

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Physics-based Refinement of Proteins in Model Systems

Permalink

<https://escholarship.org/uc/item/1tv93190>

Author

Sellers, Benjamin D

Publication Date

2008-06-13

Peer reviewed|Thesis/dissertation

Physics-based Refinement of Proteins in Model Systems

by

Benjamin D. Sellers

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Graduate Group in Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright (2008)

By

Benjamin D. Sellers

Acknowledgements

This work would not have been possible without many key collaborators. I would first like to thank Matthew Jacobson for unending support and great discussion. This work relies heavily on his previous computational innovations and hard work. I would also like to thank the other members of my thesis committee, Ken Dill and Andrej Sali for their helpful direction. In addition, principal collaborators who have been instrumental to this work are Kai Zhu, Arjun Narayanan, Sergio Wong, Jeremy Wilbur, Greg Kapp, Jerome Nilmeier and Chris Farady. I would like to thank the Jacobson lab their help and support. I would also like to thank Jan E. Gudell for bringing me to an E. O. Wilson lecture at Harvard which led to my decision to get a Ph.D. in Biology. Most of all, I appreciate the love and support from Holly, Bodhi, Niko, Mom, Dad, Liz, Michael, Gail, Heather, Matt, and my wonderful friends.

Chapter 1 is an article originally published online in Feb, 2008 in *Proteins: Structure Function and Bioinformatics*. Kai Zhu carried out half of the calculations and Suwen Zhao refined a homology model presented in the work.

Chapter 2 is a manuscript *in preparation*. Jerome Nilmeier helped with preliminary results at the start of this project.

Chapter 3 is a manuscript *in preparation*. It was written primarily by Arjun Narayanan and was partially edited by Matthew Jacobson and me. All analysis was carried out by Arjun Narayanan which used, in part, a test set, multiple alignment, and antibody

modeling and sequence identity scripts that I created. All citations and introductory remarks on kinases and domain-prediction were written by me.

Chapter 4 is a manuscript *in preparation*. It was written primarily by Sergio Wong, my collaborator, except for methods, analysis, and figures describing H3 loop minima enumeration, which I wrote. It was partially edited by Matthew Jacobson and me. All molecular dynamics calculations and analysis were carried out by Sergio while all loop minima enumeration work was carried out by me.

Chapter 5 is a methods report that is an elaboration on a manuscript *in preparation*.

Jeremy Wilbur, Peter Hwang, Michael Lane, Joel Ybe, Ben Sellers, Matthew P. Jacobson, Robert Fletterick, Frances Brodsky. "Regulation of clathrin lattice assembly by conformational switching in clathrin light chain". *in prep*

Chapter 6 is a manuscript summary that may not be published. Matthew Jacobson, Gabriela Barreiro and Chris Farady aided in discussion of this work.

This dissertation is dedicated to Holly, Bodhi and Nikolao

Physics-based Refinement of Proteins in Model Systems

Benjamin D. Sellers

Abstract

More accurate comparative (homology) models would enable greater biological understanding through structural genomics efforts as well as aid in biological and small-molecule drug development. However, improving the accuracy of comparative models beyond that of the homologous template protein has proven extremely difficult for many years. The primary aim of this work is to develop more accurate, molecular mechanics-based, computational methods for refining loops in comparative models. My approach is two-fold:

- A. Create a set of protein “model systems” that exhibit specific types of modeling error as found surrounding loops in comparative models
- B. Develop new loop-sampling methods that optimize atoms outside the loop

The types of structural errors found in comparative models can be divided into bins based on sampling degrees-of-freedom: 1. side-chain error, 2. backbone error, and 3. larger-scale structural errors, such as helix or domain orientations. In chapter 1, we perturbed crystal structures to contain side-chain errors exclusively (error type 1.) We then augmented our previous loop prediction method to simultaneously optimize side-chains surrounding the loop. Results show that our new method can recover the native state in

most of the cases where our previous method failed. In chapter 2, we chose homology models of antibodies as a test system to investigate loop prediction when the surrounding backbone atoms are incorrect (error type 2.) We predict the antibody H3 hyper-variable loop *ab initio* while the remaining five, hyper-variable loops are modeled using loop templates whose backbone atoms tend to deviate slightly from native. By increasing H3 loop sampling and performing optimizations on the surrounding loops iteratively, we were able to increase accuracy over previous methods. In chapter 3, through collaboration with Arjun Narayanan, we have taken initial steps in analyzing the determinants of variation in antibody light and heavy domain orientation (error type 3.) In chapter 4, through collaboration with Sergio Wong, I applied these new loop prediction methods to investigate a previous hypothesis that antibody H3 loops rigidify during affinity maturation which occurs within B-cells upon antigen encounter. In summary, this work constitutes a significant step towards a general method of comparative model refinement.

Table of Contents

INTRODUCTION.....	1
THE HUMAN GENOME PROJECT AND STRUCTURAL GENOMICS.....	1
COMPARATIVE MODELING AND THE NEED FOR REFINEMENT OF MODELS	2
OUR APPROACH.....	4
OUR FINDINGS	4
FUTURE DIRECTIONS.....	8
REFERENCES	10
CHAPTER 1: TOWARDS BETTER REFINEMENT OF COMPARATIVE MODELS:	
PREDICTING LOOPS IN INEXACT ENVIRONMENTS	11
ABSTRACT.....	12
INTRODUCTION.....	14
METHODS.....	18
RESULTS AND DISCUSSION	28
CONCLUSION/FURTHER DIRECTIONS	37
REFERENCES	41
CHAPTER 2: ANTIBODIES AS A MODEL SYSTEM FOR COMPARATIVE MODEL	
REFINEMENT.....	65
ABSTRACT.....	66
INTRODUCTION.....	67
METHODS.....	71
RESULTS.....	78
DISCUSSION.....	81
REFERENCES	84
CHAPTER 3: RELATIVE ORIENTATION OF HEAVY AND LIGHT CHAIN VARIABLE	
DOMAINS IS A MANIFESTATION OF STRUCTURAL DIVERSITY IN ANTIBODIES.....	97
ABSTRACT.....	98
INTRODUCTION.....	99
METHODS.....	102
RESULTS.....	107
DISCUSSION.....	117
REFERENCES	121
CHAPTER 4: HOW DO SOMATIC MUTATIONS RIGIDIFY CDR LOOPS DURING AFFINITY	
MATURATION?.....	135
ABSTRACT.....	136
INTRODUCTION.....	137

METHODOLOGY	140
RESULTS AND DISCUSSION	142
CONCLUSION	150
REFERENCES	152
CHAPTER 5: ALL-ATOM MODEL OF CLATHRIN HUB USING COMPARATIVE MODELING AND ELECTRON-MICROSCOPY DATA ENABLES X-RAY CRYSTALLOGRAPHIC SOLUTION BY MOLECULAR REPLACEMENT	168
SUMMARY	169
METHODS	170
REFERENCES	171
CHAPTER 6: INVESTIGATION OF PH-DEPENDENT INHIBITION OF MEMBRANE-TYPE SERINE PROTEASE 1 BY A PICO-MOLAR ANTIBODY INHIBITOR USING CONSTANT-PH MOLECULAR DYNAMICS SIMULATION.....	175
SUMMARY	176
METHODS	177
RESULTS.....	178
FUTURE DIRECTIONS.....	179
REFERENCES	181

List of Tables

CHAPTER 1: TABLE I	44
CHAPTER 1: TABLE II	45
CHAPTER 1: TABLE S 1	58
CHAPTER 1: TABLE S 2	59
CHAPTER 1: TABLE S 3	60
CHAPTER 1: TABLE S 4	61
CHAPTER 1: TABLE S 5	63
CHAPTER 1: TABLE S 6	64
CHAPTER 2: TABLE I	87
CHAPTER 2: TABLE S 1	95
CHAPTER 3: TABLE I	123
CHAPTER 3: TABLE I	123
CHAPTER 3: TABLE II	124
CHAPTER 4: TABLE 1	156
CHAPTER 4: TABLE 2	157

List of figures

CHAPTER 1: FIGURE 1	46
CHAPTER 1: FIGURE 2	47
CHAPTER 1: FIGURE 3	48
CHAPTER 1: FIGURE 4	50
CHAPTER 1: FIGURE 5	51
CHAPTER 1: FIGURE 6	52
CHAPTER 1: FIGURE 7	53
CHAPTER 1: FIGURE 8	54
CHAPTER 1: FIGURE 9	55
CHAPTER 1: FIGURE 10	56
CHAPTER 2: FIGURE 1	89
CHAPTER 2: FIGURE 2	90
CHAPTER 2: FIGURE 3	92
CHAPTER 2: FIGURE 4	93
CHAPTER 2: FIGURE 5	94
CHAPTER 3: FIGURE 1	125
CHAPTER 3: FIGURE 2	126
CHAPTER 3: FIGURE 3	128
CHAPTER 3: FIGURE 4	129
CHAPTER 3: FIGURE 5	130
CHAPTER 3: FIGURE 6	131
CHAPTER 3: FIGURE 7	132
CHAPTER 3: FIGURE 8	133
CHAPTER 3: FIGURE 9	134
CHAPTER 4: FIGURE 1	158
CHAPTER 4: FIGURE 2	159
CHAPTER 4: FIGURE 3	160
CHAPTER 4: FIGURE 4	161
CHAPTER 4: FIGURE 5	162
CHAPTER 4: FIGURE 6	163
CHAPTER 4: FIGURE 7	164
CHAPTER 5: FIGURE 1	172
CHAPTER 5: FIGURE 2	173
CHAPTER 6: FIGURE 1	183
CHAPTER 6: FIGURE 2	184
CHAPTER 6: FIGURE 3	186
CHAPTER 6: FIGURE 4	188

Introduction

The Human Genome Project and Structural Genomics

In 2001, just before the research described in this dissertation began, a working draft of the entire human genome was completed^{1,2}. This scientific landmark, like all discoveries, poses at least as many new questions as it answers. The Human Genome Project unlocked a blueprint that describes the parts of a complicated machine, the cell. But unlike the blueprint for a car engine which details the shape and interconnections of fans and belts, the human genome alone contains no information on the *shape or function* of the parts. DNA holds a highly-compressed and convoluted message that requires a containing cell to decipher its full meaning. With 20,000-25,000 coding genes within the human cell³, and with each species on earth expressing its own unique set of genes, how will we ever answer the questions: *What are the functions of the proteins that these genes encode? With which other proteins do they interact? How do changes in proteins lead to disease? How have these proteins evolved? How do proteins function at the molecular level?*

Researchers have been working on these questions for many decades. To address the question of how a protein functions at the molecular level, researchers may attempt to determine the three-dimensional atomic structure of a protein. Most work, however, has focused on a single protein at a time, often over many years and at great expense. For example, in 2005, the New York Structural GenomiX Research Consortium (NYSGXRC) reported spending an average of \$109,641 per protein structure it produced⁴. (Note: this number does include some indirect costs.)

NYSGXRC is part of Structural Genomics, a world-wide, post-genomic effort to determine large numbers of protein structures. In a similar high-throughput approach as the genome sequencing projects, Structural Genomics aims to streamline the process of protein cloning, expression, purification and structure determination. However, given the time and cost for each individual protein experiment, only a small percentage of all proteins will be characterized in this effort.

Comparative modeling and the need for refinement of models

The designers of Structural Genomics were well aware of this limitation and incorporated a key aim to leverage the limited number of experimentally determined protein structures in order to gain knowledge about the remaining majority that would not be characterized. *Computer modeling* holds this key by taking advantage of a known relationship: proteins that are related through evolution, known as homologs, often have similar atomic structures. Comparative modeling, also known as homology modeling, can generate virtual atomic structures of unknown proteins using known atomic structures of homologous proteins⁵. Vitkup et al⁶ estimated that 90% of all uncharacterized protein sequences within a protein family could be modeled if on average two proteins per family are rationally selected for experimental structure determination.

The accuracy of these comparative models varies, however, and consequently their usefulness to further computational analysis varies. For example, one possible application is to use computers to determine which drugs will inhibit a disease-related

protein by “docking” small molecules into a computer model of a protein. Successfully identifying *true* inhibitors is dependent on the accuracy of the protein model used⁷. Many other applications that use comparative models exist and each has a requirement for model accuracy⁸. In general, more accurate models will be more useful to research in basic science and health.

There are two primary roadblocks to more reliably accurate comparative models, namely difficulties in 1) identifying and aligning to a homolog template sequence and 2) refining regions in the initial model that potentially differ structurally from the target protein. In general, while much progress has been made in the alignment step, little has been made in the refinement step⁹.

The field of comparative modeling refinement is measured to some degree every two years in the Critical Assessment of Techniques for Protein Structure Prediction (CASP)⁹, a blind test of computational methods. Experimental structures are temporarily withheld from publication while hundreds of computational teams across the globe submit protein models based solely on the protein’s amino-acid sequence. In the template-based (comparative) modeling category, to the dismay of everyone, no team has been able to improve on the accuracy of the best homolog template protein across a majority of the test cases. On average, researchers who attempt to refine the starting model make the models worse¹⁰.

Our approach

As detailed in the following dissertation, our approach is to simplify the problem of refining comparative models. First, we only focus on the prediction of protein loops. Loops are a good choice since the protein *fold* (i.e. everything except the loops) of comparative models are structurally conserved if a homolog protein is available with >30% sequence identity⁵. Second, we focus on “model systems,” computer models of proteins that exhibit only one problem found in comparative models at a time. By reducing the complexity in refining comparative models, we aim to address each issue in turn.

Our Findings

In this work, we report new insights into the causes of model refinement failure and we have taken significant steps toward addressing these difficulties with 1. the creation of novel, simplified test sets and 2. the development of novel computational methods. These advances benefit multiple fields of research, namely, *comparative model refinement*, *loop modeling*, and *antibody modeling*.

Greater understanding of difficulties in comparative model refinement

In Chapter 1, we show that even small perturbations in side chains that surround loops greatly decreases loop prediction accuracy using our previously-published, state-of-the-art method. We were surprised at this simple result but we clearly show that small errors in surrounding residue side-chain position can create 1. sampling problems by blocking our method from testing native-like loops and 2. energy function problems where native-like loops are scored higher in energy due to subtle, misaligned energetic contacts between the loop and surroundings. In chapter 2, we show how small errors in the modeled backbone of residues surrounding loops exacerbate these sampling and energy function failures.

Novel application of simplified test systems to comparative model refinement

Our validation of refinement methods using simplified protein systems (crystal structures with perturbed side chains in chapter 1 and antibody comparative models in chapter 2) is a novel approach in the field of comparative model refinement. This field has seen little progress using “unfocused” test sets, collections of proteins that contain a wide variety of protein targets with varying structure, quality, and crystal environment (e.g. CASP). That we have made progress using our simplified test systems which aim to de-convolute the causes of refinement failure is evidence that our novel approach is working and should be a benefit to the field.

Furthermore, by expanding our loop modeling goals beyond reproduction of loops in crystal structures and on to predictions within error-prone, modeled environments, we have implicitly redefined the *loop prediction problem* for the loop modeling field. Though predicting loops within protein models is more difficult than within crystal structures, we believe this shift will be more beneficial to real-world, loop-prediction applications.

Development of novel refinement methods

We have methodically developed new prediction algorithms that refine loops when the surrounding side-chain and backbone atoms are modeled incorrectly. These methods additionally sample the loop environment simultaneously with the predicted loop. Our success in these developments is important because it is generally understood in the field that increasing sampling degrees of freedom increases risk of low-energy decoys. Most importantly, though anecdotal at this point, we have shown in blind predictions that these methods can improve the accuracy of loops in full comparative models beyond the starting template protein, a task generally understood to be difficult.

In chapter 2, though our primary focus is developing loop prediction methods when surrounding backbone atoms are incorrect, we have simultaneously developed a method for predicting the complete structure of antibodies. We created a hybrid method by utilizing knowledge-based loop modeling for five of the CDR loops followed by Physics-based, *ab initio* modeling for the sixth H3 CDR loop. While similar methods have been developed¹¹, our optimization of the surrounding CDR's simultaneously with the H3 loop

prediction is completely novel. However, further work is needed to validate this approach as a general antibody prediction protocol.

Further advances

In collaboration with Arjun Narayanan, in chapter 3, we present the first structural-bioinformatics analysis of antibody heavy and light chain orientations. Our work is a first step toward predicting antibody domain orientation, which is itself a necessary step in predicting complete antibody models. More accurate antibody models would benefit antibody engineering, humanization, phage-display library design, and epitope mapping. We found a large variation in domain orientation throughout antibodies. We also show that this variation is due to the amino-acid content and backbone structure of the domain interface and that crystal packing and antigen effects are minor contributors to domain orientation. We conclude the large variability in domain orientation is a mechanism for antibody-antigen recognition in addition to the well-known CDR loop variability.

In collaboration with Sergio Wong in chapter 4, we validate the previous hypothesis that antibodies rigidify upon affinity-maturation. This hypothesis was previously based on “macroscopic” evidence, crystal (i.e. averaged) structures of germline and mature antibodies. Here, we present a “microscopic and dynamic” view into this process using two different computational methods and show that antibody rigidification occurs using multiple physical mechanisms.

In Chapter 5, I detail methods for producing an all-atom model of the clathrin hub which (following years of failed attempts) enabled solving of phases by molecular replacement of a low-resolution crystal structure. The structure contains novel information concerning clathrin heavy chain regulation by clathrin light chain.

In Chapter 6, I describe constant-pH molecular dynamics simulations at pH 6 and 8 of a pico-Molar inhibitor in complex with MT-SP1, membrane bound serine protease implicated in multiple forms of cancer. I observe dramatic differences in the stability of the complex, providing a microscopic hypothesis for pH-dependent catalysis by correlating charge-changes of key titratable residues to complex dissociate at low pH.

Future Directions

Beyond the work described here, there are several directions to further improve accuracy in comparative model refinement. The first approach is to continue with the plan described in this work. Future loop prediction methods would be more useful if they can account for even larger structural variation in the surroundings (ex. displaced helices or domains). Loop enumeration and minima sampling methods used in this work are very fast in comparison to more rigorous sampling approaches.

In homology models where sequence identity between target and template drop below 30%, the surroundings of the loop begin to require as much refinement as the loops themselves. In these cases, methods that can sample multiple structural movements simultaneously are needed. Molecular dynamics (MD), time-based simulation of each

atom in the protein, offers a clear solution. Recent applications to comparative modeling refinement have been limited however, and successful refinement appears to require long simulation times or advanced sampling techniques such as replica exchange^{12,13}. But sampling power using MD will only improve as computers become faster. Monte Carlo (MC) sampling methods may be the answer, either exclusively or in conjunction with MD. By coordinating multiple move sets, quickly refining multiple protein regions in comparative models may be possible. Furthermore, MD and Monte Carlo sampling will enable the field to move away from the presumption that there is a single native structure for a protein and move towards producing statistical ensembles that more accurately capture the *true* flexible nature of proteins.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. Initial sequencing and analysis of the human genome. Volume 409; 2001. p 860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA. The Sequence of the Human Genome. 2001;291(5507):1304-1351.
3. Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. Nature 2004;431(7011):931-945.
4. Bonanno JB, Almo SC, Bresnick A, Chance MR, Fiser A, Swaminathan S, Jiang J, Studier FW, Shapiro L, Lima CD. New York-Structural GenomiX Research Consortium (NYSGXRC): A Large Scale Center for the Protein Structure Initiative. Volume 6: Springer; 2005. p 225-232.
5. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. COMPARATIVE PROTEIN STRUCTURE MODELING OF GENES AND GENOMES. Volume 29: Annual Reviews; 2000. p 291-325.
6. Vitkup D, Melamud E, Moulton J, Sander C. Completeness in structural genomics. Volume 8; 2001. p 559-566.
7. McGovern SL, Shoichet BK. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 2003;46(14):2895-2907.
8. Baker D, Sali A. Protein Structure Prediction and Structural Genomics. Volume 294; 2001. p 93-96.
9. Moulton J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction-Round VII. Proteins 2007;69 Suppl 8:3-9.
10. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69 Suppl 8:38-56.
11. Whitelegg NRJ, Rees AR. WAM: an improved algorithm for modelling antibodies on the WEB. Protein Eng 2000;13(12):819-824.
12. Fan H, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. Volume 13; 2004. p 211.
13. Chen J, Brooks III CL. Can Molecular Dynamics Simulations Provide High-Resolution Refinement of Protein Structure? Volume 67; 2007. p 922-930.

Chapter 1: Towards better refinement of comparative models: predicting loops in inexact environments

Benjamin D. Sellers^{1†}, Kai Zhu^{2†}, Suwen Zhao², Richard A. Friesner², Matthew P. Jacobson³

† These authors contributed equally to this work.

¹Graduate Group in Biophysics, University of California, San Francisco, California

²Department of Chemistry, Columbia University, New York, New York

³Department of Pharmaceutical Chemistry, University of California, San Francisco, California

Abstract

Achieving atomic-level accuracy in comparative protein models is limited by our ability to refine the initial, homolog-derived model closer to the native state. Despite considerable effort, progress in developing a generalized refinement method has been limited. In contrast, methods have been described that can accurately reconstruct loop conformations in native protein structures. We hypothesize that loop refinement in homology models is much more difficult than loop reconstruction in crystal structures, in part, because side-chain, backbone, and other structural inaccuracies surrounding the loop create a challenging sampling problem; the loop cannot be refined without simultaneously refining adjacent portions. In this work, we single out one sampling issue in an artificial but useful test set and examine how loop refinement accuracy is affected by errors in surrounding side-chains. In 80 high-resolution crystal structures, we first perturbed 6–12 residue loops away from the crystal conformation and placed all protein side chains in non-native but low energy conformations. Even these relatively small perturbations in the surroundings made the loop prediction problem much more challenging. Using a previously published loop prediction method, median backbone (N C α C O) RMSD's for groups of 6, 8, 10, and 12 residue loops are 0.3 / 0.6 / 0.4 / 0.6 Å, respectively, on native structures and increase to 1.1 / 2.2 / 1.5 / 2.3 Å on the perturbed cases. We then augmented our previous loop prediction method to simultaneously optimize the rotamer states of side chains surrounding the loop. Our results show that this augmented loop prediction method can recover the native state in many perturbed structures where the previous method failed; the median RMSD's for the 6, 8, 10, and 12 residue perturbed loops improve to 0.4 / 0.8 / 1.1 / 1.2 Å. Finally, we highlight three

comparative models from blind tests in which our new method predicted loops closer to the native conformation than first modeled using the homolog template, a task generally understood to be difficult. Although many challenges remain in refining full comparative models to high accuracy, this work offers a methodical step toward that goal.

Introduction

Despite the rapid increase in the rate of experimental protein structure determination catalyzed by structural genomics initiatives, the vast majority of known protein sequences will lack experimental structures for the foreseeable future. The ability to generate protein models comparable in accuracy to moderate-to-low resolution experimental structures for these proteins would have enormous utility for structure-based drug design and biological studies. Though the method of comparative (or homology) modeling is a useful tool in this regard, the resulting models vary in accuracy.¹

Vitkup et al² estimated that 90% of all uncharacterized protein sequences within a protein family could be modeled if on average two proteins per family are rationally selected for experimental structure determination. This suggests that, on average, about 100 protein sequences without any prior structural characterization could be modeled for each new experimental structure¹. The accuracy of these models, however, varies significantly. As documented by the New York Structural Genomix Research Consortium, many models accurately represent the overall tertiary structure, but relatively few (<10%, i.e., those with >50% sequence identity) are expected to be as accurate as moderate resolution experimental structures (1–2 Å RMSD)³. The majority of the models will require refinement in order to be useful for problems requiring high-resolution information, such as structure-based drug design.

There are two primary roadblocks to more reliably accurate comparative models, namely difficulties in 1) identifying and aligning to a homolog template and 2) refining regions in the initial model that potentially differ structurally from the target protein. In general,

while much progress has been made in the alignment step, little has been made in the refinement step^{4,5,6}. There are a number of approaches to refining protein models and we will not provide a detailed summary here. There has been some work to investigate the particular reasons for refinement difficulties. Fiser et al⁷ predicted loops in structures with artificially distorted backbone positions in the environment of the loop and found that predictions became worse as expected. They suggested simultaneously optimizing the environment during the loop prediction to improve accuracy but results were not shown. Qian et al⁸ found that reducing the number of degrees of freedom, through sampling along principal components derived from protein family members, avoided generation of low energy, non-native models, and generally improved accuracy beyond the starting template. Mönnigmann and Floudas⁹ performed backbone sampling of residues flanking loops to account for flexibility and variation in the loop stems. Finally, Misura and Baker¹⁰ assessed the accuracy of their refinement methods on a test set of increasingly distorted starting structures by perturbing bond lengths, side chains and secondary structure elements and finally on full de novo models. In general, they were able to refine the perturbed starting structures to lower RMSD models when compared to the native structure. They also found that most of the deviation in their models occurred in loop regions.

Two requirements for successful comparative model refinement are 1) efficient methods for sampling degrees of freedom that enable near-native configurations to be located from the starting structure; and 2) an energy function capable of identifying near-native conformations. Though some progress has been made^{11,12,13}, both of these challenges

remain unsolved in our view. In this work, we take a simplified approach and focus exclusively on the sampling problem, in particular through increased sampling within and around loop regions.

Most loop prediction algorithms have been evaluated primarily by their ability to reproduce the conformations of loops in protein crystal structures. In these tests, all portions of the protein other than the loop in question are generally retained in their native conformation, after adding hydrogen atoms and sometimes performing energy minimization. Numerous methods have been reported that achieve high-accuracy reproduction of loops in such tests^{14,15,16,17,18,19,26}. Accurate reconstruction of loops in crystal structures is an important prerequisite for the more challenging task of refining loops in homology models. The critical difference is that a given loop in a homology model will be surrounded by other portions of the protein which themselves are inaccurate. Refining the loop in this inaccurate environment, without explicitly optimizing the surroundings, frequently fails, indicating that loop refinement in homology models is much more difficult than loop prediction in crystal structures. The inaccuracies in the surroundings can be divided into three categories: 1) errors in the conformations of side chains surrounding the loop, 2) errors in the backbone flanking the loop (the loop “stems”), and 3) errors in non-adjacent portions of the backbone. In this work we consider an artificial but useful intermediate case where we isolate only the first of these types of errors. That is, we have chosen to focus on loop prediction when side chains outside the loop have inaccurate initial conformations but the backbone outside the loop is retained in the native conformation. This makes the loop prediction problem

much more challenging, although still less difficult than loop refinement in homology models. We are not addressing larger refinement problems such as surrounding backbone, helix or domain optimization. In doing so, we hope to de-convolute some of the causes of error and begin to bridge the gap between loop prediction in crystal structures and loop refinement in homology models.

We have developed a new method, Hierarchical Loop Prediction with Surrounding Side chain optimization (HLP-SS), for predicting loops in inexact environments that builds on a previously reported method, which we refer to here as Hierarchical Loop Prediction (HLP). Through the simultaneous optimization of side chains within and in the vicinity of the loop, we have increased the accuracy of our loop predictions relative to our previous protocol when applied to proteins with inaccurate surroundings. We previously applied a similar method to predicting loop conformational changes due to post-translational phosphorylation¹⁸, an application with challenges that are similar to homology model refinement, but the approach has not otherwise been extensively tested. Here we evaluate this protocol using a large and diverse test set of 80 loops, varying in length and difficulty, with artificially perturbed surroundings. We examine specific cases that illustrate successes and failures, and provide some anecdotal but encouraging results suggesting that the approach can be used successfully in blind tests of homology model refinement.

Methods

In previous work¹⁹, HLP, which is implemented in the Protein Local Optimization Program (PLOP), has been tested for its ability to reconstruct protein loops in crystal structures. The sampling algorithm and energy function have recently been improved²⁶ for long loops as highlighted below. In this current work, we augment HLP by enabling the simultaneous sampling and optimization of surrounding side chains. A full description of the previous protocol can be found here^{19,26}. We provide an overview of HLP, and discuss the features of our new method, HLP-SS.

Previous Hierarchical Loop Prediction method: HLP

The previously published method involves a hierarchy of loop prediction stages in which the lowest energy loops generated from one stage are passed to the next where more focused (constrained) sampling is performed.

Specifically, as shown in Figure 1, the initial structure is passed to two, parallel, initial prediction stages (only one is shown) labeled “Init.” The two initial stages vary with the amount of allowed steric overlap between atoms as measured by an overlap factor: 0.7, 0.6, respectively. The overlap factor is defined as the ratio of the distance between two atom centers to the sum of their van der Waals radii. The resulting lowest 5 energy structures from each of the initial stages (10 total loops) are passed as new starting structures to the parallel refinement stages (only one is shown in Figure 1). In the first refinement stage, the C α atoms of the loop are constrained during sampling to less than 4

Å from C α atoms in each starting loop. Finally, the 5 lowest energy loops from all refinement 1 stage processes are passed to 5 parallel refinement stages (only one is shown in Figure 1) labeled “Ref 2.” In this second refinement stage, the C α atoms of the loop are constrained to less than 2 Å from C α atoms in each starting loop. The lowest energy loop from all stages is taken as the predicted loop.

An all-atom force field energy with implicit solvent is calculated for each sampled loop and the loops are then ranked by energy. The energy is calculated using the OPLS all-atom force field^{20,21,22}, the Surface Generalized Born model of polar solvation²³, an estimator for the nonpolar component of the solvation free energy developed by Gallicchio et al.²⁴, and a number of correction terms as detailed in Ghosh et al.²³ and in Jacobson et al.²²

At each stage of the procedure, the sampling is performed by perturbing dihedral angles in the backbone and side chains, using knowledge-based preferences: as described previously¹⁹, backbone dihedral angles are chosen randomly from a 5° resolution library representing the well-known Ramachandran plot, and side chain rotamers are chosen randomly from a 10° resolution library developed by Xiang and Honig²⁵. All heavy-atom torsion angles between the terminal peptide bonds are sampled. All bond lengths and angles associated with these are initially set to default values, but are allowed to vary during energy minimization. Polar hydrogens (e.g., OH group on Ser/Thr/Tyr) are sampled during side chain optimization; non-polar hydrogens are not sampled other than through minimization.

To obtain greater accuracy for long loops (in this work, loops longer than 9 residues), the algorithm has been augmented as described previously²⁶. Sampling has been increased dramatically through the addition of five additional “fixed” stages where sub-segments of the loop are sampled while the remainder of the loop is held fixed. In addition, Zhu et al. have also incorporated an additional hydrophobic term adapted from the ChemScore²⁷ scoring function, which has been successfully used to describe the hydrophobic contribution to the binding free energy between ligands and protein receptors. The "long loop" protocol and scoring function were utilized in this study for the 10 and 12 residue loop cases. We did not use the augmented protocol and energy function on the 6 and 8 residue loop cases for efficiency reasons. Though, these changes have been applied to short loops in a previous study²⁸ which shows moderate improvement in accuracy. Only three and four "fixed" stages were used for the 10 and 12 residue loop cases, respectively, to improve the computational efficiency. Our experience showed this choice was sufficient to achieve convergent results.

New method incorporating surrounding side chains: HLP-SS

In this work, we modified the HLP algorithm presented above in two places: during the loop buildup and during the side chain optimization (Figure 2).

1. Removal of side chains during backbone sampling

For efficient backbone sampling, the previously published loop prediction method applies a variety of screens to rule out high energy loops as early as possible. One of these

screens checks for steric clashes between the loop backbone and the rest of the protein, as the loop is built up from either side. By default, this steric screening checks for clashes with all heavy atoms outside the loop. However, if the portions of the model surrounding the loop are inaccurate, this screen could prevent native-like structures from being sampled, e.g., if a side chain is occupying a portion of the space that the loop backbone should pass through. To avoid this problem, we created an option to ignore side chains surrounding the loop during the steric screening.

However, a significant downside to ignoring the surrounding side chains during backbone sampling is that the conformational search space increases significantly. Also, we may be discarding information because frequently, some initial side chain conformations in the surroundings are approximately correct. For any given loop refinement in a particular model, it is not *a priori* obvious whether including or excluding surrounding side chains is more likely to succeed. For this reason, in our method, we do both (in separate optimizations) and ultimately use the MM-GBSA energy function implemented in PLOP to choose the final predicted conformation. Specifically, we added a third initial stage “Init3” with overlap factor of 0.7 and used our new option to exclude surrounding side chains during the steric screening. We continue to include the surrounding side chains during steric screening in the original two initial stages, “Init1” and “Init2”. Finally, we also make certain to optimize the *same* surrounding side chains across all prediction stages so that we can compare the energies of all sampled loops.

2. Simultaneous optimization of side chains in surroundings and loop

In the HLP method, the energy for each candidate loop conformation is obtained after iteratively optimizing the side chain conformations on the loop followed by energy minimization. In the HLP-SS, we expand the list of side chains to be optimized by including the side chains of the surrounding residues. The self-consistent side chain optimization is accomplished by iteratively placing one side chain at a time while holding the others fixed until no side chain changes rotamer state. In HLP-SS, we optimize the side chains on the loop first and then side chains from surroundings, iteratively.

Data set choice and perturbation

In order to de-convolute the many compounding problems that occur in loop prediction in full comparative models, we chose to predict loops on a data set consisting of crystal structures that are perturbed to contain modeling errors in side chains only. Our current goal is neither to create a representative sampling of all loops found within proteins nor to generate all possible loop refinement scenarios found in comparative models. Rather, our intention is to create a test set with enough variety in difficulty and types of refinement problems, that we (and others) may test approaches to loop prediction in inaccurate environments.

Criteria for test set selection

We constructed a test set of 80 loops, 20 loops each of 6, 8, 10, and 12 residues in length. For each loop length, we chose a smaller number of loops from the larger, previously

published sets^{19,26} due to the computational expense of the additional sampling of the loop surroundings. The 6, 8 and 10 residue loops were taken from Jacobson et al¹⁹ and the 12 residue loops were taken from Zhu et al²⁶. The proteins in the set are diverse in sequence, observed in crystal structures with ≤ 2.0 Å resolution, and have been filtered such that the simulated loops are far from heteroatom groups. This set contains a mix of both difficult and easy loop prediction cases, similar to the previously published larger test sets. The median RMSD predictions for our subset of loops are similar to those from the larger, previously published test sets (when predicted on the native crystal structure with simulated crystal environment).

Generation of perturbed crystal structures

We perturbed crystal structures in the following way. For each of the 80 loops we performed the following.

- 1) Generate a low-energy loop far from the native conformation:** We performed a single run of loop prediction in PLOP which generates a list of sampled loops ranked by MM-GBSA energy. In general, we chose a sampled loop greater than 3 Å backbone heavy atom RMSD from the native loop and then grafted this loop onto the crystal structure. In some cases, no loops were sampled greater than 3 Å from the native and in those cases we simply selected from one of the lower-RMSD, non-native loops.
- 2) Rotamer optimization on the full protein with non-native loop:** We performed rotamer optimization and energy minimization on all side chains in the perturbed protein using the method described in Jacobson et al¹⁹. This procedure removes the memory of native χ angles and bond lengths/angles of all side chains and places the protein in a non-

native, local minimum, creating a more difficult loop prediction scenario that more closely resembles an initial comparative model.

The dataset and the relevant information are listed in tables S1, S2, S3, and S4.

Method of choosing surrounding side chains to optimize

The key new element introduced into the loop modeling procedure is simultaneous optimization of side chains on the loop and in its surroundings. In the extreme, the algorithm could optimize all side chains on the protein, but this would unnecessarily increase computational expense due to sampling many side chains distant from the loop (and also increases “noise” in the computed energy). At the other extreme, only those side chains in contact with the starting loop could be optimized. However, the initial loop may be far from its native position in a homology model, as are many of the perturbed loops in our test set. For this reason we developed a protocol to attempt to identify all side chains that could interact with any conformation of the loop. We accomplish this by first generating a coarse unbiased sampling of loops, <50, using a quick backbone buildup within PLOP. We then identify all residues with a distance cutoff of any of these loop conformations to decide which side chains outside the loop are optimized. Surrounding residues are included that have a side-chain heavy atom within a certain cutoff from C β atoms within an initial set of sampled loops. The C β atoms from the N-terminal and C-terminal loop residues are excluded in this screen. For example, at a distant cutoff of 7.5 Å, this translates to an average of 17 surrounding side chains for the 8-residue cases but the number varies considerably from 9 to 37 depending on the

solvent-exposure of each loop. We tested distance cutoffs of 5.0, 7.5, and 10.0 Å in this work.

Sampling and energy function failure analysis

In cases where the method predicted loops greater than 1.5 Å backbone heavy atom RMSD, we attempted to understand why, distinguishing between two broad classes of problems: insufficient sampling, and inability of the energy function to identify near-native states. Sampling problems were identified if no loops are sampled within 1 Å N-C α -C RMSD from the native. Energy function problems are identified by calculating E_{gap} , the difference in energy between our predicted loop and a native-like loop. We did not calculate the native energy using the conformation found in the crystal structure because the loop found in the crystal structure must relax using the same optimizations as our predicted loops in order for the energies to be comparable. The native energy is taken from the lowest energy loop with <1 Å N-C α -C RMSD from the native. For consistency, this analysis was carried out on both the unperturbed and perturbed crystal structure test cases.

RMSD calculations

Loop RMSD's are calculated using N, C α , C, and O atoms in the loop backbone with the protein aligned, excluding the loop. Side chain RMSD's are calculated using non-hydrogen atoms in the side chain. Full comparative models were first aligned to the

native crystal structures using the MatchMaker function in Chimera²⁹ with default settings.

Crystal packing simulation

We do not include crystal packing in the primary predictions presented here. However, in one case we suspected crystal packing effects might affect the results, and to investigate this possibility, we used the option in PLOP that includes all atoms found in a single asymmetric unit plus all atoms <20 Å from adjacent asymmetric units. Each asymmetric unit is identical at every stage of the calculation.

Protonation states of titratable residues

All titratable residues are placed in their standard protonation state at pH 7.0 (e.g. histidine is neutral), regardless of whether pH is specified in the PDB file. This assumption may affect accuracy in some cases, particularly when we compare to structures that were crystallized at non-physiological pH.

Generation of full comparative model test set

In order to begin to test the applicability of our new method in full comparative models, we refined loops within initial models that were generated by our team in the latest Critical Assessment of Techniques for Protein Structure Prediction (CASP7) experiment.³⁰ As the submitted models highlighted here were refined using HLP-SS, the

refinements were “blind” tests. Targets T326, T345, and T376 are highlighted in this work. Target T326 was aligned to template, 2GHR, using BLAST³¹ and constructed using the Protein Local Optimization Program¹⁹ as previously described³². T345 and T376 were aligned to templates 2F8A and 1YXC, respectively, using HMAP³³ and constructed using NEST³³. A loop was identified in the initial model as “requiring refinement” if the sequence alignment between template and target contained gaps or deletions within regions between secondary structure elements found in the template structure.

Results and discussion

Assessment of our previous method: HLP

A comparison of results applying our previous protocol, HLP, to the unperturbed and perturbed test cases illustrates how incorrect side chain conformations can degrade the performance of loop predictions when surrounding side chains are not included in the optimization (Table I). For all loop lengths, the median backbone RMSD increases by approximately a factor of 4. More specifically, HLP predicts 42 out of 80 test cases greater than 1.5 Å backbone RMSD on the perturbed test set compared to 15 out of 80 when predicted on the unperturbed crystal structures. The “easy” test cases, i.e., the ones that the previous protocol performs relatively well on, serve as controls to verify that our new protocol does not adversely affect these cases.

Using HLP on perturbed structures, we anticipated a decrease in accuracy due to sampling since the perturbed side chains in the surroundings of the loop may block the native conformation. Interestingly, the number of energy function problems also increases which indicates that native-like loop backbones are being sampled using our old protocol but the perturbed surroundings may prevent key side chain contacts from forming.

Assessment of our new method: HLP-SS

In order to test whether our new protocol, HLP-SS, is more effective at predicting loops in inexact environments, we utilized our perturbed test set and compared results using HLP-SS to results using HLP.

We first varied the number of surrounding side chains to include during the loop prediction by testing our new protocol with different cutoff distances (see Methods). The comparison of prediction accuracy to computational expense summarized in Figure 3 indicates that a radius of 7.5 Å is a good tradeoff. Interestingly, increasing the radius to 10 Å does not give a marked increase in accuracy over a radius of 7.5 Å but does increase computational expense considerably. All results in this paper are reported using the 7.5 Å cut-off unless otherwise noted.

The results in Table I show a consistent increase in accuracy when side chains surrounding the loop are optimized during the loop prediction compared to when they are held fixed. Individual predictions can be found in the supplemental Tables S1, S2, S3 and S4. The overall accuracy for each loop length is increased using HLP-SS over HLP. For example, for the 8 residue perturbed loop set, the median backbone RMSD is 0.8 Å using HLP-SS compared to 2.2 Å using HLP. By comparison, on the unperturbed 8 residue loops, the median backbone RMSD is 1.0 Å using HLP-SS and 0.6 Å using HLP. Thus, the results of using HLP-SS on the perturbed loops approaches the accuracy that can be achieved in loop reconstruction, especially for short loops (6 and 8 residues). With longer loops (10 and 12 residues), HLP-SS produces a small increase in median

backbone RMSD (by a factor of ~ 1.3) on perturbed structures versus unperturbed. Including surrounding side chains during the loop optimization clearly produces more accurate results on our perturbed test set than not including them.

The number of sampling problems is reduced using our new method: out of the 80 perturbed test cases, HLP produces 24 loop sampling problems compared to 3 using HLP-SS (see Methods for our definition of sampling and energy problems). HLP-SS produces similar numbers of sampling problems as seen in the control experiments on the unperturbed crystal structures. These results suggest 1) our perturbed test set creates more sampling difficulties than the unperturbed crystal structures, and 2) our enhanced sampling in the HLP-SS method is addressing these difficulties. However, the number of errors that can be attributed to limitations of the energy function is not reduced using HLP-SS. In loop reconstruction in native crystal structures, the number of failures attributed to the energy function increases using HLP-SS versus HLP. That is, the increased sampling due to simultaneous optimization of side chains surrounding the loop places a greater burden on the energy function in distinguishing between native and non-native configurations (i.e., many more non-native side chain contacts are sampled). Thus, the perturbed loop test set increases the difficulty of both sampling native-like configurations and identifying these among many non-native conformations, relative to loop reconstruction in native proteins.

As a control, we tested our new method on native crystal structures to see if our new method degrades accuracy when the loop and its surroundings are initially in the native

state. This control represents the “best that we can expect” using HLP-SS because it will uncover energy and sampling problems unrelated to the altered side chains in the perturbed test set. As expected, median backbone RMSD’s increase slightly using HLP-SS compared to HLP: results using HLP-SS show an increase of +0.1, +0.4, +0.4, and +0.3 Å for 6, 8, 10, and 12 residue loops respectively over HLP (Table I). Sampling surrounding side chain rotamers increases the number of degrees of freedom and thus the likelihood of energy function or sampling problems.

If we consider the set of “easy loops” among the perturbed loop test set, i.e., cases where our old protocol predicts native-like loops (better than 1.5 Å backbone RMSD), our new protocol predicts *non-native* loops (worse than 1.5 Å RMSD) in only 5 of these 39 easy cases.

Interestingly, although the average loop prediction accuracy improves with our new method, the average accuracy of the side chains in the surroundings does not improve (data not shown). Looking at averages over many surrounding residues may be hiding the role of an important few. While in some cases the role of a single surrounding residue is clear, such as in case 1CLC (see below) where a single residue blocks sampling of the native conformation, other cases are more subtle. In cases where sampling is not a problem, key energetic contributing residues, now free to move in our new method, may form incorrect contacts for reasons such as differences in the pH between crystal structure and our modeling conditions, or other problems with our energy function.

Effects of crystal packing

Because our goal is to predict loops within comparative models, where crystal symmetry information is not known, we do not simulate the crystal environment in the primary predictions presented here. However, since we are comparing to crystal structures in this intermediate step, crystal packing effects may contribute to apparent error in our predictions^{19,34}. In order to assess these effects in our test set, we performed predictions using HLP on the unperturbed crystal structures with and without simulation of crystal packing (Table S5). Simulation of crystal packing is described in Methods. Nine cases (PDB's: 1XIF, 3TGL, 1IAB, 1PRN, 1SBP, 1ARB-12 residue case, 1CNV, 1M3S, 1OTH) show potential crystal packing effects which we define as a predicted loop backbone accuracy $>1.5 \text{ \AA}$ RMSD without simulating crystal packing but $<1.5 \text{ \AA}$ RMSD with crystal packing. Loop predictions are affected by either restricting the sampling space or by changing the energy landscape through inter-chain contacts. Since HLP is sampling conformations $<1.2 \text{ \AA}$ RMSD in all nine cases without simulated crystal packing, the increases in accuracy with crystal packing are probably not due to restricting the sampling space but are enabled through inter-chain energetic contacts. See example 3TGL below for an example.

Most importantly, these errors likely propagate through our perturbed test predictions and should be taken into account in assessing the new method. However, removing the above nine cases (identified as having adverse crystal packing effects) from our statistics, we see little increase in overall accuracy for our new method (Table S6). HLP-SS median RMSD's for the 6, 8 and 10 residue perturbed cases stay within 0.1 \AA of the statistics

derived from the full test set, suggesting crystal packing is playing a minor role for the cases. Because four of the nine “crystal packing” cases are in the 12 residue test set, the statistics show moderate decreases in median and average RMSD’s for both our old and new protocol. In the 12 residue perturbed test set, median/average RMSD’s for HLP-SS are reduced from 1.2/1.7 Å to 1.1/1.3 Å in this filtered test set. Statistics for HLP are reduced from 2.3/2.6 Å to 1.6/2.4 Å.

1CLC

The benefits of our approach are clear in the 8 residue test case 1CLC, residues 313–320. In the perturbed loop starting structure, Glu322 protrudes into the space that the native loop would occupy (Figure 4). Without optimizing this side chain during the loop prediction, near-native loops will have high energies due to steric clashes. HLP-SS selects a near native loop with backbone RMSD of 0.4 Å. Without side chain optimization, the lowest energy loop is 4.3 Å RMSD. Sampling is enhanced near the native as seen in Figure 5.

1F46

Another successful prediction, the 12 residue test case 1F46 (residues, 64-75), highlights a more subtle effect of incorrect surroundings on loop refinement. HLP-SS predicts a 1.1 Å backbone RMSD loop while HLP predicts a 3.8 Å backbone RMSD loop (Figure 6). In contrast to 1CLC above, there are no surrounding side chains obstructing the backbone from sampling close to the native. As seen using HLP, backbone conformations are

sampled as low as 0.3 Å RMSD. HLP selects a 3.8 Å RMSD loop because at least one incorrect surrounding residue prevents a key loop side chain from repacking, thus leading to near-native conformations having high energies. Interestingly, HLP-SS with a 5.0 Å cutoff also failed, predicting a loop of 2.3 Å RMSD. By examining which residues are included in the 7.5 Å and 5.0 Å cutoffs, we determined that the repacking of the side chain of Met64 is blocked by nearby Arg161, a residue that is not optimized in the HLP and HLP-SS (5.0 Å cutoff) protocols. Optimizing Arg161 with the larger 7.5 Å cutoff enables repacking of Met64 and contributes to a lower energy, near-native loop.

3TGL

Including surrounding side chains did not improve our prediction of the six residue loop, residues 82–87 in 3TGL, beyond 3.1 Å backbone RMSD. Upon inspection of the original crystal structure, we found significant interactions between the loop and other chains within the asymmetric unit. We thus reran our calculations while including all atoms from crystal symmetry chains within 20 Å of the original chain. Note that the 7.5 Å cut-offs for including nearby side chains was reapplied to capture residues from the symmetry copies of the protein. Our new prediction achieved a 0.5 Å backbone RMSD and correctly forms the salt bridge between Arg86 and Glu47 of the symmetric chain (Figure 7). In a control experiment using HLP on the unperturbed crystal structure, the effect of crystal packing is clear with accuracy increasing from 3.1 Å to 0.7 Å when crystal packing is simulated. The apparent failure of HLP-SS to predict the native loop in the perturbed 3TGL structure is therefore due to crystal packing.

1ALC

HLP-SS failed to identify a native loop in the eight residue loop case, PDB 1ALC. This target was identified as problematic even when predicting the loop within the native crystal structure but the failure is not due to omission of crystal packing. Although we sampled a loop as close as 0.4 Å backbone RMSD (Figure 8, black triangles), we were not able to identify it as a near native loop because its energy is higher than decoy loops with >4.0 Å backbone RMSD from the native. In this case the failure does not appear to be due to limitations of the energy function but rather due to a particularly rugged energy landscape for this loop. Though the loop is on the protein surface, a large percentage of its surface area is buried (Figure 9). To validate this idea, we re-ran our prediction while constraining the sampling within 1 Å of the native loop and successfully identified a near-native loop lower in energy than the previously generated decoys (Figure 8, red circles). In the future, optimizing the surrounding backbone atoms in addition to side chains may improve sampling for loops like this that are tightly constrained by their environment.

Application of our new method to the refinement of full comparative models

To begin testing whether our new method can better refine full comparative models, where backbone as well as side chain atoms are inexact, we compared results from loop predictions on three homology models using our old and new methods. The resulting

models using HLP-SS were submitted to the CASP7 experiment. After the experiment, we re-ran the loop refinements using HLP to compare to our blind HLP-SS results.

The important question is whether each method predicted loops more accurate than the initial homolog-derived loop, a task generally considered to be difficult. Figure 10 and Table II clearly show that ignoring surrounding side chains produced less accurate predictions than the initial starting models: HLP results for targets T345, T326, and T376 were 5.6, 2.2, and 5.6 Å RMSD respectively in comparison to the starting conformations, 1.5, 1.7, and 4.6 Å RMSD. In contrast, HLP-SS not only produced more accurate loops than HLP, it also improved somewhat upon the starting loop in these cases: HLP-SS predicts loops to 1.4, 1.1, and 3.5 Å backbone RMSD for the three targets.

Our intention here is to highlight evidence that initial models with both backbone and side chain errors can also benefit from HLP-SS. We do not make the claim that our protocol will work on all comparative models and we have seen many cases where backbone perturbations outside of the loop are large enough that HLP-SS fails (data not shown). However, the difficulty of homology model refinement is such that any success in a blind test is encouraging. At the very least, these examples highlight how loop prediction methods that do not account for errors in the surroundings (HLP in this work) not only fail to improve homology models but can make the results much worse.

Conclusion/Further directions

Refining comparative models is difficult for two reasons: the energy landscape is rugged and the sampling space is vast. In this study, we aimed to address one important sampling difficulty that occurs in the refinement of protein models, namely the *ab initio* prediction of loop segments when surrounding residue side chain positions are incorrect. By sampling rotamer states of nearby residues simultaneously with our previous all-atom loop sampling strategy, we have shown that a simple solution can significantly improve our predictive ability in these cases.

We chose to test our protocol on perturbed crystal structures. Rationally perturbed, idealized test sets are critical to de-convolute the sources of difficulty facing full comparative modeling refinement. In this study, we show that simply perturbing loops and then scrambling side chains in crystal structures creates a much more difficult loop prediction problem, relative to reconstructing loops in unperturbed crystal structures. We then show that our new method can predict near-native loops in a large majority of these perturbed cases by simultaneously sampling the side chains surrounding the loop. Our 80 perturbed test cases are available for download (see link below).

A logical next step after perturbing side chains within crystal structures is to introduce inaccurate backbone conformations in regions of the protein surrounding the loop in question. Initial results (data not shown) suggest that optimizing backbone atoms in surrounding residues, including the loop “stem” residues, during the side chain

optimization stage can improve predictions in cases where surrounding inaccuracies cannot be corrected through side chain optimization alone.

Increasing the sampled degrees of freedom as we have done in this study implies a need for an energy function that is increasingly more robust at discerning native-like from non-native-like structures. Robust homology model refinement will thus require not only the development of new sampling methods but also more accurate energy functions. Some limitations are addressable whereas others are not. For example, because we are comparing our predictions to experiments, experimental factors that are not generally known at the time of comparative modeling can affect our accuracy. We have shown (Figure 7 and Table S5) that inclusion of the crystal symmetry chains during the loop prediction can increase accuracy. The pH at which the protein was crystallized can also dramatically affect conformations seen in the crystal structure, particularly with side chain positions relevant to this current study. Since we assumed standard protonation states for titratable residues at pH 7.0, we will not account for changes in conformation due to pH.

There are energy function limitations we can improve without knowledge of experimental crystal conditions, however. Though we can only assume a physiological pH at the time of modeling, we should be able to predict local pKa shifts within the protein. We hope to address this issue in the near future. Limitations in using the Generalized Born implicit solvent model can lead to over-stabilized salt bridges³⁵ and

will fail to predict water-mediated interactions. Using a fixed-charge, non-polarizable force field may introduce errors as well.

An assumption implicit in this work is that the loop adopts a single well-defined conformation, and that the correct answer is the single conformation reported in the PDB file. From a computational standpoint, this limitation can be addressed by recasting the algorithms presented in this work as Monte Carlo sampling, i.e., to predict an ensemble of structures rather than a single structure. Work along these lines is underway. Predicted ensembles of loop structures could be compared to experimental temperature factors from crystal structures, or preferably, to structures that were refined using an ensemble approach³⁶.

Acknowledgements

We would like to thank David Pincus for early work and discussion which guided much of this work. We would also like to thank Andy Kuziemko from the Barry Honig lab for providing the initial comparative model T376 for the “applications” section of this study. Finally, we thank the entire Barry Honig and Richard Friesner labs for their support during CASP7. This work was supported by grants from the Sandler Program in the Basic Sciences and from the Sloan Foundation (to MPJ), and by NIH grants GM52018 (to RAF) and GM81710 (to MPJ). The work of BDS was supported in part by the Genentech Scholars Program. MPJ is a member of the Scientific Advisory Board of Schrodinger Inc.

Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081).

Our test set of perturbed crystal structures can be downloaded here:

<http://jacobson.compbio.ucsf.edu>

References

- 1 Baker D and Sali A. Protein Structure Prediction and Structural Genomics. *Science* 2001; 294: 93-96.
- 2 Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. *Nat Struct Biol.* 2001; 8: 559-66.
- 3 <http://www.nysgrc.org/>
- 4 Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology* 2005;15:285–289.
- 5 Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. CASP Introduction: Critical assessment of methods of protein structure prediction (CASP) - Round 6. *Proteins* 2005; 61:3-7.
- 6 Tress M, Ezkurdia I, Graña O, López G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* 2005; 61:27-45.
- 7 Fiser A, Do RK, Sali A. Modeling of loops in protein structures. *Protein Sci* 2000; 9:1753-1773.
- 8 Qian B, Ortiz A, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *PNAS* 2004; 101:15346-15351.
- 9 Mönnigmann M., Floudas CA. Protein loop structure prediction with flexible stem geometries. *Proteins* 2005;61: 748-762.
- 10 Misura K, Baker D. Progress and challenges in high-resolution refinement of protein structure models. *Proteins* 2005; 59: 15-29.
- 11 Misura K, Chivian D, Rohl C, Kim D, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *PNAS* 2006; 103: 5361–5366.
- 12 Chen J, Brooks C. Can molecular dynamics simulations provide high-resolution refinement of protein structure?. *Proteins* 2007; 67:922-930.
- 13 Fan H, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci* 2004; 13: 211-220.
- 14 Moult J, James MNG. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* 1986; 1-2: 146 – 163.
- 15 Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* 1986; 26:137 – 168.
- 16 van Vlijmen HWT, Karplus M. PDB-based Protein Loop Prediction: Parameters for Selection and Methods for Optimization. *J Mol Bio* 1997; 267-4:975-1001.
- 17 Deane CM, Blundell DL. CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Prot Sci* 2001; 10:599.
- 18 Groban E, Narayanan A, Jacobson MP, Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS Comput Biol* 2006;2(4):e32.

-
- 19 Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A Hierarchical Approach to All-Atom Loop Prediction. *Proteins* 2004; 55:351-367.
- 20 Jorgensen WL, Maxwell DS, TiradoRives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* 1996; 118: 11225-11236.
- 21 Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 2001; 105: 517-529.
- 22 Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. *J Phys Chem B* 2002; 105: 11673-11680.
- 23 Ghosh A, Rapp CS, Friesner RA. Generalized born model based on a surface integral formulation. *J Phys Chem B* 1998; 102: 10983-10990.
- 24 Gallicchio E, Zhang LY, Levy RM. The SGB/NP hydration free energy model based on the surface generalized Born solvent reaction field and novel nonpolar hydration free energy estimator. *J Comput Chem* 2002; 23: 517-529.
- 25 Xiang ZX, Honig B. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* 2001;311:421-430.
- 26 Zhu K, Pincus D, Zhao S, Friesner RA. Long loop prediction using the protein local optimization program. *Proteins* 2006; 65:438-452.
- 27 Eldridge M, Murray C, Auton T, Paolini G, Mee R, Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* 1997; 11: 425-445.
- 28 Zhu K, Shirts MR, Friesner RA. 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. *J. Chem. Theory Comput* 2007; 3: 2108-2119.
- 29 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J Comput Chem* 2004; 25: 1605-1612.
- 30 See <http://predictioncenter.org/casp7/>
- 31 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; 215: 403-410.
- 32 Kenyon V, Chorny I, Carvajal W, Holman T, Jacobson MP. Novel Human Lipoxxygenase Inhibitors Discovered Using Virtual Screening with Homology Models. *J Med Chem* 2006; 49:1356 -1363.
- 33 Petrey D, Xiang X, Tang CL, Xie L, Gimpelev M, Mitors T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh IYY, Alexov E, Honig B. Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling. *Proteins* 2003; 53:430-435.
- 34 Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol* 2002 ; 320: 597-608.

-
- 35 Zhou R, Berne BJ. Can a continuum solvent model reproduce the free energy landscape of a β -hairpin folding in water?. PNAS 2002; 99, 12777–12782.
- 36 Levin EJ, Kondrashov DA, Wesenberg GE, Phillips Jr GN. Ensemble Refinement of Protein Crystal Structures: Validation and Application. Structure 2007; 15: 1040-1052

Chapter 1: Table I

		Crystal Structures				Perturbed Crystal Structures			
		6 res.	8 res.	10 res.	12 res.	6 res.	8 res.	10 res.	12 res.
Starting Structures	Median RMSD	0.0	0.0	0.0	0.0	2.9	3.9	4.2	4.8
	Average RMSD	0.0	0.0	0.0	0.0	3.4	4.3	4.9	4.6
HLP	Median RMSD	0.3	0.6	0.4	0.6	1.1	2.2	1.5	2.3
	Average RMSD	0.7	1.2	0.6	1.2	1.7	2.4	1.7	2.6
	Sampling Failures	1	3	0	0	3	9	6	6
	Energy Failures	1	2	1	7	4	4	4	6
HLP-SS	Median RMSD	0.4	1.0	0.8	0.9	0.4	0.8	1.1	1.2
	Average RMSD	0.8	1.4	1.0	1.4	0.8	1.3	1.5	1.7
	Sampling Failures	0	0	0	4	0	1	0	2
	Energy Failures	3	6	4	3	3	3	7	6

Table I caption

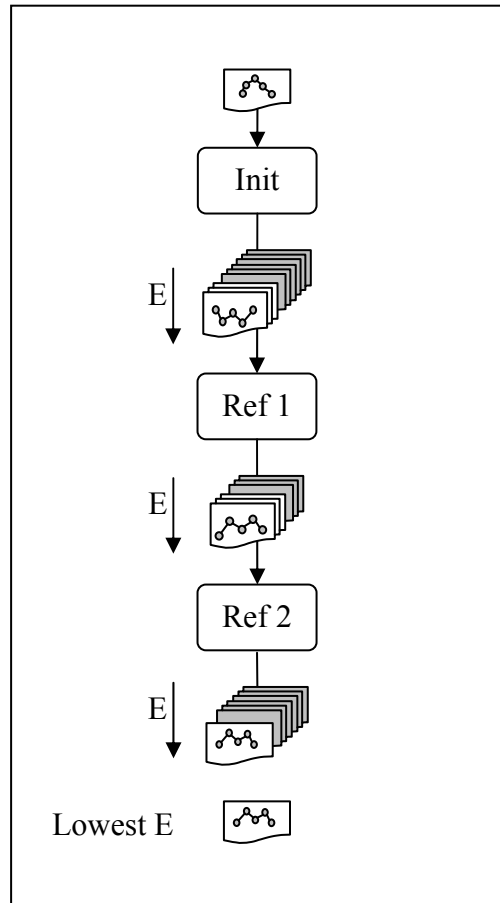
Median and average predicted loop backbone (N C α C O) RMSD's in Å using our previous and new methods on unperturbed and perturbed crystal structure test cases. Statistics are calculated over the 20 test cases in each loop-length category. The numbers of sampling and energy failures are also listed. Rows 1 and 2: the median and average RMSD of the loop before prediction. Rows 3 and 4: median and average RMSD for predictions using HLP (without optimizing surrounding side chains). Rows 5 and 6: the number of sampling and energy failures in each subset using HLP. Rows 7 and 8: median and average RMSD for predictions using HLP-SS (optimizing surrounding side chains). Rows 9 and 10: the number of sampling and energy failures in each subset.

Chapter 1: Table II

Model	Native PDB	Loop start res num	Loop end res num	Starting loop RMSD	Predloop RMSD: HLP	Pred loop RMSD: HLP-SS
T345	2HE3	11	19	1.5	5.6	1.4
T326	2H2W	173	178	1.7	2.2	1.1
T376	2HMC	273	284	4.6	5.6	3.5

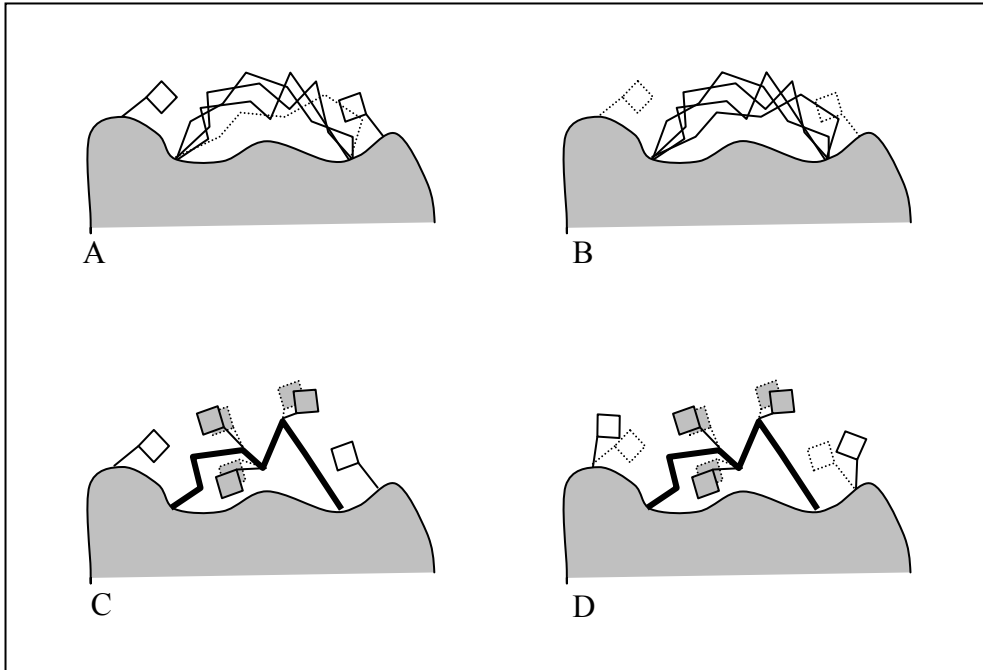
Table II caption

Loop predictions on full comparative models. Column 1: CASP model designation, Column 2: the native PDB, Column 3, 4: the loop endpoints, Column 5: the starting backbone heavy atom (N C α C O) RMSD, Column 6: the predicted backbone RMSD using our old protocol which does not optimize surrounding side chains, Column 7: the predicted backbone RMSD using our new protocol which simultaneously optimized the surrounding side chains. Each model was first globally aligned to the crystal structure using Chimera to calculate RMSD's.



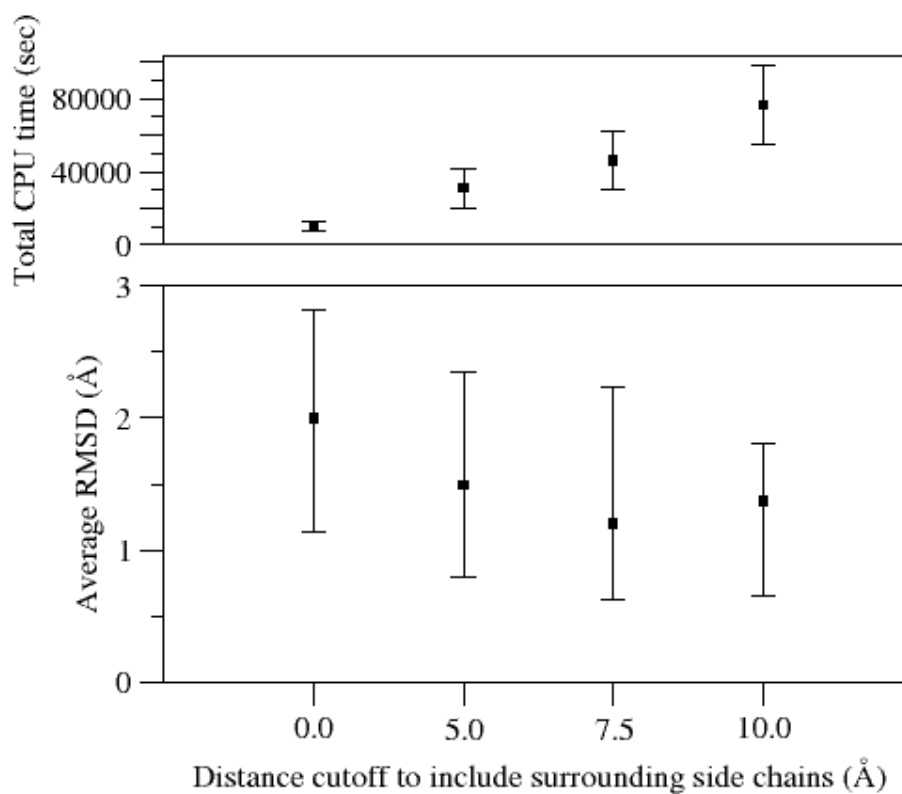
Chapter 1: Figure 1

A high level schematic of the Hierarchical Loop Prediction (HLP) protocol described here and previously. “Init” refers to the initial stage of sampling and scoring. “Ref” refers to the refinement stages where sampling is constrained around starting loop conformations. See Methods for details.



Chapter 1: Figure 2

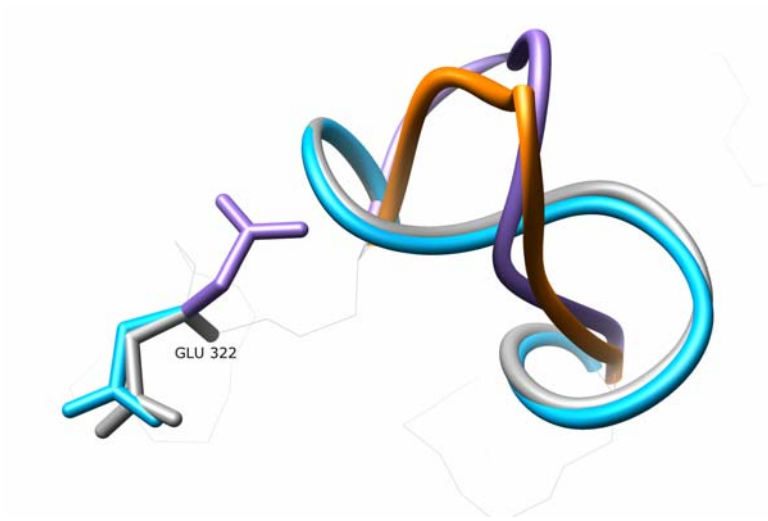
Schematic of the two ways surrounding side chains (white squares) are incorporated into each stage of our hierarchical protocol, backbone sampling (A and B) and simultaneous side chain optimization (C and D). A) Our previous protocol HLP would eliminate loop backbones (dashed line) that overlap with surrounding side chain positions. B) In an initial stage of HLP-SS, we remove the surrounding side chains (dashed squares) during backbone sampling to allow for backbone conformations that might be allowed if the surrounding side chains are given a chance to optimize. C) HLP optimizes side chains on the loop only. D) In HLP-SS the side chains are optimized on the loop as well as the surrounding residues.



Chapter 1: Figure 3

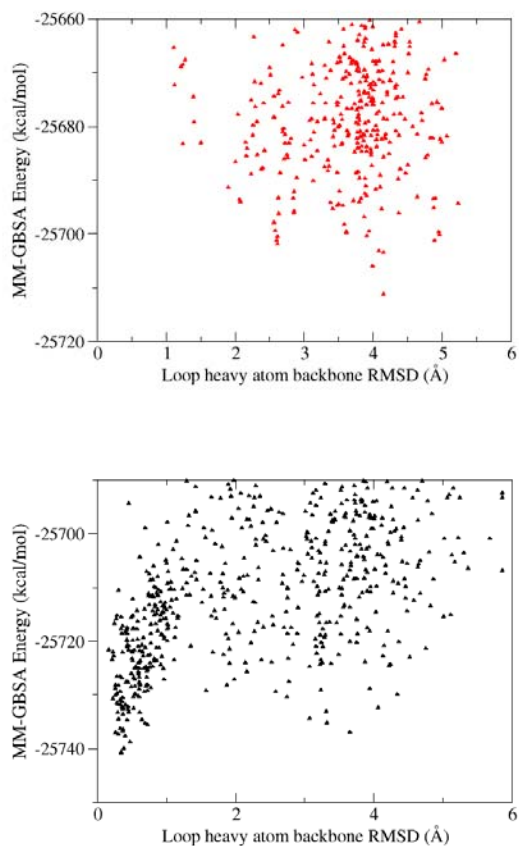
Comparison of total relative CPU time (top) and prediction accuracy (bottom) versus increasing cutoff distance for including surrounding side chains using HLP-SS. The radii

cutoffs of 0.0 (i.e. none), 5.0, 7.5, and 10.0 Å specify which surrounding residues are included in the optimization (see Methods). Top: Average total CPU times in seconds represent a relative cumulative time if our new protocol were not run in parallel. For clarity, average CPU times are shown for the 20 eight-residue test cases only. Error bars represent the standard deviation across each data set. Bottom: Average RMSD's are calculated over all test cases. Positive and negative error bars contain 34.1% of the RMSD population above and 34.1% below the average, respectively.



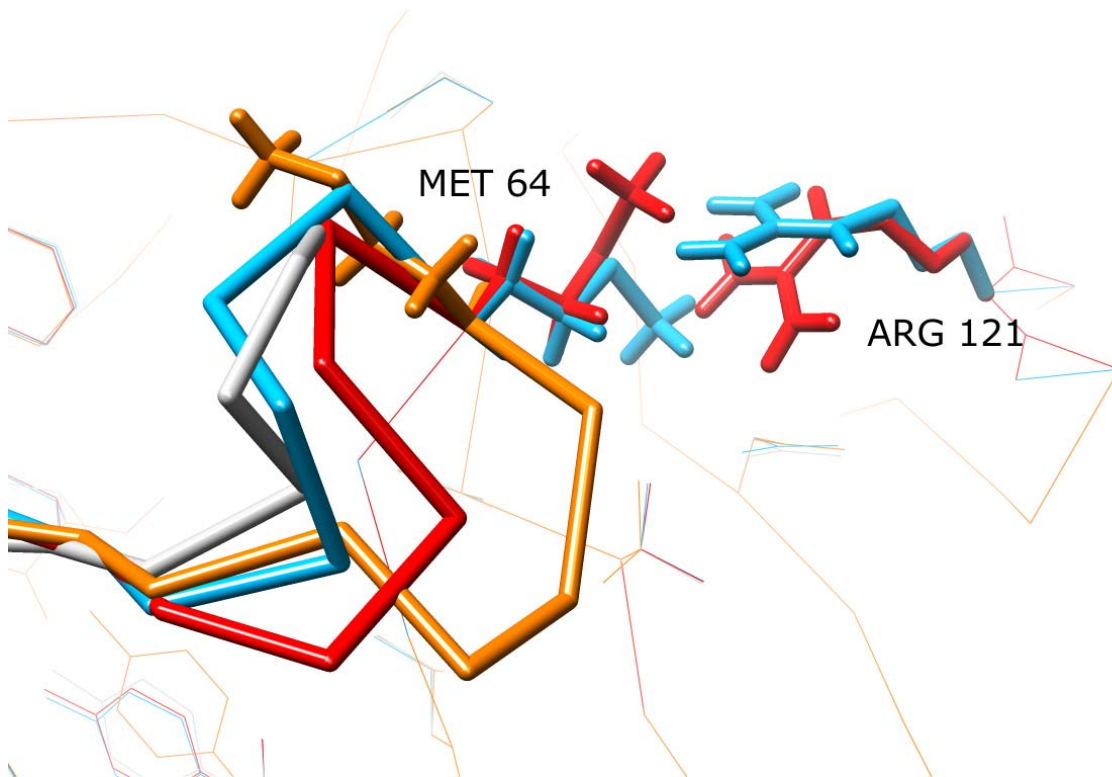
Chapter 1: Figure 4

A successful application of our algorithm in PDB 1CLC, residues 313 to 320. The crystal structure is in gray, the initial perturbed structure is in purple, the predicted loop using our old protocol is in orange and our prediction using surrounding side chain optimization during loop prediction is in light blue. Glu322 partially obstructs the native loop conformation in the initial perturbed starting structure. A near-native loop is only correctly predicted when nearby side chains including Glu322 are also optimized.



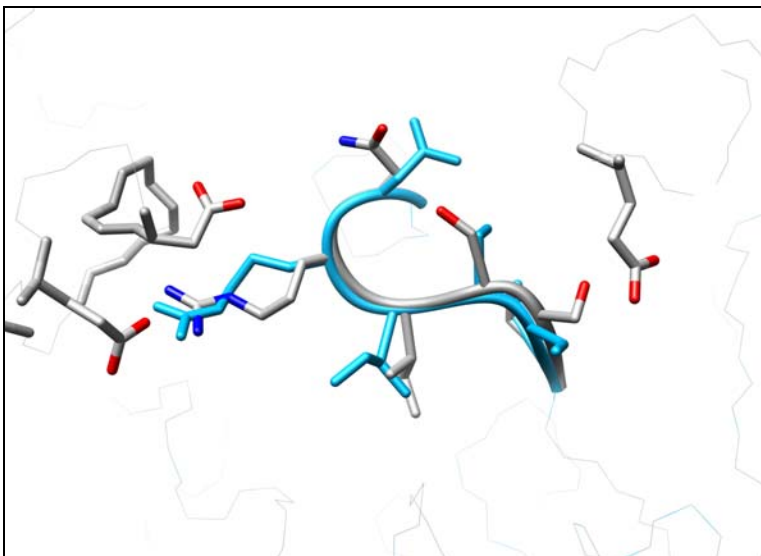
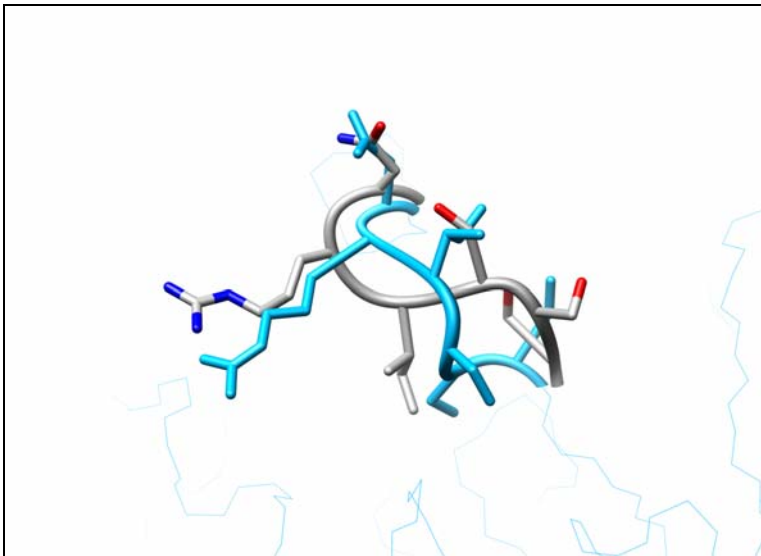
Chapter 1: Figure 5

Calculated MM-GBSA energy versus loop N-C α -C RMSD between predicted and native loop conformations for test case 1CLC. Only samples within the 50 kcal/mol of the lowest predicted energy are shown. Top: predicting loops without nearby side chain optimization (HLP method). Bottom: predicting loops with nearby side chain optimization (HLP-SS method). Note since different atoms are optimized in each example, the relative shapes, and not absolute energies, between the two plots are comparable.



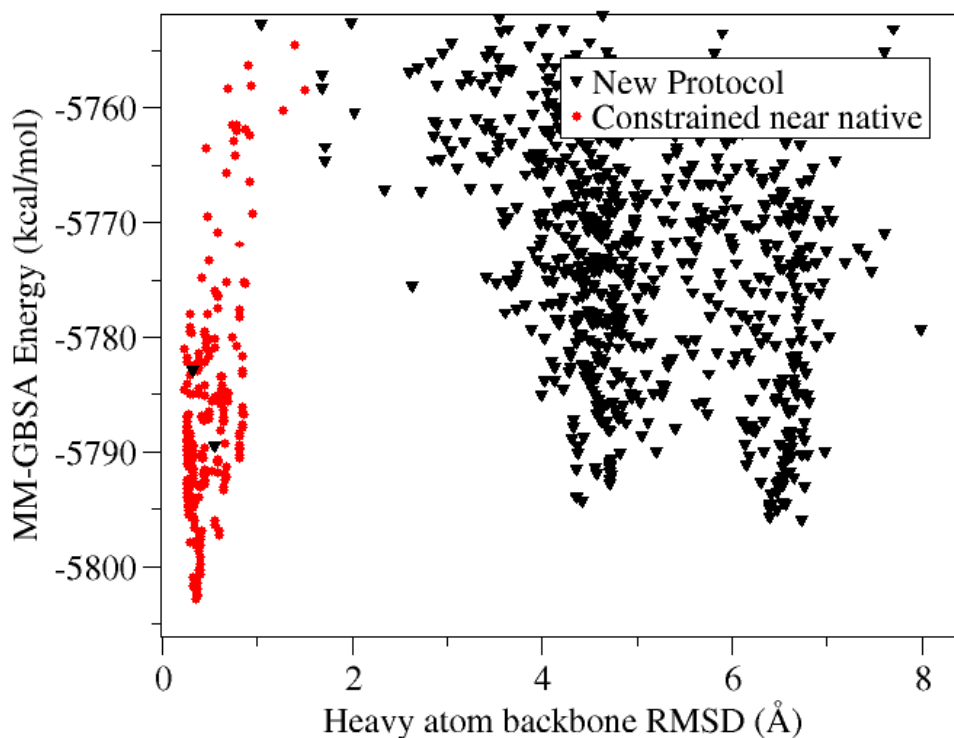
Chapter 1: Figure 6

Comparison of loop predictions on the 12 residue loop test case, PDB 1F46, residues A:64-A:75. The crystal structure is in gray, the predicted loop using our old protocol, HLP, is in orange, our prediction using HLP-SS using a side chain cutoff of 5 Å is in red and our prediction using HLP-SS with a side chain cutoff of 7.5 Å is in light blue. Heavy atom backbone RMSD's are HLP: 3.8 Å, HLP-SS (5.0 Å cutoff): 2.3 Å, HLP-SS (7.5 Å cutoff): 1.1 Å. The Arg121 residue outside the loop is not optimized in either the HLP or HLP-SS (5.0 Å) protocols.



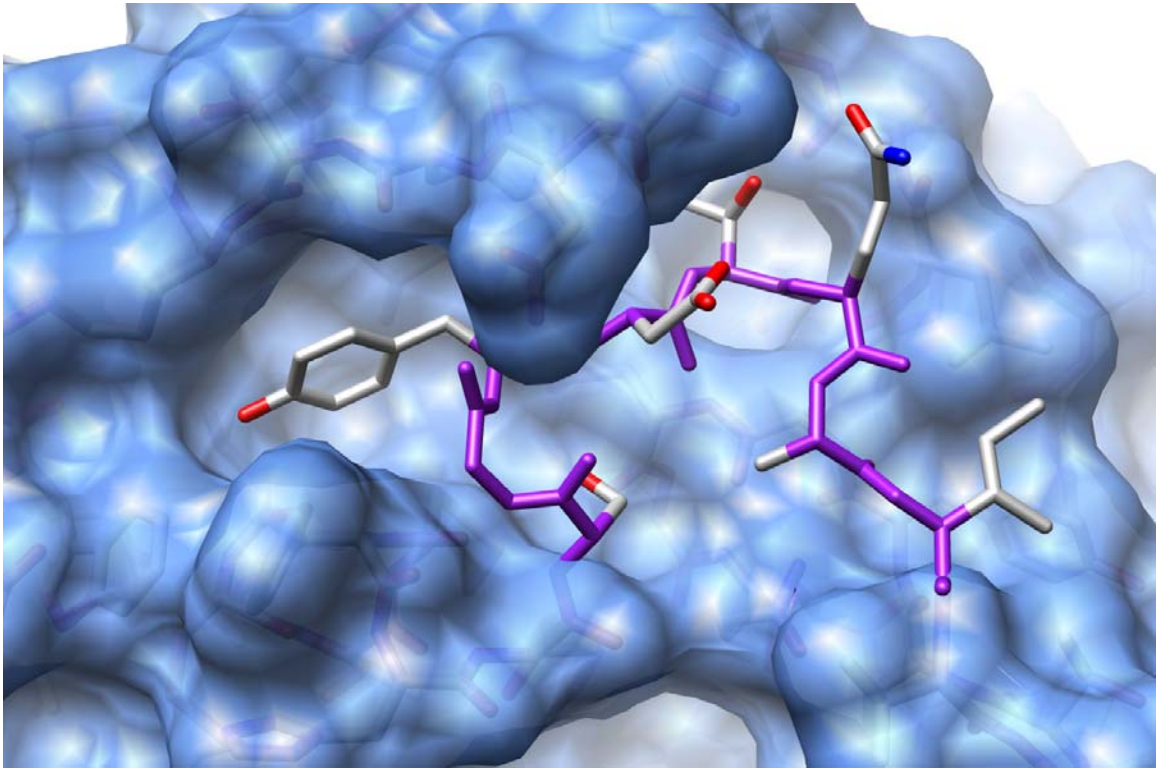
Chapter 1: Figure 7

The loop in 3TGL, residues 82-87, is shown with our predictions in light blue and crystal structure in gray. Top: our new protocol predicts a conformation with a backbone RMSD of 3.1 Å. Bottom: by including the additional atoms from the surrounding chains in the crystal, our new protocol predicts a near-native loop with 0.7 Å backbone RMSD.



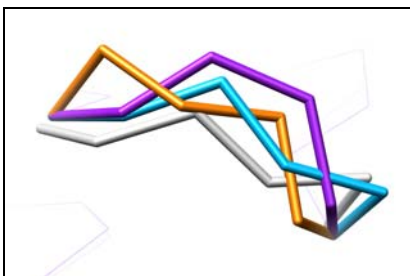
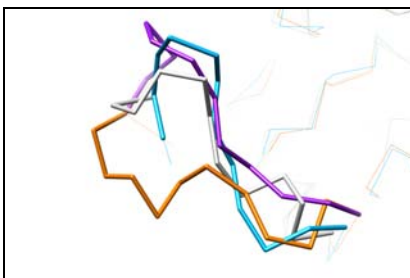
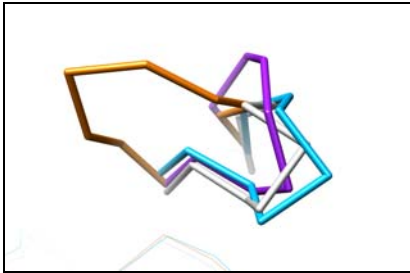
Chapter 1: Figure 8

In black triangles, the MM-GBSA energy is plotted versus N-C α -C RMSD for each sampled loop conformation using our HLP-SS applied to loop residues 34 to 41 in PDB 1ALC. Though one loop conformation with loop backbone RMSD of 0.3 Å is sampled, lower energy decoys exist with backbone RMSD greater than 4.0 Å from the native. In red, energy versus backbone RMSD for sampled loops constrained to less than 2.0 Å from the native.



Chapter 1: Figure 9

Depiction of the crystal structure, PDB 1ALC. The native loop, residues 34 to 41, sits in a tight pocket. All atoms not in the loop are represented as a blue surface map while the loop surface has been removed to show the environment in which we are sampling.



Chapter 1: Figure 10

Comparison of blind loop predictions within full comparative models to native crystal conformations. In all figures, the loops are colored as follows. Gray: native crystal structure, Purple: unrefined loop model, Orange: refined loop model without simultaneous optimization of surrounding side chains, Blue: refined loop model with simultaneous optimization of surrounding side chains. Top: CASP7 target T345, residues 11-19. Middle: CASP7 target T376, residues 273-284. Bottom: CASP7 target T326, residues 173-178.

Supplemental

Tables S1, S2, S3, S4 Description

RMSD values are listed for individual test cases within each loop length, 6, 8, 10, and 12 residues within tables S1, S2, S3, and S4, respectively. RMSD's are backbone heavy atom (N C α C O). Energy gaps (Egap) are in kcal/mol and are the difference between the predicted and native energies (see Methods). An "S" in the energy gap columns means sampling failure where no loops under 1 Å were sampled to provide a native energy. Col 1: PDB ID, Cols 2 and 3: the first and last loop residue numbers, Cols 4, 5 and 6: RMSD of predicted loop, RMSD of best sampled loop, and energy gap using old protocol on unperturbed crystal structures, Cols 7, 8, 9: RMSD of predicted loop, RMSD of best sampled loop, and energy gap using our new protocol on unperturbed crystal structures. Col 10: Backbone RMSD of loops in starting perturbed crystal structure, Cols 11, 12, 13: RMSD of predicted loop, RMSD of best sampled loop, and energy gap using old protocol on perturbed crystal structures, Cols 14, 15, 16: RMSD of predicted loop, RMSD of best sampled loop, and energy gap using new protocol on perturbed crystal structures.

Chapter 1: Table S 1

6res	Native crystal										Perturbed crystal										
	Old Protocol					New Protocol					Old Protocol					New Protocol					
	Loop start res num	Loop end res num	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Start RMSD		Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Start RMSD		Predict RMSD	Best sampled RMSD	Egap (kcal/mol)
lads	:149	:154	0.3	0.2	0.0	0.2	0.2	0.0	6.0		0.8	0.2	0.0	0.2	0.2	0.0	6.0		0.8	0.2	0.0
larp	:282	:287	0.2	0.2	0.0	0.2	0.2	0.0	4.8		0.8	0.6	0.0	0.2	0.2	0.0	4.8		0.8	0.6	0.0
lbrt	:174	:179	0.5	0.2	0.0	0.4	0.2	0.0	6.5		3.8	2.9	0.0	0.4	0.2	S	6.5		3.8	2.9	0.0
lcbs	:66	:71	0.3	0.3	0.0	0.4	0.3	0.0	2.7		2.4	0.5	0.0	0.4	0.5	-0.9	2.7		2.4	0.5	0.0
lgca	:100	:105	0.2	0.2	0.0	0.3	0.3	0.0	2.6		0.3	0.2	0.0	0.3	0.2	0.0	2.6		0.3	0.2	0.0
lhbq	:15	:20	0.4	0.2	0.0	0.4	0.2	0.0	3.6		0.5	0.4	0.0	0.4	0.3	0.0	3.6		0.5	0.4	0.0
lmp	:215	:220	0.8	0.3	0.0	0.3	0.3	0.0	1.9		1.7	0.8	0.0	0.3	0.9	-6.7	1.9		1.7	0.8	-6.9
lnoa	:57	:62	1.1	0.3	-0.6	1.1	0.3	-0.2	2.3		1.1	0.3	-0.2	0.3	0.3	0.0	2.3		1.1	0.3	0.0
lnc	:12	:17	0.3	0.2	0.0	0.9	0.2	0.0	5.0		3.4	0.7	0.0	0.3	0.3	-17.8	5.0		3.4	0.7	0.0
lppn	:144	:149	0.3	0.2	0.0	0.3	0.2	0.0	3.0		0.5	0.5	0.0	0.5	0.2	0.0	3.0		0.5	0.5	0.0
lrie	:126	:131	0.3	0.1	0.0	0.4	0.2	0.0	2.8		1.4	0.6	0.0	0.6	0.2	-5.5	2.8		1.4	0.6	0.0
ltea	:94	:99	0.3	0.3	0.0	2.3	0.3	-6.1	3.2		0.3	0.3	0.0	0.3	0.2	0.0	3.2		0.3	0.3	-6.5
ltys	:66	:71	0.2	0.1	0.0	0.1	0.1	0.0	6.7		6.8	4.1	0.0	0.2	0.2	S	6.7		6.8	4.1	0.0
lxif	:357	:362	3.3	1.3	S	3.5	0.6	-9.2	4.0		1.2	1.3	0.0	1.3	0.2	S	4.0		1.2	1.3	0.0
2cba	:82	:87	0.5	0.5	0.0	0.8	0.5	0.0	2.3		1.1	1.1	0.0	1.1	0.5	0.0	2.3		1.1	1.1	0.0
2cpl	:122	:127	0.3	0.2	0.0	0.3	0.2	0.0	0.8		0.9	0.3	0.0	0.3	0.2	0.0	0.8		0.9	0.3	0.0
2pth	:136	:141	0.3	0.1	0.0	0.4	0.2	0.0	2.8		1.9	1.8	0.0	0.4	0.2	S	2.8		1.9	1.8	0.0
3tgl	:82	:87	3.1	0.3	-7.5	3.1	0.4	-6.9	3.0		2.9	0.5	-7.1	0.5	0.3	-7.1	3.0		2.9	0.5	-3.1
5p21	:104	:109	0.5	0.4	0.0	0.5	0.4	0.0	2.3		1.2	1.2	0.0	1.2	0.4	S	2.3		1.2	1.2	0.0
7rsa	:14	:19	1.0	0.2	0.0	1.0	0.2	0.0	2.4		0.9	0.3	0.0	0.3	0.2	0.0	2.4		0.9	0.3	-0.8
Median			0.3	0.2		0.4	0.2		2.9		1.1	0.5		0.5	0.2		2.9		1.1	0.5	
Avg			0.7	0.3		0.8	0.3		3.4		1.7	0.9		0.9	0.3		3.4		1.7	0.9	

Chapter 1: Table S 2

8res	Loop start res num	Loop end res num	Native crystal						Perturbed crystal					
			Old Protocol			New Protocol			Old Protocol			New Protocol		
			Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Start RMSD	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD
135l	:84	:91	0.4	0.2	0.0	2.6	0.4	-17.4	5.2	2.1	0.9	2.5	0.3	-13.3
1alc	:34	:41	5.9	2.5	S	4.5	1.0	-37.9	5.6	5.2	2.7	6.9	0.4	-6.4
1btl	:50	:57	0.3	0.3	0.0	1.6	0.2	-1.3	7.2	3.1	2.0	0.9	0.5	0.0
1cex	:73	:80	3.9	0.6	-6.8	3.7	0.6	-3.2	4.8	3.2	1.6	2.5	1.5	S
1clc	:313	:320	0.3	0.2	0.0	3.0	0.2	-1.4	4.8	4.3	2.7	0.4	0.2	0.0
1ddt	:127	:134	1.0	0.3	0.0	0.9	0.3	0.0	3.4	0.9	0.4	1.1	0.4	0.0
1ezm	:92	:99	0.4	0.2	0.0	0.5	0.3	0.0	2.7	3.2	1.8	0.5	0.3	0.0
1hfc	:142	:149	0.2	0.2	0.0	0.3	0.2	0.0	7.3	0.4	0.5	0.3	0.2	0.0
1iab	:48	:55	1.6	0.5	-0.3	0.8	0.4	0.0	2.4	2.6	0.7	0.5	0.3	0.0
1ivd	:413	:420	0.9	0.7	0.0	1.1	0.6	-1.9	3.3	3.3	2.3	1.3	0.7	-2.9
1lst	:101	:108	0.6	0.2	0.0	0.7	0.3	0.0	7.0	0.9	0.5	0.7	0.3	0.0
1nar	:192	:199	0.6	0.4	0.0	1.5	0.4	-10.3	2.7	0.7	0.6	1.2	0.6	-4.0
1oyc	:80	:87	0.5	0.3	0.0	0.4	0.2	0.0	2.1	2.2	0.7	0.6	0.3	0.0
1prm	:150	:157	1.7	1.1	S	2.4	0.3	-10.8	4.1	0.7	0.3	2.3	0.4	-9.6
1sbp	:107	:114	3.3	3.9	S	0.3	0.2	0.0	3.8	3.4	1.9	0.3	0.2	0.0
1tml	:187	:194	0.3	0.2	0.0	1.4	0.2	-14.5	2.1	1.2	1.0	1.5	0.4	-19.3
2cmd	:270	:277	0.4	0.2	0.0	0.4	0.2	0.0	8.0	6.4	1.2	0.4	0.3	0.0
2exo	:262	:269	0.4	0.2	0.0	0.4	0.2	0.0	3.5	1.9	1.3	0.7	0.2	0.0
2sga	:32	:43	1.2	1.0	0.0	1.2	0.6	0.0	4.0	1.0	0.9	1.1	0.4	0.0
5p2l	:45	:52	0.4	0.3	0.0	0.9	0.3	0.0	2.4	2.2	0.7	0.8	0.3	0.0
Median			0.6	0.3		1.0	0.3		3.9	2.2	1.0	0.8	0.3	
Avg			1.2	0.7		1.4	0.4		4.3	2.4	1.2	1.3	0.4	

Chapter 1: Table S 3

10res	Loop start res num	Loop end res num	Native crystal						Perturbed crystal						
			Old Protocol			New Protocol			Old Protocol			New Protocol			
			Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Start RMSD	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)
lads	:170	:179	0.8	0.4	0.0	2.3	0.8	-18.8	8.8	0.7	0.4	0.0	0.8	0.6	0.0
larb	:41	:50	0.5	0.4	0.0	1.2	0.8	0.0	5.8	4.1	1.2	S	1.1	0.4	-3.1
laru	:128	:137	0.2	0.2	0.0	0.3	0.2	0.0	1.8	1.8	0.7	0.0	2.0	0.2	-1.3
ldim	:87	:96	0.2	0.2	0.0	1.1	0.3	0.0	4.0	1.1	0.6	0.0	1.0	0.4	0.0
ledg	:269	:278	0.2	0.2	0.0	0.2	0.2	0.0	9.2	3.1	1.6	S	3.0	0.9	-42.9
lgvp	:49	:58	3.3	0.7	-11.7	2.7	0.8	-10.0	5.6	2.5	1.2	S	3.1	1.0	-19.0
lixh	:84	:93	1.0	0.7	0.0	1.0	0.6	0.0	4.3	0.8	0.6	0.0	1.2	0.4	-0.1
llst	:10	:19	0.6	0.2	0.0	0.7	0.3	0.0	4.7	2.0	1.3	S	0.5	0.5	0.0
lmrj	:173	:182	0.6	0.3	0.0	0.7	0.2	0.0	5.1	0.9	0.7	0.0	0.3	0.2	0.0
lpgs	:68	:77	0.4	0.3	0.0	1.8	0.7	-12.2	3.6	2.9	0.5	-1.5	3.1	1.0	-69.6
lplc	:42	:51	1.2	0.5	-3.9	1.1	0.9	-0.7	4.4	1.2	0.6	-1.0	1.2	0.5	0.0
lsec	:65	:74	0.4	0.3	0.0	2.2	0.5	-8.3	3.1	2.8	0.3	-7.3	2.1	0.6	-30.5
ltca	:23	:32	1.0	0.2	0.0	0.7	0.3	0.0	10.6	0.8	0.3	0.0	1.8	0.2	-0.9
2alp	:90	:105	0.2	0.2	0.0	0.4	0.2	0.0	3.7	0.4	0.3	0.0	0.3	0.2	0.0
2ayh	:80	:89	0.2	0.2	0.0	0.2	0.2	0.0	2.5	0.3	0.3	0.0	0.2	0.2	0.0
2cmd	:57	:66	0.4	0.3	0.0	0.7	0.4	0.0	3.7	2.0	1.0	S	1.0	0.4	0.0
2mnr	:91	:100	0.6	0.4	0.0	0.8	0.6	0.0	3.8	1.3	0.5	-9.9	4.6	0.7	-4.9
2sil	:197	:206	0.4	0.3	0.0	0.3	0.3	0.0	3.6	2.2	0.9	-24.9	0.4	0.3	0.0
3fgl	:257	:266	0.3	0.3	0.0	0.8	0.4	0.0	6.3	1.7	1.4	S	0.8	0.5	0.0
7rsa	:33	:42	0.4	0.4	0.0	1.0	0.5	0.0	3.6	1.2	0.4	0.0	1.0	0.5	0.0
Median			0.4	0.3		0.8	0.4		4.2	1.5	0.6		1.1	0.4	
Avg			0.6	0.3		1.0	0.5		4.9	1.7	0.7		1.5	0.5	

Chapter 1: Table S 4

12res	Native crystal										Perturbed crystal					
	Old Protocol					New Protocol					Old Protocol			New Protocol		
	Loop start res num	Loop end res num	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Start RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)	Predict RMSD	Best sampled RMSD	Egap (kcal/mol)
1a8d	:155	:166	1.0	1.1	0.0	2.7	1.4	S	4.0	0.9	0.8	0.0	2.8	1.4	S	
1arb	:182	:193	2.5	0.6	-2.4	1.0	0.5	0.0	4.8	2.2	0.6	-8.4	2.6	0.5	-3.3	
1bhe	:121	:132	0.5	0.4	0.0	0.5	0.5	0.0	4.5	1.0	0.8	0.0	0.7	0.7	0.0	
1bn8	A:298	A:309	1.5	0.4	-3.2	1.3	0.6	-24.7	5.0	8.3	2.4	S	2.6	0.5	-3.9	
1e5e	A:82	A:93	0.4	0.3	0.0	0.4	0.3	0.0	5.1	1.8	0.8	-45.8	1.7	1.0	-51.3	
1eb0	A:33	A:44	0.3	0.2	0.0	0.3	0.3	0.0	5.0	0.4	0.4	0.0	0.3	0.3	0.0	
1env	:188	:199	2.2	0.6	-14.1	1.5	0.7	-9.9	4.5	3.2	0.9	-48.5	3.3	1.0	-45.0	
1es6	A:145	A:156	0.6	0.4	0.0	1.6	0.4	-7.7	5.1	3.4	1.4	S	3.5	1.4	S	
1dqz	A:209	A:220	0.3	0.3	0.0	0.7	0.3	0.0	4.8	1.2	0.4	-3.8	0.6	0.3	0.0	
1exm	A:291	A:302	0.7	0.4	0.0	4.5	1.2	S	4.6	4.0	1.2	S	0.5	0.5	0.0	
1f46	A:64	A:75	0.3	0.2	0.0	0.5	0.4	0.0	5.1	3.8	0.3	-15.4	1.1	1.2	0.0	
1i7p	A:63	:74	0.3	0.2	0.0	0.3	0.2	0.0	5.0	0.5	0.5	0.0	0.3	0.2	0.0	
1m3s	A:68	A:79	5.0	0.7	-34.5	5.1	1.4	S	5.0	5.3	1.3	S	5.6	0.6	-49.5	
1ms9	A:529	A:540	1.9	0.6	-16.8	2.8	0.5	-31.7	2.9	2.9	0.5	-23.7	2.5	0.5	-31.5	
1my7	A:254	A:265	0.5	0.4	0.0	0.9	0.6	0.0	5.1	1.4	0.8	-8.2	0.9	0.6	0.0	
1oth	A:69	A:80	1.8	0.4	-0.8	0.5	0.5	0.0	4.1	2.3	2.1	S	0.7	0.7	0.0	
1oyc	:203	:214	0.5	0.4	0.0	0.5	0.3	0.0	5.2	2.8	1.2	S	1.2	0.4	-7.5	
1qlw	A:31	A:42	1.9	0.3	-22.3	2.0	1.6	S	4.1	1.5	0.8	-10.1	1.4	1.1	S	
1tld	A:127	A:138	0.5	0.4	0.0	0.8	0.5	0.0	3.3	0.6	0.5	0.0	1.0	0.6	0.0	
2pia	:30	:41	0.6	0.4	0.0	0.5	0.4	0.0	4.6	3.7	1.0	-7.6	0.5	0.4	0.0	
Median			0.6	0.4		0.9	0.5		4.8	2.3	0.8		1.2	0.6		
Avg			1.2	0.4		1.4	0.6		4.6	2.6	0.9		1.7	0.7		

Table S5 description

Comparison of loop predictions using previously published protocol (without simultaneous optimization of surrounding side chains) with and without inclusion of crystal symmetry chains on unperturbed crystal structures. Atoms from chains within the crystal are included if they are within 20 Å of any primary-chain atom. RMSD values are calculated for backbone heavy atoms for each prediction. Differences between these two predictions enable us to identify crystal packing effects on our predictions.

Chapter 1: Table S 5

With Crystal Symmetry?	6 residues		8 residues		10 residues		12 residues		Y		N	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
lads	0.3	0.3	135l	3.6	0.4	lads	0.8	0.8	1a8d	1.1	1.0	1.0
larp	0.2	0.2	1alc	5.6	5.9	1arb	1.2	0.5	1arb	1.2	2.5	2.5
lbrt	0.5	0.5	1bt1	0.4	0.3	1aru	0.3	0.2	1bhe	0.9	0.5	0.5
lcbs	0.3	0.3	1cex	3.8	3.9	1dim	0.9	0.2	1bn8	1.5	1.5	1.5
lgca	0.3	0.2	1clc	0.3	0.3	1edg	0.2	0.2	1c5e	0.4	0.4	0.4
lhbq	0.3	0.4	1ddt	1.0	1.0	1gvp	1.9	3.3	1cb0	0.2	0.3	0.3
lmrp	0.3	0.8	1ezm	0.4	0.4	1ixh	1.0	1.0	1cnv	0.7	2.2	2.2
lnoa	2.0	1.1	1hfc	0.2	0.2	1lst	0.6	0.6	1cs6	0.9	0.6	0.6
lonc	0.7	0.3	1iab	0.4	1.6	1mrj	0.2	0.6	1dqz	0.3	0.3	0.3
lppn	0.2	0.3	1ivd	1.0	0.9	1pgs	0.6	0.4	1exm	1.0	0.7	0.7
lrie	0.3	0.3	1lst	0.6	0.6	1plc	0.6	1.2	1f46	0.5	0.3	0.3
ltca	0.4	0.3	1nar	0.5	0.6	1scs	0.3	0.4	1i7p	0.3	0.3	0.3
ltys	0.1	0.2	1oyc	0.5	0.5	1tca	0.5	1.0	1m3s	0.5	5.0	5.0
lxif	0.2	3.3	1prn	0.3	1.7	2alp	0.2	0.2	1ms9	1.9	1.9	1.9
2cba	1.0	0.5	1sbp	0.4	3.3	2ayh	0.2	0.2	1my7	0.5	0.5	0.5
2cpl	0.3	0.3	1tml	0.3	0.3	2cmd	0.4	0.4	1oth	0.5	1.8	1.8
2pth	0.5	0.3	2cmd	0.2	0.4	2mnr	0.4	0.6	1oyc	0.4	0.5	0.5
3tgl	0.7	3.1	2exo	0.2	0.4	2sil	0.5	0.4	1qlw	1.7	1.9	1.9
5p21	0.4	0.5	2sga	1.0	1.2	3tgl	0.5	0.3	1tld	1.1	0.5	0.5
7rsa	0.8	1.0	5p21	0.2	0.4	7rsa	0.2	0.4	2pia	0.4	0.6	0.6
Median	0.3	0.3	Median	0.4	0.6	Median	0.5	0.4	Median	0.6	0.6	0.6
Avg	0.5	0.7	Avg	1.1	1.2	Avg	0.6	0.6	Avg	0.8	1.2	1.2

Table S6 Description

Overall median and average predicted backbone RMSD's are listed for the unperturbed (top) and perturbed (bottom) *filtered* test sets for both the old HLP and new HLP-SS methods. The filtered statistics are across all test cases but with the 9 cases removed that were determined to be adversely affected by crystal packing (see Results).

Chapter 1: Table S 6

Loop Length	Crystal structures			
	Previous Method		New Method	
	Median RMSD	Avg RMSD	Median RMSD	Avg RMSD
6 Residues	0.3	0.4	0.4	0.6
8 Residues	0.4	1.0	1.1	1.5
10 Residues	0.4	0.6	0.8	1.0
12 Residues	0.5	0.7	0.7	1.3

Loop Length	Perturbed crystal structures			
	Previous Method		New Method	
	Median RMSD	Avg RMSD	Median RMSD	Avg RMSD
6 Residues	1.1	1.7	0.4	0.7
8 Residues	2.2	2.5	0.9	1.4
10 Residues	1.5	1.7	1.1	1.5
12 Residues	1.6	2.4	1.1	1.3

Chapter 2: Antibodies as a model system for comparative model refinement

Benjamin D. Sellers¹, Jerome P. Nilmeier¹, Matthew P. Jacobson²

¹Graduate Group in Biophysics, University of California, San Francisco, California

²Department of Pharmaceutical Chemistry, University of California, San Francisco, California

Abstract

Refining comparative models of proteins to accuracies similar to moderate-resolution (<2.5 Å) crystal structures is difficult and no published method has yet provided a general solution to improve the accuracy beyond the template protein. While previous attempts have validated protocols on full comparative models, our approach is to develop new methods using simple model systems, where causes of sampling or energy function failures are easier to dissect. We recently showed in one such system, that incorrectly modeled side chains in residues surrounding loops can drastically decrease loop prediction accuracy. We then showed that optimization of rotamers in surrounding residues during the loop prediction reliably increases accuracy. As a next step, the aims in the present study are to investigate the effects of incorrectly modeled backbone atoms in residues surrounding loops and to develop more accurate methods.

We have chosen the H3 loop in comparative models of antibody variable fragments (Fv) as a model system not only because of their tremendous biological and therapeutic relevance but also because predicting Fv structure is largely reduced to predicting a single loop, the hyper-variable H3 loop. The remaining five complementarity determining region (CDR) loops can often be modeled (with some inherent backbone error) using knowledge-based rules. H3 is thus surrounded by minor (<2Å RMSD) backbone errors in the surrounding loops. From a non-redundant library of 49 high-resolution X-ray crystal structures, we constructed a test set of 14 Fv models with 5-8 residue H3 loops, by grafting canonical CDR loop-templates from chains with <60% sequence identity from the target onto the native protein framework. By using the native

crystal structure for the conserved Fv framework, effects of heavy and light domain orientation are ignored in this study. The H3 loop was then placed in a non-optimized starting conformation as typical of homology modeling. Variation in the backbone atoms outside the H3 loop are solely due to modeling errors in the nearby CDR loops.

In crystal structures, results show we predict H3 loops to an average backbone RMSD of 1.3 Å when crystal packing and antigens are removed (0.5 Å with these elements are included.) However, failure to refine the surroundings of the loop in comparative models of the same antibodies decreases accuracy to 3.2 Å. Using our previously published protocol that refines rotamers in surrounding residues, accuracy improves to 1.8Å. Finally, we have developed a new method that additionally optimizes *backbone* atoms of residues outside H3, iteratively. Average accuracy improves to 1.4 Å which is very close to the accuracy achieved in crystal structures without crystal packing or antigens. These results suggest that our hierarchical protocol of iteratively refining surrounding backbone and side-chain atoms around loops in conjunction with a Physics-based force field can often correctly identify native-like states and offers another step toward accurate refinement of loops in error-prone comparative models.

Introduction

Reliably accurate models of proteins would be useful to biological and therapeutic studies that investigate protein function at the atomic level. Though tens of thousands of experimental protein structures exist¹³, millions of protein sequences, many with unknown function, have been generated¹⁴. As recognized by the field of Structural

Genomics¹⁵, demand for accurate computational models of proteins will be high for the foreseeable future.

To address this broad gap between the numbers of sequences and structures, comparative (or homology) models have been utilized as surrogates for moderate-resolution ($\leq 2.5\text{\AA}$) experimental structures in a variety of successful biological studies. Examples include inhibitor discovery¹⁶⁻¹⁹, enzymatic function prediction²⁰, and protein-protein docking²¹. While anecdotal successes exist, in general, comparative models are not as useful as crystal structures, particularly in applications requiring atomic-level accuracy. For example, McGovern and Shoichet⁷ compared docking results (as measured by database enrichment factor) with crystal structures and homology models and found in general, homology modeled receptors produced worse results (less enrichment.) A general method for producing high-accuracy comparative models would extend the usefulness of such models. In our view, a general method has yet to be developed.

The most recent Critical Assessment of Techniques for Protein Structure Prediction (CASP7)⁹ shows little progress in the development of high-accuracy modeling methods. For the template-based (comparative-modeling) category, an important metric for success is whether predicted protein structures are more accurate than the starting homolog template protein, that is, whether the model can be *refined* closer to the native structure. In the most recent CASP, though some submitted models were closer to the native structure than the best template, no single method improved upon the optimal template on average¹⁰.

There are a number of reasons why a comparative model will deviate from the X-ray crystal structure of the same protein. Between the target and template sequences, the modeler may choose a non-optimal sequence alignment or there may be gaps and/or insertions. With a perfect sequence alignment, parts of the target structure may simply be different from the template or the protein may be flexible. Crystal packing and the pH of the template crystal environment can create artifacts in the experimental structure, which can either be functionally relevant or not, depending on whether the crystal symmetry represents a biological symmetry or the pH is close to physiological pH. Finally, even an ideal template structure can produce an inaccurate comparative model if the modeling tools do not provide adequate sampling or an accurate energy function.

Refining protein models is in part limited by the accuracy of methods which refine loop regions within these models. Since protein fold is generally conserved between proteins >30% sequence identity²², most residues requiring refinement reside in loop regions. But predicting loops in models (as opposed to within crystal structures) is difficult. Loops are also inherently flexible and can be found in various conformations in different crystal environments or with different binding partners²³. For example, crystal packing in X-Ray structures has been shown to affect loop conformations^{24,25}.

Refining loops in comparative models is more difficult than predicting loops within crystal structures because of inherent modeling errors in residues surrounding the loop. We showed previously that a loop refinement method that does not optimize the

surrounding residues often fails²⁶. It fails because residues outside the loop may be modeled incorrectly, preventing algorithms from sampling and accurately scoring a near native loop conformation. Our approach is to first define three types of errors that can surround loops: 1. the side chains of surrounding residues may be incorrect 2. the backbone atoms of surrounding residues may be incorrect or 3. the loop stems, residues adjacent to the loop, may be incorrect.

In previous work²⁶, we addressed item (1) above by developing a new method that refines rotamers of surrounding residues while simultaneously predicting the loop. The method holds the backbone of surrounding residues fixed. On average, in perturbed crystal structures, in which the side chains in the entire protein had been scrambled, we obtained similar accuracies as when predicting the same loops in unperturbed crystal structures. We also demonstrated that the method could improve accuracies in full comparative models but the problem was not generally solved. Errors in the backbone of surrounding and adjacent residues were preventing our method from accurately predicting near native loops.

In this work, we aimed to augment our loop refinement method to account for errors in the backbone in residues around loops. We first sought a test set that would exhibit relatively small ($\sim < 2\text{\AA}$) errors in the backbone of residues surrounding the loop to be predicted. We chose to predict the H3 loop in comparative models of antibody variable fragments (Fv). We chose this protein family as our test system for several reasons: 1. antibodies are biological and therapeutic important 2. there are large numbers of antibody

structures for testing 3. five of the six CDR loop structures can often be predicted with the well known sequence-structure statistical rules²⁷⁻³⁰, leaving the sixth H3 CDR to be predicted *ab initio* by our method and 4. the sequence variability of H3 loops and the surrounding residues is high providing a diverse test set. One downside of using antibodies is that the H3 loop is routinely directly involved in antigen binding. We therefore discuss the effects of antigens (i.e. induced fit) on H3 loop prediction accuracy.

Methods

Non-redundant library of antibodies

Our first goal was to create a non-redundant set of all possible, high-resolution antibody crystal structures in the PDB. We searched the Protein Databank¹³ for antibody Fv and Fab structures. Single chain (scFv), homo-dimer and single-domain (e.g. camelid) antibodies were removed leaving an initial set of 459 structures.

In narrowing the test set further, we had competing requirements. We aimed to have a high-resolution test set that 1. maximizes the test set size in order to calculate meaningful statistics and 2. minimizes sequence redundancy. The Protein Databank contains many structures of closely related proteins. In particular the contained antibody sequences are highly redundant, due to the vast number of mutational studies, bound-unbound antigen studies and due to the genetics of antibodies. Several attempts at clustering antibody Fv sequences produced very small test sets (data not shown). Eventually, we clustered the

antibodies using Pisces^{31,32} with a cutoff of 80% sequence identity, 2.2Å resolution, and 0.3 R-value. The method clusters each chain sequence separately with at most one chain from each PDB selected. The result was a non-redundant list containing both heavy and light chains, each from a separate PDB. Since we wanted complete antibody variable fragments, for a given light chain in the non-redundant list, the associated heavy chain was included and vice versa. Since many of these bound chains were first excluded during clustering, this step introduces some redundancy but enabled us to have a larger library of 49 PDB structures.

We constructed multiple alignments of the heavy and light chain sequences from this test set using CLUSTAL³³. The sequences were then cut to Fv length that we define as eight residues after the C-terminal end of CDR3, which corresponds roughly to the end of the final beta strand in the Fv.

Force-field parameterization of antigens

To assess the effects of antigens on our loop prediction ability, we included antigens that are present in the crystal structure in control calculations (see Results). Small-molecule antigens were parameterized for the 2005 Optimized Potential for Liquid Simulations (OPLS) force-field³⁴⁻³⁶ using the *hetgrp_ffgen* program (Schrödinger, Inc.).

Antibody Fv loop prediction test set

The 14 antibodies in our non-redundant antibody library that have H3 loops of length 5 to 8 residues were used as the tests cases in this study.

Construction of antibody Fv models

Antibody Fv models were constructed as follows (see Figure 1):

1. individual models were created for each target chain sequence, one for the heavy chain and one for the light chain, using the target (i.e. native) crystal structure as the template for the non-CDR residues (here referred to as the scaffold) and using multiple loop-templates for the non-H3 CDR loops as described below. The H3 loop is left truncated initially. The multiple-template homology modeling feature in PLOP is then used to create heavy and light chain models.

2. The heavy and light chain models are then combined into a single model. Since the models are constructed using the native chains as the scaffolds, the domains are in the native orientation.

3. All residue side chains with steric clashes are optimized using the side chain optimization function in PLOP^{24,36}. Steric clashes are defined as any residue containing atoms with overlap factor <0.7 with any other atom and were calculated using the “evaluate steric” command in PLOP. Overlap factor is defined as the distance between two atom centers divided by the sum of the two atomic radii. The side-chain prediction is

executed with a minimum overlap factor of 0.3 and with 4 iterations to enhance sampling. The result is a complete Fv model, with native-like scaffold residues, template-modeled canonical loops for the non-H3 loops and an initial, non-optimized H3 loop. Further details of the canonical loop modeling follow.

4. The missing H3 loop is constructed *ab initio* and placed in a non-optimized starting conformation using PLOP's loop prediction function²⁴. Minimal sampling was used to mimic the gap-closure method used in the homology modeling function in PLOP as previously described³⁷.

Library of canonical CDR loops

We constructed a library of canonical loops from which templates are chosen for the non-H3 CDR's as follows. Each CDR loop, excluding H3, in the non-redundant antibody library was classified by canonical class as defined in Martin et al³⁰. If a loop did not match at least 75% of the residues in any sequence rule, the loop was labeled "non-canonical." By being lenient on matching, we were able to assign most of the CDR loops to canonical classes, though this introduces more backbone error in the CDR loops.

Though previous studies have identified H3 structural classes³⁸⁻⁴², we ignore them in this study to test our *ab initio* loop prediction method. A physics-based loop prediction method that does not rely on knowledge-based methods would be a more general solution for H3 loops and applicable to other protein families.

Modeling of canonical CDR loops

When possible, the H1, H2, L1, L2, L3 loops in our models were constructed using the well-known canonical loop rules, a set of mappings between key positions in the Fv chain sequence and a set of loop backbone coordinates. The CDR loops were first identified in the target sequence and assigned to a canonical class. We then chose the best loop-template from the library given the following criteria: the sequence-identity between the template *chain* and the target chain has the highest sequence identity of all possible class templates while <60% sequence-identity. The 60% cutoff was used to reduce bias towards the native CDR conformation.

Each CDR in the target structure is removed and replaced using the loop-template as follows. The loop-template structure is first structurally aligned to the scaffold-template by aligning the C α coordinates from: A. the three residues preceding the N-terminus and following the C-terminus of the H3 loop and B. the three residues centered on each Cysteines involved in the conserved disulfide bridge. Initial results (data not shown) showed that aligning to either all the residues in the conserved framework or just the stem residues produced poor structural alignment of the loop end-points.

Modeling of non-canonical CDR loops

If a non-H3 CDR of length N is not assigned to a canonical class, then it is modeled from a loop template with the same length N with the highest chain sequence identity <60%. If no loop template exists in the non-redundant library that meet these criteria, then loops with N-1 residues are checked. If no loop templates are found, then loops of length N+1

are searched. This is repeated until a template is found. If a loop of different size is used, then the loop is modeled with an appropriate number of gaps or insertions in the sequence using the homology modeling command in PLOP³⁷. For example, if a target L1 loop of length 10 cannot be modeled by a canonical loop-template and no other 10-residue loops exist with chains <60% sequence identity, then it may be modeled by a 9 residue L1 loop and the extra residue will be modeled as part of the homology modeling process in PLOP.

Prediction of H3 loop

Initial models with non-optimized H3 loops were used to test various loop prediction methods.

Hierarchical Loop Prediction – HLP

The HLP protocol (as named here and previously²⁶) was first described in Jacobson et al²⁴.

HLP with Surrounding Side-chain optimization - HLP-SS

The HLP-SS protocol was described previously²⁶ as a new method for predicting loops when the surrounding side chains are incorrect. In summary, the method is an augmentation of the HLP method in that 1. backbone sampling is increased through an additional “initial” stage in which surrounding side chains are excluded from the backbone screening process and 2. in all stages, the surrounding side chain rotamers are optimized simultaneously with the loop side chain rotamers.

HLP with Surrounding Side-chain and Backbone optimization - HLP-SSB

The new method described here augments HLP-SS by further increasing the sampling in the initial stages and by optimizing the backbone atoms of the non-H3 CDR loop residues before each refinement stage (see Figure 2). Three more initial stages were added with reduced overlap factors. The overlap factor is defined as the ratio of the distance between two atom centers to the sum of the atomic radii. Overlap factors <1.0 indicate some overlap between atoms. As described previously in detail²⁴, to remove samples with large steric clashes, we use a minimum-allowed overlap factor in deciding which loop backbones in the backbone buildup stage and side chains in the side-chain optimization stage are to be screened out. In this work, the overlap factor of the first refinement stage is reduced as well. The five lowest energy loops in each initial stage are used as starting points in the refinement stages, so the three, additional initial stages lead to 15 additional round-one refinement stages. As in HLP-SS, the third and sixth initial stages in HLP-SSB do not screen out loop backbone samples that clash with the surrounding side-chains in order to allow for native-like samples that may be initially occluded by incorrectly modeled side chains (using the “sidefrz=no” option in PLOP.)

In addition to increased sampling, a novel addition to this method is the minimization of all CDR (except H3) backbone and side-chain atoms using the Protein Local Optimization Program⁴³. This is performed before each refinement stage to enable producing an iterative protocol: minimize around loop, then predict loop, then minimize surroundings, then predict loop.

Results

Assessment of test set

Attributes of our test set are shown in Table 1. Eight of the fourteen Fv test cases contain antigens, five with small molecules and three with protein antigens. Loops range from five to eight residues in length. The average sequence identity of the Fv's (i.e. combined across heavy and light chains) between all test cases is 54% with minimum and maximum identities of 40% and 78% (see supplemental Table S1).

Predicting H3 in crystal structures

We first set out to establish baseline results by predicting H3 loops in crystal structures. Past experience suggests these results should be much more accurate than in comparative models of the same antibodies. Using our previously published method²⁴, HLP (see Methods), which does not optimize the surroundings of the loop, we predict an average and median backbone RMSD of 0.5 Å and 0.4 Å respectively. We included crystal packing from adjacent chains in the crystal and included any antigen that may be present (see Table 1). The accuracy of these results are in agreement with previous results generated across a much larger loop prediction test set²⁴.

To assess the effects of crystal packing, we predicted H3 loops using HLP without including crystal symmetry molecules while including antigens. Results do not change, with average and median backbone RMSD of 0.5 Å and 0.4 Å respectively, suggesting that crystal packing is not playing a role in determining H3 structure in this test set. This

however, may be more indicative of the large number of antibodies in complex with antigens within our test set. Antigens that interact with H3 may prevent H3 from making crystal contacts.

To assess the effects of antigens on H3 loop conformation, we reran our calculations using HLP on the apo form of all antibodies. Average and median backbone RMSD's increase to 0.9 Å and 0.6 Å, respectively, indicating that antigens are playing a role in H3 loop conformation. We have shown previously that crystal packing and antigen effects may be less extreme if all atoms outside the H3 loop are held fixed. By extending the degrees of freedom to the residues surrounding H3, these atoms may relax when other molecules from the unit cell are removed, resulting in different predictions. To quantify this, we ran further predictions on the apo antibody structures using HLP-SS which additionally samples side-chain rotamers outside the H3 loop (see Methods). Accuracy further decreased to an average and median backbone RMSD of 1.3 Å and 0.8 Å, respectively.

Predicting H3 in homology models of antibody Fv's

We first predicted H3 loops using our previous method, HLP, in homology models of antibody variable fragments to assess the accuracy of a method that does not refine the residues outside of the loop. In Table 1, the average and median RMSD's are 3.2 Å and 3.9 Å respectively (see Figures 3 and 4.) The median RMSD is much worse than the 3.2 Å median RMSD of the starting loops. Thus, failing to incorporate the surroundings of the loop in an error-prone environment such as a comparative model, leads to erroneous

results. In one case, PDB 1UB6, HLP failed to generate any loops because all loops were screened out. This can occur when the residues around the loop create a tight conformational space and particularly with HLP which screens more loops than the other methods with its large minimum overlap factors (see Methods.)

We then predicted H3 loops using our previous method, HLP-SS, in homology models of Fv's to assess the accuracy of a method that refines the side-chains, but not the backbone, of residues outside of the loop. Since the backbone of the surrounding residues contains only small structural errors, we were interested if simply including surrounding side chains would be enough to predict native-like H3 loops. The results show that accuracy using HLP-SS increases compared to results with HLP. Average and median backbone RMSD's are 1.8 Å and 1.2 Å, respectively (see Figures 3 and 4.) Six of the fourteen cases show significant improvement. All five and six-residue loops are predicted < 1.5 Å RMSD. We can hypothesize that shorter loops are less affected by errors in the environment because 1. fewer energetic contacts are required and/or 2. a reduced number of possible backbone conformations leads to fewer chances of the native being blocked by surrounding residues.

Results using our new method show an increase in accuracy to 1.4 Å and 1.1 Å average and median RMSD, respectively (see Figures 3 and 4.) . Through increased sampling and iteratively refining all the non-H3 CDR's before each refinement stage, more native-like loops are sampled and refined gradually throughout the hierarchical protocol. The energy landscapes change dramatically (Figure 3) using this new protocol. More low-

energy minima are explored and the method is able to iteratively refine into these lower energy basins. Figure 5 shows much of the surrounding residue rotamer states are predicted to near-native states. Examples of rotamers that fail to pack correctly around H3 are large, non-specific residues Phe and Trp.

Discussion

In summary, our results in crystal structures are in agreement with previous published experimental studies that suggest H3 loops are flexible depending on the chemical environment of H3. This is particularly important in comparative models, where crystal packing and antigens are not present. Moreover, predicting H3 loops in crystal structures without any molecules found in the crystal establishes a new baseline to compare to when we analyze results from H3 prediction in comparative models.

By design, the test set of comparative models we present here are not optimal starting models. For example, some of the starting loop conformations are far from the native which in turn causes the surrounding residues to “collapse” around the incorrect starting loop. This may create a more difficult refinement problem than we would want if our goal were to simply predict antibody structures to the highest accuracy. In this work, our aim was not to develop an antibody modeling tool. Rather, we aimed to develop methods that can generally recover from a variety of modeling errors in the surroundings H3.

In preliminary results for this work, we found that the orientations of the variable heavy (VH) and light (VL) chains can affect our H3 predictions (data not shown.) Because of

this, choosing an appropriate scaffold template that has a similar VL-VH orientation as the target's domain orientation will be imperative when modeling antibodies that are truly novel. We are currently analyzing the determinants of VL-VH domain orientations (not published.)

The remaining type of modeling error that can affect loop predictions is structural variation found in the stem residues, those just before and following the termini of the predicted loop (error type #3 mentioned in the Introduction.) In the present study, we used the native crystal structure for the conserved scaffold in order to focus on the surrounding residues alone. In the future, sampling multiple stem conformations may benefit H3 loop prediction as it has been shown in other proteins in a previous study⁴⁴.

Extending predictions to longer H3 loops will be required for many antibody applications. As there have been advances in predicting longer loop lengths within crystal structures using the Protein Local Optimization Program²⁵, we believe predicting long loops in comparative models is possible. In the future, we aim to combine lessons in long loop prediction with lessons from the present work. However, the additional sampling of the surroundings will increase computational requirements sharply in long loops and therefore, keeping such a protocol efficient will be challenging.

We believe our method will be beneficial to modeling complete antibody structures, particularly in scenarios where a similar template antibody exists such as in antibody humanization efforts. Our Physics-based method does not use any previously

published^{38-42,45,46} statistics-based information for H3 conformations, making our method more valuable to synthetic antibody engineering efforts that may utilize H3 sequences that are not found in natural antibodies.

Acknowledgements

Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081).

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28(1):235-242.
2. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2007;35(Database issue):D193-197.
3. Sali A. 100,000 protein structures for the biologist. *Nat Struct Biol* 1998;5(12):1029-1032.
4. Honma T, Hayashi K, Aoyama T, Hashimoto N, Machida T, Fukasawa K, Iwama T, Ikeura C, Ikuta M, Suzuki-Takahashi I, Iwasawa Y, Hayama T, Nishimura S, Morishima H. Structure-based generation of a new class of potent Cdk4 inhibitors: new de novo design strategy and library design. *J Med Chem* 2001;44(26):4615-4627.
5. Schapira M, Raaka BM, Samuels HH, Abagyan R. In silico discovery of novel retinoic acid receptor agonist structures. *BMC Struct Biol* 2001;1:1.
6. Enyedy IJ, Ling Y, Nacro K, Tomita Y, Wu X, Cao Y, Guo R, Li B, Zhu X, Huang Y, Long YQ, Roller PP, Yang D, Wang S. Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *J Med Chem* 2001;44(25):4313-4324.
7. Enyedy IJ, Lee SL, Kuo AH, Dickson RB, Lin CY, Wang S. Structure-based approach for the discovery of bis-benzamidines as novel inhibitors of matriptase. *J Med Chem* 2001;44(9):1349-1355.
8. Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP, Gerlt JA. Prediction and assignment of function for a divergent N-succinyl amino acid racemase. *Nat Chem Biol* 2007;3(8):486-491.
9. Gray JJ, Moughon SE, Kortemme T, Schueler-Furman O, Misura KM, Morozov AV, Baker D. Protein-protein docking predictions for the CAPRI experiment. *Proteins* 2003;52(1):118-122.
10. McGovern SL, Shoichet BK. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem* 2003;46(14):2895-2907.
11. Moutl J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction-Round VII. *Proteins* 2007;69 Suppl 8:3-9.
12. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* 2007;69 Suppl 8:38-56.
13. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986;5(4):823-826.
14. Wedemayer GJ, Patten PA, Wang LH, Schultz PG, Stevens RC. Structural insights into the evolution of an antibody combining site. *Science* 1997;276(5319):1665-1669.

15. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. *Proteins* 2004;55(2):351-367.
16. Zhu K, Pincus DL, Zhao S, Friesner RA. Long loop prediction using the protein local optimization program. *Proteins* 2006;65(2):438-452.
17. Benjamin D, Sellers KZSZRAFMPJ. Toward better refinement of comparative models: Predicting loops in inexact environments. *Proteins: Structure, Function, and Bioinformatics* 2008;9999(9999):NA.
18. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* 1987;196(4):901-917.
19. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* 1997;273(4):927-948.
20. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, et al. Conformations of immunoglobulin hypervariable regions. *Nature* 1989;342(6252):877-883.
21. Martin AC, Thornton JM. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol* 1996;263(5):800-815.
22. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19(12):1589-1591.
23. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* 2005;33(Web Server issue):W94-98.
24. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22(22):4673-4680.
25. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Q ReV Biophys* 1993;26:49.
26. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 2001;105(28):6474-6487.
27. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. *J Phys Chem B* 2002;106(44):11673-11680.
28. Kenyon V, Chorny I, Carvajal WJ, Holman TR, Jacobson MP. Novel human lipoxygenase inhibitors discovered using virtual screening with homology models. *J Med Chem* 2006;49(4):1356-1363.
29. Shirai H, Kidera A, Nakamura H. Structural classification of CDR-H3 in antibodies. *FEBS Letters* 1996;399(1-2):1-8.
30. Oliva B, Bates PA, Querol E, Avilés FX, Sternberg MJE. Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *Journal of Molecular Biology* 1998;279(5):1193-1210.
31. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Antibody structure, prediction and redesign. *Biophysical Chemistry* 1997;68(1-3):9-16.

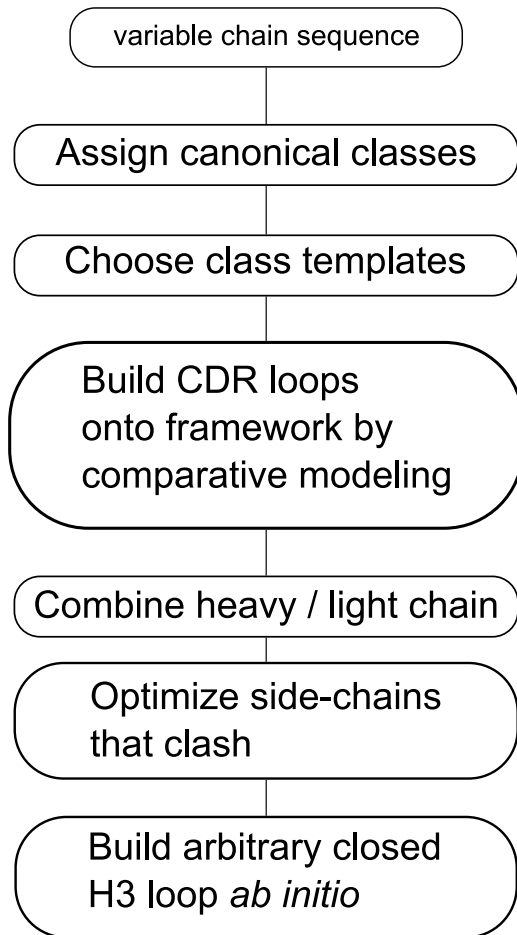
32. Koliashnikov OV, Kiral MO, Grigorenko VG, Egorov AM. Antibody cdr h3 modeling rules: extension for the case of absence of arg h94 and asp h101. JOURNAL OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY 2006;4(2):415.
33. Shirai H, Kidera A, Nakamura H. H3-rules: identification of CDR-H3 structures in antibodies. FEBS Letters 1999;455(1-2):188-197.
34. Zhu K, Shirts MR, Friesner RA, Jacobson MP. Multiscale Optimization of a Truncated Newton Minimization Algorithm and Application to Proteins and Protein-Ligand Complexes. J Chem Theory Comput 2007;3(2):640-648.
35. Monnigmann M, Floudas CA. The Role of Flexible Stem Geometries in Protein Loop Structure Prediction.
36. Daisuke Kuroda HSMKHN. Structural classification of CDR-H3 revisited: A lesson in antibody modeling. Proteins: Structure, Function, and Bioinformatics 2008;9999(9999):NA.
37. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. Journal of Molecular Biology 1998;275(2):269-294.

Chapter 2: Table I

PDBID	Antigen	H3 length	Crystal Structures (controls)				Comparative Models					
			HLP	HLP	HLP	HLP-SS	Starting RMSD	HLP	HLP-SS	HLP-SSB		
1cr9	-	5	0.8	0.3	0.3	0.5	3.6	4.2	0.6	0.6	-	-
1mex	S	5	0.2	0.2	0.6	0.7	2.0	0.4	0.3	0.9	-	-
1ngz	-	5	0.3	0.3	0.3	0.3	2.1	1.1	1.2	1.3	-	-
1uac	P	5	0.2	0.2	0.3	0.4	0.9	0.6	0.7	0.7	-	-
1ub6	-	6	0.3	0.4	0.4	0.5	5.6	-	1.2	1.1	-	-
1flr	S	7	0.3	0.3	0.4	1.0	1.0	1.5	0.9	1.4	-	-
1kcv	-	7	0.7	0.9	0.9	0.9	5.1	2.4	0.5	0.5	-	-
1mju	-	7	0.4	0.5	0.4	0.4	2.5	4.4	0.8	1.2	-	-
1yqv	P	7	0.4	0.4	0.7	2.5	1.9	4.9	1.7	1.5	-	-
2cgr	S	7	1.6	0.5	0.9	1.9	2.8	3.9	3.4	3.4	-	-
1a7q	-	8	0.7	0.6	0.6	0.4	7.0	4.7	4.3	0.5	-	-
1uj3	P	8	0.2	0.2	0.9	2.0	6.6	4.4	1.9	1.0	-	-
1uz8	S	8	1.1	1.1	4.5	4.8	5.7	3.1	3.1	1.7	-	-
2pcp	S	8	0.4	0.6	1.4	1.4	9.2	6.1	4.4	4.3	-	-
Avg			0.5	0.5	0.9	1.3	4.0	3.2	1.8	1.4		
Median			0.4	0.4	0.6	0.8	3.2	3.9	1.2	1.1		

Table I. H3 loop prediction results in crystal structures and comparative models of 14 antibody Fv using various protocols and crystal environments. Crystal packing: '+' indicates other chains in crystal are included (see Methods). Antigen: '+' indicates the binding partner, if it exists in the crystal structure (see Col 2), is included. Col 1: PDB code Col 2: 'P' indicates the antibody has a protein antigen present in the crystal structure. 'S' indicates small molecule antigen. '-' indicates no antigen present in crystal structure. Col 4-7: backbone RMSD's for H3 predictions in crystal structures using HLP. Col 8-11: backbone RMSD's in comparative models of antibodies. Col 8: RMSD of starting H3 loop. Col. 9: RMSD using HLP. Col 10: RMSD using HLP-SS. Col 11: RMSD using HLP-SSB.

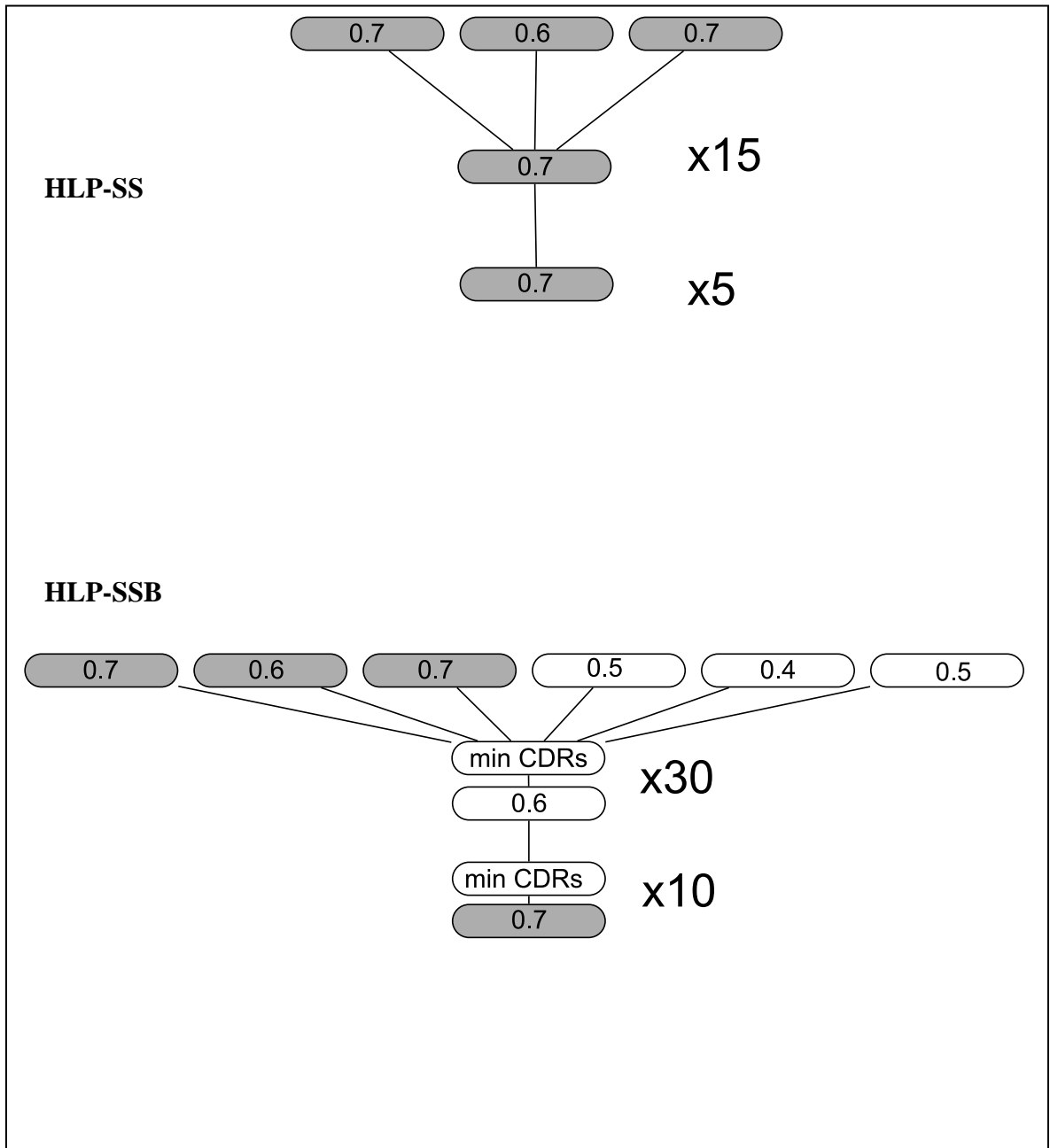
Figure 1



Chapter 2: Figure 1

Flow chart depicting process of creating the initial antibody Fv homology model.

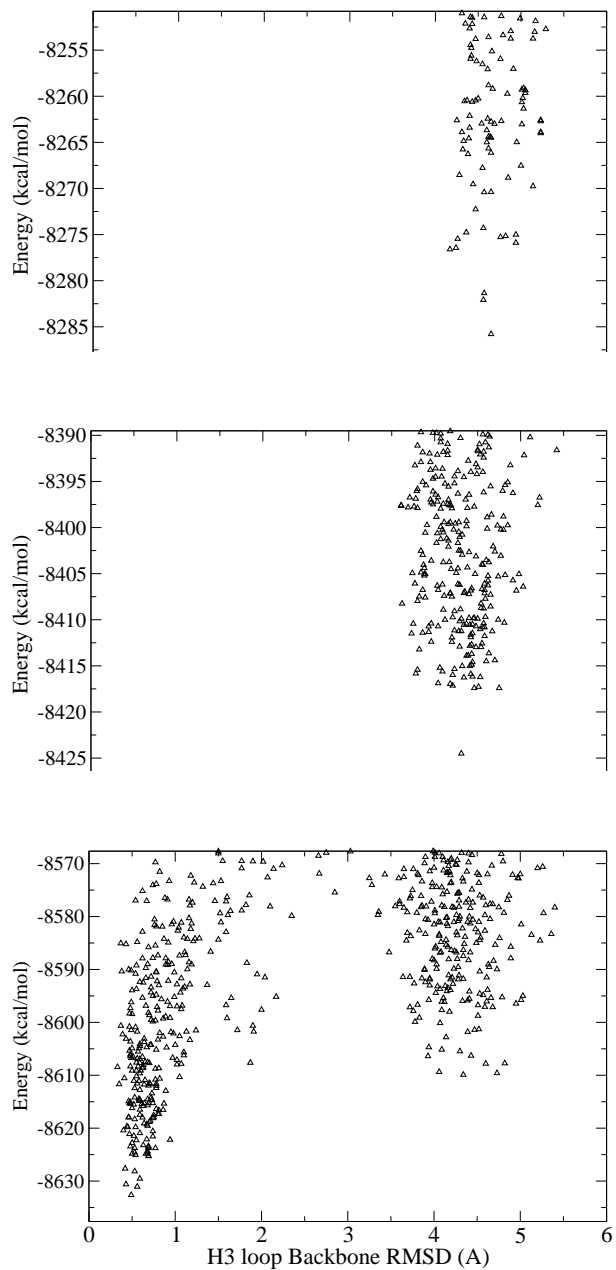
Figure 2



Chapter 2: Figure 2

Differences between the previously published HLP-SS (top) and the new method HLP-SSB (bottom). Each oval depicts a single execution of the loop prediction command in the Protein Local Optimization Program. The numbers in each oval are the overlap factor (see Methods). Three more initial stages with reduced overlap factors were added in addition to two non-H3 CDR-minimization stages before each refinement stage. Items in white are new or changed elements of the new protocol.

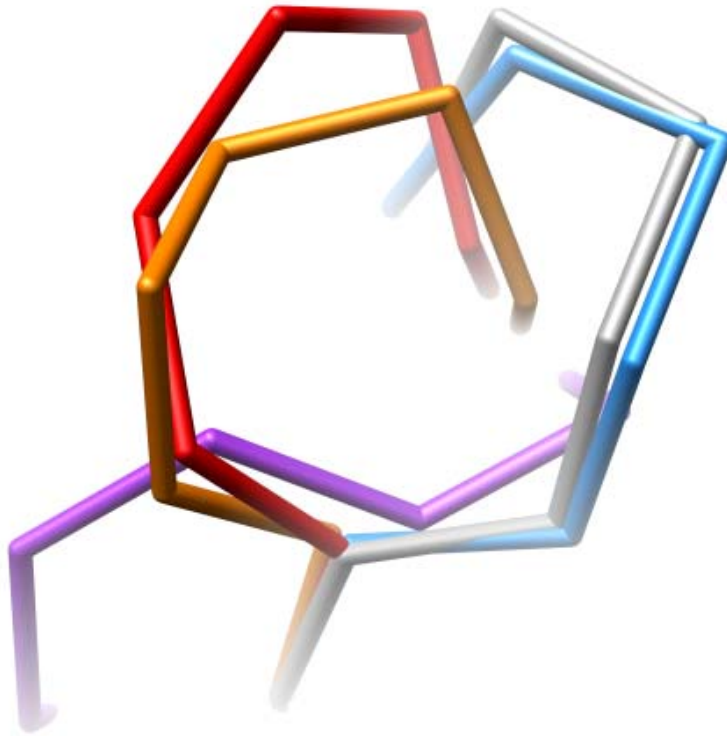
Figure 3



Chapter 2: Figure 3

Energy versus RMSD for all samples < 35 kcal/mol from the lowest predicted energy using three protocols for 8 residue loop PDB 1A7Q. Top: HLP, Middle: HLP-SS, Bottom: HLP-SSB.

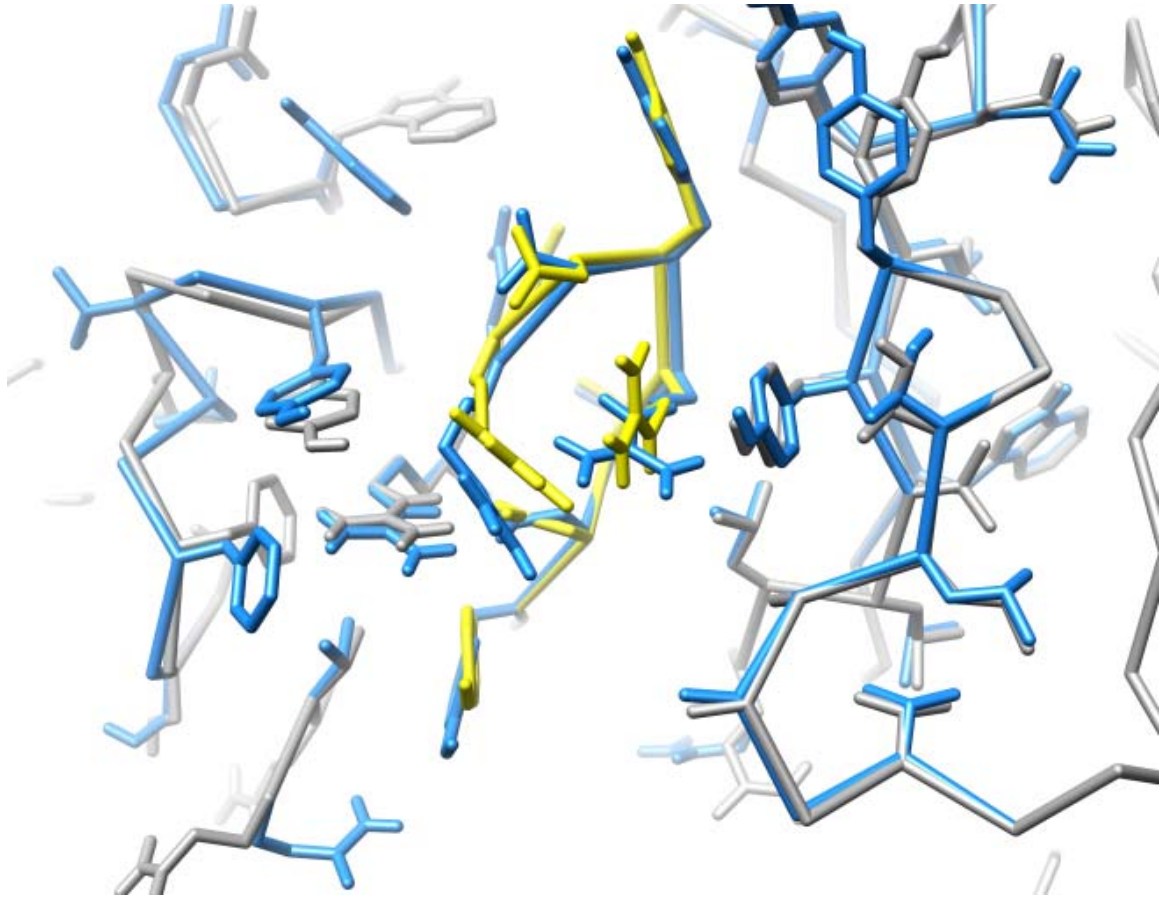
Figure 4



Chapter 2: Figure 4

Closeup of final H3 loop predictions in the same comparative model of antibody PDB 1A7Q. Gray: Crystal structure, Purple: starting H3 loop with 7.0 Å RMSD, Orange: HLP predicts to 4.7 Å RMSD, Red: HLP-SS predicts to 4.3 Å, Blue: our new protocol HLP-SSB predicts to 0.5 Å RMSD.

Figure 5



Chapter 2: Figure 5

Global view of HLPSS-B prediction (blue) in 1A7Q with crystal structure (gray and yellow H3 loop.)

Supplemental

Chapter 2: Table S 1

	1A7Q	1CR9	1FLR	1KCV	1MEX	1MJU	1NGZ	1UAC	1UB6	1UJ3	1UZ8	1YQV	2CGR	2PCP
1A7Q														
1CR9	0.40													
1FLR	0.44	0.58												
1KCV	0.55	0.44	0.46											
1MEX	0.45	0.58	0.48	0.54										
1MJU	0.44	0.63	0.57	0.49	0.60									
1NGZ	0.46	0.57	0.52	0.57	0.78	0.62								
1UAC	0.54	0.44	0.47	0.63	0.51	0.49	0.48							
1UB6	0.46	0.52	0.61	0.44	0.43	0.61	0.45	0.46						
1UJ3	0.54	0.56	0.47	0.49	0.58	0.52	0.61	0.48	0.44					
1UZ8	0.48	0.58	0.66	0.49	0.47	0.65	0.50	0.49	0.70	0.48				
1YQV	0.49	0.55	0.47	0.46	0.64	0.62	0.62	0.47	0.44	0.57	0.49			
2CGR	0.42	0.65	0.67	0.46	0.61	0.72	0.66	0.48	0.56	0.55	0.60	0.65		
2PCP	0.44	0.66	0.65	0.46	0.64	0.68	0.64	0.47	0.53	0.54	0.59	0.57	0.74	

Average 0.54
 Min 0.40
 Max 0.78

Table S1. Combined fraction sequence identity for antibody Fv domains in test set. The fraction sequence identity is across all residues in both the heavy and light chain.

Sequence identities $\geq 70\%$ are shown in gray.

Chapter 3: Relative Orientation of Heavy and Light Chain Variable Domains is a Manifestation of Structural Diversity in Antibodies

Arjun Narayanan¹, Benjamin D. Sellers¹, Matthew P. Jacobson²

¹Graduate Group in Biophysics, University of California, San Francisco, California

²Department of Pharmaceutical Chemistry, University of California, San Francisco, California

Abstract

Diversity in antibody structure is crucial to the ability of adaptive immune system to recognize the tremendously diverse set of potential antigens. The diversity in structure is most apparent in the six hyper-variable loops of the complementarity-determining regions (CDRs). However, given that these CDR loops occur at the interface of the heavy and light chain variable domains and form the paratope for antigen binding, we examined the possibility that the relative orientation of the heavy and light chain variable domains is another source of structural diversity that could lead to changes in antigen binding. Here, we show that there is a large variation in the VL:VH orientation in existing crystal structures. We find the extent of variation is much greater than that expected from effects of crystal packing, antigen binding, or the presence of antibody constant regions, and that this variation is consequently encoded by sequence. Through calculations of the energetics of different orientations, we show that side-chain mediated contacts of interface residues play a major role in defining the energy landscape with respect to VL:VH orientation. The backbone structures of the individual domains are also shown to play a role in defining the orientation. Together, this work establishes that the relative orientation of the heavy and light chain variable domains in antibodies is an important source of structural diversity, which may be important in the ability of the antibody molecule to serve as a scaffold for the recognition of a diverse set of antigens.

Introduction

The sequence diversity in antibodies generated by the immune system is crucial to the recognition of the almost infinite set of potential antigens. How this sequence diversity manifests itself in structural diversity and, consequently, in functional diversity, is an area of active research. Structurally, the overall fold of each antibody variable domains is extremely well conserved, and the differences in sequence and structure within the variable domains are largely localized to the complementarity-determining regions (CDRs), the six variable loops that are important for antigen binding. However, we show here that structural diversity is also generated via the relative orientation between the variable domains of the heavy and light chains. Since the paratope used for antigen binding is most often formed at the interface of the light and heavy chain variable domains, the relative orientation of these domains could have significant impacts on the antigen-binding properties of antibodies.

Some evidence for the idea that the $V_L:V_H$ orientation affects antigen-binding properties comes from antibody humanization efforts in which CDR loops from a high-affinity murine antibody are grafted onto a human framework. In particular, Banfield et al.¹ reported a humanized antibody with a 2-fold change in affinity. Though the conformations of the CDR loops were minimal, a significant difference in the $V_L:V_H$ orientation resulted in large changes in the relative positions of the CDR loops, and consequently the affinity for antigen. Additionally, biochemical experiments in which individual residues at the inter-domain interface were mutated to alanine showed that the dissociation rate of antigen binding increased significantly, affecting both affinity and

specificity². Since mutations at the $V_L:V_H$ interface could have significant effects on the $V_L:V_H$ orientation, these studies may suggest changes in domain orientation can alter antigen binding.

The ability to predict domain orientations would benefit homology modeling refinement not only for antibodies but in general. For example, another family of proteins, kinases, have been observed with different domain orientation dependent on the state of the protein as well as between family members. Several studies have shown that differences in kinase domain orientations are functional³⁻⁷. Furthermore, anecdotal successes exist in computationally predicting domain orientations with various proteins targets and methods^{8,9,10}.

If the $V_L:V_H$ orientation is indeed a functionally important manifestation of antibody sequence diversity, then one would expect to see a large variation in the orientation in existing antibody crystal structures. However, to our knowledge, no such large-scale analysis of the orientation has been performed. Here, we characterize the diversity of $V_L:V_H$ orientations in 142 existing antibody crystal structures and find that the orientation can vary widely, confirming that the relative orientation of the heavy and light chains are a source of structural, and potentially functional, diversity in the set of possible antibodies. We show that this variation in the $V_L:V_H$ orientation is sequence-dependent, and not purely an artifact of crystal packing or antigen-binding. We then used a physics-based scoring method coupled with a sampling scheme based on existing antibody crystal structures to begin understanding how the domain orientation is encoded in sequence.

We find that side-chain mediated contacts in the interface and the conformation of the backbone of the antibody framework are important determinants of the $V_L:V_H$ orientation.

Methods

Production of Datasets:

We first obtained all experimentally-determined structures of antibodies by searching the Protein Databank¹¹ for entries with titles containing “antibody”, “Fv”, or “Fab”. Single-domain (ex. camelid), single-chain (scFv) and homo-dimers were removed resulting in a list of 459 antibody structures. These structures were further reduced in number due to technical difficulties that were discovered in identification and sequence processing of CDR loops and framework residues, generally involving missing residues. This reduced the number of structures in our dataset to 301. As this was deemed a sufficient sized dataset, calculations were performed neglecting these other antibodies. Each structure in this dataset was cut to an Fv fragment by removing all residues more than 8 residues after the end of the third CDR on each chain.

From this original dataset, a non-redundant dataset was produced. First, 9 antibodies with less than 90% sequence identity to each other and which were crystallized as Fv fragments in the absence of antigen were chosen for physics-based analyses to avoid any questions of whether the presence of constant domains or ligand were affecting our prediction accuracy. These 9 antibodies were used as a seed to generate a set of 142 antibodies that had less than 90% sequence identity between any pair.

Generation of Multiple Sequence Alignments for the Heavy and Light Chains:

The heavy and light chain multiple sequence alignments were modified to align CDRs L1, L3, H1, and H2 according to structure-based alignments¹² Gaps in the CDR regions

for H3 were placed in the middle of the CDR region, and for L2, the alignment was preserved (only 7FAB has a deletion in L2).

Calculation of difference in domain orientation:

In the case of comparing the domain orientation between two different antibodies, antibody **A** and antibody **B**, the following protocol was applied. Structural alignments were performed using the non-CDR alpha carbons as the atoms to align unless otherwise noted. First, the heavy chains of **A** and **B** were aligned. The light chain of **A** was then aligned to the light chain of **B** to obtain **A'**, the light chain of **A** in the configuration it would adopt if it had the same domain orientation as **B**. The root mean square deviation (RMSD) of the non-CDR alpha carbons was then calculated between the light chain of **A'** and **A** to give the difference in domain orientation. Using the same alignments, the difference in the fraction of native contacts was calculated, with a contact being defined as any residue pair for which heavy atoms were within 4Å of one another. Using this method, RMSD and native contact calculations are unaffected by structural *intra-chain* differences between **A** and **B**, apart from the effect of these differences on the initial alignments themselves.

Structural bioinformatics examination of variation in domain orientation:

To examine the variation in the $V_L:V_H$ orientation of antibodies, we calculated the difference in orientation between all pairs of our 142 antibody non-redundant dataset.

Calculation of sequence identity and similarity:

Sequence identity and similarity of all pairs within our 142 antibody non-redundant dataset were calculated by combining the identity and similarity calculations for the heavy and light chains. Both the identity and similarity were calculated using the same multiple sequence alignments used to perform the structural alignments. For each position in the alignment, the number of positions that were identical or similar (as defined by a positive score in the Blosum62¹³ matrix) was kept and was divided by the total number of positions compared. Any column in which either of the sequences being compared contained a gap was not used in the comparison.

Decoy Sampling and Physics-based Scoring:

We aimed to develop a method for sampling and scoring different orientations of the heavy and light chain variable domains in order to test hypotheses about the determinants of the orientation. Due to the large number of antibody structures available, we decided to take the approach of using the V_L:V_H orientations in existing antibody crystal structures as states to sample from. We refer to these orientations as “decoys”, as in the context of the current work, these are orientations that are designed to test the ability to predict the known native orientation for a given antibody. The decoy orientations are taken from all 300 antibody structures found in our dataset before filtering for redundancy. For each of the 9 free, non-redundant Fv structures that we found in our dataset (above), we separated the crystallographic heavy and light chains variable domains and aligned these individually to the corresponding domains in each structure of our dataset to generate the decoy orientations.

Clearly, simply reorienting the domains will lead to severe clashes of the side-chains and possibly the backbone in many cases. We made the approximation of a fixed-backbone for simplicity, and while there are obvious problems with this¹⁴, we believed that it would still lead to a useful model in order to probe the interactions that are important for determining the orientation. However, the side-chains at the interface required optimization.

To define the interfacial side-chains, we determined the residues that had any side-chain heavy atom within 4 Å of any heavy atom of the opposing domain in *any* of the native or decoy structures sampled. This set of side-chains, comprising the union of the side-chains at the interface in any of the native or decoy structures, was optimized in all of the native or decoy structures, thus providing a fair comparison of each orientation. The side-chain optimization was performed using the Protein Local Optimization Program (PLOP)¹⁵ with the OPLS all-atom energy function^{16,17} and a Generalized Born solvent model with a surface-area correction^{18,19}.

Decoy Sampling and Physics-based Scoring Using Comparative Models:

In order to test whether backbone perturbations affect $V_L:V_H$ orientation, we built comparative models of the heavy and light chains of each of the 9 free, non-redundant Fv antibodies in our dataset. We built 3 comparative models of each of the 9 antibodies, with different levels of template accuracy. For each of our 9 target antibodies, possible templates in our non-redundant dataset of 142 antibodies were classified according to their difference in structure of the individual domains. One group was formed by those

antibodies for which neither the heavy nor light chain variable domains had $> 1.0 \text{ \AA}$ non-CDR C-alpha RMSD with the corresponding domain in the target antibody. Another group contained antibodies for which at least one of the variable domains had between 1.0 and 1.5 \AA non-CDR C-alpha RMSD with the corresponding domain in the target, and another group contained antibodies for which at least one of the variable domains had between 1.5 and 2.0 \AA non-CDR C-alpha RMSD with the corresponding domain in the target. In each group, the antibody with the largest orientation difference between itself and the target antibody was chosen as the template.

Results

Extent of Variation of $V_L:V_H$ Orientation in Experimentally-Observed Antibody Structures:

We are interested in the hypothesis that $V_L:V_H$ orientation in antibodies may play an important role in determining the structure of the antigen-combining site and thus the binding properties of the antibody. This hypothesis suggests that we should find a significant amount of variation in the orientation of the variable domains in existing crystal structures.

To examine the extent of variability in the $V_L:V_H$ orientation present in existing crystal structures of antibodies, we created a dataset of 300 antibody structures of which a subset of 142 structures was used as a non-redundant set to be used in structural bioinformatics studies (see Methods). The difference in domain orientation for each pair of structures was calculated by first aligning the heavy chains followed by optimal superposition of the light chains. The difference in orientation was defined as the non-CDR alpha-carbon RMSD calculated between the light chain of one of the structures in its original and superimposed orientations. This ensures that contributions to the RMSD arising from different internal structures of the light chain are discounted, resulting in a more accurate measure of the difference in orientation.

We found the differences in domain orientation to be moderate overall, with a mean of 2.3 Å RMSD. However, with a standard deviation of 1.1 Å, it appears that the

differences in orientation between antibodies can often be quite large, especially in comparison with the RMSDs between the non-CDR regions between different light or heavy chains, which is 0.9 and 1.1 Å, respectively. This indicates that the $V_L:V_H$ orientation may be an important source of structural diversity in antibodies. The difference in domain orientation in 211 out of 10011 pairs were greater than 5.0 Å, with the largest deviation observed being 7.4 Å. While these pairs obviously represent only 2% of the data points, the fact that they exist reinforces the idea that the heavy and light chain variable domains can take on significantly different orientations between different antibodies.

Crystal packing and antigen binding effects may give rise to some of the observed variation in orientation. Therefore, we calculated the differences in domain orientation between 24 pairs of antibody structures of identical sequences regardless of crystal form or antigen-binding state and filtered for redundancy at 90% sequence identity. We found that the mean orientation difference was 0.7 Å with a standard deviation of 0.4 Å and a maximum orientation difference of 2.0 Å. These smaller orientation differences in identical-sequence complexes indicate that the large differences observed in the non-redundant dataset are mostly due to sequence difference and only a small amount of the variation is attributable to effects of crystal packing or antigen binding.

Correlation of Differences in $V_L:V_H$ Orientation with Differences in Sequence:

We next investigated whether the variation between antibodies in our dataset was correlated with simple measures of the relatedness of their sequences. First, we

calculated the sequence identity and similarity over the full length of both the light and heavy chain variable regions for all pairs of antibodies. The difference in domain orientation for each pair was plotted against the sequence identity or similarity (Figure 1, A and B, respectively). The mean difference in orientation was qualitatively correlated with both sequence identity and similarity. However, the range of observed orientations seen at a given sequence identity or similarity was large.

Since the residues at the interface of the two domains may play a larger role than non-interface residues in determining the $V_L:V_H$ orientation, we examined whether differences in domain orientation were correlated with sequence identity and similarity of the interface residues.. Once again, while the mean values of differences in orientation are correlated with the sequence identity or similarity of the pairs, there is large variation for pairs of antibodies with the same identity or similarity (Figure 1, C and D) . Together, these data suggest that the domain orientation of an antibody is encoded by its sequence; however, the simple sequence identity or similarity metrics are not sufficient to determine the difference in orientation between two antibodies, even when comparing sequences with relatively high sequence conservation.

Energy-Based Prediction of $V_L:V_H$ Orientation – Case Study:

To further support the idea that the $V_L:V_H$ orientations is encoded by sequence, we developed a tool for sampling and scoring relative orientations of the heavy and light chain variable domains. The method samples from $V_L:V_H$ orientations found in existing crystal structures and scores these orientations after side-chain optimization of the residues at the interface of the variable domains. We illustrate the various calculations we

performed and the results obtained using a case study of a humanized variant of anti-p185^{HER2} antibody 4D5 (PDB ID: 1fvc).

Testing the impact of the loss of interface side-chain contacts on $V_L:V_H$ orientation:

In order to examine the consequences of the loss of interface side-chain contacts on $V_L:V_H$ orientation, we predicted the orientation of nine antibodies using their native sequences or after having mutated all interface residues to alanine. These calculations were done in the context of a perfect backbone structure of the individual domains, including all the CDR loops.. We sampled relative orientations of the heavy and light chain observed in 300 existing crystal structures of antibodies. The full dataset of crystal structures (and not the non-redundant dataset) was used because it is possible that slight variations in orientation between very closely related sequences may have significantly different energies. While there may be some bias in the fact that some areas of orientation space will be better sampled than others, this would be true with the smaller non-redundant data set as well, and the smaller sampling space could limit the prediction accuracy. All residues found at the interface (see Methods) in any of the 300 decoy structures or native structure were selected to be treated flexibly in calculating the energy of each orientation. For each decoy structure and for the crystal structure, the interface side-chains were optimized and ranked by energy using the Protein Local Optimization Program (PLOP) with the OPLS all-atom energy function¹⁵⁻¹⁷ and a Generalized Born solvent model^{18,19} with a surface-area based nonpolar correction term. The results show that we predict a near-native orientation as the lowest-energy, with a domain orientation RMSD of 0.5 Å (Figure 3a). This provides further evidence that the orientations observed in antibody crystal structures are encoded in sequence, and are representative of the global minimum in the energy landscape.

In contrast, when the same protocol was applied after all of the interface residues were mutated to alanine, the lowest-energy orientation was far from the crystallographic orientation, with a domain orientation RMSD of 3.8 Å (Figure 3a). Furthermore, while the energy landscape of the native sequence showed a distinct funnel-like shape, the energy landscape of the sequence with interface residues mutated to alanine resulted in a much flatter energy landscape with little preference for orientation (Figure 3b). It is clear that side-chain mediated contacts involving the residues at the interface are defining factors of the energy landscape, resulting in the preference for a well-defined $V_L:V_H$ orientation.

We next tested to see whether the CDR H3 loop plays a significant role in determining domain orientation. There is crystallographic evidence showing different domain orientations for antibodies with and without antigen and, though no causal effect has been established, conformational changes in the H3 loop have been linked to these domain orientation changes. Additionally, in comparative modeling studies, we have seen that the orientation of the variable domains can affect predictions in the structure of the H3 loop. If the structure of the H3 loop was also required to accurately predict the orientation, a difficult optimization problem would result.

To test the hypothesis that contacts made by the H3 loop are a significant determinant of the $V_L:V_H$ orientation, we removed the H3 loop from the 1fvc heavy chain and again ran our decoy-sampling and interface side-chain optimization protocol. The orientation RMSD increased from 0.5 Å in the calculations on the unmodified domains to 1.4 Å after

the H3 loop is removed (Figure 3a and b). There appears to be some information contained in the H3 loop structure to specify the orientation, but overall, it is not a drastic effect, as there appears to be enough orientation-determining information contained in the rest of the interface. This suggests that comparative modeling efforts can consider the orientation of the variable domains prior to structure prediction of the H3 loop and, in the absence of other errors, be able to expect a reasonably accurate prediction of the orientation.

Effects of Deviations in the Backbones of the Individual Variable Domains on the $V_L:V_H$ Orientation:

All of the above calculations were performed using the crystallographic backbone structure of the individual variable domains of 1fvc. We wanted to examine how much altering the backbone structure would affect the orientation for two reasons. First, alterations in backbone structural elements through affinity maturation may lead to changes in the backbone structure of the individual variable domains. This consequently could have an effect on the orientation. Secondly, backbone errors will be present in comparative models of the variable domains. These errors may significantly affect the ability to accurately predict the $V_L:V_H$ orientation. In order to test the effects that deviations from the native backbone conformation have on the domain orientation, we introduced error in the backbone structures by building comparative models of the heavy and light chain variable domains with varying accuracy.

We picked three antibodies to be used as templates from which to model the variable domains of 1fvc (see Methods). The accuracy of the resulting models is shown in Figure 4. The H3 loop was not built for these models, as we believe that modeling the conformation accurately may require an accurate domain orientation. Given that the absence of the H3 loop did not result in large decreases in accuracy when the crystallographic backbone atoms were used, we do not think that the absence of the H3 loop significantly affects our results.

These models of the individual domain structures were then input into our decoy sampling and side-chain optimization protocol to see the impact that these backbone differences have on the $V_L:V_H$ orientation. Our results are shown in Figure 4. Even fairly small deviations in the backbone structures of the individual domains can dramatically change the relative orientation of the heavy and light chain variable domains. Conversely, this allows the possibility for small and seemingly local changes in structure to have long-range effects and alter the structure of the paratope by causing changes in the $V_L:V_H$ orientation.

Physics-Based Prediction of $V_L:V_H$ Orientation – Overall Results:

The overall results of testing our decoy sampling protocol on nine antibodies using the crystallographic domain structures are shown in Table I. The nine antibodies were chosen as those which were a) crystallized as an Fv fragment in the absence of ligand and b) had < 90% pairwise sequence identity between them. The fact they were crystallized

as Fv fragments discounts the possibility that the constant regions of the antibodies or the presence of ligand could affect the orientation.

Overall, the results in nearly all of the cases show similar results to those presented for 1fvc. In the context of the unmodified crystallographic structures of the individual domains, our method predicts the native or near-native ($< 1\text{\AA}$ RMSD) orientation in all cases. Again, this suggests that the observed crystallographic orientation in these cases is the global minimum in the energy landscape and determined by its sequence. In all cases, we find that side-chain contacts made by residues at the interface are a key determinants of the $V_L:V_H$ orientation, as mutation of the interface residues to alanine often results in prediction of a non-native orientation (Table I) and a flattened energy landscape with respect to the relative orientation of the heavy and light chain variable domains. In addition, we see that in most cases, the removal of the H3 loop prior to prediction only has fairly small effects on the orientation suggesting that major determinants of the orientation reside outside the H3 loop.

We also tested the dependence of the orientation on differences in backbone structure as we did above with 1fvc. Again, the results show that deviations, even modest ones, of the backbone conformation from the native conformation results in altering the energy landscape such that non-native orientations are now the lowest-energy (Table II). . Minimizations that allow the backbone to relax back towards their native conformation (which is also determined by sequence) allow the variable domains to return to more native-like orientations, although the limited sampling afforded by minimization leads to

the recovery being incomplete. The fact that small deviations in the backbone conformation can significantly influence the energetics of different $V_L:V_H$ orientations suggests an avenue by which mutations could affect the domain orientation and consequently the function in an action-at-a-distance effect.

Discussion

We have shown that there is a large amount of structural diversity in the relative orientation of the heavy and light chain variable domains in antibodies, and we show that crystal packing, antigen binding, or constant region effects can account for only a small part of this variation. We show this by comparing the differences in orientation we see in existing crystal structures to the differences we see between crystal structures of identical sequence, regardless of crystal packing environment, antigen-binding state, or presence or absence of Ig constant regions (Figure 1). We also find an inverse relationship between the mean orientation difference and the sequence identity or similarity over either the full or interface sequences. We support this idea through calculations of the energy of different relative orientations of the heavy and light chain variable domains, and show that by neglecting any factors other than the sequence of the antibody and the structure of its individual domains, the native orientation is still predicted to be the global energy minimum (Table I).

Because of the large amount of variation we observe in the $V_L:V_H$ orientation, we propose that this may be an important part of the ability of the immune system to recognize a tremendously diverse set of antigens, together with other structural manifestations of sequence diversity as the amino-acid composition and structure of the CDR loops. Indeed, the diversity in antibody orientation would allow loops of the same structure (and possibly sequence) to present a different combining site to the antigen, allowing for a greater diversity of recognition. Coupled with the newly-emerging idea that the heavy chain constant region can also affect the affinity and specificity of antigen-

binding²⁰⁻²³, the notion that there are multiple ways in addition to CDR diversity which the immune system uses in order to convert the sequence diversity generated through somatic recombination and affinity maturation into functional diversity is gaining credibility.

We have shown that the domain orientation is sensitive to the amino-acid composition of the interface and that the orientation is sensitive to small backbone shifts in the individual domains. In our calculations we observe that mutation of the interface residues to alanines resulted in a flattening of the energy landscape governing $V_L:V_H$ orientation, indicating that the side-chain mediated contacts of the interface residues are key factors in shaping the energy landscape. Incorrect predictions were also observed when the correct sequences were given small backbone errors in the individual domains by building comparative models of the individual domains.

Mutations in the framework have previously been observed to have an effect on the spatial positioning of the CDR loops²⁴, and to have effects on antigen-binding properties through altering of the $V_L:V_H$ affinity. But our work suggests that additionally, mutations may change the orientation of the variable domains through a direct mechanism (e.g. mutation at the interface) or an indirect mechanism whereby influencing the backbone structures of the individual domains propagates to changes in the interface that must be accommodated by reorientation of the variable domains.

There is a possibility that some number of key contacts are essential in determining the domain orientation. We cannot, at this time, rule out this possibility. However, one would expect such a model to predict that there would be a limited number of discretized orientations based upon the combination of contacts that exist for each antibody. However, we failed in attempts at clustering the orientation by RMSD. Instead of observing discrete clusters, we observed a continuum of conformations (data not shown). Secondly, comparisons of antibodies with either very similar or very different orientations suggests that there are amino-acid differences that can compensate for each other to maintain the same orientation, and residues that can be used in different ways, adopting different conformations based on the residue context around them to result in very different orientations (data not shown). At the moment, the orientation seems to be dependent on the properties of the interface as a whole as defined by the exact sequence of the antibody.

Finally, the fact that there is such a large diversity in the $V_L:V_H$ orientation suggests that this is a degree of freedom that needs to be taken into account when building comparative models of antibodies. Our work here has shown that prediction of the orientation cannot be achieved by simply repacking side-chains on a fixed backbone. In general, in order to build quality comparative models of antibodies, backbone perturbations of the individual domains must be sampled (either in advance in the production of an ensemble of backbones or simultaneously) in addition to the domain orientations and side-chains at the interface.

References

1. M.J. Banfield DJKAMRLB. VL:VH domain rotations in engineered antibodies: Crystal structures of the Fab fragments from two murine antitumor antibodies and their engineered human constructs. Volume 29; 1997. p 161-171.
2. Khalifa MB, Weidenhaupt M, Choulier L, Chatellier J, Rauffer-Bruyere N, Altschuh D, Vernet T. Effects on interaction kinetics of mutations at the VH-VL interface of Fabs depend on the structural context. Volume 13; 2000. p 127-139.
3. Ulmer TS, Werner JM, Campbell ID. SH3-SH2 Domain Orientation in Src Kinases NMR Studies of Fyn. Structure 2002;10(7):901-911.
4. Arold ST, Ulmer TS, Mulhern TD, Werner JM, Ladbury JE, Campbell ID, Noble MEM. The Role of the Src Homology 3-Src Homology 2 Interface in the Regulation of Src Kinases. Journal of Biological Chemistry 2001;276(20):17199-17205.
5. Cai SJ, Khorchid A, Ikura M, Inouye M. Probing Catalytically Essential Domain Orientation in Histidine Kinase EnvZ by Targeted Disulfide Crosslinking. Journal of Molecular Biology 2003;328(2):409-418.
6. Wilson KP, Fitzgibbon MJ, Caron PR, Griffith JP, Chen W, McCaffrey PG, Chambers SP, Su MSS. Crystal Structure of p38 Mitogen-activated Protein Kinase. Journal of Biological Chemistry 1996;271(44):27696.
7. Cobb MH, Goldsmith EJ. How MAP kinases are regulated. Journal of Biological Chemistry 1995;270(25):14843-14846.
8. Deng M, Mehta S, Sun F, Chen T. Inferring Domain-Domain Interactions From Protein-Protein Interactions. Genome Research 2002;12(10):1540.
9. Halperin I, Ma B, Wolfson H, Nussinov R. INVITED REVIEW Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. Proteins 2002;47:409-443.
10. Xu D, Baburaj K, Peterson CB, Xu Y. Model for the three-dimensional structure of vitronectin: Predictions for the multi-domain protein from threading and docking. Proteins Structure Function and Genetics 2001;44(3):312-320.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
12. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. Volume 273: Elsevier; 1997. p 927-948.
13. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America 1992;89(22):10915-10919.
14. Bonvin AMJJ. Flexible protein-protein docking. Current Opinion in Structural Biology 2006;16(2):194-200.
15. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. J Phys Chem B 2002;106(44):11673-11680.
16. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. Q ReV Biophys 1993;26:49.
17. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with

- accurate quantum chemical calculations on peptides. *J Phys Chem B* 2001;105(28):6474–6487.
18. Gallicchio E, Zhang LY, Levy RM. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. Volume 23; 2002. p 517-529.
 19. Ghosh A, Rapp CS, Friesner RA. Generalized Born model based on a surface integral formulation. Volume 102; 1998. p 10983-10990.
 20. Cooper LJ, Shikhman AR, Glass DD, Kangisser D, Cunningham MW, Greenspan NS. Role of heavy chain constant domains in antibody-antigen interaction. Apparent specificity differences among streptococcal IgG antibodies expressing identical variable domains. *J Immunol* 1993;150(6):2231-2242.
 21. Torres M, Fernandez-Fuentes N, Fiser A, Casadevall A. The Immunoglobulin Heavy Chain Constant Region Affects Kinetic and Thermodynamic Parameters of Antibody Variable Region Interactions with Antigen. *J Biol Chem* 2007;282(18):13917-13927.
 22. Torres M, May R, Scharff MD, Casadevall A. Variable-Region-Identical Antibodies Differing in Isotype Demonstrate Differences in Fine Specificity and Idiotype. *J Immunol* 2005;174(4):2132-2142.
 23. McLean GR, Torres M, Elguezabal N, Nakouzi A, Casadevall A. Isotype Can Affect the Fine Specificity of an Antibody for a Polysaccharide Antigen. *J Immunol* 2002;169(3):1379-1386.
 24. Studnicka GM, Soares S, Better M, Williams RE, Nadell R, Horwitz AH. Human-engineered monoclonal antibodies retain full specific binding activity by preserving non-CDR complementarity-modulating residues. *Protein Eng* 1994;7(6):805-814.

Chapter 3: Table I

	RMSD of Lowest Energy Conformation in Å		
PDB ID	Reproduction	Alanine Interface	H3 Removal
1a6u	0.8	1.5	1.6
1a7n	0.2	0.8	0.2
1bvl	1.0	2.1	1.2
1dlf	0.4	0.4	0.4
1dql	1.5	4.2	0.9
1dsf	1.0	1.0	2.5
1fvc	0.5	3.8	1.4
1igm	0.6	3.6	1.6
43c9	0.3	1.3	0.7

Table I: Results of decoy-sampling calculations using crystallographic backbones.

The domain orientation RMSDs of the predicted lowest-energy orientations (excluding the native orientation) are shown. Cases in which this number is in **bold** are cases for which the native orientation would score as the lowest-energy orientation.

Chapter 3: Table II

	<i>RMSD of Lowest Energy Conformation in Å</i>		
PDB ID	<1 Å	1-1.5 Å	1.5 - 2.0 Å
1a6u	2.6	3.5	3.1
1a7n	3.4	4.5	3.2
1bvl	3.7	4.4	4.3
1dlf	3.5	2.6	2.4
1dql	1.3	2.3	4.2
1dsf	1.5	2.9	4.7
1fvc	2.4	1.4	2.8
1igm	2.0	2.0	2.5
43c9	2.0	3.4	3.4

Table II: Results of decoy-sampling calculations using modeled structures of the individual heavy and light chain domains.

The domain orientation RMSDs of the predicted lowest-energy orientations are shown in cases in which templates of < 1Å non-CDR C-alpha backbone RMSD of the individual domains, 1-1.5 Å, or 1.5-2.0 Å were used to generate the individual domain structures. These individual domain structures were generated without an H3 loop, and the results are significantly worse than predictions made using the crystallographic backbone (Table I, column “H3 Removal”).

Chapter 3: Figure 1

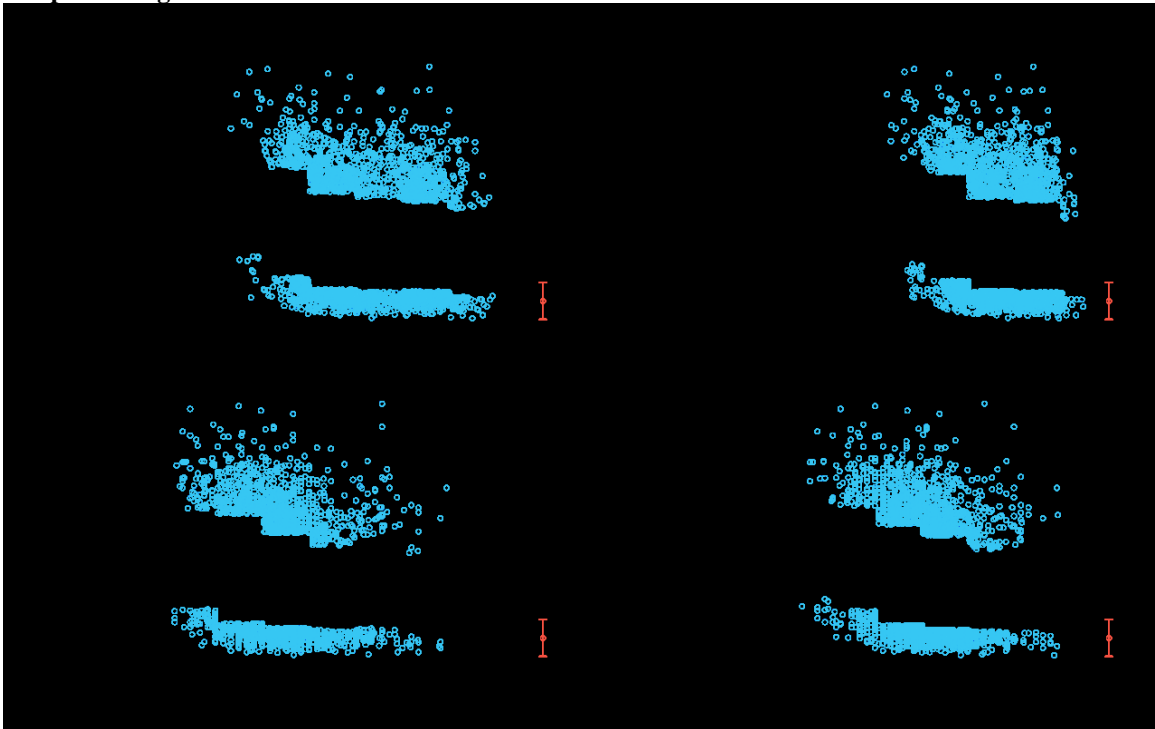


Figure 1: Mean difference in domain orientation is correlated with simple measures of difference in sequence, but there is large variability. The difference in the domain orientation (as measured by non-CDR C-alpha RMSD, as calculated in methods) for a non-redundant set of 142 antibodies is plotted against the a) combined sequence identity (a) or similarity (b) over the entirety of both the heavy and light chain sequence; or the sequence identity (c) or similarity (d) over all residues that make inter-chain contacts in any of the structures in the dataset. Each data point was binned in 10% increments of sequence identity or similarity, and the mean was calculated, and a line was plotted for the means of all bins containing greater than 50 data points (black line). Data points that were in the lowest or highest 10% of domain orientation differences are plotted as points (blue). The mean and standard deviations of comparisons involving identical antibodies from different crystal structures is shown in red at 100% identity or similarity.

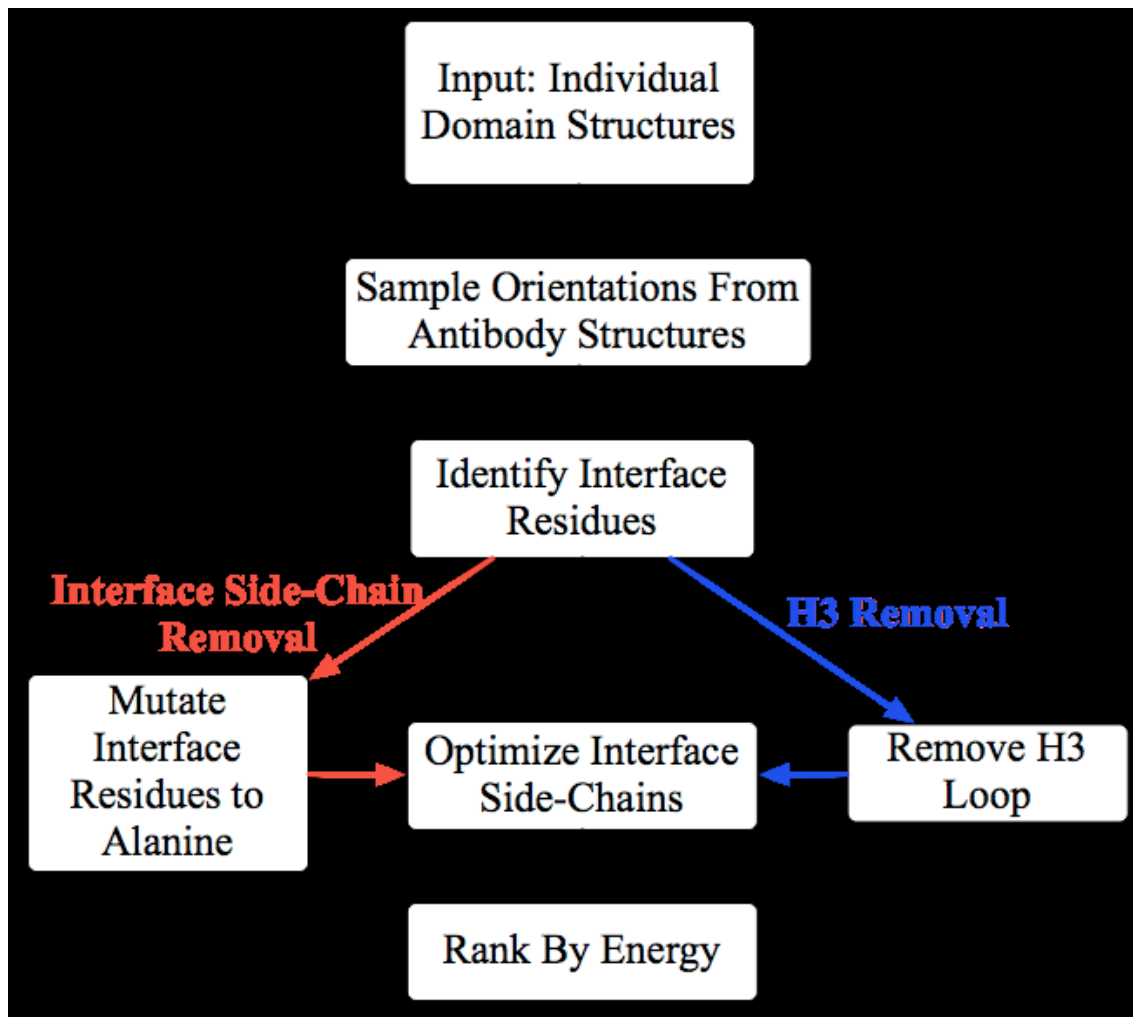


Figure 2: Schematic of different modifications of a decoy-sampling scheme for prediction of the $V_L:V_H$ orientation.

As input, the methodology takes the individual structures of each domain. In the testing of the method done here, these individual structures come from the true crystallographic structure. In the case of a true prediction and in our analysis of the tolerance of the method to backbone errors, these individual domain structures would come from comparative models. The individual domain structures are then combined to take on the orientations observed in our dataset of 300 antibody structures as described in Methods. For each of these orientations, any side-chains that are involved in interchain contacts are

added to a list of residues for which side-chain sampling will be done in all orientations. In the *Reproduction* protocol, these orientations are passed directly to side-chain optimization and scoring with the OPLS all-atom energy function and Generalized Born solvation model. This provides a test of whether this method works in the ideal case of a perfect environment for reconstruction of the crystallographic orientation. In the *Interface Side-Chain Removal* protocol, the interface residues are all mutated to alanine and then passed to the side-chain optimization and scoring. This test allows us to ensure that we are capturing the sequence-dependent nature of the orientation. In the *H3 Removal* protocol, the H3 loop is removed before the orientations are passed to the side-chain optimization and scoring. This provides a test of whether the H3 loop is required for determining the $V_L:V_H$ orientation.

Chapter 3: Figure 3

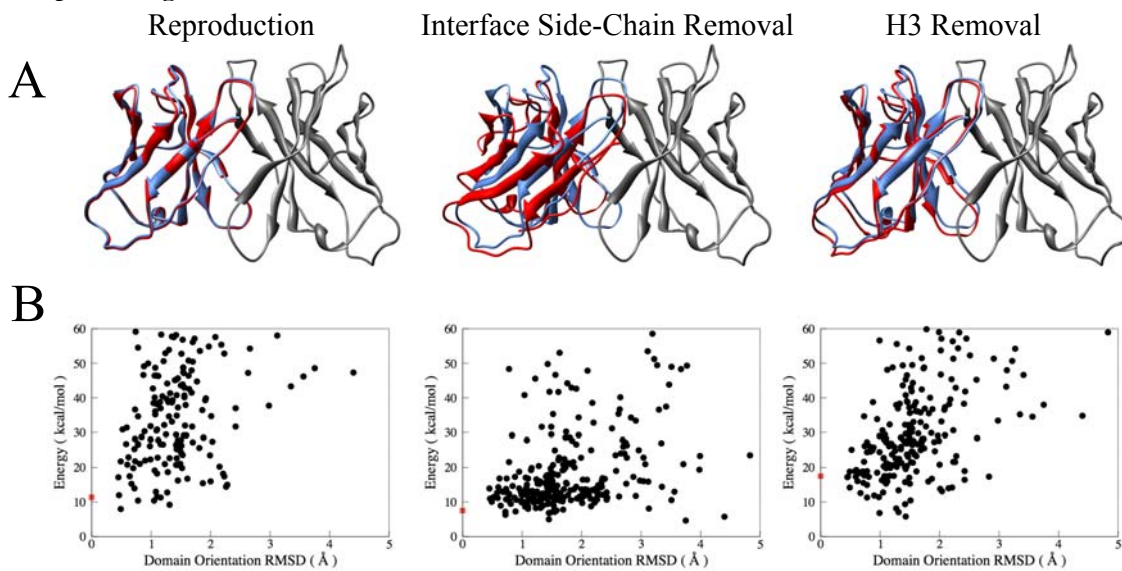


Figure 3: Results of the decoy-sampling method for antibody 1fvc.

a) The predicted orientation of the light chain (red) vs. the native orientation (blue) and the corresponding domain orientation RMSDs. The crystallographic heavy chain is shown in gray in all cases. b) Predicted energy vs. domain orientation RMSD plots of the decoy orientation sampled. The native orientation is shown as a red square on the y-axis. The decoy sampling method correctly predicts the native orientation, captures the sequence dependence of the orientation, and shows that the H3 loop is not a major determinant of the $V_L:V_H$ orientation in antibody 1fvc.

Chapter 3: Figure 4

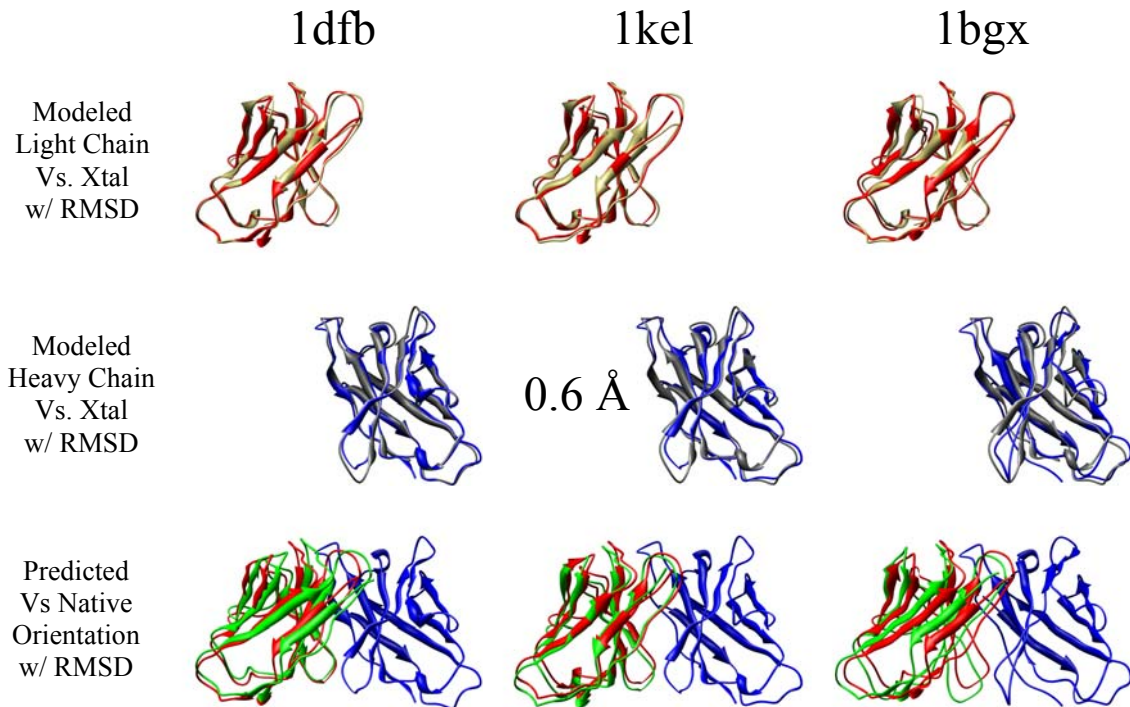


Figure 4: Small Differences in Backbone Conformation can lead to differences in orientation.

Models of the individual light (red)and heavy (blue) chain variable domains of 1fvc generated by templates 1dfb, 1kel, and 1bgx, and the associated predicted orientations (green)and RMSDs using the decoy-sampling scheme. Models built from 1dfb have a small amount of error in the backbone conformation of the individual domains, yet still result in an inaccurate prediction. Models built from 1kel have a slightly larger amount of error in their light chain, but yield a prediction comparable in accuracy to the crystallographic domains without the H3 loop. Models built from 1bgx have larger errors in the backbones of the individual domains and result in an inaccurate prediction.

Chapter 3: Figure 5

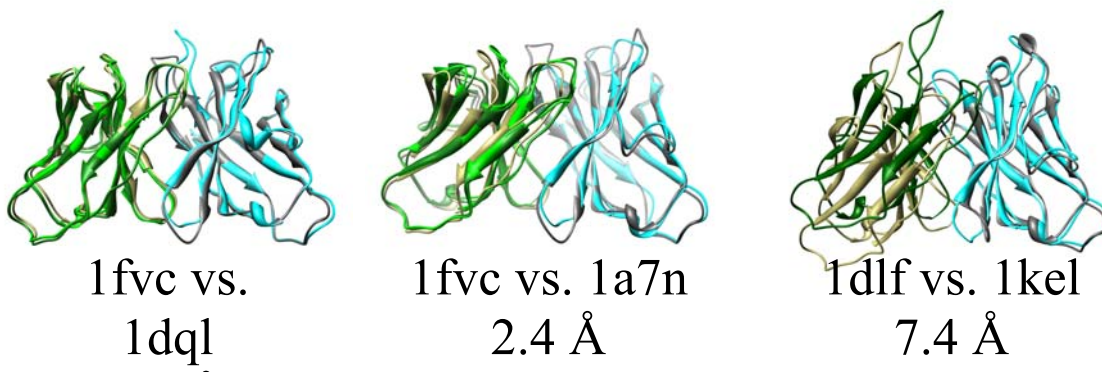


Figure 5: Overall structural and orientational differences observed in the comparative analysis of three sets of antibodies.

The antibodies being compared and the domain orientation RMSDs are indicated below each image. The first of the antibodies is shown with the heavy chain in gray and the light chain in beige. The second of the antibodies is shown with the heavy chain in cyan and the light chain in green.

Chapter 3: Figure 6

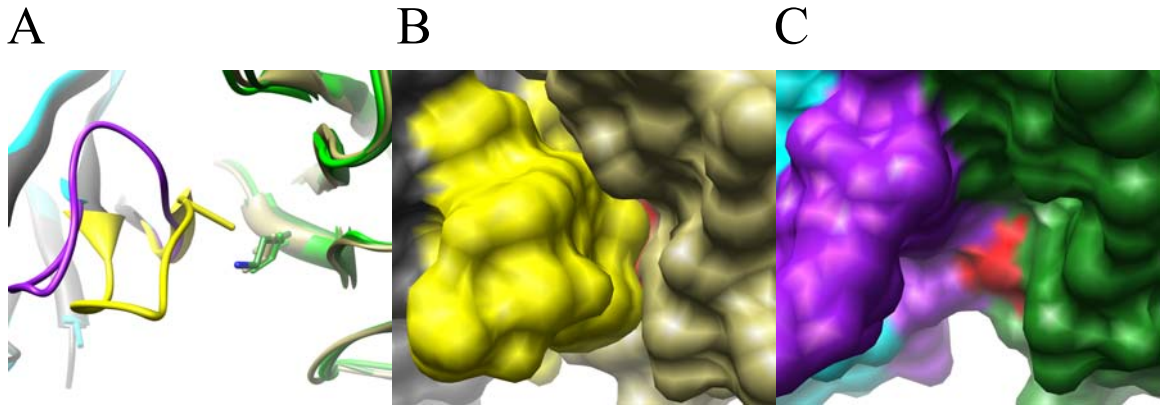


Figure 6: Differences in the H3 loop compensate for the L46K sequence difference between 1fvc and 1dql.

1fvc is shown with its heavy chain in grey, its light chain in beige, and its H3 loop in yellow. 1dql is shown with its heavy chain in cyan, its light chain in dark green, and its H3 loop in purple. a) The ribbon representation showing the conformation of the residue at position 46_L and the differing conformations of the H3 loop. B) The surface representation of 1fvc. L46 is colored red, and is largely buried by the H3 loop. C) The surface representation of 1dql. K46 is colored red, and the differing H3 loop allows this residue to be solvent exposed.

Chapter 3: Figure 7

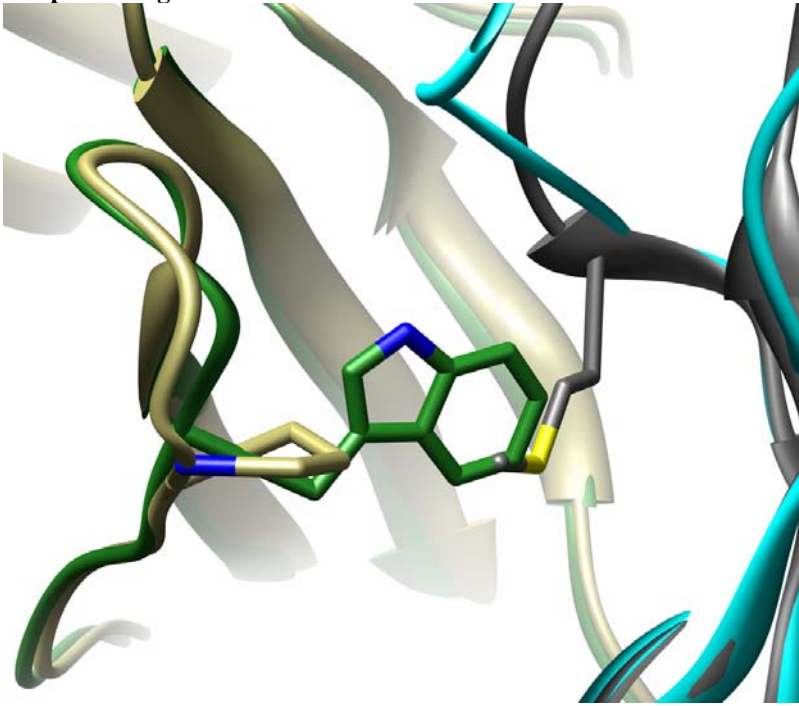


Figure 7: Compensatory mutations in the hydrophobic core of the V_L:V_H interaction.

1fvc is shown with its heavy chain in grey and its light chain in beige. 1dql is shown with its heavy chain in cyan and its light chain in dark green. M107_H in 1fvc is replaced with a glycine at the corresponding position in 1dql. To make up for the space created in this substitution, there is a corresponding difference at position 96_L where there is a proline in 1fvc and a tryptophan in 1dql, along with a 1 residue deletion in L3.

Chapter 3: Figure 8

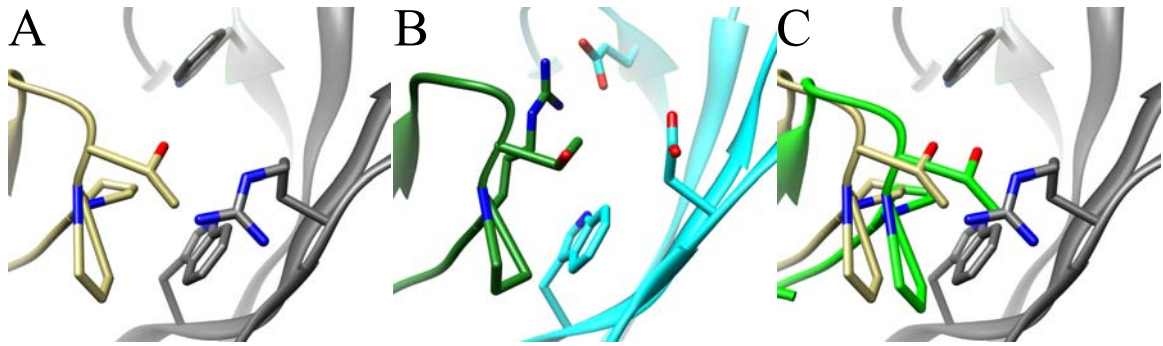


Figure 8: Differences around W47_H would lead to clashes if 1fvc adopted the 1a7n orientation.

Structural differences around W47_H are shown for a) 1fvc (heavy chain in grey, light chain in beige) and for b) 1a7n (heavy chain in cyan, light chain in dark green). Corresponding residues in both structures are shown with R59_H in 1fvc corresponding to a glutamate in 1a7n, P96_L corresponding to an arginine, and F98_H corresponding to glutamate. C) If the 1fvc light chain were to adopt the orientation found in 1a7n, (green), T94_L would clash with R59_H and P95_L would clash with the highly conserved W47_H.

Chapter 3: Figure 9

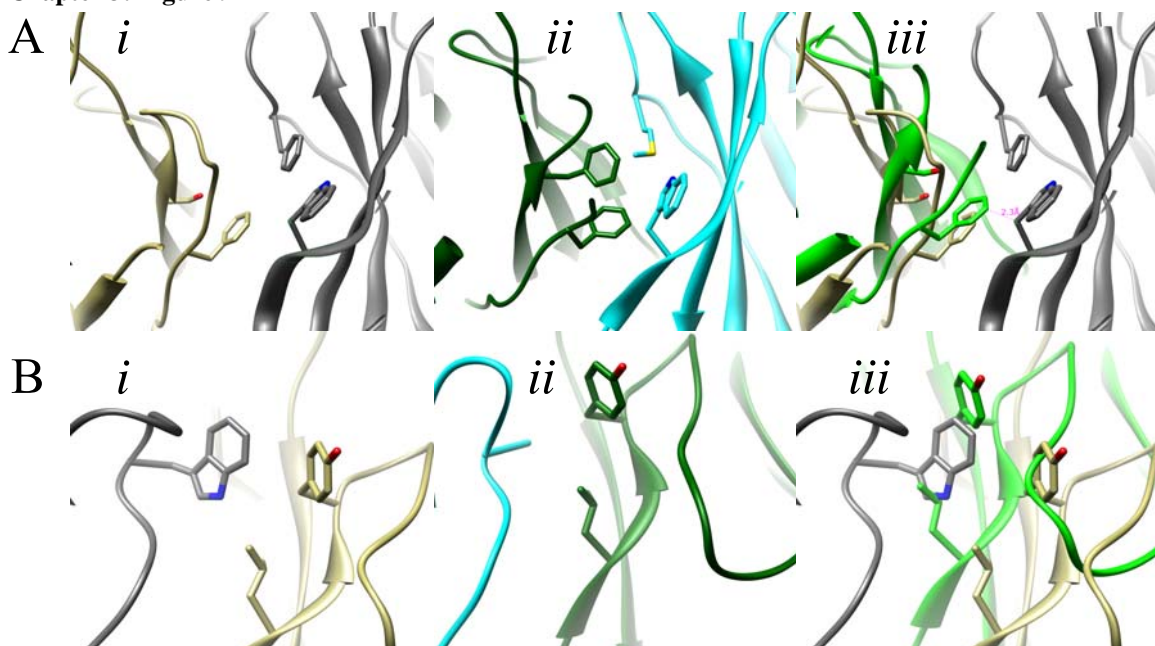


Figure 9: Substitution of large aromatics leads to inability of 1dlf to adopt the 1kel orientation.

Close-up structural examination of a) the region around W47_H and b) the region around W101_H (both in 1dlf numbering). Panel *i* shows 1dlf, with its heavy chain in grey and its light chain in beige. Panel *ii* shows 1kel, with its heavy chain in cyan and its light chain in dark green. Panel *iii* shows 1dlf, with its light chain in its native orientation and moved into the orientation taken on by 1kel (green). The large shift in orientation between these two structures would result in steric clashes between F98_L and W47_H (a), and between W101_H and Y49_L and L46_L (b). These clashes would likely require extensive side-chain and backbone conformational changes to be resolved.

Chapter 4: How Do Somatic Mutations Rigidify CDR Loops During Affinity Maturation?

Sergio Wong¹, Benjamin D. Sellers¹, Matthew P. Jacobson²

¹Graduate Group in Biophysics, University of California, San Francisco, California

²Department of Pharmaceutical Chemistry, University of California, San Francisco, California

Abstract

Prior studies have suggested that antibody CDR flexibility is lost during affinity maturation, but the physical basis for the rigidification remains unclear. Here, molecular dynamics simulations captured CDR flexibility differences between 4 mature antibodies (7G12, AZ28, 28B4, 48G7) and their germline predecessors. Analysis of their trajectories: 1) rationalized how mutations during affinity maturation restrict CDR motility, 2) captured the equilibrium between bound and unbound conformations for the H3 loop of unliganded 7G12, and 3) predicted a set of new mutations that, according to our simulations, should diminish binding by increasing flexibility. In addition, we enumerated energy minima for the H3 loop using our previously published loop sampling method. Qualitatively, three of the four pairs show an increased number of low energy basins in the mature sequence.

Introduction

The immune system is able to produce highly specific and potent antibodies for essentially any target molecule, or antigen. The process starts with germline antibodies, the set originally available at birth, binding a target antigen. These antibodies undergo cycles of somatic hypermutation and selection, for the strongest binders, to yield “affinity-mature” antibodies. Six hypervariable loops or complementary determining regions (CDR) form the antibody binding site¹.

Because antibodies can bind essentially any antigen by changing the amino acid sequence of these hypervariable loops, they are an excellent model system to study molecular recognition and binding. Understanding how and why somatic mutations modulate binding affinity may permit the design or disruption of protein-protein interactions, protein-ligand interactions and enable the rational engineering of antibodies for improved affinity. These would be useful in developing new applications in biotechnology and as therapeutic agents^{2,3}.

One hypothesis suggests that germline antibodies are inherently flexible, easily rearranging to facilitate binding to multiple antigens and that mutations during maturation restrict CDR loops to prearrange them for binding⁴⁻⁷. Binding site pre-organization eliminates the free energy cost of rearrangement upon binding and contributes to affinity maturation. Experimental data that supports this position includes: 1) crystallographic and kinetic studies of mature immunoglobulins and their germline predecessors^{4,8,9,8}; 2) the entropic lower cost of binding after maturation^{5,10}; 3) the loss of polyreactivity⁶, the

ability to bind distinctly different antigens, as antibodies mature. If this hypothesis is correct, how do mutations modulate flexibility?

Studying the factors that drive flexibility can be experimentally challenging. B-factors from crystallographic studies and NMR relaxation data^{11,12} can provide indications of flexibility; however, deciphering the mechanism by which a given mutation modulates flexibility can be difficult. This motivates a theoretical approach to examine these effects to identify atomic-resolution interactions that limit flexibility.

A few previous studies have examined the dynamics of specific antibodies. Zimmerman et al¹³ studied the anti fluorescein antibody 4-4-20 using molecular dynamics to complement nonlinear laser spectroscopy and surface plasmon resonance measurements. Their simulation complement experimental data in showing that the mature antibody is more rigid than its germline counterpart. Analysis of crystallographic coordinates suggest hydrogen bonding and packing help rigidify CDRs. Thorpe and Brooks¹⁴ performed molecular dynamics and binding free energy estimates to further study the 4-4-20 antibody. Their analysis shows a drop in the entropic cost of binding as the antibody matures. Sinha et al¹⁵ performed simulations of the HyHEL63 antibody and its target. Results of this analysis show that salt bridges, absent in crystal structures, formed during the molecular dynamics runs and play an important role in binding and specificity.

We present here a broader survey that examines four mature/germline antibody pairs (7G12, AZ28, 28B4 and 48G7) using molecular dynamics and loop sampling. We

explore how interactions introduced by somatic mutations lower CDR flexibility during antibody maturation and find common mechanisms for this process. In all four cases, the calculations capture greater flexibility, assessed by calculated B-factors, in germline H3 CDR loop than their mature counterparts. Via analysis of the simulations it was possible to: 1) identify single mutations that, even far from the paratope, can significantly restrict CDR mobility, 2) identify putative new mutations that could affect binding, 3) observe 7G12 visit its bound H3 conformation even in absence of antigen i.e. it binds through a conformational selection mechanism, and 4) observe that 48G7 CDR flexibility loss does not drive its affinity maturation. These results support previous suggestions that hydrogen bond/salt bridges and tight sidechain packing play an important role in restraining CDRs during the maturation process. In addition, they show evidence of a conformational selection binding mechanism for the 7G12 germline antibody, where somatic mutations change the population distribution towards a pre-arranged “bound” conformation. While anecdotal, the finding that single mutations can independently modulate flexibility, and affect the binding affinity, could be particularly useful in the design and optimization of binding sites.

Methodology

Molecular dynamics

Molecular dynamics of 7G12, 28B4, AZ28 and 48G7 were performed using the Amber 8 simulation package¹⁶. Simulations were carried out at 300 and 400 K. In order to avoid distortions at 400 K and make RMSD comparisons easier, C α atoms not on CDR loops were restrained harmonically (force constant =0.3 kcal/mol/Å²) to their crystallographic coordinates. Solvent effects were incorporated by: 1) using Langevin dynamics¹⁷ and 2) a Generalized Born¹⁸ (igb=5) model with a surface area correction for non-polar effects. The collision frequency parameter was set equal to 2 to maximize barrier crossing¹⁹. The parm99 force field with Simmerling's backbone corrections²⁰ were used. For computational efficiency, long range forces were only updated every 4 time-steps (nrespa=4). The inner timestep was 1 fs. The production run spanned 15 ns of simulation time.

The PDB structure files, specific chain segments used, and the residues not restrained during the simulations are listed in Table 1. Note that no ligands were included in the simulations. The starting structures underwent 2000 steps of steepest descent minimization; an equilibration run followed. The temperature was gradually increased from 0 K to 300 K over 10,000 time-steps. The temperature was kept at 300 K for the rest of the 100 ps equilibration run. If the system target temperature was 400 K, it was heated from 300 to 400 K over the 10 ps period following the initial heating (over

picoseconds 10-20 of the equilibration run). During this time, the harmonic restraints mentioned above were enforced.

The trajectories were subsequently visualized using VMD²¹. The mutated residues were observed for differences in interactions between the germline and mature species.

Generation of local minima using Protein Local Optimization Program

In order to view differences in the energy landscape of possible H3 loop conformations in the four pairs of antibodies, sampled local minima were generated using the loop prediction algorithm in our Protein Local Optimization Program (PLOP). To better represent the antibodies in solution, we first allowed the crystal structures to relax without other molecules from crystal symmetry with an all-atom minimization using PLOP²². Titratable residues were placed in standard protonation states at pH 7.0 (Histidine is neutral.) We used a hierarchical approach to iteratively optimize loop conformations as described previously²³ but with increased sampling. In order to avoid artifacts due to holding nearby residues fixed (ex. blocking alternative loop conformations), side chains with a heavy atom $<7.5\text{\AA}$ from the loop were removed during the backbone buildup stage of the loop prediction and later optimized simultaneously with the loop residues. All samples generated during the hierarchical loop prediction run were collected and backbone (N-C α -C-O) RMSD's were calculated relative to the loop conformation with the lowest energy.

Hydrogen bonds analysis

The ptraj module in Amber was used to hydrogen bond partner distances. The distance cut-off heavy-atom to hydrogen distance cut-off was 2.5 Å. In cases where equivalent atoms existed i.e., a carboxylic acid, the lowest distance to either atom was taken.

Results and discussion

Molecular dynamics

Molecular dynamics simulations of AZ28²⁴, 7G12²⁵⁻²⁷, 28B4^{28,29} and 48G7³⁰⁻³² and their corresponding germline predecessors were performed to explore the role of flexibility in antibody maturation. The H3 CDR was the focus of the investigation because it is the most diverse CDR loop in sequence and length³³ and is believed to play a central role in specificity during maturation³³⁻³⁵. Its conformation is far more variable than other CDRs^{36,37} (Figure 2). So, as a first order measure of flexibility we focused on CDR C_α B-factors of H3 and found larger deviations for the germline H3 loop than its mature counterpart (Figure 1). Visualization of the trajectories made it possible to identify contacts formed only in the mature species and that appear to restrict CDR mobility. The four examples discussed here illustrate mechanism by which somatic mutations rigidify CDR's during maturations. The two mechanisms that arise are: 1) the formation of hydrogen bonds or salt bridges and 2) restriction via pi-pi interactions (side-chain packing). Zimmerman et al¹³ previously suggested these in their analysis of crystallographic structures. Here we see these mechanisms surfacing during the simulations.

7G12

MD simulations of the 7G12 antibody and its germline predecessor help explain how somatic mutations modulate flexibility. Two examples are: 1) the Ser76^HAsn mutation anchors the H1 loop via hydrogen bonds and 2) the Ser97^HMet mutation seems to hinder H3 by restraining Met97^H via sidechain packing interactions. In addition, the simulations illustrate the advantage of a flexible germline antibody. The H3 loop of the 7G12 germline predecessor is in equilibrium between its apo and holo conformations. The Ser97^HMet mutation anchors the H3 loop into the holo conformation in the mature antibody.

While Yin et al ²⁶ ignore the Ser76^HAsn mutation in their analysis, it substantially hinders the motion of the H1 loop in our calculations. Asn76^H in the mature species simultaneously hydrogen bonds with the alcohol group of Thr28^H and the backbone carbonyl of Tyr27^H (Figure 4). This does not occur in the germline case where a serine occupies that position. Germline H1 C α B-factors (up to ~440) were roughly 3 times the magnitude of its mature counterpart (~130); Figure 4 shows snapshots of the H1 loop, which clearly show the level of restriction. In addition, the probability density of the distance between the residue centroids at positions 76^H and 28^H clearly depicts a peak around 6.5 Å for the mature simulation which is absent in the germline result; the probability density in the 6–8 Å interval is roughly twice for the germline trajectory than the mature (Figure 3). If this model is correct: 1) a significant difference in H1 flexibility should be observed between the 7G12 antibody and its germline predecessor and 2) a

Thr28^HAla mutation would increase flexibility the mature species and make no difference in the germline antibody.

While the effect of the Ser76^HAsn mutation on binding is experimentally unknown, it may play a role since H1, in the germline simulation, swings to the binding pocket vicinity and interacts with nearby residues. This will have to be experimentally verified.

The simulations of the mature antibody are consistent with the interpretation of Yin et al²⁶, that the Ser97^HMet mutation anchors the H3 loop to its holo conformation, since the methionine contacts are firmly kept throughout the simulation. It appears that this mutation in fact reduces the flexibility of H3, as predicted. An unexpected observation was the transition from the apo to holo conformation of the H3 loop in the germline simulation, in the absence of ligand. Figure 5 shows the C_α RMSD to the bound and unbound conformations. The corresponding simulation of the mature species fails to display this behavior. This observation is striking because it suggests 7G12 binds through a conformational selection mechanism and the Ser97^HMet shifts the equilibrium towards the conformation most auspicious for binding. Conformational selection in antibodies has been proposed^{5,8,9,38,39} and, in some cases, experimentally verified^{4,8}, but, to our knowledge, it had not been observed in this type of simulation. Simulating these changes would be useful in predicting putative induced fit effects in antibodies.

28B4

The maturation of 28B4 involves the Asp95^HTrp mutation which, when reversed, lowers binding affinity by 3.7 kcal/mol²⁸ (more than half of the total binding free energy gained during maturation). Trp95^H is located at the base of the H3 loop and interacts with the bound ligand through pi-stacking interactions. However, our simulations suggest it may also restrict H3. H3, in the germline simulation, drifts into the space occupied the Trp95H side-chain in the mature structure and onto the binding pocket (Figure 6). The corresponding simulation of the affinity-matured antibody does not display this behavior. So, here we again observe tight sidechain packing as a mechanism to anchor H3.

A caveat of this observation is that perhaps, given enough simulation time, the mature H3 loop would explore the same conformational space spanned in the germline trajectory. A loop prediction calculation, which is not subject to this type of kinetic trapping, helped eliminate this possibility. The loop conformation from the germline simulation was grafted onto the mature antibody crystal structure, the residue at H95 mutated manually in the pdb file, and a loop prediction calculations, where loop C α atoms were kept within 2 Å of these coordinates was carried out. This calculation failed to find a set of loop coordinates that satisfied these restraints because the Trp sidechain could not be accommodated (even if nearby side-chains were removed). Therefore, it is unlikely the mature antibody H3 loop adopts this conformation or, at the very least, it clearly has fewer accessible states than its germline counterpart, which shifts the equilibrium toward the “bound” structure in the mature case.

The energy contribution to binding of this prearrangement effect versus the pi-pi stacking interaction of the Trp side-chain and ligand are difficult to deconvolute. But it is clear that this bulky non-polar group is able to very specifically shift the conformational equilibrium of H3.

AZ28

During AZ28 maturation, only one mutation occurs at a binding site residue (Ser34^LAsn)⁴⁰, and no structural changes are apparent from the crystallographic structures. In both cases the residue at position 34^L hydrogen bonds with the hydroxyl group of Tyr100^Ha⁴⁰, which sits at the base of the H3 loop. Yet, it does cause a binding affinity difference of 0.9 kcal/mol^{40,41} and our calculations show a drop in flexibility of H3. One notable difference is that a crystallographic water molecule mediates this interaction in the germline antibody, which may indicate a weaker interaction in this case.

The differences in dynamics are far more evident. The hydrogen bond to Tyr100^Ha is quickly lost in the germline trajectory (3% occupancy), while, in the mature case, Asn34^L interacts with Asp101^H (90% occupancy) and maintains its hydrogen bond with Tyr100^H (74% occupancy) (Figure 7). The hydrogen bond between Asp101^H and Tyr100^H is absent in the germline simulation (0.4% occupancy), while in the mature simulation, it creates an electrostatically auspicious environment for the tyrosine hydroxyl group where it can make two hydrogen bonds simultaneously..

This interaction is important because Tyr100^Ha sits at the base of the H3 loop and it is the mutation most likely to affect the H3 conformation⁴¹. In the germline simulation, the H3 loop folds over onto the space that would be occupied by the ligand upon binding as Tyr100^Ha makes contacts with other residues which permits this shift (Figure 7). Thus, this mutation likely contributes significantly to the flexibility differences in Figure 1. It also suggests a mutation of Asp101H to alanine would have an effect on the binding affinity of the mature antibody and have no effect on binding for its germline predecessor.

As in the 7G12 case, a hydrogen bond network appears responsible for anchoring H3. Sinha et al¹⁵ observed the formation of similar salt bridge networks during an MD simulation, which proved to play a role in binding and specificity. So, the formation of these electrostatic contacts, even is absent in the crystallographic coordinates, is not necessarily an artifact.

48G7

48G7 is an outlier in this analysis. It appears to mature differently than the other cases because: 1) it undergoes 9 mutations (Ser30^LAsn, Ser34^LGly, Asp55^LHis, Glu42^HLys, Gly55^HVal, Asn56^HAsp, Gly65^Hasp, Asn76^HLys and Ala78^HThr)³⁰ rather than 6 or less as in the other cases, 2) the ligand bound conformation differs between the mature and germline species, 3) none of the mutated residues directly interact with the ligand, and most of them only weakly affect the binding affinity³² and 4) there is cooperativity between pairs of mutations³¹, that is, combinations of mutations have a larger effect on

binding than the sum of their individual contributions. Priscilla et al^{31,32} suggest that initial mutations may change the ligand binding-pose and subsequently introduce new contacts. Thus, mutations that optimize those new contacts are only relevant if the previous mutations are present. Because of the latter observation, the analysis of the simulations performed here focused on those mutations that, alone, most affected the binding affinity.

H3 C α B-factors from the mature simulation are only slightly lower values than the corresponding germline data (Figure 1). Most of the mutations induced no new contacts; many of their side-chains simply diffused through solvent during the simulation. The higher rigidity of the H3 loop in the affinity-matured antibody appears to be due largely to the Asp55^LHis mutation, which causes the loss of a salt bridge between 55^H and Arg46^L; the latter is then free to interact with the backbone of H3. The Asn76^HLys mutation actually increases the flexibility of H1³⁰. This flexibility, as discussed by Wedemeyer et al³⁰, permits the rearrangements necessary to optimize contacts with the ligand. Interestingly, this is consistent with the effect of Ser76^HAsn in the maturation of 7G12 and Ser34^LAsn in AZ28. That is, asparagine appears to be able to restrict motion by making two electrostatic interactions simultaneously. In the case of 48G7, its removal causes precisely the expected effect and increases flexibility.

A previous molecular dynamics study on the germline and mature species of this antibody concluded the germline complex is more flexible than its mature counterpart⁴². In that study the simulation included the ligand and was much shorter (less than a

nanosecond). At shorter simulation times (< 2 ns), we find a larger gap between the germline and mature H3 flexibility, so it is unclear if their conclusion would hold at longer simulation times and in the absence of ligand.

H3 loop local-minima energy landscape analysis

To further investigate H3 loop flexibility, we generated local minima for the H3 loops in the same germline and mature antibodies using a different sampling method and a different force field (see Methods). While in general, molecular dynamics explores low-energy basins and has difficulty crossing high energy barriers, loop enumeration methods can hop from basin to basin freely (at the expense of generating a true statistical ensemble.) We aimed to analyze qualitative differences between the germline and mature antibody energy landscapes at a large scale as an alternative to the more local fluctuations captured by MD. In Figure 8, energy landscapes are generally different between the antibody pairs. 7G12 shows a marked focusing of the energy well, though in the mature antibody there is one minima outlier ~ 8 kcal/mol higher than the global sampled minimum. The mature 28B4 and AZ28 antibodies show a similar reduction in the number of minima < 10 kcal/mol from the lowest energy antibody. The mature 48G7 antibody shows a slight increase in local minima $< 1\text{\AA}$ from the lowest energy minimum. This contradicts the slight decrease in B-factor in the molecular dynamics but does concur with our earlier conclusion that 48G7 matures differently in comparison to the other antibody pairs.

Conclusion

While the set of cases is too small to make generalizations, it was possible to rationalize how mutations, sometimes not at the active site (Ser76^HAsn in 7G12), restrict CDR mobility. Mutations to asparagine appear to have a restricting effect, especially since reverse mutations caused an increase in flexibility (48G7). Restrictions due to bulky side chains (Asp95^HTrp of 28B4) can also play a role. This type of mutation long-range effects are evident in HIV-1 protease drug resistant mutants, where mutations not in the active site affect binding⁴³ and, interestingly, seem to modulate the binding site flexibility^{44,45}.

An unexpected and interesting result from our simulations was the equilibrium between bound and unbound H3 conformations in the absence of ligand (7G12 germline), which is consistent with a conformational selection mechanism for binding. These results suggest a general model where germline antibodies span a broad conformational space and somatic mutations that introduce multiple hydrogen bonds or tight sidechain packing anchor them into conformations auspicious for binding.

Here we focused on rigidification of CDR loops during maturation, however, other mechanisms for affinity maturation exist; 48G7 appears to mature via an alternative route and experiments identified others. Sethi et al⁴⁶ found a germline antibody that 1) bound several peptides, but in different conformations and 2) its maturation involved disruption of binding contacts except for those involving the target antigen.

A further step would be to investigate the contribution of both entropy and enthalpy to the change in binding free energy upon maturation. Mutations that lead to rigidifying antibodies reduce the conformational entropy penalty upon binding. But these mutations can also lead to more favorable enthalpy upon binding. If the enthalpic change is greater than the entropic change, rigidification of antibody variable regions may be less important than previously thought. For example, Torigoe et al.⁴⁷ calculated these contributions using isothermal titration calorimetry and found that maturation lead to a predominant enthalpic reduction. More studies would need to be carried out in order to form a general theory, however. In addition, calculating the enthalpic and entropic contributions would be very difficult to do rigorously in a molecular dynamics simulation. Finally, the entropic changes of water molecules near the binding site complicate the question.

References:

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell*. New York, NY: Garland Science; 2002.
2. Huston JS, George AJ. Engineered antibodies take center stage. *Hum Antibodies* 2001;10(3-4):127-142.
3. Holliger P, Hudson PJ. Engineered antibody fragments and the rise of single domains. *Nat Biotechnol* 2005;23(9):1126-1136.
4. Foote J. Isomeric antibodies. *Science* 2003;299(5611):1327-1328.
5. Foote J, Milstein C. Conformational Isomerism and the Diversity of Antibodies. *Proceedings of the National Academy of Sciences of the United States of America* 1994;91(22):10370-10374.
6. Notkins aL. Polyreactivity of antibody molecules. *Trends in Immunology* 2004;25(4):174-179.
7. Nguyen HP, Seto NOL, MacKenzie CR, Brade L, Kosma P, Brade H, Evans SV. Germline antibody recognition of distinct carbohydrate epitopes. *Nature Structural Biology* 2003;10(12):1019-1025.
8. James LC, Tawfik DS. Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition. *Proceedings of the National Academy of Sciences of the United States of America* 2005;102(36):12730-12735.
9. Berger C, Weber-Bornhauser S, Eggenberger J, Hanes J, Pluckthun A, Bosshard HR. Antigen recognition by conformational selection. *Febs Letters* 1999;450(1-2):149-153.
10. Manivel V, Sahoo NC, Salunke DM, Rao KVS. Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity* 2000;13(5):611-620.
11. Kroon GJ, Mo H, Martinez-Yamout MA, Dyson HJ, Wright PE. Changes in structure and dynamics of the Fv fragment of a catalytic antibody upon binding of inhibitor. *Protein Sci* 2003;12(7):1386-1394.
12. Renisio JG, Perez J, Czisch M, Guennegues M, Bornet O, Frenken L, Cambillau C, Darbon H. Solution structure and backbone dynamics of an antigen-free heavy chain variable domain (VHH) from Llama. *Proteins* 2002;47(4):546-555.
13. Zimmermann J, Oakman EL, Thorpe IF, Shi X, Abbyad P, Brooks CL, 3rd, Boxer SG, Romesberg FE. Antibody evolution constrains conformational heterogeneity by tailoring protein dynamics. *Proc Natl Acad Sci U S A* 2006;103(37):13722-13727.
14. Thorpe IF, Brooks CL, 3rd. Molecular evolution of affinity and flexibility in the immune system. *Proc Natl Acad Sci U S A* 2007;104(21):8821-8826.
15. Sinha N, Li Y, Lipschultz CA, Smith-Gill SJ. Understanding antibody-antigen associations by molecular dynamics simulations: detection of important intra- and inter-molecular salt bridges. *Cell Biochem Biophys* 2007;47(3):361-375.
16. D.A Case TAD, T.E. Cheatham, III, L.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S Brozell, V. Tsui, H.

- Gohlke, J.Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, and P.A. Kollman. AMBER 8; 2004.
17. Pastor RW, Brooks BR, Szabo a. An Analysis of the Accuracy of Langevin and Molecular-Dynamics Algorithms. *Molecular Physics* 1988;65(6):1409-1419.
 18. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins-Structure Function And Bioinformatics* 2004;55(2):383-394.
 19. Loncharich RJ, Brooks BR, Pastor RW. Langevin Dynamics of Peptides - the Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-Methylamide. *Biopolymers* 1992;32(5):523-535.
 20. Simmerling C, Strockbine B, Roitberg AE. All-atom structure prediction and folding simulations of a stable protein. *Journal Of The American Chemical Society* 2002;124(38):11258-11259.
 21. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* 1996;14(1):33-&.
 22. Zhu K, Shirts MR, Friesner RA, Jacobson MP. Multiscale Optimization of a Truncated Newton Minimization Algorithm and Application to Proteins and Protein-Ligand Complexes. *J Chem Theory Comput* 2007;3(2):640-648.
 23. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. *Proteins-Structure Function and Bioinformatics* 2004;55(2):351-367.
 24. Mundorff EC, Hanson MA, Varvak A, Ulrich H, Schultz PG, Stevens RC. Conformational effects in biological catalysis: an antibody-catalyzed oxy-cope rearrangement. *Biochemistry* 2000;39(4):627-632.
 25. Romesberg FE, Santarsiero BD, Spiller B, Yin J, Barnes D, Schultz PG, Stevens RC. Structural and kinetic evidence for strain in biological catalysis. *Biochemistry* 1998;37(41):14404-14409.
 26. Yin J, Andryski SE, Beuscher AE, Stevens RC, Schultz PG. Structural evidence for substrate strain in antibody catalysis. *Proceedings of the National Academy of Sciences of the United States of America* 2003;100(3):856-861.
 27. Yin J, Beuscher AE, Andryski SE, Stevens RC, Schultz PG. Structural plasticity and the evolution of antibody affinity and specificity. *Journal of Molecular Biology* 2003;330(4):651-656.
 28. Yin J, Mundorff EC, Yang PL, Wendt KU, Hanway D, Stevens RC, Schultz PG. A comparative analysis of the immunological evolution of antibody 28B4. *Biochemistry* 2001;40(36):10764-10773.
 29. HsiehWilson LC, Schultz PG, Stevens RC. Insights into antibody catalysis: Structure of an oxygenation catalyst at 1.9-angstrom resolution. *Proceedings of the National Academy of Sciences of the United States of America* 1996;93(11):5363-5367.
 30. Wedemayer GJ, Stevens RC. Structural insights into the evolution of an antibody combining site (vol 276, pg 1665, 1997). *Science* 1997;277(5331):1423-1423.
 31. Priscilla Y, Schultz PG. Mutational Analysis of the Affinity Maturation of Antibody 48G7. *Journal of Molecular Biology* 1999;294:1191-1201.

32. Patten PA, Gray NS, Yang PL, Marks CB, Wedemayer GJ, Boniface JJ, Stevens RC, Schultz PG. The immunological evolution of catalysis. *Science* 1996;271(5252):1086-1091.
33. Davis MM, Boniface JJ, Reich Z, Lyons D, Hampl J, Arden B, Chien YH. Ligand recognition by alpha beta T cell receptors. *Annual Review of Immunology* 1998;16:523-+.
34. Arden B. Conserved motifs in T-cell receptor CDR1 and CDR2: implications for ligand and CD8 co-receptor binding. *Current Opinion in Immunology* 1998;10:74-81.
35. Xu J, Davis M. Diversity in the CDR3 Region of VH is Sufficient for Most Antibody Specificities. *Immunity* 2000;13:37-45.
36. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *Journal of Molecular Biology* 1998;275(2):269-294.
37. Chothia C, Lesk aM, Tramontano a, Levitt M, Smithgill SJ, Air G, Sheriff S, Padlan Ea, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, Poljak RJ. Conformations of Immunoglobulin Hypervariable Regions. *Nature* 1989;342(6252):877-883.
38. Bosshard HR. Molecular recognition by induced fit: How fit is the concept? *News in Physiological Sciences* 2001;16:171-173.
39. James LC, Tawfik DS. Conformational diversity and protein evolution - a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences* 2003;28(7):361-368.
40. Mundorff EC, Hanson MA, Varvak A, Ulrich E, Schultz PG, Steens RC. Conformational Effects in Biological Catalysis: An Antibody-Catalyzed Oxy-Cope Rearrangement. *Biochemistry* 2000;39(4):627-632.
41. Ulrich HD, Mundroff E, Santarsiero BD, Driggers EM, Stevens RC, Schultz PG. The interplay between binding energy and catalysis in the evolution of a catalytic antibody. *Nature* 1997;389(6648):271-275.
42. Chong L, Duan Y, Wang L, Massova I, Kollman P. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proceedings of the National Academy of Sciences of the United States of America* 1999;96(25):14330-14335.
43. Lin Y, Lin X, Hong L, Foundling S, Heinrikson RL, Thaisrivongs S, Leelamanit W, Raterman D, Shah M, Dunn BM, et al. Effect of point mutations on the kinetics and the inhibition of human immunodeficiency virus type 1 protease: relationship to drug resistance. *Biochemistry* 1995;34(4):1143-1152.
44. Lauria A, Ippolito M, Almerico AM. Molecular dynamics studies on HIV-1 protease: a comparison of the flap motions between wild type protease and the M46I/G51D double mutant. *J Mol Model* 2007;13(11):1151-1156.
45. Meiselbach H, Horn A, Harrer T, Sticht H. Insights into amprenavir resistance in E35D HIV-1 protease mutations from molecular dynamics and binding free-energy calculations. *J Mol Model* 2007;13:297-304.
46. Sethi DK, Agarwal A, Manivel V, Rao KVS, Salunke DM. Differential epitope positioning within the germline antibody paratope enhances promiscuity in the primary immune response. *Immunity* 2006;24(4):429-438.

47. Torigoe H, Nakayama T, Imazato M, Shimada I, Arata Y, Sarai A. The Affinity Maturation of Anti-4-hydroxy-3-nitrophenylacetyl Mouse Monoclonal Antibody. *J Biol Chem* 1995;270(38):22218-22222.

Chapter 4: Table 1

Antibody	Germline Structure PDB code	Mature Structure PDB code	CDR region residues
7G12	1NGZ (L1-110, H1-110)	1NGY (L1-110, H1-110)	L1: L23-34 L2: L50-56 L3: L88-97 H1: H26-33 H2: H50-58 H3: H99-103
28B4	1FL5 (L1-112, H1-113)	1KEM (L1-112, H1-117)	L1: L24-40 L2: L54-61 L3: L94-102 H1: H26-35 H2: H52-56 H3: H101-108
AZ28	1D5I (L1-110, H1-110)	1D5B (L1-110, H1-110)	L1: L24-34 L2: L50-56 L3: L89-97 H1: H26-35 H2: H52-58 H3: H95-102
48G7	2RCS (L1-110, H1-110)	1HKL (L1-110, H1-110)	L1: L24-33 L2: L48-55 L3: L89-97 H1: H26-35 H2: H51-56 H3: 95-103

Table 1. Structural information for simulations setup. PDB structure codes and specific chain segments used in the simulation for germline and mature antibodies are listed. In addition, the assignment of CDR residues is explicitly shown.

Chapter 4: Table 2

Name	Temp (K)	Average B-factor difference (germline - mature)										
		L1	L2	L3	H1	H2	H3					
48G7	300	-32	-5	-16	-21	23	4					
28B4	300	2	6	-5	2	3	14					
AZ28	300	2	-3	2	19	-38	22					
7G12	300	-3	0	1	18	-98	-2					
48G7	400	12	-282	13	118	36	20					
28B4	400	9	278	2	8	14	-1					
AZ28	400	-13	-17	5	-92	-25	73					
7G12	400	-14	-13	-20	-29	-92	44					

Table 2. Average C α B-factor differences (germline-mature) calculated from simulations at 300 and 400 K. Raw data of the individual residues is available in supplementary materials.

Chapter 4: Figure 1

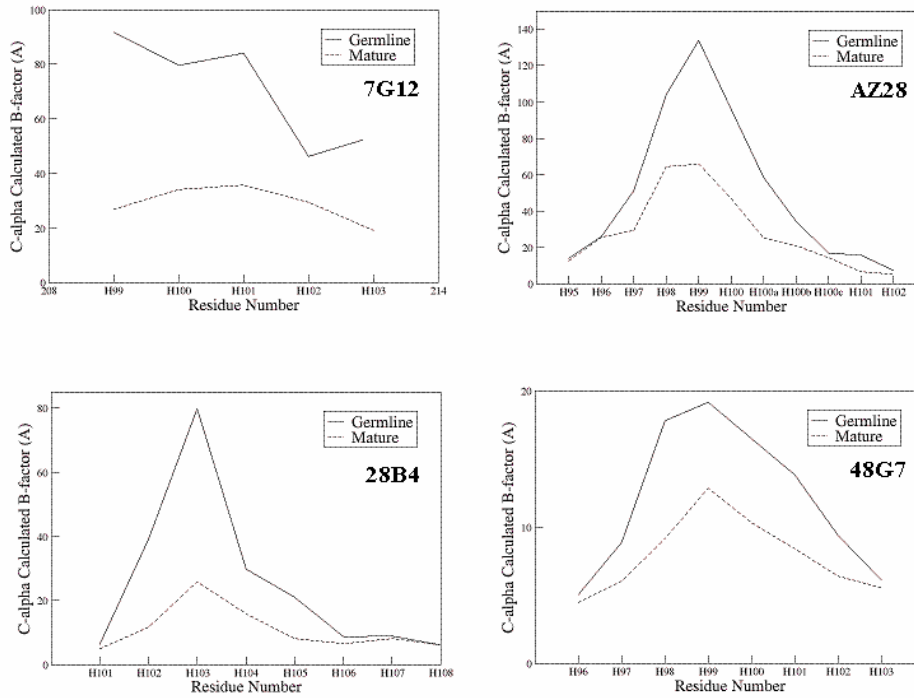


Figure 1. H3 loop Ca B-factors calculated from molecular dynamics trajectories. The germline species consistently yields higher, sometimes by a large margin, B-factors than the mature simulation.

Chapter 4: Figure 2

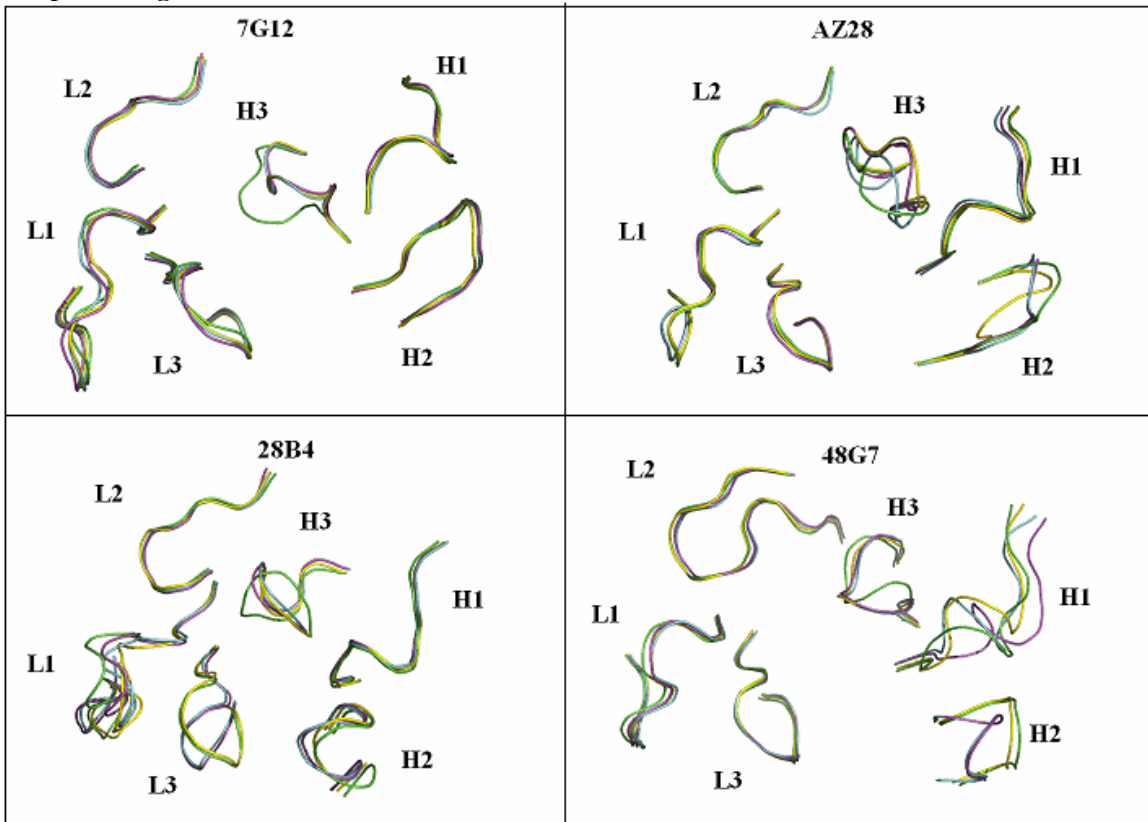


Figure 2. CDR loops for bound and unbound germline and mature antibodies. The bound germline and mature structures are colored yellow and magenta while the free germline and mature structures are colored green and cyan, respectively. It is most clear in the H3 loops that the unbound germline structure (green) is significantly different than the others.

Chapter 4: Figure 3

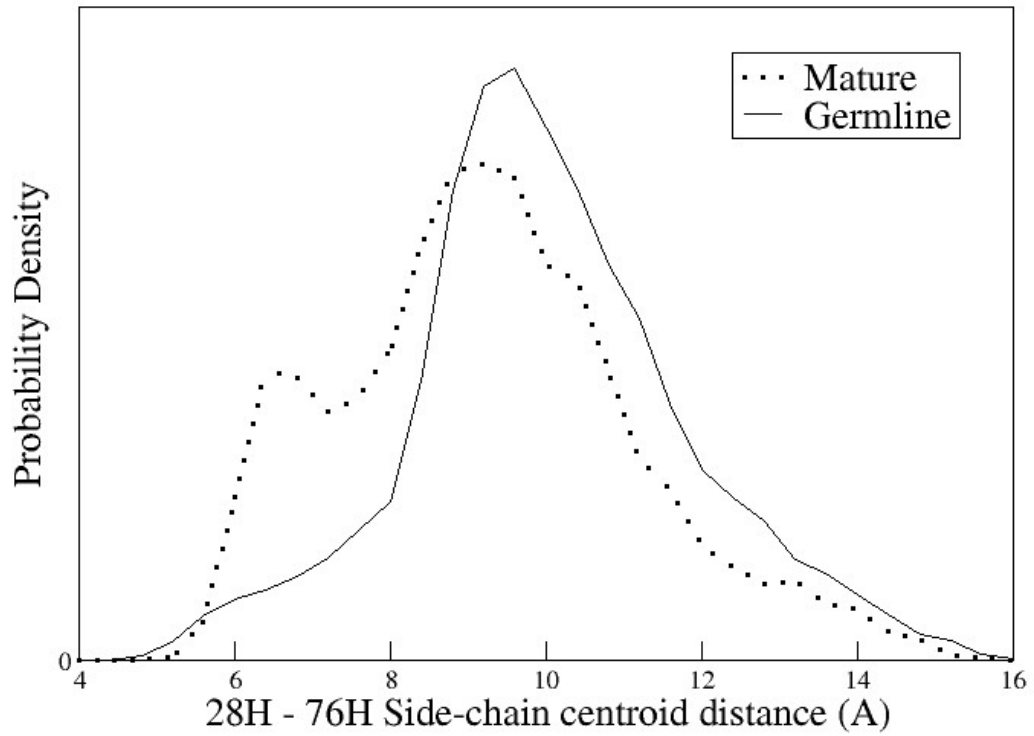


Figure 3. Probability density function of distances between residue centroid at positions 28H and 76H of the germline and mature 7G12 antibody.

Chapter 4: Figure 4

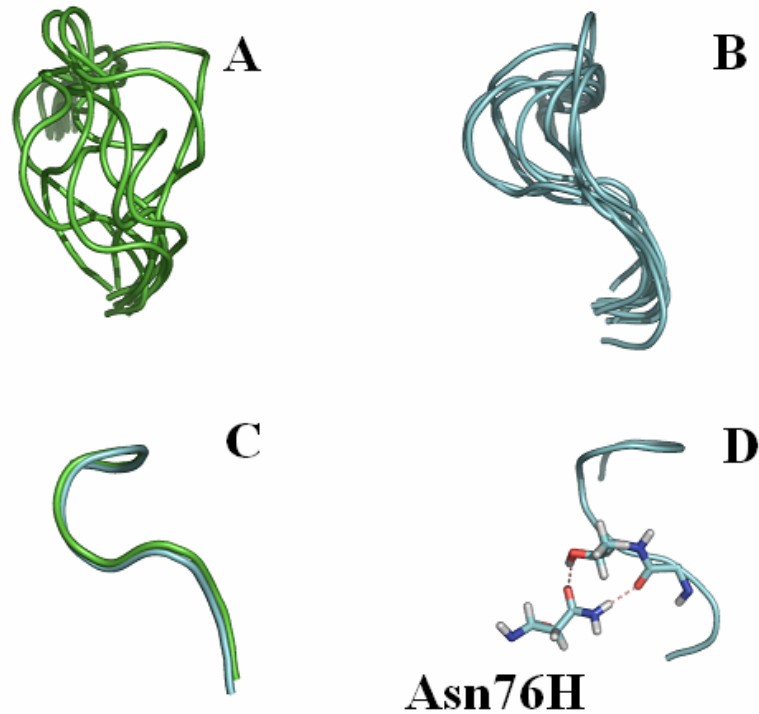


Figure 4. 7G12 H1 loop flexibility is diminished by Ser76^HAsn mutation in ways not obvious from crystallographic structures. A. H1 loop snapshots from 7G12 germline predecessor antibody simulation. B. H1 loop snapshots from 7G12 mature antibody simulation. C. H1 crystal structures of the 7G12 mature antibody and its germline predecessor. The crystallographic structures are nearly identical. D. Asn76H hydrogen bonds H1 and restricts its motion in the mature species.

Chapter 4: Figure 5

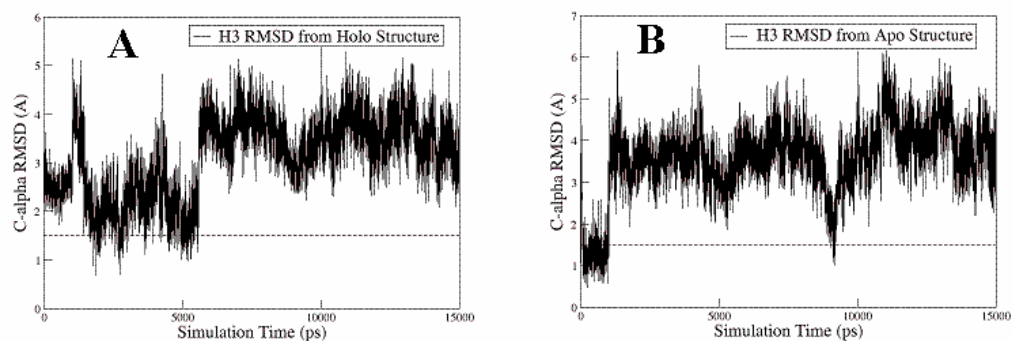


Figure 5. 7G12 germline H3 loop visits bound and unbound conformations during simulation. A. H3 loop RMSD from the germline holo structure as a function of simulation time. B. H3 loop RMSD from the germline apo structure as a function of simulation time. The dotted line marks a 1.5 Å RMSD.

Chapter 4: Figure 6



Figure 6. Simulation of the H3 loop in the germline antibody (green) explores conformational spaces occupied by a bulky Trp95H sidechain in the mature species (green). The latter is the result of a somatic mutation during maturation. The bulky Trp sidechain restricts the available conformational space of H3.

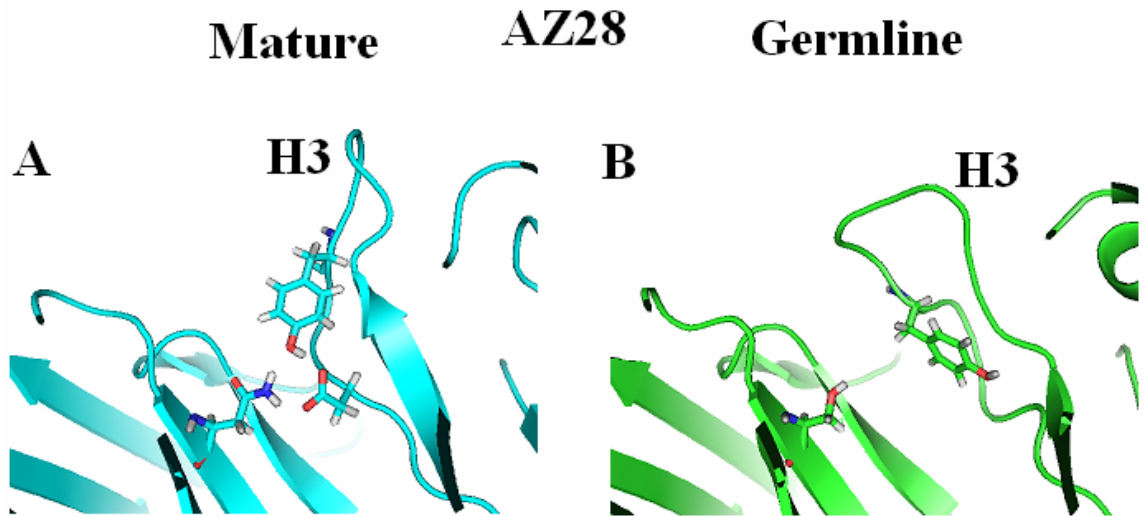
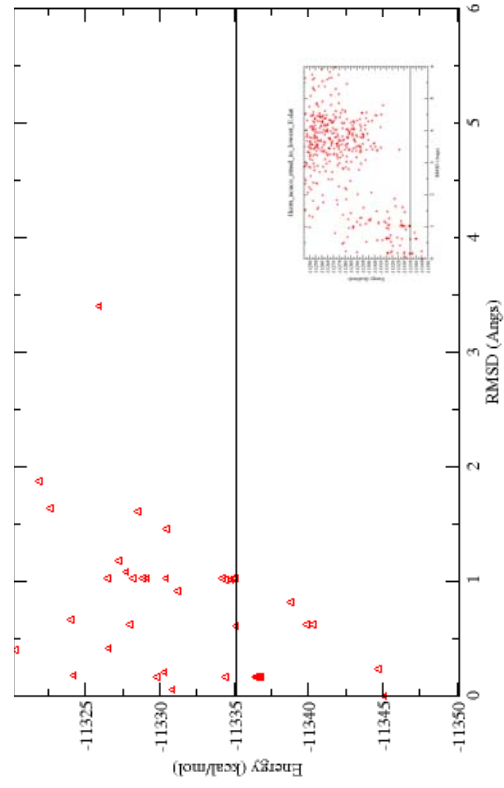
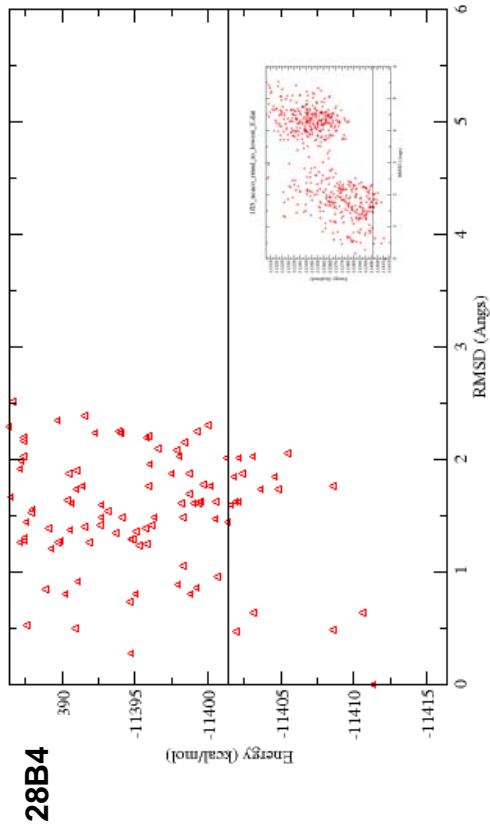
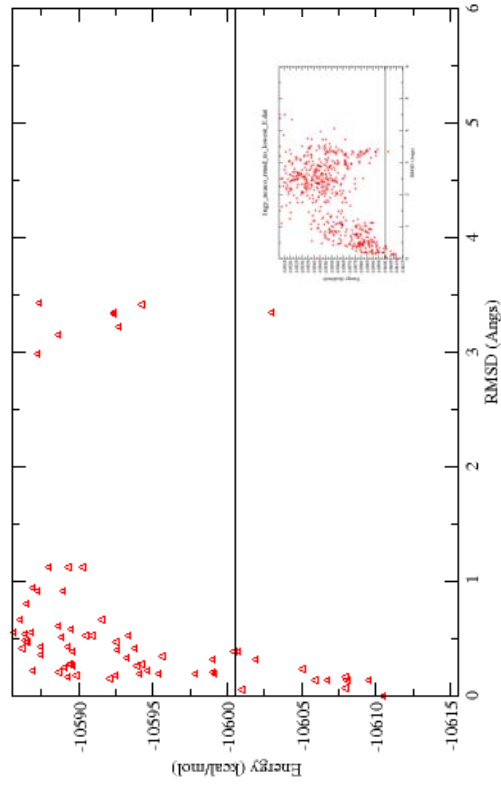
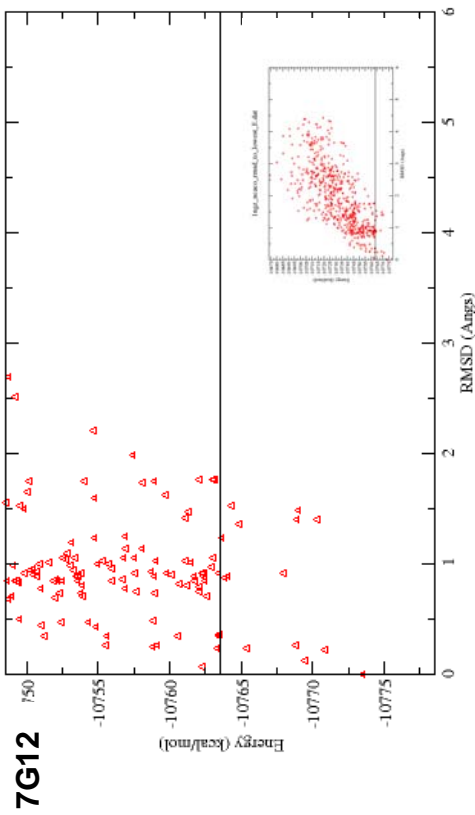


Figure 7. Simulation dynamics differences between AZ28 and its germline predecessor. A. Simulation snapshot showing the hydrogen bond contacts of Asn34L of the mature AZ28 antibody. Asn34L makes contacts with Asp101H and His100aH. B. Simulation snapshot showing the lack of a hydrogen bond between Ser34L and His100aH in the germline species simulation. These snapshots are representative of the entire simulation.

Chapter 4: Figure 8

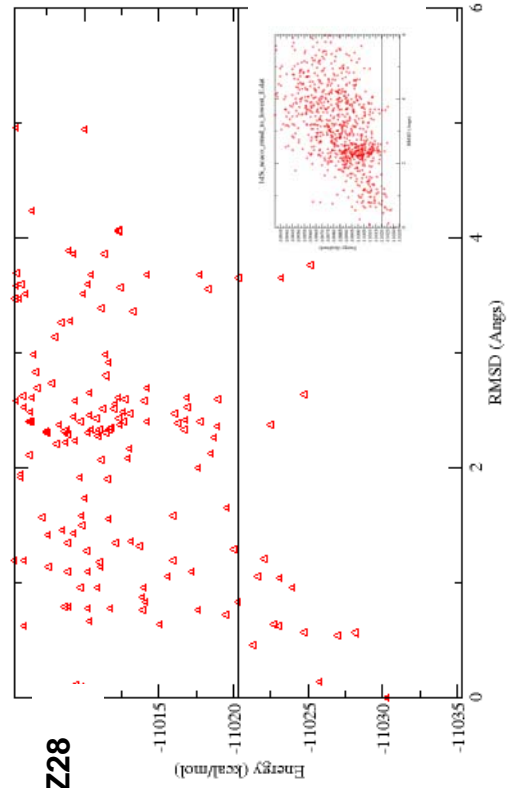
Germline

Mature

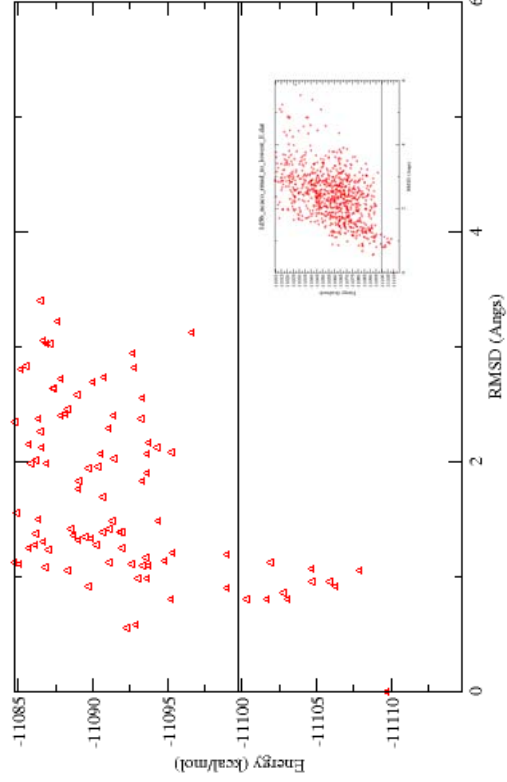


Germline

AZ28



Mature



48G7

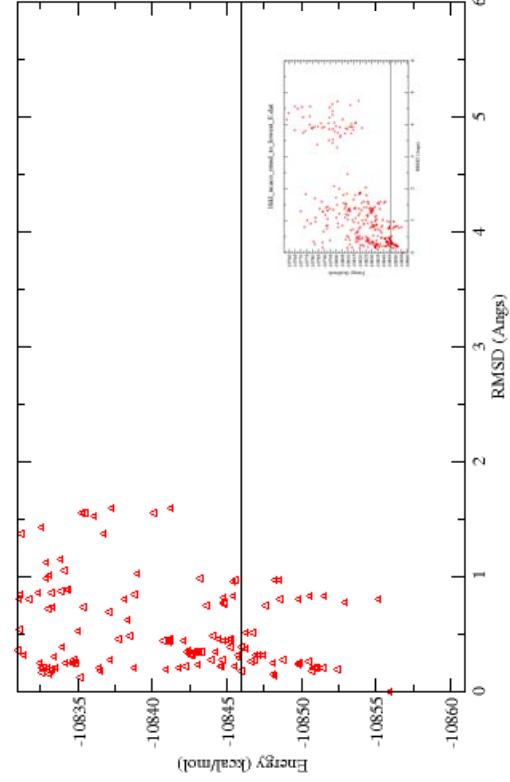
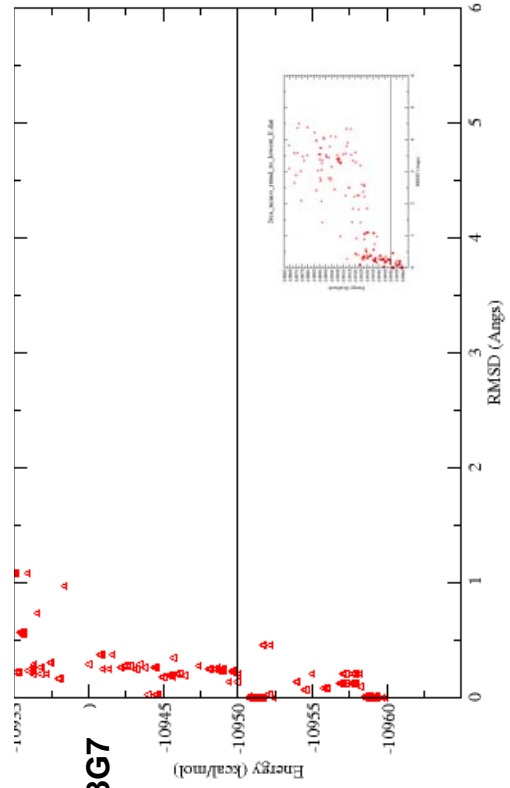


Figure 8 Plots of energy (kcal/mol) versus RMSD for H3 loop samples generated using loop minima enumeration (see Methods). Each red triangle represents a single loop conformation at an local energy minimum. Left column: germline antibodies. Right column: mature antibodies. Large plots show samples <25 kcal/mol from lowest energy sampled. Line indicates 10 kcal/mol above lowest energy sample for reference. For qualitative scale, inset plots show samples < 100 kcal/mol from lowest energy sample. RMSD's are calculated between H3 backbone atoms (N-C α -C-O) and are relative to the conformation of the lowest energy sample.

Chapter 5: All-atom model of clathrin hub using comparative modeling and electron-microscopy data enables X-ray crystallographic solution by molecular replacement

Benjamin D. Sellers¹, Jeremy Wilbur¹, Matthew P. Jacobson²

¹Graduate Group in Biophysics, University of California, San Francisco, California

²Department of Pharmaceutical Chemistry, University of California, San Francisco, California

Summary

Clathrin is a key structural protein in membrane trafficking which forms a large, three-legged triskelion. These clathrin trimers, along with adaptor proteins create large cage-like structures leading to clathrin-coated pits and vesicles found in endocytosis and membrane separation from organelles. Crystal structures of a portion of the proximal leg and beta-propeller terminal domain were reported previously^{1,2} as well as a 7.9 Å resolution cryo-electron microscopy structure of the complete clathrin trimer³. Questions remained however, regarding how clathrin is regulated, particularly by the flexible clathrin light chain which has been shown to inhibit clathrin assembly⁴⁻⁶. The Brodsky and Fletterick labs successfully crystallized a clathrin hub construct (trimerization domain plus proximal leg) with the clathrin light chain. However, several attempts failed to find a solution by molecular replacement using the previously published crystal structure of the proximal leg. Here, we report methods for creating a complete, all-atom model of the clathrin hub which lead to a successful solution by molecular replacement. The model was assembled by aligning comparative models of each clathrin heavy chain repeat (CHCR) to the previously published cryo-EM model which only contained Ca coordinates. The trimerization domain helices were also modeled and all loops between CHCR's were refined. The resulting 8.3 Å X-ray crystal solution contains novel atomic structure including multiple clathrin light chain conformations in the unit cell which are shown to be functional through biochemical analysis. This work is part of a larger manuscript by Wilbur et al⁷.

Methods

We constructed an all-atom model of the clathrin trimer hub for molecular replacement as follows. Since the previously published³ C α model was derived from low-resolution EM data (7.9 Å), we chose to build comparative models of each CHCR segment using the higher-resolution (2.6 Å) crystal structure, PDB: 1B89¹, as a template. CHCR models were generated with the homology modeling function in the Protein local Optimization Program (PLOP)⁸ using a previously published sequence alignment CHCR's¹. A subset of the alignment is shown in Figure 1. Each CHCR model was then aligned to the C α coordinates in the EM structure, PDB 1XI4³, using the protein alignment function in Chimera⁹ (see Figure 2). Loops joining CHCR segments were refined using our Physics-based, loop prediction method¹⁰ resulting in a single clathrin leg. The two loops and two helices, residues 1575-1630, found in the trimerization domain remained to be modeled. We constructed the two helices by building two small homology models using Prime (Schrodinger,Inc.) based on an arbitrary PDB structure containing a coiled coil (PDB 1VDF¹¹). The remaining loops connecting these helices to the rest of the proximal leg, were predicted *ab initio* using our Protein Local Optimization Program. The complete leg was then copied and aligned to the EM C α coordinates to form a trimer model.

References

1. Ybe JA, Brodsky FM, Hofmann K, Lin K, Liu SH, Chen L, Earnest TN, Fletterick RJ, Hwang PK. Clathrin self-assembly is mediated by a tandemly repeated superhelix. *Nature* 1999;399:371-375.
2. ter Haar E, Musacchio A, Harrison SC, Kirchhausen T. Atomic Structure of Clathrin: A Propeller Terminal Domain Joins an Zigzag Linker. *Cell* 1998;95:563-573.
3. Fotin A, Cheng Y, Sliz P, Grigorieff N, Harrison SC, Kirchhausen T, Walz T. Molecular model for a complete clathrin lattice from electron cryomicroscopy. *Nature* 2004;432:573-579.
4. Ungewickell E, Ungewickell H, Holstein SEH, Lindner R, Prasad K, Barouch W, Martini B, Greene LE, Eisenberg E. Role of auxilin in uncoating clathrin-coated vesicles. *Nature* 1995;378(6557):632-635.
5. Liu SH, Wong ML, Craik CS, Brodsky FM. Regulation of clathrin assembly and trimerization defined using recombinant triskelion hubs. *Cell* 1995;83(2):257-267.
6. Ybe JA, Greene B, Liu SH, Pley U, Parham P, Brodsky FM. Clathrin self-assembly is regulated by three light-chain residues controlling the formation of critical salt bridges. *The EMBO Journal* 1998;17:1297-1303.
7. Jeremy Wilbur PH, Michael Lane, Joel Ybe, Benjamin D. Sellers, Matthew P. Jacobson, Robert Fletterick, Frances Brodsky. Regulation of clathrin lattice assembly by conformational switching in clathrin light chain. *Nature* to be submitted.
8. Kenyon V, Chorny I, Carvajal WJ, Holman TR, Jacobson MP. Novel human lipoxxygenase inhibitors discovered using virtual screening with homology models. *J Med Chem* 2006;49(4):1356-1363.
9. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 2004;25(13):1605-1612.
10. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. *Proteins* 2004;55(2):351-367.
11. Malashkevich VN, Kammerer RA, Efimov VP, Schulthess T, Engel J. The Crystal Structure of a Five-Stranded Coiled Coil in COMP: A Prototype Ion Channel? *Science* 1996;274(5288):761.

Chapter 5: Figure 1

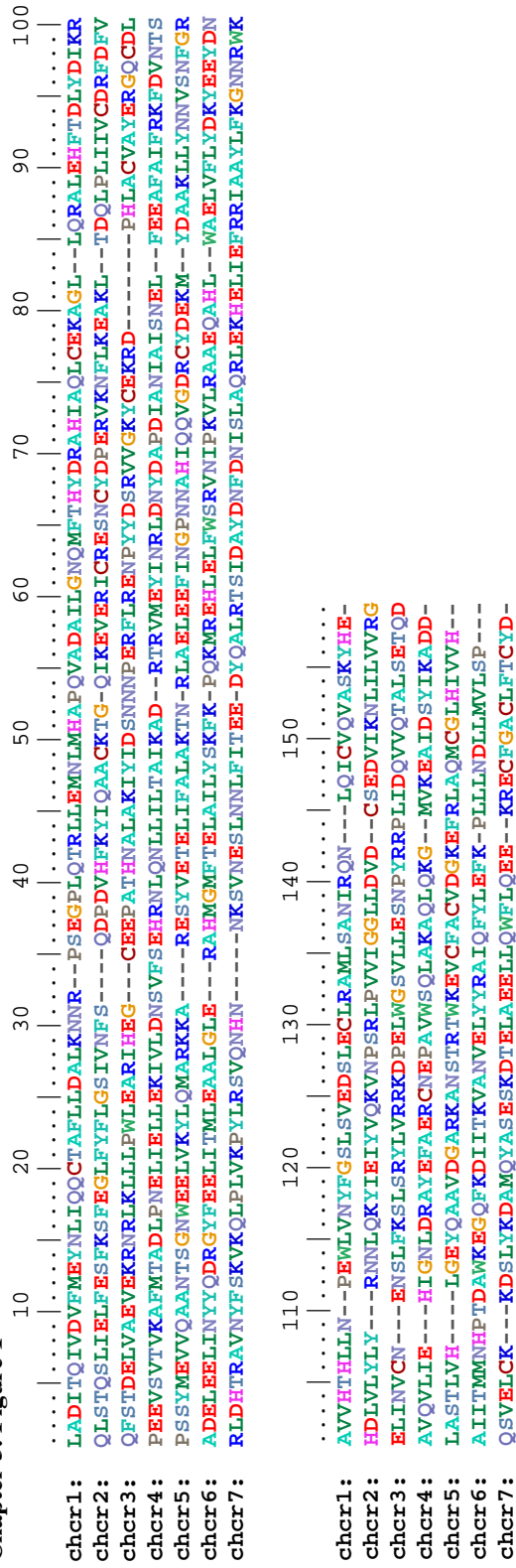


Figure 1 caption
Multiple alignment of CHCR sequences as derived from previous multi-species alignment (see Methods.)

Chapter 5: Figure 2

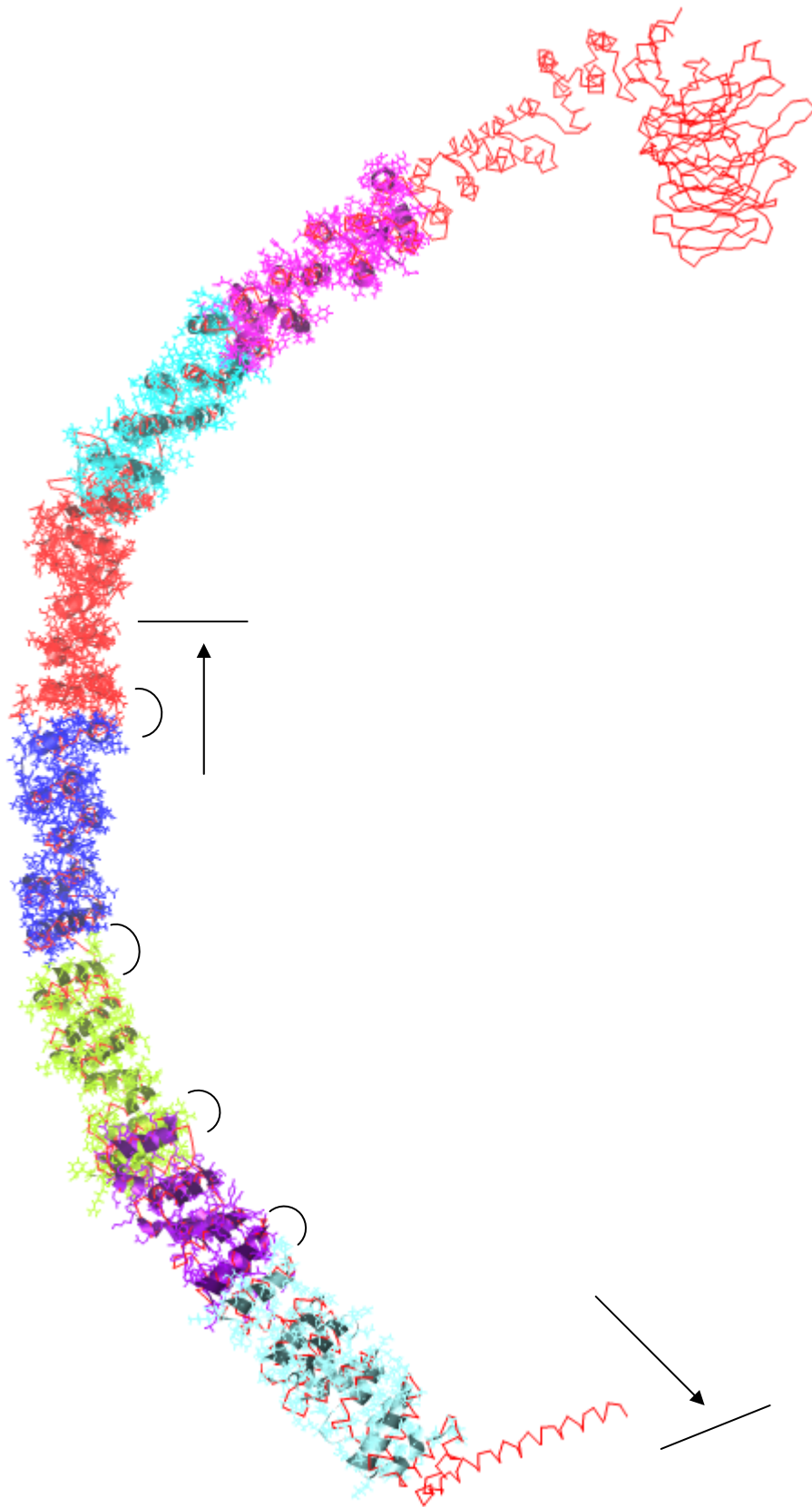


Figure 2 caption

Overlay of CHCR comparative models (multi-colored) onto previously published $C\alpha$ model derived from EM data (depicted as red wire). CHCR's 1-7 are depicted in pink, light blue, red, dark blue, yellow, purple, and cyan respectively. Arrows delineate the clathrin hub, residues 1074-1675. Black arc segments show inter-CHCR loops that were refined. Atomic representation generated using Chimera, UCSF.

Chapter 6: Investigation of pH-dependent inhibition of membrane-type serine protease 1 by a pico-Molar antibody inhibitor using constant-pH molecular dynamics simulation

Benjamin D. Sellers¹, Christopher J. Farady¹, Matthew P. Jacobson²

¹Graduate Group in Biophysics, University of California, San Francisco, California

²Department of Pharmaceutical Chemistry, University of California, San Francisco, California

Summary

Membrane-type serine protease 1 (MT-SP1) or Matriptase is a human type II transmembrane serine protease expressed in human epithelial cells. Though its multiple roles are not fully understood, MT-SP1 is known to be integral to growth factor and growth factor receptor regulation¹ as well as protease activation² leading to shedding of the extracellular matrix. Over-expression of MT-SP1 has been implicated in tumor growth and metastasis in several forms of cancer³⁻⁸ which suggests it may be a possible drug target. Recently, Farady et al.⁹ characterized two potent ($K_i = 8$ and 140 pM) single chain antibody (scFv) inhibitors of MT-SP1 and recently crystallized the 8 pM E2 antibody antigen binding fragment (Fab)¹⁰. Farady et al also found that MT-SP1 cleaves E2 in a substrate-like manner at pH 6 while E2 remains intact at pH 8. This pH dependence is common among cognate inhibitors of serine proteases but is not understood at the molecular level^{11,12}. We hypothesize that E2 exhibits a different conformation at low pH which enables the protease to cleave E2. Here, we describe unpublished, initial results using constant-pH molecular dynamics of the E2 / MT-SP1 complex at pH 8 and pH 6. The pH 8 complex is stable for the 2 nanosecond simulation while the pH 6 complex begins to dissociate after 1 nanosecond and separates further up to 2 nanoseconds. Though a drastic difference is observed, more sampling is needed. Future simulations should include a structured water molecule to prevent a presumably-artificial collapse of a key arginine into the P1 pocket of the protease. Though additional work is needed, this work presents a possible, molecular-level, hypothesis answering why there is pH-dependent catalysis. At this time, there is no plan to publish this work.

Methods

Simulation

The E2-MT-SP1 crystal structure was obtained from the Craik lab before publication and E2 was truncated to variable fragment (Fv) length to reduce the number of atoms in the simulation. The constant pH simulation, using Amber 9 with the implicit solvent model, Generalized Born with Surface Area (GBSA)¹³, at pH 6 and pH 8, included the following stages:

1. *minimization* in five stages with Cartesian harmonic restraints. Weights used were 25.0, 5.0, 4.0, 3.0, 2.0, and 1.0, respectively.

2. *equilibration* in 11 stages of 1000 steps each. Temperature increased from 0-300K in first three stages with $C\alpha$ restraint weights of 1.0, 0.8, and 0.5. Seven more equilibration stages reduce $C\alpha$ restraint weights from 0.4 to 0.0. Finally, a pH equilibration stage of 1024 steps was run at pH 6 and pH 8 for the two MD calculations presented here.

3. *production* simulation was run for 2 ns for each case.

The simulations were run at pH 6 and 8 to compare to experiments by Farady et al⁹. (The protease is most efficient at pH 8). Analysis and trajectory viewing were carried out using Amber 9¹³ and Chimera¹⁴.

Choice of titratable residues

Histidine, aspartic acid, and glutamic acid residues in the interface were considered for titration during the constant-ph simulation. The following residues were included in the titration: His42, Asp47, Asp91, Asp97, His138, Asp185, Asp190, Asp214, His332.

Results

Artifacts of equilibration and implicit solvent

The use of an implicit solvent model is required at this time for running constant-pH simulations but neglects effects of discrete water molecules. In this complex, a structured water is seen in the crystal structure in the P1 pocket which is not included in the present study. The H3 loop of E2 collapses into the P1 pocket at the end of equilibration (see Figure 1). The effect of this is not clear without further simulations but it may affect the stability of the complex in an artificial way and may prevent us from seeing relevant conformations.

Stability differences between pH 8 and pH 6

Qualitatively, the simulations at pH 8 and pH 6 are very different. The pH 8 simulation shows a stable complex throughout the 2 nanoseconds while the pH 6 simulation clearly begins to dissociate after 1 nanosecond (see Figure 2.) This difference supports our hypothesis that the different pH leads to a conformational change. The pH 6 complex may lead to eventual cleavage through reduced stability in the binding site.

At the residue level, we see that four of the nine titrating residues (three ASP and one HIS) change their charge state about one third of the way into the pH 6 simulation while all titrating residues for the duration of the pH 8 simulation stay in their initial protonation states (see Figure 3.) The pH 8 simulation is a good control as the X-ray structure was crystallized at pH 8 so we do not expect changes. In the pH 6 simulation,

the three aspartic acids are close in proximity and their charge states may be coupled. The arginine in the P1 pocket may stabilize ASP185. In figure 4, we see that the neutralizing of ASP185 is correlated with the arginine leaving the P1 pocket. The other aspartic acids follow and neutralize as well. This may then lead to the 100% protonation of HIS 138 which is already initially protonated to a large percentage because of the low pH.

Non-specific molecular mechanism for catalysis at low pH

These simulations do not propose, however, a discrete molecular mechanism for proteolytic cleavage at low-pH. Though catalysis will not be observed in this classical dynamics simulation, we did hope to see a discrete change in orientation of the H3 peptide backbone that we could propose is the mechanism for catalysis given the distance and orientation to the catalytic triad. This was not observed, however, possibly because of the short simulation time or the lack of explicit water mentioned above. The effect of low pH may be non-specific, whereby loosening of the complex enables more freedom for the H3 loop to be catalyzed. However, though the effect of pH on the complex may be non-specific, the catalysis is very specific, cleaving between Arg131 and Arg132 in the H3 loop as reported earlier using mass-spectrometry⁹.

Future Directions

Understanding the mechanism by which cognate inhibitors of proteases are cleaved at low pH but not at physiological pH would be beneficial to understanding protease

inhibition in general. Therefore, beyond this study of a synthetic antibody inhibitor, further constant-pH simulations of various proteases with their cognate inhibitors would be useful. More sampling would be required as the effects seen here may not be representative of a mechanism based on an ensemble average. Inclusion of at least the discrete water observed in the P1 pocket would also be interesting and may enable a more discrete change in conformation.

References

1. Lee SL, Dickson RB, Lin CY. Activation of Hepatocyte Growth Factor and Urokinase/Plasminogen Activator by Matriptase, an Epithelial Membrane Serine Protease. *Journal of Biological Chemistry* 2000;275(47):36720-36725.
2. Takeuchi T, Harris JL, Huang W, Yan KW, Coughlin SR, Craik CS. Cellular Localization of Membrane-type Serine Protease 1 and Identification of Protease-activated Receptor-2 and Single-chain Urokinase-type Plasminogen Activator as Substrates. *J Biol Chem* 2000;275(34):26333-26342.
3. Shi YE. Identification and characterization of a novel matrix-degrading protease from hormone-dependent human breast cancer cells. *Cancer Research* 1993;53(6):1409-1415.
4. Oberst MD, Johnson MD, Dickson RB, Lin CY, Singh B, Stewart M, Williams A, al-Nafussi A, Smyth JF, Gabra H. Expression of the Serine Protease Matriptase and Its Inhibitor HAI-1 in Epithelial Ovarian Cancer: Correlation with Clinical Outcome and Tumor Clinicopathological Parameters. *Clinical Cancer Research* 2002;8(4):1101.
5. Riddick AC, Shukla CJ, Pennington CJ, Bass R, Nuttall RK, Hogan A, Sethia KK, Ellis V, Collins AT, Maitland NJ. Identification of degradome components associated with prostate cancer progression by expression analysis of human prostatic tissues. *Br J Cancer* 2005;92(12):2171-2180.
6. List K, Szabo R, Molinolo A, Nielsen BS, Bugge TH. Delineation of Matriptase Protein Expression by Enzymatic Gene Trapping Suggests Diverging Roles in Barrier Function, Hair Formation, and Squamous Cell Carcinogenesis. Volume 168: ASIP; 2006. p 1513-1525.
7. List K, Bugge TH, Szabo R. Matriptase: Potent Proteolysis on the Cell Surface. *Molecular Medicine* 2006;12(1-3):1.
8. Parr C, Watkins G, Mansel RE, Jiang WG. The Hepatocyte Growth Factor Regulatory Factors in Human Breast Cancer. *Clinical Cancer Research* 2004;10(1):202.
9. Farady CJ, Sun J, Darragh MR, Miller SM, Craik CS. The Mechanism of Inhibition of Antibody-based Inhibitors of Membrane-type Serine Protease 1 (MT-SP1). *Journal of Molecular Biology* 2007;369(4):1041-1051.
10. Christopher J. Farady PFE, Eric L. Schneider, Molly R. Darragh and Charles S. Craik. Structure of an Fab-Protease Complex Reveals a Highly Specific Non-canonical Mechanism of Inhibition. *JMB* 2008 epub.
11. Ozawa K, Laskowski M. The Reactive Site of Trypsin Inhibitors. *Journal of Biological Chemistry* 1966;241(17):3955-3961.
12. McGrath ME, Hines WM, Sakanari JA, Fletterick RJ, Craik CS. The sequence and reactive site of ecotin. A general inhibitor of pancreatic serine proteases from *Escherichia coli*. *Journal of Biological Chemistry* 1991;266(10):6620-6625.
13. Case DA, Cheatham Iii TE, Darden T, Gohlke H, Luo R, Merz Jr KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. *J Comput Chem* 2005;26(16):1668-1688.

14. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 2004;25(13):1605-1612.

Chapter 6: Figure 1

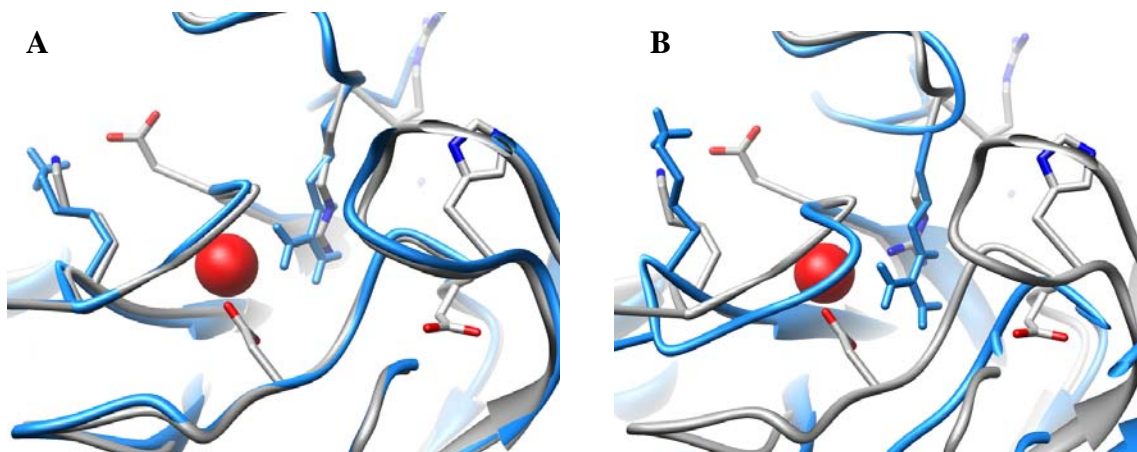


Figure 1 caption

Collapse of arginine and H3 loop towards P1 pocket due to use of implicit solvent. Pre-equilibration is shown in (A) and post-equilibration in (B) with crystal structure in gray and simulation in blue. The red oxygen is one of the structured waters found in the crystal structure that bridges between the arginine and aspartic acid. The water is not present in the simulation which uses implicit solvation.

Chapter 6: Figure 2

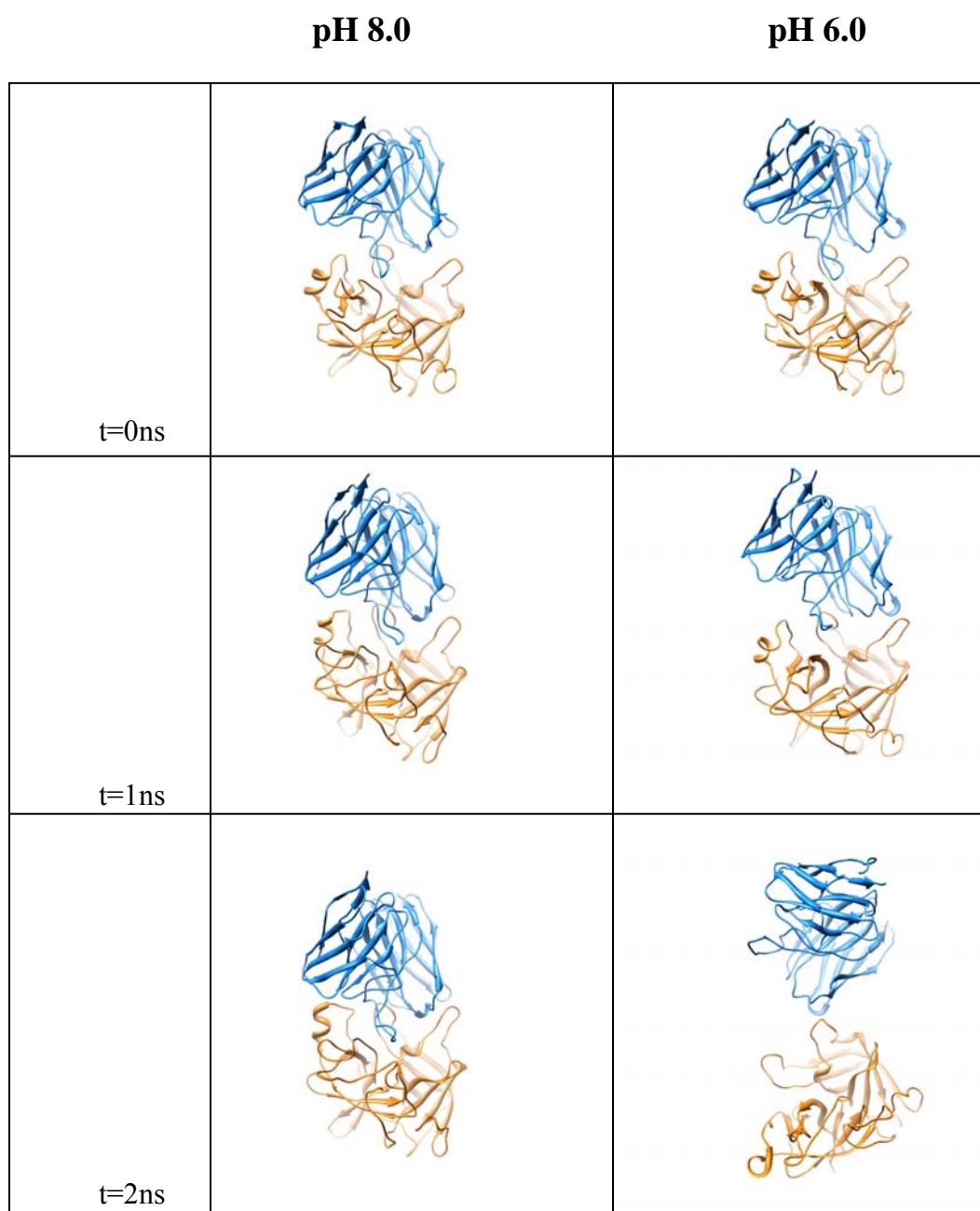


Figure 2 caption

Comparison of complex structure at 0, 1, and 2 nanoseconds for pH 8 simulation (left column) and pH 6 (right column.) Antibody is in blue and protease is in orange.

Chapter 6: Figure 3

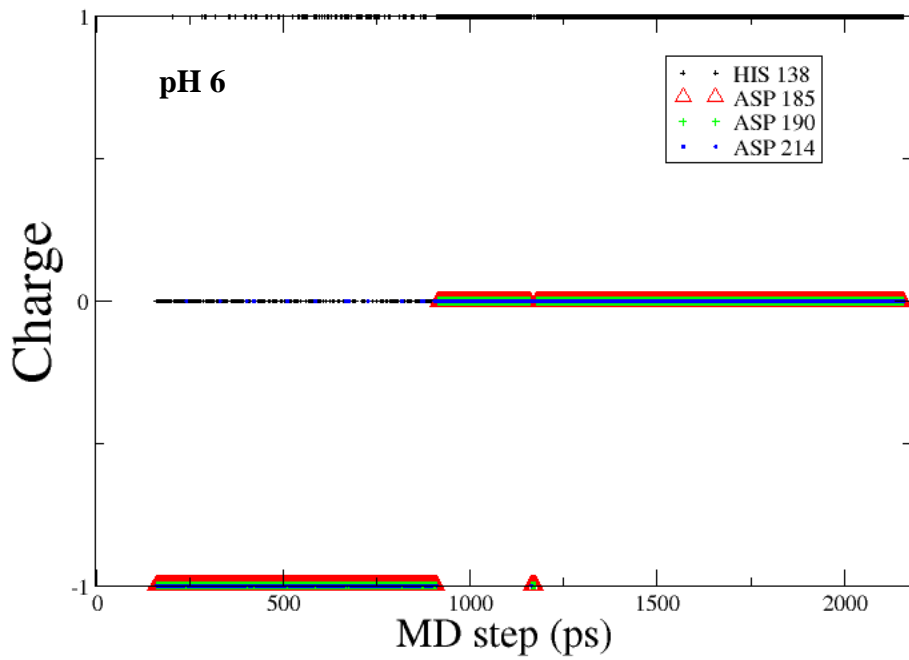
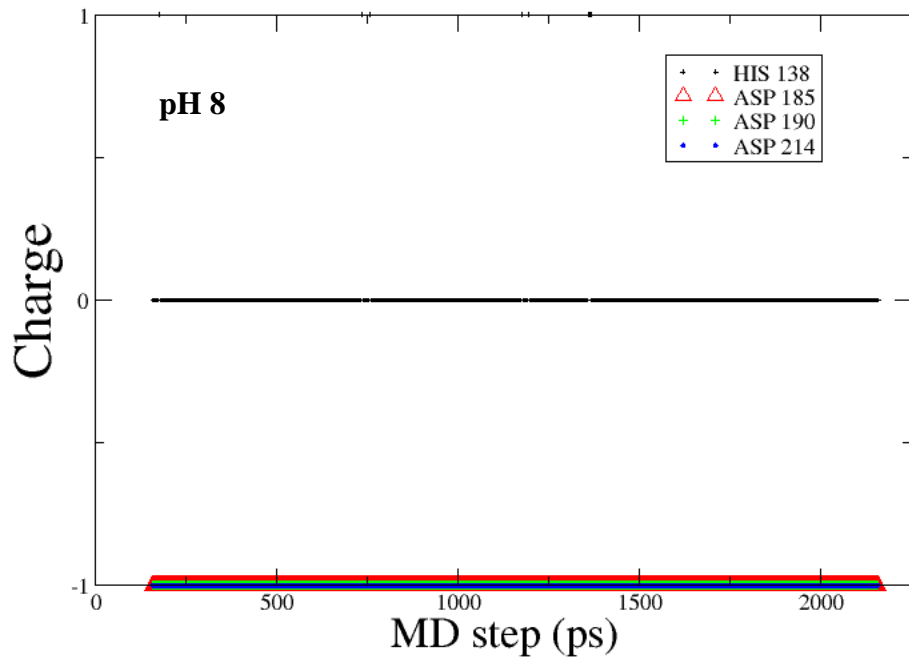


Figure 3 caption

Plot of charge state for 4 titrating residues that show differences across simulation versus simulation time. Top: simulation at pH 8. Bottom: simulation at pH 6.

Chapter 6: Figure 4

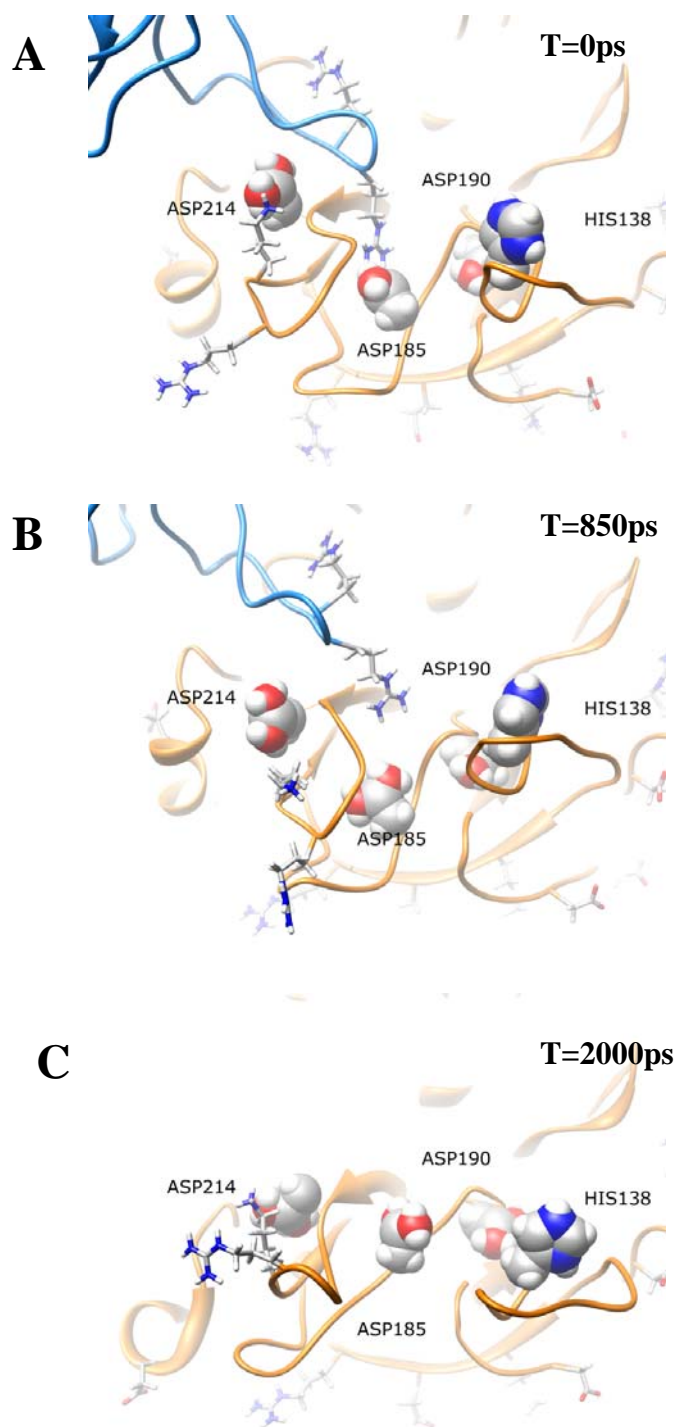


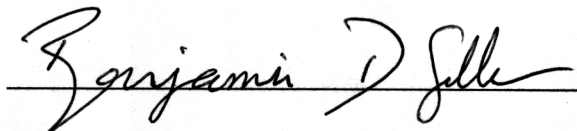
Figure 4 caption

Snapshots from simulation in at 0, 850 and 2000 picoseconds. Labeled residues in sphere representation are titrating residues. All other non-titrating residues are in stick representation. Hydrogen atoms do not represent true protonation state.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature Date