5-24-2006

Knowledge based structure modeling of the third hypervariable region of antibodies

Kevin Galens

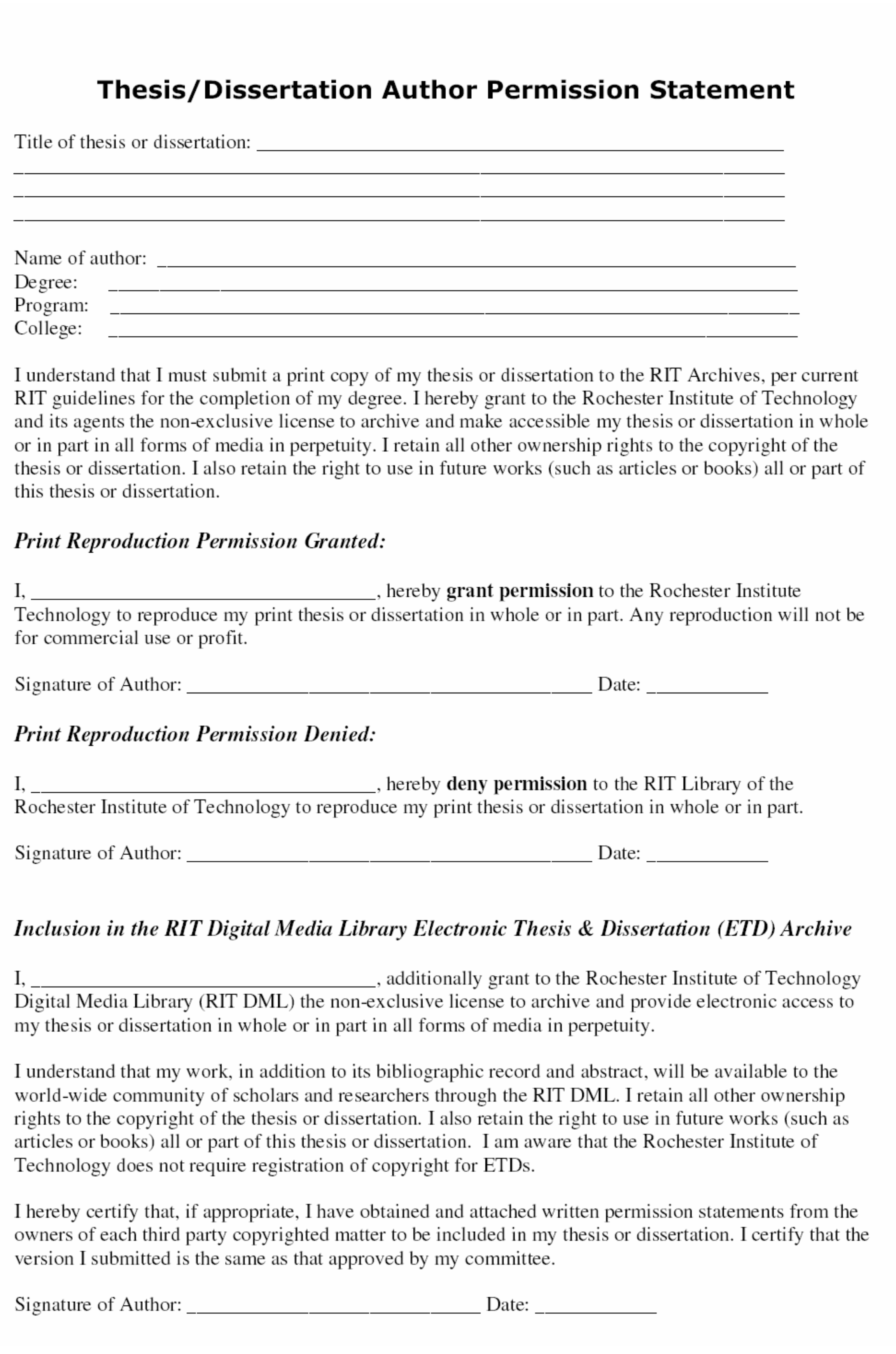
Knowledge Based Structure Modeling of the Third Hypervariable Region of Antibodies

Approved: ___________________________________
Thesis Advisor

____
Director of Bioinformatics or
Head, Department of Biological Sciences

Submitted in partial fulfillment of the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology

Kevin Galens



Thesis/Dissertation Author Permission Statement

Title of thesis or dissertation: ____
____
____
____

Name of author: ____
Degree: ____
Program: ____
College: ____

I understand that I must submit a print copy of my thesis or dissertation to the RIT Archives, per current RIT guidelines for the completion of my degree. I hereby grant to the Rochester Institute of Technology and its agents the non-exclusive license to archive and make accessible my thesis or dissertation in whole or in part in all forms of media in perpetuity. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Print Reproduction Permission Granted:

I, ______________________________, hereby **grant permission** to the Rochester Institute Technology to reproduce my print thesis or dissertation in whole or in part. Any reproduction will not be for commercial use or profit.

Signature of Author: ____________________________________ Date: ___________

Print Reproduction Permission Denied:

I, ______________________________, hereby **deny permission** to the RIT Library of the Rochester Institute of Technology to reproduce my print thesis or dissertation in whole or in part.

Signature of Author: ____________________________________ Date: ___________

Inclusion in the RIT Digital Media Library Electronic Thesis & Dissertation (ETD) Archive

I, ______________________________, additionally grant to the Rochester Institute of Technology Digital Media Library (RIT DML) the non-exclusive license to archive and provide electronic access to my thesis or dissertation in whole or in part in all forms of media in perpetuity.

I understand that my work, in addition to its bibliographic record and abstract, will be available to the world-wide community of scholars and researchers through the RIT DML. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I am aware that the Rochester Institute of Technology does not require registration of copyright for ETDs.

I hereby certify that, if appropriate, I have obtained and attached written permission statements from the owners of each third party copyrighted matter to be included in my thesis or dissertation. I certify that the version I submitted is the same as that approved by my committee.

Signature of Author: __________________________ Date: ___________

Table of Contents

List of Tables

List of Figures

Acknowledgments

Thesis Advisory Committee

Dr. Mark Paris, Ph.D.
Vaccinex, Inc.

Dr. Gary Skuse, Ph.D.
Director of Bioinformatics
Rochester Institute of Technology

Dr. Anne Haake, Ph.D.
Department of Information Technology
Rochester Institute of Technology

Dr. David Lawlor, Ph.D.
Department of Biological Sciences
Rochester Institute of Technology

Abstract

Protein structure prediction has gained increased attention over the past decades in a wide range of biological disciplines. Creating an accurate visual model of a protein can aid in protein engineering; which has implications in the creation of therapeutic molecules as is the case with antibodies. The third complementarity determining region of the heavy chain of antibodies (CDR-H3) is known to show a large degree of variation in sequence and in length, and therefore has provided difficulties for structure prediction. By separating the CDR-H3 into two logical sections, the apex and base, and using a homology modeling techniques for each section, this study attempts to predict structure for this important region of antibodies. This method also accounts for certain interactions proven to be relevant in CDR-H3 structure to select a suitable parent for modeling an unknown CDR-H3. The selection algorithm was tested using a test set of proteins, selected based on base type, length and diversity. Overall, there seemed to be a slight improvement in the prediction of CDR-H3 by this method when compared with traditional homology methods; although both drastic improvements and evident decreases in accuracy of predictions from individual molecules can be observed.

Introduction

Monoclonal antibodies have gathered increasing attention over the past decade as possible agents for the treatments for diseases, such as cancer and arthritis. There are currently ten therapeutic monoclonal antibodies on the market and there is a significant amount of research attending this area currently. In general, monoclonal antibodies can be developed more quickly and cheaply than small molecules, and therefore are deserving of such attention. In the area of protein engineering, knowledge of structure is important as it ultimately determines function. Using computational methods to predict structure will lead to a better understanding of how primary sequence determines a protein's tertiary structure.

Antibodies are multi-domain proteins that contain active sites specific for a particular antigen. Two domains, VL and VH, are positioned close together to form the scaffold upon which the antigen binding site is located. The scaffold is relatively conserved as contrasted with the sequence and length variation seen in the complementarity determining regions (CDRs) that make up the antigen binding site (Wu, *et al.* 1970). There are six such CDR loops, three on the light chain, L1, L2, L3, and three loops on the heavy chain, H1, H2, and H3.

Even though these loops are highly variable, it has been shown that five out of six of the loops can take on only a limited number of main-chain conformations, known as canonical structures, based on a limited number of residues (Chothia, *et al.* 1987; Decanniere, *et al*. 2000). Due to the large variety in length and sequence of the third CDR of the heavy chain (CDR-H3), a concrete set of canonical classes has not been assigned to this loop (Al-Lazikani, *et al.* 1997). It has also been observed that CDR-H3 plays an important role in determining the specificity and affinity of the antigen binding site due to its great degree of variability.

A main contributor to the high degree of variability of antibodies comes from the process by which they are created. An immature B-cell contains germline genes, or multiple exons, of various classes (named variable, diversity, and joining) which need to become rearranged in order for transcription to begin. A gene (or group of exons) from each of these classes will remain in the mature B-cell, which will combine to produce a complete antibody molecule. Antibodies that are produced from the same germline genes share similarities in sequence and therefore possibly share similarities in structure.

CDR-H3, the most variable loop in the antigen binding domain, can be broken down into two regions: the apex (or head of the loop) and the base (or torso), which are located distal and proximal to the framework region respectively (Shirai, *et al.* 1996). For this study, the CDR-H3 region will be defined as in Morea, *et al.* 1998, which is comprised of the amino acids from 92Cys to 104Gly following the Kabat numbering scheme (Wu, *et al*. 1970). The base of the loops are defined as the ten residues that reside proximal to the framework region, four residues from the N terminus and six residues from the C terminus (Morea, *et al.*1998) (figure 1).

It has been observed that the base of CDR-H3 regions can take on one of two classes, kinked (K) or extended (E). The conformation depends on a salt bridge commonly formed by the conserved 101Asp and an N-terminal basic residue, resulting in a bulge at the 101st residue (Shirai, *et al.*1996) (table 1). It was observed that the presence of a basic residue (Arg/Lys) at position 94 is essential for the bulged base region. A sub-class,

		Arg / Lys 94	
		Present	**Absent**
Asp 101	**Present**	Bulged	Non-Bulged
	Absent	Bulged	No examples length > 10

Table 1: Rules governing the bulged base. As propsed by Shirai *et al.* 1996.

kinked plus extra bulge (K^+), can also be formed, with rules being defined by Shirai, *et al.* (1996). The three conformations can be seen in Figure 1.

Along with molecular interactions within CDR-H3 itself, other inter-molecular interactions have been described (Morea, *et al.* 1998). One notable interaction involves residue $100b_H$ of H3, which interacts with the light chain through the $V_H – V_L$ interface. More often than not, $100b_H$ is a tyrosine which packs with the tyrosine commonly found at position 49_L. This is also the case if $100b_H$ is a threonine. If 100b is instead a tryptophan, the side chain points in the opposite direction and interacts with L3 instead of the normal framework interaction (49_L). It has also been observed that $100a_H$ Phe will interact with 49_L in the case that $100b_H$ is a glycine (Morea, *et al.* 1998). Since the side chain of $100b_H$ determines the conformation of the backbone proximal to this residue, it can dramatically change the presentation of the apex into the antigen binding site, and therefore should be considered relevant in H3 structure.

Another interaction noted between H3 and the other variable loop domains involves Arg94. Because this amino acid is also involved in determining the class of the base (E, K or K^+), its interactions with the other loops depends on the base class. If the H3 structure is in the bulged conformation, Arg94 usually packs against the aromatics found in 27 and 32 of H1. If the base is extended, 27 and 32 will pack against either 96 or 99 of H3 if either of these residues is an aromatic (Morea *et al.* 1998). It has not been determined if these interactions are relevant to the structure of H3 due to the lack of data when there is no interaction between H3 and amino acids 27 and 32 of H1.

Determining rules governing the conformation of the apex has proven to be more difficult. It has been seen that the conformation of the apex of a loop depends on the general class of the torso region it rests on. The extended base conformation usually has a short loop region which follows the general β-hairpin classification system described in Sibanda *et al.* 1989. The kinked bases, which do not conform to the normal hydrogen bonding pattern of an anti-parallel β-sheet, are often characterized by longer apex regions (Morea, *et al.* 1997).

Culler *et al.* (2004) have used clustering and entropy information in order to define characteristics regarding the importance of specific residues in CDR loops. They found that residue 95_HV commonly points into the V_H-V_L interface, but in a few cases pointed upwards into the antigen binding site. This is of importance in this study because of the residue's possible role in antigen affinity in this uncommon conformation. Another lab also found that this residue may play a role in antigen binding. (Vargas-Mardrazo, *et al.* 2003).

Due to the difficulties in identifying characteristics in CDR-H3, it has been difficult to model the entire antigen binding site (Morea, *et al.* 1998). One technique used to predict the structure of CDR-H3, and often used to predict the structure of loops in general, is

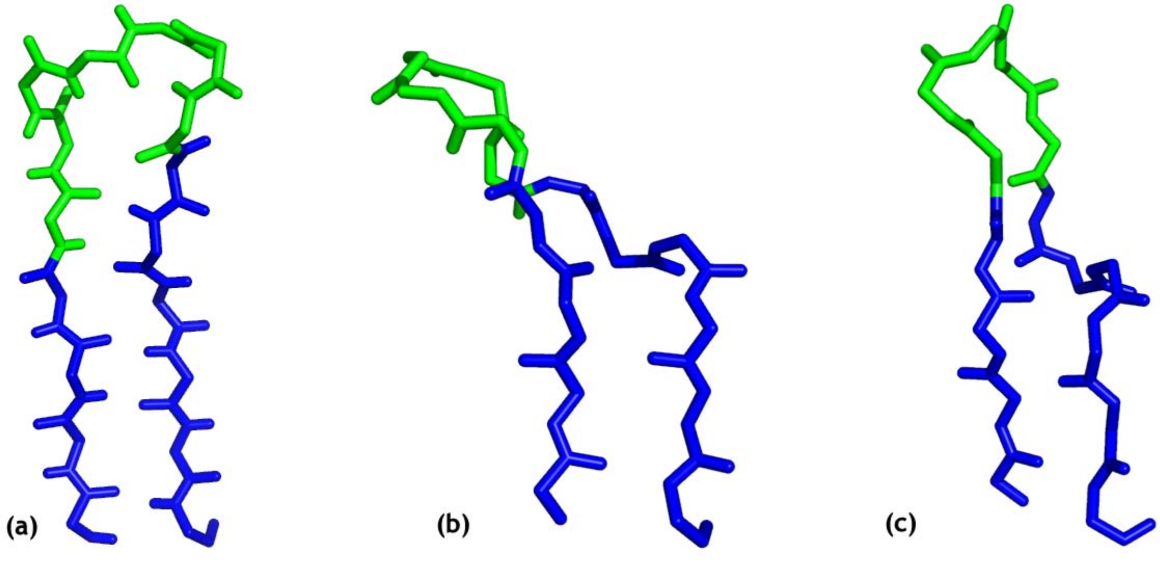


Figure 1 – The three basic conformations of base regions of CDR-H3. Base shown in blue and apex in green. (a) Extended base conformation of 1MPA. (b) Kinked base conformation of 1KB5. (c) Kinked plus bulge conformation of 1CLY. Notice the disconfiguration of the C terminal (left side of loop in picture) region of base of (b) and (c).

extensive database searching. This involves searching a database for loops that share similar sequences and similar lengths (Morea, *et al.* 1997; Shirai, *et al.* 1996) and searching for a relationship of sequence and structure. These studies have resulted in limited success, but have generated some valuable information regarding CDR-H3 structure. The lack of great success may be due to the limited number of proteins available at the time of the study.

In general, homology modeling is based on the idea that if two molecules share similar sequence, primary structure, they will also share a similar tertiary structure. By locating a parent molecule with known structure that matches the sequence an unknown one, it is possible to model the unknown molecule by using the coordinates of the former. This procedure has resulted in a high level of accuracy when compared with alternative modeling methods. Although the accuracy of the method is not yet 100%, there are benefits to using computational techniques to predict structures instead of more traditional techniques such as x-ray crystallography. Computational methods are drastically quicker and cheaper to develop and when homology modelling's weaknesses are kept in mind, this procedure can be used as an effective tool to help lead studies in a laboratory.

Traditional homology based modeling methods have been studied and it has been shown that there are three important factors regarding this procedure (Sternberg, 1996). The first is that automated methods, if they fail, usually fail due to a faulty multiple sequence alignment. This is one of the first steps in homology modeling, and if the alignment is grossly incorrect, the rest of the prediction will also be flawed. Secondly, Sternberg states that the modeling of variable loop regions is still very difficult (but improving). The third general conclusion about comparative modeling is that energy minimization steps often resulted in a prediction that was further from the correct conformation than prior to the energy minimization step (Sternberg, 1996). Energy minimization will often force the protein to adopt a local minimum, which may or may not be equal to the protein's absolute minimum energy conformation or represent the proteins native structure. Further, in the context of only modeling a section of the protein, there may be other factors affecting the energy of a molecule which are not represented in the predicted model. Without all the information, the energy minimization algorithm will not be able to incorporate all the data needed. For both these reasons, we cannot be confident that energy minimization will increase the accuracy of the homology modeling technique.

Other methods used to model variable loops, besides homology modeling, include various statistical techniques. A clustering method based on sequence similarity of seven residue CDR-H3 loops, was performed (Martin, *et al.* 1996). This resulted in identifying residues outside of the H3 loop that are commonly involved in the main chain conformation of H3. Another novel approach to modeling CDR-H3 included a neural network technique (Reczko, *et al.* 1995). This method proved successful for apices shorter than seven residues in length, while the longer loops had larger average root mean square deviations when compared with the actual crystal structure. The neural network was used to predict φ and ψ angles of the main-chain backbone.

The study described herein proposes a new technique to modeling the antigen binding site, with an emphasis on predicting CDR-H3 structure by homology modeling. By combining the rules and characteristics defined in the literature, data from recent residue/affinity experiments, along with existing technology, a semi-automated procedure for structure prediction is created. This system can be applied in a multitude of ways. In a most basic sense, it can be used to visualize a specific CDR-H3 with unknown structure, but the system set up can also be used for more functional purposes. In the context of antibody active site structure prediction, this may prove a valuable resource in constructing a confident variable loop region. This system can also be applied in proteomics studies, where changing various amino acids at different points may change the structure (and possibly function) of the loop.

Methods and Materials

The method described in this section will outline a technique that involves knowledge based homology modeling of a target unknown antibody by parts to parent CDR-H3 pieces with known structures. The hypothesis for this technique is that modeling the base and the apex separately will result in a more accurate prediction of the entire CDR-H3 loop, due to the finding that the structures of these two sections are governed by separate sets of rules (Shirai *et al.* 1996). It must also be noted that the interaction between these two sub-domains is also very important and was kept in mind throughout the project (as described in the following sections).

The methods section contains five general segments that outline distinct conceptual pieces of the project as a whole. As these sections may appear grouped or overlapped in the actual project, they will be described as individual pieces for the sake of clarity. Although the methods section describes the project in its final state, assessments were done during the development process to determine the effectiveness of the algorithm. The results and description of points of assessment can be seen in the results section.

Pre-processing

Before we can model an amino acid sequence by parts, we must first identify the parts. The base (defined as the four residues most proximal to the N-terminus of the CDR and the six residues most proximal to the C-terminus) and the apex, the remaining residues distal to the framework region, are these parts. In order to speed up the homology search to identify a proper parent sequence (the sequence on which the target, or query, will be modeled on) a pre-processing stage is performed. This involves searching the Brookhaven Protein Databank for immunoglobulin proteins and parsing out bases and apices of CDR-H3 loops. These short sequences are placed in a relational database, listed as sequence information with links relating to their three dimensional data, via their respective PDB ID. The process by which these segments of sequences will be parsed from entire sequence will be described later in this document (Materials and Methods: Identification of CDR-H3 region).

Intra-peptide interactions have been described in the literature (Morea, *et al.* 1998) and include interactions between CDRH3 with other sections of the light chain. As introduced above, a common interaction with structural consequences can be observed between $100b_H$ (found within CDRH3) and 49_L from the light chain. These two amino acids are parsed out of the sequence separately using Kabat numbering and stored in the relational database. This allows efficient access to information regarding the presence and type of interaction that may be present at these sections.

Identification of CDR-H3 region

The target sequence can be in a variety of different states so that there is a freedom of input format. It can be entered as an entire molecule (both heavy and light chains), just a heavy chain, or just a CDR-H3 region. Due to the fact that the entire molecule can be used in this algorithm, it was predicted that the more information given to the algorithm, the more accurate the result, since the majority of the algorithm is based on more than just the CDR-H3 sequence.

If the target sequence is given as an entire molecule or as a heavy chain, the CDR-H3 needs to be parsed from the rest of the heavy chain. As mentioned in the introduction, the CDR-H3 loop is highly variable, not only in sequence but also in length. Although the patterns identifying the CDR-H3 loop itself will be a difficult marker to classify the region, it is known that the CDR loops reside on a more conserved β-framework region (Oliva, *et al.* 1998). The Kabat numbering system can also be used to define the various regions and to act as a map while traversing these molecules. Important residues that do not reside within the CDR-H3 (such as residue 49_L) are also collected to compare with the parent proteins.

When the sequence is given as an entire antibody molecule (heavy and light chains, and the possibility for other molecules, e.g. antigens), the algorithm must determine which chain is heavy and which is light. This is done by using a multiple sequence alignment of previously aligned heavy sequences to create a hidden markov model (HMM) which can be used to score the peptide chain in question. The algorithm will score every chain present against the heavy chain HMM and then take a closer look at the highest scoring chain. If the highest scoring chain scores above a certain threshold, it can be confidently used as the heavy chain. The multiple sequence alignment was created by choosing antibodies that represented some diversity within the heavy chain, but conformed to the rules for heavy chains (i.e. so that extreme outliers were not used for the alignment). See the discussion section for a complete analysis of the sequences used. An HMM was created from this multiple sequence alignment and was tested against multiple other heavy chains (including outliers), light chains, sequences of similar length to heavy chains (randomly chosen) and sequences of varying lengths.

Once the heavy chain is identified, the same multiple sequence alignment is used to parse out the CDR-H3. The unknown heavy chain is aligned with the others. Since it is already known where the CDR-H3 is in the alignment, it is simply a text processing problem to retrieve the CDR-H3 from the unknown sequence (assuming the new sequence has been aligned correctly). Because of the chance of outliers and sequences that, for some reason or another, do not align very well, the score of the alignment is parsed after the addition of the new heavy chain and if it is above a certain threshold, we can be confident that the sequence aligned properly and the CDR-H3 aligns with the known sequences. If the alignment score is not above a certain threshold, we cannot be confident in the alignment and therefore must use a different technique to identify the CDR-H3.

Martin, A. (2005) describes a method for identifying the CDR's by "walking" through the sequences. This is done by identifying patterns before and after the first CDR, and then using this location to find the second CDR, and so on. If this method fails, a CDR-H3 cannot be determined confidently and the sequence is discarded.

As mentioned above, residues from other parts of the heavy chain and from the light chain also need to be parsed out at this step. Those heavy chain residues are gathered during the alignment step described above. The light chain is aligned in a very similar fashion as described for the heavy and the specific sequences can be seen in the results section.

Finding the Template

After retrieving the CDR-H3 from the target protein, this sequence is searched against the pre-processed database of antibody molecules in order to find a suitable apex and base that will be used to as templates for the target molecule. The algorithm used to identify template protein segments is a weighted search algorithm. Each potential template molecule is given a score based on a number of properties including sequence similarity, presence of important amino acids, structure class, and suspected interactions for both the base and apex separately. The highest scoring apex and the highest scoring base, not necessarily from same molecule, are then chosen as templates for modeling the target molecule.

To score the base (four residues nearest the N terminus and six residues nearest the C terminus) we first looked at sequence similarity. Since the length of the base is static, we can simply check residue to residue for similarities. As described in Shirai *et al.* (1996) and summarized in table 1, there are three base types that are governed by key residues. The algorithm awards a bonus to the base that fits the same base type. This algorithm will allow slight variation in the sequence in order to retain the general structural class of the base.

The scoring algorithm for the apex is slightly more difficult due to the varying length of this section. A simple residue comparison method cannot be employed, such as in the base scoring algorithm because of the variety of lengths and therefore an introduction of gaps, and it is important that a high-quality alignment be used to assess the sequence similarity between the target and potential template. The algorithm therefore uses a global alignment algorithm to compare sequence similarities and parses the alignment score from an external alignment program. Because an external global alignment program is used to determine the sequence similarity between two apices, the scales of the base and apex scores do not match. Therefore, these score cannot be compared to determine relative confidence in the match.

The multiple sequence alignment scores matched amino acids and gives a penalty for gaps. The default parameters for this program were used and scores were not adjusted. An interesting note about the algorithm is that gaps placed at the beginning or end of the sequence are not penalized. Therefore apices of different lengths have the ability to score higher than two apices of the same length with similar sequence identity. It has been found that β-hairpin structure can be broken into classes depending on the loop length (Sibanda *et al.* 1989). In general, there are four classes (1-4), detemined solely by sequence length following the formula of n%(modulo)4, where n is the sequence length. Therefore two apices that differ in length, even by one residue, would fall in different classes and therefore conform to a different structure. A bonus is added (and scaled to be of importance with the external multiple sequence algorithm) for sequences of the same length.

An additional alignment method was developed as an alternative to the pairwise scoring using a multiple sequence alignment, comprised of 5 sequences (selection process described later). A global alignment algorithm (Needleman, S.B. *et al.* 1970) was used in order to introduce a more controlled method of scoring sequence similarity. This method is much more computationally intensive, but the parameters and finer details of the alignment method can be altered in order to achieve an ideal score.

As mentioned above, intra-peptide interactions are also of consideration in the template searching algorithm. The interaction between residue $100b_H$ of CDR-H3 and 49_L of the light chain, in particular, is being used to help determine a suitable template molecule. Due to the high variation in length of the CDR-H3 molecule, $100b_H$ is sometimes present in the apex, sometimes in the base and sometimes is not present at all. Therefore it must be determined where, if at all, $100b_H$ is located and score a bonus in the corresponding section if this interaction is present in both the target and template molecules. When this interaction is not present in both the target and template molecules, no bonus is scored because it is unclear from the literature whether or not a similar heavy-light chain interaction is present. In other words, the lack of this interaction tells us nothing about the conformation of the molecules and is therefore left out of the scoring algorithm.

The final consideration in the scoring algorithm is to look at the overlap between the apex and base. This step is important because it is the overlap area which will determine the presentation of the apex into the antigen binding site (Morea *et al.* 1998). This is scored by a comparison of the four amino acids closest to the junction of the apex and base between the target and the template molecules, two on each side of the loop. The residues are compared using a protein weight matrix under the assumption that similar amino acids will produce a similar 'take off' point for the apex.

At the end of this set of scoring (sequence similarity, interaction bonuses, overlap) there may still be a couple of parent molecules that share the highest score. In order to break this tie, the set of germline genes is used to determine the best match. By choosing the same germline gene, we can guarantee that the rest of the heavy chain (not just the CDR-H3) shares some sequence similarity. This is important because the CDR-H3 interacts with the rest of the heavy chain; finding two molecules that share the same germline gene may also share similar interactions with relevance to CDR-H3 structure. It should also be stated that the third hypervariable region is not encoded by a VH gene. The end of the VH gene is the beginning of the third hypervariable region. Instead, CDR-H3 comes from the diversity region. This is not used in the scoring algorithm because of the high diversity seen within this region. Identifying a D germline would be uncertain at best.

Generating the model

At this point, we have identified a base and apex parent and the CDR-H3 three dimensional model can be generated. The program used in this section is Modeller, developed by Andrej Sali (1993). Modeller is used for homology or comparative modeling of protein structures. By using spatial restraints, the program will generate a model of a protein molecule from an alignment of unknown and related proteins with known structures. From this alignment, Modeller will give coordinates for all non-hydrogen molecules and output multiple predictions for its structure. In this study, only three predictions were viewed and tested for accuracy.

Conveniently, Modeller natively supports modeling of an unknown protein by using known structures from multiple proteins. An alignment is created that aligns the correct sections of the unknown protein with both the parent molecules. This is given to Modeller as input and pdb files are exported as a result.

Accuracy Evaluation

After the predictions have been made by Modeller (Sali *et al.* 1993), the accuracy of these threaded molecules must be evaluated for accuracy. A test set of molecules was created before the testing of the system began, in order to rate the success of the system. A test set of 23 antibodies was developed from a set of 231 antibodies used in the high resolution database (the pdb id's of these data can be seen in table 2). The test set was chosen randomly for diversity in apex length and base type. These antibodies have known structures and were removed from the database for the duration of testing, so that they would not be chosen for the prediction of structures.

PDB ID	*CDRH3*	*CDRH3 Length*
1BAF	CARGWPLAYWG	11
1BQL	CLHGNYDFDGWG	12
1CIC	CARGLAFYFDHWG	13
1CT8	CARYRYDEGFAYWG	14
1DBA	CTRGDYVNWYFDVWG	15
1H0D	CTRLGDYGYAYTMDYWG	17
1IKF	CTRHTLYDTLYGNYPVWFADWG	22
1MFB	CTRGGHGYYGDYWG	14
1NGW	CTRRDMDYWG	10
1Q9O	CVRDIYSFGSRDGMDYWG	18
1SBS	CTRGAYYRYDYAMDYWG	17
1UZ6	CARETGTRFDYWG	13
2H1P	CARRDSSASLYFDYWG	16
1WT5	CARSGGPYFFDYWG	14
2PCP	CGRSTWDDFDYWG	13
2A77	CARHDDYGKSPYFFDVWG	18
1XIW	CARSGYYGDSDWYFDVWG	18
1EGJ	CSRGDGIHGGFAYWG	15
1G9M	CAGVYEGEADEGEYRNNGFLKHWG	24
1KC5	CARGGTGFDYWG	12
1L7T	CARAYYGYVGLVHWG	15
1NFD	CTRAGRFDHFDYWG	14

Table 2 – The test set of 22 immunoglobulins

By varying the length of the apex we can measure the success of the system on molecules of varying length. In previous studies (Morea, *et al.* 1998) it has been stated that longer CDR-H3 molecules were more difficult to accurately predict the structure for and it will be interesting to see if this is the case with the data found here. Only 22 of these 23 antibodies were used in tests because there seemed to be a duplicate antibody in the PDB or at least a very similar antibody. When using this specific antibody and threading on the molecule it scored to be the best match for a parent (which in this case was the same molecule for the base and apex) and the prediction was completed, the RMSD (described below) was unusually small and virtually zero, due to the presence of a highly similar molecule in the database. This would have skewed the results of the testing by reducing the average distance of predicted molecules against their known structures. Because of this finding, this antibody was omitted from the test set.

When a prediction is made for a molecule in the test set, the root means square deviation from the known structure is calculated. This measurement can be defined as a means of measuring the overall distance of the molecule from its original state, in Angstroms. By using this standard measurement we can quantify the molecular distance of one known structure from the predicted molecule and compare predictions to measure the accuracy of the system.

Implementation

This section describes the specific implementation of the above described method. Various considerations were made in order to maximize the accuracy of the system during the implementation process. The system was developed on a linux system running Fedora Core 3. The operability of the prediction on other operating system types was not in the scope of this project and therefore was not tested.

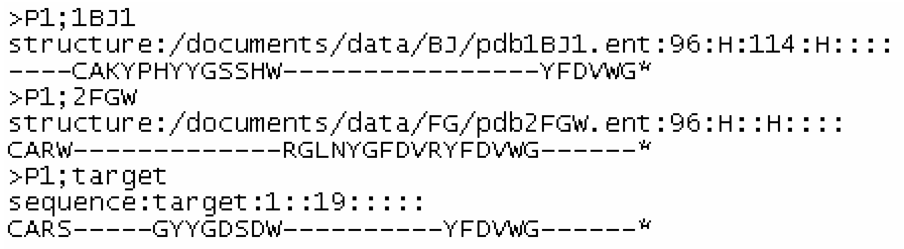
The majority of the scripts were developed in the Perl programming language. Since the majority of the process involves information related to antibodies, a Perl module was developed to handle the most commonly used antibody methods. This pseudo-object oriented approach reduces the repetition of code and also allows a more abstract view of the process, from a programming standpoint. The antibody module allows access to various antibody attributes, such as heavy and light chain, CDR-H3 sequence, resolution as reported in the pdb file etc.

The high resolution database of antibodies was constructed using the Postgresql database management system filled automatically with a perl script. The PDB ID’s of antibodies were obtained from the web resource *Summary of Antibody Structures in the Protein Databank* (Allcorn *et al.* 2002). They have developed a system titled Self-Maintaining Database of Antibody Crystal Structures (SACS for short) which automatically searches the protein databank for immunoglobulin molecules by using an advanced text search. SACS then provides an XML file of these molecules which is free to download. The system described here in this paper uses this XML file and retrieves sequences from the PDB files listed in the XML file. Since this file is regularly updated with new releases of the PDB, new structures can be added to the high resolution database easily. The PDB was downloaded in October 2005 with 254 high resolution antibodies present in that release.

In the high resolution database, N-terminal and C-terminal base sequences are stored along with the apex, pdb id, and $100b_H$, $100a_H$ and 49_L amino acids for each antibody. The structure is only stored in the original files in locally stored protein databank.

In order to find the parent molecules on which to be modeled, a perl script, using the antibody module, along with another module used in the alignment method were written in order to score bases and apices of potential parent molecules. The alignment module implements the Needleman Wunsch algorithm for finding similarities in proteins (Needleman, *et al.*, 1970). Using a dynamic programming algorithm, the module will return a score when given two sequences. For the entirety of the system, a match is worth 4, a gap is -2 and a mis-match is -1. The gap score is fixed and relatively mild because we want to encourage using a gap during the alignment (especially in searching for germline genes). This will encourage the maximum alignment of amino acids.

As mentioned above Modeller takes in an alignment of the unknown molecule to parent molecules with known structures. The program also uses a python script for instructions. An example of the alignment file can be seen in figure 2. Both the python script and the alignment are generated by a Perl script.



```
>P1;1BJ1
structure:/documents/data/BJ/pdb1BJ1.ent:96:H:114:H::::
----CAKYPHYYGSSHW----------------YFDVWG*
>P1;2FGW
structure:/documents/data/FG/pdb2FGW.ent:96:H::H::::
CARW-------------RGLNYGFDVRYFDVWG------*
>P1;target
sequence:target:1::19:::::
CARS-----GYYGDSDW----------YFDVWG------*
```

Figure 2 – Alignment file example used in Modeller.

The Modeller alignment file is a modified fasta file with more information needed by the Modeller program. Each sequence has its own entry and each contains three parts. The first line, preceeded by a '>', denotes the sequence type and identifier. The P1; is simply an identifier to modeller that lets the program know that this line contains the identifier of the molecule. The second line contains a series of information about the molecule if the structure information is available for that molecule, where the structure information can be found, the beginning and ending amino acids for the sequence and a chain identifier. The start and end amino acid numbers must come from the pdb file for that molecule. There are a couple numbering methods used for amino acid numbering of antibodies, which makes determining these data difficult. The entire pdb file must be searched for these amino acids in the correct position.

Results

In order to parse the CDR-H3 regions of the molecules in this study, a multiple sequence alignment was created and used as templates for unknown sequences to be aligned to and to confidently identify the variable region. Table 3 shows the sequences chosen, as previously described, and the alignment can be seen in Appendix A. The CDR-H3 region of the alignment is highlighted.

PDB ID	*CDRH3*	*Total Length (Heavy Chain)*
15C8	CAADPPYYGHGDYWG	217
1QBM	CAGYDYGNFDYWG	219
1OAY	CARMWYYGTYYFDYWG	122
1A14	CARSGGSYRYDGGFDYWG	120
1MRD	CANLRGYFDYWG	215
1Q9W	CVRDIYSFGSRDGMDYWG	226
1PG7	CARDTAAYFDYWG	217

Table 3 – Sequences used in alignment for parsing the heavy chain of antibodies.

Using this alignment, various sequences were added to see the range of scores that one would get and still confidently assume that the sequence used was a heavy chain. A Hidden Markov Model was created from this alignment in order to score possible heavy chains. The HMM used is not included, but was made from the alignment shown by using the program HMMer. The scores in this table are used to illustrate the range of scores obtained using said hidden markov model to search against a putative heavy chain.

PDB ID	*Description*	*Chain ID*	*HMM Score*
1ETZ	Heavy Chain	H	350
1TXV	Heavy Chain	H	489
1BM3	Heavy Chain	H	425
1OAK	Light Chain	L	29
1A8J	Light Chain	L	56
1PG7	Light Chain	L	42
13PK	Kinase	C	-81
1G53	Carbonic Anhydrase II	A	-81

Table 4 – HMM test scores. Test proteins run against hmm created from heavy chain alignment shown in Appendix A. Scores are retrieved from the output of hmmsearch.

In this test, heavy and light chains from antibodies were used, and also random unrelated sequences. The relevance of these numbers is discussed later.

A similar process was repeated for light chains. The light chains of the antibodies in table 5 were aligned and used to create hidden markov model in which to detect light chains. The program locates heavy and light chains separately because often times there are more than just the heavy and light chains described in the PDB file. By looking for each chain separately, we can be more confident in the automatic retrieval and identification of antibody chains.

PDB ID	*Length*
1CT8	214
1BAF	214
1NCW	219
1T4K	217
1AP2	113

Table 5 – Light chains used in the light chain alignment and hidden markov model used to identify light chains automatically.

A fewer number of sequences were used in the light chain alignment due to the lower amount of diversity present within this chain. It must also be added at this point, that it seems that some of the light and heavy chains seem incomplete. For example, 1AP2 used in the light chain alignment aligned very well with the other sequences, and it is known (Morea *et al.* 1998) that there is not that much variation in the sequences of light chains. The chain was included in the alignment because this would make the hidden markov model more flexible. This results in lower scores overall for light chains, but would give a shorter, incomplete chain a reasonable score. Without including one of these sequences to the alignment, the hidden markov model would score very low against such sequences, even though the sequence is a light chain. The light chain alignment can also be seen in Appendix A.

As a control, the test was ran against the high resolution database, comparing the sequence of the CDR-H3 molecule with the known structures. Only one parent was used to model the CDR-H3 of this test and the selection of a parent was based solely on sequence similarity (Table 6).

Test Antibody	***Parent Antibody***	***Run 1 (RMSD)***	***Run 2 (RMSD)***	***Run 3 (RMSD)***	***Average (RMSD)***
1BAF	1NJ9	1.611	1.089	1.217	1.306
1BQL	1MRC	2.910	2.062	1.940	2.304
1CIC	1AD0	1.339	1.323	1.740	1.467
1CT8	1FRG	1.566	2.199	1.119	1.628
1DBA	1IQW	2.046	1.956	2.572	2.191
1H0D	1TXV	3.146	3.354	2.321	2.940
1IKF	1FN4	8.507	6.754	7.715	7.659
1MFB	1BZ7	2.438	1.993	1.668	2.033
1NGW	1N7M	0.506	0.545	0.722	0.591
1Q9O	1A14	2.296	2.819	2.050	2.388
1SBS	1I8M	2.077	2.638	2.473	2.396
1UZ6	1BEY	4.157	3.852	4.442	4.150
2H1P	1A6U	2.627	2.410	3.003	2.680
1WT5	1JFQ	2.546	2.361	2.357	2.421
2PCP	1KB5	2.933	1.316	1.814	2.021
2A77	2A1W	2.584	1.925	1.453	1.987
1XIW	1HYX	3.034	3.139	3.746	3.306
1EGJ	1CLY	1.911	2.252	2.618	2.260
1G9M	1GC1	4.266	5.871	4.172	4.770
1KC5	1KCR	1.911	1.521	1.024	1.485
1L7T	1I9I	3.263	2.724	2.549	2.845
1NFD	1AD0	1.699	2.002	1.899	1.867
Average		2.751	2.583	2.503	2.612

Table 6 – RMSD values of structure prediction of test set for traditional homology modeling. The parent molecule was chosen based on sequence similarity alone and three models were made.

During each test, three models were made for each prediction. With these data, we can see if there are any trends based on which prediction is closest to the actual, if any. This data will be summarized later in this section once more test sets have been presented. Each prediction was then viewed in PyMol and aligned with the known structure for the antibody and the resulting root mean square deviation (in angstroms) was recorded.

For the remaining data, an apex and base were selected individually. The first set of data produced selected the base parent molecule by comparing sequence similarity and adding bonuses based on interactions described in the methods section. The apex for this test was selected based solely on the sequence similarity. As mentioned above, two scoring algorithms were used to compare the sequence similarity. The first approach usied ClustalW

(Thompson *et al.* 1994), aligned the two sequences and parsed the score from the output. The second algorithm is the Needleman-Wunsch (Needleman *et al.* 1970) dynamic programming algorithm, doing a more complete analysis of amino acid similarity. The latter does an exhaustive search of possible alignments and selects the optimum alignment, and therefore will produce an accurate score.

Test Antibody	***Base Parent***	***Apex Parent (ClustalW)***	***Apex Parent (NW)***	***ClustalW Average RMSD***	***NW Average RMSD***
1BAF	1CLY	1F3D	1DZB	1.587	1.472
1BQL	1AXS	1H3P	1AHW	2.540	1.737
1CIC	1LO4	1JPS	7FAB	1.699	2.227
1CT8	1FRG	1FRG	1FRG	1.628	1.628
1DBA	1MHP	1KNO	1IQW	2.967	2.277
1H0D	1KFA	1OTS	1FNS	2.714	3.107
1IKF	1B2W	1FOR	1FNS	2.676	4.085
1MFB	1BZ7	1IGT	1IGT	2.501	2.501
1NGW	1N7M	1AJ7	1AJ7	0.591	0.591
1Q9O	1QKZ	1UWX	1HYX	3.046	3.518
1SBS	1I8I	1BM3	1FNS	3.167	2.589
1UZ6	1A2Y	1IEH	1CBV	2.316	2.472
2H1P	1A6T	1I8I	1I8I	2.872	3.021
1WT5	1A14	1A3L	1IGT	3.199	2.412
2PCP	1A14	1TET	1NCW	3.246	2.518
2A77	2A1W	2A1W	2A1W	1.987	1.987
1XIW	1AY1	1FBI	1J05	3.521	2.950
1EGJ	1CLY	1A3L	1MHP	2.365	2.465
1G9M	1GC1	1GC1	1GC1	4.770	4.770
1KC5	1A14	1KCR	1KCR	1.843	1.843
1L7T	1I9I	1AY1	1I9I	3.278	2.845
1NFD	1AD0	1PLG	1OB1	2.453	2.764

Table 7 – Test set of antibodies predicted structures against known for two apex algorithms. The base parent molecule was chosen based upon sequence similarity, awarding bonuses for stated amino acids present. Apices are selected by sequence similarity alone. The average RMSD of three predicted models is shown. The ClustalW column signifies a scoring algorithm based on using ClustalW. The NW column represents those apices that are selected based on the Needleman-Wunsch alignment algorithm. The base parent is the same between both algorithms since the base selection algorithm was not changed between the two test sets. The rows highlighted in gray are those where the apex did not change between the two algorithms.

Table 7 shows the data produced using both these strategies. Only the average RMSD is given for each set of three predictions. Although only the average for these data is given here, the differences between runs one, two, and three will be reported and discussed later.

The data in Table 8 shows the results for the next stage of the assessment. During this test, again both ClustalW and NW algorithms were used in retrieving the apices. In addition, one interaction in particular was looked at. The interaction of 100b (or 100a in some cases) of the heavy chain with 49_L was tested in this set. Since this interaction is not present in all cases, and, within this test case, only affected a few of the antibodies, only the test CDR-H3 molecules whose apex or base parent changed will be presented. Those data that are omitted can be assumed to have the same data as in Table 7. Since this interaction can affect either the base or apex of the antibody (depending on the length of the antibody and therefore where 100b lies within CDR-H3) both the base parent and apex parent are shown. It must also be noted that none of the parent molecules changed (for either base or apex) for the ClustalW selection algorithm and therefore are also omitted from Table 8. This will be discussed in a later section. We can also see that none of the bases changed from the previous test set, only apices.

Test Antibody	***Base Parent***	***Apex Parent (NW)***	***NW Average RMSD***
1H0D	1KFA	1LMK	2.814
1IKF	1B2W	1DFB	3.493
1SBS	1I8I	1IBG	2.673
2H1P	1NC2	1I8I	2.865

Table 8 – RMSD values of changed predictions with 100b interaction. The test antibodies that did not change from the last assessment point to this one were omitted from this table. Also omitted are the results from the ClustalW apex scoring algorithm due to the lack of change from the last assessment. Shown is the average RMSD from three predicted models from Modeller.

Also included in the parent search algorithm to choose the best base and apex is the overlap between these two molecules. As stated above, the overlap was used to capture the angle at which the apex is presented from the base molecule. Therefore, when looking for a correct apex, an overlap from a putative apex that matches the unknown molecule's base would score a bonus. At this point in the development, the ClustalW algorithm was dropped due to the inadequacy of the prediction from this apex selection method and only the NW algorithm was used. This decision is further discussed in the next section.

The data contained in Table 9 describe the prediction method after adding the overlap bonus. The bonus is given to the apex, since the apex (in this classification) has most of the variability and should have a similar take off point. The overlap is defined as the two amino acids most near the apex for the base and the two distal amino acids for the apex. Together, these four amino acids are searched against potential base and apex molecules and the apex is given a bonus if such a match is found. Again, only the test antibodies that changed from the previous assessment are included in table 9.

Test Antibody	***Base Parent***	***Apex Parent (NW)***	***NW Average RMSD***
1CIC	1LO4	8FAB	1.387
1H0D	1KFA	1A3L	2.720
1IKF	1B2W	2A1W	5.005
1MFB	1BZ7	1A3L	1.552
1NGW	1N7M	1N7M	0.591
1SBS	1I8I	1J05	3.167
1WT5	1A14	1FOR	2.649
2PCP	1A14	1S3K	1.911
1XIW	1AY1	1BJ1	3.174
1NFD	1AD0	1MHP	1.867

Table 9 – RMSD values of changed predictions with overlap bonus. This data set includes those antibodies whose apex parents changed from the addition of the overlap bonus. This assessment includes only the NW apex scoring algorithm (the ClustalW scoring algorithm was eliminated for the remainder of the tests).

Even after the sequence similarity scoring, 100b (100a) – 49_L, overlap, and structurally important amino acid bonuses, there are sometimes ties for the highest scoring apex or base parent molecules. In the case of a tie, the algorithm will search the VH (variable heavy) germline, from which the molecule came from and pick the base or apex if the two molecules came from the same germline VH gene, the algorithm would choose this molecule. Both human and mouse VH genes are used to compare, since we are unsure from which species the antibody came from. It is also possible that the sequence is engineered. The algorithm searches through 228 VH genes and finds the sequence with the highest similarity. Although an exact match of VH genes is ideal, if there is not such a condition, the family of the VH genes are then compared and a molecule is chosen if it is in the same VH gene family. If neither of these rules apply, then the algorithm uses the old method of selecting the molecule with the highest sequence similarity to the target protein.

Table 10 shows the test set of antibodies after the addition of the 'VH gene tie selector' portion algorithm. Again, the antibodies that did not show a change at this assessment are omitted from the table.

Test Antibody	***Base Parent***	***Apex Parent (NW)***	***NW Average RMSD***
1UZ6	1A2Y	1CBV	2.064
2PCP	1A14	1S3K	1.428

Table 10 – RMSD values of changed predictions after VH gene tie rule. This data set includes the two molecules thats chosen parents were changed due to the 'VH gene' tie rule.

Discussion

The objective of this study is to look at various factors regarding the structure of the third complementary determining region of antibodies. There were many considerations and assumptions made during this study and this section will take a look at those in context of the data presented. At many stages during the study, assumptions were made based on research and experience gained while doing the experiment. Here, these considerations and assumptions are discussed through each step of the study.

Although antibodies have been studied extensively over the past few decades and the domains of heavy chains have been characterized well, automatically parsing out the CDR-H3 region is not a trivial task. Many of the structure files of antibodies publicly available contain antigen structures and the chains are not labeled with a standard; making it necessary to identify the heavy and light chains by sequence alone. In this study, a Hidden Markov Model (HMM) was used to search the putative heavy or light chain. As seen in Table 4, the heavy chain HMM was tested by searching various sequences known to be heavy chains, light chains, or a random unrelated sequence. The known heavy chains are common heavy chains that have recognizable characteristics, such as a common framework region. These scored very high on the model. The heavy chain of 1ETZ was similar to the other heavy chains, except that it was slightly shorter, like some of the publicly available sequences found. This heavy chain scored slightly lower at 350 when compared with the 450+ of the other two heavy chains.

The HMM was also compared with antibody light chains. A light chain is relatively the same length as a heavy chain when compared with the wide range of protein lengths and also share a similar overall three dimensional shape, with three antigen binding loops pointing into the antigen binding pocket. It was also necessary to be sure that the HMM could discriminate between a heavy and light chain. The scores for the three light chains shown in Table 4 were much lower than the heavy chain examples used, but slightly higher than the unrelated sequences.

The unrelated sequences were chosen completely at random with no regard for length or sequence content. Although this was a very small sample, it is clear that the HMM can discriminate between a heavy chain and another sequence, be it a light chain or otherwise. The light chain HMM was tested in much the same way, producing comparable results.

A cutoff score of 150 was chosen for using the HMM, only to be completely confident that the protein chain in question was indeed a heavy chain. If no chain scored above a 150 from an HMM search, the highest scoring chain is printed out with the score for user inspection. Many times, the sequence is truncated in the PDB file and therefore much shorter than the expected heavy chain. For example, the antibody 1AP2 contains a heavy chain that ends only a few amino acids after CDR-H3, and therefore produces a score of 106 against the HMM. This happens in a few cases, and it is often caused by a truncated heavy chain.

A remedy to this situation would be to include shorter heavy chain sequences in the alignment into the HMM. This was not done because this would also reduce the specificity of the heavy search. In this case we would much rather not include an unusual heavy chain, than to include a false positive where the CDR-H3 could not be parsed out or false data be put into the high resolution database.

The high resolution database contains all PDB files that have been identified by the “Self-Maintaining Database of Antibody Crystal Structures” or SACS that contain a resolution of 2.5 angstroms or higher. This number was chosen to follow with previous studies done on CDR-H3 structure by Shirai *et al.* (1996) and Morea *et al.* (1998). This is just above the length of a hydrogen bond (1.98 angstroms) and therefore we can be confident of the data present in these files.

The test set developed (shown in Table 2) needed to have sufficient diversity to represent a wide range of CDR-H3 regions. Therefore the test set needs to be of an appropriate size to incorporate this diversity, but larger is not necessarily better. Since this study bases its predictions on a limited set of molecules, a decrease in the number of antibodies in the high resolution database could hurt the accuracy of predictions. There were originally 255 antibodies placed in the high resolution database and 23 of those were chosen based on their CDR-H3 length and base type. This leaves 232 molecules to be chosen for putative parent molecules for structure prediction. Although this number would ideally be higher (and of course, it would be ideal for both the test set and possible known structure to be larger), this is all that there is currently available in the public domain. In the future more sequences can be added as new structures are discovered and consequently placed in the PDB.

In Table 2, there are only 22 molecules shown. The 23rd molecule that was omitted from the test set was 1FVC due to its high similarity to the molecule 1FVD. As seen in table 11, the sequences are very similar only differing in the length of the sequence. When used

PDB ID	Heavy Chain Sequence
1FVC	EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRF TISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLVTVSS
1FVD	EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRF TISADTSKNTLYLQMNSLRAEDTAVYYCSRWGGDGFYAMDVWGQGTLVTVSSASTKGPSVFPLAPSS KSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICN VNHKPSNTKVDKKVEPKSC

Table 11 – The heavy sequences for 1FVC and 1FVD molecules.

for prediction on the 1FVC CDR-H3, 1FVD was chosen for both the base and apex (as it should) and the three predictions scored very low, 0.001, 0.000, and 0.005 for the three models produced. The molecules were also produced by the same lab group (the reference papers have the same main author) and were deposited on the same date with the same description. The shorter sequence has a better resolution (2.20 vs. 2.50 angstroms for the longer heavy sequence, 1FVD). The reason these two molecules were deposited separately and under differing PDB ID's was not investigated. Instead, the molecule 1FVC was removed from the test set, so that it would not lower the averages for the RMSD scores falsely.

The test was first run under a more traditional context of homology modeling, where one parent was chosen and used for the basis upon which the unknown is threaded. This will be used as a control to compare the system to as new factors are introduced into searching algorithm.

Throughout the analysis of this project, the measurements are given as root means square deviations (RMSD) in angstroms. This can basically be explained as the average distance between the atoms of each molecule. The side chains were not included in this alignment; only the alpha carbon backbones were aligned. The general shape of the domain is tested in this case, instead of the specific side chain positions. Since the entire molecule is not being modeled, it would be difficult to enforce a certain side chain position if it were to have an interaction with another molecule not present in the structure prediction. For example, if a tyrosine in the CDR-H3 of the molecule being modeled were to natively interact with an amino acid in the light chain, there would be no information for the threading program to correctly place the tyrosine in the correct orientation. Due to this fact, they will be left out of the final alignments and therefore will not count for or against the model.

The results for the 'one parent' model set varied. The range of predictions were as low as 0.506 Å(1NGW, Run 1) and as high as 8.507 Å(1IKF, Run 1). The two molecules that produced these results definitely seem to be outliers when looking at the rest of the data. These values are in angstroms, and at this point, one may want to compare these values to known lengths in angstroms, such as the hydrogen bond length of 1.98 angstroms. It would seem then, that these predictions (an average of 2.577 angstroms) would be very close, just over the length of one hydrogen bond. But we must also take into account into our analysis that the molecules we are modeling are very short sequences, and the set of putative parents are generally of the same shape and have similar function. It would be expected for the values produced by such a method to be lower than for modeling an entire molecule. On the contrary, the CDR-H3 is the most variable out of the six hyper-variable regions and has evaded very accurate structure prediction in the past.

As seen in figure 3, the prediction of 1NGW molecule from Run 1 is shown. We can see that the prediction is quite close to the known structure of 1NGW. The predicted molecule is shown in blue, where as the known structure is shown in green. It would seem that the majority of the difference accounting for the score would be at the two termini of the molecule. For both the C and N termini, the predicted points one way while the actual goes the other. This may be a consequence of the parent molecule CDR-H3 launching off the

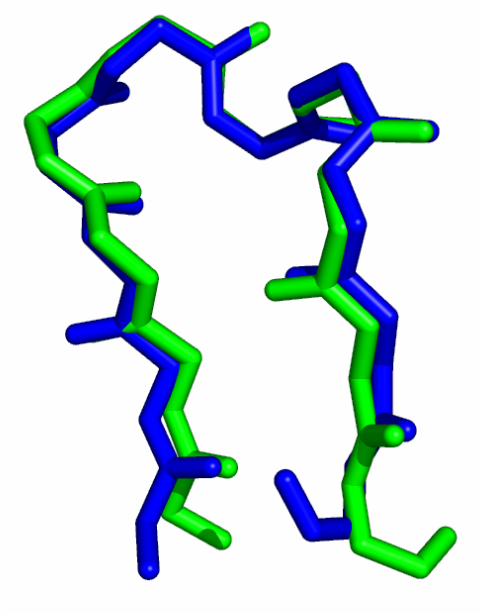


Figure 3 – The predicted 1NGW molecule aligned with its known structure. Predicture structure shown in blue, known structure in green. Model generated with PyMol.

framework region at a different angle the 1NGW CDR-H3. Since this section of the region is not looked at, it would make sense for there to be diversity in the predictions at this section of the region.

Figure 4 depicts the predicted CDR-H3 (1IKF) using 1FN4 as a parent. The score for this prediction was 8.507 from the first out of three runs for this prediciton. There is a huge difference between the two aligned molecules, as we can see. It seems that one of the major differences, which may cause the poor alignment score, is the placement of the hairpin loop. The predicted structure for 1IKF (shown in blue) is very lopsided in its CDR-H3 structure, as opposed to the known structure (shown in green). There seems to be many more amino acids on one side of the hairpin turn than on the other, and on th e known structure there seems to be an equal number. We can also see a break in the backbone chain for both the prediction and known structure. The missing amino acid in the prediction structure could have been caused in one of two ways. The first would be caused by a gap produced in the alignment given to the Modeller program. This would cause the prediction to have a break in the backbone such as seen in figure 4. In the case of the break in the known structure, this is most likely caused

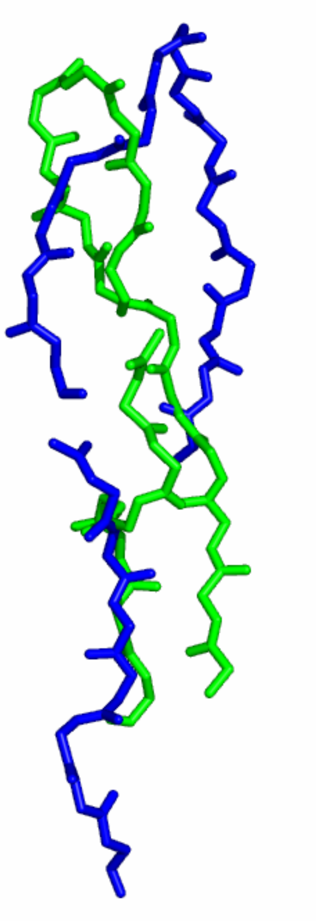


Figure 4 – 1IKF prediction aligned with known structure. The predicted 1IKF CDR-H3 using 1FN4 as a parent in blue. The known 1IKF structure depicted in green.

by missing amino acid three dimensional information. It is interesting to note that there are both breaks in near same positions in both the predicted and known structures. Anything more than coincidence as the cause for this cannot be determined.

If we look at these two predictions (for 1NGW and 1IKF, representing the best and worst) and compare the predictions with their CDR-H3 lengths, we can see that our highest and lowest scoring RMSD values were also the second to longest and shortest CDR-H3 values respectively. It has been stated that longer CDR lengths are more difficult to predict an appropriate structure (Morea *et al.* 1998). We can see a summary of the CDR-H3 lengths versus the average prediction scores in figure 5.

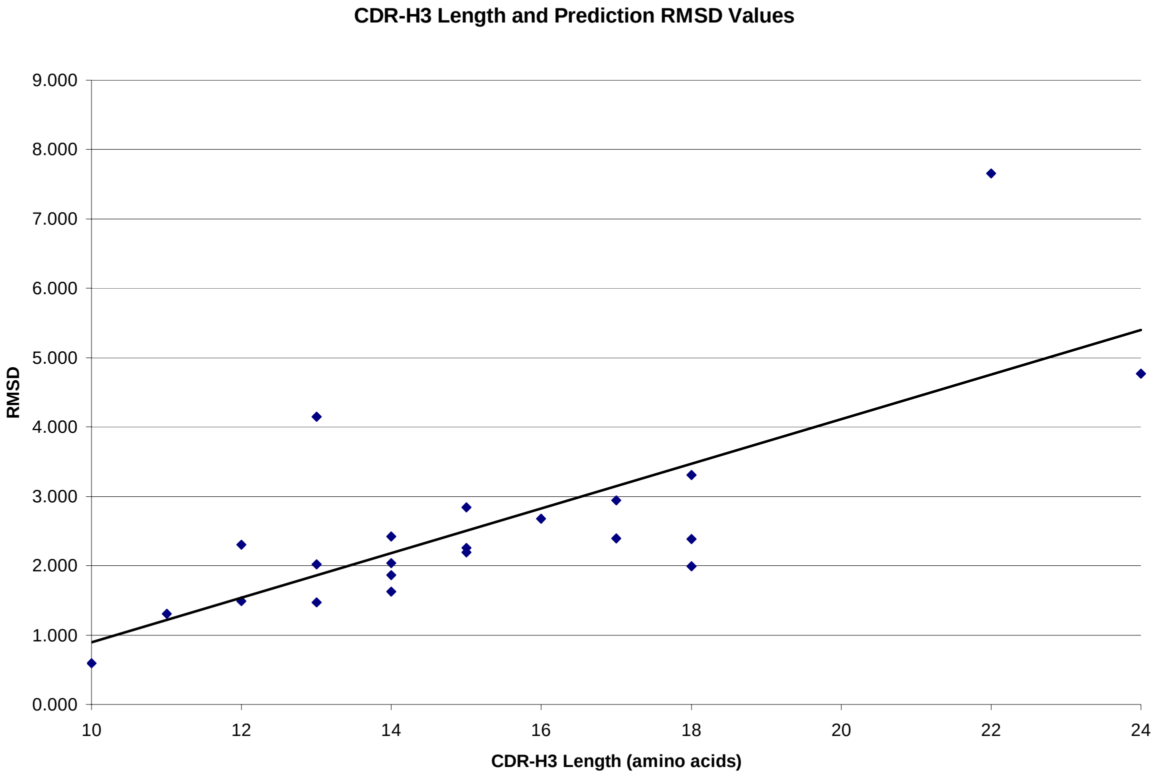


Figure 5 – CDR-H3 Length and Prediction RMSD values. Using data shown in Linear trend line was added to show general relationship of data.

Figure 5 shows the relationship between CDR-H3 length and the root means square deviation from the known structure. A general trend can be seen, such that the prediction is closer to the known structure as the CDR-H3 gets shorter. This could be due to the fewer conformations the region of protein can take as fewer amino acids are present. There were a few regions that tend to stray from the general region, such as 1UZ6 which was recorded at 4.150. We will see if these predictions can be improved as other criteria for parent selection are added.

One assumption made in this study was that choosing a parent base and parent apex region separately to model on would make the prediction method more flexible. The first set of data produced from this algorithm involved two separate searching methods, one for the base and one for the apex. The apex method involves a sequence similarity search based on two different methods as described in the Methods section. The two methods will be referred to as the ClustalW method and the other as Needleman-Wunsch method, referring to the alignment and scoring implementations. The base scoring algorithm that produced the data for this assessment involved an initial sequence similarity comparison and bonuses for base types of the same class (as described in Shirai *et al.* 1996). Figure 6 describes these data and summarizes the data seen in tables 6 and 7 (in results).

The comparison that needs to be made is that between the single parent selection algorithm against the separate base and apex parent selection algorithms. In most of the cases (16 out of 22), the single parent CDR-H3 algorithm performed slightly better than at least one of the two apex and base separated algorithms. This does not support the hypothesis that selecting separate base and apex regions will increase the accuracy at which we can predict the target's CDR-H3 region's structure. The overall average of the three algorithms (taken by averaging each run for each molecule and then combing those to find the overall average for that run) were 2.612 for the single parent, 2.589 for the ClustalW algorithm and 2.529 for the NW algorithm. This is a crude approximation for the overall performance for the system. Since there were only a few test proteins, more complicated statistical analyses would not prove useful as well. Instead, it should prove beneficial to look at places where the program has improved or gotten worse and try to account for this happening in the system.

As mentioned above, in many cases the single parent algorithm out performed the other two methods by a slight amount. A more obvious observation of the data is the drastic improvement for the CDR-H3 region for the antibody 1IKF. In the single antibody prediction the average RMSD was around 7.6 while the ClustalW and NW algorithms produced average RMSD values at 2.6 and 4.0 respectively. It should also be noted that the base parent used for the prediction of this molecule did not change between each of the methods. Only the apex selection algorithm was changed and therefore change in performance of the system can be attributed to the apex. This shows us that the majority of error (or success for that matter) can be attributed to the apex region. Since the base is fairly static and base classes have been characterized by Shirai *et al.* (1996), this would make sense.

Another molecule that showed drastic improvements was 1UZ6. For the one parent prediction algorithm, the average RMSD score was 4.2. For the ClustalW and NW algorithms the predictions score a better 2.3Å and 2.5Å respectively. Figure 7 shows the predictions aligned with their known structures for 1UZ6 for both the single parent algorithm and the NW algorithm (prediction was taken from Run 1 of the three structures output by Modeller). Interestingly, the dramatic improvement of the score can barely be noticed when looking at the actual structures. One structural detail that's difficult to see without being able

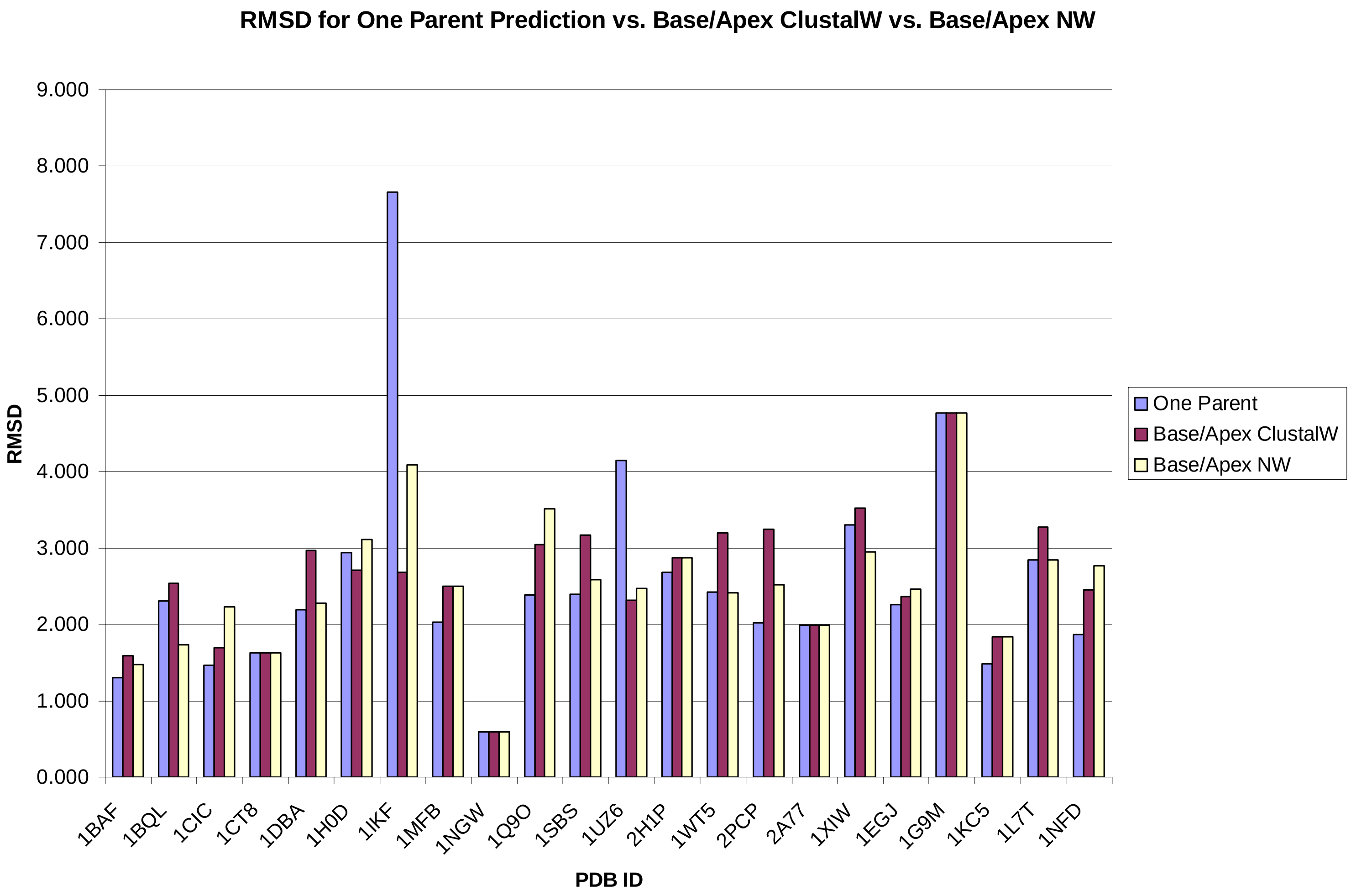


Figure 6 – RMSD of traditional homology modeling compared with two apex selection algorithm. Graph showing the change in RMSD values for the test set. The first series shows the predicted root means square deviation from the known structure for the one parent algorithm (traditional homology modeling). The second series shows the prediction algorithm involving both apex and base selection criteria with the ClustalW apex selection algorithm. The third series represents the NW algorithm.

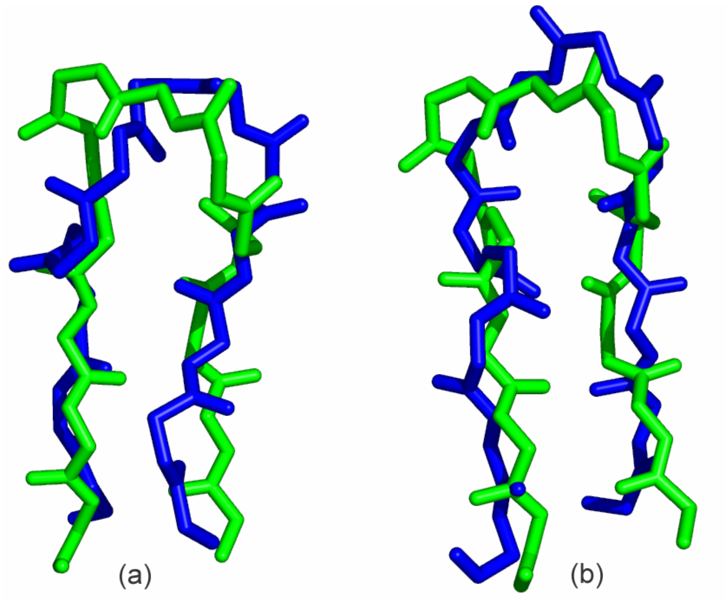


Figure 7 – 1UZ6 prediction comparing traditional homology modeling and base/apex selection. (a) CDR-H3 of 1UZ6 predicted (blue) aligned with know structure (green) using the one parent method. (b) 1UZ6 CDR-H3, predicted (blue) aligned with known (green) using NW method.

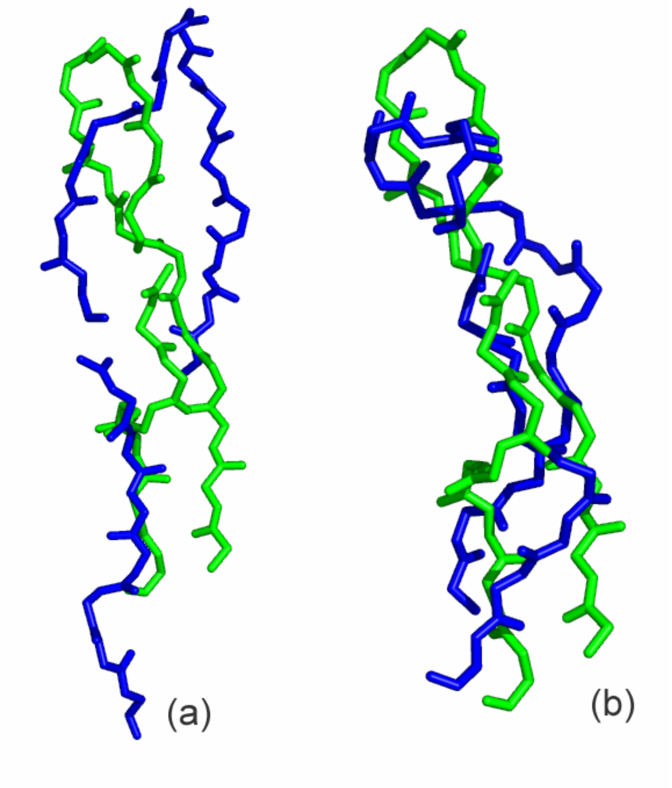


Figure 8 – 1IKF prediction for traditional and apex/base selection algorithms. (a) CDR-H3 of 1IKF predicted (blue) aligned with know structure (green) using the one parent method. (b) 1IKF CDR-H3, predicted (blue) aligned with known (green) using NW method.

to manipulate a 3D model is the angle of the apex. The single parent algorithm missed the prediction angles for the apex off of the base which may have caused a few amino acids to be placed relatively far from their known structure counterparts, which could account for the much higher scores for this prediction.

Figure 8 shows the difference in predictions from the single parent method to the NW method. This prediction caused a drastic change in the predicted molecule. We can see that the predicted molecule actually resembles an antigen binding loop which would account for the much better score.

There were not any notable times when the single parent prediction method significantly beat the other two prediction methods. It seems selecting a base and apex separately dramatically increased the accuracy of the worst predictions, but at the cost of a slight decrease in accuracy of the other predictions.

The next sets of data did not produce a change of parents for the ClustalW method, therefore the numbers were not reported for the rest of the tests, although they were conducted. This was an unwanted side effect of the alignment method and the reason a new alignment technique was developed. After just looking at sequence similarity and base type for scoring, it seems that the ClustalW method is slightly better, although it did not change over the following data sets, and the NW method did. So therefore as the NW selection algorithm got better, it seemed to improve past the prediction accuracy of the ClustalW method.

The failure of the other criteria (100b-49_L interaction, overlap) in changing predictions for the ClustalW method was a result of the ClustalW scoring. The score range was very large in comparison to the base alignment scoring, and also did not scale in a linear fashion. For the algorithm, the scores were adjusted by dividing by the number of amino acids present in the unknown CDR-H3. It seems that this score was also too inflated, and the bonuses that were given seemed to be drowned out by the alignment score alone.

Also at this point, it was realized that the ClustalW algorithm does not penalize an alignment with gaps at the end. This is an undesired way to score apices for modeling because it is known that there are classes of β-turns (the structure of the apex) that are based upon the length of the apex. If gaps at the end (or beginning) of the sequence are not penalized, an apex may be chosen that is of different length, and would have an equal or higher score than an equally similar apex of the same length. Therefore, the Needleman-Wunsch algorithm was implemented as a replacement and the scoring was therefore lower and would allow the input of the various criteria to help select a suitable apex. Gaps at the beginning and ends of sequences would also be penalized, therefore favoring an apex selection of the same length and thus a more favorable, theoretically, scoring algorithm.

With the introduction of new criteria, the 100b – 49_L interaction, only a few apices changed. Although only four predictions changed with this addition (table 8), this was higher than expected. Since this interaction is very specific (requiring a specific set of amino acids in positions 100b and 49_L, or 100b, 100a and 49_L) it was expected that only one or two, if any, predictions would contain this interaction. From the results reported, it can be seen that at least four of the molecules contain this interaction and it played a role in selecting a new parent. This does not mean that the conformation described by Morea *et al.* (1998) was present in all these molecules, but it is assumed that they would. It would seem that this interaction was more common than originally thought, being present in 18% of the test set. Again, since the test set is very small, and the entire set of antibodies was not sampled (only those that have well determined structures), this can not be seen as an accurate representation of the whole population.

Figure 9 shows compares the RMSD values of the prediction method for the one parent traditional method, along with the apex/base NW prediction with and without the 100b-49_L interaction.

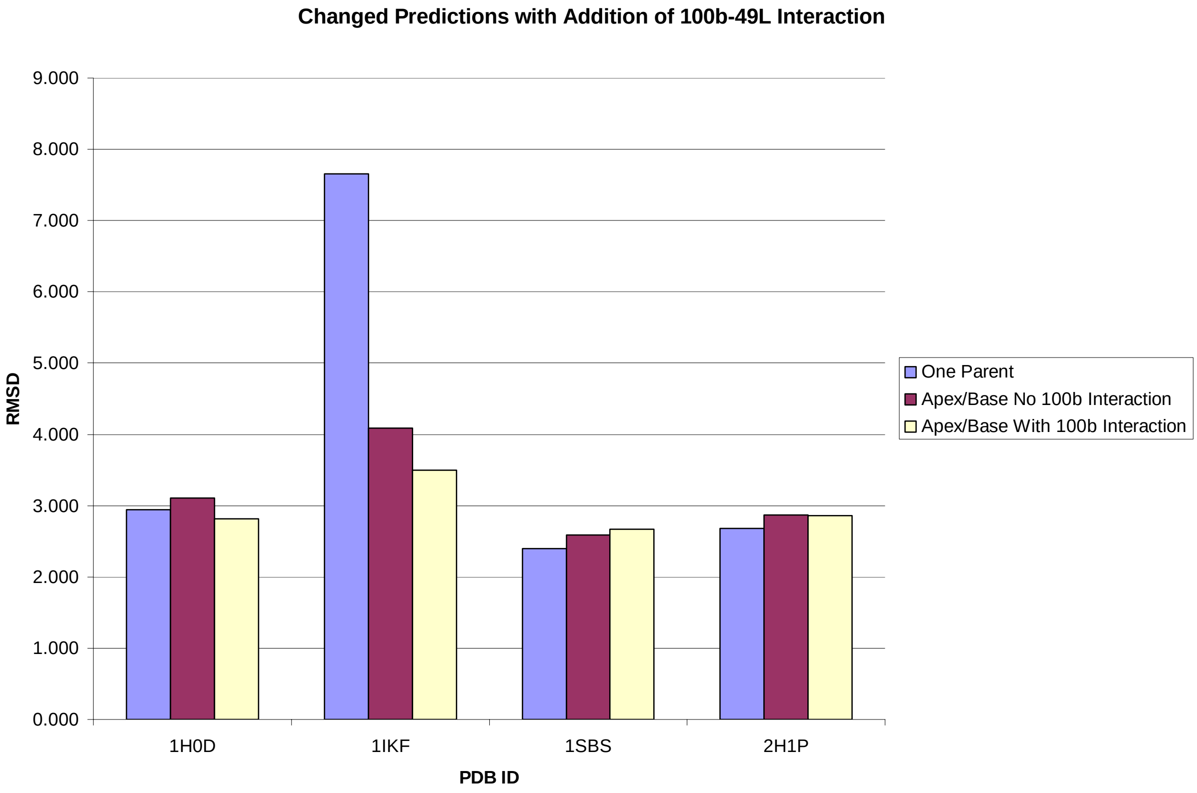


Figure 9 – CDR-H3 region predictions changed with the addition of the 100b-49L (white) interaction, shown against one parent prediction (blue), and without the 100b interaction (purple).

The amino acid in position 100b can be located in either the base or the apex of a CDR-H3. Therefore, it is possible for either the base or the apex to be changed by this step. For the first three proteins in the table (1H0D, 1IKF, 1SBS) the apex was changed. The protein 2H1P was the only CDR-H3 whose base was affected by this change in the algorithm. As seen above, the introduction of a separate base and apex scoring algorithm greatly increased the prediction accuracy for the relatively long 1IKF CDR-H3 region. The 100b – 49_L for this molecule also increase the average RMSD score by more than half an angstrom.

The new apex selected for the 1H0D prediction (1LMK) seems also to be a better choice, lowering the average RMSD value by almost half an angstrom. This model also seems to be an improvement over the parent template molecules chosen by the one parent selection algorithm, although only slightly. In the other two molecules, there was only a slight change from the other two prediction methods.

In order to put these changes in perspective, taking a look at the predicted structures compared with the known that revealed the RMSD values is important. As an example, lets take a look at molecules 1IKF (again) and 1SBS. The 1IKF prediction benefited from the addition of the 100b – 49_L interaction and 1SBS seemed to reveal a more inaccurate prediction. Figures 10 and 11 should reveal the scale that the average RMSD score changes represent.

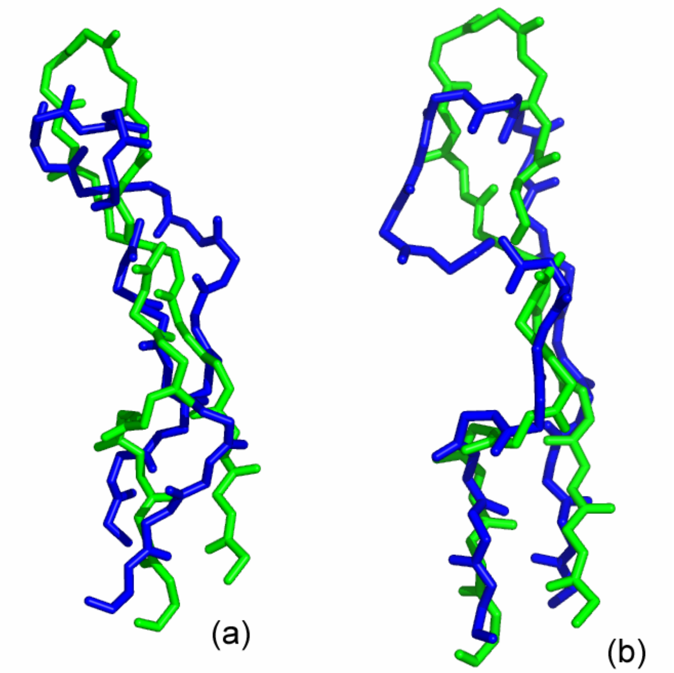


Figure 10 – 1IKF structure alignment with known with and without 100b bonus in selection. 1IKF predictions in blue, known 1IKF structure in green. (a) Prediction based on parent selection algorithm without 100b interaction. (RMSD score of 4.085) (b) Prediction base on parent selection algorithm with 100b interaction. (RMSD of 3.493)

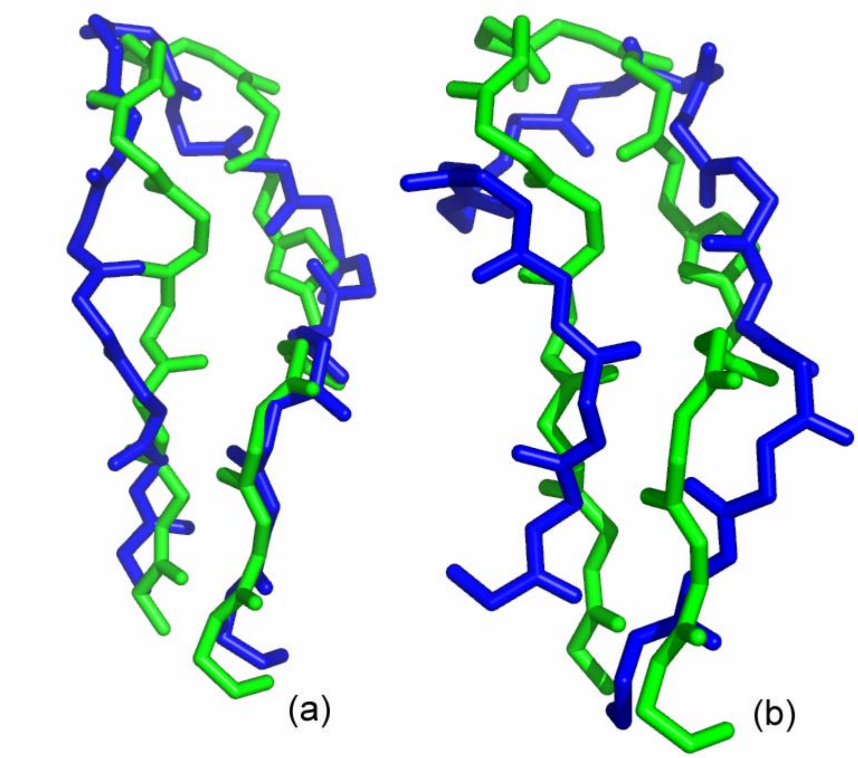


Figure 11 – 1SBS structure alignment with and without 100b interaction bonus. 1SBS predictions in blue against known 1SBS structure in green. (a) Prediction based on parent selection algorithm without 100b interaction. (RMSD score of 2.589) (b) Prediction base on parent selection algorithm with 100b interaction. (RMSD of 2.673)

The improvement in the alignment of the known molecule with the predicted for 1IKF (figure 10a) seems to be concentrated in the base of the CDR-H3 molecule. The predicted molecule (using the 100b interaction) generated a molecule with a very similar base region, but a lot of variation in the apex region. This result is exaggerated by the way the two molecules were aligned. The region of the apex nearest to the base also seem to align fairly well, with the tip of the apex slightly wider and pointed in a different direction accounting for the majority of the difference.

The 1SBS molecule, which resulted in a slightly worse RMSD value after the addition of 100b interaction in the selection algorithm, seems not to have a large change between the two prediction methods. It can be seen that the new apex selected for this prediction did change the structure of the molecule, but did not see an improvement, which we expected to see given the small change in the RMSD values. Overall, the 100b interaction did not drastically change the prediction accuracy in these selected regions, with the exception of 1IKF; although the predicted apex of 1IKF is very different from the known structure.

To ensure a good fit of the apex upon the base, the overlap between these two regions was used in the scoring algorithm. A bonus is given to the apex of a potential parent apex if

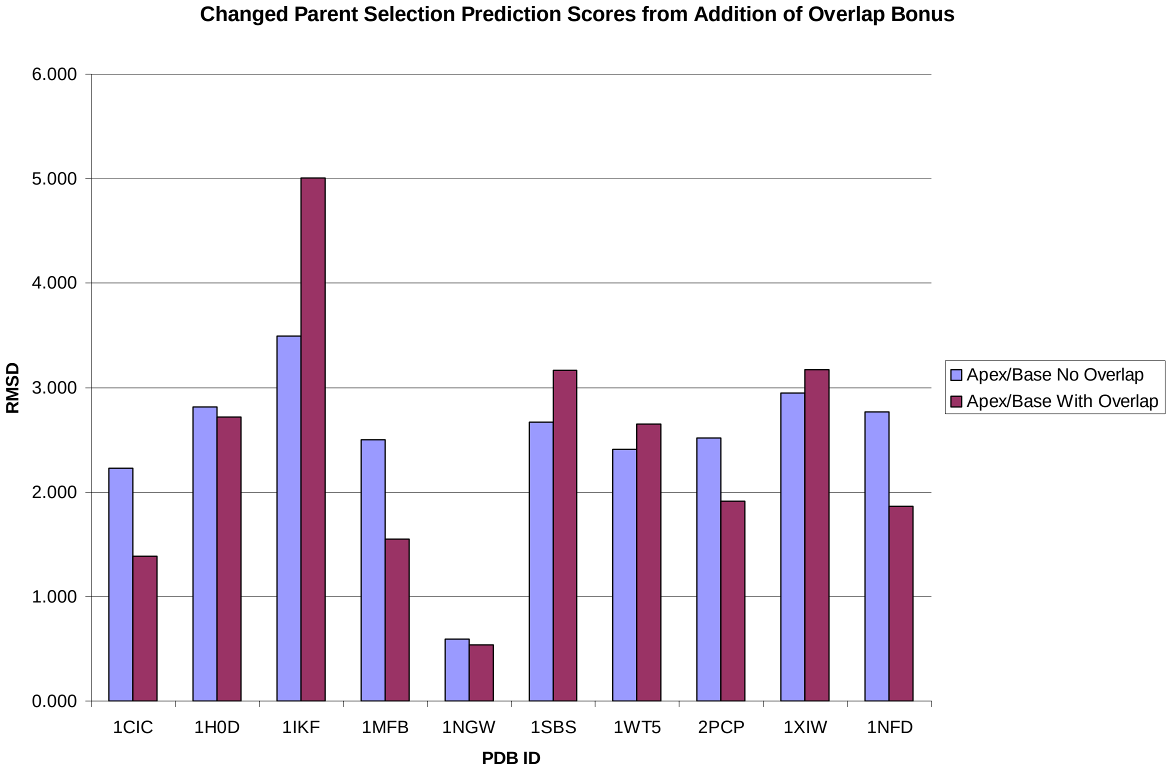


Figure 13 – Changed prediction RMSD values with addition of overlap bonus. This graph includes those predictions that were changed by the addition of the overlap bonus in the apex parent selection algorithm. Shown are the RMSD scores before and after the addition.

the overlap of this molecules base and apex match those of the unknown CDR-H3. In theory, by this would allow the selection of an apex to have a similar take off region and therefore share a similar angle that the apex is presented by the base. As seen in figure 12, the results for this section were mixed.

The parent apex for the prediction of 1IKF was changed, once again, for this section of the molecule. As discussed above, the apex of this prediction seemed to contribute the most to the high RMSD value. With the addition of the overlap and using 2A1W for the prediction of the apex, the RMSD value actually went up, resulting in a worse prediction. Figure 13 shows this new alignment with the known. Although, it is interesting to note that the average for the three predictions (5.005) for this section is very misleading. The scores for each of the three predictions output by Modeller were 3.723, 7.843, and 3.809. Shown in figure 13 is the first prediction given.

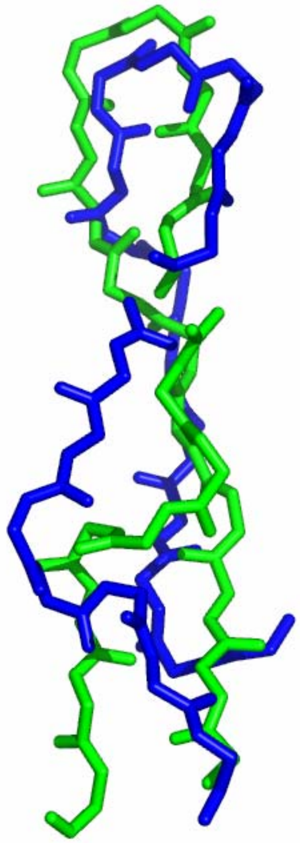


Figure 13 – Prediction of 1IKF using overlap bonus. This shows the first prediction output by Modeller, using the parents 2A1W for the apex and 1B2W for the base. The RMSD score for this alignment was 3.723.

Although the base of the prediction in figure 13 does not align very well, the apex seems to be much closer in shape. Again, if the base amino acids were aligned with known structure from the prediction, the angle at which the apex comes off from that base would be much different from the known.

The final addition to the parent selection algorithm is used to break a potential tie in the selection of a base and apex. If there are more than one apex or base selected for a prediction, the germline VH gene is identified for those high scoring parents. The parent with a matching VH germline gene, or a germline gene from the same family is then selected for modeling. Table 10 shows the two CDR-H3 regions that were affected by this tie breaking rule. Figure 14 shows the change of RMSD values for these two predictions after the new base was selected.

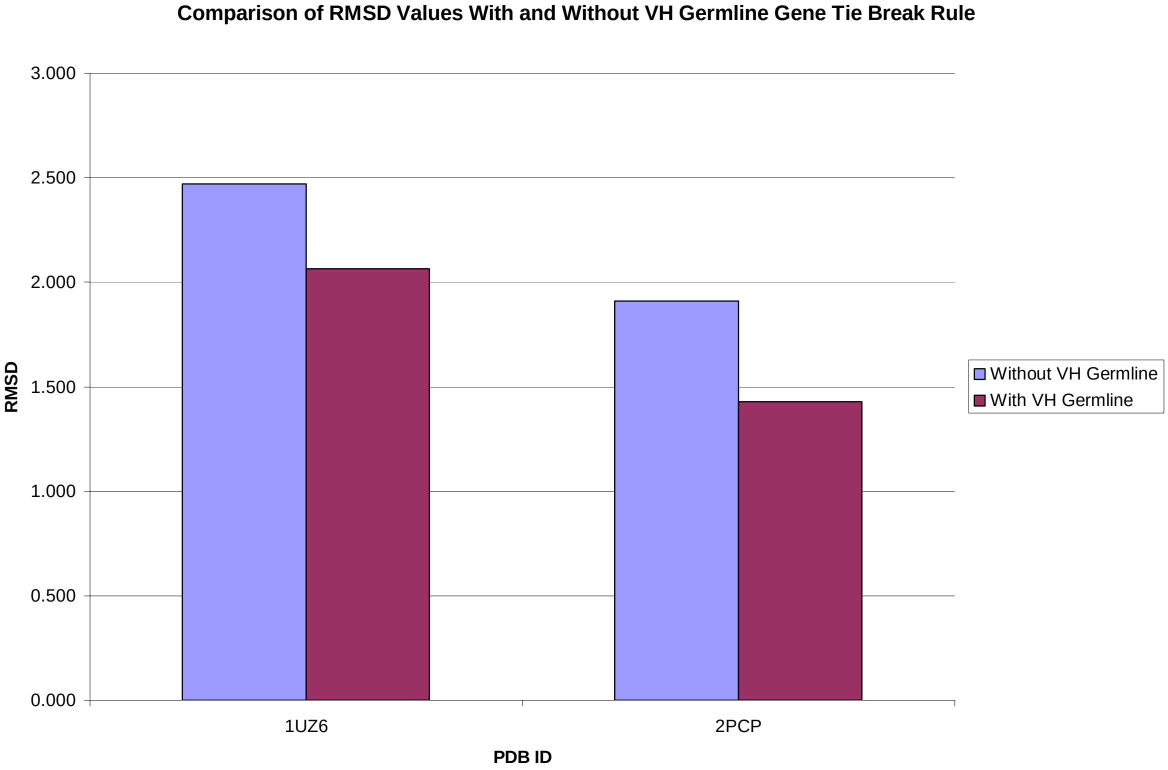


Figure 14 – Changes in prediction RMSD values with addition of VH gene tie break rule. The two CDR-H3 predictions that changed due to the VH Germline tie break rule.

In both cases the RMSD values were lowered slightly when the VH gene information was included in the selection algorithm. In a B cell, the VH gene does not code for the CDR-H3, but instead, codes for the amino acids directly preceding this region. In both cases stated above, the base parent changed and the scores improved; this may be due to some interaction with the portion of the protein that is coded for by the VH gene. This interaction would be impossible for Modeller to predict, because it only receives information about the CDR-H3 and models only based upon this region.

With some of the data presented, it is not a stretch to think that the more valid information the prediction program is given (in this case, Modeller), the better prediction it will yield. Also, we can also see from the data that the longer the region given to the prediction program, the more possible conformations the protein as a whole can take, therefore creating a much larger problem. Perhaps there is some middle ground, where more important amino acids can be given to a prediction program that have proven to be relevant in the structure of the molecule, without overloading the prediction software with masses of amino acids. This is taken from the observation that Modeller would be unable to predict many interactions on the basis that all the amino acids were not given to the program, as seen in the case of the 100b – 49_L, where the amino acid at position 49_L was not even given to the prediction algorithm, but was hoped that this conformation would be adopted from the parent.

Another interesting observation of the data discussed earlier in this section was the direct relationship of RMSD value and CDR-H3 length (which maps directly to apex length, since the base is constantly 10 amino acids long. Figure 15 shows the final RMSD values (including the 100b-49_L bonus, overlap bonus, and the VH germline tie break bonus each with using the NW alignment scoring for apex selection) against the length of the CDR-H3.

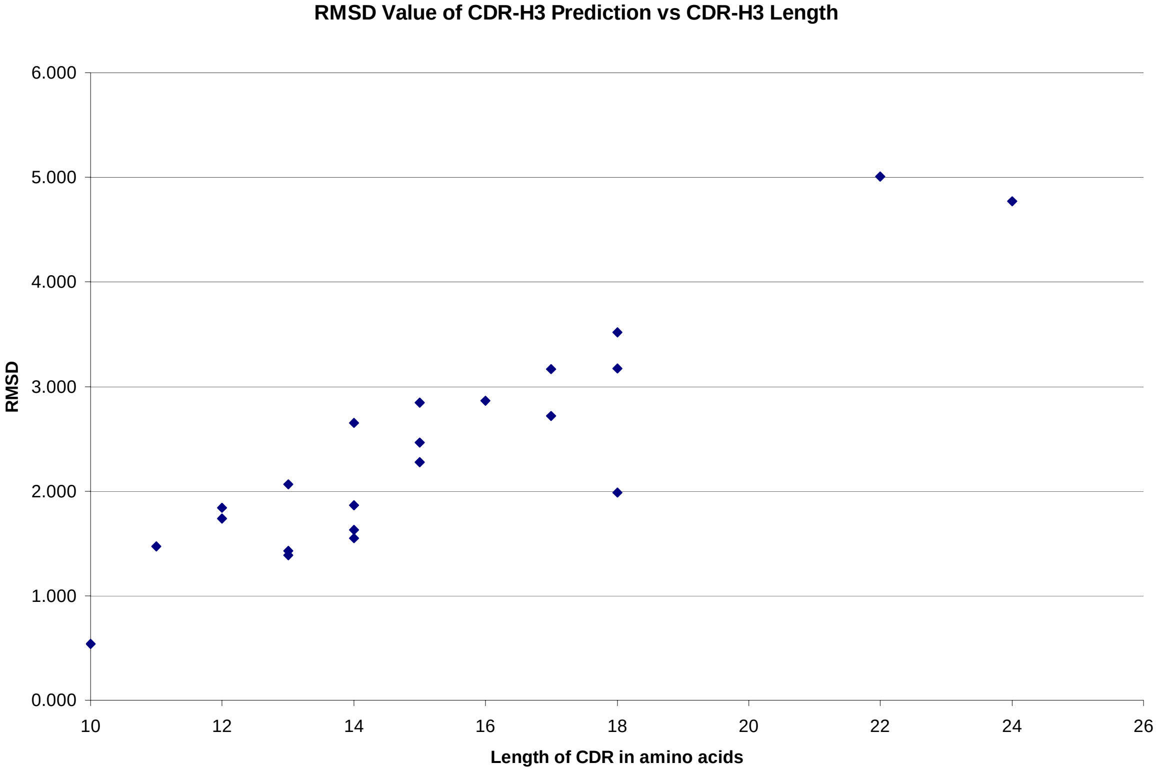


Figure 15 – Final RMSD predictions against the length of the CDR-H3.

As was expected and seen in past studies, the prediction seems to get worse as the CDR-H3 gets longer. In the final assessment we did lessen the range of RMSD from the known, by reducing 1IKF by almost 3 angstroms. The other CDR-H3 predictions resulted in relatively the same quality as with the one parent, most having a small improvement in the RMSD. As seen in the visualizations included in this section we can also see that the apex is the main hurdle in determining full CDR-H3 structure, especially when the apex is longer than 10 amino acids, as seen in molecules 1IKF and 1G9M (22 and 24 amino acids respectively). The apices for these molecules were difficult to predict, but the model of 1IKF was dramatically improved by introducing a separate apex and base selection system.

Conclusion

Many of the interactions and structurally relevant information incorporated into the parent selection algorithm for homology modeling of the CDR-H3 region have shown to improve the modeling capabilities of the Modeller program in isolated cases. There was no overall improvement throughout the entire test set, just isolated cases where the RMSD was improved (or in some cases the quality of prediction was decreased). From such a small test set, it is difficult to tell if separately selecting the base and apex as parents is of benefit. It can be said though, in 1IKF, this technique dramatically improved the prediction made by the Modeller program.

It can also be said that as CDR-H3 lengths increased, so did the RMSD values for the prediction of their molecules. This general trend (seen in figures 5 and 15) was fairly obvious given the data and also has been reported prior to this study (Shirai *et al.* 1996, Morea *et al.* 1998).

Since homology modeling relies on a large set of proteins to draw their own structure from, a larger high resolution database of antibodies would only benefit this system. A larger diversity of antibody lengths and conformations could lead to the selection of a parent more closely related to the region being modeled. This is happening over time, as more and more antibodies are being sequenced and their structures are being discovered. There were also a smaller proportion of longer CDR-H3 regions in the protein databank, which could have also contributed to the poor structure prediction of the longer regions.

The test set used, although relatively small, showed diversity in modeling accuracies, from the dramatically incorrect to very close models. If the dataset were able to be increased, this would more confidently reveal the real accuracy of the system. Due to the small number of high resolution sequences available at the time of the study, the test set could not realistically be increased in numbers. It would also be of benefit to the system if the number of antibodies to be modeled on was increased. This would also be interesting to test, to see if the results found here were due to poor selection of parent molecules or rather, if there was not a suitable parent available to be modeled on. If it were reasonable, we could find the parent apex and base that would be responsible for the best homology model view these sequences and determine characteristics, but this would take a lot of time and computational power.

A confident, automated method for parsing out the highly variable CDR-H3 region was developed using modern bioinformatics techniques. Although parsing this section out by eye has been easily done in the past, automatic computational methods for parsing this region from the rest of the sequence have proven difficult for even the most flexible regular expressions.

Overall, we did not see a huge improvement of modeling using this system when compared with the traditional homology method. Although, individual improvements, as well as declines in accuracy, have been seen and discussed. The possible reasons for the accuracy of the predictions vary for each molecule or set of molecules. Perhaps a larger, more diverse set of antibody structures are needed to accurately determine a set of rules.

Works Cited

Allcorn, L.C.; Martin, A. C. R. "SACS – A Self-Maintaining Database of Antibody Crystal Structures", *Bioinformatics* **18,** 175-181 (2002).

Al-Lazikani, B.; Lesk, A.M.; and Chothia, C.; "Standard conformations for the canonical structures of immunoglobulins.", *J. Mol. Biol.* **273**:927-948 (1997).

Chothia, C.; Lesk, A.M.; "Canonical structures for the hypervariable regions of immunoglobulins.", *J. Mol. Biol.* **196**(4):901-917 (1987).

Culler, S.; Hsiao, T.R.; Glassy, M.; Chau, P.C.; "Cluster and information entropy patterns in immunoglobulin complementarity determining regions.", *BioSystems.* **77**:195-212 (2004).

Decanniere, K.; Muyldermans, S.; Wyns, L.; "Canonical antigen-binding loop structures in immunoglobulins: more structures, more canonical classes?", *J. Mol. Biol.* **300**(1):83-91 (2000).

Martin, Andrew C.R.; "Antibodies: general" Andrew C.R. Martin's Group at UCL. University College London. October 2005.

Morea, V.; Tramontano, A.; Rustici, M.; Chothia, C.; Lesk, A.M.; "Antibody structure, prediction and redesign.", *Biophys. Chem.* **68**:9-16 (1997).

Morea, V.; Tramontano, A.; Rustici, M.; Chothia, C.; Lesk, A.M.; "Conformations of the Third Hypervariable Region in the VH Domain of Immunoglobulins.", *J. Mol. Biol.* **275**:269-294 (1998).

Needleman S.B.; Wunsch C.D; "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *J. Mol. Biol.* **48**:443-53 (1970).

Oliva, B.; Bates, P.A.; Querol, E.; Avilés, F.X.; Sternberg, M.J.; "Automated Classification of Antibody Complementarity Region 3 of the Heavy Chain (H3) Loops inot Canonical Forms and Its Application to Protein Structure Prediction.", *J. Mol. Biol.* **279**:1193-1210 (1998).

Petrey D.; Xiang Z.; Tang CL.; Xie L.; Gimpelev M.; Mitros T.; Soto C.S.; Goldsmith-Fischman S.; Kernytsky A.; Schlessinger A.; Koh I.Y.; Alexov E.; Honig B.; "Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling."; *Proteins.* **53**:430-435 (2003).

Reczko, M.; Martin, A.C.; Bohr, H.; Suhai, S.; "Prediction of hypervariable CDR-H3 loop structures in antibodies.", *Protein Eng.* **8**:389-95 (1995).

Sali, A.; Blundell, T.L.; "Comparative protein modeling by satisfaction of spatial restraints."; *J. Mol. Biol.* **234**:779-815 (1993).

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

Sibanda, B.L.; Blundell, T.L.; Thornton, J.M.; "Conformation of β-hairpins in protein structures: A systematic classification with applications to modelling by homology, electron density fitting and protein engineering.", *J. Mol. Bio.* **206**:759-777 (1989).

Sternberg, M.J.; "Tertiary Structure Prediction" in *Protein Structure Prediction: A Practical Approach*; Sternberg, M.J., Ed.; Oxford Universtiy Press: New York, 1996; chapter 6.

Thompson, J.D., Higgins, D.G. and Gibson, T.J.; " CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice." *Nucleic Acids Res.* **22**:4673-80 (1994).

Vargas-Madrazo, E.; Paz-Garcia, E.; "An improved model of association for VH-VL immunoglobulin domains: asymmetries between VH and VL in the packing of some interface residues.", *J. Mol. Recognit.* **16**:113-120 (2003).

Wu, T. T.; Kabat, E. A.; "An Analysis of the Sequences of the Variable Regions of Bence Jones Proteins and Myeloma Light Chains and Their Implications For Antibody Complementarity.", *J. exp. Med.* **132**:211-250 (1970).

Appendix A

Heavy Chain Sequence Alignment

```
15C8_H          EVQLQQSGAELVKPGASVKLSCTASGFNIKDTYMHWVKQKPEQGLEWIAQID--PANGNT
1QBM_H          EVQLQQSGAELVKPGASVKLSCTASGFNIKDTYMHWVKQRPEKGLEWIGRID--PASGNT
1OAY_H          EVQLQQSGAELVKPGASVKLSCKASGYTFTSYWMHWVKQRPGRGLEWIGRID--PNGGGT
1A14_H          QVQLQQSGAELVKPGASVRMSCKASGYTFTNYNMYWVKQSPGQGLEWIGIFY--PGNGDT
1MRD_H          QVQLQQSGAELVKPGASVKLSCKASGYTFTSYWMQWVKQRPGQGLEWIGEID--PSDSYT
1Q9W_D          EVILVESGGGLVQPGGSLRLSCSTSGFTFTDYYMSWVRQPPGKALEWLGFIRNKPKGYTT
1PG7_H          EVQLVESGGGLVQPGGSLRLSCAASGFNIKEYYMHWVRQAPGKGLEWVGLID--PEQGNT
                :* * :**. **:**.*:::** :**:.:..  * **:* * :.***:. :   *    *

15C8_H          KYDPKFQGKATITADTSSNTAYLHLSSLTSEDSAVYYCAADPPYYGH---GDYWGQGTTL
1QBM_H          KYDPKFQDKATITADTSSNTAYLQLSSLTSEDTAVYYCAGYD--YGN---FDYWGQGTTL
1OAY_H          KYNLKFKSKATLTVDKPSSTAYMQLSSLTSEDSAVYYCARMWYYGTYY--FDYWGQGTTL
1A14_H          SYNQKFKDKATLTADKSSNTAYMQLSSLTSEDSAVYYCARSGGSYRYDGGFDYWGQGTTV
1MRD_H          NYNQKFKGKATLTVDTSSSTAYMQLSSLTSEDSAVYYCANLRG-Y-----FDYWGQGTTL
1Q9W_D          EYSASVKGRFTISRDNSQSILYLQMNTLRAEDSATYYCVRDIYSFGSRDGMDYWGQGTSV
1PG7_H          IYDPKFQDRATISADNSKNTAYLQMNSLRAEDTAVYYCARDTAAY-----FDYWGQGTLV
                 *. ..:.: *:: *....  *:::.:* :**:*.***.            ******* :

15C8_H          TVSSAKTTPPSVYPLAPGSAAQTNSMVTLGCLVKGYFPEPVTVTWNSGSLSSGVHTFPAV
1QBM_H          TVSSAETTPPSVYPLAPGTAALKSSMVTLGCLVKGYFPEPVTVTWNSGSLSSGVHTFPAV
1OAY_H          TVSSAA------------------------------------------------------
1A14_H          TV----------------------------------------------------------
1MRD_H          TVSSAKTTPPSVYPLAPGCGDTTGSSVTLGCLVKGYFPESVTVTWNSGSLSSSVHTFPAL
1Q9W_D          TVSSAKTTPPSVYPLAPGSAAQTNSMVTLGCLVKGYFPEPVTVTWNSGSLSSGVHTFPAV
1PG7_H          TVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAV
                **
15C8_H          LQS-DLYTLSSSVTVPSSTWPSETVTCNVAHPASSTKVDKKIV----
1QBM_H          LQS-DLYTLSSSVTVPSSTWPSQTVTCNVAHPASSTKVDKKIVPRNC
1OAY_H          -----------------------------------------------
1A14_H          -----------------------------------------------
1MRD_H          LQS-GLYTMSSSVTVPSSTWPSQTVTCSVAHPASSTTVDKKLEP---
1Q9W_D          LQS-DLYTLSSSVTVPSSTWPSETVTCNVAHPASSTKVDKKIVPRDC
1PG7_H          LQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEP---
```

Figure A1 – Heavy chain alignment used to parse CDR-H3 regions. Also used in the creation of an HMM used to identify heavy chains.

```
1CT8_A      ELVMTQTPATLSVTPGDSVSLSCRASQSVSN------KLHWYQQKSHESP
1BAF_L      QIVLTQSPAIMSASPGEKVTMTCSASSSVY-------YMYWYQQKPGSSP
1NCW_L      DVVMTQSPKTISVTIGQPASISCKSSQRLLNSNGKT-FLNWLLQRPGQSP
1T4K_A      DIQMTQSPSSLAVSPGEKVTMSCRSSQSLFNSRTRKNYLAWYQQKPGQSP
1AP2_A      DIVMTQSPSSLTVTAGEKVTMSCKSSQSLLNSGNQKNYLTWYQQKPGQPP
            :: :**:*  ::.: *: .:::* :*. :         : *  *:. ..*

1CT8_A      RLLIKFASQSIPGIPSRFSGSGSGSDFTLSINSVETEDFGIYFCHQTHGR
1BAF_L      RLLIYDTSNLASGVPVRFSGSGSGTSYSLTISRMEAEDAATYYCQQWSSY
1NCW_L      KRLIYLGTKLDSGVPDRFTGSGSGTDFTLKISRVEAEDLGVYYCWQGTHF
1T4K_A      TKLIYWASTRESGVPDRFTGSGSGTDFTLTISSVQAEDLAIYYCKQSYDL
1AP2_A      KLLIYWASTRESGVPDRFTGSGSGTDFTLTISSVQAEDLAVYYCQNDYSY
              **   :   .*:* **:*****:.::*.*. :::** . *:* :

1CT8_A      -PLTFGAGTKLELKRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKD
1BAF_L      PPITFGVGTKLELKRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKD
1NCW_L      -PYTFGGGTKLEIKRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKD
1T4K_A      --PTFGAGTKLELKRSDAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKD
1AP2_A      -PLTFGAGTKLEPG------------------------------------
               *** *****

1CT8_A      INVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYT
1BAF_L      INVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYT
1NCW_L      INVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYT
1T4K_A      INVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYT
1AP2_A      --------------------------------------------------


1CT8_A      CEATHKTSTSPIVKSFNRNEC
1BAF_L      CEATHKTSTSPIVKSFNRNEC
1NCW_L      CEATHKTSTSPIVKSFNRNEC
1T4K_A      CEATHKTSTSPIVKSFNRN--
1AP2_A      ---------------------
```

Figure A2 - Light chain alignment used to parse residues from light chain. Also used in the creation of the light chain HMM to identify light chains.