunity to explore the effect of the helical dihedral library in complex situations. We attempted to predict the loop by supplying as helical bounds either of the two helices or treated the helices as combined, disregarding the non-helical residues dividing the helices. Table 3.2 describes the result of these loop predictions. In both cases, the helical library produced the lowest energy conformation with sub-Ångström RMSD.

PDB	1W27					2VPN			
Helical Bounds Supplied	None	4-res 3^{10}-helix	5-res α-helix	Combined 10-res "helix"	SSPro truncated α-helix	None	4-res α-helix	7-res α-helix	Combined 12-res "helix"
RMSD (Å)	2.69	1.50	0.77	1.98	0.34	0.42	0.41	0.37	0.38
ΔE(kcal/mol)	38.96	22.27	-3.43	24.32	-12.19	2.01	11.08	-9.69	0.23

Table 3.2 Prediction of multi-helical loops using various loop bounds.

When no helical bounds were supplied, loop prediction was performed using the dipeptide dihedral library. The 1W27 prediction using the 4-res 310-helix for helical bounds still employed the α-helix dihedral library described in this work. The combined helical bounds of 1W27 and 2VPN consider both helices to be one large α-helix during loop buildup. The truncated SSPro helix is equivalent to the 5-res α-helix but truncated one residue at the helical N-terminus. ΔE refers to the change in energy of the predicted loop relative to the native conformation.

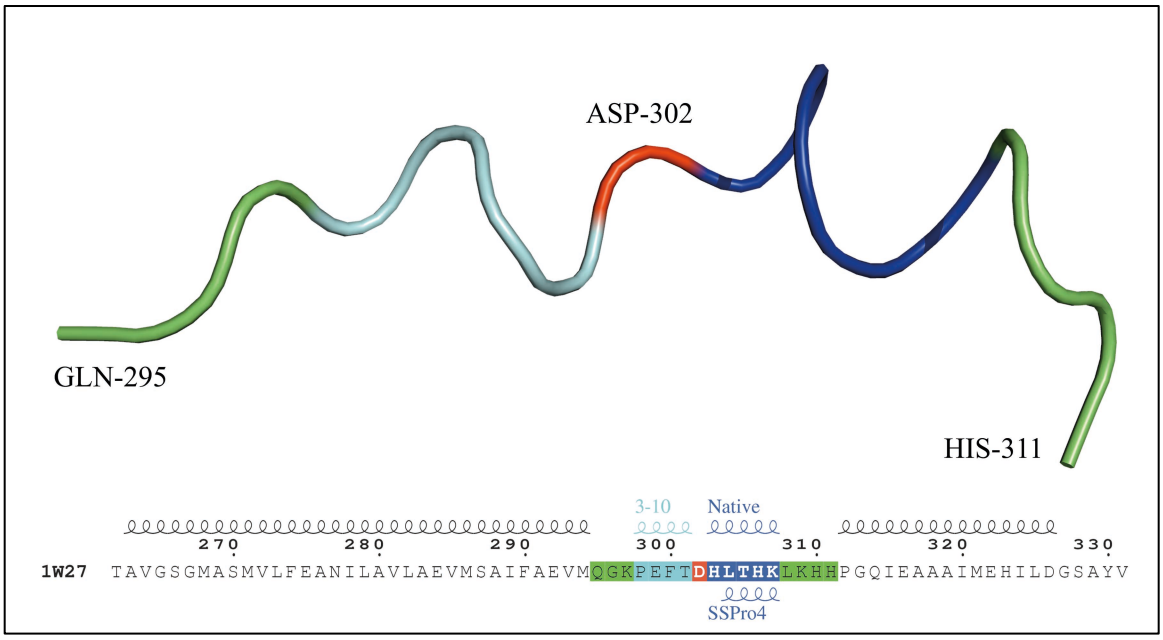

Figure 3.4 Multihelical loop in PDB 1W27.

The loop bounds are Q295 to H311. Residues preceding and following the helices are colored green. The five-residue α-helix is colored blue, while the four-residue 3_{10}-helix is colored cyan. Residue D302, the kinked residue dividing the two helices, is colored red. We attempted separately to use the helical bounds of either the α-helix or the 3_{10}-helix or treated all 10 residues as one "α-helix".SSPro4, a sequence-based secondary structure prediction program, assigned the four residues from L304-K304 as helical. The sequence annotation was generated using ESPript. This loop conformation and all other similar illustrations were produced using Pymol.

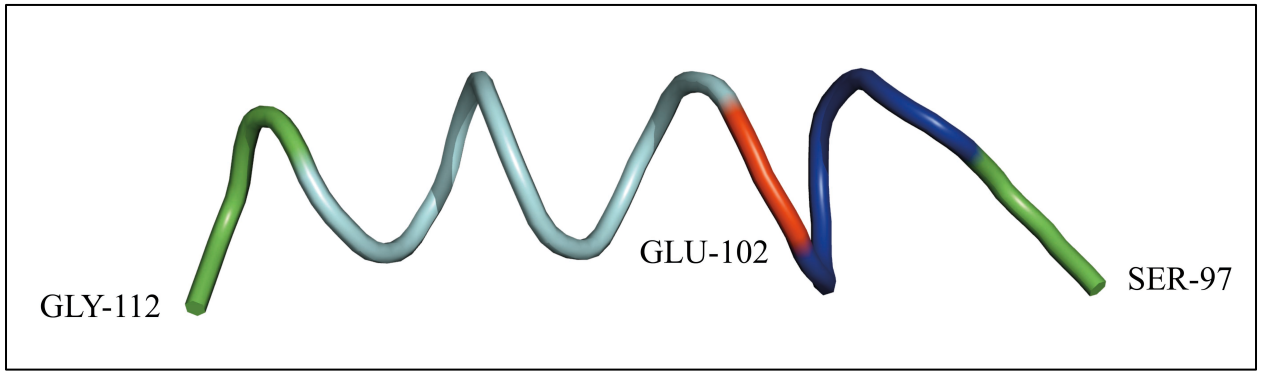

Figure 3.5 Multihelical loop in PDB 2VPN.

The loop bounds are S97 to G112. Residues preceding and following the helices are colored green. The seven-residue α-helix is colored cyan, while the four-residue α-helix is colored blue. Residue E102, the kinked residue dividing the two helices, is colored red. We attempted separately to use the helical bounds of either the seven-residue helix or the four-residue helix or treated all 12 residues as one "α-helix".

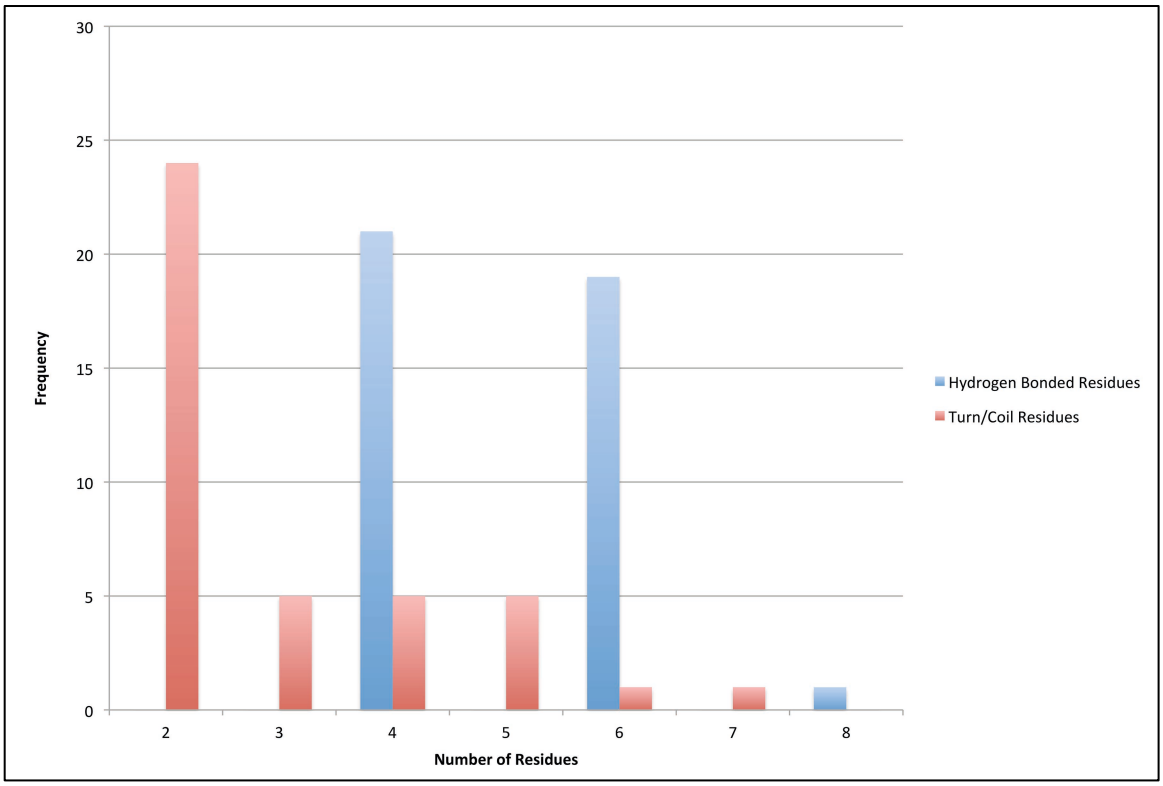

Figure 3.6 Distribution of hairpin characteristics.

Hairpins contained from four to eight hydrogen-bonded residues and with the internal turn/coil residues spanning a length from two to seven residues.

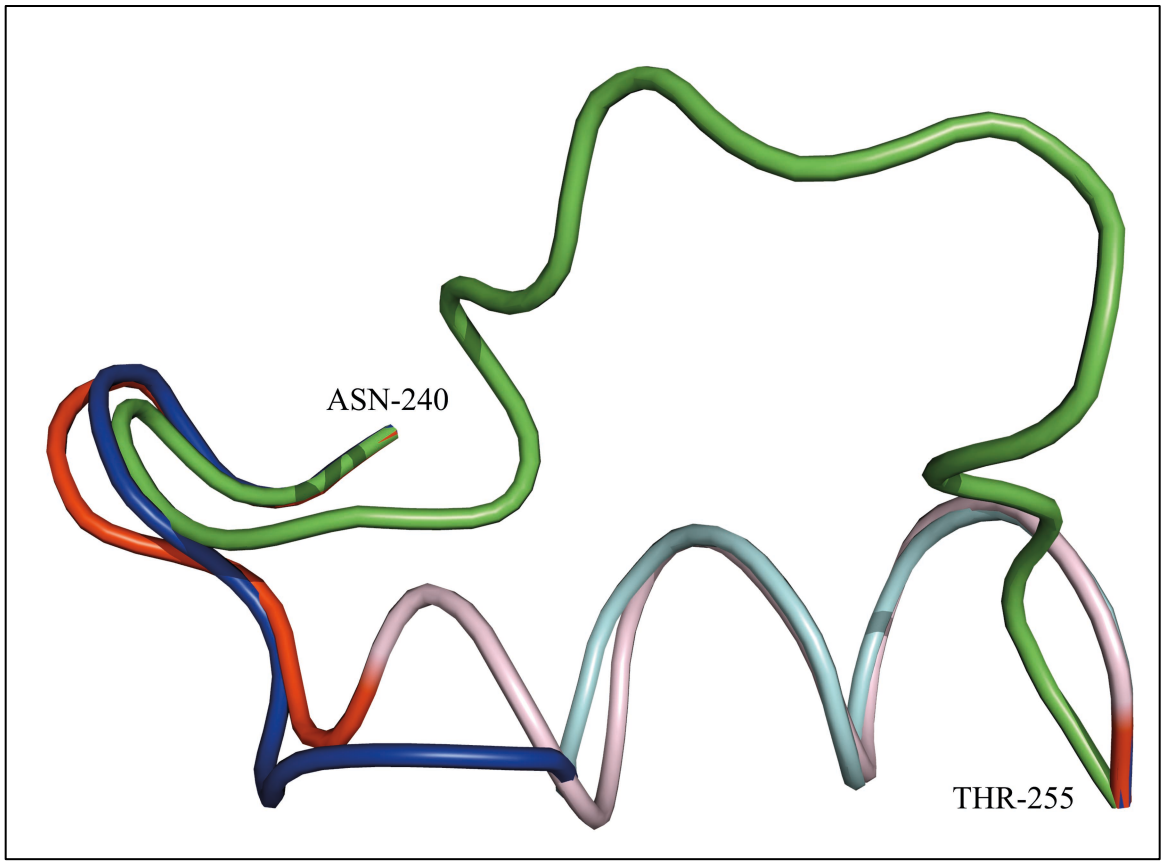

Figure 3.7 Loop-helix-loop predicted in PDB 2YR5.

The native loop coordinates are colored blue with the seven-residue α-helix colored teal. The prediction using the helical dihedral library is shown in red with the resulting nine-residue α-helix colored in pink. The loop prediction performed using the dipeptide dihedral library is shown in green. Despite supplying the exact seven-residue helical bounds during loop prediction with the helical library, what resulted was a slightly larger helix, evidently "seeded" by the small seven- residue α-helix.

Loop-Helix-Loop prediction based on helical bounds derived from SSPro4 and PSIPRED

In the previous section, exact helical bounds were used which were taken from the output of DSSP when applied to the crystal structure. Such accurate information will not be known *a priori*. Indeed, significant variability in the definition of secondary structure assignment has been known to affect the precise bounds of secondary structure, especially as the number of secondary structure assignment definitions is now legion[79]. To simulate the effectiveness of using the helical dihedral

library in more realistic computational experiments, and to further gauge the sensitivity of our method to accurate knowledge of the helical bounds, we applied the popular sequence-based secondary structure prediction packages SSPro4[41, 78] and PSIPRED[40] to our set of 35 loop-helix-loops and attempted loop prediction using these predicted helical bounds. The results from these secondary structure prediction packages, excluding the multi-helical loops of PDB 1W27 and 2VPN, are presented in Table 3.3.

Helical Bounds Predicted	PSIPRED	SSPro4
Exact	2	14
Truncated	6	2
Overlapping	6	9
Non-overlapping	1	1
No Helix	18	7
Total	33	33

Table 3.3 Results of sequence-based secondary structure prediction packages PSIPRED and SSPro4 on our set of LHLs, excluding cases 1W27 and 2VPN, the multi-helical loops.

Exact helical bounds are those that are in perfect agreement with the bounds assigned by DSSP on the crystal structure. Truncated helical bounds are those that lie within the DSSP assigned bounds. Helical bounds are considered overlapping if the secondary structure predicted helix has at least a single residue overlapping the exact bounds. No helix is considered predicted if the entire loop-helix-loop lacks any helical assignments greater than three residues.

Comparing the two packages, it would appear that SSPro4 could more reliably find exact or overlapping helical bounds compared to PSIPRED, however the two methods are complementary. For example, SSPro4 fails to find any helix in the LHL in PDB 3LY0, while PSIPRED found a truncated helix whose bounds are contained within the DSSP results. We must caution the reader that we do not attempt here to perform a rigorous evaluation of secondary structure prediction

algorithms. For that, we refer the reader to Koh *et al.* 2003[80] and Pirovano and Heringa, 2010[42]. Rather, we simply selected two popular and easily available packages for our study. Alternative secondary structure prediction algorithms may be just as valid, as is using more than two packages to find the helical bounds. However, the fact that in a large set of cases, the exact, DSSP helical bounds were identified provides some legitimacy in interpreting the results from the previous section – accurate knowledge of a helix within an LHL is not unreasonable.

For the two multi-helical loops in PDB 1W27 and 2VPN, the two secondary structure prediction methods contrast. For the LHL in PDB 1W27 (Figure 3.4), PSIPRED correctly identifies the five-residue α-helix but fails to predict the four-residue 3^{10}-helix. SSPro4 also fails to identify the 3^{10}-helix but the α-helix is incorrectly predicted to be four residues, truncated at the N-terminus. In 2VPN (Figure 3.5), PSIPRED predicts a combined helix that spans both α-helices and extends one residue further towards the C-terminus. Contrastingly, SSPro4 considers the entire LHL to be one large helix – a result that is inadequate for our helical dihedral library approach. In both of these cases, PSIPRED offers a reasonable set of helical bounds for use in our method.

Table 3.4 summarizes the results of LHL prediction using the helical bounds, when available, from PSIPRED and SSPro4. In general, the helical bounds provided by the sequence-based secondary structure prediction methods SSPro4 and PSIPRED are effective in loop-helix-loop prediction. Although the statistics might suggest that the fewer cases afforded by PSIPRED result in higher quality predictions, we refrain from making such a conclusion, as it may be necessary to also take into account the size of exact helix studied. This does illustrate, however, that sequence based secondary structure assignments are useful to our method when performing three-dimensional loop prediction.

Method	Number of Successful Cases	RMSD (Å)		ΔE (kcal/mol)	
		Median	Mean	Median	Mean
PSIPRED	13	0.44	0.49	-1.37	-1.54
SSPro4	25	0.60	0.91	1.05	0.65

Table 3.4 LHL prediction using the helical bounds available from PSIPRED and SSPro4.

Multi-helical cases 1W27 and 2VPN are included in these statistics. Cases where the helical bounds provided by sequence-based secondary-structure prediction are not useable in our method are excluded. Further, cases where no loops were able to be predicted with the supplied helical bounds are also excluded.

It should be mentioned that five cases were found where the helical bounds offered by either PSIPRED or SSPro4 resulted in failed loop predictions where not a single predicted loop was constructed. In four of these five cases (PSIPRED bounds: PDBs 1N45, 1OAO, 2YR5; SSPro4 bounds: PDB 3GWI), the sequence-based secondary structure assignment places the helix as part of the N or C terminus. It would appear that in these cases, the sequence-based assignment is extending the larger helix that forms the boundary of the loop-helix-loop into what DSSP, and the criteria used in this chapter, consider to be part of the loop. Although in practice, assigning the terminal residues of a loop to be helical is not fatal – PSIPRED and SSPro4 both place a helix on the C-terminus of the LHL in PDB 1HN0 and yet a sub-0.5 Å RMSD loop is predicted – loop prediction without any non-helical residues to precede the helix is extremely difficult.

In these situations, the lever effect, described previously in the single-loop prediction section of Materials and Methods, becomes very pronounced. As PLOP constructs the loop in a tree-based method, where the tree is split into additional branches as more loop residues are predicted, placing the helix at a loop terminus means there are no preceding branches to rely upon. Whatever few

positions the leading residue of the helix is placed at are set entirely by the sparse number of helical rotamers present in our library. In practice, this means that all the rotamers in our helical rotamer library for a given helix size are easily rejected. Although in principle, one could reduce the *ofac* parameter to permit greater steric overlap between a loop residue and the surrounding environment, in practice, the *ofac* was rarely seen as the limiting factor. The one case that permitted loop prediction after adjusting the *ofac* was the PSIPRED bounds for 1N45, however, we had to set the *ofac* to an abnormally low value of 0.20, meaning enormous steric clashes were permitted. Even still, the output of this loop prediction only produced a 5.69 Å RMSD loop with a ΔE of 9.30 kcal/mol.

In all cases, nascent loop segments were screened out when the helix placed a residue too far from the body of the protein to what has been empirically observed across published crystal structures containing protein loops. Or instead, loops were screened when the distance between the loop segment containing the helix and the opposing end of the loop is considered too great to be spanned by whatever intermediate residues remain. In other words, the helix places one half of the loop too far away for loop closure to be possible. These loop screening methods are described briefly in the Materials and Methods section, and in greater detail in Jacobson *et al.*[51] Setting the *ofac* to an arbitrary low value has no effect on these screens – the helical rotamer library simply does not contain a suitable rotamer to permit loop prediction with the supplied helical bounds. Although there is certainly an argument to be made for increasing the size of the helical library, as evidenced from our other successes, the size of the library does not appear to be an impediment to loop-helix-loop prediction. Rather, the practitioner of our method might gain insight by noting that if no suitable rotamer is present in the library, it may be prudent to consider alternative helical bounds. Indeed, none of these terminus-bounded helices are the crystal structure helical bounds – we avoided such cases by our definition of loop-helix-loops. Determining the helical bounds from the

output of our previous dipeptide-dihedral library method, as discussed in greater detail below, may be a fruitful alternative. The multi-helical loop of 2VPN (Figure 3.5) is one slight exception. In this case, PSIPRED combines the 4-residue α-helix and the adjacent 7-residue α-helix into one large helix and even extends the helical bounds further by one additional residue to produce a 13-residue helix. SSPro4 simply considers the entire loop-helix-loop to be one large helix, an outcome useless for our helical dihedral library. In this case, the helical bounds provided by PSIPRED produce independent N- and C-terminus loop segments but closure is not achieved. This result occurs regardless of how low we set the *ofac*. Again, extending the size of the helical library may offer a solution to this case, but more likely, the helical bounds provided deviate too greatly from the native structure to permit reasonable loop prediction.

Truncated helical bounds from sequence-based secondary structure prediction or derived from inspection of coordinates predicted with the standard PLOP dihedral library

In a few cases, sequence-based secondary structure prediction methods produced a helix that was truncated relative to the native helical bounds, yet these cases performed as well, if not better, than the native bounds. For example, PDB 1W27, one of the multi-helical loops, is composed of a four-residue 3^{10}-helix and an adjacent five-residue α-helix (Figure 3.4). SSPro4 fails to identify the 3^{10}-helix but predicts the α-helix to be truncated by one residue at the helical N-terminus, relative to the exact helical bounds (Figure 3.4). PLOP was able to predict this LHL with an RMSD of 0.77 Å and a ΔE of -3.43 kcal/mol when using the native, five-residue α-helix. However, the SSPro4 bounds led to a predicted LHL with a superior RMSD of 0.34 Å and a ΔE of -12.19 kcal/mol. Table 3.2 summarizes these results. These loop predictions are illustrated in Figure 3.8.

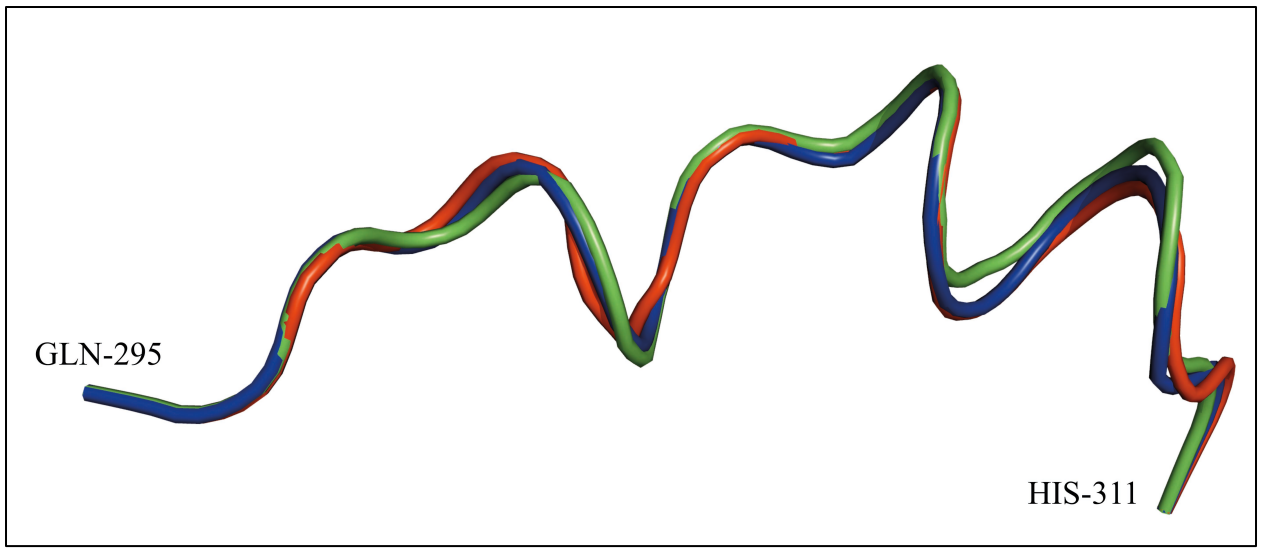

Figure 3.8 Loop-helix-loop prediction for the multihelical loop in PDB 1W27.

The native loop is shown in red. Loop prediction using the exact five-residue α-helix is shown in green. Loop prediction using the truncated, four residue α-helix provided by SSPro4 is shown in blue. Loop prediction using the truncated four-residue α-helix bounds appears to permit improved sampling of the α-helix. Notice that the greatest discrepancy between the two loop predictions occurs along the α-helix near the C-terminus.

Consistent with our past discussion, the smaller helix may permit less of a lever effect and thereby enable finer sampling of the α-helix. It should be noted, however, that the absence of any helical bounds, that is, using the previous dipeptide dihedral library from our previous work, results in a 2.69 Å RMSD prediction (Table 3.2). Thus the small helix is shown to also seed our hierarchical sampling method to more heavily explore conformational space near α-helices.

The LHL in PDB 2YR5 is another case where truncated helical bounds led to a superior prediction. However, this is one of the cases where the helical bounds provided by both PSIPRED and SSPro4 were attached to the LHL C-terminus and no loops emerged from our attempts at predicting this LHL with such helical bounds. Rather, we attempted LHL prediction using as helical bounds all

possible four-residue α-helices that lie within the 10-residue α-helix suggested by PSIPRED and SSPro4 – a set of seven possible helical bounds. Both PSIPRED and SSPro4 suggested identical helical bounds. The results from these predictions are shown in Table 3.5.

Helical Bounds	RMSD (Å)	ΔE (kcal/mol)
None	7.26	-0.9
B:248 – B:254 (Native bounds)	1.11	-18.34
Bounds derived from PSIPRED/SSPro4 Truncation		
B:246 – B:249	1.11	-6.2
B:247 – B:250	1.11	-18.72
B:248 – B:251	1.11	-18.49
B:249 – B:252	1.11	-18.43
B:250 – B:253	1.10	-18.18
B:251 – B:254	1.10	-18.28
B:252 – B:255	4.27	28.57

Table 3.5 Prediction results from the LHL in PDB 2YR5.

LHL prediction without helical bounds refers to the use of the dipeptide dihedral library exclusively. The native bounds are those provided by DSSP analysis on the crystal structure. The PSIPRED/SSPro helical bounds are from B:246 and B:255 and bracket the seven truncation attempts shown. The lowest energy prediction across all helical bounds is highlighted in red.

The predictions indicate that nearly every possible four-residue α-helix attempt produces results that are nearly identical to the LHL prediction performed using the native, seven-residue α-helix. While knowledge of the precise, native helical bounds may not be available, we demonstrate that we can still exploit information provided by sequence-based secondary structure prediction, even if that

information does not perfectly match the DSSP secondary structure identification obtained from the crystal structure of the native conformation.

In total, we attempted all possible four-residue α-helix bounds for all LHL cases where the lowest energy loop was found only by using the native helical bounds. This was performed in order to discount the concern that precise *a priori* information about a helix must be known. In many cases, information about a helix was provided by sequence-based secondary-structure prediction. However, as we show in Table 3.1, providing no helical bounds and using the dipeptide dihedral library can still lead to low RMSD predictions and the formation of a helix. From these cases where a helix four-residues or larger was produced *ab initio*, we also applied our truncation sampling method across the predicted helix and took the lowest energy loop. When the dipeptide-dihedral library simply produced a four-residue helix, we reattempted loop prediction using the helical dihedral library with this previously found 4-residue helix as bounds. The lowest energy loops predicted from these experiments are shown in Table 3.6. In general, the truncation method produces helices that, on their own, are quite accurate with sub-Ångström RMSD routinely reported.

PDB	RMSD (Å)	ΔE (kcal/mol)
1HN0	0.31	-2.77
1Q1R	0.30	-8.07
1WOV	0.95	-6.08
2EX0	1.74	0.91
2FHF	0.62	-8.02
2II2	0.35	-3.4
2J9O	1.55	2.65
2QMC	0.49	-3.05
2VPN	0.22	-11.94
2YR5	1.11	-18.72
3GWI	0.53	3.77
Mean	0.80	-4.28
Median	0.58	-3.23

Table 3.6 Result of LHL prediction using truncated helical bounds.

All possible 4-residue helical bounds that lie within bounds provided by sequence-based secondary structure prediction or by analyzing the results from the dipeptide-dihedral based predictions were used. What is shown is the lowest energy prediction across all helical bounds attempted.

Creation of a systematic method for predicting loop-helix-loop regions

We have described above a number of different approaches to predicting LHL regions, each of which exhibits significant success for a subset of test cases. We briefly enumerate these methods below:

1. Normal loop prediction, without any use of the helical rotamer library.

2. Use of the rotamer library with helical bounds specified by the results of either SSPro or PSIPRED secondary structure prediction (this leads to two separate calculations).
3. Reprediction of the loop subsequent to normal loop prediction, using as helical bounds helical regions forming spontaneously in the normal loop prediction simulation.
4. Truncated helix loop prediction where all possible four-residue helixes that can fit within previously obtained helical bounds are explored.

Our final algorithm is a composite method in which all of the above calculations are performed for each loop, and the lowest energy prediction is selected as the predicted result. The computational cost of this composite method is roughly 4X that of one normal loop prediction. In return, one achieves a remarkably high level of reliability as is shown in Table 3.7 below. The vast majority of predictions are sub-Angström, an exceptionally low level of error for loops of this length and complexity. Only one loop has an RMSD greater than 2A, the loop in PDB 2O70. We discuss this case further below, but in essence neither normal loop prediction, nor any of the secondary structure prediction methods, predict a helix in the relevant region. When the native helix is seeded into the calculation, a superior prediction is returned. Thus, this is a sampling problem, which we can hope to solve by improving the sampling algorithm. However, with the current approach, such sampling errors are very infrequent.

Arguably, the results from predictions with the native helical bounds rely on information that may not be precisely known in a homology modeling experiment. As such, we also report in Table 3.7 the RMSD of the lowest energy loop prediction across all sampling methods. For comparison, results of LHL prediction using helical bounds taken only from the native PDB are shown in the right half of Table 3.7.

PDB	Helical Bounds Identified Without DSSP			Exclusively DSSP Identified Helical Bounds	
	Method	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
1BKR	SSPro4	0.55	1.05	0.55	1.05
1E3D	SSPro4	0.55	-1.1	0.55	-1.10
1HN0	PSIPRED	0.35	-5.51	0.3	0.22
1L5W	SSPro4	0.4	-3.74	0.4	-3.74
1LLF	PSIPRED	0.44	-5.53	0.45	-5.5
1N45	Dipeptide	0.36	-4.22	2.05	12.55
1N7O	SSPro4	0.4	-1.18	0.4	-1.18
1O7E*	SSPro4	0.37	3.34	0.37	3.34
1OAO	SSPro4	0.49	-80.01	0.49	-80.01
1OX0	Dipeptide	1.35	-13.23	0.58	-8.7
1Q1R	Truncate	0.3	-8.07	0.3	-8.07
1QMY	SSPro4	1.27	-2.67	1.27	-2.67
1SU8	Dipeptide	0.43	2.24	1.45	18.65
1W27	SSPro4	0.34	-12.19	0.77	-3.43
1WOV	Truncate	0.95	-6.03	0.67	-2.88
1ZX0	Dipeptide	1.04	11.2	1.81	20.12
2DEB	Dipeptide	1.35	-10.14	1.55	8.93
2EX0	PSIPRED	0.44	-0.86	0.33	-4.4
2FHF	Dipeptide	0.54	-8.7	0.54	-10.16
2II2	Truncate	0.35	-3.4	0.36	-4.22
2J9O	Truncate	1.55	2.65	1.55	2.65
2JA2	Dipeptide	0.81	0.13	0.72	1.15
2JDI	SSPro4	0.51	15.02	0.51	15.02
2O70	Dipeptide	3.24	-12.42	1.71	-15.09

2P0W	**Dipeptide**	0.47	-7.94	0.51	-6.96
2QMC	**Truncate**	0.49	-3.05	0.49	-3.05
2RJ2	**PSIPRED**	0.31	-1.37	0.57	4.72
2V36	**Dipeptide**	0.18	-1.22	0.25	-0.37
2VPN	**Truncate**	0.22	-11.94	0.37	-9.69
2WEU	**Dipeptide**	0.91	-10.24	1.73	12.19
2YR5	**Truncate**	1.11	-18.72	1.11	-18.34
3CWW	**SSPro4**	0.28	-12.04	0.27	-10.71
3GWI	**Truncate**	0.53	3.77	0.38	7.28
3HL0	**PSIPRED**	0.93	-2.04	0.39	2.52
3LY0	**PSIPRED**	0.54	1.28	0.79	9.14
Mean		0.70	-5.91	0.76	-2.31
Median		0.50	-3.57	0.55	-1.74

Table 3.7 Results of all LHL predictions independent of helical bounds derived from analysis of the crystal structure as well as the results using bounds derived exclusively from the crystal structure.

By sampling with alternate helical bounds derived from sequence-based secondary-structure prediction and/or the truncation method, the LHL prediction statistics are slightly superior to predictions using helical bounds derived from the output of DSSP. The four cases that are inferior to LHL prediction with exact DSSP helical bounds are highlighted in red. Only one case, 2O70, has an egregiously poor RMSD. The LHL in PDB 1O7E was an exception in that the low Å RMSD reported herein was only produced by introducing the native helical dihedrals into our helical dihedral library.

Overall, by exploring helical bounds provided by sequence-based secondary-structure prediction methods, as well as using the truncation method, we were able to predict LHLs with slightly superior accuracy than if we were to rely on the DSSP identified helical bounds. However, there were four cases where we were unable to produce a prediction that was superior to approach using the DSSP-based bounds. Three of the four predictions are 0.11 Å from the DSSP results and can be left as acceptable.

The only egregiously inferior prediction was for the LHL in PDB 2O70. Here, the use of the DSSP-based helical bounds led to a 1.71 Å RMSD prediction compared to a 3.24 Å RMSD prediction performed solely using the dipeptide dihedral library – that is, without any supplied helical bounds (Table 3.7). Evidently, this LHL is a challenge for sequence-based secondary-structure prediction as well since neither PSIPRED nor SSPro4 predict there being any helix at all within the LHL. Cendron *et al.*, 2007 argue that the sequence of PDB 2O70, an OHCU decarboxylase from *Danio rerio* (zebrafish), lacks homology with other known amino acid sequences[81]. This may have been the case in early 2007 but evidently is now longer so. In June 2007, the crystal structure of *Arabidopsis thaliana* OHCU decarboxylase was published (PDB: 2Q37), and in 2010, the *Klebsiella pneumoniae* structure (PDB: 3O7I) was deposited in the PDB[82-83]. However, in these two more recent structures, the five residues comprising the α-helix are not conserved and the more homologous eukaryotic 2Q37 structure fails to form a helix at this position. It seems reasonable then that PSIPRED and SSPro4 would fail to identify this helix.

With respect to size of our helical dihedral library, the LHL in PDB 1O7E posed the only challenge. In the above Table 3.7, we report the prediction results when using an augmented helical dihedral

library containing the native dihedrals for the helix. In the absence of this addition to our library, the LHL prediction led to a sampling error with an RMSD of 2.09 Å and a 16.99 kcal/mol ΔE compared to a 0.37 Å RMSD and 3.34 kcal/mol ΔE with the augmented library. As discussed in the methods section, our helical dihedral library is populated with rotamers that conform close to ideality. This approach fails here and seems likely due to the large discrepancy from ideal ϕ,ψ angles for the two terminal residues of the helix. While we expect angles near $(\phi,\psi)=(-60^o,-40^o)$, the torsions for two of the N-terminus residues of the helix, A223 and G224, are $(\phi_{A223},\psi_{A223})=(-68^o,-20^o)$ and $(\phi_{G224},\psi_{G224})=(-104^o,1^o)$ $(\phi_{G224},\psi_{G224})=(-104^o,1^o)$. In particular, the terminal glycine residue poses the largest problem. From this limited case, there may indeed be utility in further expanding our helical dihedral library, but even in its current implementation, the difficulty in this LHL case appears anecdotal.

The ability of the energy model to robustly pick out the correct loop as being lowest in energy provides new confirmation of the quality of our latest generation model, supporting the results obtained in Li *et al.,* for long loop regions without secondary structure elements embedded[30]. It is true that phase space available to the loop is significantly restricted when the native environment is (as here) retained; nevertheless, previous results from our group and others show that it is quite easy to generate grossly incorrect predictions (with substantial energy errors) with an inferior scoring function. The results discussed below in which surrounding side chains are allowed to move, in which sub-Ångström results are uniformly obtained, provides further evidence of scoring function accuracy and robustness.

Hairpins predicted using the standard PLOP dihedral library

In addition to loop-helix-loops, we also attempted prediction of, what could be termed, loop-hairpin-loops as another challenge of loop prediction containing local secondary-structure.

The results from loop-hairpin-loop prediction, arranged by hairpin length, are shown in Table 3.8, and the complete results for all 41 hairpin predictions are provided in Table 3.9.

Hairpin Length	Number of Cases	Dipeptide Dihedral Library			
		RMSD (Å)		ΔE (kcal/mol)	
		Median	Mean	Median	Mean
6	11	0.41	1.07	-5.61	-5.05
7	2	1.13	1.13	-21.38	-21.38
8*	16	0.64	0.90	-6.47	-6.77
9	7	0.51	0.89	-5.00	-5.74
10	2	0.42	0.42	-7.32	-7.32
11	1	0.53	0.53	-10.55	-10.55
12	1	0.30	0.30	-3.06	-3.06
13	1	0.44	0.44	-0.04	-0.04

Table 3.8 Results of loop-hairpin-loop predictions using the dipeptide dihedral library.

The ΔE value compares the energy of the lowest energy loop against the crystal structure loop coordinates, minimized using our energy function. The RMSD reported is of the lowest energy loop prediction. * - Of the eight-residue hairpins, one of the cases, the loop-hairpin-loop as part of PDB 2ZBX, initially reported the best structure as that with a 17.29 Å RMSD. The results for this prediction were rescored, using the RFS, leading to a 1.02 Å prediction being considered the lowest in energy and was used in the statistics reported in this table. This rescoring is discussed in detail in the text.

PDB	Loop Length	Hairpin Length	RMSD (Å)	ΔE (kcal/mol)
1C7N	13	8	0.69	-3.34
1F0L	11	6	0.41	-1.71
1GWI	15	8	0.64	-9.05
1GYH	14	9	0.38	-3.18
1LLF	11	6	1.69	-5.61
1NVM	15	9	0.51	-9.55
1O5K	11	6	0.17	-10.79
1TC5	15	8	1.08	1.69
1U60	14	8	0.47	-12.09
1U8V	13	9	0.33	-1.05
2BS2	15	9	0.6	0.03
2C0D	11	7	0.29	-10.5
2CIU	15	10	0.29	-7.11
2IJ2	16	9	0.61	-7.83
2O36	12	9	3.61	-5
2OKX	16	8	2.88	-6.87
2PB2	13	9	0.21	-13.6
2R2N	8	6	0.24	-6.21
2RFG	11	6	0.63	-5.61
2SLI (A: 177 – 190)	14	6	0.26	-0.93
2SLI (A: 236 – 249)	14	8	0.47	-8.88
2WIY	16	8	0.63	-2.36
2WM5	15	8	1.14	-18.43
2YR5	13	6	0.63	-10.95
2YWN	17	13	0.44	-0.04
2ZBX	15	8	1.02	4.70

2ZWA	16	11	0.53	-10.55
2ZYO	8	6	0.36	-1.32
3A9S	12	6	0.18	-4.65
3BF7	11	6	0.98	-9.1
3BJE	12	8	0.34	-3.33
3CSS	17	12	0.30	-3.06
3CU2	11	8	0.38	-3.71
3EGW	12	8	0.49	-10.12
3EI9	15	8	2.12	-2.04
3EJA	15	7	1.97	-32.25
3F8T	14	10	0.54	-7.52
3FAU	13	6	6.21	1.33
3GW9	15	8	0.51	-6.06
3HVW	16	8	0.47	-11.81
3LID	10	8	1.02	-16.62

Table 3.9 Results of all loop-hairpin-loop predictions.

For PDB 2SLI, two hairpins satisfying the criteria described in Materials and Methods were found. Those predictions occurred for the chain A residues 177 - 190 and 236 - 249.

Similar to the results for loop-helix-loop predictions, we observe no correlation between the size of the hairpin and the RMSD of the predicted loop-hairpin-loop. We note however that one of the eight-residue hairpin cases produced a large discrepancy between the median and the median (Table 3.8). This case is part of PDB 2ZBX and led to an RMSD of 17.29 Å with a surprising ΔE of -177.74 kcal/mol. It should be noted that the second best case has an acceptable RMSD of 1.02 Å and a ΔE of -10.91 kcal/mol. Of course, we cannot choose this 1.02 Å loop as the best case *a priori*

as determination of the best loop is made purely on energetic grounds. The apparent lowest-energy loop and the native are shown in Figure 3.9.

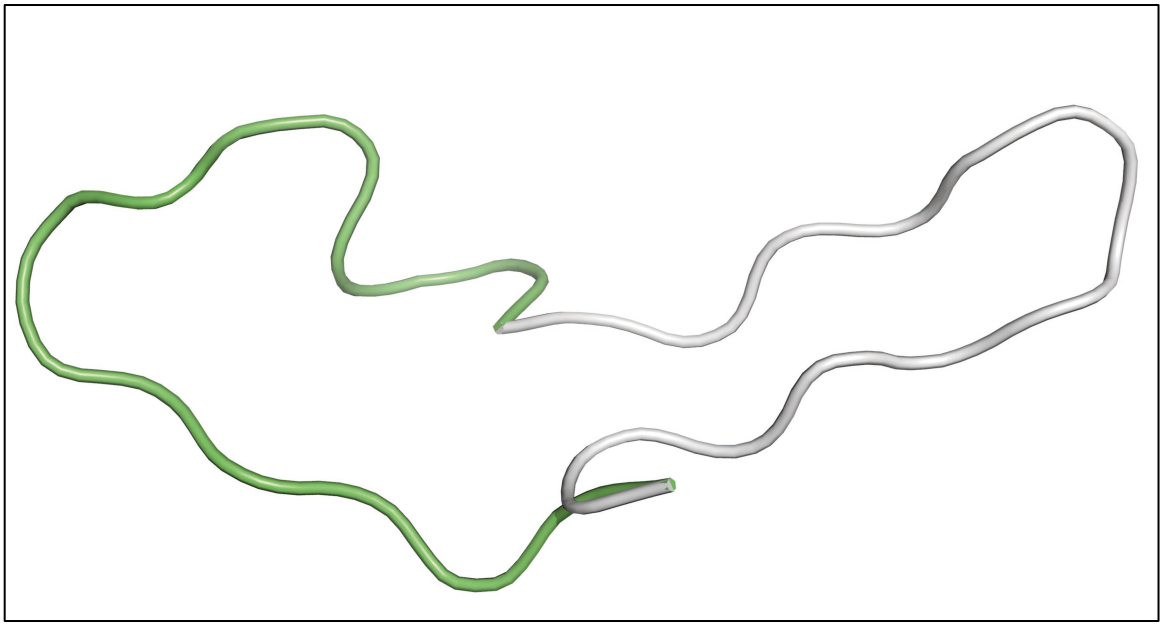

Figure 3.9 Loop-hairpin-loop prediction for PDB 2ZBX.

The native loop is shown in gray while the predicted loop is shown in green.

However, it was observed that the dihedrals in the predicted loop occupy regions of dipeptide-dihedral space $(\phi_1, \psi_1, \omega, \phi_2, \psi_2)$ that are poorly populated across a set of high quality PDB structures. It became possible in this case, and in other cases not discussed in this work, to identify the more "native-like" loop by introducing a dipeptide-dihedral rotamer frequency-based scoring (RFS) term that penalizes structures with non-native dipeptides confirmations. The details of the RFS will be discussed in a future publication. We applied this penalty term to this loop-hairpin-loop case.

Application of the penalty term ranks the 1.02 Å RMSD prediction lower in energy than the 17.29 Å RMSD prediction (Table 3.10). Aside from 2ZBX, five hairpin cases remain where the predictions remain at around 2 Å or worse. These cases are highlighted in red in Table 3.10. For these cases, we explored the use of the RFS throughout the entire loop prediction, rather than just to rescore the final loop candidates. The results for these five cases when using the RFS are shown in Table 3.11.

RMSD (Å)	Freq.-based Score (kcal/mol)	Total Energy (kcal/mol)	ΔE (kcal/mol)
0.0 (native)	9.89	−15697.1	0.0
1.02	25.65	-15692.4	4.7
17.29	4387.82	−9927.01	5770.09

Table 3.10 Energy of the 2ZBX loop-hairpin-loop predictions after application of the frequency-based penalty term.

PDB	Standard Energy Model		Standard Energy Model + RFS	
	RMSD (Å)	**ΔE (kcal/mol)**	**RMSD (Å)**	**ΔE (kcal/mol)**
2O36	3.61	-5.00	0.93	-10.71
2OKX	2.88	-6.87	3.62	6.42
3EI9	2.12	-2.04	0.36	0.24
3EJA	1.97	-32.25	1.86	-27.2
3FAU	6.21	1.33	0.51	-11.6

Table 3.11 Re-prediction of hairpin cases with initial RMSDs of around 2 Å or worse.

Re-predictions were performed by using the RFS throughout the prediction, rather than just to rescore the final putative loops.

The RFS appears successful at correcting the energy error and leading to a lower RMSD in three of the five cases. PDB 2OKX remains a difficult case. Although this case appears to exhibit an energy error before penalizing unlikely structures with the RFS, now a sampling error remains where we

appear unable to produce the native conformation. PDB 3EJA appears to remain an energy error and this case warrants further discussion.

PDB 3EJA contains a 7-residue hairpin within a 15-residue loop that satisfies the various criteria specified in the methods section. In particular the global quality criteria of having suitably highly resolution and superior R-factors was satisfied as well as the local criteria for B-factors and real-space R-factors. Inspection of the predicted loop reveals that we are able to form a reasonable hairpin (Figure 3.10a), and further that during hierarchical loop prediction we succeed in producing a near native loop with an RMSD of 0.94 Å and a ΔE of -1.16 kcal/mol, relative to the native (Figure 3.10b). This would seem to suggest the sampling is not an issue here. The fact that the lowest energy loop predicted (Figure 3.10a) was found nearly 30 kcal/mol lower in energy than the native was surprising. Inspection of the individual residues comprising the loop revealed an unusual close contact between the oxygen on the amide side chain of Q108 and an aromatic carbon on Y191. The distance between these polar and non-polar atoms was a surprising 3.0 Å. Loop minimization perturbs the hairpin such that this distance is increased to 3.5 Å where Y191, like all surrounding residues, is held fixed (Figure 3.10c). The suspicion was that these residues might have been improperly built in the crystal structure and indeed inspection of the electron density showed Y191 to be confidently placed while Q108 was modeled into sparse density (Figure 3.10d). We see no alternative positions to place Q108, however it is beyond the scope of this work to construct the necessary omit maps and attempt model refinement. In describing the structure, the crystallographers do describe a possible role for Y191 but no mention is made of Q108 and so perhaps this residue simply does not hold a stable conformation[84]. Difficulty in modeling an occasional residue in a high-resolution crystal structure is certainly not uncommon. We attempted to exclude loops that were affected by problems such as these in using a real-space R-factor cutoff of

2.0. However, this residue has a real-space R-factor of 0.185. In future studies, it appears a more stringent cutoff is required.

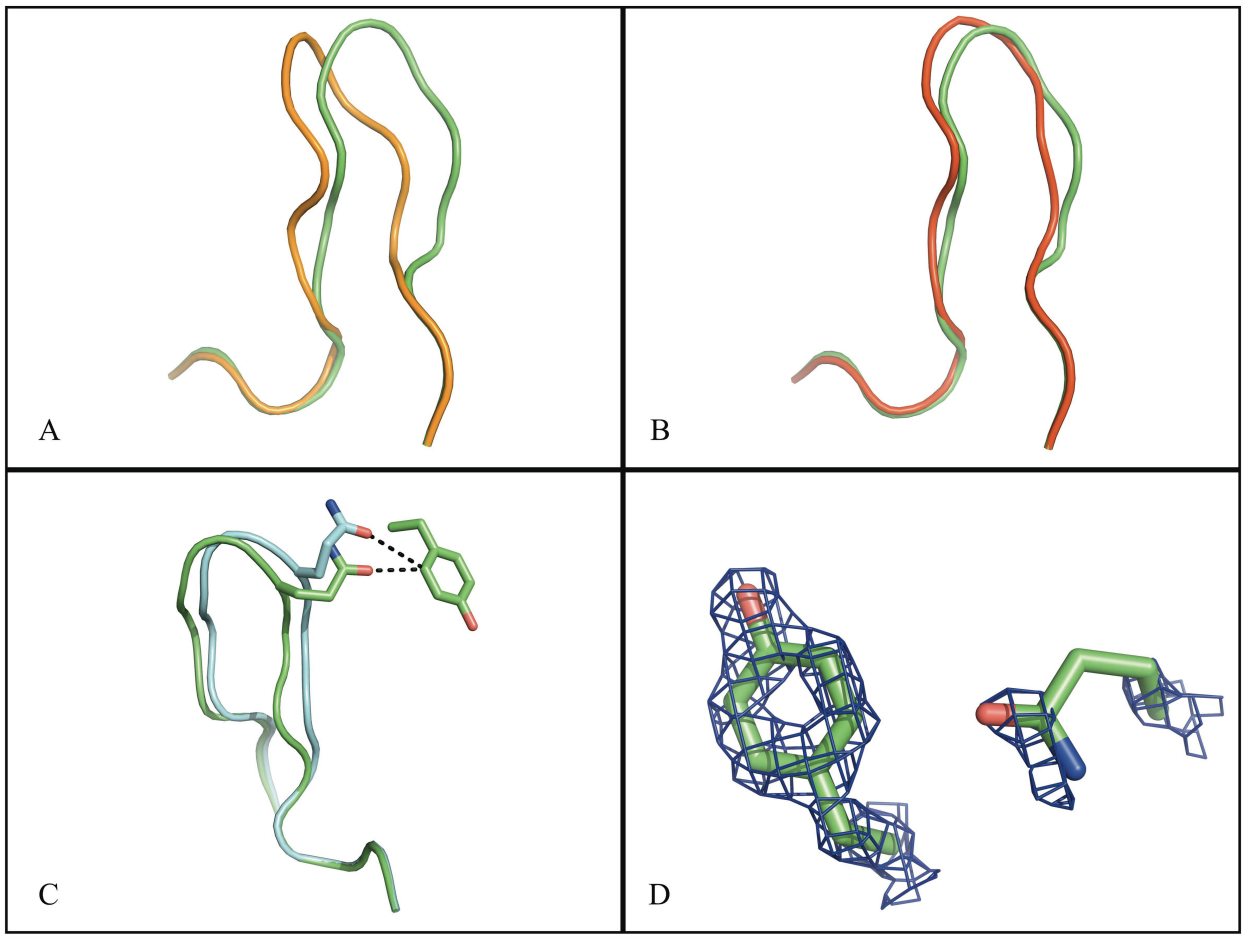

Figure 3.10 Loop-hairpin-loop predictions in PDB 3EJA.

In all panels, the native loop is shown in green. (A) Native hairpin versus the lowest energy prediction using the RFS. (B) Native hairpin versus an intermediately ranked loop. This loop has a 0.94 Å RMSD and a ΔE of -1.16 kcal/mol. (C) Native hairpin versus minimization of the native hairpin. After minimization, the distance between Q108 and Y191 increases from 3.0 Å to 2.5 Å. (D) 2Fo-Fc map contoured at 2σ around residues Q108 and Y191. Observe that while Y191 is confidently built, Q108 has very poor density.

Predictions performed in an inexact environment

Throughout all loop predictions, we have relied on the crystal structure to provide the surrounding environment of the loop. This too, like the precise knowledge of helical bounds, may not be accurately known in a homology modeling experiment. To explore the effectiveness of our sampling and energy model in a more realistic setting, we minimized the surrounding environment in the presence of a predicted, but poor, 3Å RMSD loop. This produced a non-native but locally minimized surrounding side chain environment. However, the backbone environment is still that of the native. From here, we deleted the target loop and performed loop prediction with simultaneous refinement of all surrounding residues. This approach was for both loop-helix-loops and hairpins. We repredicted in an inexact surrounding environment one loop for each secondary-structure length. The loops selected had a sub-1 Å RMSD when predicted in the native environment. For loop-helix-loops, this selection was based on the results from predictions using the exact helical bounds. As would be expected, prediction of the loop as well as surrounding side chains increases the sampling required and computational cost of these predictions. In particular, we found it necessary to introduce additional rounds of side-chain randomization (Table 3.12). Hence, we used only the exact helical bounds to avoid the added complication and expense of sampling surrounding side chains with all the combinations of alternative helical bounds. We also explored the use of the rotamer frequency score (RFS), mentioned previously when describing the improvement in hairpin case 2ZBX (Figure 3.9 and Table 3.10) and others (Table 3.11). Here, we used the RFS throughout the loop prediction, penalizing all intermediate loops as necessary so that only structures with the lowest penalty are likely to advance onto subsequent refinement. The results of these predictions for LHLs are shown in Table 3.12.

Helix Length	PDB	Native Environment		Perturbed Native		Perturbed Native with extra side-chain randomization		Perturbed Native with extra side-chain randomization and RFS	
		RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
4	1BKR	0.55	1.05	1.67	24.54	2.77	2.42	0.61	-2.11
5	1L5W	0.4	-3.74	0.78	-1.39	0.98	-8.97	0.54	-15.03
6	1WOV	0.67	-2.88	1.29	5.25	1.32	-12.85	0.66	-22.55
7	3HL0	0.39	2.52	0.62	-6.97	0.6	-16.28	0.68	-17.84
8	2EX0	0.33	-4.40	2.28	23.84	0.54	10.49	0.76	7.77

Table 3.12 Results from LHL prediction in an inexact environment.

The RMSD is relative to the native structure. The ΔE shown is relative to the energy of the native where the loop and surrounding side chains are minimized.

In all cases, we were able to recover the loop with sub-1 Å RMSD when utilizing additional rounds of side-chain randomization and the RFS. The use of additional rounds of side-chain randomization finds in all cases a lower energy structure. In 2EX0 the effect is most pronounced where a 2.28 Å prediction is improved to 0.75 Å. Still in the cases 1BKR, 1L5W, and 1WOV, additional rounds of side-chain randomization is further improved with the addition of the RFS, which brings, in the most striking example, a 2.77 Å prediction down to 0.61 Å.

Similar results were seen for hairpins as is shown in Table 3.13. As before, the use of additional rounds of side-chain randomization improves results. Most notably, this additional side chain sampling takes the perturbed native prediction for 2CIU from 6.18 Å to 0.41 Å.

Hairpin Length	PDB	Native Environment		Perturbed Native		Perturbed Native + addl. side-chain randomization		Perturbed Native + addl. side-chain randomization + RFS	
		RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
6	1F0L	0.41	-1.71	0.72	12.68	0.74	-10.01	0.73	-14.16
7*	2C0D	0.29	0.89	0.89	-14.19	2.34	-27.39	1.71	-1.54
8	2SLI	0.48	-6.85	0.54	-1.64	3.22	-8.67	0.49	-12.72
9	1GYH	0.38	-3.18	0.73	0.45	0.82	0.18	0.9	1.54
10	2CIU	0.29	-7.11	6.18	29.67	0.41	-22.61	0.57	-10.21
11	2ZWA	0.53	-10.55	0.91	-10.16	0.46	-6.17	0.77	7.68
12	3CSS	0.30	-3.06	0.57	2.05	0.4	-4.86	0.37	-3.73

Table 3.13 Results from hairpin prediction in an inexact environment.

The RMSD is relative to the native structure. The ΔE shown is relative to the energy of the native where the loop and surrounding side chains are minimized. The hairpin of length 7, 2C0D is shown before protonation of D136 in chain B. After protonation of this residue, the energy errors shown here are eliminated. Energy errors occur when predicted loops are reported substantially lower in energy than the native but have poor RMSD. This is discussed in greater detail in the text.

PDB 2C0D evidently posed a significant challenge. The lowest energy structure reported is substantially lower in energy than the native and other similar calculations on 2C0D (Table 3.13). This suggests a problem separate from sampling. Visual inspection of the predicted structure relative to the native illustrates the source of this energy error being due to incorrect protonation state assignment.

This situation is illustrated in Figure 3.11. Shown is the close contact between D136 and Y63. Both residues are part of chain B but Y63 is interacting from a crystallographically related monomer. The distance from the carboxylic oxygen in D63 to the C_β is only 3.2 Å while the distance from that same carboxylic oxygen to that residue's backbone carbonyl is 3.35 Å. Were D63 to be assigned as charged, as it originally was using our previously published algorithm[77], substantial repulsion between D136 and Y63 shown in 3.11.b is expected. D136 lies at the tip of the hairpin and so a large deviation of this residue can lead to a significant RMSD for much of the hairpin. Once D136 is assigned as protonated, the successful prediction shown in 3.11.c results. Here, a 0.56 Å loop is produced with a ΔE of -19.02 kcal/mol. The effect of protonation of this residue on all three perturbed native predictions performed for PDB 2C0D is shown in Table 3.14.

D136 Protonation State	Perturbed Native		Perturbed Native + addl. side-chain randomization		Perturbed Native + addl. side-chain randomization + RFS	
	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
Deprotonated	0.89	-14.19	2.34	-27.39	1.71	-1.54
Protonated	1.34	19.14	0.71	-22.56	0.56	-19.02

Table 3.14 The effect of protonation of D136 on the hairpin prediction in PDB 2C0D.

Remarkably, the prediction of this hairpin when the surrounding environment is native is possible with D136 left as deprotonated (Table 3.13). As shown in Figure 3.11b, incorrect protonation state assignment of D136 leads to residue Y63 being perturbed from its native conformation. Evidently, leaving Y63, and all surrounding environment residues constrained to their native position, removes the heavy dependence on correct protonation state assignment of D136. The fact that the removal of this constraint leaves our predictions sensitive to additional factors is not surprising. Additional perturbed native experiments such as these will be run in the future to expose more weaknesses in our algorithm, however for the cases presented in this work, the difficulties appear isolated to this case and are tractable.

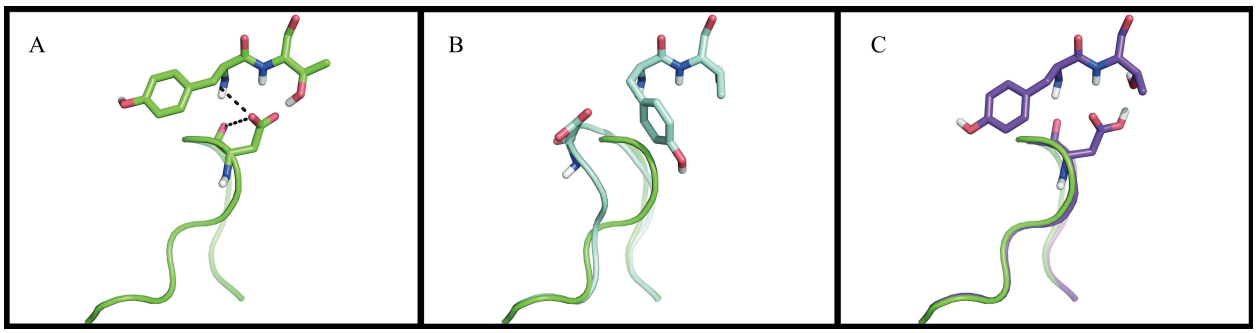

Figure 3.11 Protonation errors in the perturbed native prediction in PDB 2C0D.

In all panels the native loop is shown in green for comparison. (A). The native loop with all atoms shown for D136 and surrounding side chain Y63 and T64. The suspicious close contacts that motivated protonation of D136 are shown dotted in this panel. (B) The coordinates of the atoms in the RFS prediction with D136 deprotonated. (C) The coordinates of the RFS prediction with D136 protonated. Notice the similarity to the native loop in panel A.

Interpretation of the relative energies

Throughout this work, we have reported results comparing the geometry of our predicted structure to the native coordinates via the RMSD, and comparing the energy of our predicted structure to the minimized native via the ΔE. As mentioned in the methods section, $\Delta E = E_{prediction} - E_{native}$. In any successful energy model, the minimized native structure should be reported as being lowest in energy and yet we report negative ΔE values across various predictions. It is worth speculating on the source of this. We believe there are two general possibilities:

1. There are problems in the backbone of the crystal structure that cannot be rectified with our gradient-based minimization as our energy model places the backbone in a local minimum. This seems perfectly plausible in crystal structures, even for the high quality structures explored in this work, as hydrogen atoms positions are not experimentally known, preventing, at the least, the use of an all-atom energy model for refinement. Indeed, Bell *et al.,* report a successful reduction in non-bonded clashes in crystal structures, introduced after consideration of explicit hydrogen atoms, through the use of an all-atom refinement

procedure without any loss in adherence to the diffraction data[85]. Thus, what we may be observing instead is a slightly physically superior structure obtained during the extensive sampling performed during our *ab initio* loop prediction.

2. That negative ΔE values observed in predictions with remarkably low sub-Ångström backbone RMSD may instead be due to improper side-chain contacts being formed. For example, Table 3.12 includes a 0.33 Å prediction of an LHL in PDB 2EX0 with a ΔE of -4.40 kcal/mol. It may well be that these improper contacts are due to a flaw in our energy model, and although this is possible, our ability here to select the lowest energy structure and achieve sub-Ångström RMSDs appears unaffected. As such, in this chapter we do not investigate in greater detail the source of these errors.

We also observe systematic differences in the ΔE across methods and secondary structure. For example, Table 3.1 reports the RMSD and ΔE of LHL predictions performed using just our normal dipeptide dihedral library versus the helical dihedral library presented in this work. In this table, the mean ΔE for all helix lengths predicted is lower with the helical dihedral library than without. This suggests that without the helical dihedral library, there are sampling errors, which are removed by seeding the helix.

For the hairpin predictions, Table 3.8 and Table 3.9 show that the vast majority of predictions conclude with a structure with a negative ΔE. Referring to the first of our two speculations on the source of these negative ΔE values, it may be that the extensive sampling performed in loop prediction is producing superior backbone hydrogen bonds that are not accessible through minimization of the crystal structure.

3.4 Conclusions

We have developed a robust algorithm to exploit secondary structure prediction of small helical segments in loops to yield routinely accurate loop-helix-loops predictions to atomic accuracy. Furthermore, we have demonstrated that our previous dipeptide-dihedral library and all-atom energy model can successfully predict loops containing hairpins. By running parallel loop predictions with a systematically generated set of putative helical bounds from two secondary structure prediction algorithms (SSPro4 and PSIPRED) as well as the normal loop prediction protocol, we have demonstrated that the native loop-helix-loop can be reliably sampled and accurately scored.

This application of a separate, helical dihedral library to a subset of loop residues is at the crux of our method. It affords us increased likelihood of the formation of the coupled hydrogen bonds that define secondary structure by performing loop buildup with the coupled dihedral angles already in place, but it has also introduced a sort of lever effect, where small changes at the base of the helix lead to significant displacement of the terminal end of the loop. For smaller helices, this is obviously less of a problem but for larger helical bounds, such as the LHLs predicted in PDBs 1OAO and 2YR5 where the helical bounds were supplied by PSIPRED, it became impossible for loop buildup to be performed – all possible helix conformations produced loop halves that were considered impossible to close.

Rather than seek to expand the size of our helical dihedral library to include more rotamers, we found it more effective to attempt loop-helix-loop prediction with shorter helical bounds, one that would be less likely to demonstrate a lever effect. This led to the use of our truncated helix sampling method. We leave it up to subsequent rounds of further minimization and sampling to form the

remainder of the helix, and indeed this appears to be effective. Nonetheless, for very large helices, our limited dihedral library may fail to contain a sufficient number of rotamers to avoid a sampling error and the truncation method may leave too large of a sub-loop to correctly sample and form the remaining coupled dihedrals that are necessary to complete the helix. In practice though, this is not a very large concern for us. Such large helices are likely the well-conserved regions between homologous proteins. Knowledge of these helical bounds would likely be found with sequence-based secondary structure prediction methods, but crucially, the conformation of these large loop-helix-loops lies squarely within the purview of our previous rigid helix placement algorithm[67].

Hairpins, somewhat surprisingly, appeared as a simpler type of secondary structure to predict. The small non-locality of the hydrogen bonds deterred us from wanting to introduce a separate hairpin dihedral library as such a library would seem to produce a bias in the non-hydrogen bond turn-region of the hairpin between the two β-strands. Rather, we attempted loop-hairpin-loop prediction using only our previous dipeptide-dihedral library[37]. Low RMSD loops were successfully predicted to atomic accuracy with no significant change to our past algorithms, other than permitting a flexible *ofac* to be tried throughout all rounds of hierarchical loop prediction. For both hairpins and loop-helix-loops, it would be desirable in the future to further establish this methodology by running blind tests where the structure of a given loop is available but unknown to the researcher. However, we do not anticipate the results of such experiments to diverge from what we present here as our method is automated, using only the energy and not user input, to determine the final loop conformation.

Predictions performed in a non-native surrounding environment were successful, albeit requiring additional sampling and the use of our rotamer frequency score to accurately predict the loop. An

apparent caveat is that the additional degree of freedom now present in the surrounding environment can magnify energy errors. As shown in the hairpin in PDB 2C0D, incorrect protonation state assignment of an aspartic acid is compensated for through the coupled movement of a surrounding environment residue. Although only this case had such a problem, clearly more experiments need to performed across a large set of loops, with and without secondary structure, to expose weaknesses in our algorithm and correct them. These experiments are already underway and will be discussed in a future publication.

Chapter 4 : Improving the accuracy of homology model loop prediction in antibodies

In this chapter, we extend the application of Plop to loops in increasingly non-native environments and introduce a new scoring term that penalizes loops containing extremely uncommon backbone structure. With this scoring term, it is possible to increase sampling yet avoid the propagation of candidate loops that are energetically reasonable but never seen in nature. However, additional effort is needed to efficiently obtain the level of sampling required to reach sub-Angstrom accuracy in full homology models. We conclude this chapter with a discussion of new sampling methods in active development.

4.1 Introduction

Homology modeling applies computational methods to determine the three-dimension structure of proteins, using their known sequence and the experimentally determined structures of homologous proteins. Accurate structural knowledge is key to understanding and manipulating a protein's function and biological properties, but structure determination via X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy remains time-consuming and often challenging.

The grand promise of homology modeling is to bridge the gap between the number of known sequences and the significantly smaller number of solved structures. This is one of the explicit components of the National Institute of Health (NIH)'s Protein Structure Initiative, a program begun in 2000 with the goal of making the atomic-level, three-dimensional structure of most proteins easy to obtain from sequence[86]. Once enough structures representing all of the major

protein families are experimentally determined, the reasoning goes, homology modeling of sequence homologs can flesh out the structures within each family. Known structures could then catch up with sequences, and biochemists interested in studying the function of various proteins could draw on a wide swath of structural information.

To be as biochemically useful as a crystal structure, though, homology models must be reliably produced at high accuracy: generally, this involves predicting the position of non-hydrogen atoms within <1 Å of their actual positions.

A further complication is that some of the most interesting regions of proteins are often the most difficult to model. Accurate atomic-level knowledge of binding and active sites is desirable for studying ligand docking and protein-protein interactions, but the prevalence of loops among the secondary structure elements in these regions pose a challenge.

In template-based homology modeling, one or more sequences with known structures are aligned to a sequence of interest to build a model. This method is effective for secondary structure regions that are highly conserved across homologous proteins, but leaves something to be desired in accurately modeling loops and other poorly conserved regions. Predicting and refining loop regions is, therefore, a necessary complement to template-based homology modeling.

Skepticism abounds over the practical, biochemical utility of homology models and the scope of the Protein Structure Initiative[87]. But homology modeling and computational design has proliferated in the area of antibody modeling, beginning with humanization efforts in the 1980s[8, 88-89].

Antibodies are well suited for template-based modeling. The fragment antigen binding (Fab) fragments of an antibody consist of heavy and light chain constant domains (C_{H1} and C_L) and variable domains (V_H and V_L), with the variable domains at the N-terminus forming the antigen binding site. The variability in the V_H and V_L domains come from the hypervariable, or complementarity-determining regions (CDRs) – as with the constant domains, they have a highly conserved beta barrel as the core framework. The hypervariable region in each variable domain consists of three loops, denoted L1, L2, L3 in V_L and H1, H2, H3 in V_H. These loops are positioned near the surface of the antibody and determine the specificity of the antibody-antigen interaction[90].

The range of conformations observed for five of the six hypervariable loops fall into a set of discrete structural classes, the "canonical structures" first classified by Chothia and Lesk in 1987[91]. Using this scheme, these five loops can be classified by their sequences and modeled accurately using known structures from the same class. Yet the H3 loop persists in its extreme variability and so far defies classification – and accurate prediction via template-based modeling[8].

This diversity is a key factor in making antibodies effective, and a full and accurate representation of the hypervariable region is important for predicting and evaluating binding activity. Reliable H3 loop prediction is therefore essential for antibody homology modeling to fully achieve its potential.

In this work, we present the results of *ab initio* H3 loop predictions in a set of antibody homology models using the Protein Local Optimization Program ("Plop"). We also describe a new scoring term that has been added to Plop that penalizes loops constructed with sets of dipeptide dihedral angles that are rarely observed in the Protein Data Bank (PDB).

Up until this point, Plop has been shown to successfully re-predict loops of up to 20 residues in the native environment[37]. Recently, Plop was used to refine loops in constructing a homology model of the human β2-adrenergic G-protein coupled receptor[50]. Previous studies have also employed Plop in predicting antibody H3 loops in the crystal structure[92], and in an environment with native and nonnative elements, though with diminished accuracy[38].

With the addition of the dipeptide dihedral penalty term, we now demonstrate that Plop can predict antibody H3 loops in full homology models with accuracy comparable to or better than the current state of the art[93]. While this is discussed in the context of antibody modeling, our loop prediction method remains completely general and applicable to the broader loop modeling problem.

4.2 Methods

Overview

We predicted the H3 loop in two sets of antibodies with Plop, applying a new scoring term to penalize dipeptide dihedrals from unpopulated regions of phase space in candidate loops. The first set consists of "partial" homology models, where the six hypervariable loops were modeled from templates grafted onto the native framework region. The second set consists of "full" homology models, constructed completely from templates using Schrödinger, Inc.'s Bioluminate software package[94]. Hypervariable loops throughout were identified using the Chothia definition[91].

Test sets

The "partial" model set consists of the thirteen antibody cases containing a five to eight residue H3 loop analyzed by Sellers, et al[38]. Fourteen antibodies from the benchmark antibody set compiled by

Sivasubramanian, et al.[93] comprise the "full" homology model data set. H3 loops in these structures range from 7 to 12 residues in length. Loop sequences for each set are provided in Tables 4.1 and 4.2.

These cases were chosen because each native crystal structure, used as a reference point, does not contain any antigen. This allowed us to reliably compare the predicted H3 loop, which often undergoes structural changes upon binding[8], to its counterpart of known structure. While relevant to homology modeling in some cases, predicting an antigen's structure and/or precise orientation relative to an antibody homology model upon binding is a separate question not explored in this work.

PDB ID	H3 Length	H3 Sequence
1A7Q	8	ERDYRLDY
1CR9	5	DLHDY
1FLR	7	SYYGMDY
1KCV	7	GGTGFPY
1MEX	5	EYLDY
1MJU	7	NKLGWFP
1NGZ	5	RDSDY
1UAC	5	WDGDY
1UB6	6	GQGRPY
1UJ3	8	DSGYAMDY
1UZ8	8	ETGTRFDY
1YQV	7	GNYDFDGW

Table 4.1 H3 loop sequences for partial model test set

PDB ID	H3 Length	H3 Sequence
1A6T	8	RDDYYFDF
1CGS	7	GYSSMDY
1FGN	8	DNSYYFDY
1IGM	12	HRVSYVLTGFDS
1IGT	9	HGGYYAMDY
1JPT	8	DTAAYFDY
1KEM	8	WGSYAMDY
1MCP	11	NYYGSTWYFDV
1MLB	7	GDGNYGY
1VFA	8	ERDYRLDY
2ADG	11	HEDGNWNYFDY
2AJU	10	YDYYGNTGDY

Table 4.2 H3 loop sequences for full model test set

Model Construction

The partial models were built as described by Sellers, et al.[38] and briefly recounted here. For each variable light and heavy chain, all three CDR loops were removed from the scaffold. A model for each loop, excluding CDR H3, was selected based on sequence identity from a library of canonical antibody CDR loops, classified using Martin et al.'s definition[95]. The criteria for sequence identity were highest identity between the template and target, and a cap of 60% identity to avoid selecting the native conformation. Template structures were then grafted onto the native scaffolds using Plop's homology model tool.

For loops that either did not match the canonical sequence definitions (<75% of residues fitting the class rules) or did not find a template with a high enough sequence identity, the database was searched using different length target sequences until a template was found. That is, for an N residue loop sequence that did not fit the canonical class rules or match with a template, the sequences for

loops N-1, N+1, N-2, N+2, etc. are used to search the database until a template is found and gaps or insertions are modeled as needed using Plop when the loop is grafted onto the scaffold.

While the CDR loops were generated separately for the heavy and light chains, grafting these loops onto native scaffold ensured that the native H:L chain orientation remained constant. After the non-H3 loops were constructed on the scaffold, the H and L chain PDB files were simply concatenated. In each case, the H3 loop was left truncated and re-predicted *ab initio* using the Plop methodology discussed below.

The full homology models were built using BioLuminate, part of the Schrodinger Biologics Suite[94]. BioLuminate's antibody modeling process is described in detail by Zhu, et al.[96] and summarized here.

The model building proceeds in two primary steps: selecting templates for the framework, or scaffold, region and for each of the CDR loops. The scaffold sequence is used to search a curated antibody structure database (from PDB structures) via direct sequence alignment, using the Smith-Waterman algorithm[97] and the BLOSUM62 scoring matrix[98]. Heavy and light chain sequences are searched in pairs and the template with the highest similarity after averaging the heavy and light chain similarity scores is selected.

Templates for the CDR loops are chosen using a three-step process:

1. Loop sequences and conformations are identified separately for each of the six CDR loops.
2. For each CDR loop, the corresponding loop database is filtered based on three criteria: (a) target sequence, (b) loop length, (c) stem residue geometry. The stem residue geometry is

defined by the distance, angles, and dihedrals linking the first and last loop residues to the remainder of the protein (i.e., C-alpha and C atoms adjacent to the N-terminus of the loop and N and C-alpha residues adjacent to the C-terminus).

3. Filtered loop candidates are aligned by stem residues and clustered based on backbone RMSD. In each cluster, a representative loop is selected by highest sequence to target within that cluster. Then, starting with the largest cluster, the representative loop templates are checked against a sequence similarity cutoff. The template for each CDR loop is the first loop to exceed the similarity cutoff; in the event that no loops reach this score, the highest-similarity loop from all clusters is used.

Once templates are obtained for each component of the loop, an initial model is built by inserting conserved backbone and side chain residues into the scaffold region, mutating non-conserved scaffold residues, and grafting in the CDR loops. Finally, all non-conserved residues undergo side-chain optimization and minimization. This refined homology model thus becomes the input for H3 loop prediction.

Template framework and loop structures for each full model are listed in Table 4.3.

	Light Chain Scaffold	Heavy Chain Scaffold	L1 Loop	L2 Loop	L3 Loop	H1 Loop	H2 Loop	H3 Loop
1A6T	2OSL_L	2OSL_H	1DQD	15C8	1HIM	1I9I	2OSL	1A2Y
1CGS	1PLG_L	1PLG_H	1A4J	1M7D	1FPT	1BQL	1YQV	1PLG
1FGN	1D5I_L	1D5I_H	1PG7	1IQW	1DN0	1ZA3	1AFV	1KEL
1IGM	3BN9_C	3BN9_D	1XF4	2Q1E	1NGX	3CXD	3HI5	3MA9
1IGT	2ZUQ_E	2ZUQ_F	1RZ7	1KCS	1IGT	1FL3	1IEH	1IGT
1JPT	3EO9_L	3EO9_H	3HB3	3KLH	3MO1	1NJ9	3I75	2G60
1KEM	2CJU_L	2CJU_H	1L7T	2IGF	3IFO	1Q9K	2V17	1XGY
1MCP	3HZY_A	3HZY_B	1MVU	2MCP	1VGE	1MCP	3L1O	1MCP
1MLB	3FMG_L	3FMG_H	1DQJ	1UAC	1DQQ	1I9I	2IFF	1HQ4
1VFA	1FNS_L	1FNS_H	1AR1	1PSK	3DSF	1NC2	1P4I	2GHW
2ADG	1I8I_A	1I8I_B	1I8K	1I8I	1I8I	1OPG	3FFD	1A6U
2AJU	3LEY_L	3LEY_H	1KN2	1N0X	1AE6	1AY1	1OSP	2UUD

Table 4.3 Templates used to build antibody homology models.

Loop Prediction in Plop

The standard procedure for predicting loops of 4 to 20+ residues in Plop has been reported extensively[36-37, 39, 51]. Here, we provide a summary of the general protocol and describe the key parameters for sampling loops in a completely non-native environment.

Plop carries out *ab initio* loop prediction, generating loops with a hierarchical algorithm and scoring them with a physics-based energy model, VSGB2.0[30]. This energy model includes an optimized Generalized Born implicit solvent model and physics-based corrections to the OPLS molecular mechanics force field.

Loop prediction begins by sampling a large number of possible conformations using a backbone rotamer library, which is a discrete collection of all possible ϕ,ψ dihedral angle pairs – the backbone equivalent of widely used side chain rotamer libraries. Loop conformations with steric clashes

and/or no space for side chain atoms are screened out, and remaining loops are clustered. Cluster centroids proceed into iterative side chain addition, minimization, and energy scoring.

Many iterations of this single loop prediction procedure constitute a full loop prediction. In an initial stage, multiple single loop predictions proceed in tandem with varying overlap factor (OFAC) parameters, or cutoffs for steric clash screening. The top *n* lowest energy, non-redundant loops from this stage are passed on to a refinement stage, where the loops are predicted again, subject to a 4.0Å Cartesian constraint on the C^{α} atoms. The lowest energy loops are again passed on to subsequent predictions. Fix *N* Stages, where *N* residues at each loop terminus are fixed, allow long loops to benefit from increased sampling, and *N* can be increased with loop length. Finally, a second refinement stage is carried out with a 6.0Å Cartesian constraint on the C^{α} atoms, and the lowest energy loop across all predictions is deemed the predicted loop.

To reliably generate native-like loops in a homology model context, additional sampling is needed to compensate for a nonnative surrounding environment. In this work, we used the "HLP-SS" protocol described by Sellers, et al.[38] to iteratively sample surrounding side chains conformations while adding loop side chains in the same way. Likewise, during minimization, neighboring residues' backbone atoms were allowed to move. These considerations partially address the possibility of an inaccurate surrounding environment steering the loop away from the native conformation.

Additionally, during the initial stages, the number of clusters was increased from 4 x (number of residues) to 6 x (number of residues). While there is a small computational cost associated with optimizing more conformations, the expanded number of loops competing for refinement in successive stages increases the likelihood of a native-like loop emerging.

Dipeptide Dihedral Rotamer Frequency-Based Scoring Term

We also introduced a new term for loop scoring in Plop: the dipeptide dihedral Rotamer Frequency-based Scoring term, or RFS. It penalizes loops containing one or more dipeptides with a dihedral (ϕ, ψ, ω, ϕ, ψ) angle set found in highly unpopulated regions of phase space, based on comparisons to a library of dihedrals found in PDB structures.

To implement this term, it was necessary to construct a library of experimentally observed dihedral angles. We analyzed a slice of the PDB limited to high resolution (< 2.0 Å) x-ray structures and culled it using the PISCES web server[69] to eliminate structures with a percent identity > 95%, to eliminate duplicate crystal structures. Every backbone dihedral between consecutive residues in the remaining structures was inspected and stored, separated out by amino acid pair, and binned at five-degree increments. Each resulting discrete (ϕ, ψ, ω, ϕ, ψ) angle set is considered a 'backbone rotamer.' The number of angles binned into each rotamer was also stored in the library.

During loop scoring, this library is used to identify dipeptides whose dihedral angle sets fall well outside populated regions of phase space. Two criteria are used to determine if a dipeptide will incur a penalty, as depicted in Figure 4.1:

1. The Euclidean distance between the predicted dipeptide's dihedral angle set and the nearest library rotamer;
2. The number of rotamers within a certain radius of the predicted dipeptide's dihedral angle set.

The penalty is then added to the total energy as a pseudo-energy term.

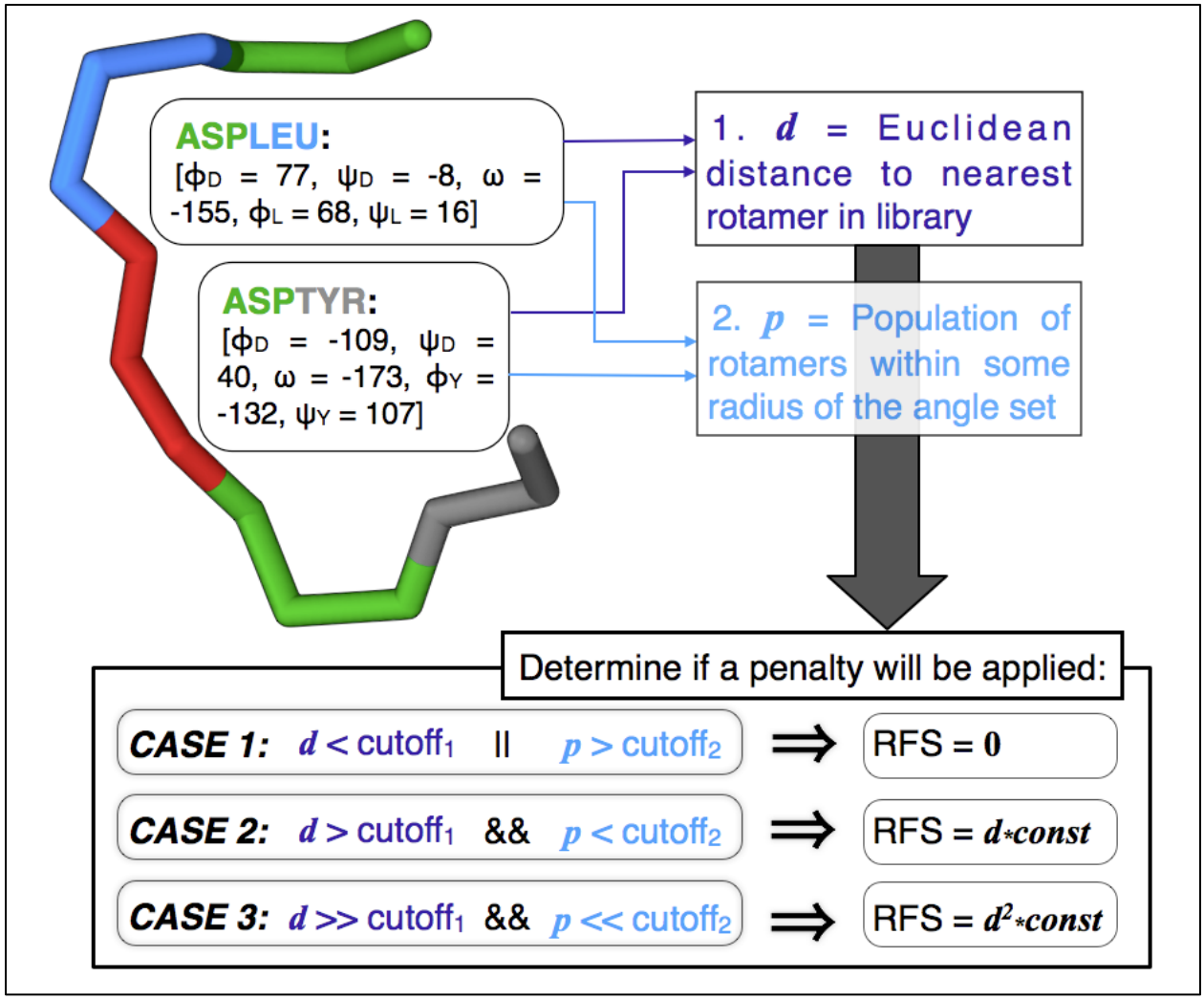

Figure 4.1 Flowchart for how the RFS is calculated.

Examples of rotamers are shown in the top left for a hypothetical loop fragment. A dipeptide rotamer is the set of five angles describing a pair of residues; the library consists of five-mers of rotamers observed, for the specific amino acid pair, in the PDB. The logic behind calculating the RFS penalty is illustrated in the box at the bottom.

The Plop user can specify each of these variables: distance *d*, radius *r*, population within the radius *p*, and scaling constant *c*. In this initial application, we used a distance cutoff of 23, a population cutoff of 40 rotamers, a radius of 23, a scaling constant of 0.4 in the Initial, Refinement 1, and early Fix stages of full loop prediction, and a scaling constant of 0.8 in the final Fix and Refinement stages.

These values were empirically chosen after analyzing a large set of predicted loops, but further optimization of these parameters via additional testing over a larger data set is pending.

Each adjacent residue pair in the loop is compared to its amino acid pair-specific library in this way. The first and last non-loop residues are included with the terminal residues, so all loop residues are scanned in two dipeptide combinations. If a penalty is incurred, it is added to final energy of the loop and therefore contributes to the energetic ranking of predicted loops, both at the end of a full Plop prediction and in individual stages when loops are being passed on to successive stages.

Measuring Accuracy

We evaluated the accuracy of our predictions using root mean square deviation (RMSD) in atomic coordinates compared to the native crystal structure. RMSD was measured, using Plop, over all backbone heavy atoms on the loop. Where needed, the model was first aligned to the native structure; alignments were carried out over all backbone heavy atoms in the protein. However, the RMSD does not factor in to the scoring of loops generated by Plop; the resulting loop reported for each Plop prediction is the final lowest energy conformation, and the RMSD is measured after selecting this loop based on the energy for purely evaluative purposes.

4.3 Results

Predicting the H3 Loop in Antibody Partial Models

We re-predicted the H3 loop in the antibody partial models using three different schemes: a.) buildup with single peptide backbone rotamers and scoring without the RFS term; b.) buildup with

dipeptide backbone rotamers and scoring without the RFS term; and c.) buildup with single peptide backbone rotamers and scoring with the RFS term.

All three of these scenarios used a full loop prediction in Plop, as described in the "Methods" section. No Fix stages were used, because the loops in this test case are relatively short (ranging from 5 to 8 residues in length). Additionally, the number of initial loops generated was not increased from the default value of 4 x (number of residues).

For comparison, we also re-predicted the H3 loop in the native crystal structure. In these predictions, the single peptide backbone rotamer library was used for buildup, no RFS was applied, Fix stages were omitted and surrounding residues were not optimized. Essentially, the Plop parameters developed for accurate homology modeling were not employed, because the surrounding environment is considered precise when starting with a crystal structure.

The full data for each case is tabulated in Table 4.4. It is clear that re-predicting native loops in the crystal structure is trivial for these loops, with a median and mean RMSD of < 0.5 Å. However, this level of accuracy disappears when the H3 loop is re-predicted in the context of CDR loops modeled from templates.

		Crystal Structure (HLP)	No RFS, Single Peptide	No RFS, Dipeptide	RFS, single peptide
PDB ID	**H3 Length**	**RMSD**	**RMSD**	**RMSD**	**RMSD**
1A7Q	8	0.8	1.3	0.6	1.0
1CR9	5	0.3	4.0	0.5	0.5
1FLR	7	0.2	4.0	2.0	0.6
1KCV	7	0.7	2.7	1.2	1.0
1MEX	5	0.3	0.6	--	0.7
1MJU	7	0.3	4.9	1.0	1.1
1NGZ	5	0.3	1.3	--	0.8
1UAC	5	0.3	0.5	--	0.6
1UB6	6	0.4	3.8	0.9	0.7
1UJ3	8	0.3	0.8	0.8	0.4
1UZ8	8	0.7	5.2	1.6	1.6
1YQV	7	0.4	2.0	1.5	1.0
	Mean	0.4	2.6	1.1	0.8
	Median	0.3	2.3	1.0	0.8

Table 4.4 H3 loop prediction results in antibody partial homology models.

Note that 6 residues is the minimum loop length for using the dipeptide rotamer library (column 5).

Without the RFS penalty and building up loops from single peptide backbone rotamers, the mean RMSD is 2.6 Å. Using dipeptide backbone rotamers improves this to 1.1 Å, but for three of the cases, every loop fragment was screened out during buildup. Applying the RFS penalty improves the mean and median RMSD to 0.8 Å, and loops are successfully built in each case.

Figure 4.2 more thoroughly demonstrates the benefit of applying the RFS penalty to these cases. Beyond an improvement in mean and median RMSD, it is able to eliminate extremely inaccurate loops and effectively shrink the range of predicted-loop RMSDs.

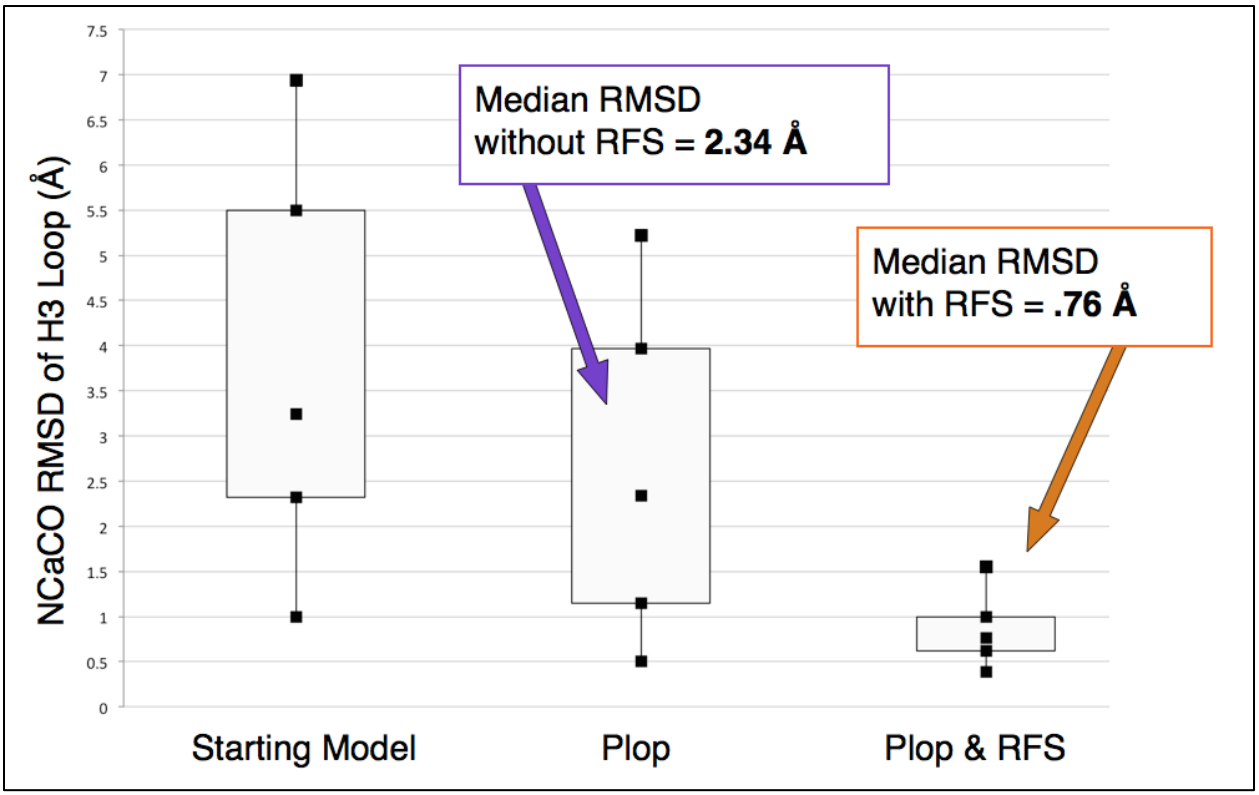

Figure 4.2 Improvement in H3 RMSD-to-native with RFS in Plop.

The maximum, minimum, and median points are shown and the box denotes the 25th and 75th percentiles. "Starting model" is the H3 loop built from template, "Plop" is the Plop H3 prediction without the RFS, and "Plop & RFS" is the Plop H3 prediction with RFS.

Many cases see a relatively insignificant improvement in RMSD with the addition of the RFS penalty. 1A7Q improves from 1.3 Å to 1.0 Å, 1UJ3 improves from 0.8 Å to 0.4 Å, and two cases get marginally worse: 1MEX goes from 0.6 Å to 0.7 Å, and 1UAC from 0.5 Å to 0.6 Å. These are trivial differences, however, and the results remain quite accurate with and without the RFS penalty.

In other cases, though, applying the RFS penalty makes a large difference in predictive accuracy. 1FLR improves from 4.0 Å to 0.5 Å, 1MJU improves from 4.9 Å to 1.1 Å, and 1UZ8 improves from 5.2 Å to 1.6 Å.

Of course, the loop prediction problem is simplified in these cases because the non-CDR regions making up the scaffold of each antibody retains the native conformation. Once we achieved accurate results with these training wheels on, we progressed to predicting H3 loops in full homology models.

Predicting the H3 Loop in Antibody Homology Models

As discussed in the "Methods" section, the starting models in these predictions were predicted entirely from templates using the BioLuminate software[94]. The cases were chosen from the benchmark set of Sivasubramanian, et al.[93], and limited to those with antigen-free crystal structures.

The H3 loops in these structures were predicted using a full loop prediction in Plop, including Fix stages and optimization of surrounding residues. The RFS penalty was applied throughout each stage and the single peptide library was used to build up loops.

Again, the H3 loop was also repredicted in the crystal structure context using the basic Plop algorithm without the RFS penalty and without optimizing surrounding residues. This can be considered the baseline accuracy level for repredicting these H3 loops.

Case-by-case results are listed in Table 4.5. The mean RMSD of predicted H3 loops in the homology models using the default parameters is 2.9 Å. More notably, almost half of the loops - 42% - were predicted with RMSD < 1.5 Å. Because we know the native loop conformation for these test cases, we were able to identify and further investigate the poorly predicted cases. For each of these cases, we ultimately identified a set of parameters to sample and select a native-like loop.

The mean RMSD is 1.1 Å when various parameters are employed and the lowest overall relative energy loop is chosen from all parameters. This level of accuracy is on par with what we can achieve in the crystal structure context and is therefore an extremely encouraging result.

Crucially, in each of these cases, the eventual native-like loop had a lower relative energy. Applying the same tests further improved two additional cases with already acceptable results. These specific parameters are outlined below.

			Any Parameters		Standard Parameters		Other Parameters		
PDB ID	# Res	Native Predicted RMSD	RMSD	Final Energy	RMSD	dE	RMSD	dE	ddE = other dE - std dE
1A6T	8	0.6	1.17	-7432.53	2.83	-91.60	1.17		
1CGS	7	1.5	1.64	54769.64	1.64	-16349.27	--		
1FGN	8	0.5	1.38	-5388.62	4.72	-105.18	1.38	-1347.32	-1242.14
1IGM	12	0.5	1.00	-6217.92	1.00	-70.56	--		
1IGT	9	1.0	0.78	-6598.23	0.78	-775.11	--		
1JPT	8	0.3	1.37	-7233.29	1.37	-795.29	--		
1KEM	8	0.8	1.38	-5969.80	1.38	-95.00	--		
1MCP	11	0.3	0.71	8982.84	3.78	-103.17	0.71	-131.67	-28.50
1MLB	7	0.2	0.67	7144.34	1.22	-80.09	0.67	-54403.2	-54323.2
1VFA	8	0.5	1.62	-7463.00	1.94	-74.12	1.62	-3833.03	-3758.91
2ADG	11	0.6	0.88	162066.5	10.35	-67.47	0.88	-134.56	-67.09
2AJU	10	0.5	0.79	1323.10	3.38	-149.89	0.79	-585.00	-435.11
Mean	**8.9**	**0.6**	**1.12**		**2.87**		**1.03**		

Table 4.5 H3 loop prediction results in antibody homology models.

Cases with nonstandard parameters are explained in detail in the text. The ddE calculated in the last column demonstrates that the "other" parameters obtained a final, lowest-energy loop that is lower in relative energy than the final, lowest-energy loop predicted using standard parameters.

Additional surrounding sidechain sampling

In the HLP-SS method, the list of surrounding sidechains to sample along with loop sidechains is generated by building in a rough set of possible loop backbones and selecting all side chains within a cutoff distance – here, 9 A. However, case 2ADG was drastically improved – from a lowest-energy loop close to 4 A to under 1 A – by sampling a specific set of surrounding side chains, some of which were not placed on the initial list. The sidechain list that allowed the native-like loop to be sampled was generated by selecting all residues within 9 A of loops generated on a fully minimized antibody scaffold. Interestingly, the full loop prediction on this structure, with this extra residue list, did not generate any native-like loop conformations. Using the native structure to generate an extra residue list, similar to the previous method of identifying interfacial residues on the native structure, also did not sample a native-like loop. It seems like some or all of the eight residues on the minimized extra residue list are critical to our ability to accurately sample this loop. Generalizing this discovery to our overall methods is examined in the following discussion.

Predicting multiple loops

While many of the non-H3 CDR loops can be accurately modeled from templates, in some cases, structural deviations in a nearby CDR loop can preclude accurate H3 loop prediction. In 1A6T, we found that simultaneously predicting the H3 loop and residues L:89-97, the L3 loop, allowed us to reach a near-native RMSD for the H3 loop, a reasonable RMSD for the L3 loop, and a lower relative energy than predicting the H3 loop alone. A full description of the multiple loop algorithm will be presented in a forthcoming publication. Briefly, we implemented multiple loop prediction by predicting each loop separately with the opposite loop coordinates deleted. Then all candidate conformations for each loop are iteratively paired and pairs with steric backbone clashes are

eliminated. Pairs of loops are then subject to further screening based on the position the C-beta atoms on loop residues, to ensure that there is sufficient room in the space of the pair of loops to fit reasonable conformations for each of the loop side chains. Remaining structures, now containing two loops, undergo minimization, sidechain optimization, and energy evaluation. These combined-loop structures are then clustered and promoted or eliminated, as in a standard full Plop prediction. This is carried out during each individual, standard Plop prediction. Predictions in Ref and Fix stages use the same protocol, adding in the constraints and shortened loops as defined previously for these stages.

Localized sampling and refinement

Improving our ability to refine candidate loops is an area of active development, and cases 1MCP and 2AJU were found to need enhanced refinement to find lowest-energy native-like loops. We tested two things: (1) whether it was possible to build a native-like loop conformation in these models and if so, (2) how the energy of such a structure would rank compared to the predicted, non-native-like loops. To answer the first question, the native loop was inserted into the homology model. Then, a localized sampling algorithm was used that made very small changes to the starting dihedral angles of the loop. After using this high-resolution sampling on the entire loop, the middle of the loop, residues 99-102, was sampled again. Many of these candidate loops were thrown out for sidechain or backbone atom clashes, and the energy was calculated for remaining loops. In both 1MCP and 2AJU, the final lowest-energy loop after localized sampling had a sub-Angstrom RMSD. But more importantly, the final energy of each of these loops was *lower* than the final energy of any predicted loop. While this is also not a generalizable protocol for homology model loop prediction, it demonstrates that a native-like loop in each of these homology models is indeed the lowest energy

conformation. Therefore, the problem is narrowed down to a sampling problem, not an energy problem, and it should be possible to *a priori* select such a loop when we are able to sample it.

Sampling the heavy:light chain interface

Three cases, with PDB IDs 1FGN, 1MLB, and 1VFA, improved significantly when the interface between the antibody heavy and light chains was sampled throughout the loop prediction. This was carried out by:

1. Identifying the list of interfacial residues to be sampled: The "structure" toolkit in Plop was used to identify residues within a certain distance of a defined interface. Here, we used a 9 A distance and defined the interface as the H3 loop, or the command "structure interface 9 H:95 H:102" in Plop.
2. Implementing the INTFC stage: Sampling of these interface residues was carried out between loop prediction stages in Plop. All loops promoted by the Init, Ref1, and the final FixN stages were passed on to an INTFC stage. All resulting structures from the INTFC stage were then passed on to the subsequent Plop stage; no additional clustering was done during the INTFC stages.
3. Sampling the interfacial residue list: During the INTFC stage, residues on the interfacial residue list each structure underwent a complete sidechain optimization, in the presence of each predicted loop structure passed on from the previous stage.

For these cases, the first step was carried out on the native PDB structure and that "real" residue list was used in the INTFC stages for homology model loop prediction. While this would not be a real

world protocol, it shows that it is indeed possible to sample a lowest-energy native-like loop in these cases, with implications discussed later on.

4.4 Discussion

In this study, we predict the highly variable H3 loop *ab initio* in two sets of antibody models using Plop, with additional sampling and a new dipeptide dihedral frequency-based penalty term. The first set retains native coordinates in the antibody scaffold but modeled the other five CDR loops from templates; these structures are more amenable to H3 loop prediction than full models. We also show the ability to predict H3 loops in completely non-native homology models with reasonable accuracy with a generalized protocol, and with high accuracy when the protocol is adapted based on knowledge of the native structure. Because we are able to pinpoint the best, most native-like loop using our energy function, the next step towards a high accuracy, generalized protocol is improving our ability to sample native-like loops over the course of the prediction.

Dipeptide Dihedral Rotamer Frequency-based Scoring Term

A key facet of the Plop algorithm is constructing candidate loops from a backbone rotamer library. Like a set of side chain rotamers, this library contains sets of possible dihedrals – in this case, ϕ and ψ angles. Previous work[37] introduced a dipeptide rotamer library to more efficiently sample conformations of very long (> 13 residue) loops. Where the single-peptide rotamer library contains one list of possible (ϕ,ψ) torsion angles for all residues except glycine and proline, the dipeptide library has a specific set of observed (ϕ, ψ, ω, ϕ, ψ) torsion angles for each of the 400 possible dipeptide pairs between standard amino acids.

The premise of this improvement was to temper the explosion of generated loop candidates in building up many-residue loops, as well as to improve the realized effective sampling resolution. However, because it is composed of dipeptide segments in the PDB, it also serves to focus the sampling towards likely dipeptide conformations.

For example, there are approximately twice as many possible dipeptide rotamers for an Ala-Ala segment as for an Arg-Arg segment, but the single peptide library uses the same set of rotamers to build each residue in both segments. The dipeptide library therefore provides a level of specificity that is not imparted via the single peptide library.

However, the single peptide library clearly provides more coverage for individual residues, and provides more exhaustive sampling for loops whose length provides no combinatorial computational challenges. The rotamer frequency-based scoring term (RFS) term uses the information encoded in the dipeptide rotamer library to penalize loop dipeptides whose torsion angles fall in highly unpopulated regions of the five-dimensional (ϕ_1, ψ_1, ω, ϕ_2, ψ_2) space, analogous to the two-dimensional Ramachandran plot.

An updated dipeptide rotamer library containing frequency information was constructed (see "Methods") and used to compute the penalty pseudo-energy term, which is included in the total energy calculated for each candidate loop.

As discussed above, each dipeptide pair in the loop is evaluated using a sliding window scheme such that each individual residue is taken twice, including the first and last residue in the loop, which are

evaluated with the adjacent non-loop residues. The set of five torsion angles for each dipeptide is compared to the new library to find two values:

1. Euclidean distance to the nearest rotamer in the library;
2. Population of rotamers within a given radius of the set of loop torsion angles.

A penalty is applied to dipeptide segments with a large distance to the nearest library rotamer and a small population of nearby rotamers. This penalty is scaled and added to the energy calculated using the physics-based force field.

Example

The mechanics of this penalty term are more easily illustrated by example. In Figure 4.3, we show two predicted H3 loops for the antibody with PDB ID 1UB6 alongside the native structure. These are the loops with the lowest energy after all stages of loop prediction starting from the same partial model; the penalty term is the only difference in the input parameters.

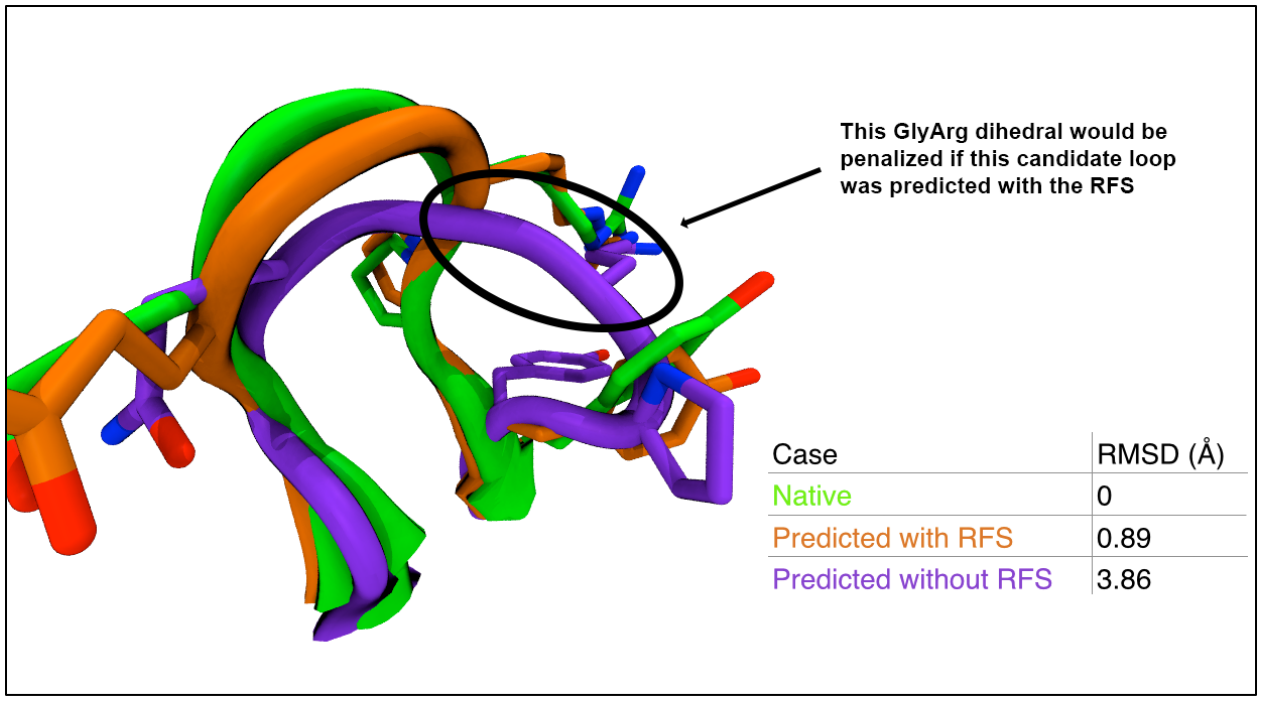

Figure 4.3 Effect of RFS term on 1UB6 H3 loop prediction.

It is clear via visual inspection that the loop predicted with the penalty is much closer to the native conformation, and the 0.9 Å RMSD-to-native is a major improvement over the 3.9 Å RMSD of the prediction without the penalty term.

Analyzing each dipeptide pair, as Plop does when applying the penalty term, we find that the Gly-Arg dipeptide dihedrals (residues H:99 to H:100) in the no-penalty prediction are far from any neighboring dipeptide rotamers in the GLYARG library. The observed dipeptide rotamer's distance from the nearest rotamer is d =174.67. In addition to exceeding the distance cutoff of 23.0, there are zero rotamers within a radius (r = 0.45) of the observed rotamer, so the formula $(\text{distance})^2$ x (scaling constant) is used for a greater penalty. Therefore, a penalty of $((d^2) \times c) = ((174.67)^2 \times 0.4) =$ 12203.4 would be applied to this loop, likely eliminating it from the lowest-energy position. For comparison, the final energy of this case without the penalty is -7917.9. It is therefore not surprising

that when this penalty is applied at each stage of a full loop prediction, this Gly-Arg rotamer is not found in the final candidate loops. In fact, a 0.9 Å RMSD loop emerges as the lowest energy loop.

Because Plop initially built up the 3.9 Å RMSD loop from acceptable single peptide rotamers, it may seem strange that a resulting dipeptide dihedral combination occurs so far from anything in its amino acid-specific library. Amino acid identity considerations (e.g., using the same library to build Val-Ala pairs as Tyr-Trp pairs) can explain some penalties, and averaging of C^{α} positions in loop closure may alter the middle dipeptides. Finally, the all-atom minimization of loop residues can nudge dipeptides away from the library-provided torsions. When this is propagated through refinement stages that constrain loop C^{α} atoms, dipeptide dihedrals from highly unpopulated regions of phase space, such as those constituting the Gly-Arg segment in the poor 1UB6 prediction, may be observed in predicted loops.

From Partial Models to Homology Models

Applying the dipeptide dihedral penalty term significantly improves the accuracy of H3 loop prediction in nonnative environments. We are able to predict H3 loops in partial models with accuracy on par with loop re-prediction in the crystal structure environment, and our results in full homology models are on par with the state of the art[99].

Successful loop prediction in partial models is a necessary but not sufficient test for homology model loop prediction. Partial models, as constructed here, isolate the problem to building a loop in an inexact environment, where nearby loops are modeled and the non-loop scaffold regions are exact. This is a relevant test in antibody models, as the six CDR loops form the variable region that

binds to antigen. The starting point for H3 loop prediction in these models is a binding pocket largely modeled from templates.

The partial model results demonstrate that we can accurately generate native-like H3 loops amid template-modeled, inexact CDR loops. With the addition of the RFS penalty term, we eliminate candidate loops containing highly unlikely dipeptide dihedral combinations, and we identify the native-like H3 loop using energy as computed by our physics-based force field, reaching sub-Angstrom accuracy.

Moving to full homology models, where the entire antibody is constructed from structures with homologous sequences, escalates the environmental challenges to predicting the H3 loop. In addition to inexact placement of backbone and/or sidechain atoms in nearby loops, the positioning of nearby loops along the non-loop scaffold can shift relative to the H3 loop endpoints, and the H3 loop stem can also deviate from its native position.

Components of Plop's sampling algorithm have been optimized, as detailed in previous work[38], to circumvent these problems and generate a swath of candidate loops including those with native-like conformations. Among these, two stand out as essential for allowing a low-RMSD loop to emerge. First, surrounding residues' backbone and side chain atoms are allowed to move during minimization. While this introduces more dimensions to the minimization, which is carried out on each candidate loop after clustering, it prevents errors in the surrounding environment from precluding correct loop positions. Second, initial predictions are run with OFAC values of 0.3 to 0.7. Extending the OFAC down to 0.3 essentially allows some steric clash between candidate loops and the surrounding backbone. In subsequent stages, a higher OFAC is used and such clashes can be

refined, but early on, this is important in sampling a diverse set of loops that ideally includes one or more candidates similar to the native loop.

However, any increases in sampling carry a computational cost – both in terms of the computing power required to carry out additional sampling and minimization, and in terms of complicating the problem of parsing out and propagating the best loops. Merely generating loops from a broader area of conformational space is not enough: the downstream steps in loop prediction must also improve to reliably convert increased coverage to accurate results.

In this work, we found that, for several cases, this extended sampling combined with the ability of the RFS to weed out egregious dihedral pairs was still not enough to generate and select a native-like loop. For each of these cases, we investigated where the algorithm fell short: was in generating promising loop conformations, selecting and propagating these conformations, or both?

An exhaustive look at the hundreds of candidate loops generated over each stage of these problem predictions (PDB IDs 1A6T, 2ADG, 2AJU, 1MCP, 1MLB) showed us that the first issue was definitely a factor. Using the standard homology modeling parameters outlined above, no sub-2A loops were predicted at any stage for these cases. The problem was not screening out good loops, or a failure of the energy model to promote good loops - the problem was that Plop did not find these loops in the first place.

To investigate where the sampling fell short, we experimented with various parameters and protocols as outlined in the Methods section. In each case, we were ultimately able to sample a native-like loop with customized parameters, guided by the ability to calculate the RMSD of

predicted loops. As mentioned previously, this would not be possible in a true homology model test – but is a crucial diagnostic tool towards developing a protocol for true homology modeling.

Yet while these customized parameters are not generally applicable to homology modeling, they show that with specific, focused sampling, Plop successfully identifies a native-like loop as the lowest energy conformation. This is important because it narrows the scope of the problem to the sampling algorithm, not the energy function. To answer the question posed previously: the issue is in generating promising loop conformations.

Developing improved sampling algorithms

Here, we explain how the outcomes of these experiments are guiding future development of Plop loop prediction with an eye towards overcoming these sampling issues in true homology model cases, where we cannot use a native structure as a guide. Three key areas have emerged for improvement: sidechain sampling, multiple loop prediction, and enhanced loop buildup.

It's clear from these test cases that occasionally, specific non-loop-residue sidechain positioning can be critical to determining a loop's structure. Further, when starting from a homology model, it may be initially unclear what residues in the surrounding region are critical to the network of hydrogen bonding interactions that support the native conformation. One possible way of ameliorating this uncertainty is simply to re-predict all sidechains in the protein, or within some large distance from the loop region.

However, as currently implemented, any brute force approach re-predicting a large number of surrounding sidechains would drastically increase the computational requirement for Plop predictions - and more crucially, run up against the limitations of the existing sidechain optimization algorithm. To compensate for this issue, we first implemented the interfacial sampling stage in Plop (see "Methods"). However, because this stage uses the same sidechain optimization algorithm, it does not solve the issue of reliably optimizing a large number of surrounding sidechains. Further, our approach in these experiments hinged on knowledge of the native structure. With a more robust sidechain optimization algorithm, we could sample a wider swath of residues along the homology-modeled H:L interface, confident that this larger radius would both include the key native-interface residues and result in a converged, fully sampled result.

A new method to reliably optimize sidechains in less ordered systems, like homology models, is currently in development in our group and will be disclosed in detail in a forthcoming publication. Briefly, this algorithm will allow us to achieve superior side chain sampling by completely disregarding the input coordinates for each side chain and explicitly pre-computing the gas-phase pairwise energy of each interacting side chain rotamer state, for each side chain, as in an Ising model. We then are able to run a large number of Monte Carlo iterations rapidly, as the cost to evaluate the energy of each side chain conformation is trivial - the process is simply a matter of reading the energy from the previously generated table. Implementing this method will make optimizing larger zones of surrounding sidechains tractable and efficient as a standard part of the homology model loop prediction protocol.

In other cases, optimizing surrounding sidechains is not sufficient – neighboring loop backbones need refinement to allow the loop of interest's native conformation to emerge. This was the case for

PDB 1A6T; predicting the nearby L3 loop was necessary to generate a native-like H3 loop. Multiple loop prediction is an important tool for cases where a neighboring loop has weak homology to known structures, is placed such that it may have substantial interactions (potentially clashes) with the loop of interest, or has possible backbone interactions with the loop of interest.

The sidechain optimization and multiple loop prediction algorithms address challenging surrounding environments, but in other cases, the sampling problem could not be isolated to either of these factors. Native-like loops for cases 1MCP and 2AJU were not build up successfully, but when the actual native loop was inserted and refined in the homology model, resulted in a lower energy than any loop built up in Plop. Here, the question is how to initially construct a truly diverse set of loops that either includes one or more native-like structures, or includes structures that can be refined to a native-like structure.

One way of addressing this current limitation is by improving the way loops are initially built up in Plop. Currently, it is possible to generate enough loops to hit memory limits in the loop buildup phase without even trying most of the possible rotamers – i.e., sampling at a very low resolution. In some cases, the upper limit of about 100,000 half-loops does not get in the way of sampling at a finer resolution, because steric clashes in the surrounding environment limit the number of possible half-loops. However, in highly solvent-exposed loops, or loops where the surrounding environment is somewhat disordered and also being sampled, the existing loop buildup process may not actually build up candidate loops from rotamers representing remotely native-like dihedrals.

A new ordered buildup routine, currently in development, will address this issue by scoring and ranking rotamers at each point in the half-loop construction. Instead of starting at the lowest (worst)

resolution and arbitrarily saving every rotamer set that does not clash with the surrounding environment until the memory limit is reached, this new algorithm will apply more nuanced screening criteria and scores to possible rotamers and build the set of loop halves from the most promising dihedral pairs. As side-chain atoms are not yet placed during backbone sampling, scoring is performed using an energy function, currently in development that treats side-chain atoms implicitly. In comparison with the hard, steric-like screens currently being used to prune rotamers during buildup, the application of even an approximate scoring function is expected to triage some of the more egregious loop candidates and improve the likelihood of finding a native-like loop within memory limitations. We are confident that these improvements will fix the sampling problem for these cases and generate loops that can be refined into the low-energy, low-RMSD structures we have identified.

Chapter 5 : Antibody Structure Determination Using a Combination of Homology Modeling, Energy-Based Refinement, and Loop Prediction

In this chapter, we present a complete, fully automated antibody homology modeling program and benchmark its performance against other methods from the community in the Antibody Modeling Assessment II.[iii]

5.1 Introduction

Computational structure prediction of antibodies is an important step in the modeling, engineering, and design of novel antibodies with desired therapeutic properties. The variable domains (Fvs) of the heavy and light chain are of special interest, as they typically impart most or all of the specificity of an antibody for its antigen target. The Fv can be further divided into the hypervariable regions and the framework regions (FRs). The hypervariable regions are so called complementarity-determining-regions (CDRs), which are composed of 6 hypervariable loops on the surface of the antibody, thereafter denoted as H1, H2 and H3 of the heavy variable domain (VL), and L1, L2 and L3 of the light variable domain (VH). Due to the highly conserved nature of the framework regions of the VL and VH domains, research into Fv structure prediction has largely focused on the prediction of the 6 CDR loops. Analysis of antibody crystal structures led to the discovery of "canonical" classes for the five non-H3 CDR loops in the 1980s and 1990s[91, 95, 100-101]. Antibody structure predictions based on this type of analysis and categorization are often qualitatively successful for loops L1-L3 and H1-H2, although predictions of H3 are more problematic. Recently, Dunbrack and coworkers performed clustering of a new expanded set of crystal structures (> 1300

[iii] Reproduced with permission from *Proteins* 2014 *82* (8), 1646-55. Copyright 2014 Wiley Periodicals Inc.

antibody structures in the PDB) and proposed 72 clusters for the five non-H3 loops[102]. Approximately 85% of the non-H3 sequences can be assigned to one of these conformational clusters based on gene source and sequence. Prediction of the H3 loops remains difficult, however. In contrast to the other CDR loops, the H3 loops are extremely diverse in the length and conformation. Although there have been a number of attempts, no satisfying classification has been possible for the H3 loops[103-107].

The modeling of non-H3 loops has traditionally relied on the canonical classes, beginning with those based on the first studies that derived loop classifications from only a small number of crystal structures[91, 100]. However, with the rapidly increasing number of the antibody structures in the PDB, the methods that use sequence similarity and other generic criteria in homology modeling instead of antibody rules have become popular and perform comparably well to the canonical classes based approach[108]. An advantage to these generic homology based approaches is that new structures can be added to the dataset more easily than with the older canonical class analyses. The large number of available structures has revealed the limitation of sequence-based analysis and its predictive power. For example, Martin and Thornton observed that a loop might be closer in sequence to one class, but structurally belongs to another[95]. Clearly, the loop conformation is not determined by its sequence alone — the interactions between a loop and its environment must be considered to make an accurate prediction.

On the other end of the spectrum of the loop modeling techniques are *ab initio* prediction methods[16, 18, 34, 36-37, 51, 109]. These methods generally use a discretized rotamer library to sample the conformational space and a scoring function or energy function to rank the candidates. One of the advantages of *ab initio* methods is that they are independent of the protein structure database and

thus can be used when no suitable template is available. With progress in sampling techniques and increased sophistication of the energy function, *ab initio* prediction methods have demonstrated high accuracy in the prediction of loops generally[30, 110-113], and in the prediction of loops for antibodies[92-93, 114] and other protein families[49, 115].

In this chapter, we describe our approach to antibody structure modeling and present our results for the Antibody Modeling Assessment II (AMA-II)[99, 116]. Antibody Modeling Assessment is a community-wide blinded test of the state of the art antibody modeling methods, and is similar in form to CASP[1]. The first assessment (AMA-I) was organized in 2011, in which four approaches (three software packages and a web server) were evaluated by predicting nine then-unpublished antibody crystal structures[108]. For AMA-II, six groups and an automated Web server attempted to predict 11 then-unpublished antibody crystal structures, and the results were used to benchmark the modeling performance. The AMA-II assessment consisted of two stages. In the first stage, the participants were asked to predict the full Fv structures of 11 antibodies from sequence. In the second stage, the crystallographic coordinates of each of the antibody structures *minus the coordinates of the H3 loops* were made available, and each participating group was then asked to predict the coordinates of the H3 loops. The second stage of the assessment is new to AMA-II, and reflects the known difficulty in predicting H3 loops.

We have developed a novel knowledge-based method to predict the CDR loops using a combination of sequence similarity, geometry matching, and conformational clustering of the database structures. The homology models are optimized with a physics-based energy function (VSGB2.0[30]), which we show significantly improves the quality and accuracy of the models. Subsequently, we present and discuss the results using our *ab inito* approach for the second stage of

the assessment--H3 loop predictions in the context of the crystallographic structure scaffolds--and compare these to the results of the same approach when carried out in the context of homology models for the remainders of the Fv regions.

5.2 Materials and Methods

Figure 5.1 depicts the flowchart of the steps in our antibody homology modeling protocol. Our protocol starts with a template search for the framework region (FR) in our curated antibody database, which is derived from the publically available crystal structures in the Protein Data Bank (PDB). Instead of a heuristic search algorithm such as BLAST[117] or PSI-BLAST[118], we do a direct alignment of the query sequence to every sequence in the database using the Smith-Waterman algorithm[97] with BLOSUM62[98] for the scoring matrix. Because there are only about 1200 antibody structures in the curated database, this direct alignment can be done relatively quickly, usually in 1-3 seconds. We select a matching pair of light and heavy chain templates from a single antibody template. There has been discussion in the literature[119-120] regarding the question of whether one should use the light and heavy chain templates from a single structure, or whether selecting them from different templates (requiring subsequent structural alignment) might be preferable. Although there is no systematic benchmark study, there is evidence that using both chains derived from a single antibody template offers some advantage[119]. The framework region and CDRs are defined according to the Chothia numbering. The templates are ranked by the average framework sequence similarity of the heavy and the light chain, and the best template is chosen as the one with the highest average similarity.

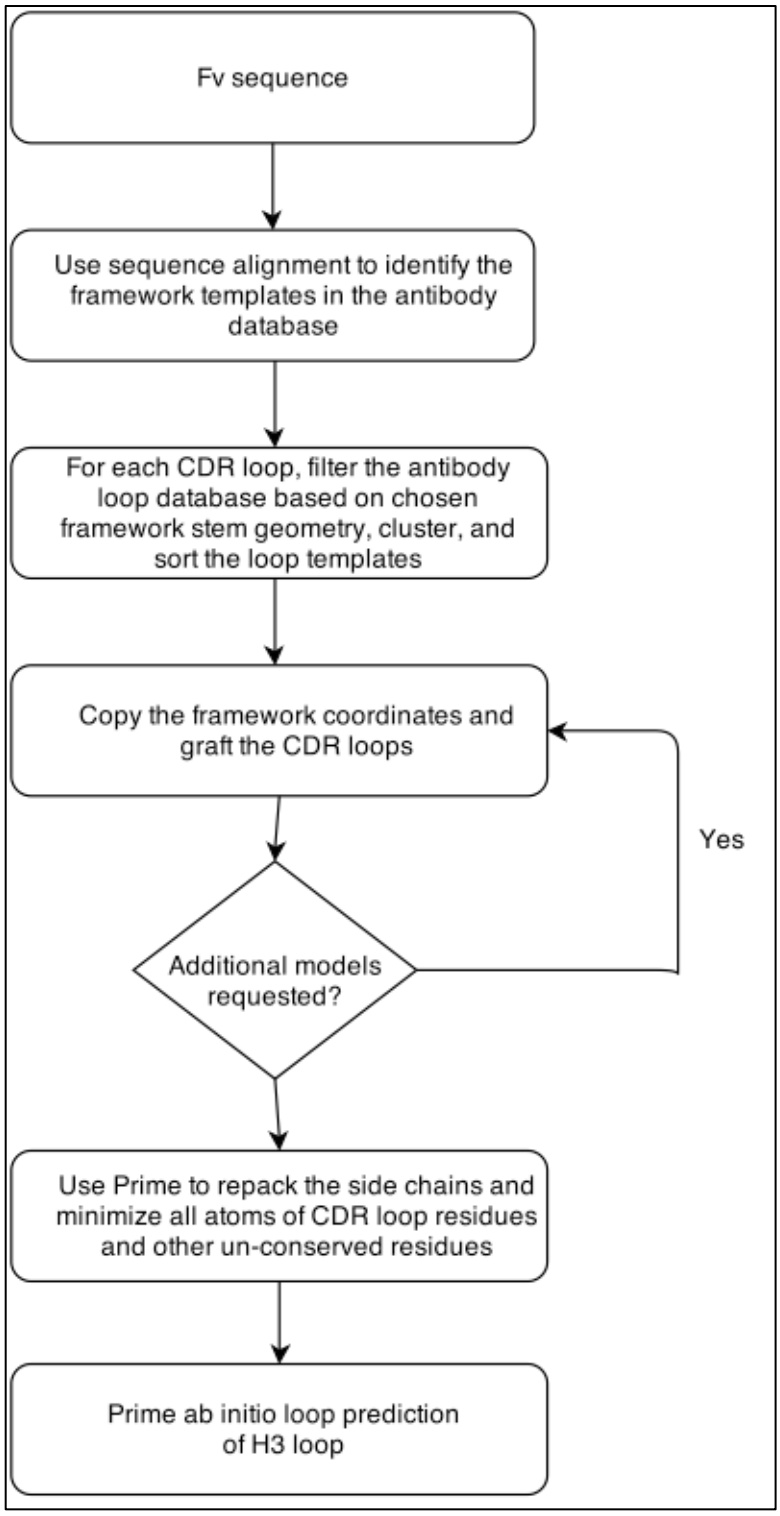

Figure 5.1 The antibody homology modeling flowchart.

The selection of the templates for CDR loops has three steps. First, a set of loop sequences and conformations is derived individually for each loop position (L1, L2, L3, H1, H2, and H3). Next, each of the six resulting loop databases is dynamically filtered based on the query sequence, loop length, and the stem residue geometry of the framework template that has been selected. The stem residues are the adjacent residues in the N- and C- termini of the loop. The stem geometry is defined using the distance, angles, and torsions by the Cα and C atoms in N-terminal stem residue and the N and Cα atoms in C-terminal stem residue[121]. After the filtering for each of the six loops, all remaining loop candidates are clustered with a complete linkage algorithm based on their backbone RMSDs with the stem residues being aligned. The clusters are ranked by the cluster size, and the "representative loop" of each cluster is defined as the one with the highest sequence similarity (as defined by BLOSUM62) to the query loop within that cluster. If the sequence similarity of the representative loop in the largest cluster exceeds a sequence similarity cutoff, this loop candidate will be chosen for the template; otherwise, the representative loop in the second largest cluster will be checked against the sequence similarity cutoff, and so on. If none of the representative loops in any cluster exceeds the similarity cutoff, then the representative loop with the highest sequence similarity will be chosen. The similarity cutoff for H3 loops and non-H3 loops are 0.3 and 0.6, respectively.

Once the templates for the framework and the six CDRs are chosen, we construct the initial homology model by first copying over the backbone coordinates and also side chains for conserved residues in the framework region, then mutating the non-conserved residues in the framework, and finally grafting the CDR loops onto the framework homology model. The non-conserved residues of the framework, and all CDR loop residues, are subject to a rotamer search to remove clashes and are then minimized with the OPLS 2005 force field[122] in vacuum. Lastly, a side-chain prediction and

minimization using the implicit solvent energy model VSGB2.0 are performed on all the non-conserved residues to further refine the model.

In the AMA-II, each modeler was asked to submit three models. The workflow, as described above, was used to produce our model #1. Our model #2 was generated by using the template loop candidate with the highest sequence similarity regardless the CDR loop clusters (framework selection was not changed). For model #3, we used Prime *ab initio* loop prediction to re-predict the H3 loops on model #1. The loop prediction follows the protocol in our previous study[92], and the side chains within 5 Å were also repacked simultaneously.

The second stage of the antibody modeling assessment was based on the well-known observation (also reflected in the aggregate results for the first stage of this assessment) that the H3 loop is generally the most difficult loop to correctly predict. Each participating modeling lab was challenged to predict the H3 loop conformations for a set of unpublished crystal structures, given the Fv crystal structure coordinates without the H3 loop. Our prediction strategy follows the protocol in our previous study[92], which is based on the Prime loop prediction method, but with some slight variations. Before the loop prediction job, the protein structure was prepared with Protein Preparation Wizard[122] available in Maestro 9.4,[123] which assigns the polar hydrogen positions, protonation states, and amide group flips. Because the starting structure does not have the coordinates for the H3 loop, we run the "preparation-then-prediction" process twice, first on the provided structure without H3 loop, and then again once a loop has been modeled in the H3 position. The loop predictions are performed on the H3 loop plus one extra residue on each end to make the loop terminal residues fully flexible. The rest of the model is kept fixed. Five models are submitted for each target, ranked by Prime energy function.

5.3 Results and Discussion

The Homology Model Accuracy

Table 5.1 and 5.2 show the PDB templates for constructing our homology models (the number 1 model submission) and the RMSD values of different regions to the crystal structure. The 11 antibody targets are denoted as AM1 to AM11. All CDRs are defined according to Chothia numbering. The structure alignment uses the Cα atoms, and the RMSD is calculated based on the backbone atoms (N, Cα and C). The first target is a rabbit antibody, which does not have high similarity to any templates in the PDB, especially for the light chain. The sequence similarity of the best template is significantly lower than other antibodies, which results in significantly worse RMSDs. (In the AMA-II assessment, no participant was able to produce an acceptable model for the rabbit antibody due to lack of homologous data in existing crystallographic databases. Prediction of this structure was deemed a failure for all participants and it was subsequently removed before the second phase of the competition[99, 116].) Excluding the rabbit antibody, the average RMSDs of the Fv region and of just the frameworks for the 10 targets are 1.19 Å and 0.74 Å, respectively. The five non-H3 CDR loop RMSDs range from 0.61 Å to 1.05 Å. Not surprisingly, the H3 loop remains to the most challenging constituent of the structure to predict, with an RMSD of 2.91 Å. In comparison, all modeling groups in AMA-II generate similar results on the FR and non-H3 loop predictions with no appreciable differences in the average RMSDs[99]. Predictions of the H3 loops exhibit a larger spread among the different groups and our results rank in 3rd place by the average RMSD. Note that the RMSDs in ref. 29 are calculated using only backbone carbonyl atoms and the values are slightly different from here.

Model1	FR		L1		L2		L3		H1		H2		H3	
AM1	2X7L	0.87	3LMJ	0.54	3L95	0.71	3MLW	0.15	2VXS	0.86	3MLX	0.60	2BDN	0.50
AM2	4H20	0.95	3O2D	1.00	2I9L	1.00	3O2D	0.88	1F11	1.00	4HK0	1.00	1EHL	0.73
AM3	2XTJ	0.99	2A6D	0.91	1RZI	1.00	3CMO	0.63	1OPG	0.86	3GBM	0.83	1A2Y	0.63
AM4	3MXV	0.92	2R0L	1.00	1T4K	0.86	3IY0	1.00	1EGJ	1.00	1UWX	1.00	1FL6	0.63
AM5	3MLW	0.95	2JB6	0.71	1Q1J	1.00	2J6E	0.73	2B1H	1.00	4DGI	1.00	2R0L	0.75
AM6	3HR5	0.96	1VGE	1.00	2ZUQ	0.71	2J4W	0.89	2VXV	0.86	3GIZ	0.67	2AAB	0.50
AM7	1F58	0.97	1I7Z	0.87	4F9L	0.71	2ZCH	0.67	2AJU	1.00	3J1S	0.60	1QFW	0.63
AM8	2I9L	0.92	1AP2	1.00	1XCT	1.00	1MCP	1.00	1P2C	0.86	1P2C	0.83	1A6U	0.64
AM9	3HMW	0.98	3HI5	0.91	2JIX	1.00	1KCV	0.89	1C5C	1.00	3MLT	0.83	4DN3	1.00
AM10	2I9L	0.94	3W11	1.00	1XCT	1.00	1JRH	0.75	3IY3	0.86	1Z3G	0.83	1E6J	0.55
AM11	2OZ4	0.93	4DGI	0.91	4F2M	1.00	3D9A	0.89	4HK3	1.00	4FQH	0.83	2UUD	0.80

Table 5.1 The PDB templates for constructing the homology models (model 1, stage 1) and the corresponding sequence similarities.

The light and heavy chain templates are taken from a common template framework. The CDR loop templates are chosen according to a combination of stem residue geometry, sequence similarity and database loop clustering (see main text). For each column, the template PDB ID and sequence similarity are listed.

Align	ALL	FR	L	H	FRL	FRL	FRL	FRL	FRH	FRH	FRH	FRH
RMSD	ALL	FR	L	H	FRL	L1	L2	L3	FRH	H1	H2	H3
AM1	2.96	2.51	3.76	1.54	3.45	5.20	3.91	4.91	0.83	1.12	2.05	4.74
AM2	1.64	0.81	0.64	1.90	0.65	0.40	0.39	1.04	0.61	0.45	0.82	6.49
AM3	0.92	0.64	0.45	1.04	0.39	0.59	0.50	0.71	0.55	3.12	1.18	1.41
AM4	1.25	1.14	0.71	1.32	0.72	0.54	1.12	0.43	1.07	0.93	1.23	3.31
AM5	1.12	0.89	1.03	0.80	0.54	1.82	0.58	2.54	0.70	0.84	0.40	1.80
AM6	1.06	0.72	0.46	1.14	0.33	0.49	0.77	1.07	0.54	0.92	0.51	3.04
AM7	0.78	0.54	0.64	0.70	0.44	1.31	0.55	0.84	0.38	1.12	0.86	1.88
AM8	1.21	0.66	0.49	1.18	0.45	0.82	0.35	0.62	0.64	0.61	0.83	3.34
AM9	0.70	0.50	0.44	0.78	0.37	0.47	0.85	0.56	0.38	0.60	0.84	2.50
AM10	2.06	0.69	0.75	0.82	0.40	1.99	0.32	0.50	0.44	1.10	0.62	2.27
AM11	1.12	0.81	0.51	1.16	0.44	0.39	0.68	0.82	0.75	0.82	1.57	3.10
Avg	1.19	0.74	0.61	1.08	0.47	0.88	0.61	0.91	0.61	1.05	0.89	2.91

Table 5.2 The backbone RMSDs between the predicted antibody models and those of the associated crystal structures, divided into various structural elements.

The structure alignment uses the Cα atoms and the RMSD is calculated on the backbone atoms (N, Cα and C) in Å. The average RMSD for antibody target AM2-AM11 is given in the last row of the table. ALL: whole Fv; FR: framework; FRL: light chain framework; FRH: heavy chain framework.

The CDR Loop Homology Modeling Protocols: Clustering and Sequence Similarity

CDR loops can have identical sequences with very different conformations. This poses an issue for methods that rely on sequence alone to select the loop template. To overcome this issue, our primary model (model #1) is produced using a method that combines sequence similarity, stem geometry matching, and conformation clustering, as detailed in the Methods. Our second submitted model uses only sequence similarity to select the CDR loop templates. This experiment provides a

blind assessment of the two methods. Figure 5.2 compares the backbone RMSDs of six CDR loops generated by these two methods. The clustering method shows some advantage over the similarity method with smaller RMSDs on all six CDR loops. The biggest improvement is for L3 with an improvement of 0.4 Å in the RMSD, followed by H2 with an improvement of 0.3 Å. On average, the RMSDs are about 0.1-0.2 Å better with the clustering approach. It should be pointed out that this observation is based on a relatively small data set, which is not sufficient to make a definitive conclusion about these two methods. A large-scale test should be conducted in the future to compare these methods.

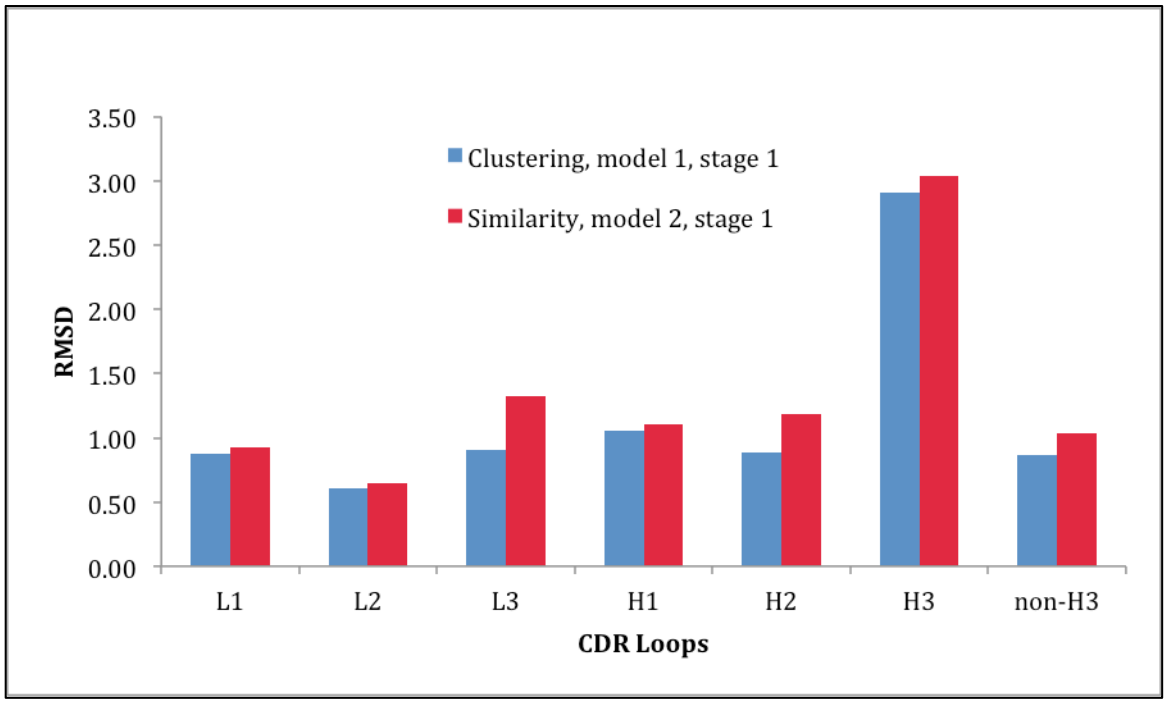

Figure 5.2 The backbone RMSDs of six CDR loop predictions using loop clustering and sequence similarity.

"Non-H3" is the average of five CDR loops excluding H3.

H3 Loop Prediction

In the first stage of this assessment, our third submitted model is different from the first model only in the H3 loop, which is based on *ab initio* prediction by Prime in the context of a homology model for the remainder of the Fv. Analysis of the blinded predictions for these loops in model #3 versus model #1 allows us to evaluate the ability this *ab initio* method to improve the predictions derived using standard homology templates. The loop prediction methodology in Prime has been extensively validated and recently we have shown encouraging H3 loop predictions in the context of crystal structure scaffolds, as well as when a highly homologous scaffold is available[92]. Table 5.3 shows the backbone RMSDs of H3 loop homology models and Prime predictions. On average, Prime refinement improves the backbone RMSD by 0.2 Å. This improvement, although small, would have put us in the 2nd place among AMA-II participants if we had submitted the *ab initio* models as the #1 models. The accuracy of *ab initio* loop prediction in the context of homology models is heavily influenced by the quality of the homology model itself, but this influence is largely local instead of global (i.e. the structures near the loop in question) in a way that is hard to quantify.

H3 RMSD	AM2	AM3	AM4	AM5	AM6	AM7	AM8	AM9	AM10	AM11	Avg
Homology	6.49	1.41	3.31	1.80	3.04	1.88	3.34	2.50	2.27	3.10	2.91
Ab initio	2.29	1.49	2.04	1.78	4.69	1.50	3.99	3.78	2.05	3.10	2.67

Table 5.3 The comparison between Prime ab initio H3 loop predictions and the predictions made by knowledge based homology modeling.

The homology models are the submitted model #1 in stage 1, and the Prime ab initio models are the submitted model #3. The accuracy is measured by backbone RMSDs in Å.

MolProbity

In constructing our homology models, we keep the side-chain conformations from the template for all conserved residues. The side chains of non-conserved residues and CDR loop residues are first optimized by a simple rotamer search to minimize the steric clashes. Then a Prime side-chain prediction is performed to optimize the side-chain conformations. Finally, all atoms on non-conserved residues and CDR loops are minimized with the VSGB2.0 energy function in Prime. In Table 5.4, we compare the MolProbity assessment of the homology models before and after Prime refinement. MolProbity score and clash score are improved significantly by the refinement. We should note that MolProbity is a measurement of structure model quality or self-consistency of the model. It does not necessarily correlate with the "correctness" of a model. Nevertheless, the side-chain accuracy after Prime refinement is also improved and for some models the improvements are substantial. The backbone accuracy does not change significantly (RMSD data not shown), which can also be seen from the relatively minor change in Ramachandran favored backbone torsions.

	Before Prime Refinement				After Prime Refinement			
	Clash score (Perc)	Rama favored	MolP score	Side Acc (%)	Clash score (Perc)	Rama favored	MolP score	Side Acc (%)
AM1	19.3(34)	87.6	3.2	38	7.3(85)	89.4	2.8	38
AM2	17.4(40)	93.7	2.8	45	8.7(78)	94.2	2.0	54
AM3	6.4(89)	91.6	2.4	51	4.6(95)	93.0	2.2	51
AM4	9.6(74)	94.4	2.4	49	1.8(99)	95.3	1.3	50
AM5	22.2(27)	93.6	3.0	36	5.7(92)	93.6	2.4	43
AM6	10.3(70)	95.5	2.6	41	5.9(91)	95.5	2.1	52
AM7	16.3(44)	94.5	2.6	54	3.8(96)	95.0	1.8	56
AM8	24.5(22)	93.7	3.1	44	15.5(48)	92.9	2.7	49
AM9	19.2(35)	93.1	3.0	48	11.7(64)	94.1	2.7	62
AM10	25.7(20)	92.4	3.2	46	15.9(46)	92.8	2.8	48
AM11	12.8(59)	94.0	2.6	42	4.5(95)	95.4	1.8	39

Table 5.4 Prime side chain prediction and minimization improves MolProbity score (model 1, stage 1).

Compared with the crystal structure, the refinement also improves the side chain accuracy. A correct side chain prediction is defined as the $\chi 1$ angle within 30 degrees of the corresponding crystal structure. "Side Acc" is the percentage accuracy of all side chains on the model. "Clash score" shows both the raw score and the percentile in parenthesis. "Rama favored" is the percentage of backbone torsional angles falling within favored Ramachandran region.

H3 Loop Prediction in the Context of the Crystal Structure Scaffold

In the second stage of the assessment, all modelers were given the 10 crystal structures (the first rabbit antibody was excluded) with the H3 loops removed and asked to predict the conformations

of the missing H3 loops. The purpose of this phase of the assessment is to determine how much the prediction depends on the scaffolds and how much improvement can be made if the perfect scaffolds are available. Figure 5.3 shows graphical illustrations of our predicted H3 loop (the first model among 5 submitted models) and the comparison with the crystal structure. Table 5.5 provides the backbone RMSD and side-chain $\chi 1$ angle accuracy of all 5 submitted models for each antibody, as well as our #1 model performance relative to that of other AMA-II participants. The average backbone RMSD of the first models is 1.28 Å. The 5 models are ranked by their Prime energy, and the model backbone accuracy is generally consistent with their ranking. Notably, our predictions rank the best among all AMA-II participants for seven of the ten targets in this category, and are within a fraction of an Å to the best model for two more of targets. In only one case (target 5) is our method significantly worse than the best approach by another group. Interestingly, in this case, one of our alternative models (#4 model) would have placed this prediction as the best among all groups.

	Model1		Model2		Model3		Model4		Model5		Model1 rank vs. other methods [29]
	RMSD	χ1	RMSD	χ1	RMSD	χ1	RMSD	χ1	RMSD	χ1	
AM2	2.78	63%	1.78	63%	2.54	50%	2.14	25%	2.81	38%	2
AM3	0.37	63%	1.28	75%	0.80	75%	1.14	75%	0.89	50%	1
AM4	0.65	75%	1.09	63%	1.27	63%	1.18	63%	0.74	75%	4
AM5	2.37	63%	3.42	38%	2.10	38%	0.36	50%	3.92	50%	6
AM6	3.11	46%	4.53	54%	4.04	31%	8.54	15%	6.08	31%	1
AM7	0.45	43%	1.67	29%	1.69	86%	1.15	57%	0.37	57%	1
AM8	1.25	57%	1.61	71%	3.57	57%	1.77	43%	1.33	43%	1
AM9	0.54	75%	3.40	38%	3.14	25%	3.72	50%	3.10	63%	1
AM10	0.85	63%	1.74	63%	3.30	63%	3.54	50%	5.49	63%	1
AM11	0.40	50%	1.46	75%	2.17	50%	0.89	38%	0.45	50%	1
Avg.	1.28	60%	2.20	57%	2.46	54%	2.44	47%	2.52	52%	

Table 5.5 Top 5 models for the H3 loop prediction in the context of the crystal structure scaffold.

For each model, the backbone RMSD and side chain χ1 angle accuracy are reported. Side chain accuracy is defined as the χ1 angle within 30 degrees of the corresponding crystal structure. The model 1 ranks among all participating groups in AMA-II are also reported in the last column based on Almagro, et al[99].

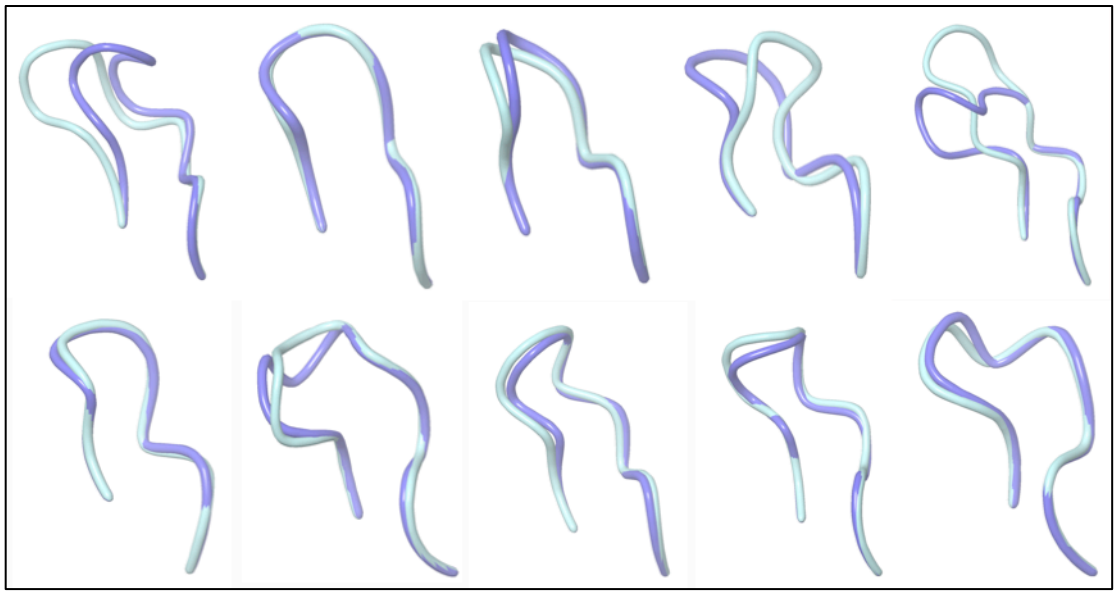

Figure 5.3 Graphical illustrations of the predicted H3 loop structures (model 1, stage 2) and corresponding crystal structures.

The crystal structures and the predictions are colored turquoise and blue, respectively. From left to right: top: AM2-AM6; bottom: AM7-AM11.

Considering the high accuracy of loop backbone predictions, it is interesting to examine how well the side chains are predicted. From Table 5.5, however, we do not see strong correlations of side chain prediction accuracy and the backbone RMSDs. For example, the side chain accuracy of model 1 averaged over the six H3 loops where the backbone RMSD is less than 1.0 Å is 63%; for the three H3 loops where the backbone RMSD is greater than 2.0 Å, the average side chain accuracy is 57%. The side-chain prediction accuracy for very accurate loop backbone predictions is not much better than that for the incorrectly predicted loops. A possible explanation for this is that the side-chain conformations are highly opportunistic — they can take distinctively different states depending on slight variation of the backbone positions. For buried side chains, the flexibility may be very limited, but surface residues can have much freedom to take different conformations without much energy costs. Figure 5.4 shows the side-chain predictions of AM7 and AM9. Both predictions have excellent

backbone RMSDs (0.45 Å and 0.54 Å, respectively), but side-chain predictions are very different: 43% vs. 75%. The buried side chains are always predicted correctly according to the crystal structure, but the surface residues may not necessarily assume the crystal structure conformations, although the loop backbones are predicted very accurately. This suggests that the observed difference in the side chain fine detail may not reflect a weakness of the prediction algorithm as much as the ability of these residues to adjust based on their environment.

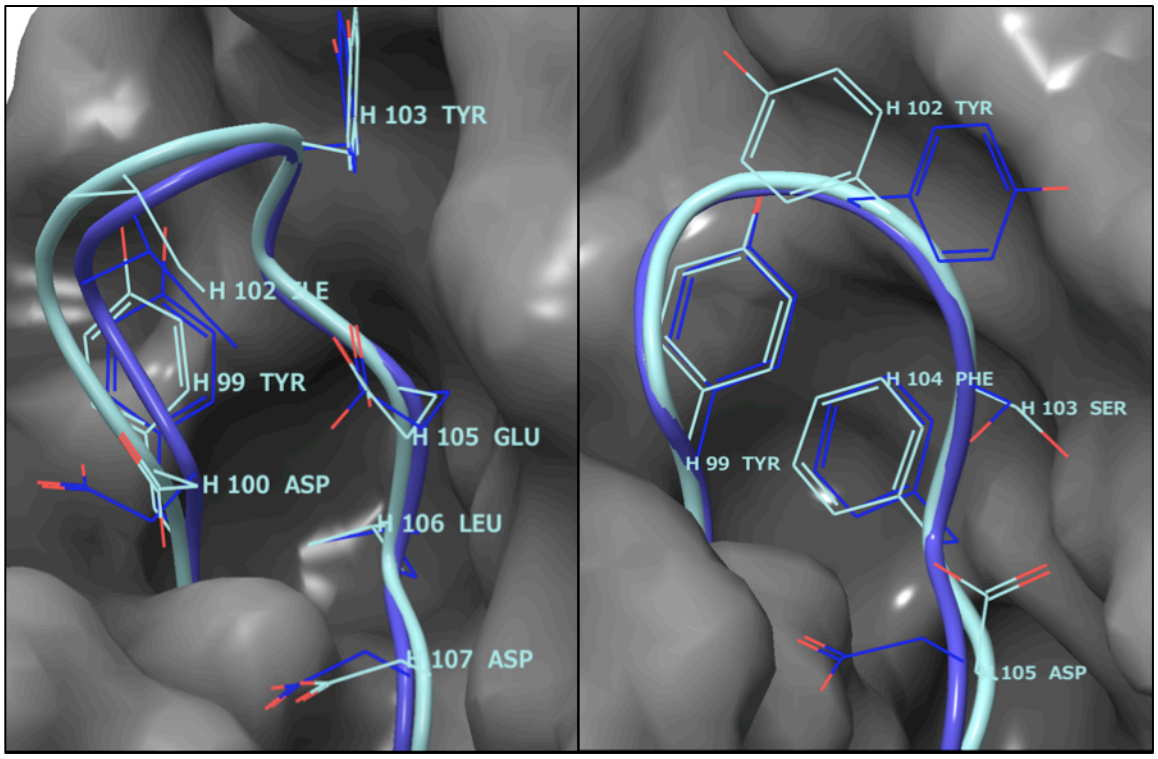

Figure 5.4 H3 loop Side chain prediction accuracies of AM7 (left) and AM9 (right).

The H3 loops are shown on the surface of the rest of the protein. The turquoise is the crystal structure and the blue is the prediction. The backbone RMSD and side chain accuracy for AM7 are 0.45 Å and 43%, respectively. The correct side predictions are H99, H100 and H104, and the incorrect predictions are H102, H103, H105 and H106. The backbone RMSD and side chain accuracy for AM9 are 0.54 Å and 75%, respectively. The correct side predictions are H99, H102, H103, H105, H106 and H107; the incorrect predictions are H100 and H108. Some of the side chains are omitted for clarity.

Knowledge-Based and Energy-Based Methods on H3 Loop Prediction

Table 5.6 shows a direct comparison of H3 loop predictions made using both homology modeling and the Prime *ab initio* method. As more antibody structures become available, the chances of finding a good template for H3 loop in the database will continue to improve. However, there is still much room for improvement for homology modeling, as demonstrated by the fact that homology model predictions are often significantly worse than the best available template (columns 3 and 4). The *ab initio* prediction method in Prime does not depend on the structure database, and its accuracy on crystal structure scaffolds is in most cases as good as the best template in the database. In fact, in several cases the *ab initio* H3 prediction is significantly better than *any* template in the PDB. On the other hand, Prime performance is sensitive to the loop environment. The relative advantages to homology modeling decrease as we move from a "perfect" crystal structure to an inaccurate homology model for the remainder of the structure. One of the reasons is that the Prime energy function is sensitive to structural errors in the "fixed" regions, such as shifted backbone positions, misplaced side-chain rotamers, and incorrect protonation states. A direct Prime energy evaluation of the knowledge based homology model with minimization usually does not yield favorable energy. Furthermore, the sampling problem is more challenging for homology models, as a small change in the nearby environment can greatly influence the generation of a structural candidate ensemble.

H3 RMSD (Å)	H3 Length	Best in Database	Using Crystal Structure		Using Homology Model	
			Homology Prediction	Prime Prediction	Homology Prediction	Prime Prediction
AM2	11	1.69	4.35	2.78	6.49	2.29
AM3	8	0.88	1.48	0.37	1.41	1.49
AM4	8	0.72	2.20	0.65	3.31	2.04
AM5	8	1.00	2.35	2.37	1.80	1.78
AM6	14	2.60	3.12	3.11	3.04	4.69
AM7	8	1.46	2.33	0.45	1.88	1.50
AM8	11	1.68	3.30	1.25	3.34	3.99
AM9	10	0.73	1.89	0.54	2.50	3.78
AM10	11	1.53	2.79	0.85	2.27	2.05
AM11	10	0.37	2.56	0.40	3.10	3.10
Average		1.26	2.64	1.28	2.91	2.67

Table 5.6 Comparison of H3 loop predictions with homology modeling and the Prime *ab initio* method.

Best in Database: The H3 loop in the PDB database with the best backbone RMSD to the crystallographic conformation of the target H3 loop. Using Crystal Structure: H3 loops built in the context of the remainder of the antibody structure taken from crystallographic coordinates. Using Homology Model: H3 loops built in the context of the remainder of the antibody structure taken from our "Model #1" submission from the first part of this assessment.

5.4 Conclusions

AMA-II has offered an opportunity for blinded testing of our approach to antibody homology modeling and refinement, as implemented in the programs BioLuminate and Prime within

Schrodinger Suite. Our homology modeling features a novel knowledge based approach to modeling the CDR loops, using a combination of sequence similarity, geometry matching, and the clustering of database structures. This method does not rely on the antibody canonical classes or other specific rules, and performs on par with the state of the art antibody modeling methods. Our homology models benefit significantly from the energy-based refinement, as demonstrated by the side-chain placement and H3 loop prediction. The *ab initio* loop prediction method in Prime performs very well when applied to repredicting the H3 loops in the context of crystal structures. Its accuracy on homology models degrades, but the method still performs better than the best database approach presented here. The refinement of homology models, in terms of reducing backbone RMSDs, still remains a very challenging problem. Its success depends heavily on the starting homology models and it should be used with caution. One particular situation where we have shown Prime works well[24] is when the template and target structure are highly homologous and the structural differences are relatively isolated (e.g. two antibodies with only differences in H3 loops and with minor changes in other CDR loops). This is a relatively common real-world scenario in antibody optimization, for which we expect our methods will be useful.

II. Applications in anti-HIV antibodies

Chapter 6 : Introduction

In this section, I discuss the application of the loop prediction methods outlined in the preceding chapters to study anti-HIV antibodies.

Human immunodeficiency virus-1 (HIV-1), the precursor to acquired immunodeficiency syndrome (AIDS), is a major cause of death worldwide. In 2013, the World Health Organization reported 2.1 million people newly infected with HIV – including 240,000 children – and a total of 35 million people living with HIV[124]. While antiretroviral therapy has drastically improved life outcomes for many HIV-positive individuals, there is still a great need for effective ways of preventing infection.

Currently, the only non-behavior-based preventative treatment option for HIV is pre-exposure prophylaxis, a daily course of antiretroviral therapy that can prevent viral replication and permanent infection upon exposure to HIVC[125]. This course of treatment has been shown to decrease the incidence of HIV in high-risk populations[126-129], but only when the daily protocol is consistently followed.

Developing a preventative approach that provides immunity against HIV therefore remains a major goal, both for at-risk populations and in preventing "vertical" transmission from mothers to infants.

The first step in this approach is identifying antibodies that can neutralize a broad spectrum of HIV strains. These antibodies would have therapeutic potential for "passive immunization," transferring the antibodies directly to the patient immune system, as well as targets for developing a vaccine that could elicit such antibodies.

The structural region of the HIV-1 virus most relevant to binding and infection is the outer envelope spike, which consists of glycoprotein gp120 on the surface and gp41 as the transmembrane domain; gp120 binds to the CD4 T-cell receptor in the cell which causes conformational changes to gp120 and instigates viral entry[130]. Based on their position and key roles, these surface glycoproteins seem to be an ideal target for antibodies to block binding to CD4 or subsequent receptors and prevent viral entry. But while this strategy has worked in other viruses, elements of HIV-1 envelope glycoprotein structure pose unique challenges to eliciting a neutralizing immune response[130-131].

Three main structural elements make HIV-1 a challenging neutralization target: (1) glycan shielding, (2) entropic barrier masking, and (3) antigenic variation[132]. Much of the gp120 molecule is covered with glycans, a type of carbohydrate that is not recognized as a non-self "attacked" by the immune system and therefore functions as a shield for the attacking virus. The multi-domain organization and intrinsic flexibility of gp120 may create a high entropic/conformational barrier that could prevent antibodies from targeting the receptor binding region[133]. Substantial sequence variation over particular regions in isolated HIV-1 structures may also prevent neutralization[134].

Despite this, many HIV-1-infected individuals have been found to develop broadly neutralizing antibodies that can provide immunity against many different strains of HIV-1[135]. While these

antibodies do not provide a meaningful clinical benefit to these patients, studies have shown that they are effective at passive immunization in animal models of HIV-1 transmission[136-137].

Successful monoclonal antibody therapeutics must be able to achieve therapeutic response in humans at a reasonable concentration in the blood after administration via injection. An antibody's effectiveness is dependent on its binding affinity to its antigen (a portion of the HIV virus, in this case). The antibody's concentration in the blood serum must be above K_D, the dissociation constant for the antibody-antigen binding reaction. To reduce the required concentration, and by extension the therapeutic dose, K_D must be also reduced. This can be achieved by increasing the binding affinity through optimizing the design of the antibody and/or epitope.

Here, we focus on computational tools for optimizing and designing more potent antibodies against HIV-1. Leveraging computational methods can save time and money in the antibody development process: modifications and alterations to the structure can be screened rapidly and cheaply *in silico* and the best performing leads can be passed along for synthesis and experimental testing. However, in order to optimize an antibody's binding affinity, we must have an accurate and robust method for predicting the change in binding affinity upon modification (typically, residue(s) mutation) to antibody-antigen complexes.

This technology exists in the form of free energy perturbation (FEP) simulations. FEP simulations calculate the free energy difference between two highly similar systems by taking an alchemical path between the two systems. To measure the relative binding affinity, or free energies of binding, between two protein-protein systems, an alchemical path transforms the wildtype protein into the mutated protein in both the complex form and the uncomplexed form. The intermediate changes in

free energy (dG) are calculated between (1) the complexed wildtype and mutated type and (2) the uncomplexed wildtype and mutated type. While these dG values have no physical corollary, the difference in dG (ddG) is equivalent to the relative free energies of binding that are experimentally measured separately for each of the wild and mutated types.

McCammon, Kollman and Jorgensen originally applied FEP methods to calculating protein-ligand relative binding affinities beginning in the 1980s[59, 138-143]. Through improvements to force fields, conformational sampling, and cost-effective parallel computing, advanced FEP implementations are now being successfully deployed in small-molecule drug discovery lead optimization workflows[144]. Protein ligands, however, pose additional challenges: namely, they are much larger than small molecules. This demands greater sampling and additional computational power. Focused sampling methods developed by Wang, et al.[145-147] as well as the ability to run simulations on graphics processing units (GPUs) brought small molecule ligand FEP to mainstream[144]. Now, we are focusing on extending these tools to protein-protein systems.

The pathway to designing more potent anti-HIV antibodies begins with developing a protein-protein FEP protocol that consistently produces accurate (under ~1 kcal/mol root mean square error between computation and experiment) predictions for the relative change in free energy. The methods can be tested in two ways: first in retrospect, by replicating existing experimental data, and then by prospectively predicting new candidate modifications to be experimentally validated.

In addition to tuning the parameters of the FEP simulation itself, a successful protocol will require accurate input structures for the simulation, and methods to evaluate the sampling carried out over

the course of the simulation. Non-free-energy-based, time-independent structure prediction methods can contribute to these tasks and supplement FEP in the overall optimization protocol.

The system we are using to build out and test this protocol is the VRC01 class of anti-HIV antibodies, which have been isolated and optimized by researchers at the NIH Vaccine Research Center (VRC)[134, 148]. VRC01 is one of many antibodies generated against the CD4-binding region of viral envelope glycoprotein gp120. It has unique structural features that go beyond simply mimicking CD4 binding, make it a highly potent and broadly neutralizing agent against HIV-1[134].

Experimental optimization has been able to further improve the binding affinity of the VRC01 antibody, but it would still require treatment on the order of grams per month to be effective. Improving the binding affinity by one to two orders of magnitude would mean only milligrams of the antibody would need to be delivered as a treatment, which is a more therapeutically and economically reasonable quantity.

VRC01 is also an ideal test system because of the wealth of structural and binding affinity data available. Nineteen crystal structures have been published for various VRC01-class antibodies alone and in complex with gp120 proteins from a range of HIV-1 clades, and binding affinity has been measured for multiple different mutations over several orders of magnitude. With both input structures and binding affinities to compare to the FEP output, this is a complete and biologically relevant data set for validating the FEP protocol.

In the following chapters, I will focus on the role of structure refinement and validation in protein-protein FEP simulations. In Chapter 7, I will discuss data confirming our ability to model key loop

regions in the VRC01-class systems. Then, in Chapter 8, I will present several case studies incorporating structure prediction and loop modeling into a complete FEP protocol. This work is a critical first step towards practical deployment of FEP simulations to study the VRC01-class system.

Chapter 7 : Loop prediction in VRC01-class antibodies

7.1 Introduction

The HIV-1 virus relies on its outer envelope spike to instigate binding and infection, consisting in part of glycoprotein gp120, which binds to the CD4 receptor in the target cell and leads to viral entry. Antibodies can therefore block HIV infection by targeting this CD4 binding site (CD4bs) on gp120. Researchers have recently shown a small subset of HIV-infected patients produce broadly neutralizing antibodies[149-151], and researchers at the NIH Vaccine Research Center (VRC) have identified broadly neutralizing antibodies in donor sera as binding to the CD4bs of gp120[152-154].

The first of these highly effective, broadly neutralizing antibodies, VRC01, was identified and reported by Wu, et al.[148, 155]. Since then, several similar antibodies have been identified and added to the VRC01 class[156]. VRC01-class antibodies share common features, in particular a heavy chain that mimics the CD4 receptor, five-residue CDR L3 loops, and a high degree of somatic mutation[155]. Typical antibodies see 5-15% of their variable region sequence change throughout the affinity maturation process[157], but the mature VRC01 sequence differs by over 40% from the germline[148].

This class of antibodies has now been studied extensively from many angles, including ontogeny of the class by Zhou, et al.[158], antigenic optimization by Joyce, et al.[159], identification of unusual somatic framework mutations by Klein, et al.[160], comparing epitope specificity and neutralization patterns by Georgiev, et al.[161], and crystallizing additional VRC01-like antibodies to understand maturation and binding by Wu, et al.[155] Over the course of these investigations, many VRC01-class antibodies have been crystallized and deposited in the PDB, creating a valuable corpus of crystal structures that can be used for computational optimization efforts.

In order to use rational, structure-based tools to further optimize VRC01 and related antibodies, we must first be able to re-predict structural features of the crystal structures. To that end, we present and discuss results of *ab initio* H3 loop predictions over the entire set of VRC01-class antibody crystal structures.

7.2 Methods

VRC01 Structures

Nineteen crystal structures of VRC01-class antibodies have been deposited in the Protein Data Bank (PDB)[68] and comprise the cases used for these predictions. Within this set there are eleven distinct antibody-antigen combinations. Two structures contain the unbound antibody and the remaining seventeen are in complex with envelope glycoprotein gp120 or gp160.

Numbering and identification of CDR loops

All structures were renumbered with the Chothia numbering scheme[91] using the Abnum antibody numbering server[162].

Antibody CDR loop H3 was identified, where possible, using the abYsis webserver's Key Annotation tool[163]. Manual identification was used for structures that the webserver could not annotate. Both methods defined the H3 loop as beginning after the residue sequence CAR and ending before the residue sequence WGXG where X is any residue[162]. The resulting loops ranged from ten to fifteen residues in length.

Loop Prediction

Loop prediction was carried out using the Protein Local Optimization Program (PLOP). Previous chapters of this thesis and past publications[36-37, 164-165] describe Plop's features in great detail, and the protocol used here is presented in depth in chapters three and four.

As the goal of this study was to restore the loops as crystallized in their native structure, all non-loop surrounding residues were treated as fixed atoms and were not sampled in loop buildup or refinement. Fix stages, where loop termini are frozen in sections and sub-loops are built up to focus sampling, were used in all loops containing six or more residues; the number of Fix stages was set so that the shortest loop predicted was four residues. The loop input file, or "looplist" format and the wrapper script that controls Fix stages was modified to read insertion codes as unique residues and iterate over, for example, residues 100, 100A, and 100B separately in setting up the Fix stage predictions.

Protonation states were assigned and hydrogens were added to the input structures using the Prepwizard and PROPKA tools through the Schrodinger Suite[123], based on the crystallization pH reported in the PDB. Two cases were found to be particularly sensitive to protonation state assignment and will be discussed later in this chapter.

Plop Parameters

The following key parameters were used in these predictions:

- Energy function: VSGB2.0 (See Li, et al.[30]).
- Stages: Init, Ref1, Fix1 … Fix8, Ref2.
- Buildup library: the single peptide dihedral library was used (keyword 'segment no').

- Rotamer frequency score: applied with default scaling constant in all stages except Ref2, where the scaling constant was doubled for final refinement. (See Chapter 3 of this thesis.)
- Side chain optimization: only loop side chains were sampled.

7.3 Results

CDR H3 was re-predicted in the crystal structure surroundings for all nineteen crystallized VRC01 class antibody systems. In each prediction, the native coordinates for the loop of interest were deleted from the structure and the loop was built up fully *ab initio* using Plop. The ultimate predicted loop structure for each case is the final lowest-energy loop structure generated by the Plop sampling algorithms, and scored using the VSGB2.0 all-atom physics-based force field, as described in Chapter 1 of this thesis.

The average root mean square deviation (RMSD) between the predicted H3 loop and the native structure, calculated over loop backbone atoms, was 0.98 Å. Eleven of the nineteen cases were predicted at an accuracy of < 1.00 Å, and seventeen out of nineteen were under 1.5 Å. There was no significant difference in results based on loop length – most of the 14-residue loops were predicted to the same level of accuracy as the 10-12 residue loops.

PDB ID	H3 Length	Heavy chain	Light chain	Plop H3 RMSD	Plop H3 dE
4GW4	10	3BNC60 with P61A	3BNC60	0.51	-17.14
4JPV	10	3BNC117	3BNC117	0.33	-4.05
4LSV	12	3BNC1171	3BNC1171	0.85	-84.82
3NGB	12	VRC01	VRC01	1.09	-21.18
4LSS	12	VRC01	VRC01 with N72T	0.93	-23.83
4LST	12	VRC01	VRC01	0.37	-9.71
4JPI	12	VRC01 germline	VRC01 germline	0.82	-14.69
4JPK	12	VRC01 germline	VRC01 germline	0.71	-20.07
4J6R	12	VRC23	VRC23	0.36	-12.05
4LSP	13	VRC-CH31	VRC-CH31	1.76	-35.50
4LSQ	13	VRC-CH31	VRC-CH31	1.21	-25.89
4LSR	13	VRC-CH31	VRC-CH31	1.06	-130.16
4JPW	13	12A21	12A21	0.87	-165.27
4LSU	13	VRC-PG20	VRC-PG20	0.96	-103.85
3SE9	14	VRC-PG04	VRC-PG04	1.37	-23.57
4I3S	14	VRC-PG04	VRC-PG04	1.08	-14.92
4I3R	14	VRC-PG04	VRC-PG04	1.00	-8.05
3SE8	14	VRC03	VRC03	0.80	-15.12
4JB9	15	VRC06	VRC06	2.58	-33.62

Table 7.1 Results of VRC01 H3 loop prediction

Four H3 cases required slight modifications to the default protocol:

Additional sampling:

Under the default parameters, the cases with PDBIDs 4LST and 4LSQ resulted in sampling errors – that is, the lowest energy predicted loop was higher in energy than the native loop. This indicates that the native conformation was never sampled. These cases were rerun with two changes to the parameters in the initial "Init" Plop stage: (1) the number of side chain optimization iterations was doubled to a total of 6 iterations per loop prediction, and (2) the number of loop clusters in buildup for this stage was doubled. These parameters ensure that a large number of loops are generated early on. While it is more time-intensive to run more side chain optimizations on a larger number of loops (from the additional clusters), it succeeded in identifying candidate loops that, when refined in

subsequent stages, were appropriately native-like and slightly lower in energy than the native structure.

Glycan molecules

While many of the VRC01 complexes are heavily glycosylated, H3 loop residues in PDBs 4JPI and 3SE9 interact directly with glycans in the crystal structure. We found that the native H3 loop conformation was sensitive to the presence of a glycerin molecule in 4JPI and an n-acetyl d-glucosamide molecule in 3SE9. These molecules were included in the 4JPI and 3SE9 predictions (with Plop parameter 'load het yes') reported here.

7.4 Discussion

We demonstrated Plop's ability to successfully restore the native H3 loop in all crystallized VRC01-class antibodies, which include antibodies alone and in complex with gp120. These results mean that we can confidently move forward to predict H3 loops with mutations or other modifications that do not have crystal data but may be important for optimization efforts. Additionally, these results give us a baseline accuracy level for predicting modified loops. Overall, we can predict the H3 loop in these cases to an accuracy of about 1 Å in the crystal structure context.

In re-predicting these loops, we also developed insight about various aspects of these systems. Here, we will discuss our findings in two key areas: protonation state assignment and glycosylation.

Protonation

We discovered that protonation of a histidine residue near the H3 loop is critical to accurately re-predict the H3 loop in antibody VRC20.

This structure, PDB ID 4LSU, was not yet deposited when we used it as a blinded loop prediction test. In preparing the system, we found that histidine residue H:35 was near the empty loop pocket and, as such, its protonation state may be linked to the native loop conformation. Additionally, the final loop residue is also a histidine. We sampled all possible protonation states for both histidines: protonating the delta position only (HID), protonating the epsilon position only (HIE), or protonated both positions and forming a charged residue (HIP).

Based on our predictions with each protonation state, we determined that in order to restore the native loop conformation, H:35 must be doubly protonated (charged) to form a salt bridge with the loop, and H:102 should be protonated in the epsilon position to avoid forming a competing salt bridge. At first, we were surprised that H:35 could be doubly protonated because it would appear to clash with a neighboring tryptophan residue (H:47). However, we established from the crystallographers that the TRP residue is accurately placed and the density is good in this region.

Looking closer at the native loop, we realized that this charged histidine forms a strong salt bridge with a glutamic acid residue on the loop, residue H:100B, that likely overwhelms the slight H-H clash with the TRP residue. We predict this interaction when H:35 is set to doubly protonated and H:102 is set to be epsilon protonated. Indeed, this prediction restores the nativelike loop altogether, with heavy atom RMSD to native of 0.96 Å. This network of interactions is shown in Figure 7.2.

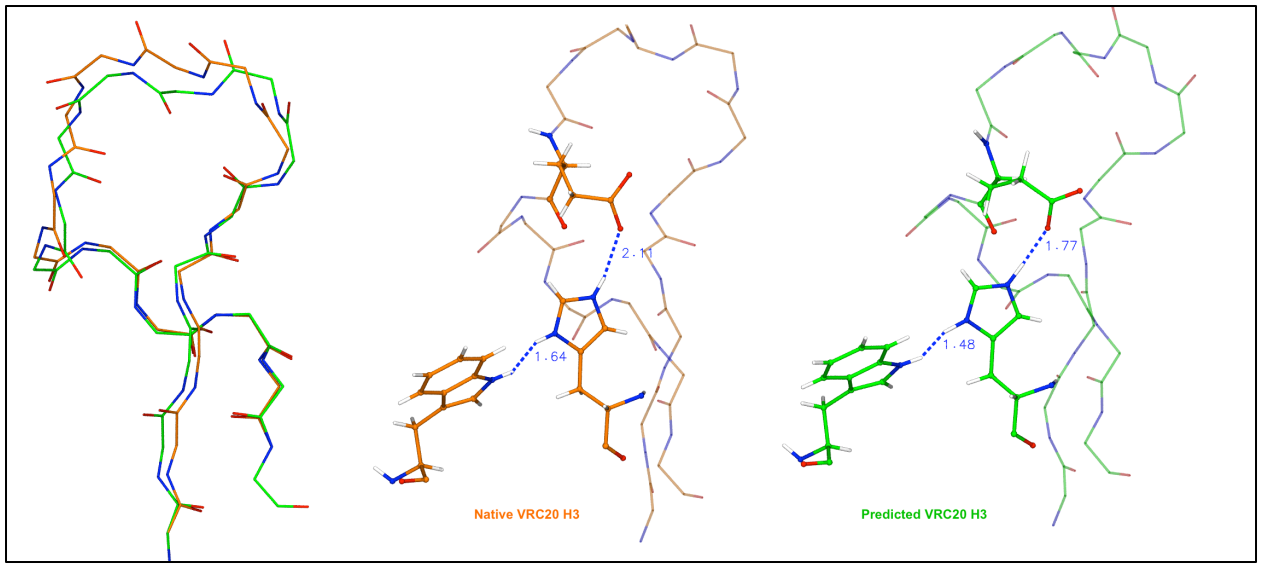

Figure 7.1 VRC20 prediction compared to native

Backbone trace showing accurate reprediction (left); Native salt bridge between H:35 and H:100B (center); Predicted salt bridge with H:35 HIP (and H:102 HIE) (right).

While some of the crystal structure refinement data suggests that the H:102 loop residue was doubly protonated, we found that HIE protonation here was necessary to restore the native loop. When H:102 is doubly protonated, we predicted a loop that forms a salt bridge between H:102 and H:100D ASP. This salt bridge is not found in the native loop, and the predicted loop's conformation is highly distorted compared to the native loop (RMSD > 5 A). Alternatively, this alternate conformation is a low-energy well, and more sampling could identify a nativelike loop making the H:35-H:100B salt bridge with even lower energy. However, there is also no compelling interaction in the native structure that would motivate a charged H:102.

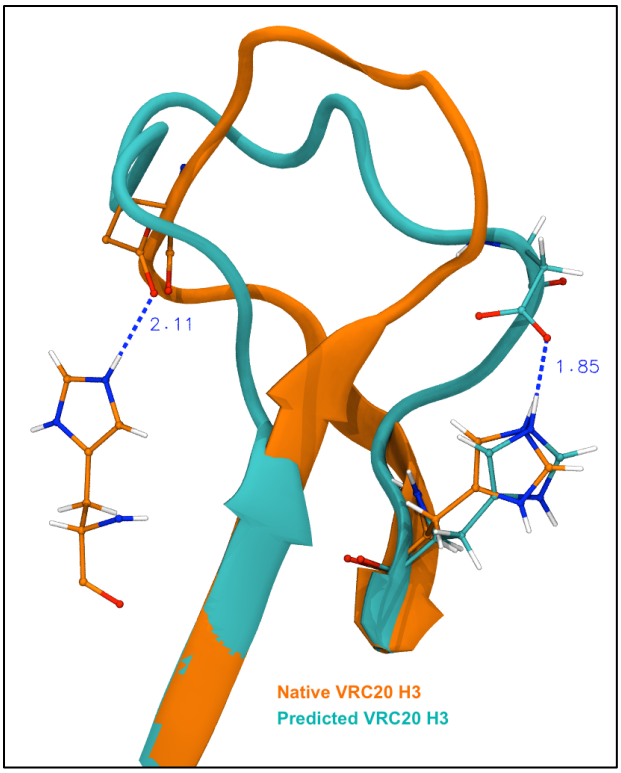

Figure 7.2 The 5.5 Å prediction with H:102 doubly protonated (blue) compared to the native (orange).

This antibody experiences a high degree of interplay between protonation states and H3 loop conformation. When this may be the case, due to titratable residues on or near the loop of interest, it's clearly necessary to test loop prediction under different protonation state assignments, and carefully examine the resulting structures to help identify the conformation with the most favorable interactions.

Antibody CH31, PDB ID 4LSR, has the same neighboring and loop-terminal histidines, and as such, we re-predicted the H3 loop with each possible protonation state assigned. Results of this prediction as well as the VRC20 predictions are shown in Table 7.2. Here, we did not find any predictions for a salt bridge with H:35, which matches the native behavior. The interactions for both this and the

histidine at H:102 are consistent with the native for all predictions with H:35 protonated in the epsilon position or both positions.

Protonation		VRC20 (4LSU)		CH31 (4LSR)	
H:35	H:102	H3 RMSD	Final E	H3 RMSD	Final E
HID	HID	6.8	7175.62	6.2	5449.14
HIE	HIE	4.7	7180.36	1.2	5427.68
HIP	HIE	1.0	7191.54	1.1	5457.34
HIP	HIP	5.5	8185.79	1.0	6825.24

Table 7.2 VRC20 and CH31 H3 loop predictions for each histidine protonation state.

Glycosylation

As mentioned in the methods section, we found that neighboring glycans are crucial to restoring the native loop in two cases: the uncomplexed VRC01 putative germline antibody (PDB ID 4JPI) and antibody VRC03 (PDB ID 3SE9). For 4JPI, including glycans in loop prediction substantially improved the H3 RMSD to native from 2.49 Å to 0.82 Å. Including glycans in 3SE9 improved the result by 0.75 Å, going from 2.31 Å RMSD to native to 1.56 Å. This data is also presented in Table 7.3.

In 4JPI, a glycerin molecule near the H3 loop forms a hydrogen bond with the sidechain of asparagine residue H:97, depicted in Figure 7.3 (a). When this molecule is included in its crystal structure position, loop prediction identifies a nativelike (<1 Å) lowest-energy loop, shown in green in Figure 7.4 (a), that also restores this hydrogen bond. Without glycerin, the loop diverges significantly from the native conformation, shown in orange in Figure 7.4 (a). It appears that native loop conformation is highly dependent on the glycerin as positioned in the crystal, and this must be a consideration in any further computational experiments involving this protein and loop region.

The 3SE9 H3 loop interacts with an n-acetyl d-glucosamide (PDB ligand code NAG) molecule, forming a hydrogen bond between the carbonyl on glycine H:100A and NAG, residue G:776. Loop predictions with and without including glycans show that NAG primarily improves our result by blocking a low-energy but extremely non-native loop (>6 Å RMSD to native), shown in orange in Figure 7.3 (b). This loop clearly clashes with NAG in the crystal structure and, unsurprisingly, is not built as a candidate loop when NAG is included.

But we would expect that including the glycan would not only block sterically impossible conformations but also form a hydrogen bond with the predicted loop. Yet our final predicted loop in the presence of NAG does not form the native hydrogen bond at residue 100A. This loop is also on the upper end of what we consider an acceptable RMSD to native. We hypothesize that additional sampling would eventually identify and refine a more nativelike loop with the NAG interaction and a lower RMSD.

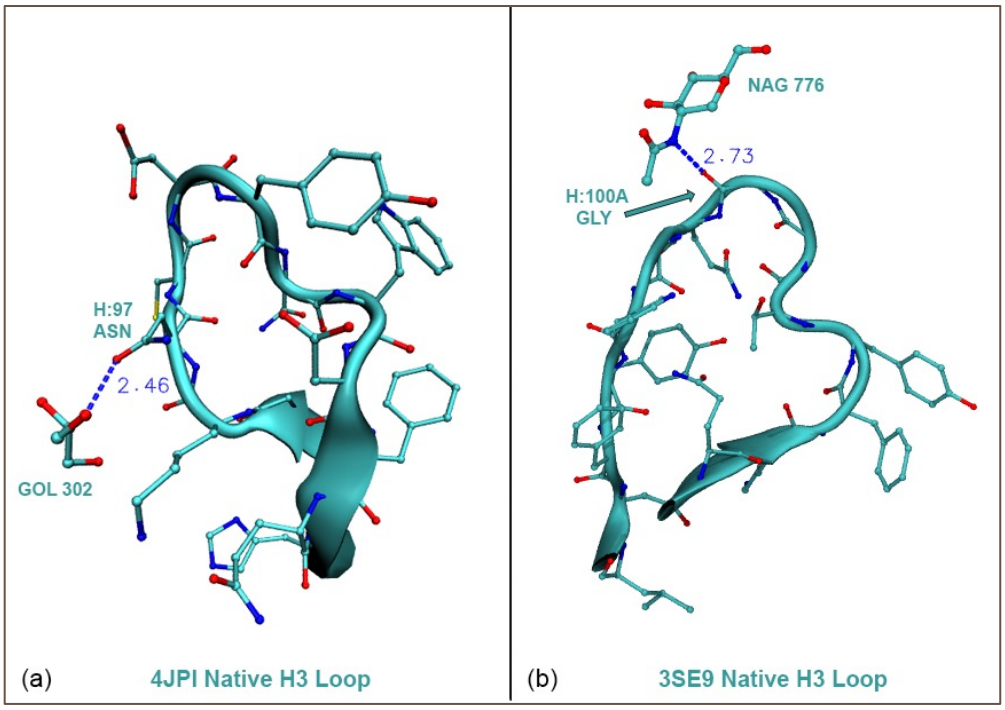

Figure 7.3 Glycan interaction with native H3 loop in (a) PDB 4JPI and (b) PDB 3SE9

	4JPI		3SE9	
	H3 RMSD	dE	H3 RMSD	dE
Without Glycans	2.49	-13.89	2.31	-20.37
With Glycans	0.82	-14.69	1.56	-23.82

Table 7.3 Loop prediction results with and without glycans for PDB 4JPI and PDB 3SE9

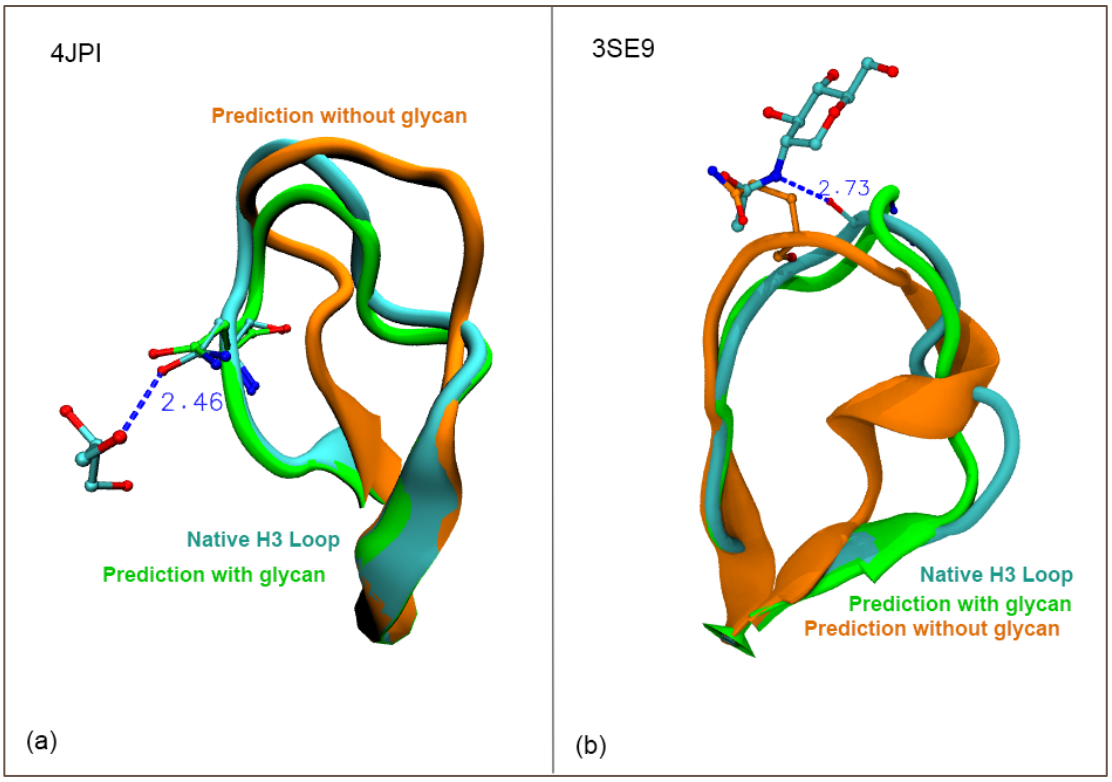

Figure 7.4 Loops predicted with and without glycan molecules for (a) PDB 4JPI and (b) PDB 3SE9

These loops are highly sensitive to the presence of glycans nearby. At a minimum, further computational experiments relying on these structures and considering interactions or perturbations on or near the H3 loops should carefully include these molecules.

Chapter 8 : Structure prediction to inform computational binding affinity predictions in a VRC01 antibody system

8.1 Introduction

The overall goal of this study is to develop a robust computational protocol for predicting the change in binding affinity upon mutation to a protein-protein antibody-antigen system. The first step, which we will focus on here, is developing our methods by comparing to preexisting experimental data over a set of simple perturbations to a VRC01-class antibody.

As in the previous chapter, we are using the VRC01 antibody as the test case. VRC01 is a broadly neutralizing, highly potent antibody against HIV-1 that has been found to target the CD4 binding site on gp120[134]. Structural studies of VRC01 in complex with gp120 indicate that VRC01 mimics some elements of CD4-gp120 binding but differs from CD4 in key ways, allowing it to exceed the neutralization potency of other CD4-binding-site (CD4bs) antibodies. Briefly, the VRC01-gp120 interaction surface is approximately 2500 A^2 in area, with roughly equivalent contributions from both proteins. As with CD4, which has a CDR2-like region centrally involved in gp120 binding, the largest part of VRC01's interaction surface comes from CDR H2[134]. But comparing both CD4 and VRC01 structures each bound to gp120, it is evident that the domains have different orientations. Zhou et al. showed that VRC01 in complex with gp120 rotates 43 degrees and translates 6 A away from the corresponding CD4 position relative to gp120[134]. VRC01 also makes contacts between its light chain and a glycan molecule; in CD4, this glycan shields a contact site[134]. Overall, VRC01 distinguished by both its similarities and differences to CD4 binding, and the depth of structural understanding of its binding surface and activity make it a promising candidate for computational studies with the ultimate goal of improving its binding affinity for gp120.

To computationally determine the change in binding affinity upon perturbing or mutating parts of gp120, we use free energy perturbation (FEP) simulations. As mentioned in the previous chapter, to measure the relative binding affinity, or free energies of binding, between two protein-protein systems, FEP simulations take an alchemical path from the wildtype protein to the mutated protein in both the complex form and the uncomplexed form. The intermediate changes in free energy (dG) are calculated between (1) the complexed wildtype and mutated type and (2) the uncomplexed wildtype and mutated type, and subtracted to obtain:

$$ddG = dG_{complex} - dG_{solvent}$$

This value can be compared with an experimental ddG value, calculated from:

$$ddG = dG_{mutant\ binding} - dG_{wildtype\ binding}$$

The perturbation we focus on here is individual residue mutations to alanine, or "alanine scanning." This is a useful starting point, both computationally and for understanding the antibody-antigen contacts that contribute to binding affinity. Computationally, the perturbation going from any side chain[iv] to alanine is straightforward and attainable in a standard-length simulation. This data also provides useful insight to the system: when we essentially remove side chains, the resulting change in binding affinity tells us how important the native interaction is to binding.

In this chapter, I discuss two experiments: first, preliminary FEP change in binding affinity data for mutating a set of contact residues on VRC01 to alanine, as compared to the experimental alanine scanning data; and second, case studies in using structure prediction as input or a guide for these FEP simulations.

[iv] Excluding charged residues, which must be treated differently and are not investigated here.

8.2 Alanine scanning experiment

The experimental data we are working to replicate come from an alanine scanning experiment conducted by collaborators at the NIH vaccine research center (VRC). The antibodies used in this study, VRC01, VRC03, and VRC-PG04, target the CD4 binding side on the envelope glycoprotein gp120 to potently neutralize up to 90% of HIV-1 strains. In order to evaluate the role of individual contact residues in binding to the gp120 antigen, the VRC researchers individually mutated each contact residue to alanine, expressed each mutated antibody, measured its affinity for HIV-1 gp120 and determined the change in Gibbs free energy (dG). dG was calculated using the formula

$$dG = RT \ln (K_d)$$

where R is the gas constant, T is the temperature, and K_d is the dissociation constant equal to K_{off} / K_{on}.

This was calculated for each unmutated wild type antibody and the mutated antibody and used to calculate $ddG = dG_{mutant} - dG_{wildtype}$. Therefore, a positive ddG suggests that the alanine mutation makes binding less favorable, typically suggesting that the wildtype residue is important in the binding reaction or structural features involved in binding. A negative ddG indicates that mutating to alanine – essentially, removing the side chain atoms in all cases but glycine – actually improves the binding. This overall ddG value is what we can compare against in benchmarking the accuracy of the FEP binding affinity calculations. Results of this experiment are listed in Table 8.1 in section 8.5.

8.3 FEP-REST simulations

FEP simulations are a method for calculating the change in protein-ligand binding free energy, initially in systems with small molecule ligands and moving into systems with protein ligands (as in this work). FEP carried out through molecular dynamics (MD) simulations, here simply referred to

as "FEP simulations," is one of several approaches to computationally determine binding free energy. It stands out for its ability to completely characterize the binding energetics and thermodynamics of the system, which should result in the most accurate binding energy prediction. However, the attainable accuracy hinges on the accuracy of the underlying force field and the degree to which the phase space of the system is sampled.

To address the phase space sampling factor, Wang, et al. developed an enhanced sampling approach: Free Energy Perturbation / Replica Exchange with Solute Tempering or FEP/REST[147]. This implementation of REST increases the sampling power of the FEP simulation, so that more varied binding site conformations can be explored. The FEP/REST methodology has been detailed in the literature[147] and applied to protein-small molecule ligand systems[146], generating reliably good results for systems of pharmaceutical interest[144]. Here, we will briefly outline how FEP/REST improves sampling and therefore the accuracy of binding free energy calculations, as described in detail by Wang, et al.[147]

Alone, FEP simulations calculate the free energy difference between two systems – for our purposes here, a wildtype antibody and a mutated antibody – by alchemically transforming the wildtype system into the mutated system. This transformation is executed in a series of discrete steps, or λ windows which range from $\lambda=0$ in the initial, wildtype state to $\lambda=1$ in the final mutated state. In FEP/REST, this general protocol is preserved, and the sampling power is cranked up in the intermediate λ windows by scaling the potential energy in a localized region by a factor < 1, so that the energy barriers between conformations states are lowered in this region. This is done by increasing the effective temperature (by scaling the Hamiltonian) incrementally in the first half of λ

windows, then reducing back down to the initial effective temperature over the second half of λ windows in this "lambda hopping" workflow[147].

Lowering energy barriers allows the system to more freely explore phase space, sampling more conformations than would be possible with a "true" potential energy and avoiding "trapping" the system in the initial conformation. These different conformations are then exchanged between neighboring λ windows and propagated to the final mutated state through the Hamiltonian replica exchange method[147].

The localized region targeted for additional sampling is called the "hot" region, and in protein-protein systems, typically consists of the residue being mutated. Selection of the "hot" region is outlined by Wang, et al.[147] and requires balancing the degree of sampling (by increasing the hot region size) with the precision of free energy results (which degrades as the difference between subsequent λ windows increases).

Our FEP/REST simulations were carried out using Desmond[166], as implemented to run on graphics processing units (GPUs). We used the OPLS2.1 force field[27, 29], solvated the system in a water box of buffer 5.0 A around the molecule, and ran the simulation for a total of 10 ns in the production (lambda hopping) stage.

8.4 Modeling gp120

We chose to initially focus on antibody VRC01 featured in the alanine scanning experiment. However, while a crystal structure has been determined for VRC01 in complex with gp120 (PDB

ID: 3NGB[134]), the gp120 used in the alanine scanning experiment differed from the gp120 reported in 3NGB.

The alanine scanning experiment studied VRC01 in complex with a gp120 resurfaced stabilized core 3 (RSC3) probe. This molecule was designed by Wu, et al. to identify CD4-binding site-specific neutralizing antibodies, and it maintains the core structure of the CD4 binding site's neutralizing surface in the absence of other antigenic regions of HIV-1[148].

We used the complexed gp120 in PDB 3NGB as the template for constructing a homology model with the RSC3 probe's sequence, and used this model as the "ligand" with the VRC01 antibody in the FEP simulations discussed here. The model was built in BioLuminate using the standard homology model-building tool[167]. The gp120 "G" chain of PDB 3NGB was the only structure template provided. As Figure 8.1 shows, regions closest to the CD4 binding site were able to be modeled with residue side chain mutations; most loops with substantial reconstruction were on the opposite side of the protein. The model was then subject to five rounds of total side chain re-optimization, using Plop, while in complex with the VRC01 antibody. Finally, protonation states were optimized using the Protein Preparation Wizard and ProPKA tools in the Schrodinger Suite[123].

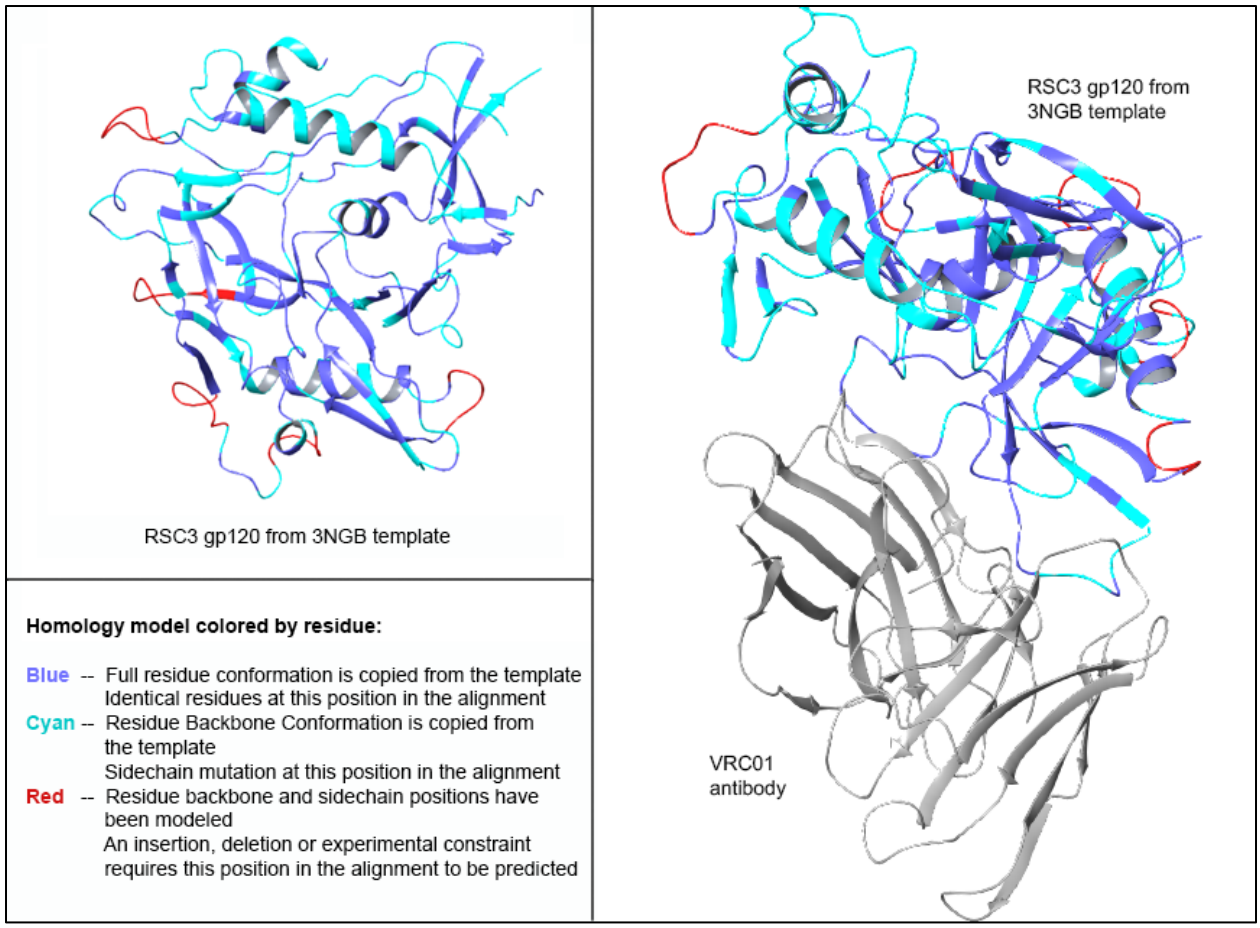

Figure 8.1 RSC3 gp120 homology model from 3NGB gp120 template

8.5 Change in binding affinity predictions

In our initial FEP simulations on this system, we obtained results in good agreement with the experimental data, as shown in Table 8.1. For the twenty highest-magnitude ddG mutations, excluding charged residues, our simulations predicted a binding affinity within 1.0 kcal of the experimental value for 70% of the cases. All but one mutation obtained a ddG within 1.5 kcal of the experiment. This is also depicted in Figure 8.2, which shows the correlation between experimental ddG and FEP-predicted ddG.

The complete alanine scanning data set shows an overall experimental noise of 0.45 kcal/mol for each measurement, and the cumulative error when subtracting the two measurements to obtain ddG gives an experimental accuracy of 0.9 kcal/mol. Nearly 70% of our simulations are therefore accurate to within the experimental limitations, and this noise explains much of the overall deviation from center in Figure 8.2.

Residue	Wildtype	Mutation	Experimental ddG (kCal/mol)	Simulation ddG (kcal/mol)	Absolute Error
H:30	ILE	ALA	-0.01	1.09	1.10
H:33	THR	ALA	0.70	0.29	0.41
H:47	TRP	ALA	1.19	4.43	3.24
H:50	TRP	ALA	1.28	0.41	0.87
H:54	GLY	ALA	-1.28	-1.31	0.03
H:55	GLY	ALA	1.26	2.34	1.08
H:57	VAL	ALA	1.35	1.21	0.14
H:58	ASN	ALA	1.76	0.71	1.05
H:59	TYR	ALA	0.54	0.59	0.05
H:64	GLN	ALA	-0.35	0.35	0.70
H:69	MET	ALA	0.52	0.31	0.21
H:73	VAL	ALA	1.08	0.41	0.67
H:100	TYR	ALA	1.42	-0.04	1.46
H:100A	ASN	ALA	1.11	0.42	0.69
H:100B	TRP	ALA	4.45	4.47	0.02
L:27	GLN	ALA	-0.78	0.69	1.47
L:28	TYR	ALA	0.97	0.22	0.75
L:30	SER	ALA	-0.15	-0.26	0.11
L:91	TYR	ALA	2.01	1.41	0.60
L:97	PHE	ALA	0.85	0.40	0.45
			Mean Absolute Error		**0.75**
			Root Mean Square Error		**1.04**

Table 8.1 FEP Alanine Scanning Results

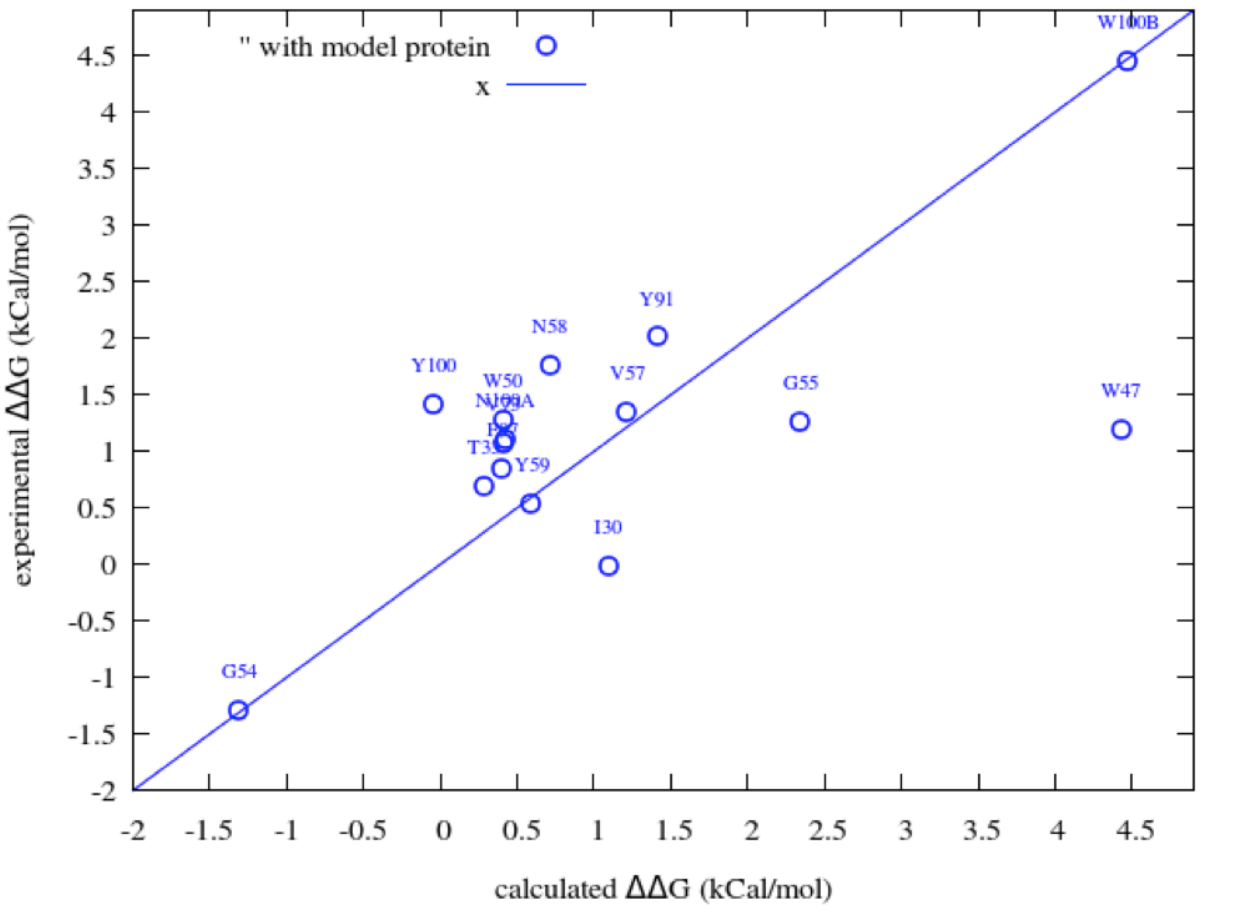

Figure 8.2 FEP Alanine Scanning Deviation from Experimental Results

However, there is clearly one outlier beyond the reasonable error range. The tryptophan to alanine mutation at residue H:47, with a simulated ddG of 4.43, is well above the experimental ddG of 1.19 and as such overestimates the penalty for removing the tryptophan side chain. Figure 8.3 shows the location of this residue near the H:L interface and adjacent to a gp120 loop.

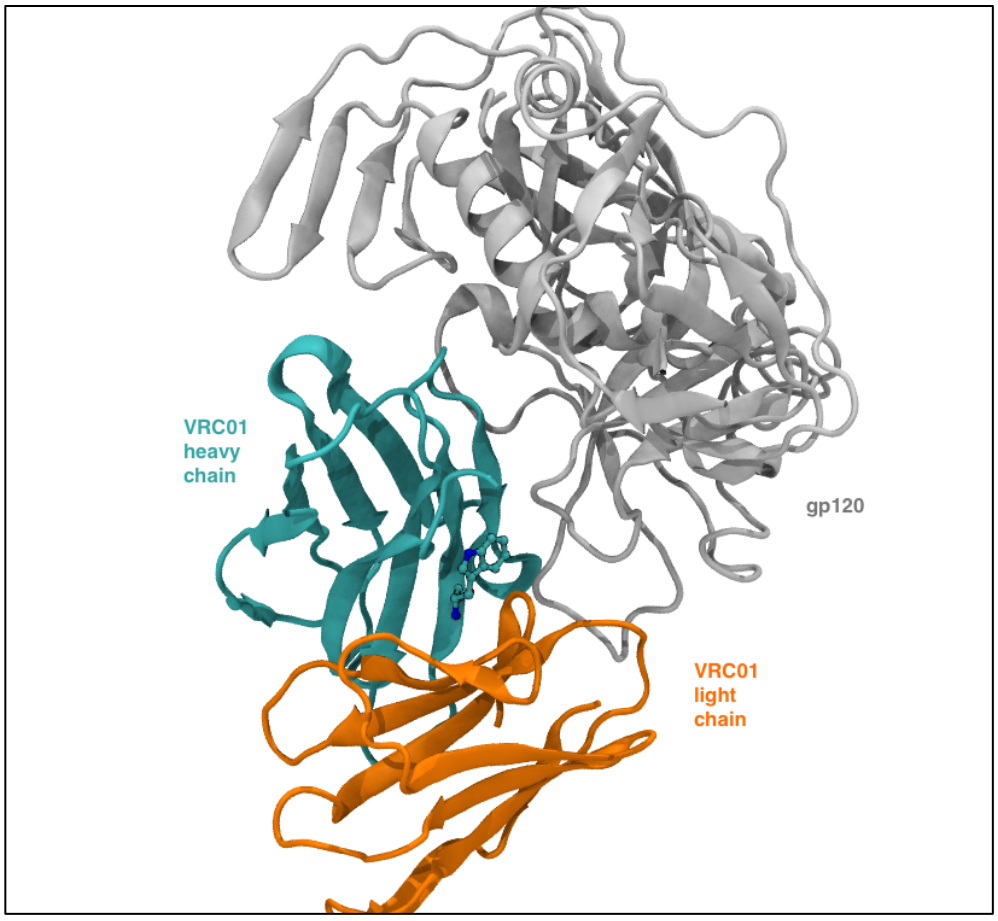

Figure 8.3 Position of tryptophan residue H:47 in the VRC01 antibody-gp120 complex

Experimenting with slight modifications to the input parameters continued to over-predict this mutation by 1-3 kcal. We are currently investigating the source of this error. One possible factor is the position of a nearby loop on gp120.

This could be affecting the simulation in multiple ways:

- Homology model loop: This loop was conserved on the homology model constructed for the RSC3-gp120. However, it is possible that in the context of the RSC3 sequence modifications, this loop would have a different conformation that could affect the H:47 interaction.

- Glycosylation: In the VRC01-gp120 complex crystal structure, 3NGB, residues G:275-6 on this loop form contacts with a nearby n-acetyl glucosamine molecule, shown in Figure 8.4. The RSC3 probe used in the alanine scanning experiment was heavily glycosylated but the positioning of the glycans is not confirmed. If the crystal structure loop conformation is dependent on a glycan that was not present in the experimental complex, our simulation may not capture the actual interactions with the current input structures.
- Loop flexibility: It is also possible that the gp120 loop is more flexible than we have so far captured in the REST-stage sampling, particularly in the absence of the large tryptophan side chain.

We are actively investigating these possible complications and their effect on the simulation results.

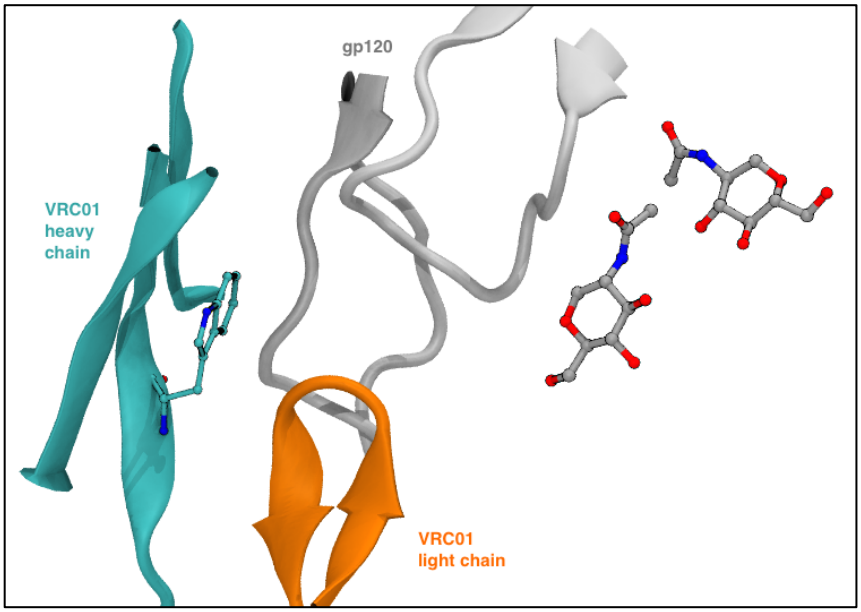

Figure 8.4 Glycosylation on the gp120 loop near the H:47 mutation in crystal structure 3NGB

8.6 Endpoint modeling

Ideally, the set of test cases for these simulations would include several mutated antibody crystal structures, both alone and in complex with the antigen. If these structures were available, we could conclusively determine whether the simulation fully sampled the actual mutated antibody conformation. In the absence of crystal structures, though, we can use loop prediction to suggest possible mutated conformation(s) and compare simulation frames to the predicted mutant structures. The work discussed here provides a starting point for further study, especially as we move on to attempt more complicated mutations than mutating to alanine.

In the case of VRC01 (PDB 3NGB), our primary test case so far, we used structure prediction to understand the result of our reverse mutation experiment on residue H:100B. In the alanine scanning experiment and simulation, this residue was mutated from tryptophan (W) to alanine (A). The reverse mutation mutates this residue back to W from A.

First, we used the output structure from the forward (W->A) simulation as input to the reverse mutation. This approach has some similarities to the cycle closure for mutation paths in small molecule ligand mutation simulations[146]. The reverse mutation should result in a ddG of similar magnitude to the forward mutation (and experimental alanine scanning result) but with the reverse sign. The hysteresis, or sum of the forward and reverse mutations, should be zero in a perfectly converged simulation using a perfect force field[146]. In practice, of course, this is not the case. Systematic force field errors compared to the true energy of the system, combined with sampling errors from incompletely examining phase space contribute to a nonzero hysteresis in the best case scenario; typically, hysteresis around 1.0 kcal is considered acceptable[146] and the magnitude of the

hysteresis gives us a hint as to the possible convergence or energy errors in simulations, especially as we extend these methods to protein-protein systems.

The ddG for the reverse mutation starting from the final forward mutation frame was calculated to be 3.82 kcal, compared the target of -4.45 for the A->W mutation (from the W->A experiment) as shown in Table 8.2. Clearly, error(s) came into play in this simulation, and we employed loop prediction to identify the source of the error – the critical first step in improving first, this case and then, the overall methodology.

	Starting Structure	Res-idue	Mutation	Complex dG	Solvent dG	Simulation ddG	Experimental ddG	Absolute Error
1	Native VRC01 (PDB 3NGB)	H:100B	W -> A	5.31	2.03	3.28	4.45	1.17
2	Final structure from forward mutation	H:100B	A -> W	5.87	2.05	3.82	-4.45	8.27
3	Native VRC01 with W->A mutation	H:100B	A -> W	-5.09	-1.26	-3.83	-4.45	0.62
4	Predicted H3 in native with W->A mutation	H:100B	A -> W	3.08	1.00	2.08	-4.45	6.53

Table 8.2 Change in binding affinity from FEP simulation of H:100B mutation in VRC01.

Data presented here: 1. A successful forward mutation simulation - note that this simulation was run with the gp120 crystallized in 3NGB, not the RSC3 model, and is provided here for reference. 2. Result using output from (1) as input for the reverse mutation back to tryptophan. 3. Result starting from native structure with tryptophan sidechains deleted ("Alanine mutation") and mutated back to tryptophan over course of FEP simulation. 4. Result from native structure, alanine mutation at H:100B, and full H3 loop re-prediction with alanine at H:100B.

This mutation is part of the CDR H3 loop and has a contact with the gp120 antigen. Looking at the H3 loop and surrounding region in the crystal and mutated W100BA output structure, it's evident that the loop and region undergo conformational shifts over the course of the simulation, including the domain shifting shown in Figure 8.5. First, we wanted to determine whether this was a legitimate loop conformation for the mutant structure, or if the simulation was affected by a degree of bias towards a disordered state. To test this, we re-predicted the H3 loop in the crystal structure context,

but with 100B mutated to alanine. If this loop predicted recovered a nativelike loop, we would investigate how the force field allowed the observed, possibly disordered structure to be propagated in the simulation. Alternatively, if loop prediction found a nonnative loop lowest in energy, we would investigate whether this conformation is observed in the FEP simulation, and look at the FEP sampling in more detail.

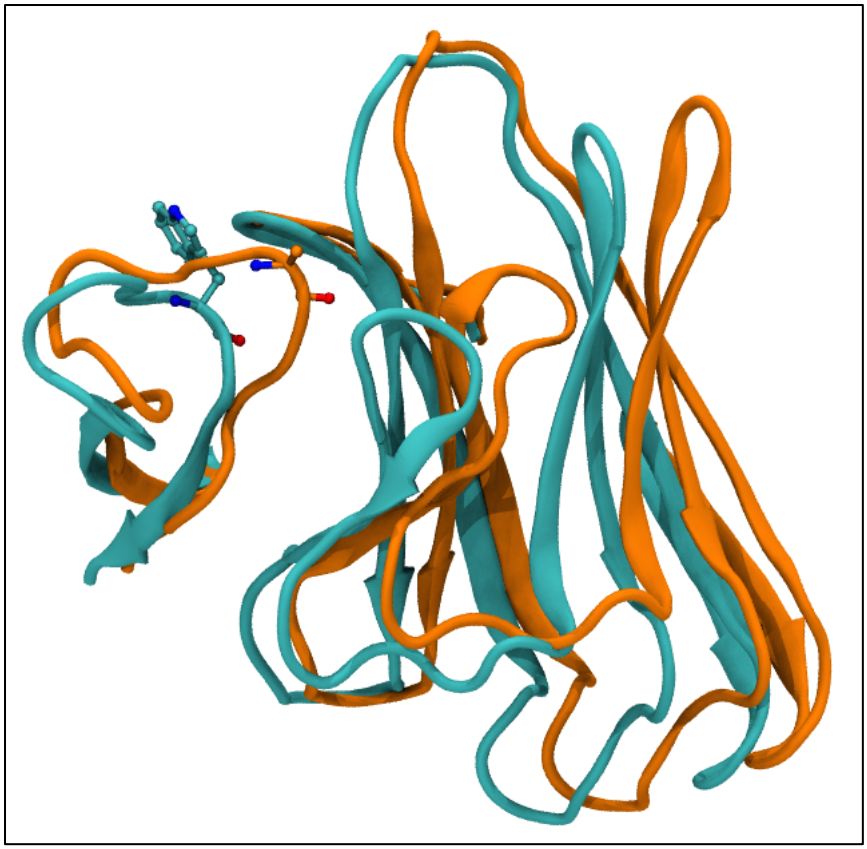

Figure 8.5 Comparison of VRC01 FEP simulation starting and ending structures.

The input, shown in cyan, is the native structure and the output, shown in orange, is after mutating residue H:100B to alanine. The H3 loop is shown on the left and the light chain is on the right. The light chain's orientation relative to the heavy chain shifts substantially over the course of the simulation phases.

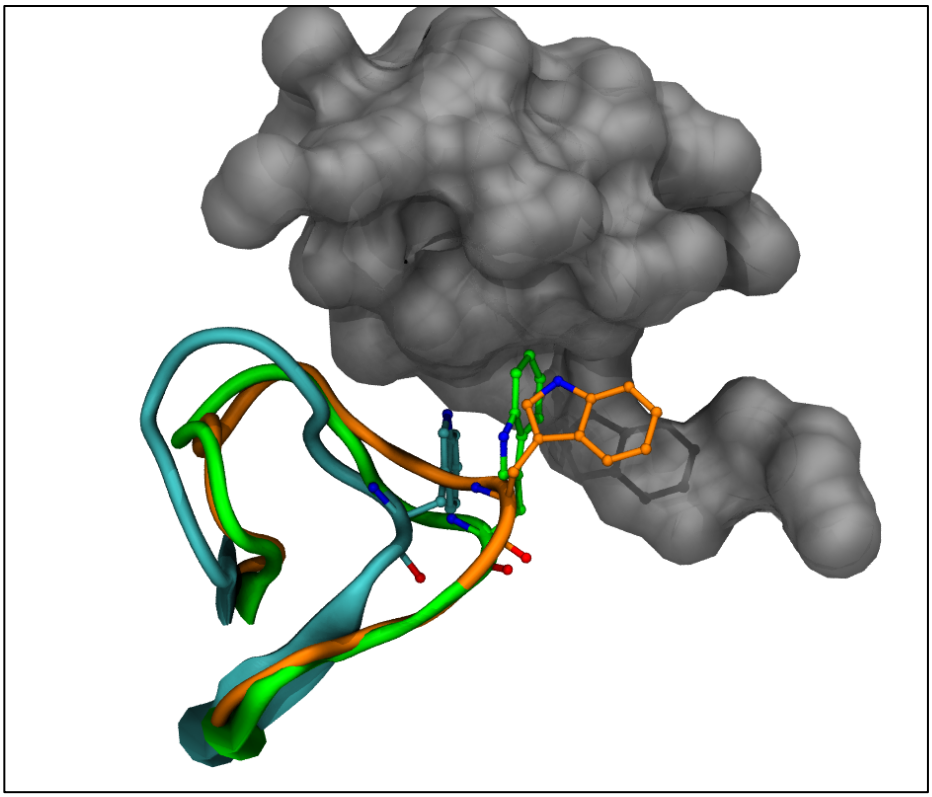

Figure 8.6 VRC01 H3 loops observed in forward and reverse FEP simulations.

The native H3 loop is shown in cyan and the gp120 residues within 12 Å from the H:100B mutation are shown in gray surface representation. The H3 loop after mutating to alanine is shown in orange. The alanine side chain is replaced here with the native tryptophan for visibility. The green loop is the H3 loop from the FEP simulation mutating H:100B back to tryptophan, with the forward mutation output (the orange structure) as the starting structure.

We found that the crystal structure mutated loop prediction identified a lowest-energy H3 loop with conformation in between that of the native and the simulation output structure. RMSDs are shown in Table 8.3. In the absence of a crystal structure with this mutation, this result tells us that the starting conformation for the reverse mutation is reasonable. Further, we can conclude that the primary error source is incomplete sampling in the reverse simulation, not force field errors. The challenge is getting over the energy barrier between this state and a nativelike conformation for the "mutant" 100BW.

Structure	H3 RMSD to Native	dE
Wildtype VRC01 - H:100BW	1.28	-18.21
Mutated VRC01 - H:100BA	2.54	-34.01

Table 8.3 Results of H3 loop prediction in VRC01 with wildtype and mutated residue H:100B.

The mutated (H:100BA) loop RMSD is also calculated in reference to the native, H:100BW-containing H3 loop. The dE reported is the relative energy to the starting structure, defined as dE = (final loop energy) – (minimized starting structure energy).

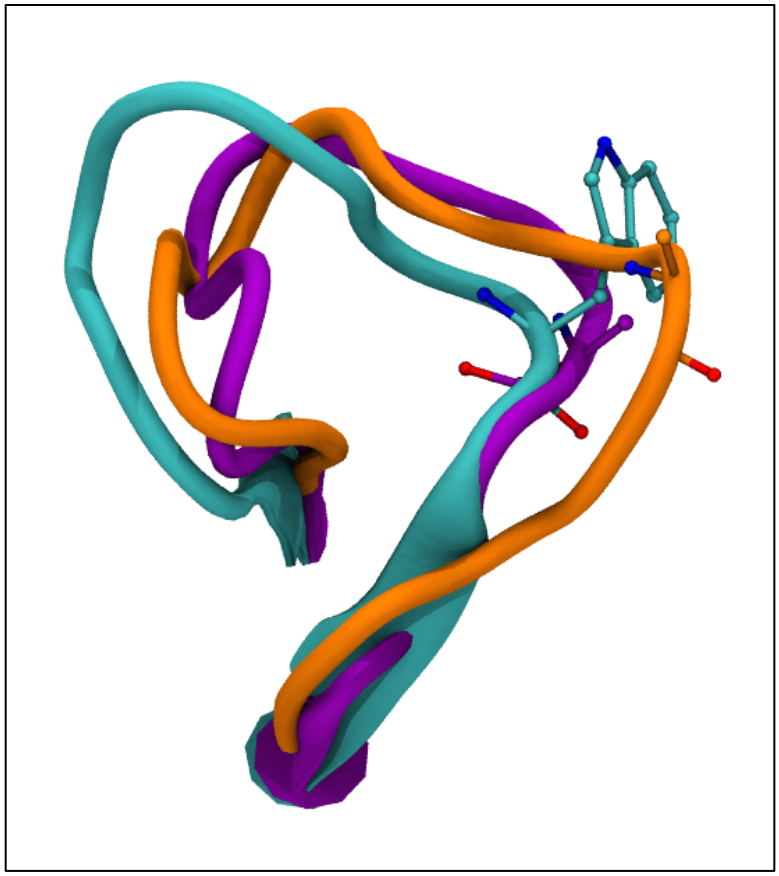

Figure 8.7 Predicted mutated H3 loop compared to native and FEP-mutated loops.

The purple loop was predicted using Plop with H:100B ALA. The orange loop is the final frame of the FEP simulation mutating H:100B to ALA. The cyan loop is the native structure.

Next, we ran reverse mutation simulations with two additional starting configurations: (1) the crystal structure with 100B mutated to alanine (by simply removing the tryptophan side chain above C-

beta) and (2) the predicted H3 loop with 100BA, from the loop prediction in the native context discussed above. The results of these simulations are shown in Table 8.2. Both calculations support the A->W sampling barrier hypothesis. Starting with a native H3 loop and surrounding context, we obtain the expected ddG with low hysteresis compared to the forward mutation. The simulation from the predicted 100BA-containing H3 loop returns the wrong sign for ddG, as with the initial reverse mutation, and is similar in magnitude to that initial reverse mutation. These tests show us that the H3 loop conformation is interdependent with this mutation, and that we are not sufficiently sampling the A->W pathway to overcome the energy barrier between the mutated and wildtype 100B H3 conformations.

Before addressing the sampling specifically, though, we are using this result to consider the best method for gleaning insight from the "reverse" mutations and other attempts to recreate cycle closure in protein-protein FEP simulations. Here the combined protein-"ligand" phase space is much larger and subject to more random fluctuations than in the well-studied small molecule ligand systems.

One approach is to use loop prediction to determine the mutated region's conformation and use that structure as the input to the reverse simulation. In this case, such an approach would not succeed out of the box but does lower the error somewhat. Another, possibly complementary, approach is specific to cases like this one, where the mutation builds up a much larger residue, like tryptophan, than the wild type. In these cases, we could use an intermediate mid-sized mutation. Instead of mutating directly from A->W, we could run two successive simulations A->X->W, where X is a residue larger than alanine but smaller than tryptophan. For this approach to effectively improve the A->W sampling, we would need to carefully avoid intermediate mutations that

substantially alter the loop conformation and introduce new challenging energy barriers, so part of this approach would involve predicting the H3 (or other relevant) loop with multiple possible intermediate mutations and selecting a mutation that supports a relatively nativelike loop conformation.

8.7 Future directions

We have shown successful proof-of-concept results for calculating binding affinity using FEP-REST simulations on a VRC-class antibody in complex with gp120. Further, we have shown the key roles homology modeling and loop structure position can play in constructing an input model, when the exact crystal structure for a set of experimental data is unavailable, and in understanding the source of error in simulations.

Moving forward, there are several active and pending areas of inquiry to bring this research closer to predicting novel mutations to optimize VRC01's binding affinity, in addition to the ongoing challenge cases discussed above.

Currently, we are getting a handle on the precision of these results. By running multiple trials of each mutation with different random seeds and/or slightly different hardware, we are evaluating the effect of averaging multiple runs on overall accuracy. We will also expand the data set for testing our protocols. First, we are simulating mutations besides X->A, as introduced previously. The additional considerations in building up larger resides from smaller wildtype residues may require adjustments to the current protocol, whether large scale, like inserting one or more intermediate structures in the cycle, or parameter modifications like increasing simulation time or expanding the size of the REST region. We are also looking to validate our protocols in other VRC01-class antibodies for which

there is experimental data. By leveraging our structure prediction toolkit to better understand systems and results over the course of these studies, we are confident that protein-protein FEP simulations will provide useful information to guide prospective VRC01 optimization efforts.

Bibliography

1. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Tramontano, A., Critical assessment of methods of protein structure prediction (CASP)--round IX. *Proteins* **2011,** *79 Suppl 1*, 1-5.

2. Kryshtafovych, A.; Fidelis, K.; Moult, J., CASP9 results compared to those of previous CASP experiments. *Proteins* **2011,** *79 Suppl 1*, 196-207.

3. Oshiro, C.; Bradley, E. K.; Eksterowicz, J.; Evensen, E.; Lamb, M. L.; Lanctot, J. K.; Putta, S.; Stanton, R.; Grootenhuis, P. D. J., Performance of 3D-database molecular docking studies into homology models. *Journal of medicinal chemistry* **2004,** *47*, 764-7.

4. Enyedy, I. J.; Lee, S.-L.; Kuo, A. H.; Dickson, R. B.; Lin, C.-Y.; Wang, S., Structure-Based Approach for the Discovery of Bis-benzamidines as Novel Inhibitors of Matriptase. *Journal of Medicinal Chemistry* **2001,** *44*, 1349-1355.

5. Selzer, P. M.; Chen, X.; Chan, V. J.; Cheng, M.; Kenyon, G. L.; Kuntz, I. D.; Sakanari, J. A.; Cohen, F. E.; McKerrow, J. H., Leishmania major: molecular modeling of cysteine proteases and prediction of new nonpeptide inhibitors. *Experimental parasitology* **1997,** *87*, 212-21.

6. Que, X.; Brinen, L. S.; Perkins, P.; Herdman, S.; Hirata, K.; Torian, B. E.; Rubin, H.; McKerrow, J. H.; Reed, S. L., Cysteine proteinases from distinct cellular compartments are recruited to phagocytic vesicles by Entamoeba histolytica. *Molecular and Biochemical Parasitology* **2002,** *119*, 23-32.

7. Thiel, K. A., Structure-aided drug design's next generation. *Nature biotechnology* **2004,** *22*, 513-9.

8. Kuroda, D.; Shirai, H.; Jacobson, M. P.; Nakamura, H., Computer-aided antibody design. *Protein engineering, design & selection : PEDS* **2012,** *25*, 507-22.

9. Jacobson, M.; Sali, A., Comparative Protein Structure Modeling and its Applications to Drug Discovery. *Annual Reports in Medicinal Chemistry* **2004,** *39*, 259-276.

10. Hou, S.; Li, B.; Wang, L.; Qian, W.; Zhang, D.; Hong, X.; Wang, H.; Guo, Y., Humanization of an anti-CD34 monoclonal antibody by complementarity-determining region grafting based on computer-assisted molecular modelling. *Journal of biochemistry* **2008,** *144*, 115-20.

11. Staelens, S.; Desmet, J.; Ngo, T. H.; Vauterin, S.; Pareyn, I.; Barbeaux, P.; Van Rompaey, I.; Stassen, J.-M.; Deckmyn, H.; Vanhoorelbeke, K., Humanization by variable domain resurfacing and grafting on a human IgG4, using a new approach for determination of non-human like surface accessible framework residues based on homology modelling of variable domains. *Molecular immunology* **2006,** *43*, 1243-57.

12. Zheng, L.; Manetsch, R.; Woggon, W.-D.; Baumann, U.; Reymond, J.-L., Mechanistic study of proton transfer and hysteresis in catalytic antibody 16E7 by site-directed mutagenesis and homology modeling. *Bioorganic & medicinal chemistry* **2005,** *13*, 1021-9.

13. Skolnick, J., In quest of an empirical potential for protein structure prediction. *Current opinion in structural biology* **2006,** *16*, 166-71.

14. Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E., Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* **2012,** *80*, 2071-9.

15. Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., How Fast-Folding Proteins Fold. *Science* **2011,** *334* (6055), 517-520.

16. Rohl, C. A.; Strauss, C. E.; Chivian, D.; Baker, D., Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **2004,** *55* (3), 656-77.

17. Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B., On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002,** *320* (3), 597-608.

18. Fiser, A.; Do, R. K.; Sali, A., Modeling of loops in protein structures. *Protein Sci* **2000,** *9* (9), 1753-73.

19. Bruccoleri, R. E.; Karplus, M., Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **1987,** *26*, 137-68.

20. Shenkin, P. S.; Yarmush, D. L.; Fine, R. M.; Wang, H. J.; Levinthal, C., Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers* **1987,** *26*, 2053-85.

21. Tomasi, J.; Persico, M., Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chemical Reviews* **1994,** *94*, 2027-2094.

22. Nicholls, A.; Honig, B., A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *Journal of Computational Chemistry* **1991,** *12*, 435-445.

23. Cortis, C. M.; Friesner, R. A., Numerical solution of the Poisson-Boltzmann equation using tetrahedral finite-element meshes. *Journal of Computational Chemistry* **1997,** *18*, 1591-1608.

24. Bashford, D.; Case, D. A., Generalized born models of macromolecular solvation effects. *Annual review of physical chemistry* **2000,** *51*, 129-52.

25. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **1990,** *112*, 6127-6129.

26. Dominy, B. N.; Brooks, C. L., Development of a Generalized Born Model Parametrization for Proteins and Nucleic Acids. *The Journal of Physical Chemistry B* **1999,** *103*, 3765-3773.

27. Jorgensen, W. L.; Tirado-Rives, J., The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **1988,** *110*, 1657-1666.

28. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L., Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides †. *The Journal of Physical Chemistry B* **2001,** *105*, 6474-6487.

29. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996,** *118*, 11225-11236.

30. Li, J.; Abel, R.; Zhu, K.; Cao, Y.; Zhao, S.; Friesner, R. a., The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins* **2011,** *79*, 2794-812.

31. Ghosh, A.; Rapp, C. C. S.; Friesner, R. R. A., Generalized Born Model Based on a Surface Integral Formulation. *The Journal of Physical Chemistry B* **1998,** *102*, 10983-10990.

32. Yu, Z.; Jacobson, M. P.; Friesner, R. a., What role do surfaces play in GB models? A new-generation of surface-generalized born model based on a novel gaussian surface for biomolecules. *Journal of computational chemistry* **2006,** *27*, 72-89.

33. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C., The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. **1997,** *5639*, 3005-3014.

34. Zhu, K.; Shirts, M. R.; Friesner, R. A., Improved methods for side chain and loop predictions via the protein local optimization program: variable dielectric model for implicitly improving the treatment of polarization effects. *Journal of Chemical Theory and Computation* **2007,** (3), 2108–2119.

35. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins* **2003,** *52*, 609-23.

36. Zhu, K.; Pincus, D. L.; Zhao, S. W.; Friesner, R. A., Long loop prediction using the protein local optimization program. *Proteins: Struct., Funct., Bioinf.* **2006,** *65* (2), 438-452.

37. Zhao, S.; Zhu, K.; Li, J.; Friesner, R. A., Progress in super long loop prediction. *Proteins* **2011,** *79* (10), 2920-35.

38. Sellers, B. D.; Zhu, K.; Zhao, S.; Friesner, R. A.; Jacobson, M. P., Toward better refinement of comparative models: predicting loops in inexact environments. *Proteins* **2008,** *72*, 959-71.

39. Miller, E. B.; Murrett, C. S.; Zhu, K.; Zhao, S.; Goldfeld, D. A.; Bylund, J. H.; Friesner, R. A., Prediction of Long Loops with Embedded Secondary Structure using the Protein Local Optimization Program. *Proteins* **2012,** *Submitted.*

40. Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* **1999,** *292*, 195-202.

41. Pollastri, G.; Przybylski, D.; Rost, B.; Baldi, P., Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Struct., Funct., Bioinf.* **2002,** *47* (2), 228-235.

42. Pirovano, W.; Heringa, J., Protein secondary structure prediction. *Methods Mol. Biol.* **2010,** *609*, 327-348.

43. Dorsam, R. T.; Gutkind, J. S., G-protein-coupled receptors and cancer. *Nature reviews. Cancer* **2007,** *7*, 79-94.

44. Evers, A.; Klabunde, T., Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *Journal of medicinal chemistry* **2005,** *48*, 1088-97.

45. de Graaf, C.; Foata, N.; Engkvist, O.; Rognan, D., Molecular modeling of the second extracellular loop of G-protein coupled receptors and its implication on structure-based virtual screening. *Proteins* **2008,** *71*, 599-620.

46. Pierce, K. L.; Premont, R. T.; Lefkowitz, R. J., Seven-transmembrane receptors. *Nature reviews. Molecular cell biology* **2002,** *3*, 639-50.

47. Rosenbaum, D. M.; Rasmussen, S. G. F.; Kobilka, B. K., The structure and function of G-protein-coupled receptors. *Nature* **2009,** *459*, 356-63.

48. Kobilka, B.; Schertler, G. F. X., New G-protein-coupled receptor crystal structures: insights and limitations. *Trends in pharmacological sciences* **2008,** *29*, 79-83.

49. Goldfeld, D. A.; Zhu, K.; Beuming, T.; Friesner, R. A., Successful prediction of the intra- and extracellular loops of four G-protein-coupled receptors. *Proceedings of the National Academy of Sciences of the United States of America* **2011,** *108*, 8275-80.

50. Goldfeld, D. A.; Zhu, K.; Beuming, T.; Friesner, R. A., Loop prediction for a gpcr homology model: Algorithms and results-Loop prediction for a GPCR homology model. *Proteins* **2012**.

51. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A., A hierarchical approach to all-atom protein loop prediction. *Proteins* **2004,** *55* (2), 351-67.

52. Go, N.; Scheraga, H. A., Ring Closure and Local Conformational Deformations of Chain Molecules. *Macromolecules* **1970,** *3* (2), 178-187.

53. Palmer, K. A.; Scheraga, H. A., Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid-geometry approximation. I. Chain closure through a limited search of "loop" conformations. *J. Comput. Chem.* **1991,** *12* (4), 505-526.

54. Moult, J.; James, M. N. G., An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins: Struct., Funct., Bioinf.* **1986,** *1* (2), 146-163.

55. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983,** *4* (2), 187-217.

56. Bassolino-Klimas, D.; Bruccoleri, R. E., Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. *Proteins: Struct., Funct., Bioinf.* **1992,** *14* (4), 465-474.

57. DePristo, M. A.; de Bakker, P. I. W.; Lovell, S. C.; Blundell, T. L., Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles. *Proteins: Struct., Funct., Bioinf.* **2003,** *51* (1), 41-55.

58. de Bakker, P. I. W.; DePristo, M. A.; Burke, D. F.; Blundell, T. L., Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins: Struct., Funct., Bioinf.* **2003,** *51* (1), 21-40.

59. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995,** *117* (19), 5179-5197.

60. Kolodny, R.; Guibas, L.; Levitt, M.; Koehl, P., Inverse Kinematics in Biology: The Protein Loop Closure Problem. *Int. J. Robot. Res.* **2005,** *24* (2-3), 151-163.

61. Petrey, D.; Honig, B., Protein Structure Prediction: Inroads to Biology. *Mol. Cell* **2005,** *20* (6), 811-819.

62. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C., High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007,** *318* (5854), 1258-1265.

63. Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F., Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **2008,** *454* (7203), 486-491.

64. Knight, S.; Andersson, I.; Branden, C. I., Crystallographic analysis of ribulose 1,5-bisphosphate carboxylase from spinach at 2.4 A resolution. Subunit interactions and active site. *J. Mol. Biol.* **1990,** *215* (1), 113-160.

65. Nikiforovich, G. V.; Taylor, C. M.; Marshall, G. R.; Baranski, T. J., Modeling the possible conformations of the extracellular loops in G-protein-coupled receptors. *Proteins: Struct., Funct., Bioinf.* **2010,** *78* (2), 271-285.

66. Zhu, J.; Xie, L.; Honig, B., Structural refinement of protein segments containing secondary structure elements: Local sampling, knowledge-based potentials, and clustering. *Proteins: Struct., Funct., Bioinf.* **2006,** *65* (2), 463-479.

67. Li, X.; Jacobson, M. P.; Friesner, R. A., High-resolution prediction of protein helix positions and orientations. *Proteins: Struct., Funct., Bioinf.* **2004,** *55* (2), 368-382.

68. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28* (1), 235-242.

69. Wang, G. L.; Dunbrack, R. L., PISCES: a protein sequence culling server. *Bioinformatics* **2003,** *19* (12), 1589-1591.

70. Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M., Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991,** *47 (Pt 2)*, 110-119.

71. Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wahlby, A.; Jones, T. A., The Uppsala Electron-Density Server. *Acta Crystallogr D* **2004,** *60*, 2240-2249.

72. Kabsch, W.; Sander, C., Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983,** *22* (12), 2577-2637.

73. Hartigan, J. A., *Clustering algorithms*. John Wiley: New York, 1975.

74. Hartigan, J. A.; Wong, M. A., Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979,** *28* (1), 100-108.

75. Xiang, Z.; Honig, B., Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **2001,** *311* (2), 421-430.

76. Jacobson, M. P.; Kaminski, G. A.; Friesner, R. A.; Rapp, C. S., Force Field Validation Using Protein Side Chain Prediction. *J. Phys. Chem. B* **2002,** *106* (44), 11673-11680.

77. Li, X.; Jacobson, M. P.; Zhu, K.; Zhao, S.; Friesner, R. A., Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. *Proteins: Struct., Funct., Bioinf.* **2007,** *66* (4), 824-837.

78. Cheng, J.; Randall, A. Z.; Sweredoski, M. J.; Baldi, P., SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* **2005,** *33* (suppl 2), W72-W76.

79. Tyagi, M.; Bornot, A.; Offmann, B.; de Brevern, A. G., Analysis of loop boundaries using different local structure assignment methods. *Protein Sci.* **2009,** *18* (9), 1869-1881.

80. Koh, I. Y.; Eyrich, V. A.; Marti-Renom, M. A.; Przybylski, D.; Madhusudhan, M. S.; Eswar, N.; Grana, O.; Pazos, F.; Valencia, A.; Sali, A.; Rost, B., EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* **2003,** *31* (13), 3311-3315.

81. Cendron, L.; Berni, R.; Folli, C.; Ramazzina, I.; Percudani, R.; Zanotti, G., The structure of 2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase provides insights into the mechanism of uric acid degradation. *J. Biol. Chem.* **2007,** *282* (25), 18182-18189.

82. Kim, K.; Park, J.; Rhee, S., Structural and functional basis for (S)-allantoin formation in the ureide pathway. *J. Biol. Chem.* **2007,** *282* (32), 23457-23464.

83. French, J. B.; Ealick, S. E., Structural and mechanistic studies on Klebsiella pneumoniae 2-Oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase. *J. Biol. Chem.* **2010,** *285* (46), 35446-35454.

84. Harris, P. V.; Welner, D.; McFarland, K. C.; Re, E.; Navarro Poulsen, J. C.; Brown, K.; Salbo, R.; Ding, H.; Vlasenko, E.; Merino, S.; Xu, F.; Cherry, J.; Larsen, S.; Lo Leggio, L., Stimulation of lignocellulosic biomass hydrolysis by proteins of glycoside hydrolase family 61: structure and function of a large, enigmatic family. *Biochemistry* **2010,** *49* (15), 3305-3316.

85. Bell, J. A.; Ho, K. L.; Farid, R., Significant reduction in errors associated with nonbonded contacts in protein crystal structures: automated all-atom refinement with PrimeX. *Acta Crystallogr D* **2012,** *68* (Pt 8), 935-952.

86. Protein Structure Initiative: Mission Statement. http://www.nigms.nih.gov/research/specificareas/PSI/background/pages/missionstatement.aspx (accessed 03/30/2015).

87. Moore, P. B., Let's call the whole thing off: some thoughts on the protein structure initiative. *Structure* **2007,** *15* (11), 1350-2.

88. Almagro, J. C.; Fransson, J., Humanization of antibodies. *Front Biosci* **2008,** *13*, 1619-33.

89. Riechmann, L.; Clark, M.; Waldmann, H.; Winter, G., Reshaping human antibodies for therapy. *Nature* **1988,** *332* (6162), 323-7.

90. Brändén, C. I.; Tooze, J., *Introduction to Protein Structure*. Garland Pub.: 1999.

91. Chothia, C.; Lesk, A. M., Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* **1987,** *196* (4), 901-17.

92. Zhu, K.; Day, T., Ab initio structure prediction of the antibody hypervariable H3 loop. *Proteins* **2013,** *81* (6), 1081-9.

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

94. *BioLuminate*, 1.1; Schrödinger, LLC: New York, NY, 2013.

95. Martin, A. C.; Thornton, J. M., Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol* **1996,** *263* (5), 800-15.

96. Zhu, K.; Day, T.; Warshaviak, D.; Murrett, C.; Friesner, R.; Pearlman, D., Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins* **2014,** *82* (8), 1646-55.

97. Smith, T. F.; Waterman, M. S., Identification of common molecular subsequences. *J Mol Biol* **1981,** *147* (1), 195-7.

98. Henikoff, S.; Henikoff, J. G., Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **1992,** *89* (22), 10915-9.

99. Almagro, J. C.; Teplyakov, A.; Luo, J.; Sweet, R. W.; Kodangattil, S.; Hernandez-Guzman, F.; Gilliland, G. L., Second antibody modeling assessment (AMA-II). *Proteins* **2014,** *82* (8), 1553-62.

100. Chothia, C.; Lesk, A. M.; Tramontano, A.; Levitt, M.; Smith-Gill, S. J.; Air, G.; Sheriff, S.; Padlan, E. A.; Davies, D.; Tulip, W. R., Conformations of immunoglobulin hypervariable regions. *Nature* **1989,** *342* (6252), 877-83.

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

102. North, B.; Lehmann, A.; Dunbrack, R. L., A new clustering of antibody CDR loop conformations. *J Mol Biol* **2011,** *406* (2), 228-56.

103. Kuroda, D.; Shirai, H.; Kobori, M.; Nakamura, H., Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins* **2008,** *73* (3), 608-20.

104. Shirai, H.; Kidera, A.; Nakamura, H., Structural classification of CDR-H3 in antibodies. *FEBS Lett* **1996,** *399* (1-2), 1-8.

105. Shirai, H.; Kidera, A.; Nakamura, H., H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Lett* **1999,** *455* (1-2), 188-97.

106. Morea, V.; Tramontano, A.; Rustici, M.; Chothia, C.; Lesk, A. M., Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* **1998,** *275* (2), 269-94.

107. Oliva, B.; Bates, P. A.; Querol, E.; Avilés, F. X.; Sternberg, M. J., Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J Mol Biol* **1998,** *279* (5), 1193-210.

108. Almagro, J. C.; Beavers, M. P.; Hernandez-Guzman, F.; Maier, J.; Shaulsky, J.; Butenhof, K.; Labute, P.; Thorsteinson, N.; Kelly, K.; Teplyakov, A.; Luo, J.; Sweet, R.; Gilliland, G. L., Antibody modeling assessment. *Proteins* **2011,** *79* (11), 3050-66.

109. Xiang, Z.; Soto, C. S.; Honig, B., Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci U S A* **2002,** *99* (11), 7432-7.

110. Mandell, D. J.; Coutsias, E. A.; Kortemme, T., Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* **2009,** *6* (8), 551-2.

111. Stein, A.; Kortemme, T., Improvements to robotics-inspired conformational sampling in rosetta. *PLoS One* **2013,** *8* (5), e63090.

112. Das, R., Atomic-accuracy prediction of protein loop structures through an RNA-inspired Ansatz. *PLoS One* **2013,** *8* (10), e74830.

113. Adhikari, A. N.; Peng, J.; Wilde, M.; Xu, J.; Freed, K. F.; Sosnick, T. R., Modeling large regions in proteins: applications to loops, termini, and folding. *Protein Sci* **2012,** *21* (1), 107-21.

114. Whitelegg, N.; Rees, A. R., Antibody variable regions: toward a unified modeling method. *Methods Mol Biol* **2004,** *248*, 51-91.

115. Goldfeld, D. A.; Zhu, K.; Beuming, T.; Friesner, R. A., Loop prediction for a GPCR homology model: algorithms and results. *Proteins* **2013,** *81* (2), 214-28.

116. Teplyakov, A.; Luo, J.; Obmolova, G.; Malia, T. J.; Sweet, R.; Stanfield, R. L.; Kodangattil, S.; Almagro, J. C.; Gilliland, G. L., Antibody modeling assessment II. Structures and models. *Proteins* **2014,** *82* (8), 1563-82.

117. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** *215* (3), 403-10.

118. Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997,** *25* (17), 3389-402.

119. Chailyan, A.; Tramontano, A.; Marcatili, P., A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res* **2012,** *40* (Database issue), D1230-4.

120. Marcatili, P.; Rosi, A.; Tramontano, A., PIGS: automatic prediction of antibody structures. *Bioinformatics* **2008,** *24* (17), 1953-4.

121. Michalsky, E.; Goede, A.; Preissner, R., Loops In Proteins (LIP)--a comprehensive loop database for homology modelling. *Protein Eng* **2003,** *16* (12), 979-85.

122. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* **2013,** *27* (3), 221-34.

123. *Maestro*, 9.4; Schrödinger, LLC: New York, NY, 2013.

124. Global summary of the AIDS epidemic | 2013. World Health Organiztion - HIV department: 2014.

125. Pre-Exposure Prophylaxis (PrEP). http://www.cdc.gov/hiv/prevention/research/prep/ (accessed 03/30/2015).

126. Grant, R. M.; Lama, J. R.; Anderson, P. L.; McMahan, V.; Liu, A. Y.; Vargas, L.; Goicochea, P.; Casapía, M.; Guanira-Carranza, J. V.; Ramirez-Cardich, M. E.; Montoya-Herrera, O.; Fernández, T.; Veloso, V. G.; Buchbinder, S. P.; Chariyalertsak, S.; Schechter, M.; Bekker, L. G.; Mayer, K. H.; Kallás, E. G.; Amico, K. R.; Mulligan, K.; Bushman, L. R.; Hance, R. J.; Ganoza, C.; Defechereux, P.; Postle, B.; Wang, F.; McConnell, J. J.; Zheng, J. H.; Lee, J.; Rooney, J. F.; Jaffe, H. S.; Martinez, A. I.; Burns, D. N.; Glidden, D. V.; Team, i. S., Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *N Engl J Med* **2010,** *363* (27), 2587-99.

127. Thigpen, M. C.; Kebaabetswe, P. M.; Paxton, L. A.; Smith, D. K.; Rose, C. E.; Segolodi, T. M.; Henderson, F. L.; Pathak, S. R.; Soud, F. A.; Chillag, K. L.; Mutanhaurwa, R.; Chirwa, L. I.; Kasonde, M.; Abebe, D.; Buliva, E.; Gvetadze, R. J.; Johnson, S.; Sukalac, T.; Thomas, V. T.; Hart, C.; Johnson, J. A.; Malotte, C. K.; Hendrix, C. W.; Brooks, J. T.; Group, T. S., Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana. *N Engl J Med* **2012,** *367* (5), 423-34.

128. Baeten, J. M.; Donnell, D.; Ndase, P.; Mugo, N. R.; Campbell, J. D.; Wangisi, J.; Tappero, J. W.; Bukusi, E. A.; Cohen, C. R.; Katabira, E.; Ronald, A.; Tumwesigye, E.; Were, E.; Fife, K. H.; Kiarie, J.; Farquhar, C.; John-Stewart, G.; Kakia, A.; Odoyo, J.; Mucunguzi, A.; Nakku-Joloba, E.; Twesigye, R.; Ngure, K.; Apaka, C.; Tamooh, H.; Gabona, F.; Mujugira, A.; Panteleeff, D.; Thomas, K. K.; Kidoguchi, L.; Krows, M.; Revall, J.; Morrison, S.; Haugen, H.; Emmanuel-Ogier, M.; Ondrejcek, L.; Coombs, R. W.; Frenkel, L.; Hendrix, C.; Bumpus, N. N.; Bangsberg, D.; Haberer, J. E.; Stevens, W. S.; Lingappa, J. R.; Celum, C.; Team, P. P. S., Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. *N Engl J Med* **2012,** *367* (5), 399-410.

129. Choopanya, K.; Martin, M.; Suntharasamai, P.; Sangkum, U.; Mock, P. A.; Leethochawalit, M.; Chiamwongpaet, S.; Kitisin, P.; Natrujirote, P.; Kittimunkong, S.; Chuachoowong, R.; Gvetadze, R. J.; McNicholl, J. M.; Paxton, L. A.; Curlin, M. E.; Hendrix, C. W.; Vanichseni, S.; Group, B. T. S., Antiretroviral prophylaxis for HIV infection in injecting drug users in Bangkok, Thailand (the Bangkok Tenofovir Study): a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet* **2013,** *381* (9883), 2083-90.

130. Wyatt, R.; Sodroski, J., The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science* **1998,** *280* (5371), 1884-8.

131. Mascola, J. R.; Haynes, B. F., HIV-1 neutralizing antibodies: understanding nature's pathways. *Immunol Rev* **2013,** *254* (1), 225-44.

132. Pantophlet, R.; Burton, D. R., GP120: target for neutralizing HIV-1 antibodies. *Annu Rev Immunol* **2006,** *24*, 739-69.

133. Kwong, P. D.; Doyle, M. L.; Casper, D. J.; Cicala, C.; Leavitt, S. A.; Majeed, S.; Steenbeke, T. D.; Venturi, M.; Chaiken, I.; Fung, M.; Katinger, H.; Parren, P. W.; Robinson, J.; Van Ryk, D.; Wang, L.; Burton, D. R.; Freire, E.; Wyatt, R.; Sodroski, J.; Hendrickson, W. A.; Arthos, J., HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **2002,** *420* (6916), 678-82.

134. Zhou, T.; Georgiev, I.; Wu, X.; Yang, Z. Y.; Dai, K.; Finzi, A.; Kwon, Y. D.; Scheid, J. F.; Shi, W.; Xu, L.; Yang, Y.; Zhu, J.; Nussenzweig, M. C.; Sodroski, J.; Shapiro, L.; Nabel, G. J.; Mascola, J. R.; Kwong, P. D., Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* **2010,** *329* (5993), 811-7.

135. Kwong, P. D.; Mascola, J. R., Human antibodies that neutralize HIV-1: identification, structures, and B cell ontogenies. *Immunity* **2012,** *37* (3), 412-25.

136. Balazs, A. B.; Chen, J.; Hong, C. M.; Rao, D. S.; Yang, L.; Baltimore, D., Antibody-based protection against HIV infection by vectored immunoprophylaxis. *Nature* **2012,** *481* (7379), 81-4.

137. Shingai, M.; Nishimura, Y.; Klein, F.; Mouquet, H.; Donau, O. K.; Plishka, R.; Buckler-White, A.; Seaman, M.; Piatak, M.; Lifson, J. D.; Dimitrov, D. S.; Nussenzweig, M. C.; Martin, M. A., Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* **2013,** *503* (7475), 277-80.

138. Wong, C. F.; McCammon, J. A., Dynamics and design of enzymes and inhibitors. *Journal of the American Chemical Society* **1986,** *108* (13), 3830-3832.

139. McCammon, J. A.; Gelin, B. R.; Karplus, M., Dynamics of folded proteins. *Nature* **1977,** *267* (5612), 585-90.

140. Merz, K. M.; Kollman, P. A., Free energy perturbation simulations of the inhibition of thermolysin: prediction of the free energy of binding of a new inhibitor. *Journal of the American Chemical Society* **1989,** *111* (15), 5649-5658.

141. Kollman, P., Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews* **1993,** *93* (7), 2395-2417.

142. Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A., Calculation of the relative change in binding free energy of a protein-inhibitor complex. *Science* **1987,** *235* (4788), 574-6.

143. Jorgensen, W. L.; Ravimohan, C., Monte Carlo simulation of differences in free energies of hydration. *The Journal of Chemical Physics* **1985,** *83* (6), 3050-3054.

144. Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R., Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc* **2015,** *137* (7), 2695-703.

145. Wang, L.; Friesner, R. A.; Berne, B. J., Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J Phys Chem B* **2011,** *115* (30), 9431-8.

146. Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R., Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *Journal of Chemical Theory and Computation* **2013,** *9* (2), 1282-1293.

147. Wang, L.; Berne, B. J.; Friesner, R. A., On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. *Proc Natl Acad Sci U S A* **2012,** *109* (6), 1937-42.

148. Wu, X.; Yang, Z. Y.; Li, Y.; Hogerkorp, C. M.; Schief, W. R.; Seaman, M. S.; Zhou, T.; Schmidt, S. D.; Wu, L.; Xu, L.; Longo, N. S.; McKee, K.; O'Dell, S.; Louder, M. K.; Wycuff, D. L.; Feng, Y.; Nason, M.; Doria-Rose, N.; Connors, M.; Kwong, P. D.; Roederer, M.; Wyatt, R. T.; Nabel, G. J.; Mascola, J. R., Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* **2010,** *329* (5993), 856-61.

149. Stamatatos, L.; Morris, L.; Burton, D. R.; Mascola, J. R., Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? *Nat Med* **2009,** *15* (8), 866-70.

150. Sather, D. N.; Armann, J.; Ching, L. K.; Mavrantoni, A.; Sellhorn, G.; Caldwell, Z.; Yu, X.; Wood, B.; Self, S.; Kalams, S.; Stamatatos, L., Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. *J Virol* **2009,** *83* (2), 757-69.

151. Simek, M. D.; Rida, W.; Priddy, F. H.; Pung, P.; Carrow, E.; Laufer, D. S.; Lehrman, J. K.; Boaz, M.; Tarragona-Fiol, T.; Miiro, G.; Birungi, J.; Pozniak, A.; McPhee, D. A.; Manigart, O.; Karita, E.; Inwoley, A.; Jaoko, W.; Dehovitz, J.; Bekker, L. G.; Pitisuttithum, P.; Paris, R.; Walker, L. M.; Poignard, P.; Wrin, T.; Fast, P. E.; Burton, D. R.; Koff, W. C., Human immunodeficiency virus type 1 elite neutralizers: individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. *J Virol* **2009,** *83* (14), 7337-48.

152. Li, Y.; Svehla, K.; Louder, M. K.; Wycuff, D.; Phogat, S.; Tang, M.; Migueles, S. A.; Wu, X.; Phogat, A.; Shaw, G. M.; Connors, M.; Hoxie, J.; Mascola, J. R.; Wyatt, R., Analysis of neutralization specificities in polyclonal sera derived from human immunodeficiency virus type 1-infected individuals. *J Virol* **2009,** *83* (2), 1045-59.

153. Li, Y.; Migueles, S. A.; Welcher, B.; Svehla, K.; Phogat, A.; Louder, M. K.; Wu, X.; Shaw, G. M.; Connors, M.; Wyatt, R. T.; Mascola, J. R., Broad HIV-1 neutralization mediated by CD4-binding site antibodies. *Nat Med* **2007,** *13* (9), 1032-4.

154. Wyatt, R.; Kwong, P. D.; Desjardins, E.; Sweet, R. W.; Robinson, J.; Hendrickson, W. A.; Sodroski, J. G., The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **1998,** *393* (6686), 705-11.

155. Wu, X.; Zhou, T.; Zhu, J.; Zhang, B.; Georgiev, I.; Wang, C.; Chen, X.; Longo, N. S.; Louder, M.; McKee, K.; O'Dell, S.; Perfetto, S.; Schmidt, S. D.; Shi, W.; Wu, L.; Yang, Y.; Yang, Z. Y.; Yang, Z.; Zhang, Z.; Bonsignori, M.; Crump, J. A.; Kapiga, S. H.; Sam, N. E.; Haynes, B. F.; Simek, M.; Burton, D. R.; Koff, W. C.; Doria-Rose, N. A.; Connors, M.; Mullikin, J. C.; Nabel, G. J.; Roederer, M.; Shapiro, L.; Kwong, P. D.; Mascola, J. R.; Program, N. C. S., Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **2011,** *333* (6049), 1593-602.

156. Scheid, J. F.; Mouquet, H.; Ueberheide, B.; Diskin, R.; Klein, F.; Oliveira, T. Y.; Pietzsch, J.; Fenyo, D.; Abadir, A.; Velinzon, K.; Hurley, A.; Myung, S.; Boulad, F.; Poignard, P.; Burton, D. R.; Pereyra, F.; Ho, D. D.; Walker, B. D.; Seaman, M. S.; Bjorkman, P. J.; Chait, B. T.; Nussenzweig, M. C., Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **2011,** *333* (6049), 1633-7.

157. Wrammert, J.; Koutsonanos, D.; Li, G. M.; Edupuganti, S.; Sui, J.; Morrissey, M.; McCausland, M.; Skountzou, I.; Hornig, M.; Lipkin, W. I.; Mehta, A.; Razavi, B.; Del Rio, C.; Zheng, N. Y.; Lee, J. H.; Huang, M.; Ali, Z.; Kaur, K.; Andrews, S.; Amara, R. R.; Wang, Y.; Das, S. R.; O'Donnell, C. D.; Yewdell, J. W.; Subbarao, K.; Marasco, W. A.; Mulligan, M. J.; Compans, R.; Ahmed, R.; Wilson, P. C., Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *J Exp Med* **2011,** *208* (1), 181-93.

158. Zhou, T.; Zhu, J.; Wu, X.; Moquin, S.; Zhang, B.; Acharya, P.; Georgiev, I. S.; Altae-Tran, H. R.; Chuang, G. Y.; Joyce, M. G.; Do Kwon, Y.; Longo, N. S.; Louder, M. K.; Luongo, T.; McKee, K.; Schramm, C. A.; Skinner, J.; Yang, Y.; Yang, Z.; Zhang, Z.; Zheng, A.; Bonsignori, M.; Haynes, B. F.; Scheid, J. F.; Nussenzweig, M. C.; Simek, M.; Burton, D. R.; Koff, W. C.; Mullikin, J. C.; Connors, M.; Shapiro, L.; Nabel, G. J.; Mascola, J. R.; Kwong, P. D.; Program, N. C. S., Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **2013,** *39* (2), 245-58.

159. Joyce, M. G.; Kanekiyo, M.; Xu, L.; Biertümpfel, C.; Boyington, J. C.; Moquin, S.; Shi, W.; Wu, X.; Yang, Y.; Yang, Z. Y.; Zhang, B.; Zheng, A.; Zhou, T.; Zhu, J.; Mascola, J. R.; Kwong, P. D.; Nabel, G. J., Outer domain of HIV-1 gp120: antigenic optimization, structural malleability, and crystal structure with antibody VRC-PG04. *J Virol* **2013,** *87* (4), 2294-306.

160. Klein, F.; Diskin, R.; Scheid, J. F.; Gaebler, C.; Mouquet, H.; Georgiev, I. S.; Pancera, M.; Zhou, T.; Incesu, R. B.; Fu, B. Z.; Gnanapragasam, P. N.; Oliveira, T. Y.; Seaman, M. S.; Kwong, P. D.; Bjorkman, P. J.; Nussenzweig, M. C., Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell* **2013,** *153* (1), 126-38.

161. Georgiev, I. S.; Doria-Rose, N. A.; Zhou, T.; Kwon, Y. D.; Staupe, R. P.; Moquin, S.; Chuang, G. Y.; Louder, M. K.; Schmidt, S. D.; Altae-Tran, H. R.; Bailer, R. T.; McKee, K.; Nason, M.; O'Dell, S.; Ofek, G.; Pancera, M.; Srivatsan, S.; Shapiro, L.; Connors, M.; Migueles, S. A.; Morris, L.; Nishimura, Y.; Martin, M. A.; Mascola, J. R.; Kwong, P. D., Delineating antibody recognition in polyclonal sera from patterns of HIV-1 isolate neutralization. *Science* **2013,** *340* (6133), 751-6.

162. Abhinandan, K. R.; Martin, A. C., Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol Immunol* **2008,** *45* (14), 3832-9.

163. Martin, A. C. R., abYsis: a fully integrated antibody discovery system. 2014.

164. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A., A hierarchical approach to all-atom protein loop prediction. *Proteins* **2004,** *55*, 351-67.

165. Miller, E. B.; Murrett, C. S.; Zhu, K.; Zhao, S.; Goldfeld, D. A.; Friesner, R. A., Prediction of Long Loops with Embedded Secondary Structure using the Protein Local Optimization Program. *Submitted* **2012**.

166. Bowers, K. J.; Chow, E.; Huageng, X.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Yibing, S.; Shaw, D. E. In *Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters*, SC 2006 Conference, Proceedings of the ACM/IEEE, 11-17 Nov. 2006; 2006; pp 43-43.

167. *BioLuminate*, 1.5; Schrödinger, LLC: New York, NY, 2014.