Oxford University

Computational Antibody Design

by

Konrad Krawczyk

in the

Department or Statistics

December 2013

Oxford University

Abstract

Department or Statistics

Doctor of Philosophy

by Konrad Krawczyk

Antibodies are a class of proteins vital in mediating immune responses in vertebrates. Their binding site is highly malleable, allowing them to bind virtually any antigen. The versatility of antibody binding sites has received much attention from the pharmaceutical industry, marking them out as the most important category of biopharmaceuticals. The development of antibodies which bind to a specific antigen has thus far been achieved by costly and time-consuming experimental screening campaigns. However, in recent years computational approaches to antibody design have started to emerge, which offer an alternative. Computational antibody design techniques focus on determination of the binding site on the antibody, antibody-modelling, antibody-antigen docking and prediction of the binding site on the antigen. Here, we explore aspects of computational antibody design with the aim of gaining a better understanding of antibody-antigen interactions and improving existing artificial antibody design tools. We start by demonstrating our structural antibody database which has become a primary resource for antibody structural information. This is followed by a detailed analysis of the antibody-antigen interactions. The information gathered from this analysis allowed us to create an antibody contact site prediction tool, Antibody i-Patch. This tool was then employed to develop a local antibody-antigen docking pipeline, which used knowledge of the binding site of the antigen. We then tackled the global antibody-antigen docking problem by developing EpiPred, antigen binding site predictor which was employed in our global antibody-antigen docking pipeline.

Acknowledgements

I would like to thank my family, Krawczyks and Schlackows, for all the support I have received from them. Special thanks go to Rita who supported me throughout this effort and provided me with a sense of purpose. This work would not be possible without my supervisors: Prof Deane, Dr Shi and Dr Baker. They have all invested a considerable amount of time in this endeavour and I hope that they are satisfied with the outcome. This D.Phil has been very far from the usual notion of a lonely, isolated scientist and it is chiefly due to the nature of the Oxford Protein Informatics Group. The group made me feel very welcome from the very beginning and it has been a pleasure being part of it.

Contents

List of Figures

List of Tables

For those who wholeheartedly supported this work but are not here to experience its coming to fruition.

Chapter 1

Introduction

1.1 Introduction

This introduction is aimed at giving the reader a sense of orientation in the broad scientific field of computational antibody design. In order to achieve this, we firstly give a brief review of the protein-structure concepts. This is followed by an introduction of the main ideas relating to antibodies and their structural aspects. After establishing the concept of antibodies in the context of structural biology, we give a review of current methods of computational antibody design. Finally, we present an outline of this thesis.

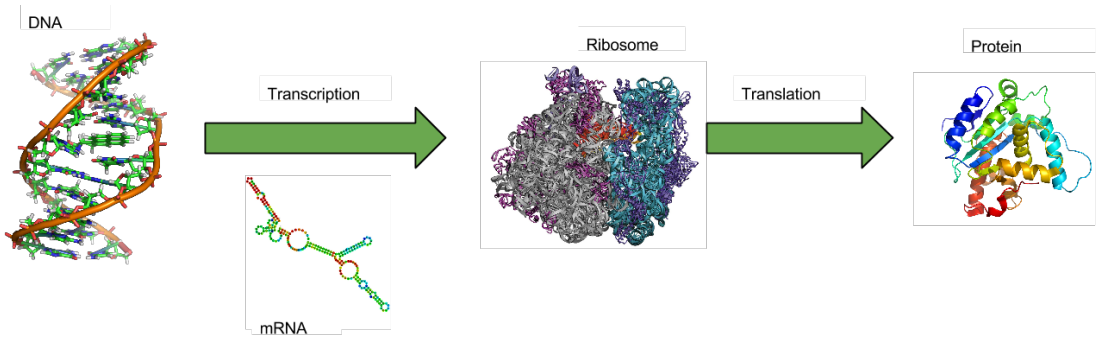


FIGURE 1.1: **Central Dogma of Biology.** The first step in creating a protein is the transcription of DNA into RNA. The RNA molecule which carries the protein blueprint is called the messenger RNA (mRNA). The mRNA is translated by ribosome into a linear chain of amino acids linked by peptide bonds. The linear chain of amino acids (polypeptide) assumes a specific three-dimensional structure that is intimately linked to its function (Reece et al. [2011]).

.

1.1.1 Structural Protein Biology

Proteins are one of the polymers underpinning the functioning of complex biological systems. They are a product of translation of messenger RNA (mRNA) by the ribosome into a linear polymer of amino acids (see Figure 1.1).

Protein structure can be divided into four levels of complexity (see Figure 1.2). The primary structure is the linear sequence of amino acids.

There are two secondary structure types: alpha helices and beta sheets. These are regular structures that come about as a result of hydrogen bonding between the backbone residues. Alpha helices are coiled structures which are formed by hydrogen bonds between amide hydrogens and carbonyl oxygens on the backbone of amino acids. In an alpha helix the hydrogen bond is formed between the n and the $n + 4$ residue in the sequence. There exist variations on the alpha helix where the hydrogen bond is formed between n and $n + 3$ residues (3_{10} helix) or the n and $n + 5$ residues (pi helix). Beta sheets are planar arrangements of amino acid strands 3-10 amino acids long which form hydrogen bonds with each other. There are two distinct arrangements of beta sheets: parallel and anti-parallel (see Figure 1.3). In an anti-parallel beta sheet, the two adjacent strands run in opposite directions (one towards N-terminus, the other to C-terminus). In a parallel beta sheet, the two adjecent strands run in the same direction. In rare cases one can observe mixed beta sheets with both parallel and antiparallel segments.

These secondary structure elements are linked by more irregular structures, termed loops or coils. Links between the secondary structures allow the polypeptide to assume a distinct three-dimensional structure termed the tertiary structure.

Proteins can exist as single polypeptide chains (monomers) or they can form multimeric structures composed of several polypeptides. Such arrangements of proteins are referred to as quaternary structures.

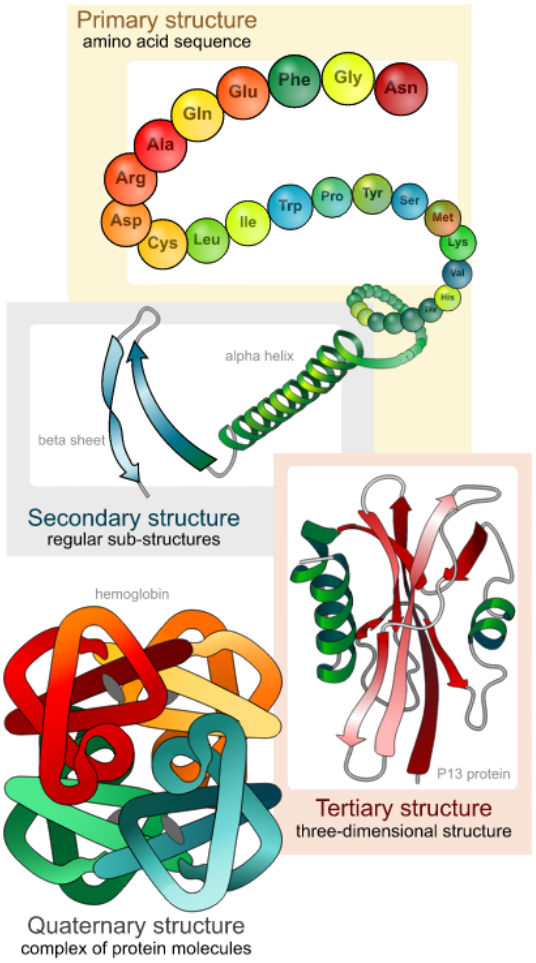


FIGURE 1.2: **Four levels of protein structure.** Protein structure is divided into four levels of complexity. The primary structure is the linear sequence of amino acids that form the polypeptide. Secondary structure is divided into alpha helices and beta sheets - regular structures formed through hydrogen bonding between backbone amide hydrogens and carbonyl oxygens. Tertiary structure is the specific three-dimensional arrangement of the secondary structure and the loops which link them. Quaternary structure is the arrangement of multiple polypeptide chains, reproduced from (Junqueira et al. [1998]).

1.2 Why antibodies?

Antibodies are the the key protein actors in the acquired immune responses in vertebrates. The most common human antibody isotype is the IgG, one of the main mediators of secondary immune responses (Kindt et al. [2007], Junqueira et al. [1998]). Antibodies have a conserved structure with more than 1700 solved structures available in the Protein Data Bank as of October 2013 (Abola et al. [1984], Dunbar et al. [2013b]). Most of the variability in antibodies (both sequence and structure) is found in its binding site

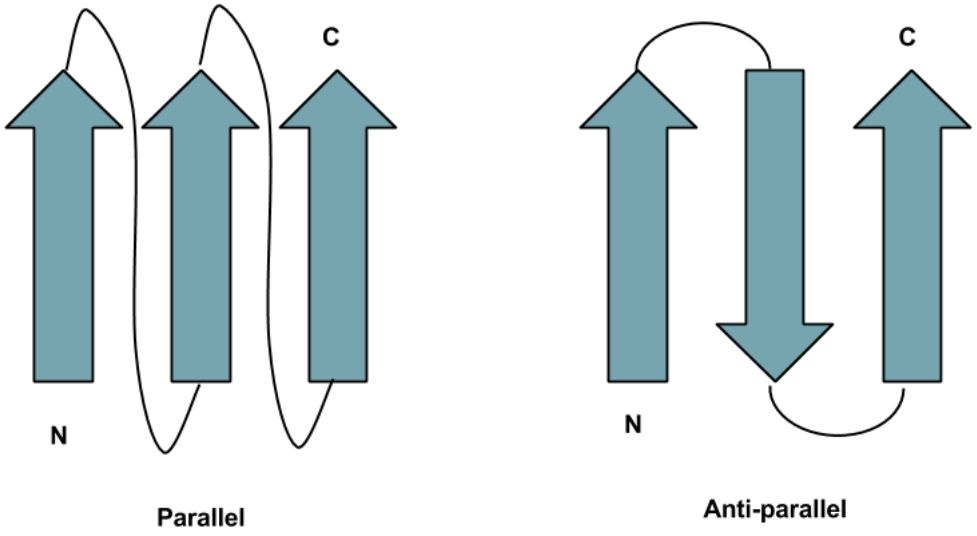


FIGURE 1.3: **Parallel and anti-parallel beta sheets.** If adjacent beta-strands have the same biochemical direction (N-terminus towards C-terminus) than the beta sheet is parallel. In the other case, the beta sheet is anti-parallel.

.

which is chiefly comprised of its Complementarity Determining Region loops (CDRs) (Raghunathan et al. [2012]). The affinity and specificity of the antibody's cognate antigen can be effectively modulated by only a few mutations to the CDRs (Raghunathan et al. [2012]). Due to their malleable binding properties, antibodies are currently one of the most important biopharmaceuticals (Murad et al. [2012a], Wark and Hudson [2003]). The number of antibody-based drugs has been rising steadily over the last twenty five years since the first such approved drug, a transplant rejection drug, muromab CD3 in 1986 (Murad et al. [2012a], Wark and Hudson [2003]). For instance, monoclonal antibody therapies have been developed against diseases like osteoporosis (Murad et al. [2012b]) and rheumatoid arthritis (Feldmann [2002]).

The majority of the technologies employed for artificial antibody design are based on costly experimental screening campaigns. There is however a growing number of computational methods aimed at aiding the process of artificial antibody design (Kuroda et al. [2012]). In this thesis, we will build on the current knowledge of the antibody-antigen interactions and computational antibody design techniques and describe the set of tools which we have developed to facilitate virtual screening. In the following sections, we will

outline the current state of understanding of the antibody-related binding mechanism. We will then review current methods for antibody design, with a focus on computational techniques

1.3 Introduction to Antibodies

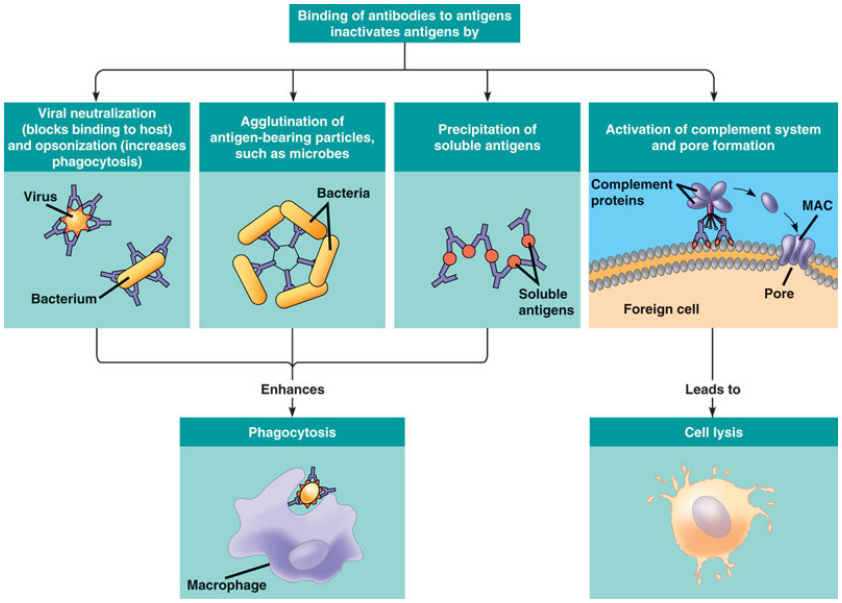


FIGURE 1.4: Different ways in which antibodies help in removing antigens from the organism. Reproduced from (Reece et al. [2011]).

The vertebrate immune system uses two sub-mechanisms: *innate* and *acquired immunity*. *Innate immunity* is the inborn mechanism of mounting a general but moderate response to a wide spectrum of antigens, while *acquired immunity* is a strong reaction against previously encountered antigens. Acquired immunity is not inborn and is developed by exposure to pathogens.

Acquired immunity can be further partitioned into cell-mediated and humoral immunity. Cell-mediated immunity does not use antibodies, but rather relies on activity of natural killer cells, macrophages and cytotoxic T cells. Humoral immunity on the other hand is mediated by antibodies and is crucial to an organism's capacity to maintain cell-memory of noxious antigens.

Humoral immunity is brought about by B-cells - antibody secreting cells. Each B-cell is capable of producing a different antibody. Those B-cells whose antibodies successfully bind to an antigen are signalled to proliferate. The signal to such activation is a function of the specificity and affinity of that B-cells antibody towards an antigen and thus it provides the basis for the selection of those cells whose antibodies have more favourable qualities.

When a B-cell is activated it divides into two groups of cells: plasma and memory cells. Plasma cells secrete antibodies against a particular antigen that prompted the activation. Antibodies then bind to the antigen in question at which point other cells and proteins help in neutralizing the threat (see Figure (1.4) for an overview of antibody effector functions). Memory cells on the other hand divide and remain in the organism for long enough to be able to mount another immune response against the antigen if it ever infects the organism again.

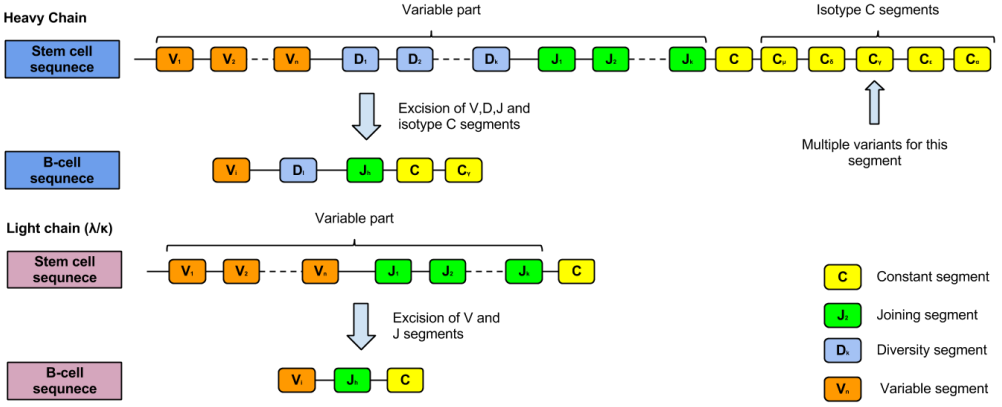


FIGURE 1.5: Antibodies are composed of four chains: two heavy and two light chains. The variable portion of the sequence of each chain, as present in the genome of a B-cell is composed of V, D and J segments (light chains do not have the D segment). The stem cell which gives rise to the B-cell, has multiple copies of each of V, D, J and C segments. Only one copy is passed to a B-cell, meaning that the remaining segments are excised from the original stem-cell genome. The numbers of each of the V, D and J segments vary in the literature, but in humans there are approximately 50-100 V, 10-20 D and 5-10 J segments (Matsuda et al. [1998], Li et al. [2004]).

1.4 Antibody genetics and structure

There are five antibody isotypes: IgA, IgD, IgE, IgM and IgG. Three of them (IgG, IgD and IgE) function as the Y-shaped *immunoglobulin monomer* while IgA and IgM are their polymers (dimer and pentamer respectively). The *immunoglobulin monomer* is composed of two heavy (450-550 amino acids) and two light (211-217 AAs) chains. On each chain there are portions which are denoted the *constant* (F_c) and the *variable* (F_v) region. The constant region acts as the binding site for many effector cells of the immune system whereas the variable region houses the antigen binding site. The variability of the antibody repertoire comes about due to the transcription mechanism by which an antibody sequence is created.

The genetic sequence of an antibody chain can be divided into the variable and constant parts (Figure 1.5). On the heavy chain one of five segments, C_μ, C_γ, C_η, C_α or C_δ (with several possible variants for C_γ), is used for the constant part. Each of the segments corresponds to one of the five antibody isotypes described above. The sequence of the variable portion is composed of multiple segments, three for the heavy chain and two for the light chain. The three segments constituting the variable part of the heavy chain are termed V, D and J. The corresponding variable sequence for the light chain does not have the D segment. Each antibody-producing B-cell has a particular combination of the V(D)J segments determined at the time of its maturation. Originally, the stem-cell sequences for heavy and light chains have multiple copies of V, D and J sequences. Thus, when a stem-cell matures into a B-cell only one of the V, D, and J segments is selected for a B-cell genome, excising the rest.

The heavy and light chain sequences are assembled in their final form using two genetic mechanisms: immunoglobulin gene rearrangement (IGR) of the V(J)D segments (Sakano et al. [1980]) and somatic hypermutation (SH) (Besmer et al. [2004], Neuberger [2008]). During the development of the antibody producing B-Cell, IGR is responsible for assembling combinations of available gene segments (V-D-J for heavy chain and V-J for the light chain) to form particular light and heavy chain templates. SH introduces

point mutations to the hypervariable region of the antibodies, the CDRs. These genetic events result in changes in the make-up of the binding site by varying the length and composition of the CDRs.

As described above, an antibody consists of two heavy and two light chains. The variable domain (F_V) which contains the antigen binding region is situated at the N terminus of each chain. An antibody contains two F_V domains as shown in Figure 1.6. This domain determines the antibody's binding specificity - hence the neighbouring *variable* domains together with first light and heavy constant domains are known collectively as the F_{ab} region, or the *antigen binding region*. The binding site of the antibody is the amalgam of six hypervariable loops, the Complementarity Determining Regions (CDRs). Three of these loops are found on each of the heavy and light chains (Wu and Kabat [1972]). The three hypervariable loops on the light chain are called L1, L2 and L3, while those on the heavy chain are called H1, H2 and H3. A simple nomenclature is used to describe CDRs specifying its type and length, e.g. H1-8, meaning H1 loops of length 8.

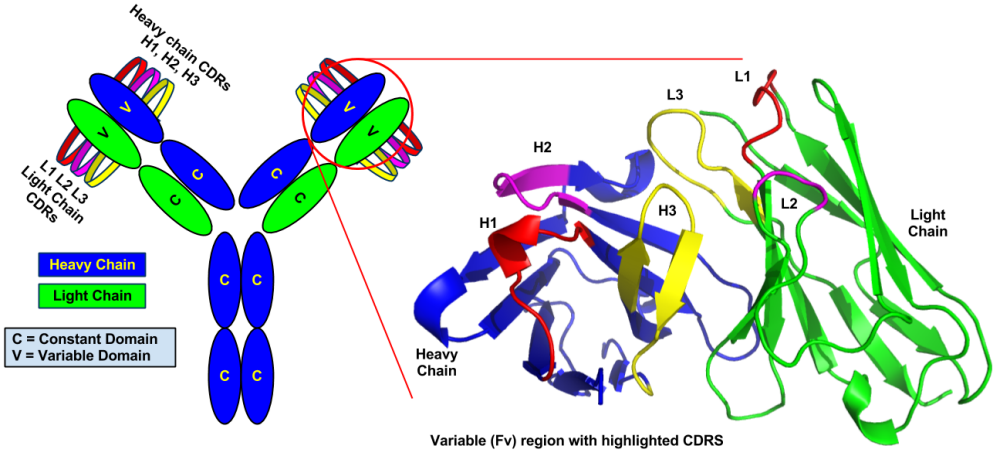


FIGURE 1.6: Schematic of an antibody molecule. There are two symmetric antibody binding sites (paratopes), each holding six CDRs denoted H1, H2, H3 (heavy chain) and L1, L2, L3 (light chain). The constant and variable domains are shown using ovals annotated C and V respectively. In magnification, the antibody variable region is shown (F_v) with the CDR loops highlighted using colours. (PDB 1A2Y).

Since the subject of this thesis is computational antibody design, below we give a more detailed account of the current understanding of the antibody binding site. In particular

we discuss antibody CDRs and previous findings concerning their ability to attain high affinity and specificity to such a broad range of antigens.

1.4.1 Compositional and structural features of antibody CDRs

Here we discuss two areas of antibody CDR research: amino acid composition and loop structure. In the former we focus on the compositional differences between species and on the residues that appear to bear more importance for antibody binding process. For the latter, structural part, we discuss the apparent conformational uniformity of the CDRs and the implications of the outlier to this trend - CDR-H3.

1.4.1.1 Composition trends in CDRs

Hypervariable loop mutation trends. Since antibodies undergo accelerated and ad hoc evolutionary process that other proteins do not, the trends in mutational patterns between antibodies and other proteins have been previously investigated (Clark et al. [2006]). The substitution patterns in Somatic Hypermutation were studied by comparing the probabilities of mutations from 23116 heavy chains and 11095 light chains with the conventional BLOSUM matrix (Henikoff and Henikoff [1992]). The trends of mutations introduced in the process of Somatic Hypermutation are not radically different from those of the evolutionary process for general proteins.

CDR composition across species. Two further aspects of the hypervariable loops that have been addressed previously are the compositional difference between CDRs and other soluble loops, as well as CDR composition difference across species. The amino acid composition of CDRs was found to be statistically significantly different from that in the loops of other proteins (Collis et al. [2003]). Human and mouse H3 CDRs were also shown to be of statistically different compositions, yet they still retain

a common feature of over-expressed serine and tyrosine (Zemlin et al. [2003]). The over-representation of serine and tyrosine, despite statistically different compositions between the two organisms, suggests that these residues may play a significant role in antibody binding.

CDR composition over time. Given this overall consistency of over-expression of serine and tyrosine, it has been tested whether it persists over time. Organisms are believed to be born with a largely predetermined initial antibody repertoire (Schroeder Jr [2006], Schroeder Jr et al. [1998]). This commonality across organisms tends to disappear as they mature, although not completely, as a study of the zebrafish repertoire revealed that certain antibodies might be produced by two distinct adult organisms (Weinstein et al. [2009], Jiang et al. [2011]). For instance, in a study of Mexican axolotl antibodies the authors noted that tyrosine and serine maintain high concentrations throughout the animals life (Golub et al. [1997]). The over-expression of certain residues as an organism matures supports the idea that intrinsic antibody binding mechanisms might be conserved.

Role of serine and tyrosine in antibody binding. Given that serine and tyrosine appear to be over-expressed in humans, mice and zebrafish and that this phenomenon appears to persists over an organism's life, the role of those residues has been investigated in antibody binding. Significance of these residues was confirmed experimentally, when antibody binders were designed using only a binary code of tyrosine and serine (Fellouse et al. [2005]). The study concluded that using tyrosine and serine alone in the binding site was sufficient to create an antibody that bound an antigen. One suggestion arising from this work was that tyrosine, being large, cyclic and offering a protected hydrogen bond mediates contacts whereas serine contributes to loop flexibility.

It was further argued both theoretically and experimentally that tyrosine, and additionally tryptophan, are good candidates for participation in antibody binding sites (Birtalan et al. [2008], Mian et al. [1991], Fellouse et al. [2005]). Both residues are large and

cyclic and can engage in a variety of interactions (Mian et al. [1991]). It was suggested, however, that tyrosine and tryptophan might provide only the initial low-affinity, low-specificity binding that is then tuned by somatic hypermutation through introduction of other residues with more favourable properties for a given binding site (Mian et al. [1991]). For instance, a previous study into trends in Somatic Hypermutation revealed the mutation patterns during affinity maturation (Clark et al. [2006]). Most notably, the frequencies of tyrosine, serine and tryptophan decrease as the antibody becomes more fine-tuned towards its antigen while there is a noted increase in the frequencies of histidine, proline and phenylalanine (Clark et al. [2006]).

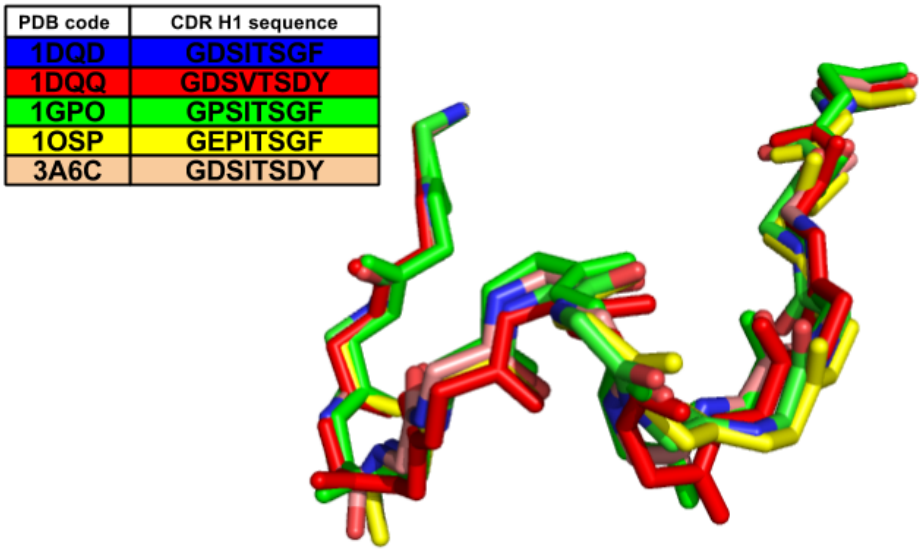


FIGURE 1.7: Five superimposed CDR-H1 structures of length 8. All structures presented above are no more than RMSD 1Å from each other, despite differences in sequence composition. Structures are presented with their anchor regions, comprising three residues on each side; these residues are not given in the sequence table above.

1.4.1.2 Antibody binding site structure

Antibodies have a very well conserved structure with the majority of the variability being concentrated in the antibody binding site composed of the CDRs (Raghunathan et al. [2012], Wu and Kabat [1972], Clark et al. [2006]). The main paradigm relating to the structural features of the antibody binding site are the canonical classes - structurally conserved conformations of CDRs.

CDR canonical classes Most CDRs appear to adopt only a limited number of shapes, termed canonical conformations (for an example see Figure 1.7). The conformational similarity appears to be brought about by key residues in certain positions (particularly certain prolines and glycines), providing an important link between the sequence and structure. The canonical groupings have been studied extensively since the 1980s (Chothia and Lesk [1987], Al-Lazikani et al. [1997], Lara-Ochoa et al. [1996], Chothia et al. [1989], North et al. [2011], Martin and Thornton [1996]). Early work on the subject identified a relatively small (with respect to all possible conformations) geometric space to which CDRs are restricted.

The most recent published clustering concluded that the original canonical classes still existed but the increase in data necessitated the creation of several new classes (North et al. [2011]). The set of conformations grew from 25 canonical classes drawn from 17 antibody structures by Martin and Thornton in 1997 to 72 classes from 337 heavy and 311 light chains used by North et al. in 2010. This suggests that structural patterns for CDRs exist but that they might not yet give the complete picture of available conformations.

The aforementioned studies focused chiefly on establishing the link between CDR sequence and structure. Another aspect of the hypervariable loops which also appears to be intimately linked to the canonical classes, and thus structure, is CDR length (number of residues). This property is known to affect the type of antigen that can be bound, with different length CDRs being responsible for binding lipids or virus molecules (Collis et al. [2003]).

It has previously been investigated whether there are correlations between CDR canonical classes drawn from any antibody binding site and the geometry of the antibody interface. It was shown that antibodies appear to use only a small fraction of the possible canonical class combinations, restricting the conformation space of the binding site (Lara-Ochoa et al. [1996]). Since geometry of the combining site has been shown to correlate with different antigens (MacCallum et al. [1996]), antibodies might use such length/structure constraints as specificity determinants.

CDR-H3 The only antibody CDR that does not appear to abide by obvious canonical rules is H3, whose conformations bear little similarity with each other. One of the earliest observations on the nature of H3 was made by Morea et al. who divided the loop into the *head* and *torso* region (Morea et al. [1997, 1998]). The torso appears to take one of the two conformations: extended or kinked (bulged) β sheet. The conformation of the head part is then constrained by the torso region. However, CDR-H3 does not appear to have a limited set of conformations such as those found for other CDRs (Morea et al. [1998], Shirai et al. [1996], Oliva et al. [1998]).

Several studies have shown that in some cases (H3 in particular) the hypervariable loops can change their conformation in order to accommodate the antigen (James et al. [2003], Stanfield et al. [2007], Rini et al. [1992]). Such induced fit requiring flexibility might be the reason that H3 cannot be structurally classified like other CDRs. It has been suggested that H3 alone is sufficient for antigen binding (Xu and Davis [2000]), making it the most important of hypervariable loops. It must be noted though that the antibody binding site appears to be more rigid in its mature form as opposed to the germline variety (Wedemayer et al. [1997], Patten et al. [1996]). Furthermore, it has been argued that rigidification of the antibody occurs across the entire variable domain not just the CDRs (Zimmermann et al. [2010], Thielges et al. [2008], Jimenez et al. [2003]).

Antibody binding site geometry The general geometry of the antibody binding site does not appear to show the same degree of shape complementarity as in enzyme complexes (Lawrence and Colman [1993]). Some trends in the topography of the binding site have been identified as they appear to correlate with the antigen being bound (MacCallum et al. [1996], Lee et al. [2006]). For instance, antibodies which target proteins have relatively flat binding sites whereas the anti-peptide ones tend to have a groove which can house the peptide. The contrast between antibodies and enzymes can be explained by the different evolutionary pressures these proteins are subject to. Whereas enzymes are allowed to undergo correlated mutations with respect to their binding partner across several generations, antibodies must adapt to their antigen within

hours. The antibody needs to adapt its shape and the residue make-up of its binding site to best match an arbitrary antigen, while the antigen remains unchanged.

In further contrast to enzymes, where the polar and hydrophobic interactions play an important role in complex formation (Vajda [2005]), antibody binding sites are less hydrophobic (Conte et al. [1999]). Nevertheless, as mentioned earlier some hydrophobic residues notably tyrosine, and to a lesser extent tryptophan, are present in the antibody binding site. On the other hand, the electrostatic complementarity of the antibody binding site is similar to that of other proteins (McCoy et al. [1997]). Furthermore, the greater contribution of electrostatic interactions appears to be correlated with higher specificity whereas hydrophobic interactions account for cross-reactivity to a higher degree (Sinha et al. [2002], Mohan et al. [2003]). For instance, in studies of the antibody 48G7, it was reported that the maturation process was driven by the optimization of the electrostatic profile of the antibody (Patten et al. [1996], Wedemayer et al. [1997]).

CDR definitions and numbering Owing to the conservation of the antibody structure, methods have been developed that standardize the way these molecules are presented - sequence numbering and CDR definitions (see Figure 1.8). The former is given only a brief overview below, followed by a more detailed account of the latter since it is of more relevance to this thesis.

Numbering of the residues in the antibody sequence is aimed at standardizing the representation of each antibody sequence. If one referred to a particular residue, say the 25th residue of the light chain, it would immediately be possible to tell the approximate position of the amino acid with respect to the structural features of an antibody. The schemes introduced by Kabat and Chothia, each assign an ordinal number to every framework and CDR residue. In both schemes one refers to a particular residue specifying its chain and ordinal number - for instance H53 stands for the 53rd residue of the heavy chain. Since the framework regions are constant, this annotation was constructed by aligning multiple antibody sequences and assigning the aligned columns a single ordinal number. CDRs pose more of a problem since one has to account for the insertions

into the hypervariable loops (as there exists a length variability between the loops of the same type). Thus there exist residues termed H72A or H72B in the CDR regions in order to account for this phenomenon. Abhinandan and Martin proposed a new scheme, Abnum, which used a greater number of solved antibody structures to improve on both previous schemes by taking more CDR structural features into account (Abhinandan K R [2008]).

The locations of the CDRs can be approximately identified in an antibody sequence using any one of several sets of rules: Kabat (Wu and Kabat [1972]), Chothia (Chothia and Lesk [1987], Chothia et al. [1989]), AbM (Abhinandan K R [2008]), Contact (MacCallum et al. [1996]) or IMGT (Lefranc [2011]). In general these rules are based on identification of key residues after which CDRs are said to begin. For instance cystines in certain positions act as such boundaries. The differences between the schemes are depicted in Figure 1.8. Even though the current standard definition according to the World Health Organisation is that proposed by IMGT, the other definitions are still widey used by both the scientific community and the pharmaceutical industry.

1.5 Artificial antibody design methodology

Artificial antibody design to date has focused chiefly on the development of experimental techniques for synthesizing specific and high-affinity antigen binders. These methods have their limitations though, and it is becoming increasingly clear that computational antibody design techniques could facilitate this process. Computational antibody design methodology thus far have focused on artificial affinity maturation, defining the antibody binding site, antibody modelling, B-cell epitope prediction and antibody-antigen docking. An overview of each of these research areas is given below.

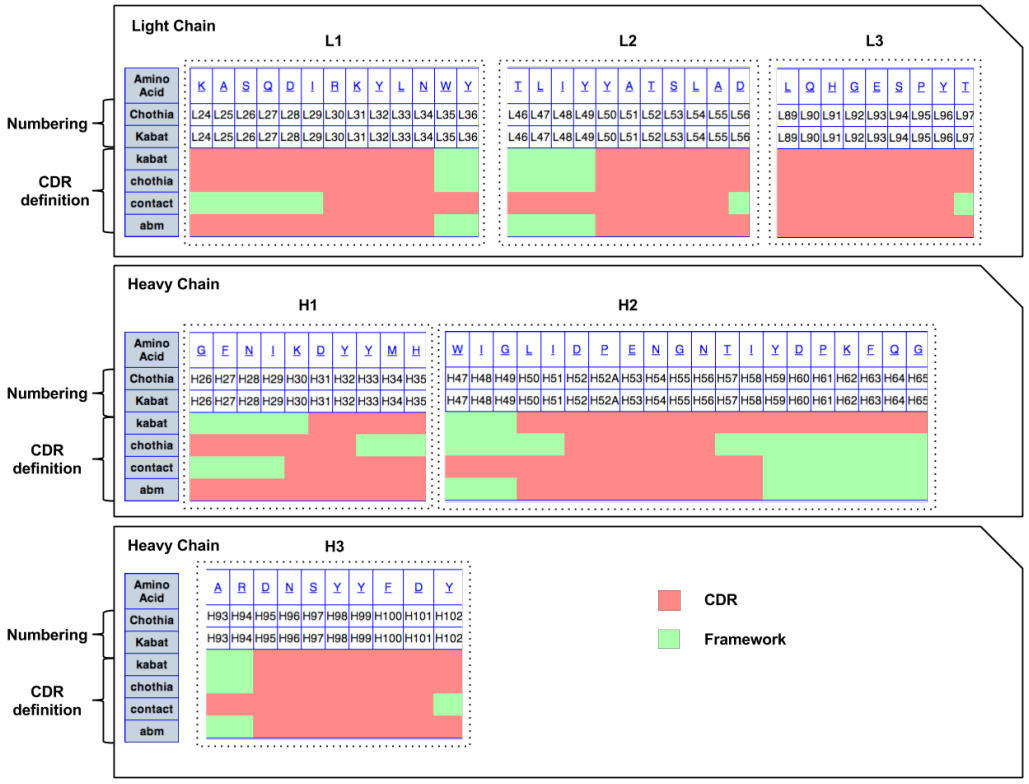


FIGURE 1.8: Visualization of differences between different numbering and CDR definition schemes in antibodies. The Figure was constructed using the results of a query for 1AHW in Abysis (Martin [2010]).

1.5.1 Conventional methods to design antibodies

The goal of antibody design (Carter [2006]) is to produce antibodies which would bind specifically to a certain antigen with a very high affinity (i.e. artificial affinity maturation). The two most widely used methods to achieve this goal are the humanization-based techniques and phage-display, summarized in Figure 1.9. The former relies on raising antibodies in an animal, say mouse, and then engineering those molecules so that they do not elicit an immune response in the host species (human) but still bind to their respective antigen. The latter method is a two-tier process where the two steps are antibody library construction and an iterative process of antibody mutation and good binder selection. The two technologies are described briefly below.

Humanization technology. Humanization technology is a set of techniques to engineer antibodies from different organisms so that they do not elicit immune responses in

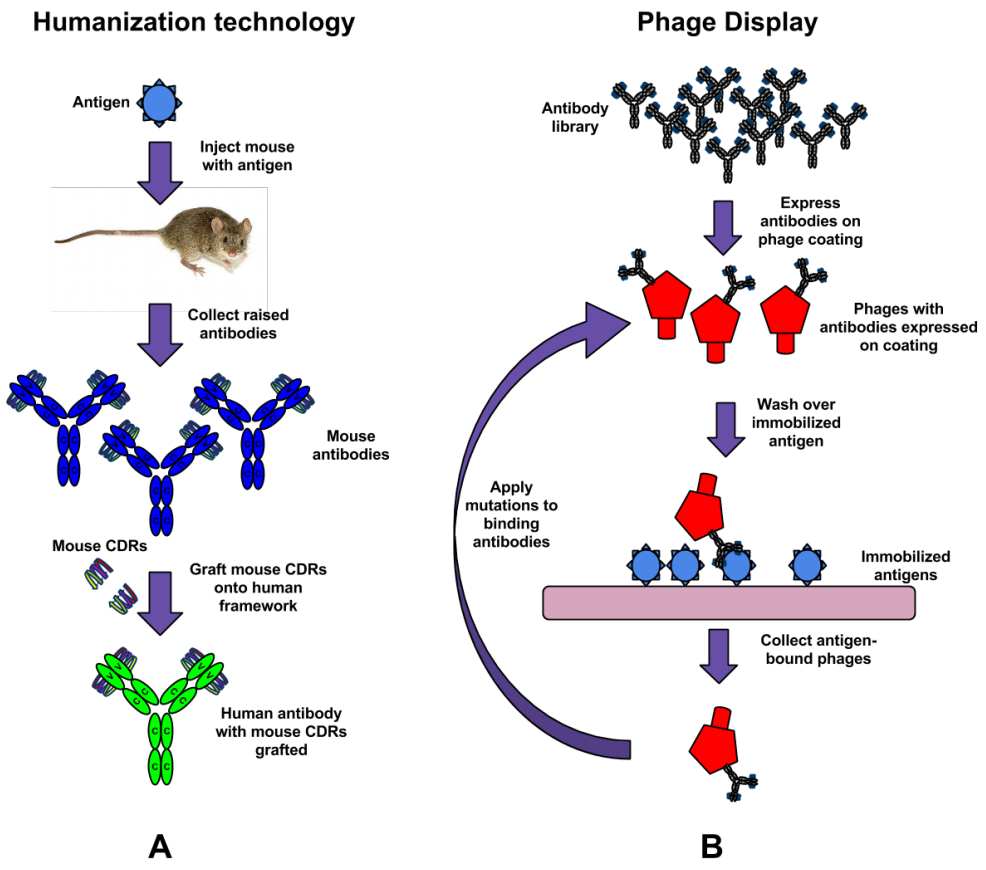


FIGURE 1.9: **A** Humanization technology. A mouse is injected with an antigen, prompting it to raise antibodies against it. The antibody producing B-cells are collected for in-vitro processing, where mouse CDRs are grafted onto a human framework. **B** Phage display. An appropriate antibody library is selected. The antibodies are then expressed on the phage coating, and panned against immobilized antigens. Those antibodies that did not bind are washed off. The remaining, binding, antibodies are mutated and re-expressed on phage coating, starting another round of the process.

humans (Köhler and Milstein [1975], Larrick and Fry [1991]). The animal, usually mouse, is injected with an antigen forcing its immune system to raise antibodies against it. The antibody producing B-cells are then harvested from mice and fused with a myeloma (B-cell cancer). The monoclonal antibodies which are produced by such hybridoma cells are all of single specificity.

This approach had to be streamlined as mouse antibodies are short lived in human serum (Khazaeli et al. [1994]). This was partially overcome by creation of hybrid antibodies with human constant regions and murine variable regions (Morrison et al. [1984], Boulianne et al. [1984]). Unfortunately, even such chimeric antibodies can still trigger immune responses (Bell and Kamm [2000]). Finally it was discovered that it is possible

to graft CDRs alone, if certain key residues are retained (Jones et al. [1986]) leaving the rest of the antibody intact. This procedure was made possible in part by the discovery of CDR canonical structures (Riechmann et al. [1988]). Such minimal intervention into the structure solved many, but not all, of the earlier problems associated with the method.

Phage display. Phage display is one of the most widely used techniques for screening antibodies for specificity (Wark and Hudson [2003], Smith [1985]). It exploits the fact that some phages like M13 can express certain protein fragments on their protein coating. By packing the plasmid encoding the F_v region in M13s capsid, the protein fragment is translated and presented on the surface. Next, phages are screened against immobilised targets that the antibody should bind to. All the non-binding phages are washed off while those that remain (i.e. they bind) are used to infect E.Coli so that the antibody-encoding fragment can be expressed and magnified through Polymerase Chain Reaction (PCR), producing input for a further iteration of the process.

Phage display technology relies on good initial collections of antibodies to be screened for specificity, termed antibody libraries. Antibody libraries are divided into two types: naive and immune libraries. Naive libraries are collections of antibodies obtained from an organism that did not have contact with the target antigen, i.e. it was never immunized. Such sets of antibodies provide a lot of initial variety but are not very specific. The second type of immune libraries consist of antibodies harvested after an organism had been immunized. These antibodies are more specific against the target antigen since the organisms immune system was allowed to perform an initial round of affinity maturation.

1.5.2 Computational antibody affinity maturation

Antibody in silico maturation is a sub-problem of the general field of protein design. Its aim is to re-engineer the antibody binding site (in most cases only CDRs) so as to modulate its specificity and affinity. Ideally, given the sequence of the antigen, a computational affinity-maturation method would produce a corresponding antibody that binds

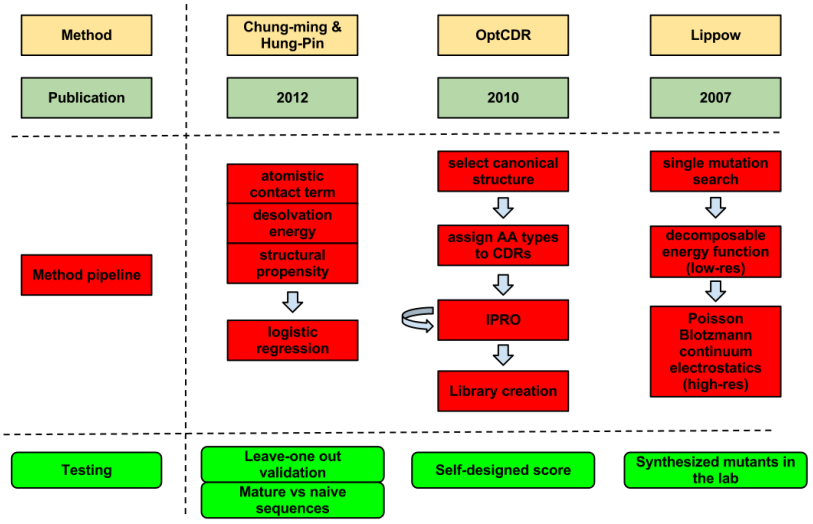


FIGURE 1.10: There currently exist three automated methods to design antibody binding sites: regression model proposed by Chung-Ming and Hung-Pin (Yu et al. [2012]), OptCDR by Pantazes et al. (Pantazes and Maranas [2010]) and the CHARMM-based mutagenesis designed by Lippow (Lippow et al. [2007]). The pipelines of the methods are presented as the computational milestones each algorithm has to pass.

the given antigen. However, as of yet no method exists that works when given sequence information alone, instead all current protocols require a structure of the antigen and most require a base structure of the antibody or even a complex structure. As stated previously, there are currently three published methods that automatically perform in silico affinity maturation (Lippow et al. [2007], Pantazes and Maranas [2010], Yu et al. [2012]) as opposed to others which require human intervention (Barderas et al. [2008], Kirkham et al. [1999]). The three methods are summarised in Figure 1.10, with a more detailed overview given below. These methods act primarily as proofs of concept, with limited availability.

The most recent algorithm, proposed by Chung-ming and Hung-Pin et al. 2012, used antibodies binding to vascular endothelial growth factor (VEGF) as the model system (Yu et al. [2012]). They describe a machine learning method which aims to inform the mutation choices during the artificial affinity maturation process. They have designed a logistic regression model to correlate a set of antibody attributes with the log-odd preference of a particular amino acid at a specified position in a given CDR. Phage display distribution of amino acids in CDRs served as background for this analysis.

In this paper the authors chose three attributes on which to base their predictions: an atomistic contact term (the probability of adopting a given rotameric conformation), maximal desolvation energy upon contact formation and a structural propensity term for a given amino acid at a specified position. Performing leave-one-out cross validation on their dataset resulted in mutations that would be favourable for creating a higher-affinity antibody. They have further tested their method by ranking the amino acids of two optimized anti-VEGF complexes and three other anti-VEGF structures from the phage library with only a limited variation of amino acids. Their ranking appeared to favour the affinity-matured antibodies amino acid choices, meaning that the algorithm correctly chooses residues that lead to higher-affinity binders. Of course, under this testing procedure, all other potential favourable mutations are false negatives.

OptCDR, by Pantazes and Maranas, is another method to fully redesign the antibody binding site to better suit its partner antigen (Pantazes and Maranas [2010]). They use a library of CDR canonical structure backbones to design an antibody binding site of optimal shape. Next, the backbone coordinates are given amino acid types, together with rotamer optimization of their side-chains. Finally, the model is optimized through rigid body docking and several rounds of structural optimization through backbone perturbation by the program IPRO (Saraf et al. [2006]). This method was tested using computational metrics alone indicating whether the resulting complex is more energetically favourable, thus the actual experimental viability of the method remains to be demonstrated.

The third of the methods, Lippow et al 2007, is so far the strongest proof of principle in computational design of antibody affinity. The strength of this analysis lies mostly in the fact that the mutants selected by the algorithm were synthesized and shown to be of improved binding affinity to that of the original structure. Lippow's algorithm simulates affinity maturation by an exhaustive search of the mutation space of CDRs, divided into two steps.

In the first round of calculations, each single mutant is considered, using a fast computational method. This allows the algorithm to consider all the possible combinations of

amino acids at a single position and to single out those most promising for more in-depth but costly computations. A pairwise decomposable energy function which allows for application of A* and dead-end elimination is used. The energy function is the CHARMM PARAM22 (MacKerell et al. [1998]) all-atom parameter set with all the energy terms. The objective function is the difference between the bound and unbound state energy. The lowest energy of each mutant is evaluated and it is admitted to the candidate list only if it has lower energy than the wild-type.

In the second round, the energies for the set of residues selected in step one are recomputed using a more costly Poisson-Boltzmann continuum electrostatics (using DELPHI (Nicholls and Honig [1991])), continuum solvent van der Waals and side chain conformation search. Penalties are applied every time flexible conformations would negatively affect binding. The sequences are ranked according to this scheme and the top-scoring mutants are given as potential antibodies to be synthesized in the lab.

Affinity of several existing antibodies was enhanced using the method presented above. Analysis of the structure of antibody D1.3, targeting the hen egg-white lysozyme produced 17 single mutations which were selected for experiments. Three of the mutations improved affinity with respect to wild type with the best one achieving 2.4 times higher affinity. In antibody D44.1, nine best scoring single mutations were selected for synthesis with 6 of them improving the overall affinity in subsequent experiments. The best single mutation achieved 8-fold improvement in binding affinity. The favourable single mutations were combined to produce a set of double and triple mutants. After synthesizing selected double mutants, the best one achieved 140-fold affinity improvement. Another multiple mutant was found for the antibody-drug Cetuximab where a triple mutant achieved a 10-fold affinity improvement.

These results show that antibody affinity maturation can be at the very least assisted computationally. Its biggest advantage is the fact that multiple affinity-improving mutations can be detected in a tractable way - something that is challenging experimentally due to the immense search space. Its only drawback is that it needs a solved antibody-antigen complex. It cannot produce an antibody on the basis of a sequence alone.

Nevertheless, combining it with a method that would create an ab initio set of decoys (artificial antibody library) could yield a very powerful antibody design method.

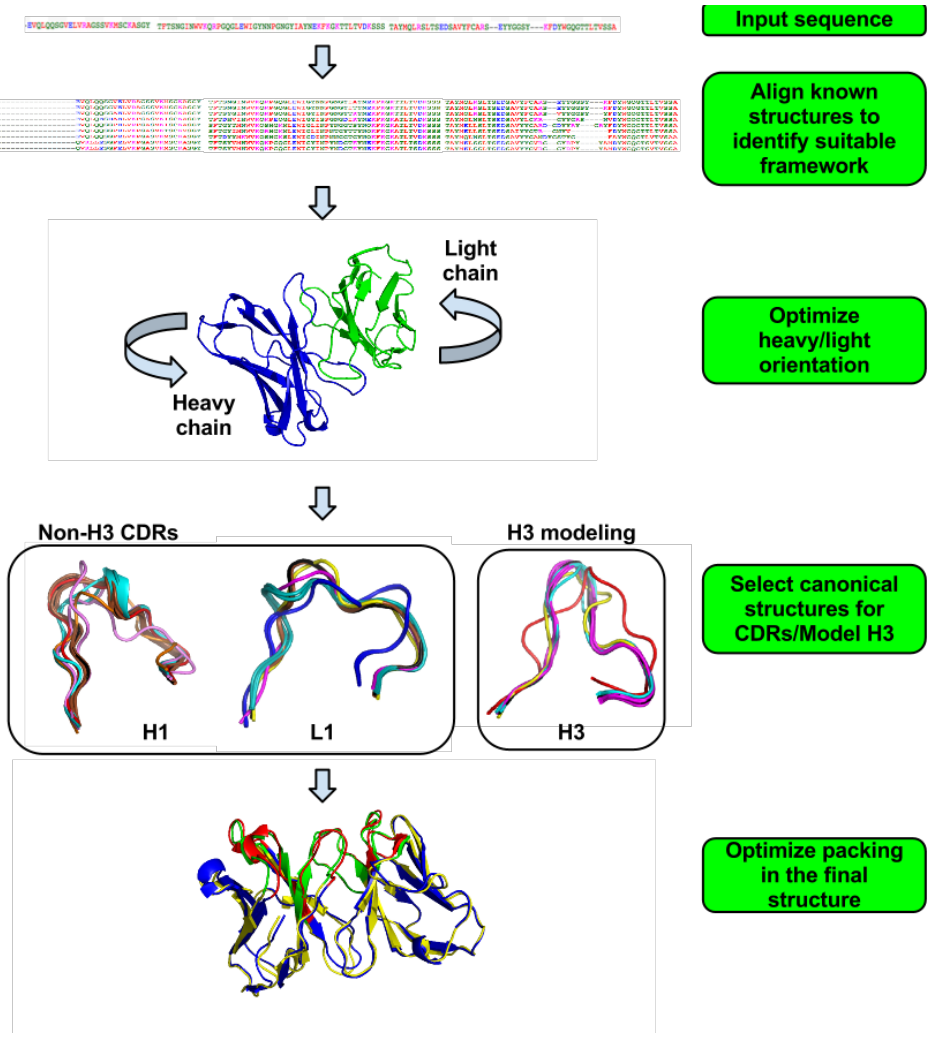


FIGURE 1.11: Typical pipeline for antibody modelling. The input sequence is aligned to antibodies with known structures. The best-matching framework is selected and the heavy and light chain orientations for the input sequence are determined. Finally, the CDRs are modelled, using canonical classes for non H3 loops and database or ab initio techniques for H3 itself. The final step involves optimization of the packing of the antibody as a whole and the binding site in particular.

1.5.3 Antibody modelling

Protein modelling is one of the most challenging problems in biology (Dill et al. [2008]). Given a sequence of a protein one would like to determine its structure (Moult [2005]).

There exist two main approaches to protein modelling: de novo and knowledge-based methods. Successful de novo techniques, like Rosetta, use protein fragment libraries to assemble parts of the query sequence, followed by successive rounds of energy-function optimization (Bonneau et al. [2001]). Knowledge based modelling, is chiefly represented by homology based approaches which lead to far better models than their de novo counterparts (Hildebrand et al. [2009], Fiser and Šali [2003], Schwede et al. [2003]). Here one aligns the query sequence to a set of proteins of known structure. Given that structures sharing more than 30% sequence similarity tend to have similar folds, one would use such structures as modelling templates for the query. The coordinates of the residues aligned to the template are copied over. The residues whose coordinates could not be determined, chiefly loops, are modelled ab initio (e.g. RAPPER (de Bakker et al. [2003])) or using similar homology approaches specific for protein loops such as FREAD (Choi and Deane [2010]).

Since antibodies have a well conserved fold, homology modelling can produce high-quality models as demonstrated by WAM (Whitelegg and Rees [2000]), PIGS (Marcatili et al. [2008]) and RosettaAntibody (Sivasubramanian et al. [2009]). The pipeline of these methods is fairly similar in that they align the query sequence to a library of known structures in order to identify the most suitable framework, followed by antibody-specific feature modelling (the process is summarised in Figure 1.11). The two biggest challenges in antibody modelling are the prediction of the structure of H3 and the V_h/V_l orientation. Current approaches to tackle those problems are outlined below.

As mentioned earlier, H3 is believed to be the most important of all the CDRs and also the most flexible. It is the longest loop with median a length of 12 residues (Zemlin et al. [2003]). One attempt to rationalize the observed set of conformations of H3 was by developing rules similar, if more complex, to those used for other CDRs. It was observed that the C-terminal regions of H3 tend to adopt either kinked or extended forms. This led to creation of the first set of rules for H3 (Morea et al. [1997, 1998], Shirai et al. [1996], Oliva et al. [1998]). The most recent attempt at defining the conformations of H3 has been carried out in 2010 (North et al. [2011]). However, as mentioned in the

earlier sections about canonical classes, these rules are still lacking the predictive ability of the other CDRs.

Another approach to model the structure of H3 is by database search of the loop structures available in the PDB (Fernandez-Fuentes et al. [2006], Choi and Deane [2010]). The most recent method, proposed by Choi and Deane, achieves higher accuracy than current ab initio methods (2Å vs 10Å for loops of length 20 or more). However, it only makes predictions for a small number of H3 CDRs.

The second issue in antibody modelling is that of the V_h/V_l orientation. The orientation of the two chains will significantly affect the geometry of the binding site. It has been shown that the orientation of the heavy and light chains may change upon binding (Stanfield et al. [1993]). One of the earliest works to address this problem was by Chothia et al. (Chothia et al. [1985]) where key residues for chain orientation were identified. Similar studies were conducted as more antibody structures became available (Abhinandan and Martin [2010], Chailyan et al. [2011]). These studies identified a fairly diverse set of residues that are considered important for the framework orientation. The most recent work on this topic, Dunbar et al. 2013, introduced five angles and a distance to describe the orientation between the two antibody domains (Dunbar et al. [2013a]). This new methodology provided a consistent way to characterize the orientation between the two domains and thus explained why the previous approaches identified diverse sets of residues affecting it. The authors further demonstrate that the unbound anti-protein antigens are more flexible than the corresponding unbound anti-hapten antibodies.

Even though the models which can be constructed using current methods are of good quality, there still exists a need for improvement. For instance, docking antibodies as a tool to inform artificial affinity maturation requires very high quality models. In such a pipeline, one would generate a model of an antibody differing from the germline variety only by a limited number of mutations. So as to be able to distinguish the sometimes tiny differences between wildtype and the mutants would require very high-accuracy modelling tools.

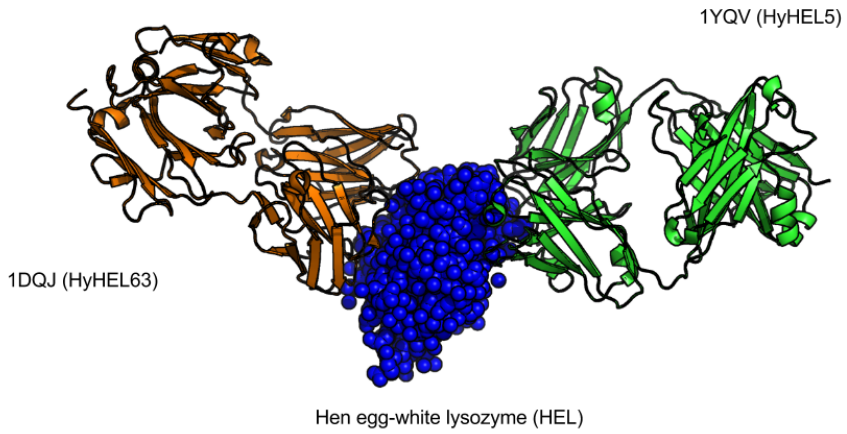


FIGURE 1.12: **Lysozyme and its binding antibodies**. Above are only two of the many different antibodies found in the PDB which have been shown to bind to distinct sites on the hen egg-white lysozyme. The many possible binding sites suggest that antibodies can bind to virtually any site on a target protein. This is in resonance with findings that epitope surfaces are virtually indistinguishable from general protein surfaces, which further complicates the task of characterizing immunogenic sites (Kunik and Ofran [2013], Sun et al. [2011], Kringelum et al. [2013]).

1.5.4 Epitope Prediction

Identifying regions on the antigen which are capable of binding an antibody is an important problem (Idrees and Ashfaq [2013], Gershoni et al. [2007], Irving et al. [2001]). It is also a difficult problem since epitope patches appear to be barely distinguishable from general protein surfaces (Kunik and Ofran [2013], Sun et al. [2011], Kringelum et al. [2013]). The most reliable epitope identification technique is X-ray crystallography which shows precisely the complex between the antibody and an antigen. There exist other experimental methods to identify epitope residues but all of them are also costly in time and resources. For this reason, the field of computational B-cell epitope prediction was developed, which aims to provide information on potentially immunogenic structures and sequences.

Computational epitope predictions can be divided into two general groups: linear epitope predictions and conformational epitope predictions. The linear epitope predictions aim to identify contiguous stretches in the antigen sequence which constitute the epitope. Conformatoinal epitope predictions aim to identify patches of sequence on the

antigen which, when folded, constitute the linearly discontinuous epitope. Around 90% of all known epitopes are conformational (Sun et al. [2013]). Nevertheless, most of the methods developed over the last 20 years addressed the easier problem of linear epitope identification (Reimer [2009]). In this thesis we consider conformational epitope prediction only.

Conformational B-cell epitope predictions can be classified into two types: those that ignore antibody information and those that do not. The methods that ignore the antibody information constitute the vast majority of conformational B-cell predictors (e.g. CEP, DiscoTope, ElliPro, PEPITO, PEPOP, SEPPA and EPITOPIA (Sun et al. [2013])). Consensus based methods like EPCES or the meta-server EPSVR/EpMETA are currently among the best performing algorithms in this area (Sun et al. [2013]).

The main aim of antibody-ignoring methods is to identify epitope-like sites on proteins as a means to improved vaccine design. Their mode of operation is fairly consistent. They use information from the PDB (Abola et al. [1984]), AntigenDB (Ansari et al. [2010]), the Conformational Epitope Database (CED) (Huang and Honda [2006]) or the Immune Epitope Database (IEDB) (Vita et al. [2010]) to identify a non-redundant set of epitopes determined by experimental methods. This is followed by the construction of a scoring function based on physicochemical properties such as surface accessibility, electrostatic contributions, hydrophobicity, electronegativity etc. There are two kinds of statistics that those methods collect: sequence based and structure based. It was noted that the structure-based descriptors are better at distinguishing epitope residues (Kringelum et al. [2012]). Moreover, methods with more descriptors do not tend to perform better than methods which use simpler techniques and less descriptors (Kringelum et al. [2012]).

Antibody-ignoring methods can provide valuable insight into what structural features can elicit immune responses. Nevertheless there exists the caveat that certain antigens can have many different antibodies binding at distinct positions. For instance, there are several solved complex structures of hen-egg white lysozyme (HEL) in the PDB. Superimposing those structures reveals that almost every part of HEL constitutes some epitope (see Figure 6.5). For this reason, it was argued that antibody information should

be included so as to achieve more specific conformational B-cell epitope predictions (Sun et al. [2011], Sela-Culang et al. [2013]).

The field of antibody-specific conformational B-cell epitope predictors is considerably underdeveloped in comparison to the methods which ignore the antibody. Currently, only three methods have been developed which address this problem (Rapberger et al. [2007], Soga et al. [2010], Zhao et al. [2011]). The earliest used only 26 antibody-antigen complexes (those available in 2007) to produce its predictions. They used the program FADE (Mitchell et al. [2001]) to quantify the complementarity between the paratopes and candidate epitope patches. The shape complementarity was augmented by physicochemical descriptors and binding energy computations, calculated using FastContact (Camacho and Zhang [2005]). On their test set of 26 antibody-antigen complexes they achieve 18% sensitivity and 87% specificity. The authors extended their method to run the predictions using many non-native antibodies towards a single antigen so as to identify all possible epitopes. Thus even though this method had an antibody-specific component, its main goal was the same as the antibody-ignoring mehtods - identification of all epitopes.

Another method which attempted to obtain antibody-specific predictions is the coupling of the Antibody-Specific Epitope Potential (ASEP) and DiscoTope (Soga et al. [2010]). ASEP was computed by counting residue-residue interface preferences from a non-redundant set of antibody-antigen complexes from the PDB. This potential was then used to constrain general epitope predictions made by DiscoTope, with respect to a single antibody.

The most recent antibody-specific epitope prediction method was developed in 2011 by Zhang et al. (Zhao et al. [2011]). Following their study of antibody-antigen complexes (Zhao and Li [2010]) they developed a method which treats an antibody-antigen interactions as a Hidden Markov Model (HMM). They used 80 antibody-antigen complexes to train their method. They achieve 43% sensitivity and 71% specificity. Nevertheless, the testing procedure has been performed using leave-one-out validation, which, as authors

admit, given the redundancy of their dataset, might have led to over-fitting (Zhao et al. [2011]).

Conformational B-cell epitope predictions are still very unreliable. Ponomarenko et al. showed that the maximum precision and recall of the antibody-ignoring conformational B-cell epitope predictors do not exceed 40% (Ponomarenko and Bourne [2007], Sun et al. [2013]). There was no comprehensive study benchmarking the antibody specific methods. The antibody-specific methods did not perform comparison with respect to one another and as they are all unavailable it is impossible to tell how they would perform on average (Sun et al. [2013]).

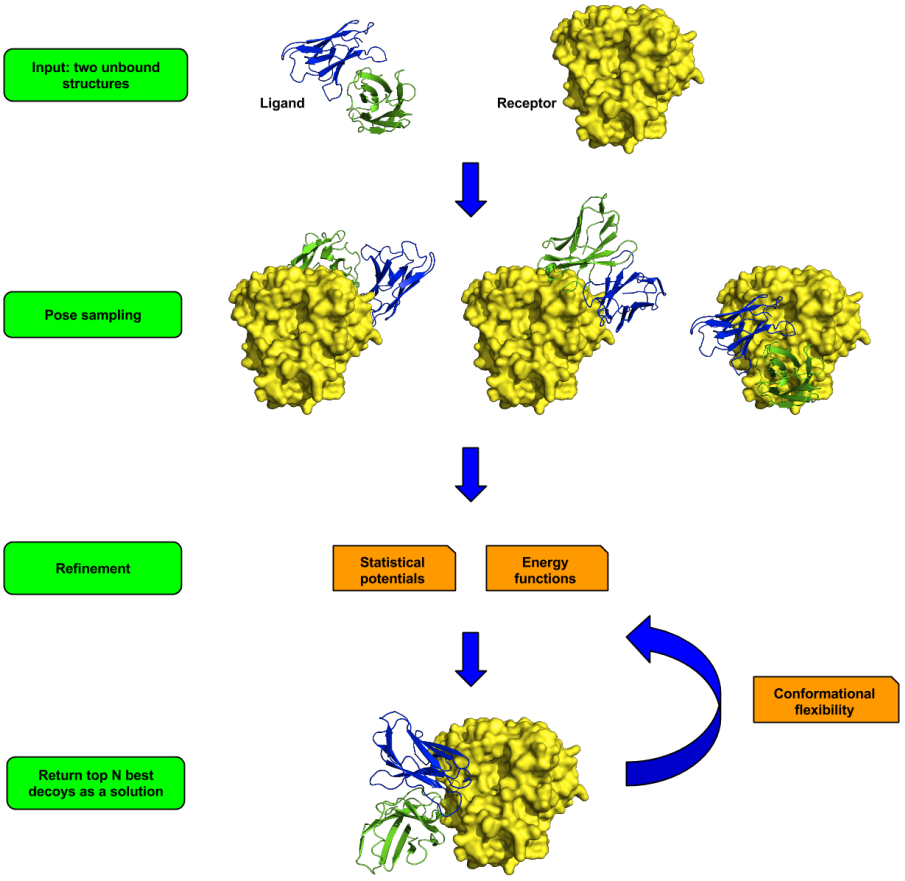


FIGURE 1.13: A typical docking pipeline is divided in two steps. Firstly, a set of poses of the ligand with respect to the receptor will be generated, using shape complementarity methods like Fast Fourier Transform (FFT) or geometric hashing. Next, the poses are scored using statistical potentials, energy functions or a combination of both.

1.5.5 Antibody docking

Crystallizing protein complexes is a more complex process than crystallizing single proteins. As a result, the number of possible complexes that can be formed using only the single structures in the PDB greatly surpasses current capacity to create reliable models for bound proteins. One way to approach this issue is to model the complex between the two proteins algorithmically using their unbound solved crystal structures - this technique is termed protein-protein docking (Ehrlich and Wadey [2003]).

In protein-protein docking, one is given as input the structures of two unbound proteins which are assumed to interact. One of the proteins, usually the larger one, would be termed the receptor and the other the ligand. The docking algorithm attempts to generate a set of poses of the ligand with respect to the receptor which would ideally approximate the bound complex of the two proteins.

There exists a myriad of approaches to tackle this problem (Dominguez et al. [2003], Gray et al. [2003], Chen et al. [2003], Duhovny et al. [2002], Schneidman-Duhovny et al. [2005], Kozakov et al. [2006]), which can generally be classified into two groups: template-based and template-free. Template based docking relies on the existence of solved complexes, homologous to the query proteins (Aloy and Russell [2002]). The antibody-antigen complexes undergo an ad-hoc adjustment of an antibody with respect to an antigen rather than the evolutionary process where both interacting partners are allowed to be optimized for binding. Therefore, methods that address the specific issue of antibody-antigen docking are template-free since such evolutionary data is unavailable and the antibody-antigen complexes significantly differ in their binding mode from general proteins (Krawczyk et al. [2013], Kozakov et al. [2006], Sircar and Gray [2010]).

The general pipeline of template-free docking algorithms is divided into two stages: sampling of the docking poses followed by refinement of such results (see Figure 1.13). The possible poses can be sampled probabilistically using Monte Carlo or Molecular Dynamics techniques or by methods which explore the discretization of the query proteins by Fast Fourier Transform or geometric hashing (Mendez et al. [2005], Chen et al. [2003],

Duhovny et al. [2002], Schneidman-Duhovny et al. [2005], Kozakov et al. [2006]). The sampling algorithms explore a large conformational space and thus, this stage is usually performed in rigid-body mode wherein backbone residues and side-chains are not allowed to be displaced. The more computationally expensive conformational flexibility can be introduced at the refinement stage.

The sampling stage produces a large number of possible poses, which usually contain a small number of close-to-native decoys (Halperin et al. [2002]). Therefore, the refinement stage poses the biggest challenge as it needs to distinguish these poses from the non-native conformations. The refinement techniques can broadly be classified into scoring functions and the introduction of conformational flexibility. The scoring functions include statistical potentials and energy functions. Statistical potentials evaluate the quality of a decoy by reference to atomic or amino acid counts in solved crystal structures or by the likelihood of a docking algorithm to be correct in a given atomic pairing (Chuang et al. [2008]). Energy functions estimate a number of physico-chemical descriptors such as electrostatics, Van der Waals interactions, desolvation or hydrophobicity (Kastritis and Bonvin [2010]).

The refinement stage might introduce flexibility to the structure so as to accommodate the conformational changes upon complex formation (Bonvin [2006]). The flexibility can be incorporated by randomly distorting the backbone residues/heavy chains or by ensemble docking (Bonvin [2006], Sircar and Gray [2010]). The latter approach explores a reduced conformational space around a group of amino acids through methods such as Monte Carlo sampling, aiming to minimize its energy with respect to a certain energy function (Sircar and Gray [2010]). The former approach relies on the generation of conformational ensembles of the target from NMR or by performing Molecular Dynamics. The elements of the conformational ensemble are then docked one by one or consensus decoys are generated by mean-field approaches (Chaudhury and Gray [2008]).

The field of protein-protein docking has made impressive progress recently, as demonstrated by the consecutive rounds of the CAPRI blind test experiment (Smith and Sternberg [2003], Mendez et al. [2005]). Currently, automated methods to recreate protein-protein complexes performed better than human predictors (Janin [2013]). Despite this impressive progress, in real terms, docking is still far from practical applications. The current standard of quantifying success is to consider the best decoys generated out of the top ten poses.

Currently docking algorithms are generally not optimized for a particular type of protein (Ponomarenko and Bourne [2007], Hwang et al. [2010], Mendez et al. [2005]). Some methods aim to incorporate protein-specific information in their pipelines, by specific input constraints, but at the core they remain unoptimized for the particular protein type.

Antibody-antigen specific docking methodology has been previously demonstrated to provide valuable insight into the development of an anti-Dengue antibody (Simonelli et al. [2013]) and the nature of the Influenza A/H1N1 virus (Cherian et al. [2011]). Antibody-antigen docking needs to be specific for this kind of protein as they have a radically different binding mode from that of general proteins (Krawczyk et al. [2013], Brenke et al. [2012]). There currently exist two sub-problems in antibody-antigen docking: local antibody-antigen docking when the epitope information is available and the much harder problem of global docking when no epitope information is known. The only two methods that have previously tackled these two problems are SnugDock (Sircar and Gray [2010]) for local antibody-antigen docking and the antibody mode of ClusPro for global antibody-antigen docking (Brenke et al. [2012]).

SnugDock is a local flexible docker, whose primary achievement is the ability to dock homology models generated using RosettaAntibody (Sircar and Gray [2010], Sivasubramanian et al. [2009]). They achieve it with a combination of successive rounds of backbone and rotamer relaxation and CDR packing optimization. Their best results are obtained by submitting the best models from SnugDock for post-processing by EnsembleDock. This ensemble methodology generates a model of the antibody-antigen

complex by taking a structural consensus of the best scoring decoys, thus feeding it several sub-optimal models might be sufficient to create a reliable model of the complex.

The development of the antibody mode of ClusPro consisted of the incorporation of an asymmetric statistical potential customized for antibody-antigen complexes. The statistical potential function for general proteins is calculated using a technique called *decoys as reference state* (Chuang et al. [2008]). The difference with respect to the conventional statistical potentials comes from the distinct way the background distribution of the residues is computed. The constituents of the complexes in the training set would be treated as inputs for the docking algorithm with no re-scoring step, only pose generation. The majority of poses generated in such a way are assumed to be incorrect and thus, in the same way as the solved complexes act as a distribution of correct contacts, those serve the opposite purpose of providing the incorrect, background distribution.

The main purpose of the ClusPro antibody mode is to tackle the global antibody-antigen docking. Due to the complexity of this problem, ClusPro is only capable of providing poses that very roughly approximate the native conformation of an antibody-antigen complex. As such, these poses can be used as starting points for refinement by more complex local docking algorithms such as SnugDock (Sircar and Gray [2010]).

These two methods for docking antibodies suffer from several limitations. It is still to be verified if the low-resolution decoys produced by ClusPro can serve as an input to other docking methods, creating a reliable global pipeline. SnugDock on the other hand is very expensive in computing resources. This could hinder any potential affinity maturation pipeline that would aim to dock dozens of mutants into the same epitope. The need for better quality antibody models is exemplified by SnugDock results. Even though the antibody models generated by RosettaAntibody are of good quality, docking them using SnugDock produces far poorer results than docking the unbound crystal structures using RosettaDock. The major drawback of SnugDock or EnsembleDock is that they take hundreds of CPU hours to produce decoys for a single target.

1.6 Outline of this thesis

In order to contribute to the field of computational antibody-antigen design, we have divided the work in the following order: we have collected antibody-related data and analysed the nature of antibody antigen interactions. We have used the information gathered from this analysis to design Antibody i-Patch - a method which predicts the precise residues on the antibody which form antigenic contacts. Those predictions were then used to develop a local antibody-antigen docking pipeline, which relies on the knowledge of the epitope. A logical continuation from local antibody antigen docking was to address the corresponding global problem wherein the epitope information is unknown. In order to tackle the global docking issue, we have developed a novel antibody-specific B-cell epitope prediction method, EpiPred. We have used the epitope predictions obtained by EpiPred to constrain the results of global antibody-antigen docking which enriched the top poses with more close-to native decoys. The work outlined here is divided into five research Chapters. A brief description of each of the Chapters is given below.

The first research Chapter covers the development of tha Structural Antibody Database (SAbDab) and is a joint work with James Dunbar and Jinwoo Leem. The main goal of this new database was to present antibody-related data from the PDB in a consistent and up-to-date manner. SAbDab provided the underlying data we have used in the studies in the following Chapters and for this reason those are presented first.

The second research Chapter of this thesis covers the first stage of this D.Phil project, which aimed to determine which residues if any play key roles in the antibody-antigen binding. In an attempt to achieve this the focus was given to the main constituents of the antibody binding site, namely CDRs. Firstly, the most frequent residues in hypervariable loops across several species were compared, concluding that the broad trends of over-representation of tyrosine and serine are conserved even if individual compositions between organisms are statistically different. Secondly, it was confirmed that those trends are well maintained throughout the organisms life. Thirdly, structural

analysis was carried out, exploring length and conformation of CDRs. It was found that lengths of the hypervariable loops recombine uniformly at random, contrary to previous results. The small number of observed combinations appears to originate from the fact that there are dominating CDR lengths which introduce combination biases. Fourthly, it was found that there is no strong tendency for hypervariable loops to leave their canonical class conformations upon binding. Fifthly, the roles of all CDRs for antigen binding were studied, with conclusion that all hypervariable loops are heavily involved in antibody-antigen contacts. Finally, the propensities of the antibody binding site residues to be in contact with the antigen were analyzed. Comparing the binding propensity of antibody molecules against other proteins, it was concluded that antibodies achieve binding by a distinct mechanism. Moreover, antibodies appear to have strong preferences for amino acids that are involved in antigen binding. These residues are tyrosine, tryptophan and histidine, which have a significantly higher binding propensity in antibodies than in other proteins.

The third research Chapter of the thesis concerns the prediction of the antibody binding sites. Here, we present Antibody i-Patch, a method to annotate the most likely residues to be in contact with the antigen. Using the sequence of the antibody as the input, we achieve better prediction results than the current leader in the field, Paratome. Moreover, our binding annotations are not binary yes/no as in Paratome and the CDR definition methods, but rather a contact likelihood score. Therefore, by specifying a higher cut-off for such a likelihood, one can trade recall for increased precision and vice-versa. We show that our predictions correlate with energetic importance and thus we argue that these may be useful in guiding mutations in the artificial affinity maturation process.

In the fourth research Chapter, we discuss the results from the study of antibody-antigen local docking. We use our predictions for the antibody binding site to constrain the search space for two popular rigid-body docking algorithms, ZDOCK and Patch-Dock. We also develop a scoring algorithm which re-orders the decoys from ZDOCK

and PatchDock, offsetting their internal biases for general proteins. Docking either homology models or solved crystal structures, we obtain high quality results which are arrived at several orders of magnitude faster than using the current state of the art method in the field of local antibody-antigen docking: SnugDock. We thus propose that the results from our pipeline can be used as good starting poses for more complex methods such as SnugDock that could further improve the quality of local antibody-antigen docking.

The fifth research Chapter covers the work carried out on B-cell epitope prediction and global antibody-antigen docking. In the previous Chapter, we were performing local antibody-antigen docking since the epitope information was unavailable. In this Chapter we have developed a novel antibody-specific B-cell epitope prediction method, EpiPred. We use our epitope predictions to constrain the results from global antibody-antigen docking. We demonstrate that including our epitope predictions enriches the top results of the state of art global antibody-antigen docker ClusPro in antibody mode (Brenke et al. [2012]) and ZDOCK (Chen et al. [2003]) with more close to native decoys. Just as in the case of local antibody-antigen docking we demonstrate that our epitope-prediction and global docking pipeline can receive a homology model of an antibody and the unbound form of the antigen to produce satisfying results.

The final Chapter concludes the work demonstrated in this thesis and presents its prospects for the future.

Chapter 2

SAbDab: The Structural Antibody Database

2.1 Introduction

As mentioned in the previous Chapter, there already exist several antibody-specific databases (Johnson and Wu [2001], Martin [1996], Allcorn and Martin [2002], Retter et al. [2005], Lefranc et al. [2009], Martin [2010], Ansari et al. [2010], Ponomarenko et al. [2011], Chailyan et al. [2012]). Most of those services focus on the sequences of antibodies, e.g. Kabat (Johnson and Wu [2001]), Kabatman (Martin [1996]), DIGIT (Chailyan et al. [2012]) and Vbase2 (Retter et al. [2005]). Another set of databases address epitope data specifically, often ignoring antibody information: AntigenDB (Ansari et al. [2010]), Conformational Epitope Database (CED) (Huang and Honda [2006]) or Immune Epitope Database (IEDB) (Vita et al. [2010]).

There are databases which do provide antibody structural data including the PDB (Abola et al. [1984]), IMGT/3DStructure-DB (Lefranc et al. [2009]), Summary of Antibody Crystal Structures (SACS) (Allcorn and Martin [2002]) and Abysis (Martin [2010]). The PDB is not dedicated to antibodies specifically but remains the biggest resource for antibody structures to date. IMGT/3DStructure-DB and Abysis are discovery tools which allow inspection of individual antibodies but lack the functionality which would facilitate the generation of larger structural subsets given a set of constraints. SACS is a summary list of antibody structures in the PDB with a very limited set of constraints to sort the list of structures available through the service. In this Chapter, we will describe the development of an antibody database centred around their structure.

Due to the dynamic nature of the field of antibody research, structural antibody data is made available at an ever increasing pace. For example, over the course of this D.Phil around 400 antibody structures were deposited in the PDB (see Figure 2.1). Antibodies now comprise about 1.75% of the PDB (Dunbar et al. [2013b]). In a data-driven study extended over several years, as presented in this thesis, it is necessary to keep adjusting the datasets to be as accurate and as up-to-date as possible. In this initial research Chapter we will present the data underlying the analysis in the following Chapters as well as describe the tools made available to the research community that resulted from the effort.

2.1.1 Contributions

Since the work presented in this Chapter was a joint effort, the contributions of each participant are outlined below.

In the Structural Antibody Database (SAbDab), James Dunbar was primarily responsible for antibody data collection and storage. He developed a back-end functionality that allows access to the data from the command line. He also contributed to the development of the web-interface in the later stages of its development. Jinwoo Leem was responsible for collection and verification of experimental affinity data associated with

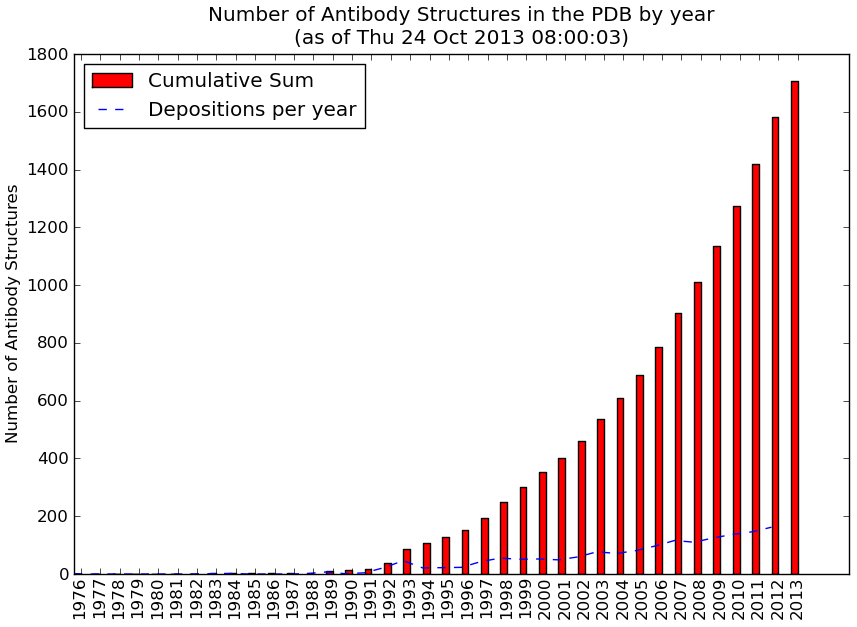


FIGURE 2.1: **Antibody structures in the PDB.** This image was taken from SAbDab (Dunbar et al. [2013b]), showing the rapid growth in the number of antibody structures in the PDB. Over the last four years (2010-2013) around 400 antibodies were deposited in the PDB. The majority of the antibody entries in the PDB are X-ray diffraction structures (1688 at the time of creation of this plot).

some structures in the database. I developed the web front end which handled the data available through the command line tool developed by James. I also implemented the CDR database and the CDR clustering tools. The majority of this work is described in (Dunbar et al. [2013b]).

We have also incorporated several of our anitbody-related tools into the SAbDab platform. James Dunbar made his ABngle (Dunbar et al. [2013a]) and antibody Template Search programs available through the SAbDAb and I developed the web interface for ABangle and Template Search along with James. I have made my Antibody i-Patch (Chapter 4) and DockSorter (Chapter 5) tools (Krawczyk et al. [2013]) available through SAbDab.

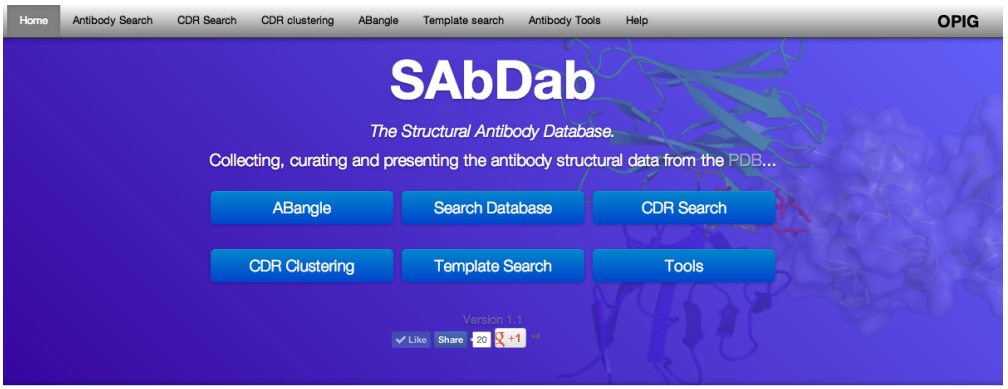


FIGURE 2.2: **Structural Antibody Database (SAbDab).** The main page of SAbDab. The new database offers up-to-date and consistent representation of the antibody data available in the PDB. It can function as a discovery tool for antibodies and CDRs. It also offers the functionality to download large and up-to-date datasets of antibody and CDR structures for analysis.

Tools

We provide a selection of tools allowing researchers to study the structures of antibodies:

Database search	Search antibody structures by complex, name, organism etc.
CDR Search	Search CDR structures by complex, name, organism etc.
ABangle	Explore the V_H/V_L orientation angles
Antibody-Antigen docking	Download software for decoy re-scoring
CDR contact prediction	Predict CDR residues to be in contact

SAbDab

Below you can find the fundamental statistics from our database:

Structures with antibody:	1712
Structures with at least one paired V_H/V_L:	1509
Number of F_V regions:	3207
Structures with defined affinities:	191

CDR clusters

UPGMA clustering of the Chothia CDRs, allowing elements in the cluster to be at most 1.5Å apart.

CDR type	Number of available lengths	Non-singleton clusters
H1	12	107
H2	5	96
H3	25	257
L1	11	58
L2	3	82
L3	9	99

FIGURE 2.3: **SAbDab statistics.** The number of antibody structures available through our service is updated weekly (middle). We also present the current number of observed CDR conformations (right). We also present the antibody tools currently available through the service (left). The service has already been visited by 275 unique users since its launch in the summer 2013.

2.2 The Structural Antibody Database (SAbDab)

The Structural Antibody Database was developed so as to provide an up-to date and consistent access to antibody structure data (see Figures 2.2 and 2.3). The structure-centric nature of the service distinguishes it from sequence-based databases such as Kabatman or DIGIT. SAbDab is updated weekly, thus ensuring that the research community has access to the latest structures deposited in the PDB.

The service offers a range of services which allow for filtering the data, including general structure descriptors such as source organism, resolution as well as antibody-specific

characteristics such as paired heavy/light chains, antibody-antigen affinity value, presence of antigen etc. The service can function as a discovery tool, allowing inspection of individual antibodies. It is distinct however from other discovery tools such as IMGT/3DStructure-DB or Abysis since it also allows the search results to be downloaded in bulk for large-scale analysis.

The database further offers a range of CDR-related tools. These include a CDR search facility which allows user to filter CDRs according to their definition, length, presence of antigen, resolution etc. Results from the search can be inspected individually or, as is the case with antibody structures, downloaded in bulk. The CDR-clustering functionality supplements the CDR database by providing an up-to date grouping of the conformations assumed by CDRs.

SAbDab also functions as a platform for antibody-related tools developed in the Oxford Protein Informatics Group. In this function, it currently offers the following functionality which is presented in this thesis: CDR-Clustering (Chapter 3), Antibody i-Patch (chapter 4) and DockSorter (Chapter 5). The epitope prediction and global docking pipeline presented in Chapter 6 are currently in the process of being integrated into SAbDab.

2.2.1 Antibody Data Collection

Antibody data in the PDB is not deposited in a completely consistent manner. Although it is possible to mine for keywords in the titles and headers of entries such as *IgG*, *antibody* etc. this approach is unreliable. Therefore our solution for extracting antibody entries from the PDB is to compare the sequences constituting a structure to highly conserved immunoglobulin sequence profiles (Kunik et al. [2012], Zhao et al. [2011]).

Structures of antibodies are mined directly from the PDB using Abnum (Abhinandan K R [2008]). Abnum is a program developed by Abhinandan and Martin in 2006, which numbers a sequence of an antibody using the improved Chothia system, identifying

light and heavy chains. If the program cannot number a given amino-acid chain, it means that the sequence is unlikely to be that of an antibody. Therefore sequences of new structures in the PDB are screened using Abnum and any that can be numbered successfully is considered to contain an antibody and thus is added to SAbDab. Since some of the structures contain single chain F_v entries (scFv), Abnum is applied to each chain recursively to account for this. If Abnum identifies a given chain as being both heavy and light, it means that it is an scFv structure. Heavy and light chains are paired if a conserved cystine at position 92 is within 22Å of the conserved cysteine at position 88, as numbered by the Chothia system.

If only some of the chains in a PDB entry are identified by Abnum as antibody chains, the remainder are treated as potential antigens. Those potential antigen sequences are aligned to antibody profiles using MUSCLE (Edgar [2004]). A sequence is classified as antigen if it has less than 35% sequence identity to any antibody profile. Any sequence that cannot satisfy this condition is marked for manual inspection. Molecules which are common solvents used for crystallization are discarded (Weichenberger et al. [2013]). An antibody chain is paired with its cognate antigen if the antibodies CDR residues are within 7.5Å of the antigen. If more than one antibody satisfies this condition with respect to the same antigen, the structure is flagged for manual inspection.

2.2.2 Development of the front-end for the antibody database

The database developed by James Dunbar and the associated antibody analysis tools were made available to the scientific community by the interface I developed. Since we envisage SAbDab to be the primary structural resource for antibody information, the web-service had to be developed in a modular and extensible way which would be robust with respect to the dynamic data is represents.

The web-service has been developed in a two tier system: the user-facing web framework and the module which communicates with the database. For the web-side of SAbDab, we used Twitter Bootstrap, which recently has become a popular framework for the

development of HTML5/Css/Javascript websites. The popularity of the library ensures the continuing support from its developers as well as ease of immersion for others to adjust or maintain the service.

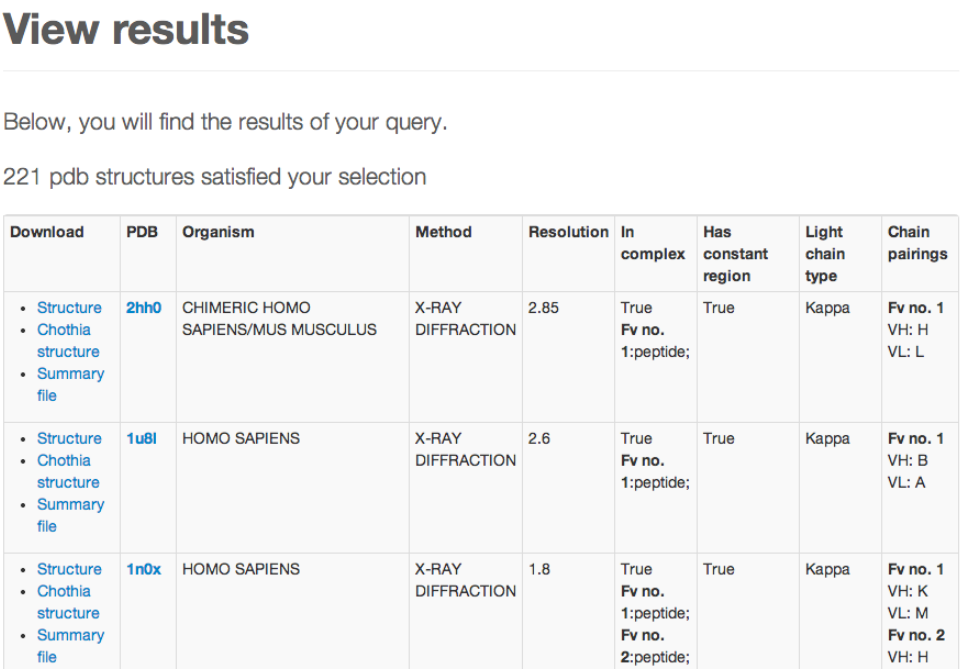
The web-framework communicates with the back-end of SAbDab through a python module designed specifically for this task. This module has been designed in a modular and extensible way in order to facilitate the connection between the web-framework and the back-end functionality. This module handles calls not only to the antibody database developed by James Dunbar but also to the CDR database, CDR Clustering database and other antibody tools which have already been integrated into the service. Each new antibody tool or database will be integrated into SAbDab through the API offered by this module as it was designed specifically for ease of extensibility.

Due to the inconsistent manner in which some PDB entries are deposited, it is possible that the service will contain errors. For this reason, I have implemented a misannotation functionality which allows an expert user to easily flag the erroneous structures. A link is displayed next to each structure and upon clicking, the expert user is taken to a form which allows her to either flag the particular structure or to give more detail about the nature of the error.

2.2.3 Database Search

The main focus of the web-interface is the Database Search tab on the main page (see Figure 2.2). Here, we offer a search functionality which is capable of displaying certain subsets of the database defined using a variety of characteristics. It is possible to define the subsets using general protein descriptors such as source organism, resolution, experimental technique etc. The web-interface also offers a range of antibody-specific tools including: light chain gene, antigen information, presence of the constant domain etc.

An example result for a search of all antibodies *in complex with a peptide* and with *resolution better than* 3.0Å is given in Figure 2.4. There were 221 antibodies satisfying



View results

Below, you will find the results of your query.

221 pdb structures satisfied your selection

Download	PDB	Organism	Method	Resolution	In complex	Has constant region	Light chain type	Chain pairings
• Structure • Chothia structure • Summary file	2hh0	CHIMERIC HOMO SAPIENS/MUS MUSCULUS	X-RAY DIFFRACTION	2.85	True **Fv no. 1**:peptide;	True	Kappa	**Fv no. 1** VH: H VL: L
• Structure • Chothia structure • Summary file	1u8l	HOMO SAPIENS	X-RAY DIFFRACTION	2.6	True **Fv no. 1**:peptide;	True	Kappa	**Fv no. 1** VH: B VL: A
• Structure • Chothia structure • Summary file	1n0x	HOMO SAPIENS	X-RAY DIFFRACTION	1.8	True **Fv no. 1**:peptide; **Fv no. 2**:peptide;	True	Kappa	**Fv no. 1** VH: K VL: M **Fv no. 2** VH: H

FIGURE 2.4: **Example search result.** The filter consisted of the structure in complex with a peptide with resolution of 3Å or better. There were 221 structures which satisfied this condition but because of space constraints we only show the top three.

this condition as of November 2013. Such results can be examined using our antibody-specific discovery tool (see Figure 2.5) or downloaded in bulk.

In order to facilitate the large-scale analsis of antibodies we have also provided a non-redundant search facility. Using this service it is possible to create a subset of antibodies or antibody-anitgen complexes which satisfy certain sequence identity cutoffs. The sequence identity is computed directly using CD-HIT (Li and Godzik [2006]). This service has been used in Chapters 4, 5 and 6 of this thesis.

2.2.4 CDR Database

Our CDR database has been constructed so that it accommodates the three major CDR definitions: Kabat, Chothia and Contact (see section 1.4.1.2 for more details). Each of these three definitions can be defined with respect to the Chothia numbering coordinates. Thus the CDR database is constructed by firstly annotating each antibody chain in SAbDab according to the Chothia system and then extracting each CDR directly

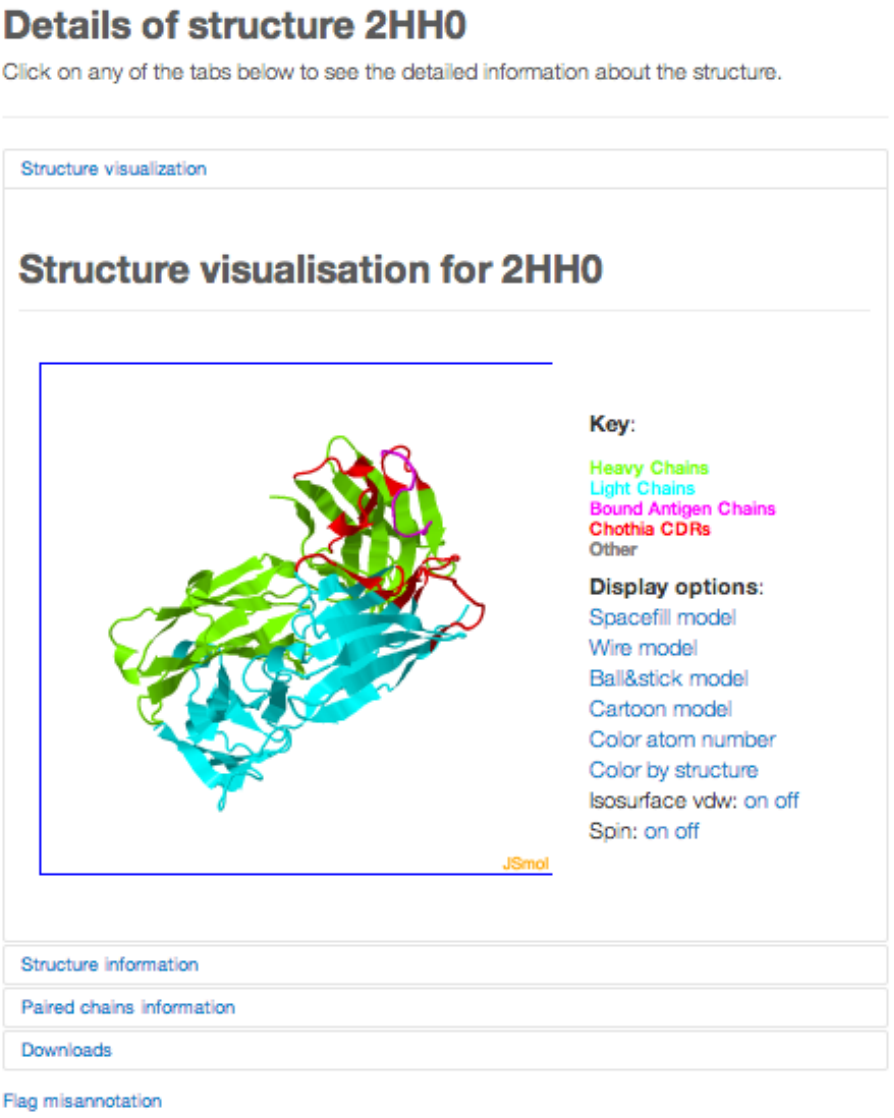


FIGURE 2.5: **Antibody structure discovery tool.** Individual structures can be visualized and analysed using the antibody discovery tool. Here, one can find the information relating to the structural parameters of the PDB entry, antibody-specific information such as paired antigens as well as links to downloads of the processed versions of the antibody (i.e. Chothia-numbered structure).

from the appropriately numbered sequences. Loops with missing residues are identified by aligning the structural sequence to the actual sequence of the antibody chain. The fourth CDR definition, namely this of IMGT, can only be identified by a set of more complex rules, often requiring manual inspection. Therefore we provide links to the IMGT website where these are available.

Using SAbDab, it is possible to create subsets of CDRs using the same characteristics as for antibodies: presence of the antigen, light chain gene, resolution, source organism

PDB	CDRs	Organism	Method	Resolution	In complex
4lsv	CDR_type: L1; Parent V_H/V_L: HL; Sequence: QANGYLN; More details: Details	HOMO SAPIENS	X-RAY DIFFRACTION	3.0	True
3rpi	CDR_type: L1; Parent V_H/V_L: -L; Sequence: QANGYLN; More details: Details CDR_type: L1; Parent V_H/V_L: -B; Sequence: QANGYLN; More details: Details	HOMO SAPIENS	X-RAY DIFFRACTION	2.648	True
4jpv	CDR_type: L1; Parent V_H/V_L: -L; Sequence: QANGYLN; More details: Details	HOMO SAPIENS	X-RAY DIFFRACTION	2.827	True

FIGURE 2.6: **Results of a search in the CDR database.** The search filter was defined as CDR structures according to the Chothia definition, type L1, length 7, in complex with the antigen and being of resolution quality of 3Å or better. The search returned many more structures but because of space constraints we only demonstrate the top three results shown here.

PDB	CDRs	Organism	Method	Resolution	In complex
1ghf	CDR_type: H1; Parent V_H/V_L: HL; Sequence: YTFTDY; More details: Details	MUS MUSCULUS	X-RAY DIFFRACTION	2.7	False
1qok	CDR_type: H1; Parent V_H/V_L: AA; Sequence: SSSVSY; More details: Details	MUS MUSCULUS	X-RAY DIFFRACTION	2.4	False

FIGURE 2.7: **Example of a non-redundant search on the CDR database.** The filter consisted of the Kabat H1 CDRs of length 6. We only had two non-redundant representatives in our CDR database shown above.

etc. Additionally, it is possible to select CDRs using characteristics particular to the hypervariable loops, namely the CDR definition, loop length etc. For instance, Figure 2.6 indicates the search results of CDRs defined by Kabat, of type L1 and length 7. We also offer a functionality to apply a sequentially non-redundant filter which should further facilitate large-scale structural analysis of CDRs (see Figure 2.7 for an example). The subsets of CDRs constructed in this fashion can be inspected using a CDR-customized discovery page (see Figure 2.8) or downloaded in bulk for large-scale analysis.

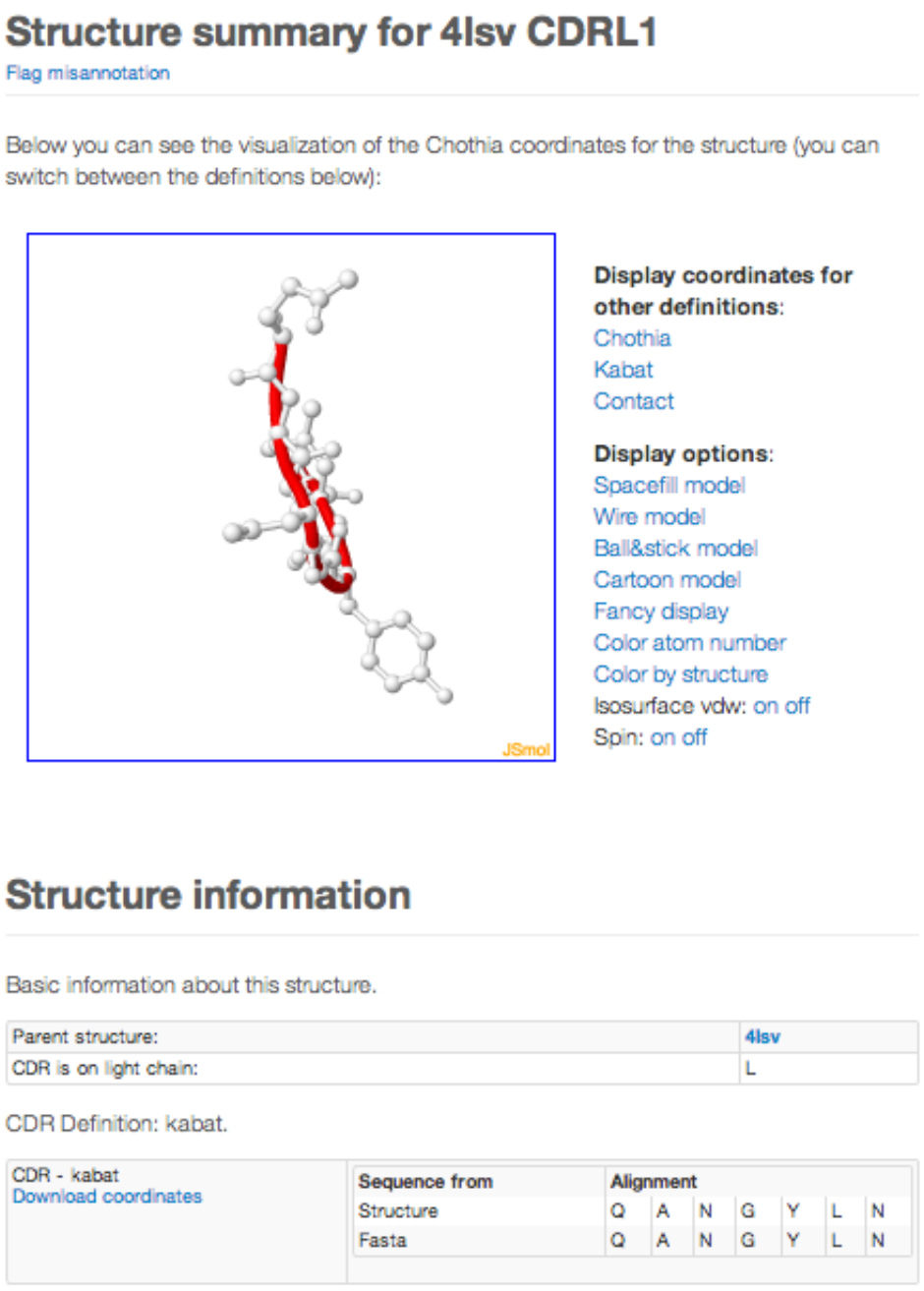
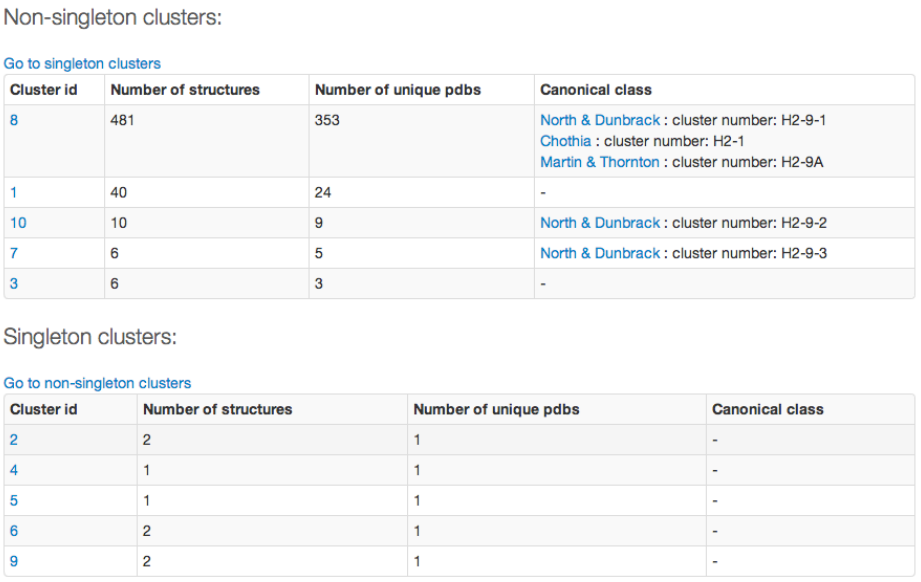


FIGURE 2.8: **CDR structure discovery page.** We offer a functionality to inspect individual structures of the CDRs and their attributes.

2.2.5 CDR Clustering

Non-H3 CDRs adopt a limited number of structural conformations that have been studied extensively over the last 25 years (Chothia and Lesk [1987], Al-Lazikani et al. [1997], Lara-Ochoa et al. [1996], Chothia et al. [1989], North et al. [2011], Martin and Thornton [1996]). It was noted by North et al. in 2010 that with the higher number of



Non-singleton clusters:

Go to singleton clusters

Cluster id	Number of structures	Number of unique pdbs	Canonical class
8	481	353	North & Dunbrack : cluster number: H2-9-1 Chothia : cluster number: H2-1 Martin & Thornton : cluster number: H2-9A
1	40	24	-
10	10	9	North & Dunbrack : cluster number: H2-9-2
7	6	5	North & Dunbrack : cluster number: H2-9-3
3	6	3	-

Singleton clusters:

Go to non-singleton clusters

Cluster id	Number of structures	Number of unique pdbs	Canonical class
2	2	1	-
4	1	1	-
5	1	1	-
6	2	1	-
9	2	1	-

FIGURE 2.9: **CDR clustering example results.** Clustering of CDRs according to Chothia definition of type H1 and length 5. The corresponding clusters from previous publications are given in the rightmost column. We distinguish between singleton and non-singleton clusters in order to highlight structural outliers.

antibody structures available, there is a rise in the number of canonical classes (North et al. [2011]). Each of the studies in the recent 25 years only focused on the snapshot of antibody structures available at the time. In order to obtain an objective and up-to-date view of the conformational space of CDRs, there exists a need to perform CDR clustering as new structures become available.

Using the CDR dataset described in the previous section, we create our own CDR clustering which is updated as new structures are added to the database. The clustering is performed on each group identified by: CDR type (H1, H2, L1, L2, L3), CDR definition (Kabat, Chothia or Contact) and length (according to the CDR definition). For each group, we create a Root Mean Square Deviation (RMSD) distance matrix using the Kabsch algorithm (Kabsch [1976]). The algorithm performs a pairwise structural optimal alignment of the heavy atoms of the backbones of the two loops. Each entry in the distance matrix corresponds to an RMSD between two CDR loops after optimal structural superposition. We perform UPGMA clustering on each distance matrix (Sokal [1958]). UPGMA creates a tree with leaves corresponding to CDR loops and branch distances proportional to the Root Mean Square Deviations (RMSD) between the structures. In

order to create the clustering, one descends down the UPGMA tree starting from the root and when all the pairwise distances in the subtree are less than a certain RMSD cutoff, the CDR loop-leaves are assigned to a single group. In the database we offer clusterings for a range of RMSD cutoffs: 0.5Å, 0.75Å, 1.0Å and 1.5Å.

In order to maintain a link to previous work in this field, we have created a mapping from our clustering to previous ones, namely: North/Dunbrack (North et al. [2011]), Martin/Thornton (Martin and Thornton [1996]) and Chothia (Chothia et al. [1989]). The mapping was created using the representative structures given for each canonical class in the previous clusterings, identifying which SAbDab cluster it was assigned to. For instance, Figure 2.9 presents search results of Chothia CDR clusters of H1, length 5 at UPGMA cutoff of 1.5Å and their corresponding canonical structures from the previous publications.

The previous clusterings only defined a handful of groups in comparison to ours, mostly because they were ignoring structures that were not obviously assignable to any bigger group. For instance, Martin et al. defined 25 canonical structures whereas North et al. 72, for all the CDR types excluding H3. In comparison, SAbDab currently has 107 non-singleton clusters for H1 according to the Chothia definition alone at UPGMA cutoff of 1.5Å (see Figure 2.10).

CDR type	Number of available lengths	Non-singleton clusters
H1	12	153
H2	5	151
H3	25	277
L1	11	133
L2	3	114
L3	9	149

1.0Å

CDR type	Number of available lengths	Non-singleton clusters
H1	12	107
H2	5	96
H3	25	257
L1	11	58
L2	3	82
L3	9	99

1.5Å

FIGURE 2.10: **Example comparison of clusters in our databases.** We contrast the number of clusters available for the Chothia definition, using UPGMA cutoffs 1.0Å and 1.5Å.

2.3 Conclusions

The Structural Antibody Database facilitated the creation of up-to-date reliable datasets for computational antibody design. The datasets created using the service benefited the analysis in most of the Chapters in this thesis, ensuring that we were using all the data available. Furthermore, SAbDab constituted a platform to make our tools for computational antibody design available to the scientific community. The service has become a primary resource for structural antibody data and as such we expect that it is going to have an impact on the broader scientific community beyond the work presented in this thesis.

Chapter 3

Characterisation of the antibody binding site

3.1 Introduction

In this Chapter we describe an analysis of antibody-antigen interactions. We embarked upon the task so as to identify features particular to antibody-antigen complexes which we could later exploit to build tools for computational antibody design. We begin by providing a brief summary of the antibody-antigen binding mechanism which we built upon, followed by our findings.

3.1.1 Antibody-Antigen interactions - previous work

In the previous Chapter we have provided an overview of the current state of understanding of antibody-antigen interactions. Here we review the main concepts which are relevant to the findings presented later in this Chapter.

The antibody binding site is chiefly composed of the Complementarity Determining Regions (CDRs), alternatively termed the hypervariable loops (Wu and Kabat [1972]). There are three CDRs on the light chain, called L1, L2 and L3, and as many on the heavy chain, called H1, H2 and H3. Typically, one refers to a loop by specifying its type and length, e.g. H1-8, meaning H1 loops of length eight. These loops are instrumental in antibody's ability to bind diverse antigens.

The antibody binding site has been previously explored from the compositional perspective. It was noted that the CDRs have a statistically significantly different residue composition from other soluble protein loops (Collis et al. [2003]). Even though the compositions of sequences of CDR-H3 were demonstrated to be statistically significantly different between mouse and human, tyrosine and serine appeared to be over-expressed in both organisms (Zemlin et al. [2003]). This over-representation was also observed under a temporal constraint, when tyrosine and serine maintained high concentrations in the CDR sequences of the Mexican axolotl throughout its life (Golub et al. [1997]). The special role of tyrosine and serine in antibodies was further corroborated experimentally when binding sites composed only of tyrosine and serine achieved specific and high-affinity binding (Fellouse et al. [2005]).

Another aspect of the CDRs which appears to be intimately linked to their function is their length. It was demonstrated that the length distributions of the CDRs appear to be associated with the type of antigen being bound (Collis et al. [2003]). Furthermore, hypervariable loop length is a major factor in determining its structure, as demonstrated through many attempts at creating the CDR canonical classes (Chothia and Lesk [1987], Chothia et al. [1989], North et al. [2011], Lara-Ochoa et al. [1996], Martin and Thornton [1996], Al-Lazikani et al. [1997]). The only hypervariable loop that appears not to form any canonical groupings is H3, which is believed to be the most important of all CDRs (James et al. [2003], Stanfield et al. [2007], Rini et al. [1992], Xu and Davis [2000], Schroeder Jr et al. [1998]).

The uniformity of CDR canonical classes has also been investigated in the broader context of the entire antibody binding site. It was noted that only a very limited number

CDR Definition	Type	Source				
		Human	Mouse	Rabbit	Bovine	Chicken
Kabat (Martin [1996])	H1	864	522	139	33	33
	H2	1533	1363	307	33	43
	H3	2222	1997	387	23	49
	L1	1212	694	85	55	61
	L2	768	361	46	37	37
	L3	1738	884	138	56	78
Chothia (Martin [1996])	H1	755	376	98	20	22
	H2	1029	768	230	28	37
	H3	2222	1997	387	23	49
	L1	1212	694	85	55	61
	L2	768	361	46	37	37
	L3	1738	884	138	56	78
AbM (Martin [1996])	H1	1202	740	191	33	41
	H2	1367	1145	284	32	43
	H3	2222	1997	387	23	49
	L1	1212	694	85	55	61
	L2	768	361	46	37	37
	L3	1738	884	138	56	78
IMGT (Giudicelli et al. [2006])	H1	3647	917	211	n/a	n/a
	H2	4091	1455	270	n/a	n/a
	H3	8136	5710	301	n/a	n/a
	L1	1944	495	51	n/a	n/a
	L2	508	154	18	n/a	n/a
	L3	3794	924	133	n/a	n/a

TABLE 3.1: Statistics for sequence dataset A1. Shown are the numbers of unique CDR sequences for each CDR definition, type and organism we used in the analysis. The different numbers of sequences across datasets stem from the associated distinct definitions of CDRs.

of available combinations of canonical class loops are observed in the X-ray structures of antibodies, significantly restricting the shape of the antibody binding site (Lara-Ochoa et al. [1996]).

Summarizing, the antibody binding site is chiefly made up of the six CDRs. The most important of the six loops is theorized to be H3. Geometry of the antibody binding site is restricted by the canonical structures of the CDRs and evidence suggests that only some combinations of these canonical forms are observed. It has previously been demonstrated that the hypervariable loops have significant over-representation of tyrosine and serine, a property which appears to be maintained across species and during an organism's life. It is thought that tyrosines in antibody CDRs mediate Ab-Ag contacts while serine confers loop flexibility. The sufficiency of serine and tyrosine for Ab-Ag binding was confirmed experimentally where a combining site composed solely of tyrosine and serine achieved complex formation (Fellouse et al. [2005]).

3.2 Materials and Methods

3.2.1 Data

Our analysis of antibody binding sites used three core datasets: CDR sequences, zebrafish D region sequences (the most variable region of the heavy chain, usually part of H3) and CDR-annotated antibody structures. These datasets are described in the list below and explained further in the text.

A1 Two CDR sequence sets, Kabatman (Martin [1996]) and IMGT (Giudicelli et al. [2006]) were downloaded. The Kabatman set was annotated in three different ways: Kabat (Kabat et al. [1992]), Chothia (Chothia and Lesk [1987], Al-Lazikani et al. [1997]) and AbM (Abhinandan K R [2008]) definitions. IMGT sequences were annotated using the IMGT definitions (Lefranc [2011]).

A2 Data from a high-throughput zebrafish study (Weinstein et al. [2009], Jiang et al. [2011]). This contains the D region sequences for five time points over the lifetimes of 14 zebrafish.

A3 IMGT CDR-annotated antibody structures (Ehrenmann et al. [2010]) extracted from the PDB (Abola et al. [1984]).

There are several definitions of CDRs but the one accepted by the World Health Organization is that of IMGT (Lefranc [2011]). Due to the breadth of such definitions (Abhinandan K R [2008]), others are still used for completeness and comparison purposes.

In the composition analysis of dataset A1, only complete and unique sequences were retained and the total numbers used in this analysis are presented in Table 3.1. Background data of 46682 $\beta - \beta$ anti-parallel protein loop sequences from non-antibody proteins were provided by Yoonjoo Choi who curated the dataset for the purposes of FREAD loop modelling software (Choi and Deane [2010]). Out of those 46682 loops,

only the 17022 sequence-unique and complete loop sequences from this dataset were retained.

Dataset A2, produced by Weinstein et al. (Weinstein et al. [2009], Jiang et al. [2011]), was employed for composition analysis over time, consisted of sequences of zebrafish heavy chains coming from 14 fish at ages of 2 weeks, 1 month, 3 months, 6 months and 1 year. The D regions were extracted from all sequences and only the sequence-unique D region sequences were retained for the study.

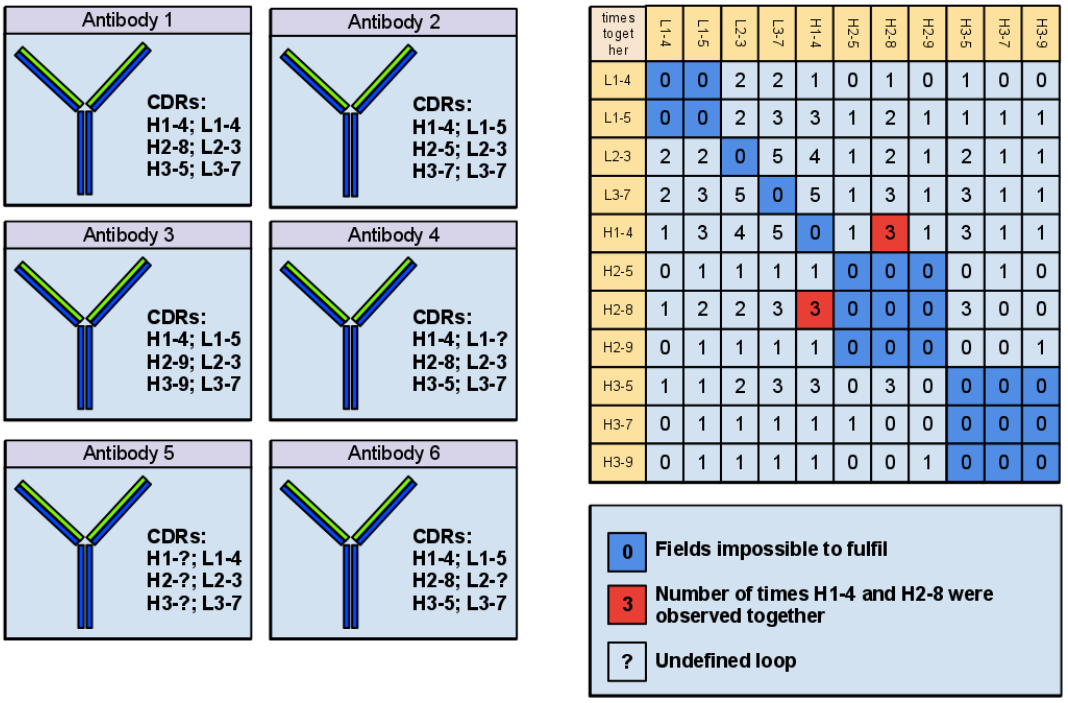
For the structural dataset A3, a total of 1064 antibody X-ray structures were extracted from the PDB. The PDB codes came from a dataset provided by Dr Jiye Shi, augmented with structures from SACS (Allcorn and Martin [2002]). Those structures were then constrained to those with resolution better than 2.8Å. This structural dataset served as a basis for three studies: CDR clustering, CDR length correspondences and the antibody-antigen binding analysis.

For the CDR clustering study, hypervariable loops were extracted from all structures in A3. To accommodate the breadth of CDR definitions in the literature (Abhinandan K R [2008]) the loops used in further analysis (as defined by IMGT (Ehrenmann et al. [2010])) are taken together with their surrounding anchor regions of three residues on either side. CDRs with high (> 80) or zero B-factor and missing or poorly defined residues were removed. In order to create a non-redundant set of CDRs, whenever two hypervariable loops had identical sequences and their Root Mean Square Deviation (RMSD) after structural superposition was no more than 1Å, only one of them was retained (the one with the better resolution).

For the CDR length correlation study it was necessary to obtain the paired heavy and light chain sequences together with their CDR annotations so as to count the number of times certain CDR lengths were observed in the same molecule. For this purpose the sequence (A1) and the CDRs from the structural dataset (A3) were used. IMGT CDRs were annotated in the IMGT 3D DB (Ehrenmann et al. [2010]). Kabatman sequences

were mapped back to their heavy and light chains using the online interface (Martin [1996]).

The structural dataset A3 was used for a binding site analysis study. All structures which contained an antibody-antigen complex were extracted. This set was then filtered to contain only those structures with complete CDRs. No two antigens were allowed to have sequence identity higher than 90% as calculated by CD-HIT (Li and Godzik [2006]). The corresponding cut-off sequence identity for the antibodies was 99%. A final filter requiring both the heavy and light chain of the antibody to be present reduced this set to 121 complexes.



times together	L1-4	L1-5	L2-3	L3-7	H1-4	H2-5	H2-8	H2-9	H3-5	H3-7	H3-9
L1-4	0	0	2	2	1	0	1	0	1	0	0
L1-5	0	0	2	3	3	1	2	1	1	1	1
L2-3	2	2	0	5	4	1	2	1	2	1	1
L3-7	2	3	5	0	5	1	3	1	3	1	1
H1-4	1	3	4	5	0	1	3	1	3	1	1
H2-5	0	1	1	1	1	0	0	0	0	1	0
H2-8	1	2	2	3	3	0	0	0	3	0	0
H2-9	0	1	1	1	1	0	0	0	0	0	1
H3-5	1	1	2	3	3	0	3	0	0	0	0
H3-7	0	1	1	1	1	1	0	0	0	0	0
H3-9	0	1	1	1	1	0	0	1	0	0	0

FIGURE 3.1: **A**: Example dataset for the length correlation study consisting of six antibodies with partially defined binding sites ('?' indicates an undefined CDR). Assume one calculates the correlation between H1-4 (t_1=H1, l_1=4) and H2-8 (t_2=H2, l_2=8). In this case $N(t_1, l_1) = 5$, $N(t_2, l_2) = 3$ and $T = 5$. The value of T equals five because Antibody 5 does not have the H1 and H2 loops defined.

3.2.2 CDR length independence (datasets A1 and A3)

In order to assess correlations between CDR lengths (in particular, whether certain combinations are favoured), for a selected set of molecules the observed number of times two CDR loops of a given type (denoted t_1 and t_2) have a particular combination of lengths (denoted respectively l_1 and l_2) in a single molecule was compared to the expected number. The expected number was calculated conditional upon the observed number of loops of each type and length (denoted $N(t_1, l_1)$ and $N(t_2, l_2)$), the total number of molecules (T) and under the assumption that each length for one type is equally likely to appear with each length of the other type.

Under these conditions, we let X be the number of molecules with type t_1 and length l_1 and type t_2 and length l_2. Then the range of X is from N_0 to $\min\{N(t_1, l_1), N(t_2, l_2)\}$ where $N_0 = N(t_1, l_1) + N(t_2, l_2) - T$ if $N(t_1, l_1) + N(t_2, l_2) > T$, and $N_0 = 0$ otherwise.

Then the expected value $E(X)$ of X is given by 3.1 (see Figure 3.1A for an example).

$$E(X) = \sum_{x=N_0}^{\min\{N(t_1,l_1),N(t_2,l_2)\}} xPr(x) \tag{3.1}$$

The probability of having CDRs of type t_1 length l_1, and t_2 length l_2 occurring together in x out of T antibodies in total, is given by $Pr(x)$, given in 3.2:

$$Pr(x) = \frac{\binom{T}{x}\binom{T-x}{N(t_1,l_1)-x}\binom{T-N(t_1,l_1)}{N(t_2,l_2)-x}}{\binom{T}{N(t_1,l_1)}\binom{T}{N(t_2,l_2)}} \tag{3.2}$$

For each organism and CDR definition in our datasets (A1 and A3), a χ^2 test was performed, assessing the statistical difference between the corresponding expected and observed values for combinations of hypervariable loops with defined type and length. For each two CDR types (e.g. H1 and L3), every possible length pair was treated as an entry in the contingency table divided between expected and observed (for instance number of times H1-8 was observed together with L3-7 would create one column in the

contingency table). Whenever an entry in the table was smaller than five (minimal viable figure for the χ^2 test), it would be merged with the next minimal entry. In this way, the entries with the smallest counts were merged together until they would create a column whose entries would be at least five or until it was decided that there is not enough data. The χ^2 test was performed on each contingency table created like that for the Kabatman sequence dataset comprising four organisms (human, mouse, rabbit and chicken - there were not enough bovine sequences for the χ^2 test) each with three CDR definitions (Kabat, Chothia and Abm). Out of 72 cases where there was enough data (possible to merge columns so that there are more than 5 counts in each entry), in only 16 cases the difference between expected and observed was statistically significantly different.

3.2.3 Structural analysis (dataset A3)

The structural difference between two loops of the same length is defined as the RMSD of the backbone C_α, C, N and O atoms after structural superposition using the Kabsch algorithm (Kabsch [1976]). This is different from the most recent CDR clustering study which used differences in dihedral angles (North et al. [2011]).

Distance matrices of pairwise CDR RMSDs after structural superposition were constructed and used to group loops in our structural dataset with a three-tier hierarchy: CDR type, basic length (as given by IMGT (Lefranc [2011])) and anchor-extension length (0-7 residues on either side). Histograms of distance matrices were plotted for each loop type and length, with anchor extensions 0 to 7. After examination of those histograms, we chose loops with anchor extension of three amino acids to constitute a dataset for further study. This decision was motivated by the fact that RMSD values tend to rise as more anchor residues are added, a trend which slows significantly after three anchor residues had been added. This suggested that from that point (more than three anchor residues), the high uniformity of the framework region starts playing a dominating role.

Structures were clustered so that the maximum RMSD between any two structures within a single cluster was no greater than a certain cutoff - α (e.g. α=1Å). In order to group the structures using this constraint, UPGMA (Sokal [1958]) was used. This was motivated by the fact that UPGMA creates a tree with branch heights corresponding to maximal distances between the child nodes, hence grouping the closest structures together.

UPGMA trees were calculated for all RMSD distance matrices defined by CDR type and length (augmented by anchor length 3). Given such UPGMA trees, the clusters were computed by cutting the trees at nodes closest to the root whose maximal distance between children in the subtree does not exceed the boundary α. For example, a constraint of $\alpha = 1$Å means that the maximal distance between members of one group is 1Å.

A mapping from our clustering to the published clustering results from other groups was created with varying degrees of the α constraints (Table 3.2). It appears that there is a degree of correspondence if clusters are allowed maximal distances in the order of $\alpha = 1.5$Å.

Type	length	North *et al.*	Chothia *et al.*	M & T	α =1.6	α =1.5	α =1.4	α =1.3	α =1.2	α =1.1	α =1.0
L1	5	2	1	1	1(0)	1(0)	1(1)	1(1)	1(1)	1(2)	1(2)
L1	6	3	2	2	1(1)	1(1)	1(1)	1(1)	1(3)	1(3)	1(3)
L1	11	1	1	3	1(7)	1(7)	1(7)	1(8)	2(12)	3(12)	3(13)
L1	12	1	1	1	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)
L2	3	5	1	1	2(5)	2(5)	2(5)	1(7)	1(8)	1(8)	1(9)
L3	8	3	1	2	1(1)	1(1)	1(1)	1(1)	1(1)	1(2)	1(3)
L3	9	6	3	6	2(5)	3(5)	3(5)	4(10)	4(15)	4(15)	4(25)
H2	7	3	1	1	1(2)	1(4)	1(4)	1(8)	1(9)	1(11)	3(13)
H2	8	9	2	5	3(24)	3(24)	3(24)	3(24)	4(34)	4(34)	6(42)
H2	10	1	0	0	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	2(1)
H1	8	12	1	4	5(37)	5(47)	5(51)	5(56)	5(58)	6(62)	7(69)

TABLE 3.2: Comparison of the CDR clustering carried out in this Chapter with previous such groupings (mapping to previous clustering was done through the work of North et al.North et al. [2011]). Clustering is given for different values of maximum distance within a cluster (α). Value in brackets gives the number of singleton clusters and the number outside of the bracket is the number of non-singleton clusters. Note that the correspondence with other clusterings (when disregarding singletons) happens tentatively at at the level of $\alpha = 1.5$Å. Also note that reducing the maximal distance within the cluster results in the most dramatic rise of singletons, hinting at individual rather than shared divergences. This mapping was created in 2010 - the corresponding up-to-date mapping is available through SAbDab (Dunbar et al. [2013b]).

3.2.4 Contact frequencies of CDR loops (dataset A3)

For each CDR in our complete Ab-Ag complex structural dataset (121 structures with both heavy and light chains defined), the number of times a loop was in contact with an antigen was recorded (at least one CDR residue within 4.5Å of an antigen residue, and an accessible surface change of 5%). Frequency was taken as the ratio over only those structures where the given CDR was defined.

In a similar fashion, the combinations of the CDRs in contact were compiled. For each structure that had all six hypervariable loops defined, the number of times combinations of CDRs were in contact with the antigen at the same time were counted (e.g. how many times L2 was not in contact when all other CDRs were in contact).

3.2.5 Binding propensities of residues (dataset A3)

The binding-propensity background data together with calculation methodology was adapted from i-Patch (Hamer et al. [2010]). Using this method, propensity to be in contact was calculated for each surface exposed residue in our set of antibody structures. For a given amino acid a, its propensity to be in a contact site, p_a, is given as the ratio $p_a = p_a^{\text{con}}/p_a^{\text{non}}$ as defined in 3.3 and 3.4.

$$p_a^{\text{con}} = \frac{f_a^{\text{con}}/f_a^{\text{all}}}{\sum_b f_b^{\text{con}}/\sum_b f_b^{\text{all}}} \tag{3.3}$$

$$p_a^{\text{non}} = \frac{f_a^{\text{non}}/f_a^{\text{all}}}{\sum_b f_b^{\text{non}}/\sum_b f_b^{\text{all}}} \tag{3.4}$$

where f_a^{con} is the frequency of amino acid a on the surface of proteins at contact sites and f_a^{non} is the frequency of amino acid a on the surface of proteins at non-contact sites. The frequency of amino acid a on the surface of proteins at all sites is given by f_a^{all}.

3.3 Results

In this Chapter, several aspects of the antibody binding site are studied with the main aim being to establish which residues are preferred for antibody binding. Here we sumarize the results of this analysis.

Firstly, we demonstrate that there exist CDR compositional similarities across species and that they appear to be maintained as the organism matures. Secondly, it is concluded that CDR lengths are not inherently correlated, meaning that there is no underlying tendency to observe particular combinations of CDR lengths. Thirdly, there also appears to be no significant tendency for hypervariable loops to change their canonical conformations upon contact with an antigen. Fourthly, it is demonstrated that all CDR loops are involved in binding. This suggests that though H3 is of primary importance, other CDR loops play strong auxiliary roles. Finally, the binding propensities of amino acids in antibodies and other proteins were contrasted. It is shown that the amino acids used for binding by immunoglobulins are distinct from those employed by other proteins. In particular, it is demonstrated that tyrosine, tryptophan and histidine are strongly preferred for establishing contacts in antibody-antigen complexes.

3.3.1 CDR composition differences

CDR compositions across different organisms were compared and their relationship to general anti-parallel $\beta - \beta$ loops was established.

3.3.1.1 CDR composition difference with respect to general protein loops

Compositions of CDRs are known to be statistically significantly different from general protein loops (Collis et al. [2003]) (this includes $\alpha - \alpha$, $\beta - \beta$ and $\alpha - \beta$ loops). Here,

we establish that an even stronger statement is true, namely that CDRs, always anti-parallel $\beta - \beta$ loops, differ significantly from general protein loops with this particular configuration for a range of organisms.

We have compared the distributions of amino acids for each organism and CDR type in dataset A1 with those from non-antibody anti-parallel $\beta - \beta$ loops from the PDB. Pearson's χ^2 test was used to determine whether CDR amino acid compositions from different organisms come from the same distribution and whether they are distinct from all $\beta - \beta$ antiparallel loops. Contingency tables for the test were created by compiling amino acid counts from our sequence dataset for each CDR type, organism and definition. Additional contingency tables were also created where each group was further divided according to the loop length. Pairwise χ^2 tests between each CDR oragnism compositions and the background composition rejected the null hypothesis that they came from the same distribution. This result indicates that antibodies have statistically significantly different amino acid composition distribution from non-antibody anti-parallel $\beta - \beta$ loops.

3.3.1.2 CDR composition correspondences between different species

Some residues, notably tyrosine, serine and glycine have previously been observed to be more common in the CDRs of human, mouse and axolotl (Zemlin et al. [2003], Golub et al. [1997], Abergel and Claverie [1991]). In the light of this consistent over-representation of tyrosine and serine in CDRs, we have checked if the result holds across a wider range of organisms as well, namely: human, mouse, chicken, bovine (cow) and rabbit.

For each species, absolute frequencies of amino acids in CDRs across all types were clustered into three groups using the K-means algorithm. We refer to those groups as *low*, *med* and *high*, with obvious connotations. Motivation for this approach was to see if the average composition of the entire binding site of an antibody bears any resemblance from one species to another and what is its relation to the background composition

of all other non-antibody anti-parallel $\beta - \beta$ loops. Final results of this clustering for each species are presented in Figure 3.2. These results demonstrate that the general trends of over-representation of certain residues like serine and tyrosine appear to be well conserved across species, regardless of the CDR definition used. Serine and glycine appear to be consistently over-represented, being classified in the high frequency cluster in each case. Tyrosine did not have quite as high absolute frequencies as serine and glycine and thus in most cases it appears in the medium frequency cluster.

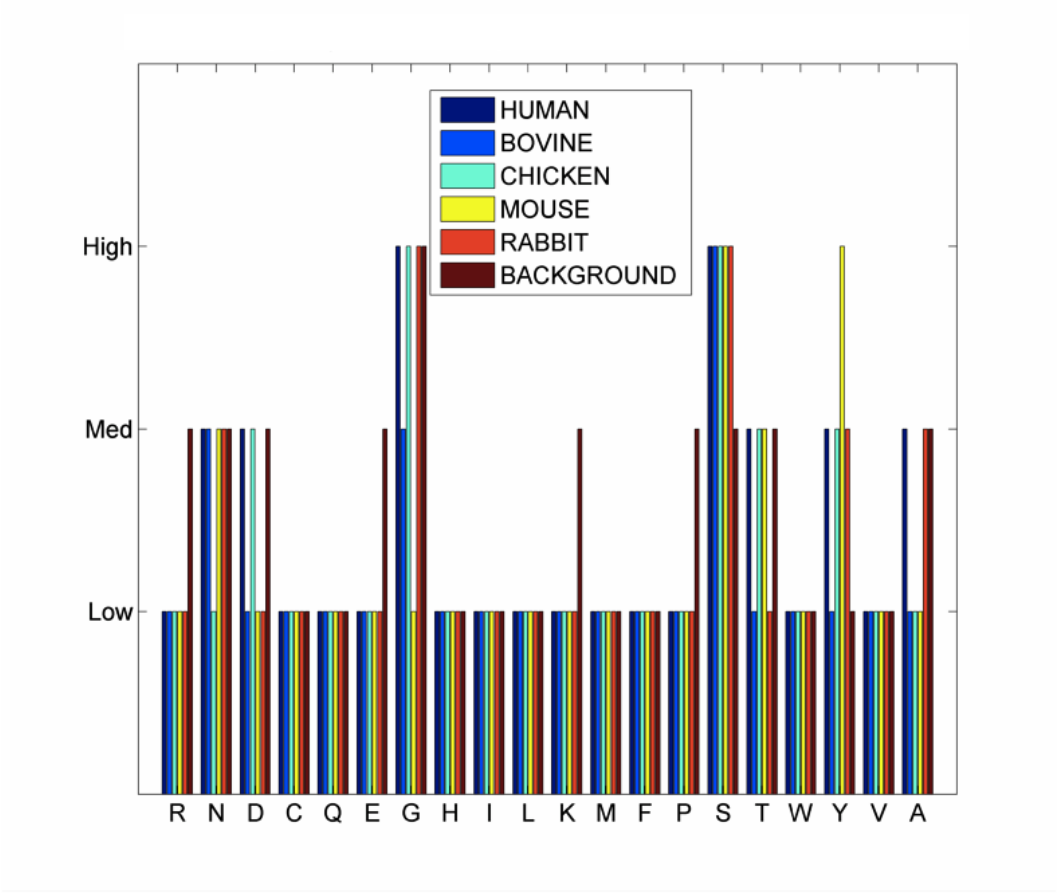


FIGURE 3.2: Clustering of absolute frequencies of amino acids of all CDRs for five species. Definition used here was Chothia. Background data indicates the distribution for the non-antibody anti-parallel $\beta - \beta$ loops. Clusterings using different CDR definitions demonstrated similar trends

Based on the frequency clustering between organisms we have manually developed the consensus clustering given in Table 3.3. Tyrosine had a non-trivially higher frequency than other residues in the middle clusters and thus we assigned it to its own cluster. The consensus clustering reflects the distinction between the general protein and antibody loops.

Frequency	Antibody CDRs	Background
High	**S**,G	G
High/Medium	**Y**	-
Medium	N,D,T	R,N,D,T,E,K,P,S,A
Medium/Low	A	-
Low	Rest	Rest

TABLE 3.3: The consensus clustering of relative frequencies of antibody residues was developed using single species clusterings presented in Figure 3.2. Residues in bold are those that appear to be of special interest for antibodies because their high concentrations in CDRs are consistently maintained across different species, which is not the case for non-antibody loops.

The over-representation of tyrosine, serine and glycine in the consensus grouping is in agreement with previous studies (Zemlin et al. [2003], Golub et al. [1997], Abergel and Claverie [1991]). Of these three, only tyrosine and serine appear to bear intrinsic significance for antibodies as glycine is usually highly represented in all antiparallel $\beta-\beta$ loops, as indicated by our background data. This might provide a further (Fellouse et al. [2005], Schildbach et al. [1993]) indication of the special roles of these residues in immunoglobulins.

3.3.1.3 CDR composition over time

After investigating the degree of commonality in CDR composition across species in the form of over-representation of serine and tyrosine, we also analysed the persistence of such similarities over an organism's lifetime. In order to do this, the sequence composition of the zebrafish D region (a highly variable region, usually part of H3) were investigated, taken at five time points during the animal's life.

Differences between the compositions of D regions from the zebrafish sequence dataset, for any two ages, were examined using the χ^2 test. Contingency tables for the test were created by comparing the counts of amino acids in the extracted D regions across all lengths.

The χ^2 tests rejected the null hypothesis that CDR composition is sampled from the same distribution at every age. In this case, the χ^2 test is capable of detecting minute

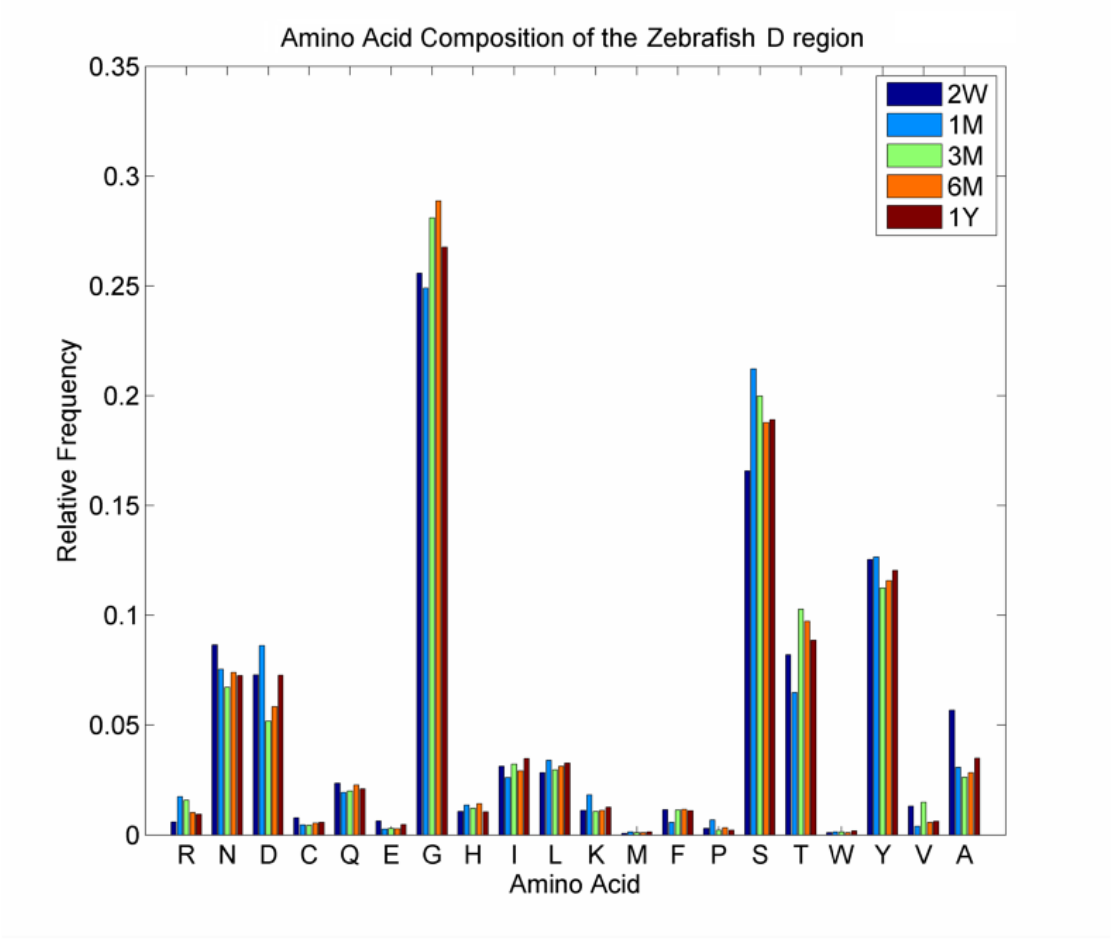


FIGURE 3.3: Distribution of relative frequencies of amino acid usage in zebrafish antibodies over time taken at an age of two weeks (2W), one month (1M), two months (2M), three months (3M), six months (6M) and one year (1Y). Note that the compositional discrepancies over time are not large, even though they are statistically significantly different according to the χ^2 test.

dissimilarities because of the large number of observations available for each amino acid, despite the evident similarity of relative frequencies of amino acids for all ages, shown in Figure 3.3.

It was previously noted that tyrosine and glycine were over-represented over the life of mexican axolotl (Golub et al. [1997]). Our results from zebrafish appear to strengthen this point by demonstrating a similar tendency. Though the composition distribution of the D region changes with time, serine, tyrosine and glycine appear to be over-represented throughout an animal's life. These results from two different organisms strengthen the argument that tyrosine and serine play a special role in CDRs.

3.3.2 CDR Length Independence

We next investigated the correlation between CDR lengths. The aim was to establish whether there is a tendency for certain CDR length to be observed together in a single antibody. Since the CDR lengths are intimately linked to canonical classes, the argument presented here applies to canonical classes as well.

The number of times a pair of CDRs of a given type and length were observed together in a single antibody molecule was compared with the expected number, assuming random association. The observed and expected values appear to be in good correspondence, corroborated by the χ^2 test which found no statistically significant difference between the two. This result suggests that CDRs associate randomly and there is no correlation between their lengths.

This conclusion does not agree with previous results (Lara-Ochoa et al. [1996]) which stated that combinations of canonical classes (which are essentially dictated by length) of CDRs are very restricted. However, certain CDR lengths make up a considerable proportion of available CDR sequences, restricting the number of observed combinations. Authors of the previous publication suggested a model introducing evolutionary pressure on certain CDR shapes, in an effort to explain the small number of observed canonical class combinations. Our results suggest that even though the high frequencies of some hypervariable loop lengths might be a feature of antibodies, there appears to be no inherent mechanism causing certain length combinations to be favored.

3.3.3 Structural Features

Structural analysis of the CDRs was carried out, aimed at elucidating the extent to which hypervariable loops diverge from canonical classes upon binding. For this purpose, a clustering of CDRs was created. This initial clustering served as the starting point for the clustering that is now available as a part of the Structural Antibody Database described in the previous Chapter.

It was observed that non-H3 loops of a given length tend to form one large structural cluster with smaller clusters and numerous singleton outliers. This property has been maintained over the last three years as demonstrated by the latest version of the clustering available in SAbDab. This uniformity for each length-type speaks in favour of length being a strong factor in determining the set of possible CDR conformations.

Given that non-H3 CDRs tend to form one big cluster and many smaller ones, it was expected that the singleton sets would chiefly constitute structures in complex. This hypothesis was motivated by the assumption that outliers would have changed conformation from that of their canonical classes.

In order to check if this hypothesis was true, the RMSD between each CDR and its closest structural neighbour was plotted against the mean RMSD of all other CDRs of the same type and length (see Figure 3.4). This was designed to reveal the putative outliers, since they should have a larger distance to the closest different structure and be divergent from the majority of CDRs of the same type and length which are, generally, structurally uniform. The plots appear not to separate out the complex and non-complex loops.

Finally, structures of bound and unbound non-H3 CDRs with the same sequence according to the IMGT definition were compared. CDR-H3 was excluded from this study as it has previously been shown to change its shape upon binding and it does not form canonical classes in the same manner as the remaining CDRs (Shirai et al. [1996], Oliva et al. [1998]). For each subset of loops with identical sequences and RMSD within 0.5Å from each other, a representative was selected so as not to bias the dataset (retaining the one with the best resolution). Results from this procedure for each bound dataset for non-H3 loops are presented in Figure 3.5.

In the majority of cases, the RMSD difference between bound and unbound versions of a single CDR were below 1Å. This suggests that binding does not significantly influence the canonical structure of non-H3 CDRs.

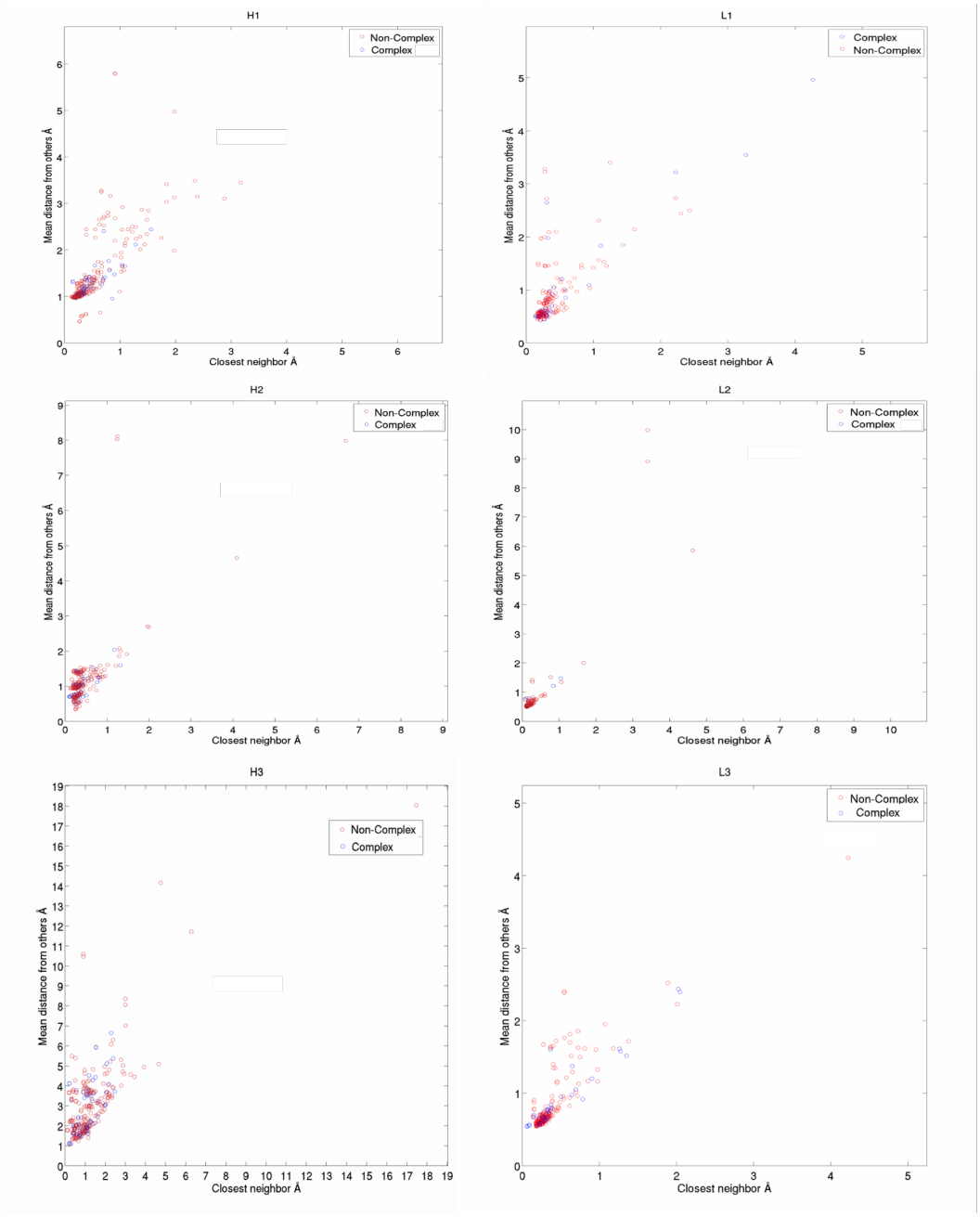


FIGURE 3.4: For each CDR of a given length and type, the RMSD distance to the structurally closest loop of the same length was plotted against the mean distance to all other loops of the same length.

We demonstrated that with the exception of H3, CDR loops are structurally similar to each other for a given length and type. Moreover, exceptions did not appear to be caused by Ab-Ag complex formation, suggesting that there is little tendency for non-H3 CDRs to deviate from their canonical conformations upon binding.

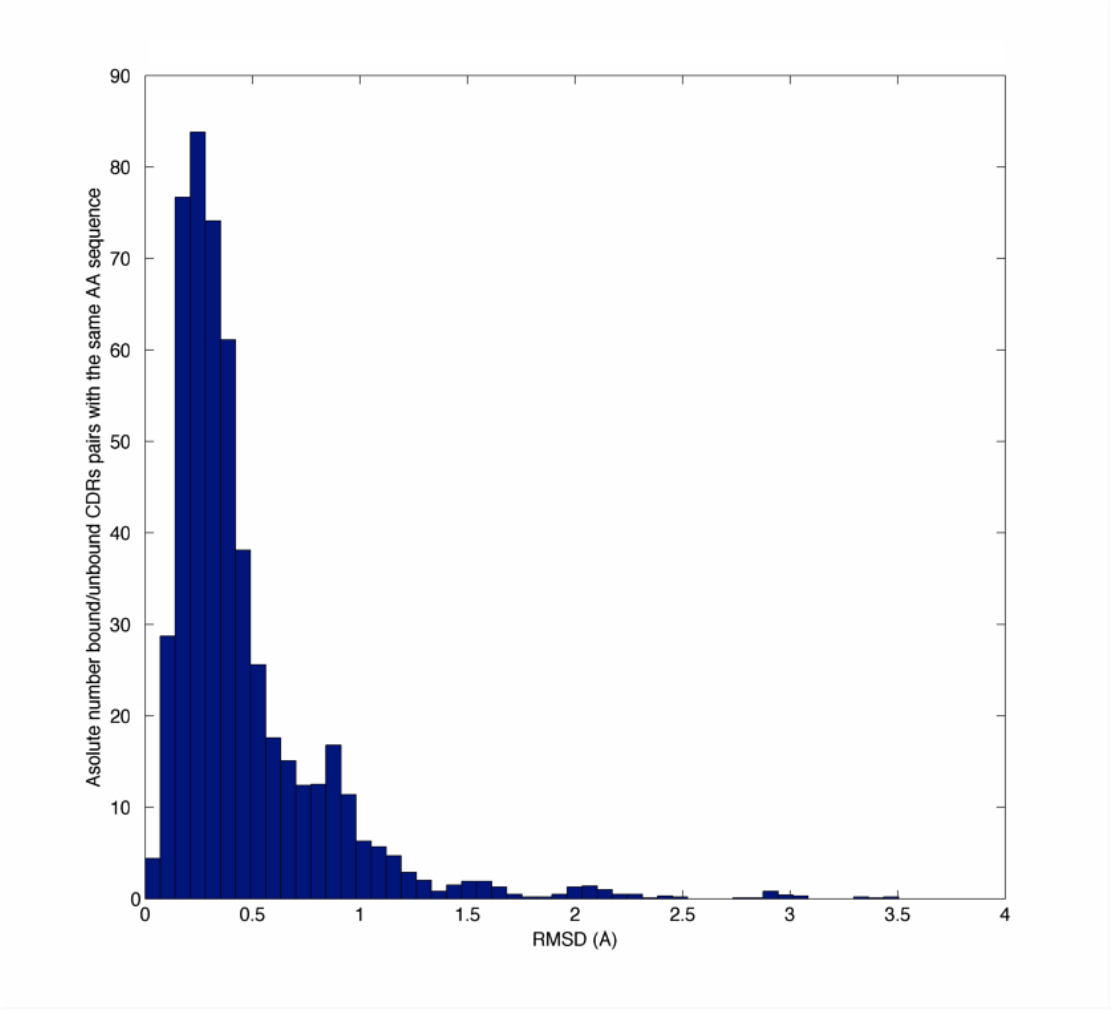


FIGURE 3.5: RMSD differences between bound and unbound versions of the same non-H3 CDR. Note that majority of the loops remains largely unchanged upon binding.

Frequency	H1	H2	H3	L1	L2	L3
77	1	1	1	1	1	1
20	1	1	1	1	0	1
5	1	1	1	0	0	1
5	0	1	1	1	0	1
3	0	1	1	0	0	1

TABLE 3.4: The top five observed combinations of CDRs in contact with an antigen, accounting for majority of cases in our complex structural dataset. The CDR combinations are indicated in binary fashion with 1 for presence and 0 for absence. Note that the most frequent configuration is having all the CDRs touching the antigen, representing well over the half of the dataset (77 out of 121 structures in the reduced dataset A3). The three most persistent binders appear to be H2, H3 and L3 and the easiest one to shed is L2.

3.3.4 CDR involvement in binding

Using the 121 identified distinct complete (having both light and heavy chain) structural Ab-Ag complexes we investigated the role of different CDRs in antigen binding. This was stimulated by earlier research suggesting that H3 is usually sufficient to attain complex formation with the antigen (Xu and Davis [2000]). The number of times each CDR was in contact with the antigen was calculated, and the top configurations are presented in Table 3.4 (a CDR was defined to be in contact if at least one residue on the CDR was within 4.5Å of the antigen and its surface accessible area changed by at least 5% with respect to the unbound state).

The configuration where all CDRs are bound to the antigen is the most common and it accounts for more than half (77 out of 121) of the structures. These results indicate that all CDRs are heavily involved in binding, suggesting that all hypervariable loops play a role in antigen complex formation.

Nevertheless, we noted instances where not all CDRs are in contact with the antigen, including cases where CDR-H3 does not contact the bound antigen. However, CDR-H3 was the loop most frequently in contact with the antigen. Table 3.4 suggests that the loops which have the strongest tendency to be in contact are H2, H3 and L3. Of those three only the role of H3 has been widely commented on in the literature (Xu and Davis [2000]). On the other hand, L2 is least frequently involved in antigen contact, which might be attributed to its size - L2 is generally the shortest CDR so it might not always reach the antigen.

In summary, all the CDRs are heavily involved in antigen binding. This conclusion does not diminish the significance of H3 which still appears to be the most important CDR for Ab-Ag complex formation, but rather suggests that the role of other CDRs is not unimportant.

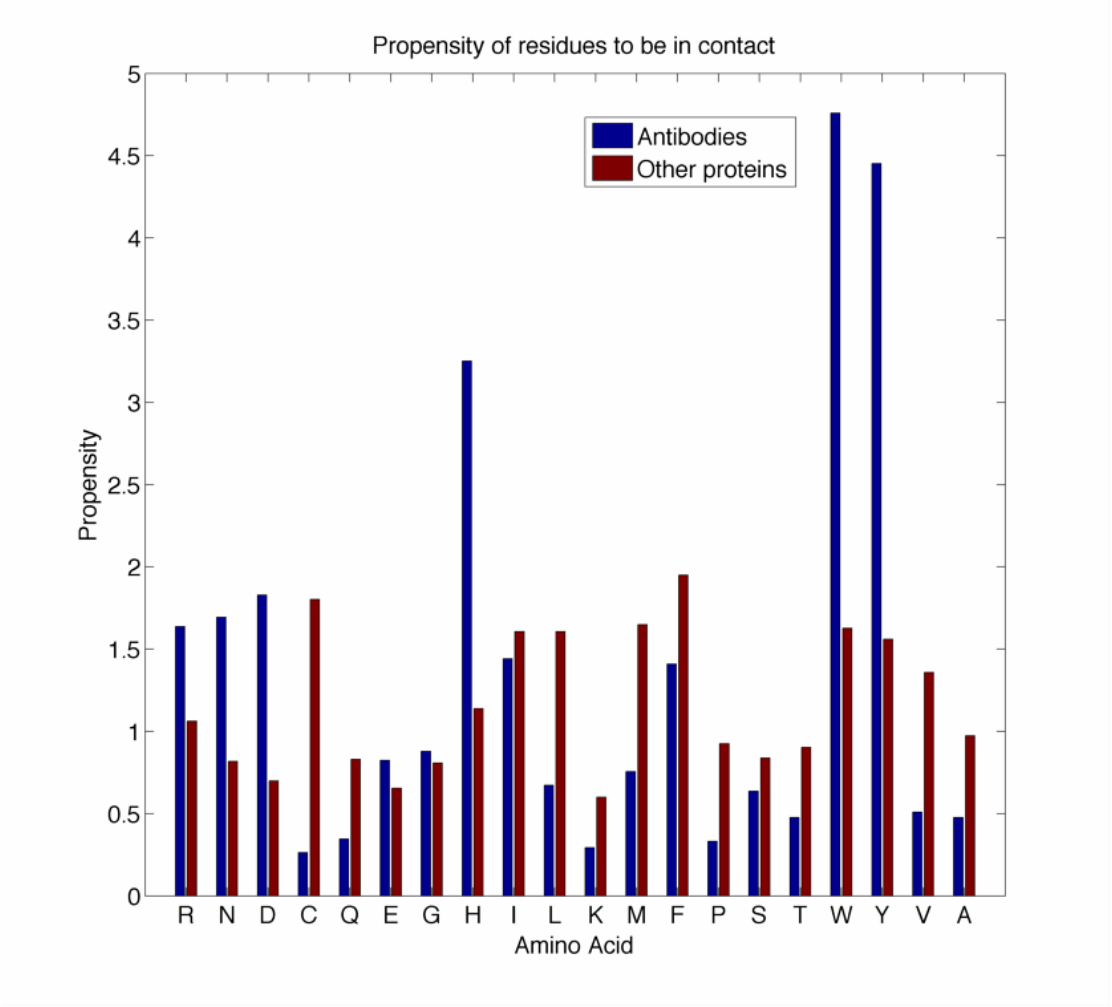


FIGURE 3.6: Comparison of binding propensities of individual residues in antibodies and in other proteins. Tyrosine (Y), tryptophan (W) and histidine (H) have much higher binding propensities in antibodies than in other proteins.

3.3.5 Residue binding propensity

The residues used for binding by antibodies were compared with those used by other proteins (Figure 3.6) so as to shed light on the main aspects of the antibody binding mechanism.

A comparison of the binding propensities of individual residues in antibodies with those in other proteins reveals that the former have a strong preference for tyrosine, tryptophan and histidine in contact sites (Figure 3.6). The high binding propensities of tyrosine and tryptophan in antibodies are in agreement with earlier theoretical (Mian et al. [1991], Wilson and Stanfield [1993]) and experimental (Fellouse et al. [2005], Birtalan et al.

[2008]) findings. In a previous study of somatic hypermutations, it was noted that the frequencies of tyrosine and tryptophan tend to fall during the affinity maturation process and the frequencies of proline, phenylalanine and histidine tend to rise (Clark et al. [2006]). Our results indicate that only histidine has a significantly higher binding propensity in antibodies as both proline and phenylalanine have lower binding propensity in antibodies as opposed to general proteins.

As indicated by our earlier composition study and in Figures 3.7 and 3.8, tryptophan and histidine, in contrast to tyrosine, do not have high relative frequencies in antibody binding sites, suggesting that they might play an auxiliary role during the affinity maturation process.

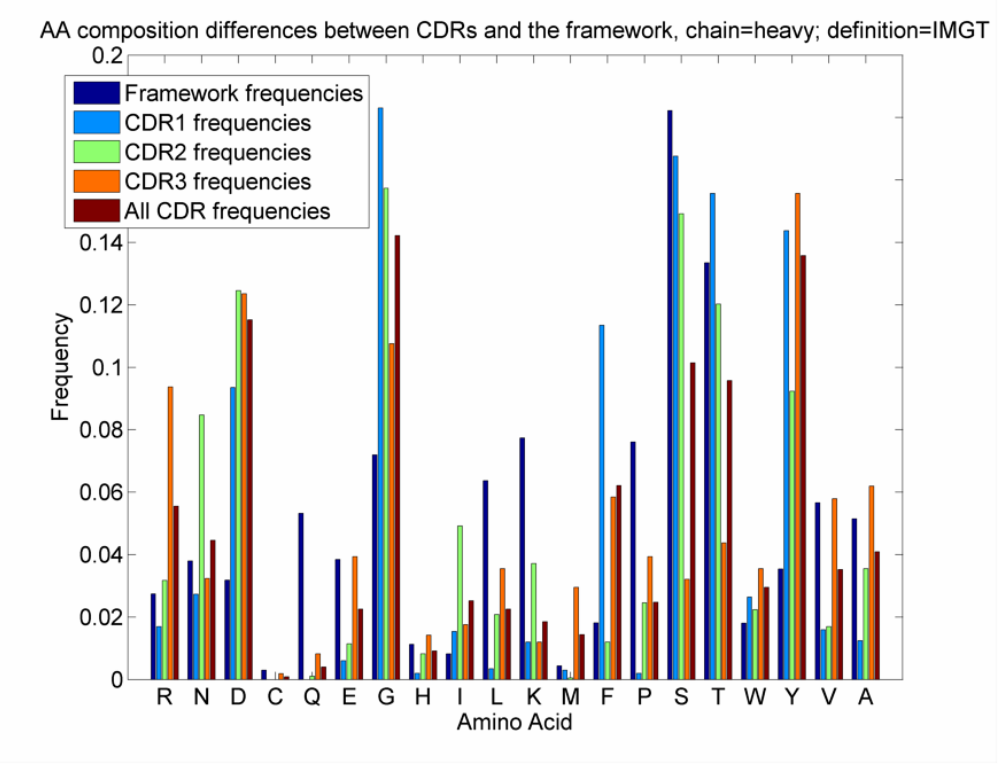


FIGURE 3.7: Residue relative frequency comparison between the framework region (excluding CDRs) and the individual CDR regions of the heavy chain according to the IMGT definition. The data from the figure is the structural dataset A3.

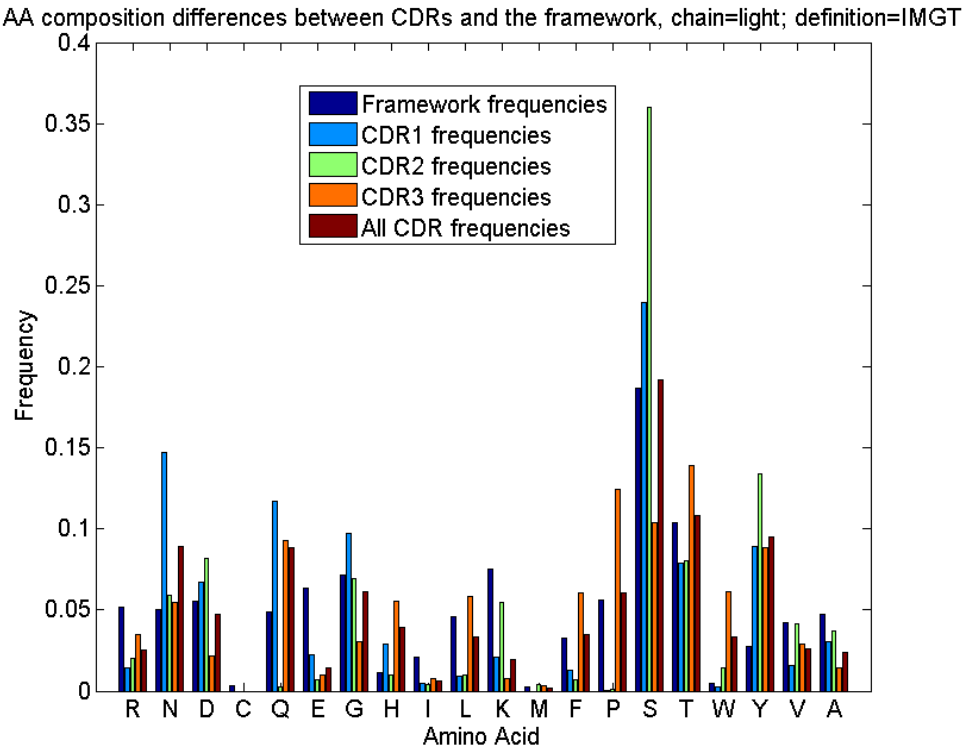


FIGURE 3.8: Residue relative frequency comparison between the framework region (excluding CDRs) and the individual CDR regions of the heavy chain according to the IMGT definition. The data from the figure is the structural dataset A3.

The high binding propensity of tyrosine in antibodies, coupled with its over-representation in CDRs (Figures 3.7 and 3.8), further suggests its pronounced role in establishing contacts with the antigen. This conclusion is strengthened by contrasting serine with tyrosine. Serine is also over-represented in the CDRs but its binding propensity in antibodies is comparable with that in general proteins. However, serine does not appear to have a significantly higher relative frequency in CDRs as contrasted to the framework region as indicated in Figures 3.7 and 3.8. The sharp contrast between these two amino acids might follow from earlier suggestions that tyrosine could be responsible for binding while serine plays an auxiliary role (Fellouse et al. [2005]).

Chemically similar tyrosine and tryptophan have been described as residues with properties favourable for facilitating Ab-Ag complex formation (Mian et al. [1991]). It was suggested that aromaticity permits both residues to participate in a variety of interactions, whereas their large size facilitates engagement of more binding partners. Histidine is also a bulky group with aromatic nature, suggesting that these properties are of special

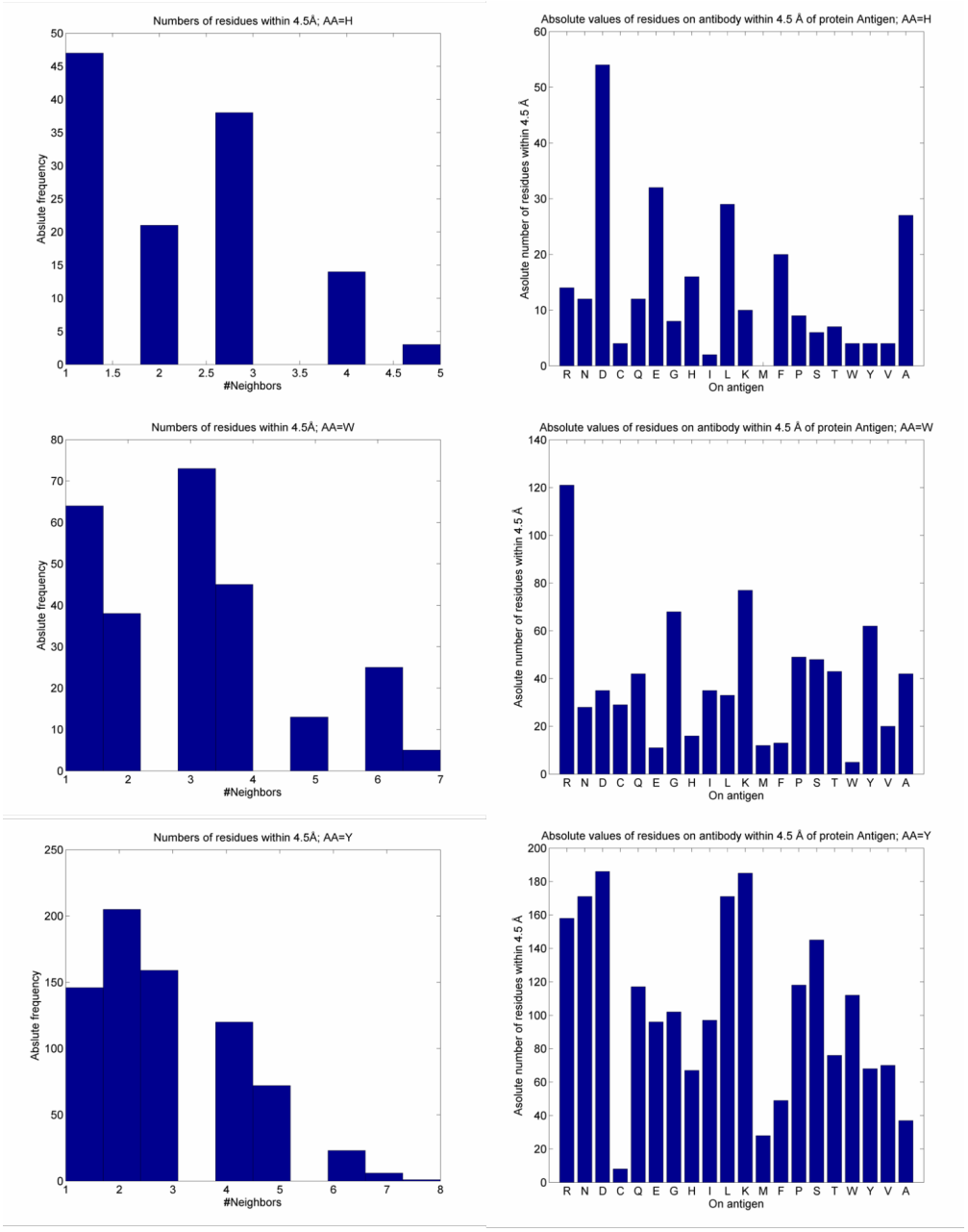


FIGURE 3.9: Absolute numbers of neighbours in contact with **Top** Histidine (H) **Middle** Tryptophan (W) **Bottom** Tyrosine (Y), on the structural dataset A3. In the left column, the total numbers of neighbouring residues on the antigen are presented. In the right column the absolute numbers of types of neighbouring amino acids are given.

importance in Ab-Ag complex formation, since all three residues have higher binding propensities in antibodies.

In order to investigate what type of interactions might be important for antibody binding, we have studied the binding partners of our three highest scoring residues: tyrosine, tryptophan and histidine. Firstly, it was investigated whether the size of these three

amino acids allows them to engage multiple binding partners on the antigen simultaneously. This was checked by plotting the numbers of neighbouring residues on the antigens for each of tyrosine, histidine and tryptophan on the antibody (see Figure 3.9). The results demonstrate that any of tyrosine, tryptophan on CDRs are usually within 4.5Å of more than one residue on the antigen.

Investigation of binding partners for each of tyrosine, tryptophan and histidine on CDRs revealed multiple contacts of the three residues with charged amino acids of the antigen (Figure 3.9). From Figure 3.9 it appears that histidine on the antibody is often paired up with aspartate on the antigen, antibody tryptophans with antigen arginines and antibody tyrosines with antigen aspartates, arginines and lysines. This propensity towards charged residues is corroborated by the fact that two other charged residues (aspartate and arginine, see Figure 3.6), appear to have higher binding propensity in antibodies than in other proteins. The significance of electrostatic interactions for antibodies was previously suggested by Lippow et al. (Lippow et al. [2007]). The authors remarked that the electrostatic components of their energy function were found to be among the best approximators of binding energy change between the antibody and antigen. The electrostatic complementarity also appears to be the driving force behind antibody-antigen specificity (Sinha et al. [2002], Mohan et al. [2003]).

In summary, we demonstrate that tyrosine, tryptophan and histidine have higher binding propensity in antibodies in comparison to that in other proteins. This is in accord with earlier results and conjectures (Fellouse et al. [2005], Mian et al. [1991], Birtalan et al. [2008]). Tryptophan, tyrosine and histidine are all big and aromatic, suggesting an important role for these properties in antigen binding. Moreover, the residues on the antigen which appear to be favoured for binding tend to be charged ones, indicating that electrostatics also plays an important role in antibody-antigen interactions.

3.4 Conclusion

In this Chapter, we present work which was aiming to elucidate the antibody-antigen binding mechanism by exploring several aspects that distinguish it from protein-protein interactions in general.

We were able to confirm that CDR sequence compositions are distinct from non-antibody $\beta - \beta$ antiparallel loops and that they differ between species. However, the over-representation of tyrosine and serine appears to be consistent across species. Moreover, our zebrafish antibody study confirmed that serine and tyrosine maintain their high relative frequencies in CDRs throughout the animal's life.

Contrary to previous results, we found that the CDR lengths associate randomly and that the small number of their observed combinations most likely stems from length preferences in the datasets. Such length biases might be an inherent property of antibodies, potentially constraining the binding site, as previously suggested (Lara-Ochoa et al. [1996]).

Analysis of the structural clustering has shown that there is no clear tendency for CDRs to deviate from their preferred canonical conformations upon binding. We demonstrated that all hypervariable loops are heavily involved in binding, with H3 playing the most significant role, followed by L3 and H2.

We show that antibodies use tyrosine, tryptophan and histidine as binding residues and that these residues are not preferentially used in other proteins. These three amino acids favour contact with charged residues on the antigen, suggesting a significant role for electrostatic interactions in antibody-antigen complex formation. Moreover, lower relative frequencies of histidine and tryptophan in the CDR region, as opposed to tyrosine, suggest that the first two might play a fine-tuning role during the affinity-maturation process, with the last allowing for low-affinity initial binding.

The analysis carried out in this Chapter provides a far from complete description of the mechanism of antibody binding. Nevertheless, by identifying aspects of the antibody

binding site which are distinct from other proteins we offer avenues for improved artificial antibody design. In particular, information obtained in this Chapter informed our antibody-related tools which we develop in the following Chapters.

Chapter 4

CDR contact prediction

4.1 Introduction

The antibody-antigen complex analysis carried out in the previous Chapter provided a solid basis for the development of computational antibody design tools. Specifically, we exploit the fact that the binding sites of antibodies are radically different from those of general proteins. We used this information to adjust a general-protein binding site prediction tool for use on antibody-antigen complexes in particular. The work presented in this Chapter was published in (Krawczyk et al. [2013]).

4.1.1 Motivation

Identification of the antibody-antigen contact residues is particularly important because only a small number of mutations to the antibody-antigen binding site can lead to a radical change in specificity and affinity towards an antigen (Murad et al. [2012a]). The antibody binding site is composed of six hypervariable loops known as the Complementarity Determining Regions (CDRs) (Wu and Kabat [1972]). Since the framework region

(FR) on which the CDRs are situated is relatively conserved, the general antibody binding site is known *a priori*. Therefore identifying the residues in contact with the antigen generally involves the analysis of CDRs.

The CDR definition methods can be regarded as a form of an antibody binding site predictor - since given a sequence of an unknown antibody they annotate it with the region which is believed to contain the majority of contacts. This was challenged by Kunik et al in 2012 as they have shown that in fact the regions returned by the four CDR definitions contain only about 80% of all the contact residues. They have also shown that the remaining 20% residues that fall outside of the traditional CDRs are just as energetically important.

Based on these findings, Kunik et al. proposed a CDR contact annotation method, Paratome (Kunik et al. [2012]). Given a sequence or structure of an antibody, Paratome annotates the region where the binding site residues are, by comparing the input sequence or structure to experimentally determined antibody structures with annotated contact sites. Using the definition of binding site residues as all antibody residues within 4.5Å of the bound antigen they achieve 31% precision at 96% recall.

In comparison, precision of current CDR annotation methods all lie in the region of 30% with recall reaching around 80% (Kunik et al. [2012]). The aim of those methods was to maximize recall. Given that minor mutations to the binding site might lead to significant changes to the specificity and affinity profile of an antibody, knowing fewer binding residues but with a higher precision might be beneficial for guiding mutations in antibody engineering (Raghunathan et al. [2012]).

In this Chapter, we develop Antibody i-Patch, a method which predicts antibody contact residues. In contrast to Paratome and the CDR definition methods which indicate the extent of the general binding region, Antibody i-Patch assigns a contact likelihood score to each residue, allowing the user to choose a cutoff so as to achieve higher precision or better coverage (see Figure 4.1 for an example). By doing so, one can differentiate between higher and lower confidence predictions which might provide a better guide for

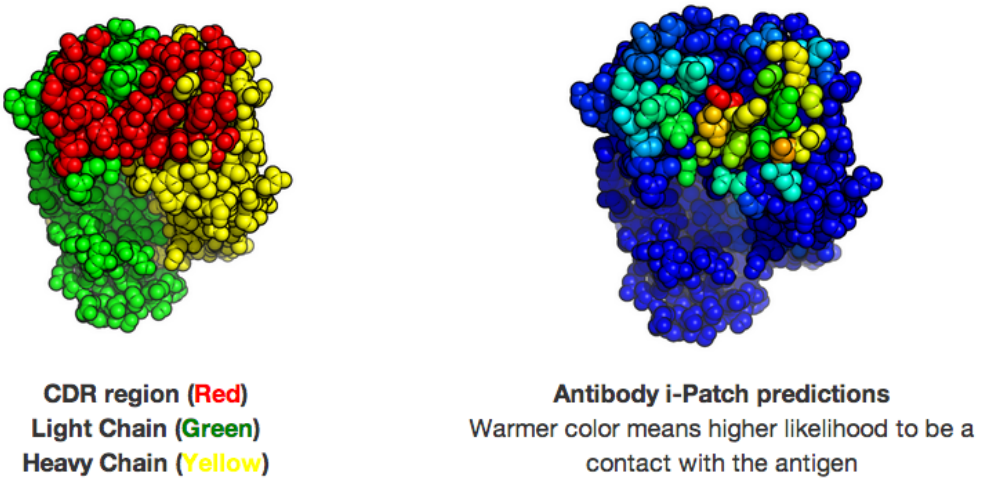


FIGURE 4.1: Example of Antibody i-Patch annotations for antibody 1AHW.

introducing mutations to the CDR region. Using an *in silico* alanine scanning protocol, we show that residues with a higher score are more important energetically. In the next Chapter, we show the applicability of Antibody i-Patch by using the predicted contact residues as constraints for local antibody-antigen docking.

4.2 Materials and Methods

4.2.1 Data

4.2.1.1 Dataset NR-full

NR-full is a non-redundant set of antibody-protein complexes extracted from SAbDab (Dunbar et al. [2013b]). Only antibodies with both VL and VH present and of resolutions 3Å or better were selected. Structures are reduced to a non-redundant set using CDHIT (Li and Godzik [2006]). In NR-full, no AB sequence was more than 99% identical and no antigen more than 90% identical. NR-full contains 148 non-redundant antibody-antigen complex structures (full list in Appendix A.1).

Comparisons to other methods (e.g. Paratome (Kunik et al. [2012])) were carried out on their datasets. A few of the structures could not be used in our analysis due to inconsistencies. For instance, 3O30 and 2E58 and 2CXD contain no antibody. Several antibody antigen chain pairings in the Paratome dataset were too far away from each other (even if we consider a minimal distance in the order of 11 Å), rendering them unusable for predicting binding sites. In most cases the distance error stemmed from the fact that there was more than one antibody antigen complex in the structure and only the pairing was incorrect. Wherever possible such pairings were amended. There was also one case of wrong assignment of light and heavy chains. A complete list of our version of the Paratome dataset is given in A.2.

4.2.1.2 Dataset RA (RosettaAntibody)

Out of 54 antibodies modelled in the RosettaAntibody study (Sivasubramanian et al. [2009]), 27 had an antigen present. Out of these 27, 1IQD, 1YNT, 1Z39, 2H1P and 1KB5 were removed as there were certain sequence differences between the native structure and the model. Since Antibody i-Patch is residue-type based it would not be sound to compare results from the model to the native structure in such cases. The remaining 22 models constituted the RA test set and are presented in A.3.

There were two variants of the dataset RA - the homology models of the 22 structures (RA-h) and the 22 corresponding crystal structures (RA-x). The available RosettaAntibody (Sivasubramanian et al. [2009]) homology models of the antibodies did not contain coordinates for CDR-H3. These loops were modelled using FREAD (Choi and Deane [2011, 2010]). Since FREAD is a database search method, in some instances the best structure returned came from the query structure itself. In such cases this result was removed and the second best structure was used.

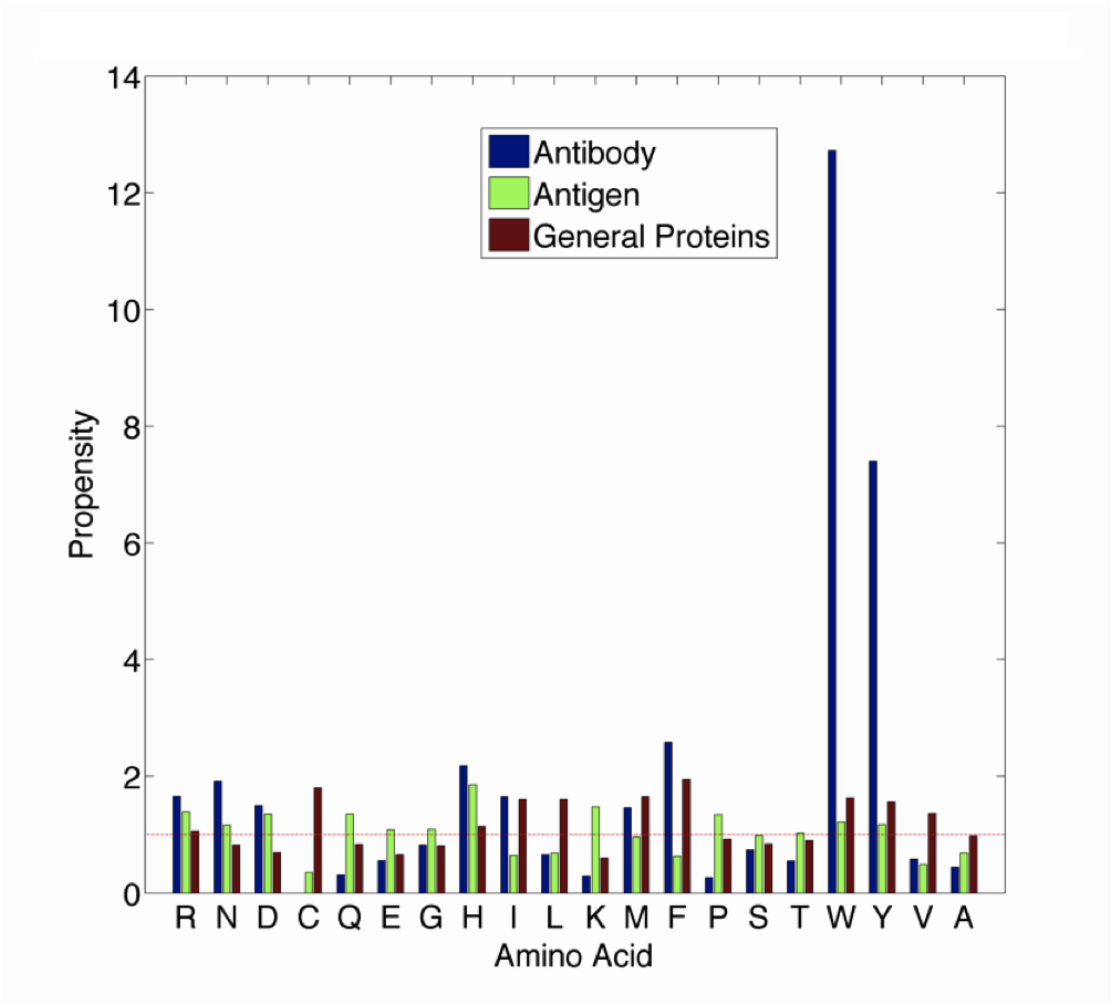


FIGURE 4.2: Binding propensity differences between antibody, antigen and general proteins. Binding propensities of each residue type were calculated for antibodies (Ab) and antigens (Ag). Values above '1' indicate a preference to be in contact while those below '1' correspond to a preference not to be in contact (red line in the figure above). The propensities were calculated in the identical fashion to those presented in the previous Chapter. These propensities are contrasted to those of other proteins reported by the authors of i-Patch (Hamer et al. [2010]). Antibodies appear to have radically different binding preferences from antigens and general proteins.

4.2.2 Antibody i-Patch

Antibody i-Patch is derived from the original i-Patch. The binding site prediction algorithm, i-Patch requires the structures of two proteins that are assumed to interact together with the multiple sequence alignment of their homologs. Two modifications are needed to generate Antibody i-Patch.

The first is the removal of the need for a Multiple Sequence Alignment (MSA). As

antibody-antigen complexes are formed under very different processes than standard protein-protein interactions, creating corresponding MSAs is impossible. In order to establish whether i-Patch can operate without an MSA we tested i-Patch on its original protein protein binding site prediction dataset (Hamer et al. [2010]) with and without MSA. For the latter, reduced dataset, the MSAs from the original test set for the i-Patch algorithm were adjusted to consist only of the reference sequences of the input structures. In order to evaluate the statistical significance of the difference between the results of the two types of input an area under ROC curve method was employed (Fawcett [2006], Chen et al. [2008]) (see A.4). We found the results from this test are not statistically significantly different from those using full MSAs, persuading us that such a modification may be possible.

The second is an asymmetry in residue interaction potential. Using the same propensity calculation methodology as in the previous Chapter, we have computed the propensities for antibodies and antigen residues to be in contact on dataset NR-full (results are in Figure 4.2). As shown in Figure 4.2, antibodies use different residues in their binding sites from proteins in general. Thus, the i-Patch scoring system needs to become asymmetric between the antibody and antigen. Below, we delineate the procedure of converting i-Patch to Antibody i-Patch.

The aim of i-Patch is to calculate a per residue score for the surface residues of two interacting proteins, A and B. This score predicts the likelihood that a residue on protein A is involved in binding to a residue on protein B and vice versa.

The i-Patch score of a residue is based on a statistical calculation of the likelihood that two patches, one on each protein, are in contact with one another. A patch consists of the residue of interest and all other residues within 4.5Å of this residue and for a given residue i it is denoted by $\Pi(i)$. The patch is described by the profile of its residue taken from the multiple sequence alignment. Below we give a brief outline of the i-Patch procedure, the full algorithm is described in (Hamer et al. [2010]).

Let a_{ji} denote the residue type of the ith amino acid in the jth sequence in the MSA. The i-Patch score is calculated using a triangle of amino acids, say a_{ji}, a_{jk} and a_{jl} corresponding to three residue categories: C_{ji}, C_{jk} and C_{jl}. The categories are a reduced alphabet consisting of seven groupings of the twenty amino acids (Hamer et al. [2010]).

The i-Patch score for a residue i is the average triangle propensity, S_i among all possible triangles between site $i_t \in \Pi(i)$ on protein A and the k surface exposed sites on protein B and a residue l which is a structural neighbour of either i_t or k.

$$S_i = \frac{1}{\Pi(i)} \Sigma_{i_t \in \Pi(i)} (w_{ii_t}^{intra} S_{i_t}^{Triangle}) \tag{4.1}$$

$$\begin{aligned} S_{i_t}^{Triangle} = \frac{1}{|\{\text{k exposed on B}\}|} \Sigma_{\text{k exposed on B}} \frac{1}{|\Pi(i_t) \cup \Pi(k)|} \times \\ \Sigma_{l \in \Pi(i_t) \cup \Pi(k)} \frac{1}{M - |G(i_t)|} \times \\ \Sigma_{j=1}^{M-|G(i_t)|} w^{pair}(C_{jl}|C_{ji_t}, C_{jk}) \frac{p^{con}(C_{jl}, C_{ji_t}, C_{jk})}{p^{non}(C_{jl}, C_{ji_t}, C_{jk})} \end{aligned} \tag{4.2}$$

In the above equations, M stands for the number of sequences in the MSA. The function $G(i_t)$ is the set of sequences from the MSA that do not have a gap at position i_t. The expression $w_{ii_t}^{intra}$ is the weight for the intra-protein interaction between the residues i and i_t. The pair weight $w^{pair}(C_{jl}|C_{ji_t}, C_{jk})$ scales the pair interaction of residues with categories C_{ji_t} and C_{jk} when a residue with category C_{jl} is found within 4.5Å of either of them. The expressions $p^{con}(C_{jl}, C_{ji_t}, C_{jk})$ and $p^{non}(C_{jl}, C_{ji_t}, C_{jk})$ stand for the propensity of the given triangle of residues to be in contact or not to be in contact respectively. The full details of the weights $w_{ii_t}^{intra}$ and $w^{pair}(C_{jl}|C_{ji_t}, C_{jk})$ are given below.

$$w_{ii_t}^{intra} = \frac{1}{M - |G(i_t) \cup G(i)|} \Sigma_{j \in G^c(i_t) \cup G^c(i)} w^{intra}(C_{ji_t}|C_{ji}) \tag{4.3}$$

$$w_{intra}(C_2|C_1) = \frac{w_{intra}^{con}(C_2|C_1)}{w_{intra}^{non}(C_2|C_1)} \tag{4.4}$$

$$w_{intra}^{non}(C_2|C_1) = \frac{f_{C_2 \in N(C_1)}^{non} / f_{C_2 \in N(C_1)}^{all}}{f_{C_1}^{non} / f_{C_1}^{all}} \tag{4.5}$$

$$w_{intra}^{con}(C_2|C_1) = \frac{f_{C_2 \in N(C_1)}^{con} / f_{C_2 \in N(C_1)}^{all}}{f_{C_1}^{con} / f_{C_1}^{all}} \tag{4.6}$$

$$p^{con}(C_{jl}, C_{ji_t}, C_{jk}) = \frac{f_{C_{jl}, C_{ji_t}, C_{jk}}^{con} / f_{C_{jl}, C_{ji_t}, C_{jk}}^{all}}{\Sigma_{T \in Triangles} f_T^{con} / \Sigma_{T \in Triangles} f_T^{all}} \tag{4.7}$$

$$p^{non}(C_{jl}, C_{ji_t}, C_{jk}) = \frac{f_{C_{jl}, C_{ji_t}, C_{jk}}^{non} / f_{C_{jl}, C_{ji_t}, C_{jk}}^{all}}{\Sigma_{T \in Triangles} f_T^{non} / \Sigma_{T \in Triangles} f_T^{all}} \tag{4.8}$$

$$w^{pair}(C_2|C_1, C_3) = \frac{w_{intra}^{con}(C_2|C_1, C_3)}{w_{intra}^{non}(C_2|C_1, C_3)} \tag{4.9}$$

$$w_{intra}^{non}(C_2|C_1, C_3) = \frac{f_{C_2 \in N(C_1, C_3)}^{non} / f_{C_2 \in N(C_1, C_3)}^{all}}{f_{C_1, C_3}^{non} / f_{C_1, C_3}^{all}} \tag{4.10}$$

$$w_{intra}^{con}(C_2|C_1, C_3) = \frac{f_{C_2 \in N(P)}^{con} / f_{C_2 \in N(C_1, C_3)}^{all}}{f_{C_1, C_3}^{con} / f_{C_1, C_3}^{all}} \tag{4.11}$$

In the above equations, $f_{C_1}^{con}$ is the absolute frequency of residues in category C_1 in the training dataset that are in contact. Corresponding $f_{C_1}^{non}$ is the absolute frequency of residues of category C_1 in the training set that are not in contact. The frequency of all the residues in the training set of category C_1 is denoted $f_{C_1}^{all}$. The frequencies having $C_2 \in N(C_1)$ as the subscript, are analogous but here the residue with category C_1 is further constrained to have one of category C_2 within 4.5Å of itself. Equations 4.7 and 4.8 give the propensities of a given triangle of residue categories to be in contact. Frequencies analogous to those for a single category were employed here, now computing the number of times a given triangle of residue categories was observed in contact or not in contact. The expressions in the denominators of equations 4.7 and 4.8 are sums over triangles (or triples) of residue categories (T) over all possible unordered combinations thereof ($Triangles$).

The expressions for $f^{all}_{C_1,C_3}$, $f^{con}_{C_1,C_3}$ and $f^{non}_{C_1,C_3}$ are defined as previously, with the exception that they denote the frequency of specific contact pairs with categories C_1, C_3. The subscript $C_2 \in N(C_1, C_3)$ again stands for the frequency of the pair C_1, C_3 but with C_2 within 4.5Å of either C_1 or C_3.

In order to convert equations 4.1 and 4.2 for use on antibody antigen complexes, we must introduce a directionality of interaction (to take into account the different binding propensity of antibodies) and remove the multiple sequence alignment (as an MSA is not available in the antibody-antigen case). Thus if C_{ij} in the original i-Patch stood for the classification of the jth residue in the ith MSA sequence, in antibody i-Patch C^{Ab}_{1k} means the antibody classification of the kth residue in the reference sequence (since there is only one sequence rather than an entire MSA it is referred to it as '1' in the subscript so as to preserve the original notation). The new pair propensities for particular residues to be in contact were calculated using dataset NR-full. This produced a 20-element vector of propensities for each residue type (asymmetric for antibody and antigen), where each entry corresponds to the propensity for a given amino acid towards a particular amino acid type on the other molecule. These vectors were used for K-means clustering of the 20 amino acids into seven groups

After successive rounds of K-means clustering with K=7, we have curated the grouping of amino acids presented in Table 4.1. Note that majority of the groups are singletons, corresponding to the more acute preferences for those amino acids to act as contact residues in antibodies. All the residues that fall in the biggest cluster (cluster I), are those that do not interact strongly with the antigen and hence are grouped together even though they are chemically distinct from one another.

The intra-protein weight annotation with AB or AG indicates the molecule type for which it was calculated (e.g $w^{ABintra}$ for antibodies). In the case of antigens, The same propensities and weightings as those for general proteins were used. In triangle propensities, consisting of a pair of residues on different molecules and a third one within 4.5Å of one of them, annotation with either AB or AG indicates the molecule to which the third, neighbouring, residue belongs. Thus $p^{AGcon}(C^{AG}_{1l}, C^{AG}_{1k}, C^{AB}_{1i})$ stands for the

Cluster	Amino Acids
I	P I L M T V A C Q G H K S
II	D R
III	F
IV	E
V	Y
VI	W
VII	N

TABLE 4.1: Clustering of amino acids in antibodies into seven groups, according to their 20-element contact propensity profile vectors.

contact propensity of C_{1k}^{AG} on the antigen, and C_{1i}^{AB} on the antibody with C_{1l}^{AG} being in the neighborhood of C_{1k}^{AG}.

Here we give the formulas for the scores computed for the antibody since the antigen case is symmetric. The Antibody i-Patch score for the antibody molecule is given by equations 4.12 and 4.13.

$$S_i^{AB} = \frac{1}{\Pi(i)} \Sigma_{i_t \in \Pi(i)} (w_{ii_t}^{ABintra} S_{i_t}^{ABTriangle}) \tag{4.12}$$

$$\begin{aligned} S_{i_t}^{ABTriangle} = \frac{1}{|\{k \in Ag\}|} \Sigma_{k \in AG} \frac{1}{|\Pi(i_t) \cup \Pi(k)|} \times \\ [\Sigma_{l_{ab} \in \Pi(i_t)} (w^{ABpair}(C_{1l_{ab}}^{AB} | C_{1i_t}^{AB}, C_{1k}^{AG}) \frac{p^{ABcon}(C_{1l_{ab}}^{AB}, C_{1i_t}^{AB}, C_{1k}^{AG})}{p^{ABnon}(C_{1l_{ab}}^{AB}, C_{1i_t}^{AB}, C_{1k}^{AG})}) + \\ \Sigma_{l_{ag} \in \Pi(k)} (w^{AGpair}(C_{1l_{ag}}^{AG} | C_{1i_t}^{AB}, C_{1k}^{AG}) \frac{p^{AGcon}(C_{1l_{ag}}^{AG}, C_{1i_t}^{AB}, C_{1k}^{AG})}{p^{AGnon}(C_{1l_{ag}}^{AG}, C_{1i_t}^{AB}, C_{1k}^{AG})})] \end{aligned} \tag{4.13}$$

The scores for the antibody were calculated for each CDR region plus two residues on either side according to the IMGT definition (for the details on the decision to take two residues and not more please see 4.2.3). This was done as the majority of the binding site of an antibody is known to be composed of the CDRs and a few residues on either side. The entire antigen surface was used. Antibody i-Patch was then evaluated on the dataset NR-full, to compare with existing CDR definition methods and on datasets RA-x and RA-h to validate it's applicability to models.

4.2.3 Contact data for framework residues

We have chosen to calculate the Antibody i-Patch scores using the IMGT (Lefranc [2011]) definition, augmented by two framework residues on either side. This decision was motivated an analysis where we have calculated average distances from framework residues to the protein antigen (more than 50 residues - ie no peptides and no haptens). The dataset for this study consisted of the non-redundant antibody-antigen complexes (distinct antibody and antigen sequences) from SAbDab.

For each complex in the non-redundant dataset of antibody-antigen structures, we have calculated the minimal distance from any residue to the antigen (taking 4.5Å as the distance cutoff between any two heavy atoms). The antibody sequences were aligned to one another so that it was possible to calculate average distance for each position in the antibody framework, and the results of this analysis are presented in Figures 4.3 and 4.4. We decided that two residues on either side of the CDR are usually close enough to be in contact with the antigen and thus we employ the CDRs using IMGT definition extended by two anchor residues on either side.

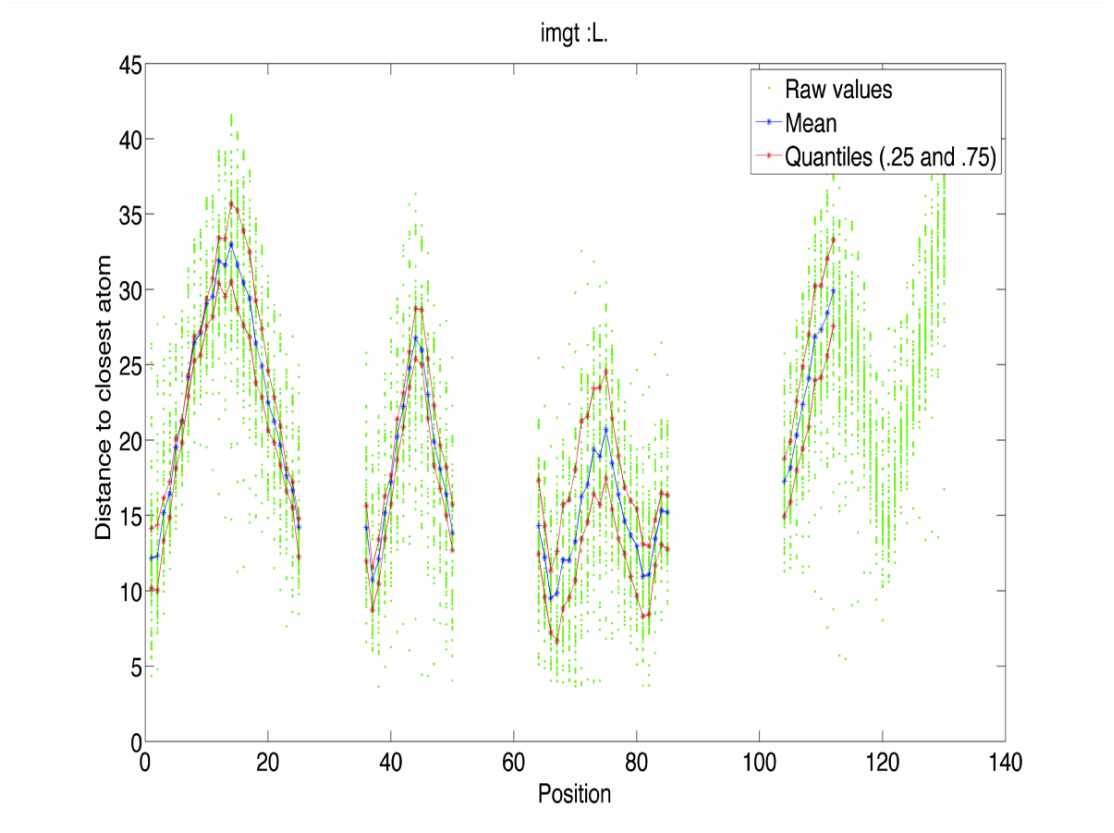


FIGURE 4.3: **Average framework distances from the antigen to the light chain of F_{ab}.** The CDR chains according to the IMGT definition were removed from the graph resulting in the three blank spaces for CDR-L1, CDR-L2 and CDR-L3 respectively. The sequences of the frameworks without the CDRs were aligned, allowing us to compute average distances to the antigen over all structures.

4.2.4 Evaluating performance of contact prediction

The evaluation procedure consisted of 2000 jackknife tests on NR-full.

One iteration of our evaluation procedure consisted of randomly splitting NR-full into training and test datasets, consisting of 118 and 30 structures respectively. Antibody i-Patch was trained using the 118 structures in the training set and then applied to each of the 30 entries in the test set. Thus each antibody in the training set would have its residues in the IMGT CDR region together with two residues on each side annotated with an Antibody i-Patch score. The performance of the method was then assessed by predicting all residues with an Antibody i-Patch score above a certain cutoff as contacts. A subset of these will be the true positives (TP), and the rest false positives (FP). Those residues below the cutoff will be either false negatives (FN) or true negatives (TN). Thus

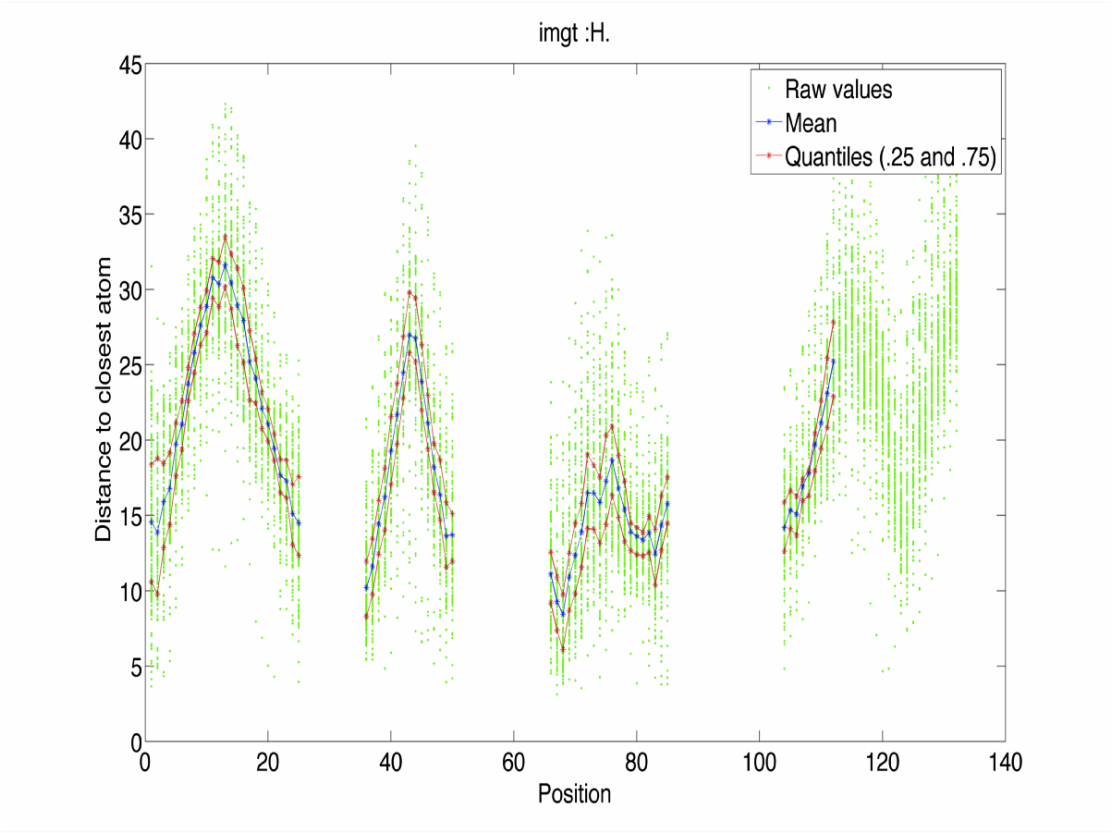


FIGURE 4.4: **Average framework distances from the antigen to the heavy chain of F_{ab}.** The CDR chains according to the IMGT definition were removed from the graph resulting in the three blank spaces for CDR-H1, CDR-H2 and CDR-H3 respectively. The sequences of the frameworks without the CDRs were aligned, allowing us to compute average distances to the antigen over all structures.

for each score cutoff we calculate its corresponding precision (TP/(TP+FP)) and recall (TP/(TP+FN)). All cutoffs between 0.1 and 300 at intervals of 0.1 were tested.

The evaluation procedure described above was repeated 2000 times on distinct partitions of NR-full into 118 training structures and 30 test structures. The precision and recall scores corresponding to a particular cutoff from each run were averaged over the 2000 iterations. The predictive power of the static CDR loop and region definitions was carried out by annotating the antibody sequences with the corresponding Kabat, Chothia and Contact regions as defined using the Chothia numbering scheme. Antibody sequences were numbered using Abnum (Abhinandan K R [2008]). The IMGT definitions were copied directly from the IMGT website (Lefranc [2011]) or annotated manually for those structures not yet in the database.

4.2.5 Computational alanine scanning

Following the protocol outlined by the authors of Paratome (Kunik et al. [2012]) for alanine scanning in silico, we have introduced alanine mutations to each binding site residue (of the IMGT CDRs together with two anchor residues on each side) which was in contact with the antigen in the native structure (closest heavy atoms within 4.5Å from each other). FoldX (Schymkowitz et al. [2005]) was used to calculate the interaction energy between the wild type antibody and antigen and each alanine mutated antibody and antigen, allowing us to calculate the interaction energy change ($\Delta\Delta G$) between the wild type and antigen and the mutant and antigen. Following Paratome (Kunik et al. [2012]), an interaction energy change $\Delta\Delta G \leq$ -0.25 kcal/mol was defined as stabilizing whereas if it was the case that $\Delta\Delta G \geq 0.25$ kcal/mol, it was destabilizing.

In order to assess the energetic importance of the Antibody i-Patch score, we took all residues with an Antibody i-Patch score greater than a cutoff and calculated the per-centage of these residues that are destabilizing ($\Delta\Delta$G ≥ 0.25 kcal/mol by computational alanine scanning). As we have 2000 independent runs, we calculate the average percentage of destabilizing positions above a given score. Thus, if the Antibody i-Patch score is related to energetic importance, a larger per-centage of residues will be destabilizing at higher scores.

4.3 Results

4.3.1 Antibody i-Patch

Antibody i-Patch is an adaptation of protein-protein contact predictor i-Patch (Hamer et al. [2010]). The original i-Patch algorithm receives on input the structures of two proteins which are assumed to interact, together with the multiple sequence alignment (MSA) of interacting homologs (Hamer et al. [2010]). The algorithm uses the contact propensity data contained in its training dataset of domain-domain interactions and

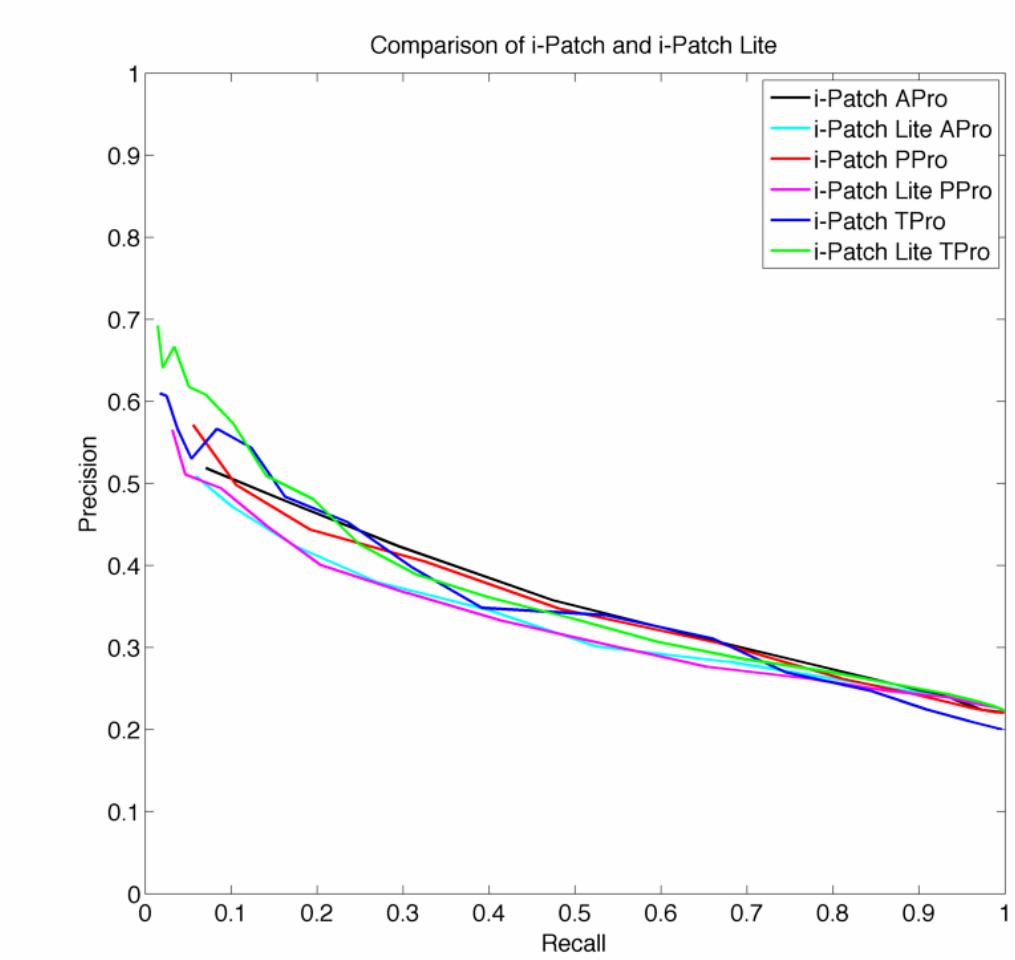


FIGURE 4.5: Comparison of performance of i-Patch (using MSAs) and i-Patch Lite (not using MSAs) on the test set used in the original i-Patch paper (Hamer et al. [2010]). Results are presented for all three scores, APro, PPro and TPro. Results of i-Patch are not statistically significantly different from these of i-Patch Lite meaning thus the i-Patch algorithm can be applied using structural input alone, without resort to MSAs.

protein complexes and the correlated mutations in the input MSA to generate a statistical score for each surface residue of the input proteins. The score for each residue is calculated using a patch of residues on the proteins surface centred on the residue of interest. The standard i-Patch score is relatively accurate at annotating contact residues (59% precision at 20% recall on a blind test-set of 31 targets).

The i-Patch methodology is not immediately applicable to the antibody-antigen complexes for two reasons. Firstly, because it is impossible to create the necessary MSAs. As antibody-antigen complexes do not undergo correlated mutations, but rather the

antibody becomes fine tuned to bind a particular antigen, there is no evolutionary information (MSA) available for either the antibody or antigen. Adaptation of the algorithm to the special case of antibodies involves removing the MSA from the input, which we have shown to be possible. Figure 4.5 demonstrates results of the original i-Patch and its reduced version, i-Patch Lite which does not use MSA information. Note that for any pair of corresponding scores between the two methods, the results appear to be similar. In fact, we have demonstrated that the results of the original i-Patch are not statistically significantly different from these obtained by i-Patch Lite.

Secondly, standard i-Patch uses contact propensities from protein-protein binding sites which may not be applicable to antibody antigen complexes. Similarly as in the previous Chapter, we compared the binding propensity of amino acids in antibodies and antigens to those in general proteins on a large dataset of antibody-protein complexes (148 structures) - Figure 4.2. In accord with previous analysis, we find that antibodies appear to have a very different binding profile from general proteins. The antigen binding profile on the other hand is not so profoundly different from that of general proteins.

Figure 4.2 shows the propensities of the different residue types to be involved in binding in antibodies, in antigens and in proteins in general. A propensity value greater than '1' indicates that if this residue is seen on the surface of, say antibodies, it is likely to be involved in binding to an antigen and conversely for scores less than '1' (for full details see materials and methods section). We observed a strong preference for tyrosine and tryptophan to be in the binding sites of antibodies, not seen for proteins in general. These high contact propensities of tyrosine and tryptophan in antibodies are in agreement with earlier theoretical (Mian et al. [1991], Wilson and Stanfield [1993]) and experimental (Fellouse et al. [2005], Birtalan et al. [2008]) findings. Serine shows a different behaviour, because even though it is known to be prevalent in the antibody binding site (Zemlin et al. [2003], Golub et al. [1997]), its binding propensity profile in Figure 4.2 does not differ significantly from that of general proteins. In contrast to analysis in the previous Chapter we notice a diminished propensity for histidine to form contacts in antibodies. This difference might be the result of a higher number of antibody-antigen complexes

used for this analysis (148 structures here, 121 in the previous Chapter), which provides a better approximation of the preferences in such binding sites, especially since the relative frequency of histidine in antibodies are relatively small.

Given that antibodies appear to have binding preferences distinct from those of general proteins, we recalculated the contact scores used in i-Patch so as to accommodate those differences. The propensity information used in Antibody i-Patch was calculated using a jackknife procedure on the NR-full dataset whilst keeping the original protein propensities for the antigen.

Using the fact that the general position of the antibody binding site is known, we constrained the region which we consider for the analysis to the IMGT CDR definition augmented by two further residues on each side. This augmentation of the IMGT definition was motivated by an analysis of which framework regions were most likely to be in contact with the antigen, presented in 4.2.3.

4.3.2 Antibody i-Patch predicts antigen binding residues

We have used jack-knife procedure on a large non-redundant dataset NR-full downloaded from SAbDab (Dunbar et al. [2013b]) so as to obtain a comparison to other contact site annotation methods.

The performance of Antibody i-Patch is evaluated using a P-ROC graph (e.g. 4.6). Here the precision (or accuracy) is plotted against the recall (or coverage). A perfect predictor would give precision of 1.0 for all values of recall. In our case, if we impose a high score cutoff, we achieve precision values up to 77% but at a recall of only 10% (Figure 4.6). At a low score cutoff Antibody i-Patch precision falls to 42% with a coverage of 93%. The results presented here are the averages of P-ROC plots of 2000 jackknife tests on NR-full.

The results indicate that Antibody i-Patch achieves both higher precision and recall as compared to the static CDR definitions. The recall of the different methods is limited

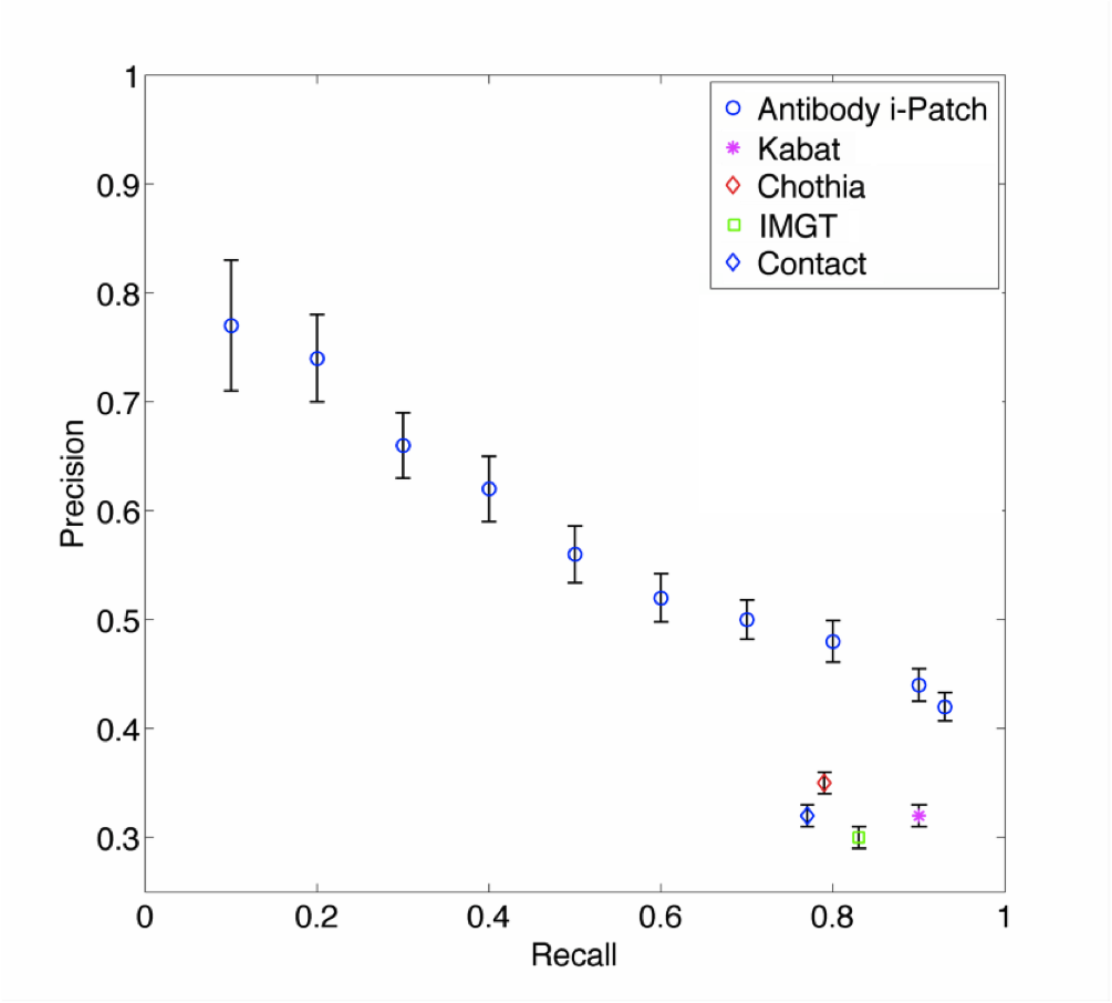


FIGURE 4.6: P-ROC plot of Antibody i-Patch results averaged over 2000 runs. Comparison of the performance of Antibody i-Patch with static antibody binding site annotation methods for the contact distance of 4.5Å. Standard error is shown for the values of precision, (recall errors can be found in A.5). In contrast to the static antibody binding site annotation methods of Kabat, Chothia, Contact and IMGT, Antibody i- Patch produces results for a wide spectrum of precision and recall values. As all the residues outside the window of IMGT definition augmented with two framework residues on either side are considered to be non-binding, the recall starts at 93%. If the original IMGT definition had been used the same graph would be truncated to the point corresponding to that of IMGT, i.e. recall of 83%.

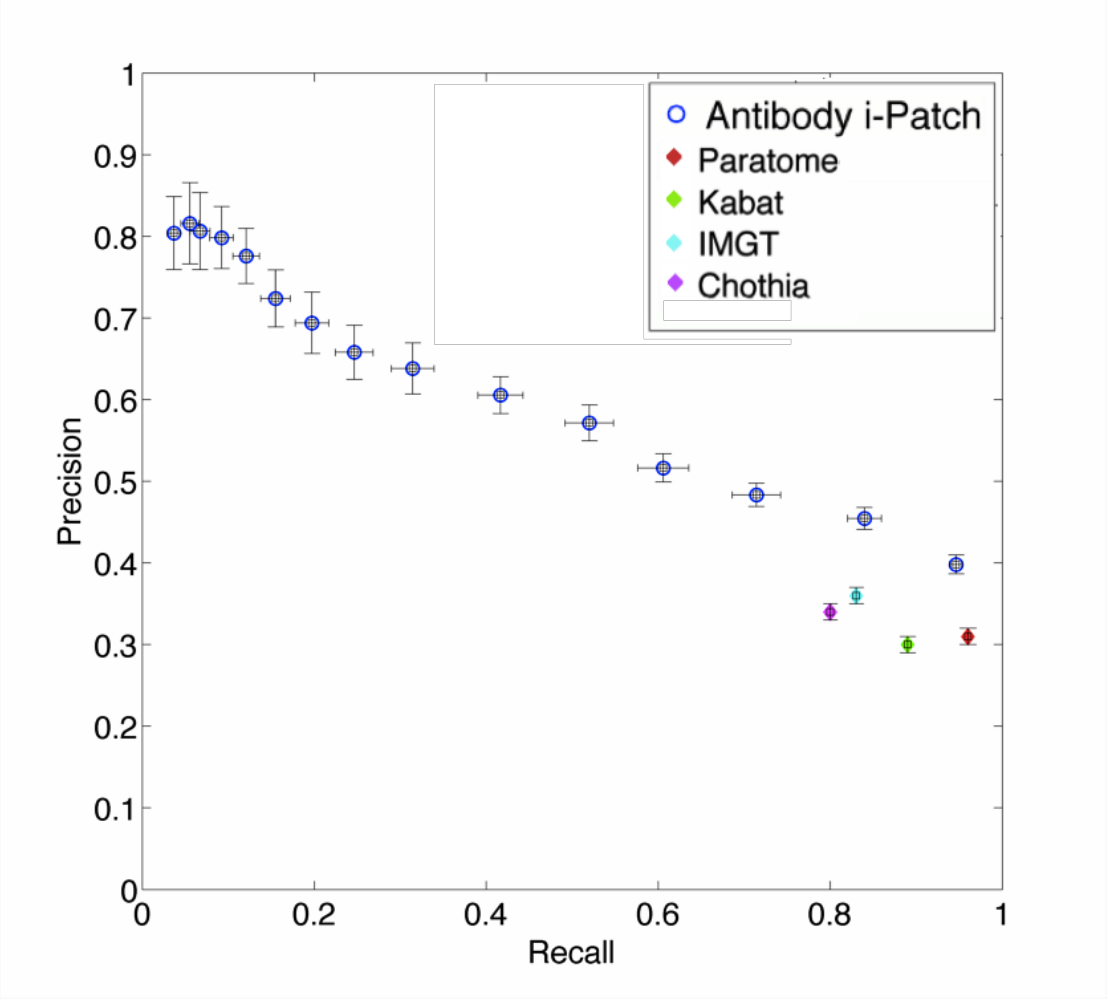


FIGURE 4.7: **Comparison of the different antibody contact site predictors.** The results for Antibody i-Patch are those trained on dataset in Table A.2 and applied to test set in Table A.3. The values for Paratome and other methods are taken from Kunik et al. 2012.

by the residues included in the predictions of each of the CDRs. This can be seen in a comparison between Antibody i-Patch and IMGT. Antibody i-Patch reaches the highest recall of 93% whereas the recall of IMGT is only around 83%. This increase is entirely due to the inclusion of the two extra framework residues on either side of each of the CDRs in the Antibody i-Patch predictions.

Out of the static CDR region and loop annotation methods, Kabat regions achieve highest recall with a value of 90%. The Contact definition (MacCallum et al. [1996]), which was designed based on antibody-antigen contacts underperforms in this experiment, achieving the lowest recall of all 77%. This is probably due to the small size of

its original training set of 26 antibody antigen complexes (not all of which were protein antigens). All the static methods achieve precision in the region of 30%.

We have also tested the Antibody i-Patch algorithm on the Paratome test set (see Figure 4.7). On this dataset Antibody i-Patch achieves highest recall of 94% at 40% precision compared to 96% recall and 31% precision by Paratome. This indicates that i-Patch does not achieve recall quite as high as Paratome but is more precise.

Unlike the other methods including Paratome, Antibody i-Patch does not just give a yes-no result for each residue, instead each residue is given a score related to its likelihood to be in the binding site. Thus if a user was not interested in identifying the entire binding region (which all methods achieve with relatively low precision) but instead wished to know with great certainty a small number of residues involved in the binding site (low recall, high precision), they could pick a high Antibody i-Patch score cutoff, say 160, which corresponds to recall of 40% and a precision of 63% (see Figure 4.6).

4.3.3 Evaluating the performance on homology models.

In the previous section we have established that Antibody i-Patch predictions are superior to the static CDR annotation techniques. Those methods however require only the sequence of the antibody as input, whereas Antibody i-Patch requires structural neighbour information to operate. Here we demonstrate that Antibody i-Patch can produce satisfying results by using a homology model only, making it equivalent with the static annotation techniques which require sequence only. In order to evaluate Antibody i-Patch on homology models, we execute the program on dataset RA.

The RA dataset consisted of 22 crystal structures (dataset RA-x) and 22 homology models (dataset RA-h). It was ensured that there was no redundancy between datasets NR-full and RA by removing those entries from NR-full that were too similar to those in RA (more then 90% sequence identity for the antigens and more than 99% for the antibody). The homology models were generated by RosettaAntibody (Sivasubramanian

et al. [2009]). The coordinates for CDR-H3 were re-modelled using FREAD (Choi and Deane [2010]) as it was shown that it produced better models for this particular loop (Choi and Deane [2011]). Since FREAD is a database search method, in some instances the best structure returned came from the query structure itself. In such cases this result was removed and the second best structure was used. We also tested the performance of the algorithm on the homology models without co-ordinates for CDR-H3. In the case where there were no co-ordinates for CDR-H3 we created the patch of Antibody i-Patch from the immediate sequence neighbours of a given residue.

The performance of Antibody i-Patch on homology models (RA-h) is very similar to that on crystal structures (RA-x). The P-ROC graphs are presented in Figure 4.8.

There are no statistically significant differences between any pair of the three tests (crystal structures, models and models with no CDR-H3 coordinates), as tested by area under the curve analysis of ROC plots (for details please see A.4). However, there appears to be a difference between the plots for homology models with modelled CDR-H3 loops and those with no H3 at low recall, high precision values. As one would expect, using a modelled H3 appears to improve precision in the low recall region. We also tested the use of an entirely sequence based Antibody i-Patch where the patch was defined by the sequence neighbours rather than structural neighbours, which considerably weakened the performance of the algorithm, thus the construction of homology models is an essential step in the prediction pipeline.

The fact that Antibody i-Patch performs as well on homology models as on crystal structures probably arises from the fact that it requires structural information only for the purpose of determining neighbour patches on the surface. Therefore, even fairly low-resolution homology models could provide an acceptable approximation of the native structure.

The above results indicate that Antibody i-Patch outperforms current CDR annotation techniques and it needs only the sequence of an antibody as input. The corresponding

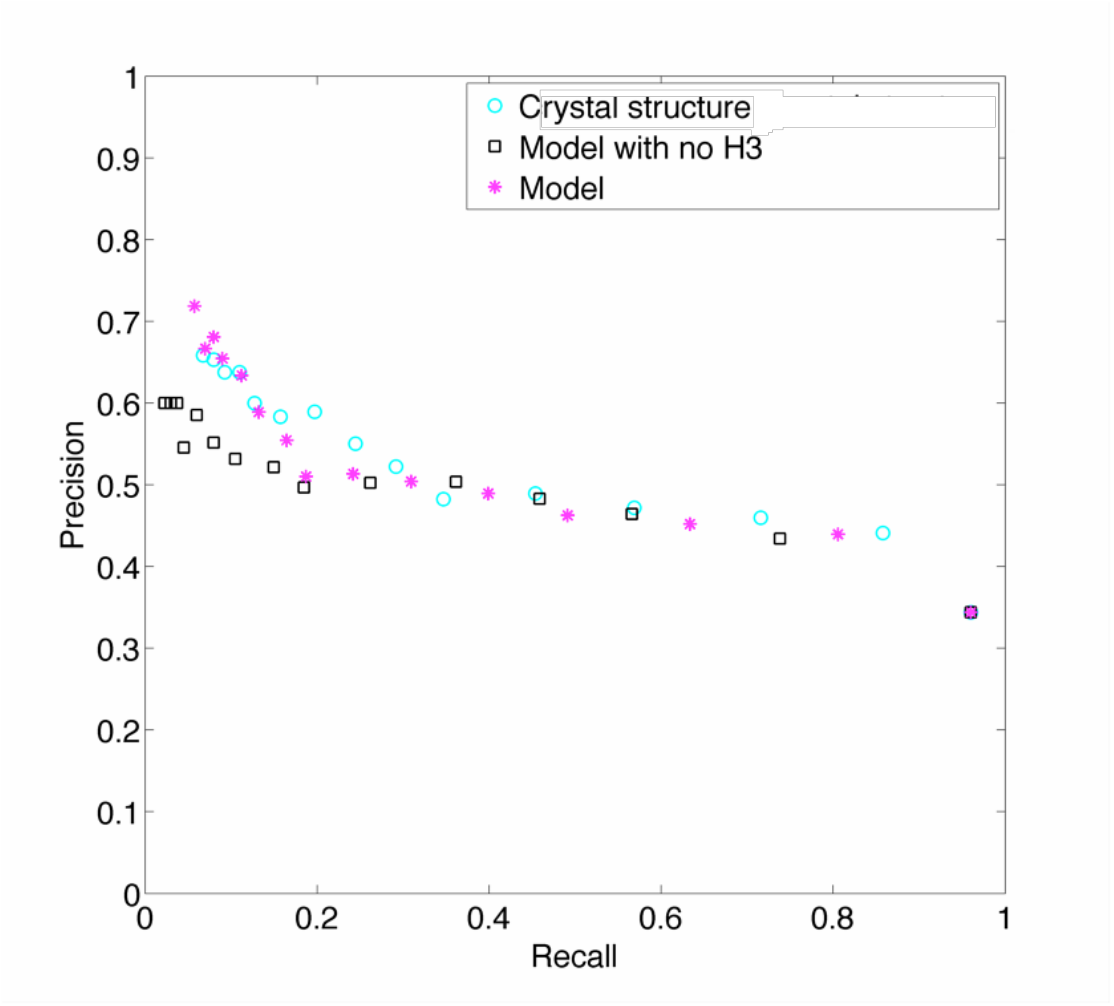


FIGURE 4.8: **Performance of i-Patch on dataset RA.** Performance on crystal structures (RA-x) is similar to this on homology models with H3 modelled by FREAD (RA-h). There is a slight drop in performance when H3 is not modelled and instead immediate sequence neighbours are used as the patch for Antibody i-Patch. The corresponding standard errors can be found in A.6.

homology model can be created using RosettaAntibody, augmented with the CDR-H3 provided by FREAD.

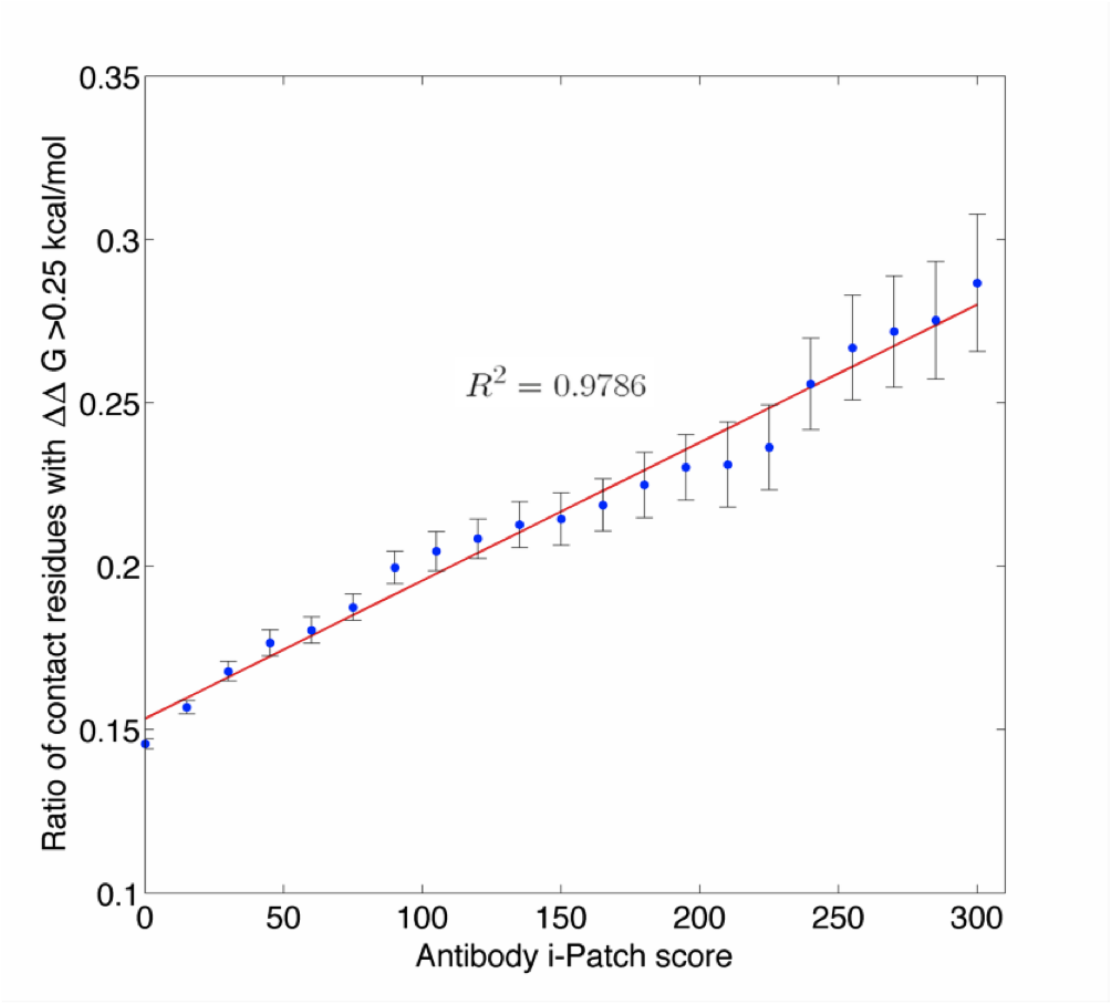


FIGURE 4.9: The energetic importance of antibody-antigen contact residues is correlated with the Antibody i-Patch score. The number of contact residues with an Antibody i-Patch score greater than a cutoff which lead to a $\Delta\Delta G > 0.25$ kcal/mol when mutated to alanine compared to the number of contact residues in general that lead to a $\Delta\Delta G > 0.25$ kcal/mol when mutated to alanine. As the Antibody i-Patch score cutoff is increased, the ratio of residues which cause an energetic change upon alanine mutation increases. In other words residues with a high Antibody i-Patch score tend to be energetically more important.

4.3.4 Residues with higher Antibody i-Patch scores are more energetically significant

We have investigated the energetic significance of the results returned by Antibody i-Patch. As a background distribution we use the residues from across the entire IMGT-CDR regions augmented by two residues on either side of each CDR. We contrast the effect of mutating residues in this background set with mutating residues which achieve

high Antibody i-Patch scores.

We follow the procedure delineated by Kunik et al. (2012) and excluded the non-binding residues from this analysis because our aim was to explore the link between the Antibody i-Patch score - which as demonstrated correlates with higher likelihood to be a contact site - and energetic importance. Even though there are non-binding residues that are crucial for the stability of the antibody-antigen complex, our aim was to establish if the higher-likelihood binding residues are energetically important, thus inclusion of non-contacts would be meaningless and could potentially skew the energetic difference distribution.

Using FoldX (Schymkowitz et al. [2005]), we have performed in silico alanine scanning of the anti-body contact residues from the region of IMGT-CDRs augmented by two framework residues on either side for each structure in dataset NR-full. In each of the 2000 runs of Antibody i-Patch on dataset NR-full, we have noted the fraction of destabilizing residues above each score cutoff. We have averaged the proportions over those 2000 runs for each score cutoff. As in the previous analysis (Kunik et al. [2012]), the majority of the mutations lie in the neutral region of -0.25 $\leq \Delta\Delta G \leq$ to 0.25 kcal/mol. For increasing values of the Antibody i-Patch score, the fraction of the contact residues with destabilizing mutations ($\Delta\Delta G \geq$0.25 kcal/mol) rises steadily, as shown in Figure 4.9.

In silico alanine scanning is not a substitute for the actual experimental data, but it acts as an indicator of the general energetic behaviour of the mutations. Based on this FoldX analysis we suggest that residues receiving higher Antibody i-Patch scores not only have a higher likelihood of being in contact but also tend to be more energetically important.

4.4 Conclusions

In this Chapter we have used the information from our prior antibody-antigen contact analysis to develop an antibody contact prediction tool Antibody i-Patch. We have demonstrated that Antibody i-Patch has superior predictive capacity to this of static methods. We have shown that our software can operate using homology models, making the only required input the sequence of the antibody. Finally, we have shown that the residues with higher Antibody i-Patch scores are both more likely to be antigen contacting residues as well as be more energetically important for the complex formation.

In the course of the Antibody i-Patch analysis we have checked the capacity of this software to predict the contact sites on the antigen. Unfortunately, we did not obtain satisfying results for the antigen in the same way as we have done for the antibody. Since the knowledge of precise pairwise contacts between the antibody and antigen are important for potential applications of computational antibody design, we address this issue in the next Chapter using the expertise developed here.

Chapter 5

Local Antibody-Antigen Docking

5.1 Introduction

Identification of the antibody contact residues carried out in the previous Chapter does not give any information about the antigen binding site (epitope). Specifically, it does not provide any information pertaining to the pairwise contacts established between the antibody and the antigen. Therefore, as a natural continuation of our work in developing tools for computational antibody design, we decided to use our Antibody i-Patch predictions to find out which pairwise contacts are formed in a given antibody-antigen interface through the medium of antibody-antigen docking.

5.1.1 Motivation

In many cases, docking of antibody-antigen complexes can be reduced to a local docking problem. The binding site of the antibody is known a priori, in the form of the CDRs and its surrounding regions (Kunik et al. [2012], Sircar and Gray [2010]). The epitope location on the antigen is also sometimes known either through experimental data or a

desire to target an antibody to a particular location on an antigen (Smith and Sternberg [2003], McKinney et al. [2007], Covaceuszach et al. [2008]). In such cases local docking would be appropriate.

Even though there has been progress in improving the docking of general proteins, as demonstrated by successive rounds of the CAPRI experiment (Mendez et al. [2005]), there are currently only two methods which specifically address antibodies - SnugDock and ClusPro in antibody mode (Sircar and Gray [2010], Brenke et al. [2012]). This may be due to the specific biology of the antibody binding site, meaning that general docking tools are not ideal for this problem (Brenke et al. [2012], Sircar and Gray [2010], de Vries and Bonvin [2011], Li and Kihara [2012]). Other methods, like ZDOCK, attempt the problem by constraining the binding region to the CDRs.

Often docking is further hampered by the lack of solved crystal structures (in our case antibodies and antigens) to serve as input (Sircar and Gray [2010]). One solution to this problem is to create a model structure of the molecule in question (Sircar and Gray [2010], Tovchigrechko et al. [2002], Mosca et al. [2009]). In the case of antibody-antigen complexes the problem is somewhat easier than the general protein case as antibodies have a very well conserved overall structure (Sivasubramanian et al. [2009]) and the CDRs (with the notable exception of CDR-H3) often adopt similar structures (Choi and Deane [2011], Chothia and Lesk [1987], Chothia et al. [1989], North et al. [2011], Martin and Thornton [1996], Lara-Ochoa et al. [1996], Al-Lazikani et al. [1997]). This facilitates modelling, as demonstrated by RosettaAntibody, WAM and PIGS (Sivasubramanian et al. [2009], Whitelegg and Rees [2000], Marcatili et al. [2008]). Nonetheless, as currently no protein model is perfect, docking methods often struggle when they are used as input (Tovchigrechko et al. [2002]). This issue was tackled by SnugDock which produces acceptable quality complexes through flexible docking of homology models, thus mitigating modelling errors and induced fit issues (Sircar and Gray [2010]). Coupling SnugDock with EnsembleDock (Chaudhury and Gray [2008]), achieves even better results as it does not have to rely on a single model. According to the CAPRI rating (Mendez et al. [2005]) SnugDock with EnsembleDock achieves (5* 9** 0***) over the

top 10 decoys of each of 15 antibody-antigen targets used in their study, as compared to (6* 5** 0***) using standard RosettaDock (Sircar and Gray [2010], Gray et al. [2003]). Here the number of stars indicates the quality of the decoy with three stars being close-to-native. The drawback of using SnugDock is that it takes hundreds of CPU hours to achieve such results.

Here, we show the applicability of Antibody i-Patch by using the predicted contact residues as constraints for local antibody-antigen docking. Using fast docking algorithms, ZDOCK (Chen et al. [2003]) and PatchDock (Duhovny et al. [2002], Schneidman-Duhovny et al. [2005]) and using our constraints and antibody-specific re-scoring scheme we can achieve results in the range of (7* 4** 0***), (6* 5** 0***) or (3* 8** 0***) for the top 10 decoys on the SnugDock homology model test set, but in minutes rather than hundreds of hours per target required for SnugDock.

5.2 Materials and Methods

5.2.1 Data

5.2.1.1 Dataset NR-subset

30 structures were chosen at random from dataset NR-full, presented in the previous Chapter, to constitute the crystal structure test set NR-subset. Full list is available in B.1.

5.2.1.2 Dataset SnugDock-H

The dataset for testing our docking pipeline (SnugDock-H) was identical to that used in the RosettaAntibody (Sivasubramanian et al. [2009]) and SnugDock (Sircar and Gray [2010]) studies. It consists of 15 antibody-antigen pairs. The dataset SnugDock-H is a

subset of RA from the previous Chapter, the former is simply a version of the latter constrained to 15 antibody-antigen pairs. In SnugDock-H, the models are the same as those in RA-h (RosettaAntibody with FREAD predictions for CDR-H3). A full list of these is given in Table B.2.

5.2.2 Docking methods

NR-subset and SnugDock-H were used for this analysis and two docking methods were tested: ZDOCK (Chen et al. [2003]) and PatchDock (Duhovny et al. [2002], Schneidman-Duhovny et al. [2005]). ZDOCK was used in its default mode to produce 2000 decoys. PatchDock was also run in its default mode, producing an indeterminate number of decoys. The antibody was treated as the receptor in both pieces of software. The top 200 scoring decoys for each target, according to each method were collected for further analysis. ZDOCK and PatchDock both allow for docking constraints to be specified in the input. We used this facility to enter constraints based on the Antibody i-Patch scores. In the manuscript the results are shown for ZDOCK (PatchDock results which are similar are given in Appendix B.3).

5.2.2.1 Antibody constraint

The binding constraint for the antibody is the set of residues that are assumed to be involved in binding (see Figure 5.1 for an example). There were three versions of the antibody constraint given to the docking algorithms: the IMGT CDRs (referred to as C-CDR), the residues that were predicted by Antibody i-Patch to be binding sites (referred to as C-Antibody i-Patch) and the actual residues constituting the paratope (referred to as C-Native). The residues for C-Antibody i-Patch were generated by applying Antibody i-Patch to antibody-antigen pairs in datasets NR-subset or SnugDock-H and taking everything above the cut-off of 40.0 so as to achieve good balance between precision and recall. Antibody i-Patch was trained on versions of NR-full with redundancy removed

by CD-HIT (Li and Godzik [2006]) with respect to either NR-subset or SnugDock-H (no more than 99% antibody sequence identity and no more than 90% antigen identity). The value of 40 was motivated by the average results of Antibody i-Patch on the training set portions of the Antibody i-Patch runs. The set of residues provided as the antibody constraint is referred to as C_{ab}.

5.2.2.2 Antigen constraint

An extended epitope was provided as input to the docking algorithms which consisted of the actual binding residues together with all other residues within 4Å, 5Å or 6Å (see Figure 5.1 for an example). Using any of these constraints, included far more residues than there are in the the actual antigen epitope (see Figure 5.2). The use of augmented epitope aims to simulate the case when the epitope position on the antigen is already known, but the precise residues involved in the interaction are not. For comparative purposes, the constraint approximates the initial position of the antibody and antigen to a similar extent to that employed in SnugDock (Sircar and Gray [2010]) where the antibody is pointed towards the antigen and then rotated to add randomness. Residues in this constraint are referred to as C_{ag}.

5.2.2.3 The precision score

Let $Pr(T_{ab}, T_{ag})$ denote the precision of a contacting pair of residues with types T_{ab} on the anti-body and T_{ag} on the antigen to be correct if observed in a decoy from ZDOCK. The precision $Pr(T_{ab}, T_{ag})$ was estimated by executing ZDOCK on each antibody-antigen complex in dataset NR-full, which was not in NR-subset. As the constraint for the local docking we gave the paratope and epitope residues on each molecule, together with those within 5Å away from them. Decision to take this cut-off was an arbitrary choice acting as a middle ground with respect to the epitope cutoffs we use in this manuscript.

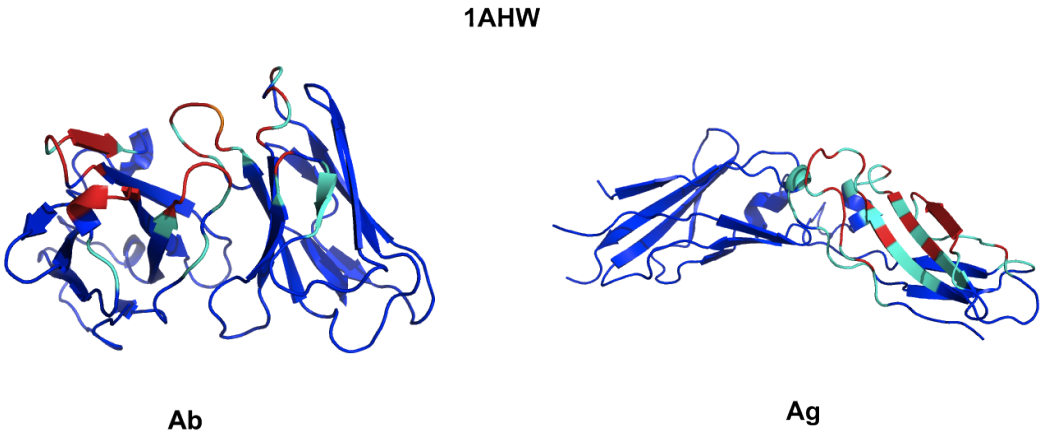


FIGURE 5.1: **Example of the constraints submitted to the docking algorithms shown on the PDB entry 1AHW.** Non-dark blue residues are those submitted as the binding constraint. For the antibody molecule, residues with Antibody i-Patch score above 40.0 were given as the binding constraint. Red residues are true positives, teal are false positives, orange is false negative and dark blue are true negatives. In the case of the antigen, red residues constitute the true epitope while teal ones are those within 5Å from them.

For each of the 118 targets in NR-full that were not in NR-subset we collected the top 200 decoys as ranked by ZDOCK. Over all the decoys collected in this manner we counted the number of true positives and false positives for amino acid types, T_{ab} and T_{ag}, being observed within 4.5Å. We denote the number of true positives and false positives collected in this manner $TP(T_{ab}, T_{ag})$ and $FP(T_{ab}, T_{ag}))$ respectively. This leads to the estimate of ZDOCK pairing up T_{ab} and T_{ag} given below:

$$Pr(T_{ab}, T_{ag}) = \frac{TP(T_{ab}, T_{ag})}{TP(T_{ab}, T_{ag}) + FP(T_{ab}, T_{ag})} \tag{5.1}$$

For example if we consider T_{ab} as cysteine residues on the antibody and T_{ag} as tyrosine residues on the antigen, we count in the 23600 decoys (118 targets times 200 decoys) the number of times cysteine residues on the antibody are in contact with tyrosines on the antigen and then partition these into true positives ($TP(T_{ab}, T_{ag})$) and false positives ($FP(T_{ab}, T_{ag})$) in order to calculate the precision ($Pr(T_{ab}, T_{ag})$). An analogous procedure was carried out for PatchDock.

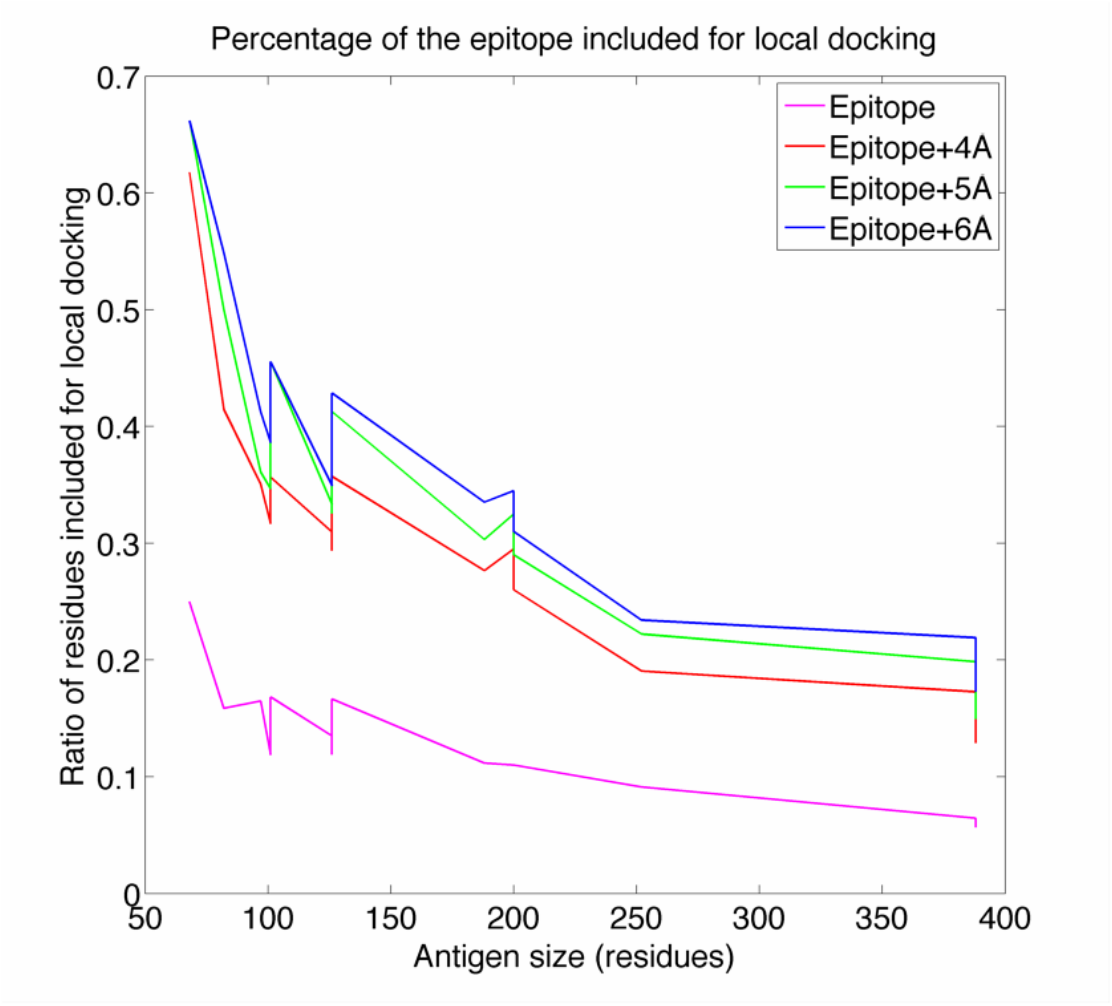


FIGURE 5.2: **Percentage of the antigen included for different cutoffs on dataset SnugDock-H.** The residues used for the antigen constraints were those within 4.5Å from the antigen and those within 4Å, 5Å and 6Å from those contacting residues. Since we appear to be supplying many more residues than are in the original epitope we consider either of the three cutoffs an accurate model for the approximate initial guess of the epitope for the local antibody-antigen docking.

We have also applied this procedure for use on the SnugDock-H dataset. Here, we used CDHIT to remove those complexes from NR-full that had more than 90% antigen sequence identity and 99% antibody sequence identity between NR-full and SnugDock. The resulting set of antibody-antigen complexes was used in an analogous fashion to create the precision score for re-scoring the decoys in SnugDock-H.

5.2.2.4 Reordering decoys

For the top 200 decoys generated by ZDOCK for a given target, a new score is computed. For each decoy, all pairs of residues within 4.5Å where one is on the antibody (r_{ab}) and the other on the antigen (r_{ag}) are selected. Those interacting residue pairs (r_{ab},r_{ag}) that belong to the initial constraints (C_{ab}, C_{ag}) are then used to re-score the decoy. This is achieved by summing the precision values given in equation 5.1 for every interacting residue pair (r_{ab},r_{ag}) in the initial constraints set (C_{ab}, C_{ag}). The 200 decoys are then reordered using this score. An analogous procedure was carried out for PatchDock.

5.2.3 Evaluating docking performance

5.2.3.1 Capri criteria for classifying docking decoys

Docking decoys are classified according to the CAPRI criteria. There are three quality classes: low (*), medium (**) and high (***) quality. Decoys that do not qualify for any of the classes are considered to be incorrect. Classification of a given decoy to any of those groups relies on three factors: f_{nat}, I_{RMSD} and L_{RMSD}. The first one, f_{nat} is the ratio of the native interface pairs that are recreated in the decoy complex, taken at a 5Å cutoff. The other two are the root mean square deviation (RMSD) measures of how well the heavy atoms of the decoy ligand (in the case of the Ab-Ag complexes, the antigen) is superimposed with respect to the native complex. This obviously penalizes the flexibility effects of binding in the case of rigid body docking that cannot be rectified by the algorithm. The L_{RMSD} measure superimposes the receptors of the native structure with this of the decoy and calculates the RMSD of the ligands. Since this measure does not give justice to different size antigens, another, more permissive measure was introduced: I_{RMSD}. In this second RMSD measure, the interface residues (¡10.0Å apart between antibody and antigen) are superimposed and their backbone heavy-atom RMSDs are calculated. Table 5.1 shows how f_{nat}, I_{RMSD} and L_{RMSD} are used to classify every decoy.

Score	Conditions
* (low quality)	$((f_{nat} > 0.1$ AND $f_{nat} < 0.3)$AND $((L_{rmsd} < 10)$ OR $(I_{rmsd} < 4)))$OR$((f_{nat} > 0.3)$ AND $((L_{rmsd} > 5)AND(irmsd > 2)))$
** (medium quality)	$((f_{nat} > 0.3$ AND$f_{nat} < 0.5)$AND$((L_{rmsd} < 5)$OR$(I_{rmsd} < 2)))$ OR $((f_{nat} > 0.5)A((L_{rmsd} > 1)AND(I_{rmsd} > 1)))$
*** (good quality)	$f_{nat} > 0.5$ AND $L_{RMSD} < 1$Å or $I_{RMSD} < 1$Å

TABLE 5.1: Evaluation criteria for docking decoys according to CAPRI.

5.2.3.2 Scoring individual decoys

Decoy quality was evaluated according to the CAPRI criteria (Mendez et al. [2005]). Four parameters were calculated for each decoy: f_{nat}, I_{RMSD}, L_{RMSD} and n_{clash}. f_{nat} is the fraction of the native pairwise contacts re-established by the docking algorithm as compared with the native complex. I_{RMSD} and L_{RMSD} estimate the quality of the overall structural fit of the decoy with respect to the native structure. In both cases the structure of the antibody is superimposed on that of the native structure. After the superposition, L_{RMSD} is calculated as the $RMSD$ of the heavy atoms of the native and decoy ligands. I_{RMSD} is the $RMSD$ of the optimally superimposed interfaces, being all the residues within 10Å from the contacting residues. Finally, n_{clash} is the number of Ab-Ag residue pairs that are less than 3Å away from each other.

We discard decoys whose number of clashes n_{clash} exceeds $\mu + 2\sigma$, where μ is the mean of n_{clash} over decoys produced for the given target and σ is the corresponding standard deviation. Quality of each decoy is assigned one of four classifications: incorrect, one star (*) for low quality decoys, two stars for medium quality and three stars for good quality.

Since it was computationally feasible to perform multiple runs of ZDOCK on the target set, we have sampled the results for each combination of inputs over 20 runs. The results for each run were in the form of vectors with the CAPRI ratings e.g (1,2,3) corresponding to 1*, 2** and 3***. A result for each input combination consists therefore of the average of the result vectors over 20 runs. Since the multiple runs do not affect the results of PatchDock, only a single run was performed for each target.

5.3 Results

5.3.1 Antibody Antigen Docking

We chose two fast rigid-body docking algorithms for our analysis: ZDOCK and PatchDock. The aim here is to generate a very rapid, relatively accurate local docking pipeline for antibody antigen complexes. We have tested the performance of our method on a large set of crystal-structures, NR-subset (Figure 5.3) and homology models, SnugDock-H (Figure 5.4). The dataset NR-subset consists of 30 structures from NR-full, described in the previous Chapter. Dataset SnugDock-H consisted of the 15 homology model targets used in the SnugDock analysis (Sircar and Gray [2010]). For each target in the datasets NR-subset and SnugDock-H, we used two docking constraints, one for the antibody and one for the antigen. Each constraint consisted of a set of residues which describe the binding site to a greater or lesser degree.

The constraint for the antigen consisted of the actual binding residues together with residues on-the antigen within 5Å. In the case of the antibody we tested three different constraints: C-Native, C-CDR and C-Antibody i-Patch. The first, C-Native, consisted of the correct binding site. The second, C-CDR, consisted of the IMGT-CDR residues. The last set, C-Antibody i-Patch consisted of the binding residues predicted by Antibody i-Patch with a score cut-off greater than 40. This Antibody i-Patch score cut-off was chosen as it gave an acceptable trade off between the values of recall and precision. According to our results on dataset NR-full in the previous Chapter, it gave an average precision of 45% and an average recall of 90%. In the case of the 30 targets in the crystal-structure docking test set (NR-subset) this cut-off of 40 gives precision of 46% and recall of 91%. In the case of the 15 targets in the docking test set (SnugDock-H) this cut-off of 40 gives precision of 47% and recall of 88%.

Figure 5.3 shows the results of running ZDOCK with different antibody constraints on the crystal structure test dataset NR-subset. The docking results are evaluated using the CAPRI criteria (Mendez et al. [2005]). Each decoy is placed in one of four categories

from zero stars up to three stars, with three stars denoting higher quality results and the best result among the top 10 decoys is reported for each target. As ZDOCK is non deterministic, but very rapid, we decided to check whether our results were consistent across multiple runs. For each test case in the set we sampled the results from 20 runs, thus the results shown in Figures 5.3 and 5.4 are the averages of these runs.

The Antibody i-Patch predictions and our antibody-antigen specific scoring system compared to the general protein score of the docking software enriches the top 10 predictions with useful decoys (Figure 5.3). Specifically, using our docking pipeline we achieve better results than using native contacts as constraints for the docking algorithm. We argue that this is due to the fact that the docking algorithms like ZDOCK and PatchDock are trained on general proteins datasets which have distinctly different binding profiles from those of antibodies, as shown in the two previous Chapters. These results demonstrate that our antibody-specific docking pipeline consistently selects better decoys, approaching the maximum possible score given in the top 200 results of ZDOCK.

On the 30 crystal structures in the NR-subset even given native contacts, ZDOCK only achieves a CAPRI vector of (6.5***, 10.2**, 5.2*), whereas using the Antibody i-Patch constraints and scoring system a vector of (8***, 14.8**, 5.1*) is given (Figure 5.3).

Thus using the antibody specific scoring improves our ability to identify correct docking poses. In fact, most of the good quality docks in the top 200 results are sorted into the top 10 by the antibody specific scoring system (Figure 5.3).

Figure 5.4 shows the docking results on antibody homology models (dataset SnugDock-H). Once again reasonable results are found in the top 200 for any of the constraints. However it is only in the antibody specific procedure that these decoys are placed in the top 10.

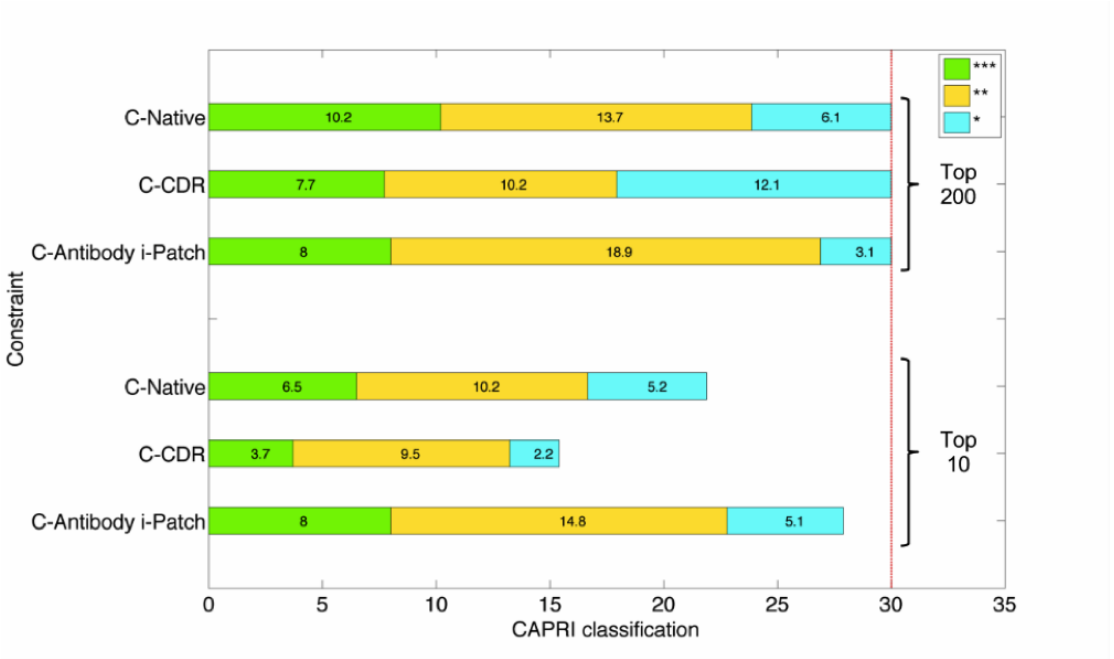


FIGURE 5.3: **Results from running our docking pipeline on the test dataset of crystal structures docking targets NR-subset.** : The results are the averages of the three element vectors given by the CAPRI classification into three quality groups: satisfying (*), medium (**) or good (***) quality. The individual three element vectors were collected over 20 runs of the pipeline on the dataset NR-subset. The standard errors are given in B.3.

5.3.2 Comparing the Antibody i-Patch rigid docking pipeline to other methods

The results of our pipeline are not as good as those obtained by more complex docking procedures such as SnugDock and EnsembleDock, but are similar to those of standard RosettaDock. For example, as shown in Figure 5.4, the top ten results for the Antibody i-Patch procedure on the 15 homology model test cases lead to a CAPRI vector of (0.7*** 3.1** 7.0*) whilst Ensemble Dock with SnugDock on the same set has a CAPRI vector of (0*** 9** 5*) and Standard RosettaDock achieves (0*** 5** 6*) (Sircar and Gray [2010]). In general the flexibility in the more complex methods allows them to generate higher quality results (e.g. move decoys from being of * quality up to **) something which cannot be achieved in a rigid docking scenario.

Another comparator between the methods however is the time taken to achieve these

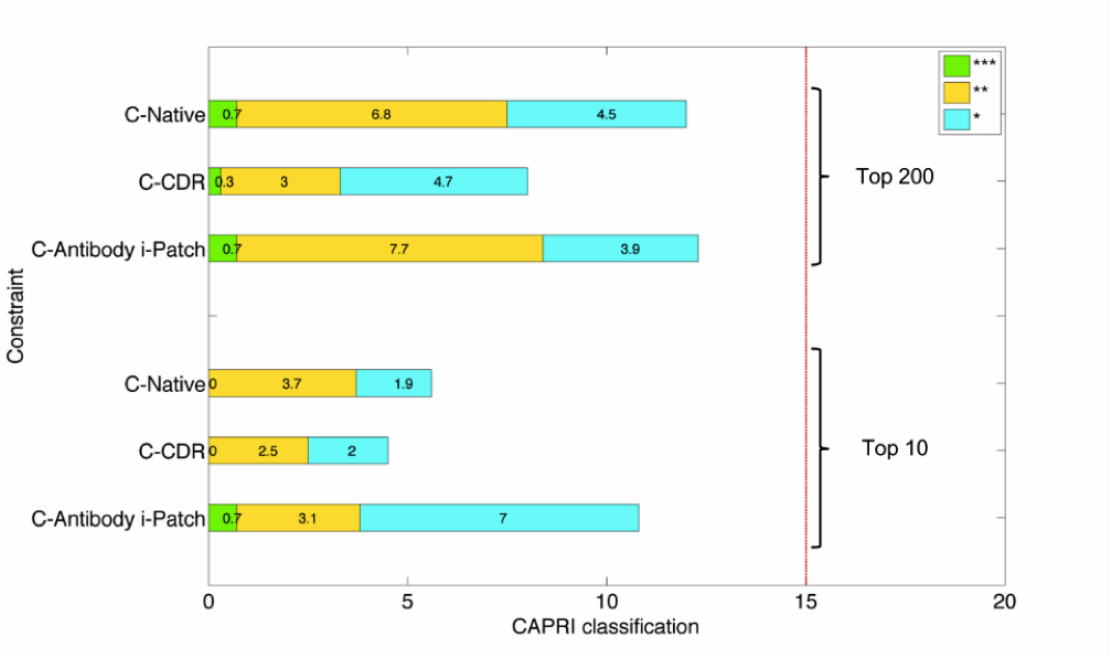


FIGURE 5.4: **Results from running our docking pipeline on the test dataset of homology models docking targets SnugDock-H.** : The results are the averages of the three element vectors given by the CAPRI classification into three quality groups: satisfying (*), medium (**) or good (***) quality. The individual three element vectors were collected over 20 runs of the pipeline on the dataset SnugDock-H. The standard errors are given in B.3.

results. The Antibody i-Patch ZDOCK pipeline can produce results for our targets in minutes rather than the hundreds of CPU hours needed for the flexible methods (Pierce et al. [2011], Sircar and Gray [2010]). Therefore, our pipeline could provide good initial poses at little computational cost for ensemble or flexible docking methods such as SnugDock for further refinement. Thus, the method could be used where speed is critical to tackling the problem, for example, all against all antibody/antigen screening.

5.4 Conclusion

In this Chapter we have demonstrated how using Antibody i-Patch constraints together with antibody-antigen specific scoring scheme can be used to achieve satisfying results for local antibody-antigen docking. We have demonstrated that our pipeline can achieve results comparable to the current state of art flexible dockers but much faster. Since

our software produces satisfying results for homology models, we argue that it might be applicable to rapidly provide an initial set of poses in a virtual screening campaign where many variants of a single antibody might be modelled and docked to a specific epitope.

Even though the epitope information might be available in some cases, local antibody-antigen docking remains a special case of the global docking. For this reason, using expertise developed here, we tackle the global antibody-antigen docking in the next Chapter.

Chapter 6

Global Antibody-Antigen Docking

6.1 Introduction

Local antibody-antigen docking work presented in the previous Chapter is a special case of global protein-protein docking. Building on the knowledge and tools developed in the previous Chapter we attempted this more difficult problem. The global problem is currently unattainable for most docking programs (Sircar and Gray [2010], Brenke et al. [2012]). For this reason, our approach to global antibody-antigen docking is by employing B-cell epitope prediction software (Ponomarenko and Bourne [2007], Kringelum et al. [2012]).

Given a sequence or structure of an antigen, in silico B-cell epitope prediction aims to identify a set of residues on the antigen capable of binding an antibody (Kringelum et al. [2012]). Computational identification of B-cell epitopes can provide an initial set

of potential binding sites on the antigen which can elucidate the immunogenicity of the molecule in question (Idrees and Ashfaq [2013], Gershoni et al. [2007], Irving et al. [2001]). The majority of the methods operate without antibody information, aiming to identify all potential antibody binding sites (Sela-Culang et al. [2013], Kuroda et al. [2012]). However attempting to map all epitopes might not be optimal since some antigens, like hen egg white lysozyme, are capable of forming complexes with many different antibodies meaning that most of its surface constitutes a part of some epitope (Sela-Culang et al. [2013]). In this paper we create antibody-specific epitope predictions as we believe these might be of better use for the development of therapeutic antibodies (Zhao and Li [2010], Zhao et al. [2011], Soga et al. [2010]).

Computational B-cell epitope prediction provides information about the immunogenic regions of the antigen but it does not directly contribute to the knowledge of the particular antibody residues that need to be mutated so as to modify its function. Such predictions however can provide insight into the pairwise contacts between the antibody and antigen when combined with protein docking. It has been demonstrated that combining epitope prediction and docking for the corresponding problem concerning T-cells, improved the prediction quality of T-cell epitopes (Zhang [2013]).

It was argued in the previous Chapter that antibody-antigen docking requires different methodology than that employed for the corresponding problem concerning non-antibody targets. This is because antibodies use very different residues in their binding sites when compared to both general proteins and to antigens and thus an asymmetric scoring system is required which accounts for these discrepancies (Krawczyk et al. [2013], Brenke et al. [2012]).

In this manuscript we focus on epitope prediction and global docking and how those two methods in concert can facilitate computational artificial antibody design. We develop an antibody-specific epitope prediction method EpiPred, which uses geometric matching of the antibody and antigen interfaces coupled with an antibody-antigen specific knowledge-based potential. We build on the fact that structural information is crucial in identification of non-linear epitopes (Kringelum et al. [2012]). We use our epitope

predictions to re-score the global docking results of two fast rigid body docking algorithms, ZDOCK and ClusPro in antibody mode (Chen et al. [2003], Brenke et al. [2012]). We demonstrate that including the epitope information in our global docking pipeline enriches the top decoys with more close to native poses.

6.2 Materials and Methods

6.2.1 Data

A non-redundant dataset of crystal structures (X-dataset) was downloaded from the Structural Antibody Database (SAbDAb) (Dunbar et al. [2013b]) in August 2013. The complexes were selected such that no two antibodies shared more than 99% sequence identity and no part of antigens shared more than 90%. All antigens were proteins and the complexes had to be of resolution 3Å or better. The final dataset consisted of 149 structures (X-dataset), 30 of which were chosen at random to consist the test set, referred to as X-test. The PDB codes and the corresponding chains of structures used in this study are given in Appendix C.1

In order to evaluate our pipeline on homology models, we re-use the homology dataset SnugDock-H from the previous Chapter.

6.2.2 Epitope prediction

Our epitope prediction algorithm is a combination of geometric fitting and a knowledge based, asymmetric antibody-antigen scoring. The algorithm is divided into three steps.

Firstly, epitope-like surface patches on the antigen are enumerated. These are designed to be roughly the same size as the approximate epitopes used in our earlier local docking study (Chapter 5). These candidate epitope patches are then scored using geometric

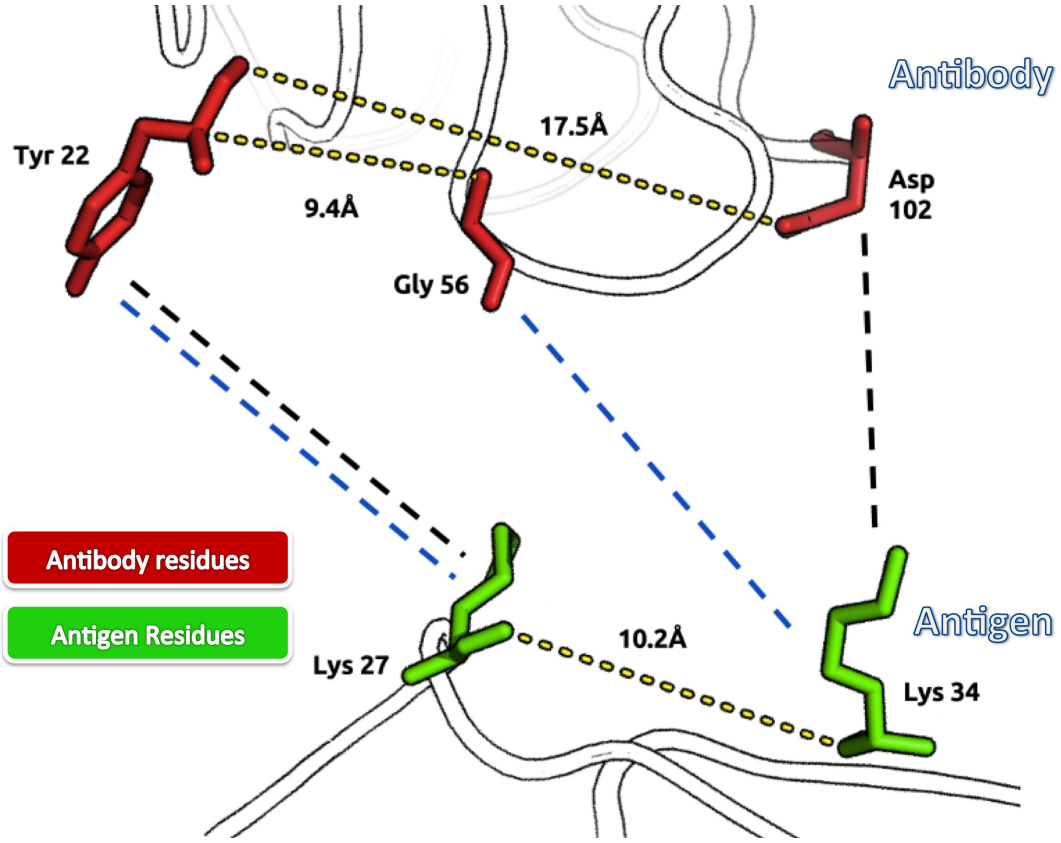


FIGURE 6.1: Example of a case when intra-molecular distances can provide information about which inter-molecular contacts can exist. The antibody-antigen contacts between Tyr-22 and Lys-27 and Gly-56 and Lys-34 (blue dashes) can exist as the intra-molecular distance between Tyr-22 and Gly-56 is 9.4Å and the distance between the two Lys residues is 10.2Å . The difference between those two intra-molecular distances is 0.8Å which is below the cut-off of 1Å. As a counterexample, the contacts between Tyr-22 and Lys-27 and Asp-102 and Lys-34 (black dashes) cannot be satisfied simultaneously since the intermolecular distance between Tyr-22 and Asp-102 is 17.5Å.

fitting and a specific antibody-antigen score. The geometric fit is calculated by enumerating all possible contacts between the set of putative epitope residues and the CDRs and evaluating which pairs of antibody-antigen contacts can be satisfied simultaneously (see Figure 6.1 for an example). The final epitope score for each patch is a sum of all possible contacts between the given epitope and CDRs, weighted by the number of other contacts they can satisfy simultaneously as well as the antibody-antigen Precision Score for the particular amino acid contact pair. The Precision Score has been adapted from our earlier work on local antibody antigen docking presented in the previous Chapter. In the final ranking of the candidate epitopes we only keep patches with less than 30% of their residues in common.

6.2.2.1 Sampling the patches on the antigen surface

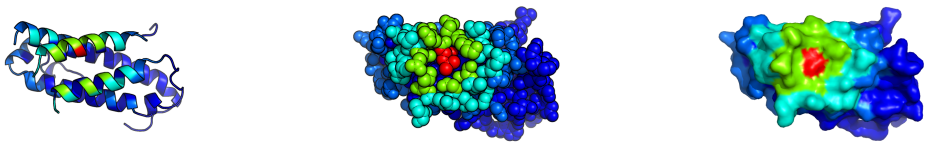


FIGURE 6.2: The visualization shows a single candidate patch on the surface, presented in three ways (cartoon, spheres and surface). This exemplifies the depth sampling used in this work: 4.5Å cutoff and depth of 3. The red residue is the central amino acid which initiates the patch (depth = 1). The green residues correspond to the neighborhood of the first residue (depth = 2). The teal residues are those within 4.5Å from the residues in green (depth = 3).

Our method of sampling candidate surface patches consisted of extending each surface exposed residue on the antigen with the surface neighbourhood. A surface neighbourhood is created by successive additions of the surface residues within a certain cut-off distance to those already in the patch. Thus there are two parameters in this procedure: neighbour cutoff and the number of iterations of extending the neighbourhood (depth). See Figure 6.2 for an example.

We have estimated the best parameter configuration for the patch sampling on our training set consisting of crystal structures (X-dataset excluding X-test). Using several values for these parameters, we have looked at average precision and recalls of the best, top five and top ten sampled patches given in Tables 6.1, 6.2 and 6.3 respectively.

We concluded that average precision in the region of 50% and recall in the region of 80% was achieved for the cut-off distance of 4.5Å and depth 3 which are used as standard parameters in this work.

Neighbuor cutoff (Å)	Depth	Average Precision (%)	Average Recall (%)
2.0	2	97	26
2.0	3	92	32
2.5	2	96	27
2.5	3	91	34
3.0	2	87	37
3.0	3	79	48
3.5	2	79	52
3.5	3	67	65
4.0	2	74	61
4.0	3	59	75
4.5	2	70	67
4.5	**3**	**52**	**80**

TABLE 6.1: Best sampled patches: the bast patch on X-dataset minus X-test.

Neighbour cutoff (Å)	Depth	Average Precision (%)	Average Recall (%)
2.0	2	92	23
2.0	3	87	29
2.5	2	91	24
2.5	3	86	31
3.0	2	84	33
3.0	3	76	43
3.5	2	75	47
3.5	3	63	62
4.0	2	70	56
4.0	3	54	73
4.5	2	66	62
4.5	**3**	**48**	**79**

TABLE 6.2: Best sampled patches: averages of 5 best on X-dataset minus X-test.

Neighbour cutoff (Å)	Depth	Average Precision (%)	Average Recall (%)
2.0	2	86	20
2.0	3	83	25
2.5	2	85	21
2.5	3	81	27
3.0	2	79	29
3.0	3	72	39
3.5	2	71	42
3.5	3	59	59
4.0	2	64	52
4.0	3	50	71
4.5	2	60	58
4.5	**3**	**45**	**78**

TABLE 6.3: Best sampled patches: averages of 10 best on X-dataset minus X-test.

6.2.2.2 Precision score for the epitope prediction

The Precision Score $Pr(T_{ab}, T_{ag})$ was calculated in an identical manner to that in the previous Chapter. Thus $Pr(T_{ab}, T_{ag})$ denotes the likelihood of the docking algorithm to correctly pair a residue of type T_{ab} on the antibody and a residue of type T_{ag} on the antigen (for instance glycine on the antibody and serine on the antigen). The precision score $Pr(T_{ab}, T_{ag})$ was estimated by executing ZDOCK on each of the 118 targets in X-dataset that were not in X-test and counting how many times a given pair of residues was matched correctly with respect to the native structures.

In order to ensure we have not over-trained the precision score for the SnugDock-H dataset, we have removed all members of the X-dataset that had more than 90% sequence identity with any antigen and more than 99% with any antibody in the SnugDock-H. The sequence identity was calculated using CD-HIT (Li and Godzik [2006]).

6.2.2.3 Scoring putative epitopes

Let *Epi* denote the set of residues in a putative epitope and *Ab* the set of residues supplied as the binding site on the antibody. We create a graph G where each node n, corresponds to an element of the cartesian product of *Epi* and *Ab*: $Epi \times Ab$. Thus if there was a tyrosine (Y) residue in *Epi* and a histidine (H) residue in *Ab*, there will be

a node n' in G which corresponds to this pair - (Y, H). Each of the nodes represents a possible inter-molecular contact between antibody and antigen residues.

We add an edge between any two nodes in G if the antibody-antigen contacts defined by those nodes can be geometrically satisfied at the same time. Take node n_1 which stands for a contact between antibody residue r_{ab1} and antigen residue r_{ag1} and node n_2 with antibody residue r_{ab2} and antigen residue r_{ag2}. Define $dist(r_{ab1}, r_{ab2})$ as the intra-molecular distance between the two residues r_{ab1} and r_{ab2}. We place an edge between n_1 and n_2 only if the difference in intra-molecular distances on the antibody and the antigen is below 1Å cut-off as given by 6.1.

$$|dist(r_{ab1}, r_{ab2}) - dist(r_{ag1}, r_{ag2})| < 1\text{Å} \tag{6.1}$$

Let $d(n)$ denote the degree of node n. The final score for a putative epitope Epi is given by 6.2.

$$EpitopeScore(Epi, Ab) = \sum_{n \in G} d(n) Pr(T_{ab}, T_{ag}) \tag{6.2}$$

where T_{ab} and T_{ag} are the amino acid types of the antibody and antigen residues respectively which belong to node n.

The epitopes are ordered by their score and the top three non-overlapping epitopes are kept. Overlapping epitopes are defined as those which share more than 30% of the same residues with respect to the epitope with the higher epitope score.

We use our epitope prediction algorithm on each of the targets in X-test and SnugDock-H.

6.2.3 Global Docking

The global docking pipeline we have developed is divided into three steps. Firstly, up to three candidate epitope predictions from EpiPred are computed. Secondly, we perform global docking using a fast rigid body algorithm (ZDOCK or ClusPro). We do not provide any epitope information at this point, only supplying the CDR residues to be masked. The final step consists of re-scoring the poses produced by the docking algorithms using the Precision Score.

The input to the Precision Score consists of a single antibody-antigen pose supplied by the docking algorithm, the set of Chothia CDR residues and a set of residues for one of the predicted epitopes. For each pose, the Precision Score is computed for each of the top three predicted epitopes as given by EpiPred. The final score of a pose is the highest of the three scores. The poses for a given target are then re-ranked by this score.

6.2.3.1 Docking algorithms

ZDOCK was run on all of the targets in the X-test set with the constraint on the antibody of the Chothia CDRs and with no epitope information. The software was executed using its default parameters, with the exception that the number of poses to probe was set to 10000. As it was computationally feasible we performed 5 runs of ZDOCK for each target, using different random seeds each time. An analogous procedure was applied to the targets in SnugDock-H.

The targets in X-test and SnugDock-H were also submitted to the ClusPro server in antibody mode, using automatic CDR masking.

6.2.3.2 Re-scoring decoys

Consider a set of decoys D returned by either ZDOCK or ClusPro for a given target. We collect the top N decoys from D as ordered by the docking method. For a given

decoy d in the set of top N decoys from D, let Ab denote the set of residues used as the antibody constraint and Epi a set of predicted epitope residues. Let (r_{ab}, r_{ag}) be any pair or residues in d, where $r_{ab} \in Ab$ and $r_{ag} \in Epi$. If the distance between r_{ab} and r_{ag} is observed to be less than 4.5Å in the decoy d, this pair of residues contributes the value of $Pr(T_{ab}, T_{ag})$ to the score for this decoy where T_{ab} is the type of the amino acid type of r_{ab} and T_{ag} is the amino acid type of r_{ag}. If we let $dist(r_{ab}, r_{ag})$ denote the distance in angstroms between residues r_{ab} and r_{ag}, the score for decoy d using antibody constraint Ab and epitope prediction Epi can be formalized by 6.3.

$$DecoyScore(d) = \sum_{\substack{r_{ab} \in Ab \\ r_{ag} \in Epi \\ dist(r_{ab}, r_{ag}) < 4.5\text{Å}}} Pr(T_{ab}, T_{ag}) \tag{6.3}$$

The top N decoys for a given target are given scores using our three epitope predictions. For each decoy, we retain the highest score out of the three. We then use those scores to re-order the top N decoys for a given target.

In the case of ZDOCK for both X-test and SnugDock-H, we re-score the top 30 decoys for each target. We use the top 20 predictions for ClusPro as this is the maximum number of decoys returned by ClusPro in most cases.

6.2.3.3 Evaluation criteria for docking

In order to evaluate the quality of each decoy we use the the interfacial root mean square deviation (I_{rmsd}), one of the metrics used in the CAPRI experiment (Mendez et al. [2005]). The value of I_{rmsd} is the root mean square deviation between the interface region of the decoy and the native structure when those regions are optimally superimposed. The interface regions are defined as those within neighborhood of 10Å from any residue on the binding partner.

We define a close to native decoy in the same way as the authors of the ClusPro antibody study (Brenke et al. [2012]). A close to native decoy is defined as having I_{rmsd} less than 10Å from the native complex. For each target in our test sets, we have the raw list of top N decoys as ordered by the docking algorithm and our re-scored version thereof. So as to evaluate which ordering is better we count the number of close to native decoys in the top one, top five and top ten entries in the raw and re-scored lists. For instance if using our re-scored list one finds three close to native decoys in the top five and only one in the top five decoys in the raw list, we consider the re-scoring to have improved the result. If in the top N of both lists no close to native decoy is found we state there was no suitable decoy.

Since multiple runs were performed for ZDOCK, the reported results are derived from the comparison of the averages of close to native decoys in the raw and re-scored lists for each target.

6.2.4 Blind test case

The sequence provided by UCB Pharma was modeled using PIGS (Marcatili et al. [2008]). Three EpiPred epitope predictions were obtained for the antigen using the identical configuration as described in the previous sections. The homology model of the antibody together with the crystal structure of the antigen were submitted to ClusPro using the same configuration as previously: run in antibody mode with CDR masking. Chothia CDR residues were identified by numbering the sequence of the antibody using Abnum (Abhinandan K R [2008]) and extracting the CDR regions according to their definition. Those annotations were inspected manually in the model structure with minor adjustments. Re-scoring of the decoys obtained from ClusPro was performed in an identical fashion as described in the previous sections.

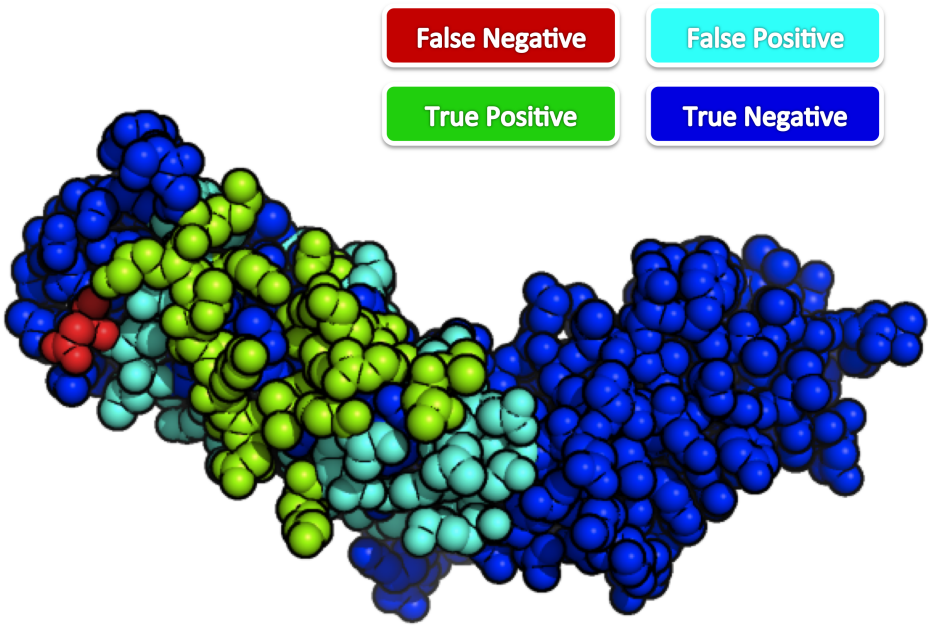


FIGURE 6.3: The top epitope prediction for the antigen 1boy (human tissue factor, the unbound form of the antigen complexed in 1ahw in SnugDock-H). The prediction consists of a set of residues which are considered to constitute the general area of the epitope. The true positives are shown in green, false positives in teal, false negatives in red and true negatives in dark blue. This prediction achieved 36% precision and 94% recall. (The target comes from the dataset SnugDock-H, thus the antibody used in the prediction was a homology model and the corresponding antigen was in the unbound form).

6.3 Results

6.3.1 Epitope prediction

The epitope prediction algorithm presented here was inspired by our earlier work on local antibody-antigen docking presented in the previous Chapter where we showed that it was possible to select close to native decoys when docking the antibody into an approximate region of the epitope.

In order to extend our local docking methodology to global docking, we have developed an epitope prediction algorithm that identifies surface patches on the antigen similar to the approximate epitopes used in our earlier work (see Figure 6.3). Our epitope

		Epitope Prediction					
		One		Two		Three	
PDB	Ag size	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
4hj0	92	32	90	0	0	-	-
1tzh	94	1	6	72	81	-	-
4am0	96	13	70	0	0	-	-
2ih3	97	16	64	0	0	-	-
4i77	97	23	55	31	50	-	-
3q1s	113	19	81	-	-	-	-
1p2c	129	0	0	34	87	-	-
4ht1	131	5	14	50	95	-	-
3ab0	136	33	73	41	73	0	0
1v7m	145	26	77	5	11	6	11
4g3y	148	3	8	51	82	0	0
2vxt	156	4	9	45	95	7	14
3u9p	169	31	100	0	0	0	0
3o2d	178	32	64	12	20	10	16
1fns	196	0	0	0	0	8	27
3ma9	197	0	0	12	27	11	22
3rvv	223	25	93	2	6	7	20
3raj	230	0	0	19	75	15	50
1nfd	239	7	23	0	0	0	0
3i50	273	0	0	0	0	25	100
3gjf	276	15	66	0	0	12	27
3liz	329	26	68	0	0	34	78
3pgf	358	2	4	8	18	11	22
3zkm	375	32	88	0	0	0	0
3r1g	381	37	100	0	0	0	0
4jr9	409	19	85	0	0	0	0
4ene	442	0	0	0	0	0	0
3o0r	449	8	70	0	0	0	0
3t3p	453	0	0	2	5	0	0
1n8z	581	0	0	0	0	0	0

TABLE 6.4: Table summarizing the results of epitope prediction on the X-test set. We present the top three epitope predictions returned by EpiPred. For smaller antigens only one or two epitope predictions may be returned as only epitope predictions that share less than 30% overlap are considered. In those cases a dash (-) is shown in place of precision and recall. Precision and recall were computed by the following formulas: $precision = TP/(TP + FP)$, $recall = TP/(TP + FN)$ where TP stands for true positives, FP for false positives and FN for false negatives.

prediction algorithm receives as input the structure of an antibody and an antigen and returns a ranked list of epitope-like regions.

In our case the aim is to generate epitope predictions specific for a given antibody in order to facilitate docking. Thus, the use of antibody information is crucial. However, as shown below, the antibody structure can be a homology model.

		Epitope Prediction					
		One		Two		Three	
PDB	Ag size	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
2bdn (B)	68	21	52	29	41	-	-
2jel (U)	82	25	92	-	-	-	-
1k4c (U)	97	12	37	0	0	44	75
1ztx (B)	101	4	11	23	58	25	52
1wej (U)	101	10	66	-	-	-	-
1jhl (U)	126	11	40	29	93	0	0
1mlc (U)	126	29	82	20	52	0	0
1bql (U)	126	10	29	27	70	0	0
1vfb (U)	126	18	38	5	9	24	42
2b2x (B)	188	0	0	21	47	25	52
1jps (U)	200	22	59	20	45	0	0
1ahw (U)	200	36	94	10	26	28	68
1ynt (B)	252	0	0	6	13	29	52
2aep (B)	358	8	18	11	22	0	0
1nca (U)	388	42	84	38	68	0	0

TABLE 6.5: Table summarizing the results of epitope prediction on the H-test set. Letters next to the PDB codes indicate whether the antigen used was bound (B) or unbound (U). We present the top three epitope predictions returned by EpiPred. For smaller antigens only one or two epitope predictions may be returned as only epitope predictions that share less than 30% overlap are considered. In those cases a dash (-) is shown in place of precision and recall. Precision and recall were computed by the following formulas: $precision = TP/(TP + FP)$, $recall = TP/(TP + FN)$ where TP stands for true positives, FP for false positives and FN for false negatives.

6.3.1.1 Evaluation of the performance of epitope prediction

In order for our epitope prediction method to be applicable in virtual screening, it must be able to produce results given nothing more than the sequence of the antibody and the structure of the antigen. So as to verify this claim, we have evaluated the results of our epitope prediction algorithm on two datasets: crystal structure dataset (X-test) and homology model dataset (SnugDock-H). The first dataset (X-test) consisted of 30 non-redundant solved crystal structures of antibody-antigen complexes. The second dataset (SnugDock-H, as used in the previous Chapter) consists of 15 antibody-antigen targets where the antibody is a RosettaAntibody (Sivasubramanian et al. [2009]) model with FREAD prediction for the H3 loop (Choi and Deane [2011, 2010]) and where ten out of the fifteen antigens are in the unbound form.

The crystal structure dataset, X-test, constitutes the simpler of two test sets cases where all the structures are in their bound conformations. The results on the crystal structure

dataset show the performance of the algorithm given close to perfect information for both structures and as such, serve as the contrast for the homology dataset, SnugDock-H. The homology dataset poses the actual challenge as it represents the realistic input that might be given to the algorithm in the course of virtual screening.

The results from evaluating our epitope prediction algorithm on the crystal structure dataset are presented in Table 6.4. We report the precision and recall scores for the top three epitope predictions.

In some cases (e.g. 4hj0) the selected patches cover a significant portion of the antigen surface so it is impossible to select another patch with less than 30% overlap. In these cases a dash (-) is shown in Tables 6.4 and 6.5.

As one would expect, it is easier to obtain good predictions on small antigens, as there are less candidate patches to enumerate.For example 4hj0 reports only two epitope predictions. In this case these two predictions were ordered correctly as the first one achieves 90% recall and 32% precision as opposed to 0% recall and precision for the second prediction. Similarly, one would expect a considerable drop of performance on the biggest antigens since the method needs to distinguish between many more candidate patches. In some cases, such as 3t3p and 3pgf, which have 453 and 358 residues respectively, no acceptable predictions were obtained. However, for 3liz, 3jr9, 3zkm and 3r1g which all have more than 300 residues, reasonable epitope predictions were generated, suggesting that the method retains a degree of predictive power even for larger antigens.

To give an indication of the background random distribution, we have executed EpiPred on each target in the crystal structure test set 500 times, randomizing the epitope score given to each candidate patch. For each run, we have averaged the precision and recall metrics of the top epitope of the 30 targets in X-test. The mean precision and recalls averaged over 500 random-score runs are 23% recall and 15% precision. The corresponding average values for the first epitope prediction from Table 6.4 are 45% recall and 15% precision. The recall is considerably higher using our method indicating its predictive power. Furthermore, the score correctly identifies better epitopes since the

corresponding values for the second epitope are recall 25% and precision 13% and recall 18% and precision 7% for the third epitope.

We have not performed the comparison with other B-cell epitope prediction methods since majority of those, e.g. ElliPro, DiscoTope or PEPITO (Kringelum et al. [2012]), use only the structure or sequence of the antigen without any antibody information. Comparison to those methods would thus be unsound and the few methods that address the issue of antibody-specific predictions are currently unavailable (Sun et al. [2013], Yao et al. [2013]).

The epitope prediction results are very similar for the unbound homology cases presented in Table 6.5. The average precision and recall for the top predicted epitope are 16% and 47% respectively.

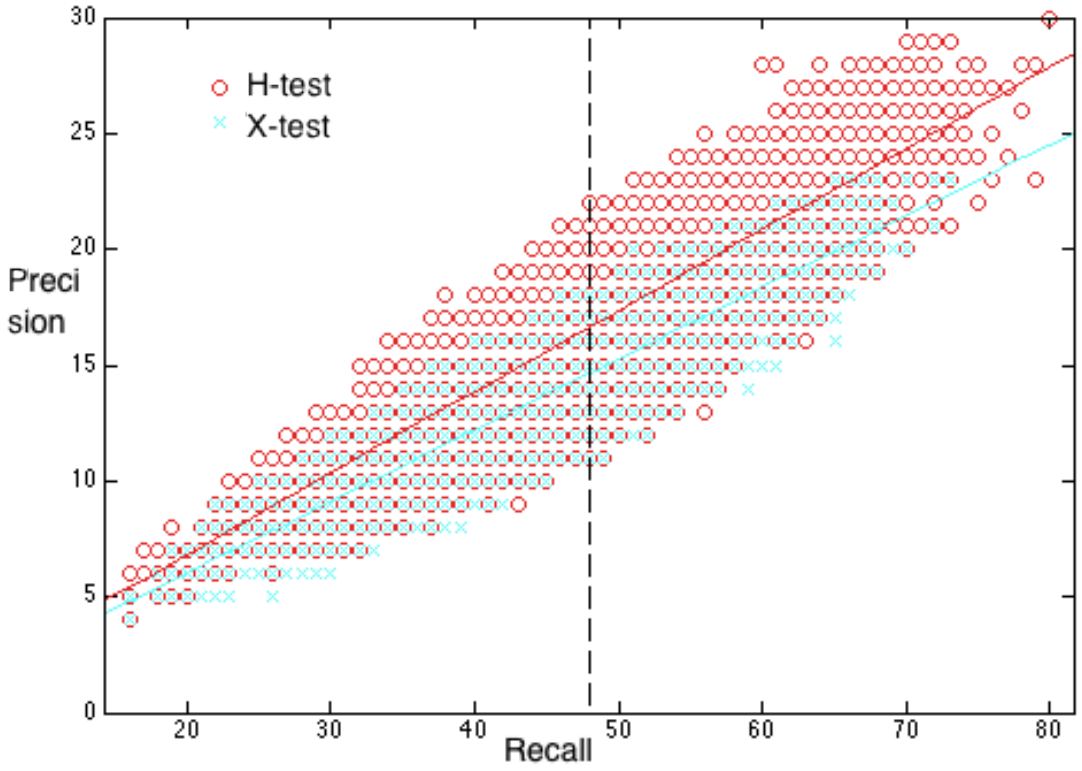


FIGURE 6.4: The plot of samples of mean precision and recall values for X-test and H-test with best-fit lines indicated.

6.3.1.2 Difference between the performance of EpiPred homology model and crystal structure datasets

In order to evaluate how different performance of EpiPred is on X-test and H-test, we have compared their sample average precisions and recalls achieved by the top predictions. In order to achieve the estimates of the average means and precision, we have sampled precision and recalls values from both datasets. For instance, one sample for the average precision and recall of X-test would consist of 30 precisions and recalls sampled at random with replacement form the top precision-recall pair values available in X-test. For each sample of 30 precision-recall pairs we have recorded the average precision and recall from those 30 pairs. Similar procedure was applied to H-test.

In total we have sampled 10^7 averages from X-test and H-test. We have fitted a line through the precision-recall pairs for the averages of X-test and H-test (see Figure 6.4). The lines plotted in this way for X-test and H-test cannot be called statistically significantly different when their slope and intercept are compared.

Since the results of EpiPred on H-test are not statistically significantly different from those achieved on the crystal structure set, we conclude that the imprecise structural information from homology models and unbound antigens does not adversely affect the method.

6.3.1.3 Evaluating the performance of specificity of predictions.

Since we have developed EpiPred to be antibody-specific, we checked if it is capable of distinguishing epitopes of different antibodies.

There are eight binding modes of antibodies to lysozyme found in the PDB (see Figure 6.5). We have picked eight representatives for each of the binding mode, motivated by resolution and low B-factor values on the antibody CDRs (see Table 6.6 for the representatives). We have re-trained EpiPred using our original set of antibody-antigen

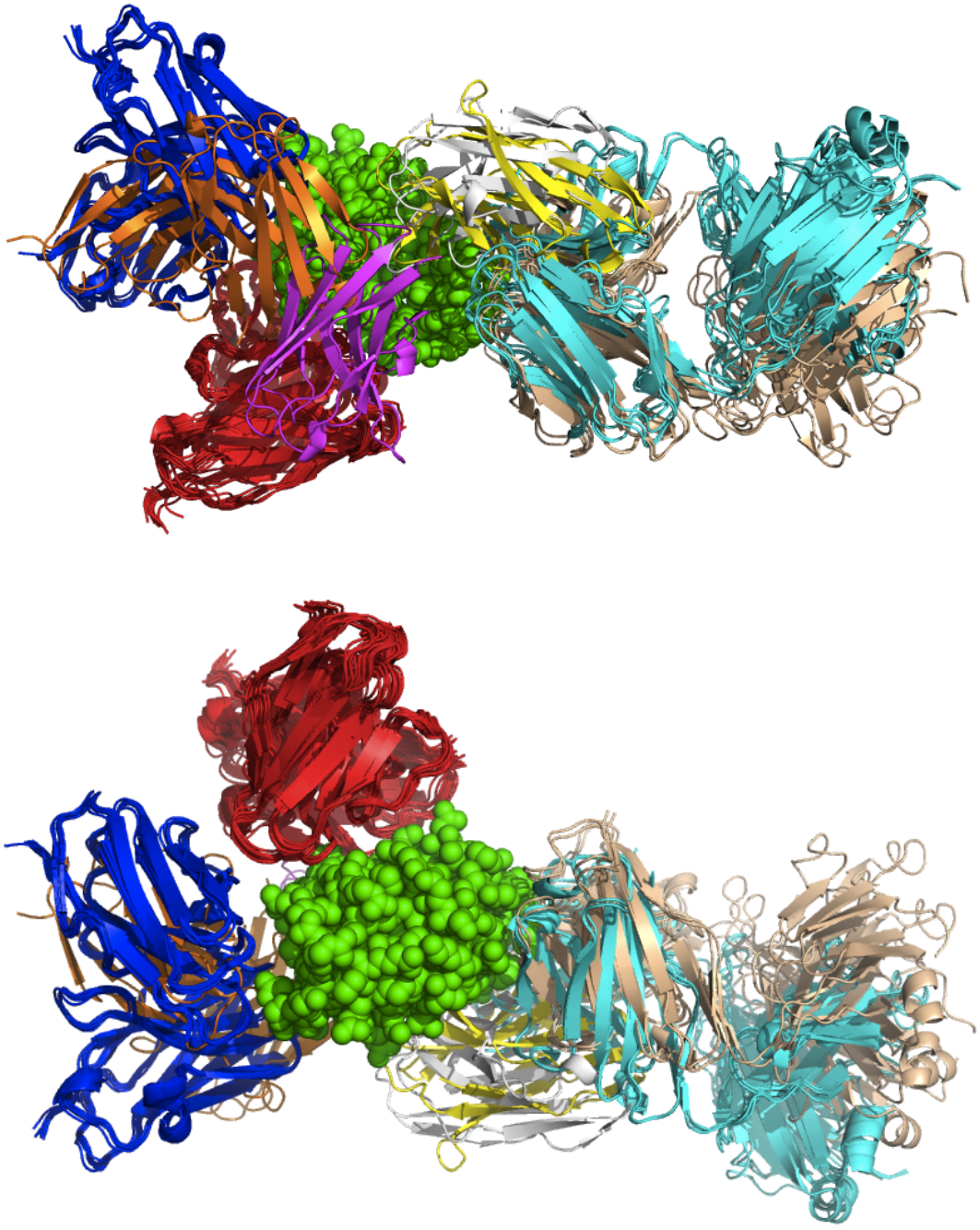


FIGURE 6.5: The 55 lysozyme-binding antibodies superimposed. The eight different binding modes/antibodies are indicated by colors. Lysozyme is the green molecule at the center. **Top:** Front face. **Bottom:** Back face.

complexes, ensuring that no antibody had more than 99% sequence identity and antigens no more than 90% sequence identity to any of the eight targets. We have than run EpiPred on each of the eight cases. We have picked the first epitope prediction for each of the eight antibody binding modes and recorded the precision and recall in each case.

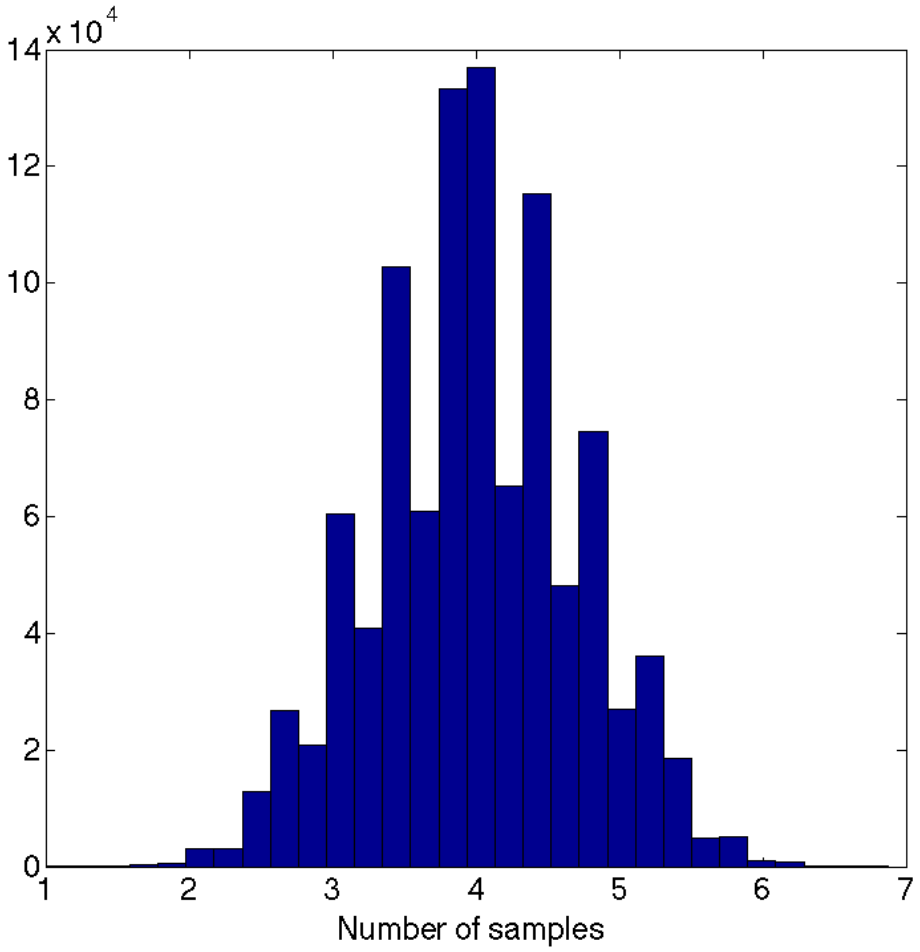


FIGURE 6.6: The distribution for the random association of epitopes sampled on the surface of lysozyme with the eight antibody binding modes. EpiPred achieves average of 2.5 epitope predictions for other antibodies being better which corresponds to a p-value of 0.0124.

For each epitope prediction returned in this way, we have also recorded the precision and recall with respect to the other seven antibody binding modes. This served as an indicator of the power of EpiPred to distinguish epitopes for different antibodies. The full results are given in Table 6.6.

If EpiPred was a perfect predictor, the best result in this case would be achieved by the diagonal entries in Table 6.6. Since this was the case only for two out of the eight predictions, we checked the ability of EpiPred to distinguish different binding modes by quantifying how many predictions that belong to other antibodies have better precision and recall than the one returned for the particular antibody. From Table 6.6, on average, 2.5 epitope predictions for other antibodies achieved better results than the one returned

PDB	1a2y	1j1x	1jhl	1p2c	1ri8	1zv5	1zvy	2iff
1a2y	44/75	33/42	25/46	0/0	0/0	0/0	11/21	0/0
1j1x	33/56	0/0	14/26	0/0	22/42	11/23	0/0	0/0
1jhl	34/50	0/0	15/27	0/0	26/42	13/23	0/0	0/0
1p2c	8/12	44/52	8/13	20/27	12/21	16/30	32/57	12/18
1ri8	30/50	0/0	11/20	0/0	23/42	11/23	0/0	0/0
1zv5	8/12	8/9	0/0	8/11	8/14	0/0	0/0	8/12
1zvy	0/0	28/38	0/0	35/55	25/50	28/61	21/42	35/62
2iff	0/0	30/38	0/0	42/61	23/42	23/46	23/42	42/68

TABLE 6.6: Precision and recall values for the top EpiPred prediction for each of the eight lysozyme binding modes calculated with respect to all eight antibodies (in the format *precision/recall*). If EpiPred was a perfect antibody-specific predictor, the best performing predictions out of each row would be on the diagonal.

for the particular antibody. We have checked how significant this result is by sampling epitope candidates on the surface at random and assigning them to either of the eight antibodies. In each case we have calculated the average number of epitopes belonging to other predictions that had better score than the one assigned to the particular antibody. Results of this sampling are given in Figure 6.6. According to this sampling procedure, the average value of 2.5 achieved by EpiPred corresponds to p-value of 0.0124 suggesting that our software has a degree of antibody-specificity in its predictions.

6.3.2 Improving global docking using epitope predictions

We have used EpiPred to constrain the results of two fast rigid body docking algorithms: ZDOCK and ClusPro (Chen et al. [2003], Brenke et al. [2012]). ZDOCK is not optimized for docking antibody-antigen complexes beyond CDR masking, but as we have shown in our earlier work, its results can be re-scored to enrich the top poses with close to native antibody-antigen complexes (Krawczyk et al. [2013]). ClusPros's global antibody-antigen docking mode has been shown to be currently the best method in this area (Brenke et al. [2012]).

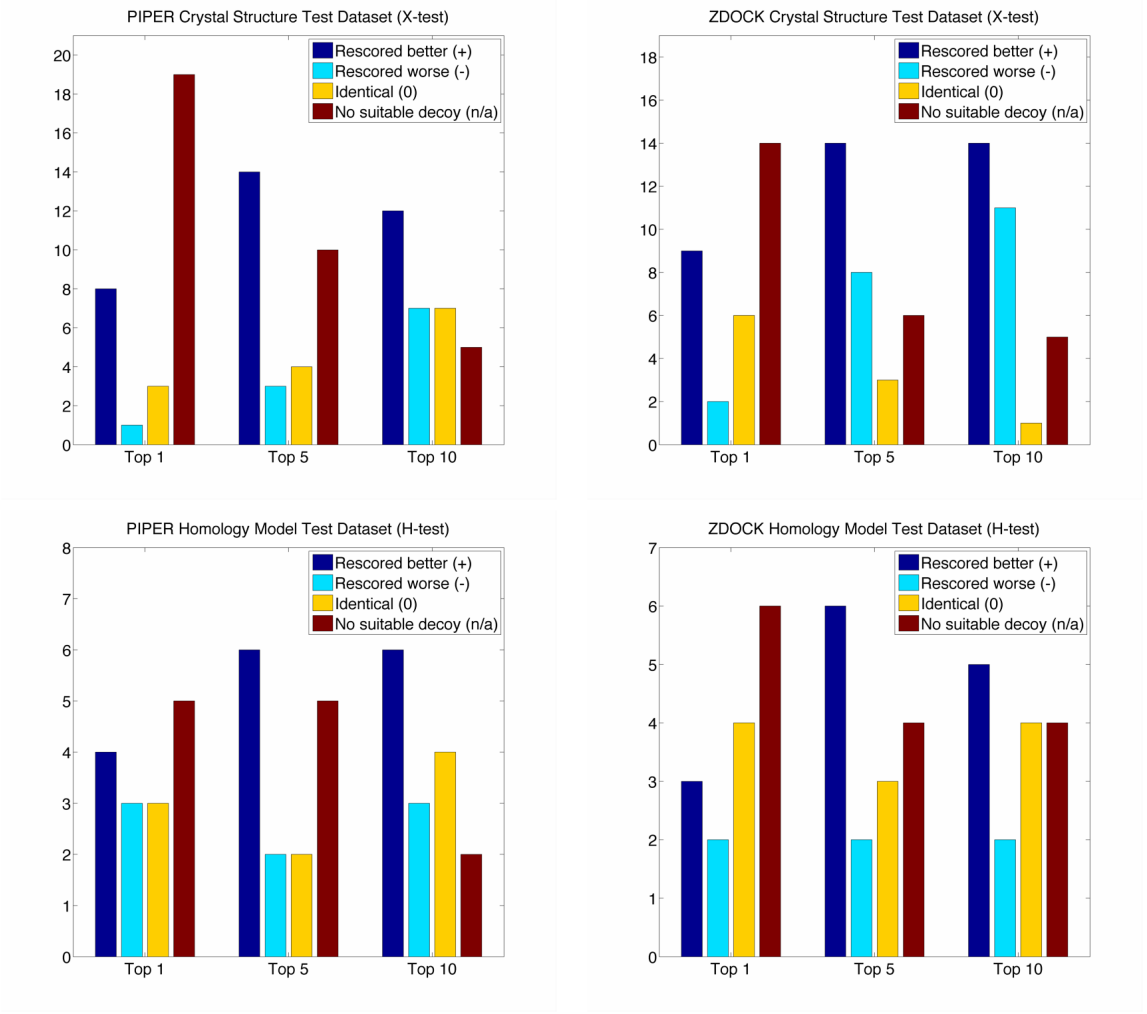


FIGURE 6.7: Success rates of re-scoring compared with the raw decoy lists given by the docking algorithms. We show results for each docking program (ZDOCK or ClusPro) on each test set (X-test or SnugDock-H) for top one, five and ten results. The rightmost bars are the number of times our global docking pipeline improved results. Bars which are second to left are the corresponding number of cases when including epitope information made the results worse. Bars which are second to right are the number of times including the epitope information did not change the raw result. The rightmost bars are the number of cases for which both procedures reported no close-to-native decoys. See supplementary information for the per-complex detailed information. **A:** Success rate of ClusPro on dataset X-test. **B:** Success rate of ZDOCK on dataset X-test. **C:** Success rate of ClusPro on dataset SnugDock-H. **D:** Success rate of ZDOCK on dataset SnugDock-H.

6.3.2.1 Evaluating the performance of our global pipeline

We have evaluated the performance of our global docking pipeline based on the criteria introduced in the ClusPro study (Brenke et al. [2012]). We call an antibody antigen pose close to native if its interfacial root mean square deviation (I_{rmsd}) is less than 10Å.

We focus on the number of near native poses found in the top N predictions, as these poses could then be passed on for further refinement by flexible methods.

In Figure 6.7 we show how our methodology improves that given by the standard docking procedures. For instance, suppose that for a target, in the top five poses as ordered by ClusPro there are three close to native decoys and in the corresponding top five results re-scored using our pipeline we obtain four close to native decoys. In such a case our global docking pipeline produced an improvement by enriching the top results with more close to native decoys. If the number of close to native decoys is zero in both lists, we state that there were no suitable decoys.

As shown in Figure 6.7 on average, including epitope information improves over the raw results obtained by either ZDOCK and ClusPro. The improvement is particularly dramatic when considering the top scoring pose in the crystal structure dataset. Our re-scoring brings eight close to native structures to the top position for ClusPro and nine for ZDOCK. This means that in almost a third of the cases, the top result is a close to native pose if re-scored using our method.

The performance of the top prediction is not as pronounced in the homology model dataset as our pipeline improves only slightly more cases than it deteriorates. The most pronounced increase in performance here is for the top five predictions where we improve upon six of the 15 cases for both ZDOCK and ClusPro. This suggests that using our method on a dataset closely resembling realistic input for virtual screening improves upon the standard algorithms.

6.3.2.2 Evaluating the performance of the EpiPred and the global docking pipeline on a blind test case.

Our collaborators from UCB Pharma provided us with a blind test case to evaluate our epitope prediction and global docking pipelines. We were given a sequence of the

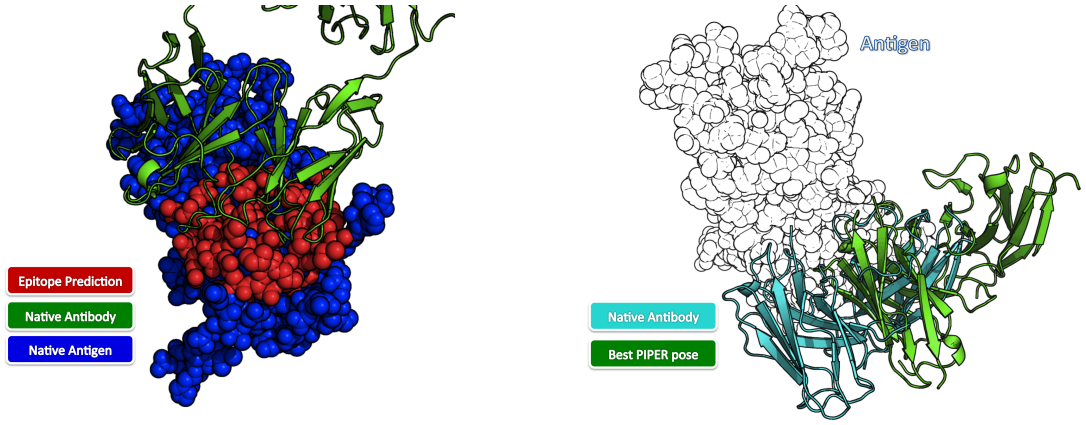


FIGURE 6.8: **Left:** The top second epitope prediction on the blind test-case is shown in red. Note that it covers the region where the actual epitope is as indicated by the native contacting antibody (in green). This epitope prediction achieved 52% precision and 69% recall. **Right:** The best decoy returned by ClusPro (green) contrasted with the native position of the antibody (teal). Notice that the antibody is rotated correctly and the discrepancy is only due to lateral translation.

antibody and the crystal structure of the antigen it forms a complex with. The antigen structure was an asymmetric homo-dimer.

For both epitope prediction and docking, we needed a structure of the antibody. This structure was modeled using PIGS (Marcatili et al. [2008]). We have not used RosettaAntibody (Sivasubramanian et al. [2009]) due to its unavailability at the time. We were unable to model H3 using FREAD (Choi and Deane [2010]) as it produced no high-scoring fragments.

We have predicted the top three epitope patches using standard parameters of EpiPred used throughout this manuscript. The top epitope prediction is incorrect placing the candidate epitope on the wrong end of the asymmetric homo-dimer structure of the antigen. Nevertheless, the second one overlaps with the actual epitope (see Figure 6.8).

We have performed docking of the antibody homology model to the antigen structure using ClusPro (Brenke et al. [2012]). The closest decoy of the 28 returned had an I_{rmsd} of 10.6Å which is low enough to be tentatively classified as close to native . All the other decoys had considerably higher I_{rmsd} values. This decoy was at seventh position as ordered by ClusPro, but was brought to the third position using our re-scoring pipeline. This top decoy superimposed on the native complex is shown in Figure 6.8.

6.4 Conclusions

In this Chapter, we have demonstrated that our antibody-specific epitope prediction method is able to improve the global docking of antibodies and antigens.

The method EpiPred, when given structures of the antibody and the antigen, annotates the likely epitope regions specific to the supplied antibody. In that respect EpiPred differs from other methods like DiscoTope or PEPITO which annotate general immunogenic/epitope-like regions on the antigen, without any antibody information required on input (Kringelum et al. [2012], Sela-Culang et al. [2013]). We demonstrate that the top epitope predictions obtained using our method have a considerably higher average recall (45%) than that expected at random (23%). We further demonstrate that EpiPred can receive homology models on input without a negative effect on performance. Thus it can be concluded that EpiPred only requires the sequence of an antibody and the structure of the antigen to produce meaningful results.

We have used the epitope predictions from EpiPred to re-rank the outputs of two fast rigid-body docking algorithms and find that re-scoring the decoys in this manner significantly enriches the number of close to native poses among the top one, five and ten results. This result holds for targets where the antibody is a homology model and the antigen is in the unbound structure which is a realistic setup for virtual screening.

We have also tested our global docking pipeline on a blind test case supplied by UCB Pharma. Even though the best epitope prediction was only second, the resulting re-scoring of the top ClusPro poses, brought the best decoy from the seventh to the third position. This is a further indication that including epitope prediction information enriches the top docking poses with more close to native conformations.

In conclusion, our global pipeline increases the confidence that the close to native decoy will be among the top five poses. This is already a significant reduction of the potential set of possibilities experimentalists need to cope with when deciding on how to adjust the antibody sequence against the antigen. A researcher might choose to infer information

by examining these top poses or they could further refine the results using more time consuming flexible docking procedures.

Chapter 7

Conclusion and future directions

In this thesis, we have presented work carried out on computational antibody design. We have analysed the antibody-antigen interactions as well as developed tools which could facilitate computational antibody design. Future directions for each of the projects described in this thesis are presented in the sections that follow.

7.1 Databases

The evolution of the antibody structure database presented in this work demonstrates the need to adapt the tools so as to keep up with a dynamic research field. The current version of the database, SAbDab, has been designed so as to serve as a portal for structural antibody information for years to come. The antibody collection and presentation mechanisms are robust and thus should be able to handle the rapidly increasing number of antibodies deposited in the PDB.

In the light of the growing datasets of antibody structures, many studies over the last 30 years have performed very similar analyses but on ever larger datasets. An example of this is the case of the CDR clustering, where many studies over the last 25 years

re-clustered the CDRs using a snapshot of structures available at the time (Chothia and Lesk [1987], Chothia et al. [1989], North et al. [2011], Martin and Thornton [1996], Lara-Ochoa et al. [1996], Al-Lazikani et al. [1997]). The aim of such studies was to validate if the conclusions drawn for the smaller datasets would hold when a more statistically relevant sample size became available. For instance, in this manuscript, the 121 complexes used in the analysis of the antibody-antigen contact preferences indicated that histidine is a very important binding residue. This result became refuted in subsequent analysis when a larger dataset of 148 antibody-antigen complexes was used. Such standardized analysis can be implemented on the fly in SAbDab using up-to-date datasets, allowing researchers to gain insight into which residues appear to have increased binding preferences. For instance, our CDR clustering deployed in SAbDab is an example of such standardized procedures being applied to new entries in the database, providing researchers with an up-to-date view of CDR conformational space.

One piece of information currently missing from SAbDab is the epitope data beyond the complex annotation. The structural epitope information could be collected directly from the complexes. The data might consist of the residues that participate in antibody-antigen contacts, individual residue distances from the binding partner etc. Alternatively, we could provide links to other epitope databases, in particular to the structural entries of AntigenDB (Ansari et al. [2010]), the Conformational Epitope Database (CED) (Huang and Honda [2006]) or the Immune Epitope Database (IEDB) (Vita et al. [2010]).

Integration with other databases could go beyond epitope resources. Even though SAbDab is primarily a structure portal, integrating it with sequence databases could increase its functionality. For instance, if one wished to carry out analysis of antibody contacts sites in the context of the somatic hypermutations, it would be necessary to map each complex to their respective germline varieties in order to identify mutation points. Such link could be provided by mapping the structures to the appropriate entries of IMGT gene database (Lefranc [2011]).

Finally, we have developed SAbDab to serve as a platform for the antibody design tools available in the Oxford Protein Informatics Group (OPIG). As such we will keep

integrating the tools as they become available in order to increase the range of services provided by SAbDab.

7.2 Antibody-Antigen Binding Site Analysis

The analysis of antibody-antigen interfaces described in this thesis provided a view of the antibody binding site available in 2010. Each of the studies presented in this part of our work would merit from renewed analysis on more up-to-date datasets. This was delineated in the previous section and one standardized analysis stemming from this work, namely CDR clustering is already available through SAbDab.

One particular aspect of the antibody binding site that has not been explored in this work was the binding affinity. At the beginning of this D.Phil we attempted analysing the correlation between antibody-antigen contacts and the corresponding binding affinities. We found that the affinity data available for experimentally determined antibody-antigen complexes at the time was very unreliable - experimental conditions were absent and many identical complexes had radically different binding affinities. We decided at that point that our efforts were better placed in developing Antibody i-Patch and the subsequent docking and epitope prediction work. The issue of binding affinity in the context of antibody structure has been revisited by another member of our research group, namely Jinwoo Leem. Building on our previous work, he has manually curated antibody structural affinity benchmark consisting of about 200 proteins which now form part of SAbDab.

Given the structural antibody affinity dataset now available, it would be worthwhile to analyse the antibody binding affinity in the context of somatic hypermutations. For this analysis, one would need a large enough dataset of affinity-annotated structures mapped back to their germline sequences. The binding affinity dataset is currently available through SAbDab, and the germline mapping can be obtained from services like IMGT (Lefranc [2011]). Using such a dataset one could group the mutations that form part of

antibody-antigen contacts for each structure and contrast it with the associated binding affinity of the complex. Such analysis would be the first of its kind, providing a first glimpse at the mechanism of structural affinity maturation of antibodies on a diverse set of complexes.

7.3 Antibody - Antigen Contact Prediction: Antibody i-Patch and EpiPred

The most obvious way to improve the quality of predictions of both EpiPred and Antibody i-Patch is to increase the training set of antibody-antigen complexes. The software was created in a way that facilitates re-training using new data. Since Antibody i-Patch has been already integrated with SAbDab, we hope that it will keep on improving the quality of its predictions. We hope to integrate EpiPred with SAbDab in a similar fashion in the near future.

It remains to be tested if increasing the complexity of the algorithm could improve the quality of predictions. One possible way of improving the working of either algorithm is to feed the results of one to the other. At the point of creating the Antibody i-Patch score, the corresponding propensities for the antigen could be augmented with the EpiPred scores for the likelihood of the given patch to be an epitope for this particular antibody. In a similar fashion, the antibody portion of he EpiPred score for each patch can be augmented with Antibody i-Patch scores for this particular antibody.

Currently, Antibody i-Patch only uses the intra-molecular structural information to obtain its predictions. Introducing a complementarity metric with respect to the antigen, as in the case of EpiPred, could potentially increase the predictive power as the algorithm would be better placed to exclude the contacts which could not possibly be made because of steric constraints. Other improvements might include an explicit electrostatic term, as it has been argued previously that it is one of the most important determinants of the

antibody-antigen complementarity (Lippow et al. [2007], Brenke et al. [2012], Krawczyk et al. [2013]).

The main improvement to EpiPred could be in the way the final list of N patches is created. Currently, candidate epitopes which share more than 30% residues with any other candidate with a higher EpiPred score are excluded. It would be beneficial to test if an alternative approach consisting of structural clustering of the high-scoring patches and choosing representatives improves results. Such clustering is the principle of operation of PIPER/ClusPro, where the best scoring poses are structurally clustered and the final ranking consists of the representatives from the groups with highest number of elements. Another improvement could be introducing explicit scoring terms akin to some docking algorithms, describing the electrostatic complementarity between the candidate patch and the antibody.

7.4 Antibody-Antigen Docking

In this thesis we have not developed our own docking algorithms. Instead we relied on poses supplied by antibody-unoptimized ZDOCK and PatchDock or by antibody-optimized PIPER/ClusPro. Given our success in deploying Antibody i-Patch and EpiPred, it would be beneficial to develop an algorithm which would use the predictions provided by these two algorithms for pose generation.

As the first approximation, such an algorithm could take the form of an FFT procedure with explicit antibody-antigen statistical potential scoring term. The scoring function could be extended by explicit terms incorporating the Antibody i-Patch and EpiPred predictions. Such a protocol could perform the re-scoring we usually carry-out after the docking at the point of decoy generation which could improve the running time of the algorithm. Positions similar to those with low scores would not be considered and thus allow exploration of more relevant poses. The algorithm could explore potential scoring

maxima at the time of decoy generation, perhaps improving the quality of returned poses.

We have suggested that the poses generated by our docking pipelines could benefit from refinement by more complex, flexible docking methods such as SnugDock. This is one venue of exploration which remains to be validated, especially with respect to the global antibody-antigen docking.

7.5 Final Words

As demonstrated in the previous sections, the work presented here is by far not complete and there are still many potential areas of further exploration. We hope that our contribution will have a positive impact on the field of computational antibody design and that some ideas described in this Chapter will be explored in the future.

Appendix A

Supplementary information for CDR contact prediction

A.1 PDB codes for the structures used in dataset NR-full

The PDB codes and chains used in NR-full are presented in Table A.1.

Table A.1: Dataset NR-fulll

PDB	Ab Heavy Chain	Ab Light Chain	Ag
1ahw	E	D	F
1dee	D	C	G
1e6j	H	L	P
1egj	H	L	A
1eo8	H	L	A
1fns	H	L	A
1fsk	C	B	A
1h0d	B	A	C
1iqd	B	A	C
1jhl	H	L	A
1jrh	H	L	I
1kb5	H	L	B
1lk3	H	L	A
1mhp	H	L	A
1nca	H	L	N
1nfd	F	E	B
1nsn	H	L	S
1oaz	J	N	B
1ob1	B	A	F
1ors	B	A	C
1osp	H	L	O
1pkq	G	F	J
1rjl	B	A	C
1v7m	H	L	V

1w72	H	L	D
1wej	H	L	F
1xiw	H	G	E
1yjd	H	L	C
1ymh	B	A	E
1yy9	D	C	A
1yym	R	Q	P
1ztx	H	L	E
2adf	H	L	A
2aep	H	L	A
2arj	B	A	R
2fd6	H	L	U
2ghw	D	D	C
2h9g	B	A	R
2hmi	D	C	B
2ih3	A	B	C
2j6e	H	L	B
2j88	H	L	A
2jel	H	L	P
2nr6	F	E	B
2nyy	D	C	A
2oz4	H	L	A
2q8b	H	L	A
2qqk	H	L	A
2qqn	H	L	A
2r29	H	L	A
2r56	I	M	B
2vh5	H	L	R
2vxq	H	L	A
2vxs	J	N	D
2vxt	H	L	I
2w9e	H	L	A
2xqb	H	L	A
2xqy	G	L	A
2xtj	D	B	A
2xwt	A	B	C
2yc1	D	E	F
2ypv	H	L	A
2zch	H	L	P
3ab0	B	C	D
3b9k	D	C	F
3bdy	H	L	V
3cvh	H	L	M
3cx5	J	K	E
3d85	B	A	C
3dvg	B	A	Y
3gbm	I	M	D
3gi9	H	L	C

3grw	H	L	A
3h3b	C	C	B
3hb3	C	D	B
3hi1	B	A	J
3hi6	X	Y	B
3hmx	H	L	A
3iu3	H	L	I
3jwd	P	O	B
3k2u	H	L	A
3kr3	H	L	D
3ks0	H	L	B
3l5w	H	L	I
3l95	H	L	Y
3ld8	C	B	A
3lev	H	L	A
3lh2	I	M	U
3lzf	H	L	A
3ma9	H	L	A
3mj9	H	L	A
3mxw	H	L	A
3nh7	J	N	C
3o0r	H	L	C
3o2d	H	L	A
3pgf	H	L	A
3pnw	H	G	I
3q1s	H	L	I
3q3g	H	F	I
3qwo	H	L	P
3r1g	H	L	B
3raj	H	L	A
3rkd	H	L	A
3ru8	H	L	X
3rvv	D	C	A
3s35	H	L	X
3sdy	H	L	B
3skj	H	L	E
3so3	C	B	A
3sob	H	L	B
3t3p	H	L	D
3tt1	H	L	B
3u4e	A	B	J
3u7y	H	L	G
3uc0	I	M	B
3ux9	D	D	C
3uze	B	B	D
3v6o	C	E	B
3vg9	C	B	A
3vi3	H	L	D

3ztn	H	L	B
4aei	I	M	B
4ag4	H	L	A
4am0	H	L	R
4d9q	E	D	B
4dkf	H	L	B
4dn4	H	L	M
4dqo	H	L	C
4dtg	H	L	K
4dw2	H	L	U
4ene	E	F	B
4ers	H	L	A
4etq	H	L	C
4f2m	C	D	F
4f3f	B	A	C
4ffv	D	C	B
4ffy	H	L	A
4fqj	H	L	A
4g3y	H	L	C
4g6j	H	L	A
4hc1	H	L	A
4hf5	H	L	A
4hkx	A	B	E
4hlz	G	H	B
4hwb	H	L	A
4i3s	H	L	G
4i9w	E	D	B
4jpk	H	L	A

A.2 Paratome datasets

The test and training datasets used in the Paratome study were downloaded Kunik et al. [2012]. A few of the structures could not be used in our analysis due to inconsistencies. A complete list of our version of the Paratome dataset is given in Tables A.2 and A.3. The structures in Table A.2 were used to train Antibody i-Patch. This version of Antibody i-Patch was then applied to the Paratome test set from Table A.3. The results of the comparison between Antibody i-Patch and Paratome are shown in Figure 4.7.

Table A.2: Paratome training dataset

PDB	Ab Heavy Chain	Ab Light Chain	Ag
2HH0	H	L	P
1NAK	H	L	P
2B1A	H	L	P
1JHL	H	L	A
1XGY	H	L	P
1CFT	B	A	C
1AR1	C	D	B
2A6I	B	A	P
1SY6	H	L	A
1W72	H	L	A
2R4S	H	L	A
1ZA3	B	A	S
2CMR	H	L	A
3C09	H	L	D
1U8N	B	A	C
2JEL	H	L	P
1TZH	B	A	W
1H0D	B	A	C
2QR0	B	A	C
2J88	H	L	A
1RJL	B	A	C
1TPX	B	C	A
2OQJ	B	A	C
2FJH	H	L	V
1ACY	H	L	P
1RZK	H	L	G
2QAD	D	C	A
1ORS	B	A	C
1N8Z	B	A	C
1PKQ	B	A	E
1E4X	H	L	P
2J4W	H	L	D
2QSC	H	L	P
3C2A	H	L	P
2NZ9	D	C	A
2B2X	H	L	A

2JIX	H	L	E
1P4B	H	L	P
2R0L	H	L	A
1U95	B	A	C
3BKY	H	L	P
1QFW	H	H	A
1NDG	B	A	C
1U8M	B	A	C
1HYS	D	C	B
2QQL	H	L	A
1FSK	C	B	A
2NR6	D	C	A
1U8Q	B	A	C
1JPS	H	L	T
1TZI	B	A	V
1A3R	H	L	P
1PZ5	B	A	C
1UJ3	B	A	C
1DQJ	B	A	C
1CZ8	H	L	W
1CU4	H	L	P
2OZ4	H	L	A
1YQV	H	L	Y
2B0S	H	L	P
2ADF	H	L	A
2OR9	H	L	P
1ZTX	H	L	E
1CFS	B	A	C
1AFV	H	L	A
1BGX	H	L	T
1HI6	B	A	C
1NFD	H	G	D
3CK0	H	L	P
2UZI	H	L	R
1EO8	H	L	A
2R0W	H	L	Q
2FJG	B	A	W
1MHP	H	L	A
2CK0	H	L	P
2VWE	E	C	A
2BOC	A	B	C
1NSN	H	L	S
2I9L	B	A	I
1EGJ	H	L	A
2GSI	B	A	W
1S78	D	C	A
1QKZ	H	L	A
1XGU	B	A	C

1BVK	B	A	C
1U92	B	A	C
1KC5	H	L	P
2NY7	H	L	G
1FPT	H	L	P
1FBI	H	L	X
1GGI	H	L	P
1KTR	H	L	P
1P2C	B	A	C
1I8K	B	A	C
2R0K	H	L	A
1JRH	H	L	I
1KEN	H	L	A
2J5L	C	B	A
2HFG	H	L	R
2B1H	H	L	P
1YNT	B	A	F
3B2U	H	L	A
1G9N	H	L	G
1BOG	B	A	C
1E4W	H	L	P
1OAZ	H	L	A
2Q8A	H	L	A
2OSL	H	L	P
1FJ1	B	A	F
1KCS	H	L	P
1CE1	H	L	P
1FE8	H	L	A
1AHW	B	A	F
1CFN	B	A	C
1XIW	D	C	A
1ADQ	H	L	A
2VDK	H	L	A
2B4C	H	L	C
1QFU	H	L	A
1N6Q	H	L	B
1WEJ	H	L	F
3BKJ	H	L	A
1ORQ	B	A	C
2HRP	H	L	P
2IFF	H	L	Y
2EH8	H	L	P
2QQN	H	L	A
1UWX	H	L	A
1MLC	B	A	E
1FRG	H	L	P
2J6E	H	L	A
1E6J	H	L	P

1N64	H	L	P
1KCR	H	L	P
2DD8	H	L	S
1N0X	H	L	R
2BDN	H	L	A
2OTU	B	A	P
1U8P	B	A	C
1FNS	H	L	A
1YJD	H	L	C
2AEP	H	L	A
1F90	H	L	E
2FX7	H	L	P
2VIS	B	A	C
1IGC	H	L	A
2IGF	H	L	P
2H1P	H	L	P
1NCD	H	L	N
1Z3G	H	L	A
1TET	H	L	P
2QHR	H	L	P
1XCQ	B	A	P
1J1P	H	L	Y
1YY9	D	C	A
1U8O	B	A	C
1HIM	L	H	P
2ZCK	L	H	P
1OTS	C	D	A
1U93	B	A	C
2V17	H	L	A
1V7M	H	L	V
1F58	H	L	P
3D85	B	A	C
1UAC	H	L	Y
1TJI	H	L	P
2ARJ	H	L	Q
1IQD	B	A	C
1XCT	B	A	P
3CVH	H	L	A
2HKF	H	L	P
1SM3	H	L	P
1NL0	H	L	G
3BT2	H	L	U
1NDM	B	A	C
1KB5	H	L	A
1HH9	B	A	C
2H9G	B	A	R
1NMA	H	L	N
1U8L	B	A	C

2BRR	H	L	P
2G5B	B	A	I
2R56	H	L	A
2IPU	H	L	Q
1OB1	B	A	F
2R29	H	L	A
1I9R	H	L	A

Table A.3: Paratome test dataset

PDB	Ab Heavy Chain	Ab Light Chain	Ag
2XQY	G	L	A
3PNW	B	A	C
3MLV	H	L	P
3NIF	H	L	A
3O6M	H	L	C
3L5W	H	L	I
306L	H	L	C
3LEY	H	L	P
3MLU	H	L	P
3LEX	H	L	P
3MOD	H	L	P
3MLY	H	L	P
3O0R	H	L	C
3O0R	H	L	B
2XTJ	D	B	A
3LDB	C	B	A
3A6B	H	L	Y
3MA9	H	L	A
3AB0	B	C	A
3L95	B	A	X
3A67	H	L	Y
3K2U	H	L	A
3NH7	H	L	A
3MNW	B	A	P
3MOA	H	L	P
3L5X	H	L	A
3LD8	C	B	A
3NID	H	L	A
2WUC	H	L	A
3L5Y	H	L	A
3A6C	H	L	Y
3KJ6	H	L	A
3MNZ	B	A	P
3LOH	A	B	E

3LOH	C	D	E
3IXT	H	L	P
3JWD	H	L	A
3MAC	H	L	A
3N85	H	L	A
3LIZ	H	L	A
3MLX	H	L	P
3MLS	H	L	P
3MXW	H	L	A
3LEV	H	L	A
3OR7	A	B	C
3O41	H	L	P
3MLZ	H	L	P
3KLH	D	C	B
2XRA	H	L	A
3KJ4	H	L	A
3LQA	H	L	C
3LQA	H	L	G
2WUB	H	L	A
3MOB	H	L	P
3MLR	H	L	P
3NIG	H	L	A
3O45	H	L	P
3JWO	H	L	A
3O2D	H	L	A
3IVK	H	L	M
3LHP	H	L	S
3LH2	H	L	S
2XQB	H	L	A
3MLT	H	L	P
3NGB	H	L	G

1ZTX	1K4C		
2FD6	2ADF	1BQL	2AEP
2BDN	1JHL	2FJH	2JEL
1NCA	1FPT	2FJG	1TET
2B2X	2B4C	2G5B	1BJ1
1FBI	2H2P	1G9M	1F58

TABLE A.4: The RosettaAntibody models used to test Antibody i-Patch' ability to make predictions from homology models.

A.3 PDB codes for the homology model dataset RA

PDBs for the homology dataset RA are given in Table A.4.

A.4 Statistical significance of ROC AUC difference

For each score offered by i-Patch, (APRO, PPRO, TPRO), we have compared their performance with or without MSAs by evaluating the statistical significance between are under curve (AUC) of their respective ROC plots.

$$specificity = \frac{TN}{TN + FP} \tag{A.1}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{A.2}$$

Similarly to P-ROC curves, the average *specificity* (precision) and *sensitivity* would be calculated for each cutoff so as to make the ROC plot. The ROC curve is created by plotting $1 - sensitivity$ against *specificity* (an example plot for the three i-Patch scores, APRO, PPRO and TPRO Hamer et al. [2010], with and without MSAs are given in Figure A.1). The difference between running i-Patch on the reduced input and with full MSAs was quantified by comparing the areas under ROC curves, where higher area indicates better performance. Due to the disproportionate number of contact versus non-contact residues, for high enough cut-offs the rate of true negatives could artificially affect the *specificity*. In order to overcome this issue, we sample the prediction for each of three i-Patch scores, APRO, PPRO and TPRO by equating in number the set of non-contact residues and contact residues. In practice this means that for a given complex, if n_c residues were in contact and n_{nc} not in contact with $n_c << n_{nc}$, we would sample n_c residues uniformly at random from the non-contact (n_{nc}) set. *Specificity* and *sensitivity* would be calculated for this reduced set of residues, ignoring the remaining $n_{nc} - n_c$ residues. An are under curve would be recorded for each sample.

For each score we have produced 1000 AUC samples for the reduced input, denoted by vector x and as many for the full MSA input, denoted by y with their respective means $\bar{x}$ and $\bar{y}$. Assuming that the number of samples we chose was large enough and that

the samples were independent, by central limit theorem, $\bar{x}$ and $\bar{y}$ should follow normal distributions. Let μ_x and μ_y denote the actual AUC means that $\bar{x}$ and $\bar{y}$ are estimating respectively. We use a two-sample z-test with the assumption of unequal variances to test the null hypothesis $H_0 : \mu_x = \mu_y$ against alternative $H_a : \mu_x \neq \mu_y$ with the test statistic given by A.3.

$$z = \frac{\mu_x - \mu_y}{\sigma_{x-y}} \tag{A.3}$$

Where σ_{x-y} denotes the standard deviation of the difference $x - y$. For H_0, $z \sim N(0, 1)$ Chen et al. [2008]. According to this test the null hypothesis cannot be rejected for comparisons between the versions with or without MSAs for any of the three i-Patch scores meaning that the MSA is not contributing significantly to the predictive power of i-Patch.

An identical procedure was applied to ROC plots corresponding to the P-ROC Figures of Antibody i-Patch performance on dataset RA. In this case, the pairwise AUC differences were not statistically significantly different, indicating the equivalent performance of Antibody i-Patch on crystal structures and homology models alike.

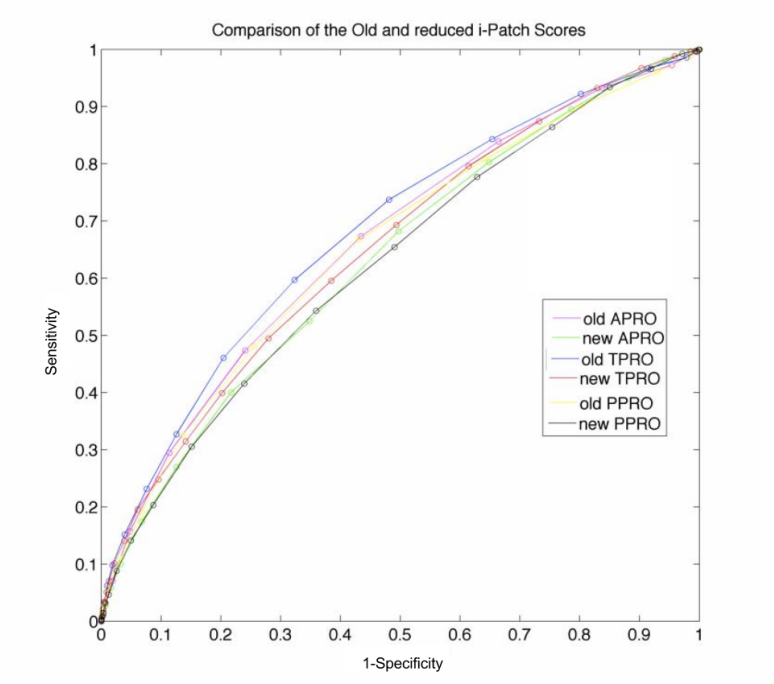


FIGURE A.1: **ROC of the i-Patch scores with and without MSAs.** The scores termed *old* use full MSAs whereas those labelled with *new* present the scores without the MSAs

A.5 Recall errors for Antibody i-Patch

The recall error bars for Figure 4.6 in the main text are given in Figure A.2.

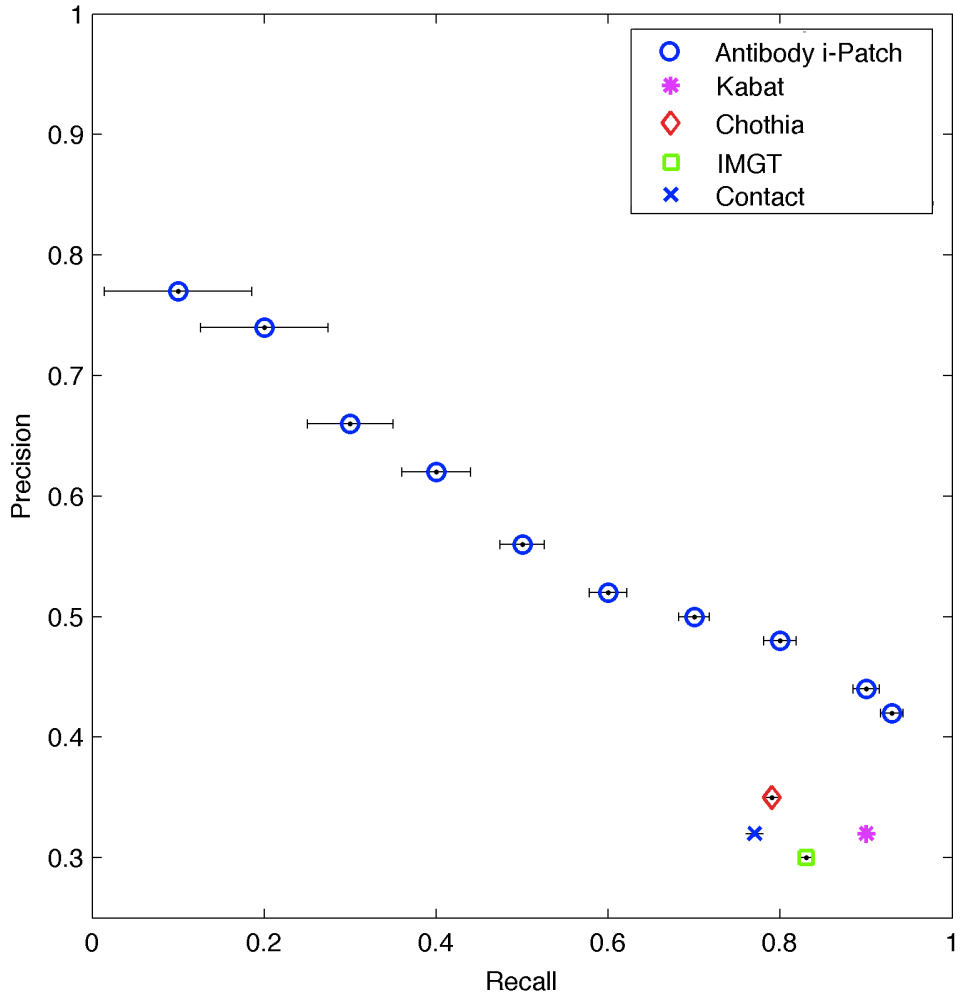


FIGURE A.2: **Recall standard error for Figure 4.6 in the main text.**

A.6 Standard errors for the RA dataset.

The standard errors corresponding to the Figure 4.8 are given in Figures A.3, A.4 and A.5.

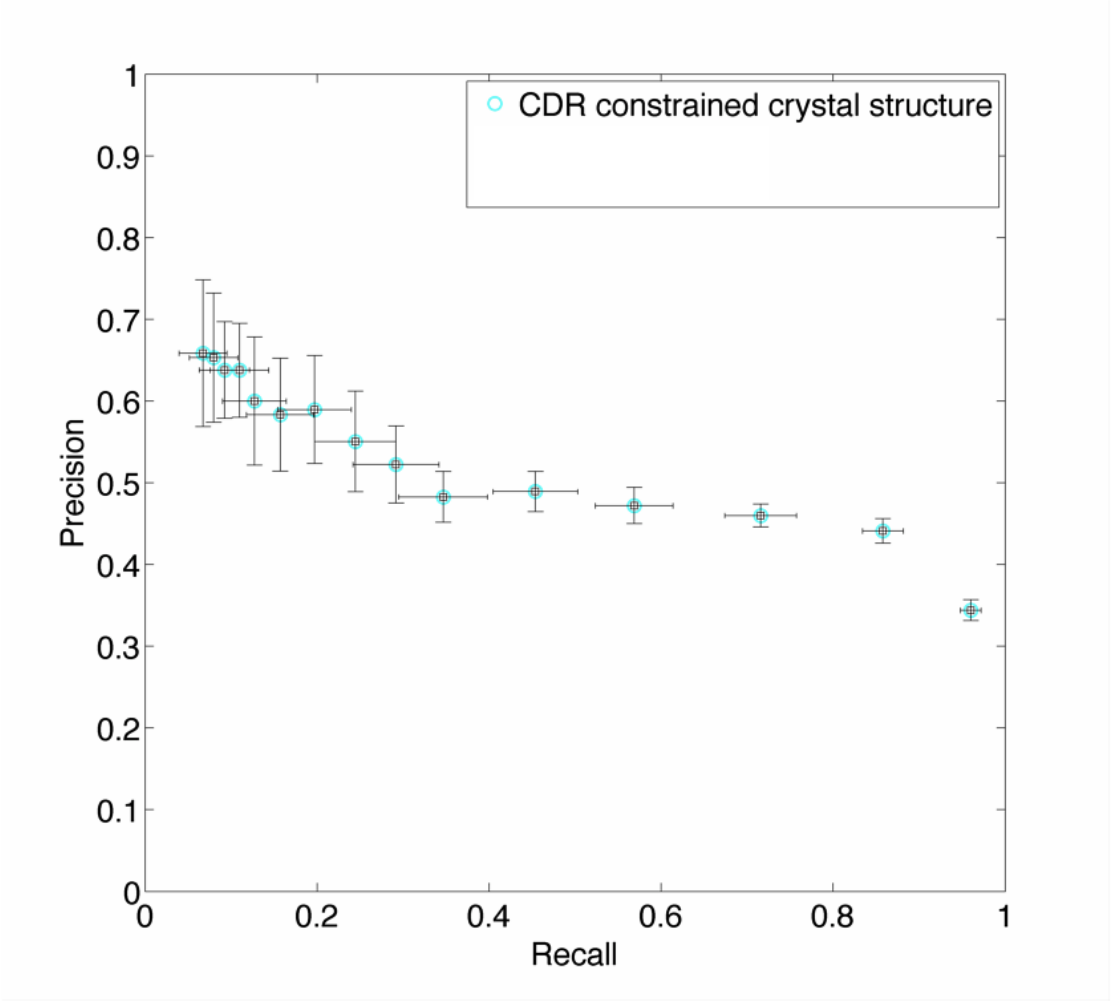


FIGURE A.3: **Performance of i-Patch on crystal structure dataset RA-x.**

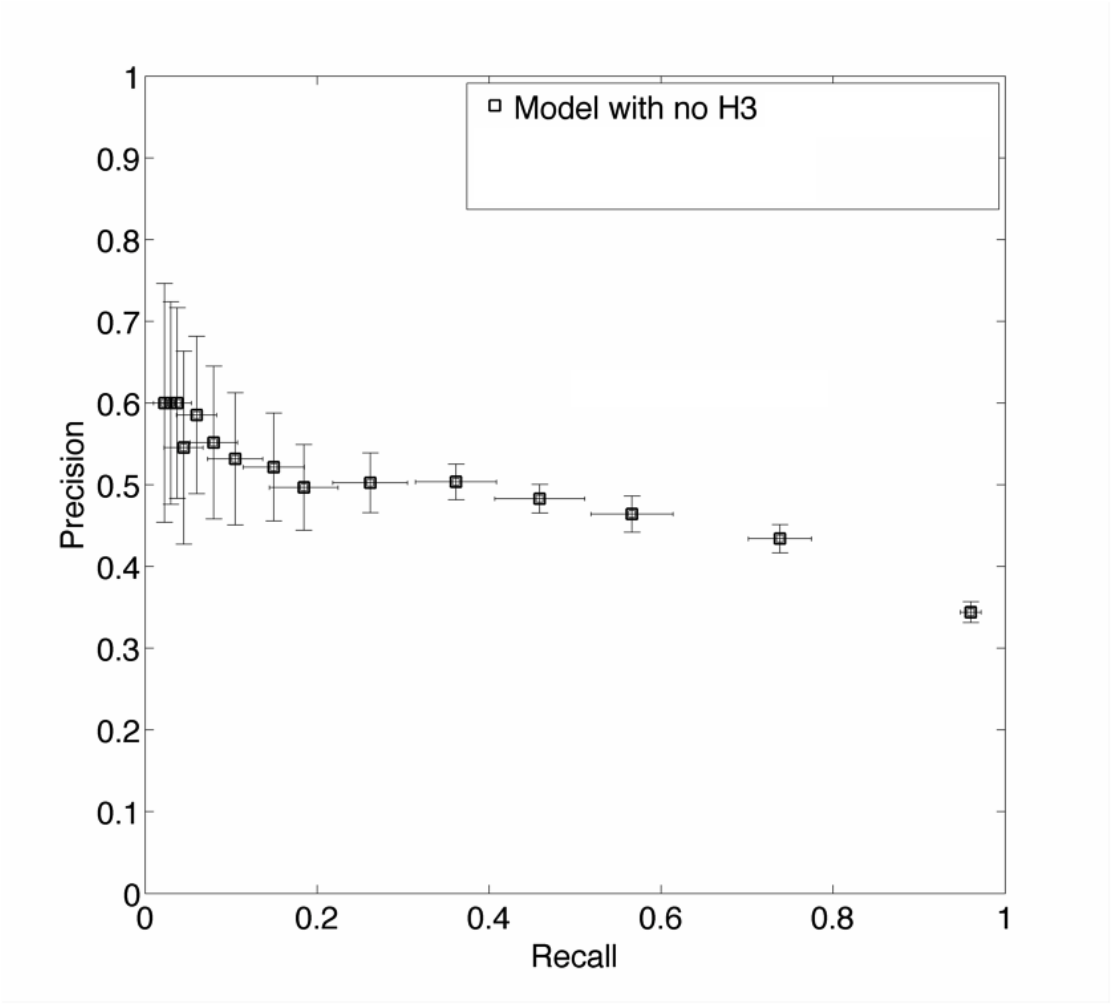


FIGURE A.4: **Performance of i-Patch on homology model dataset RA-h without models of H3.**

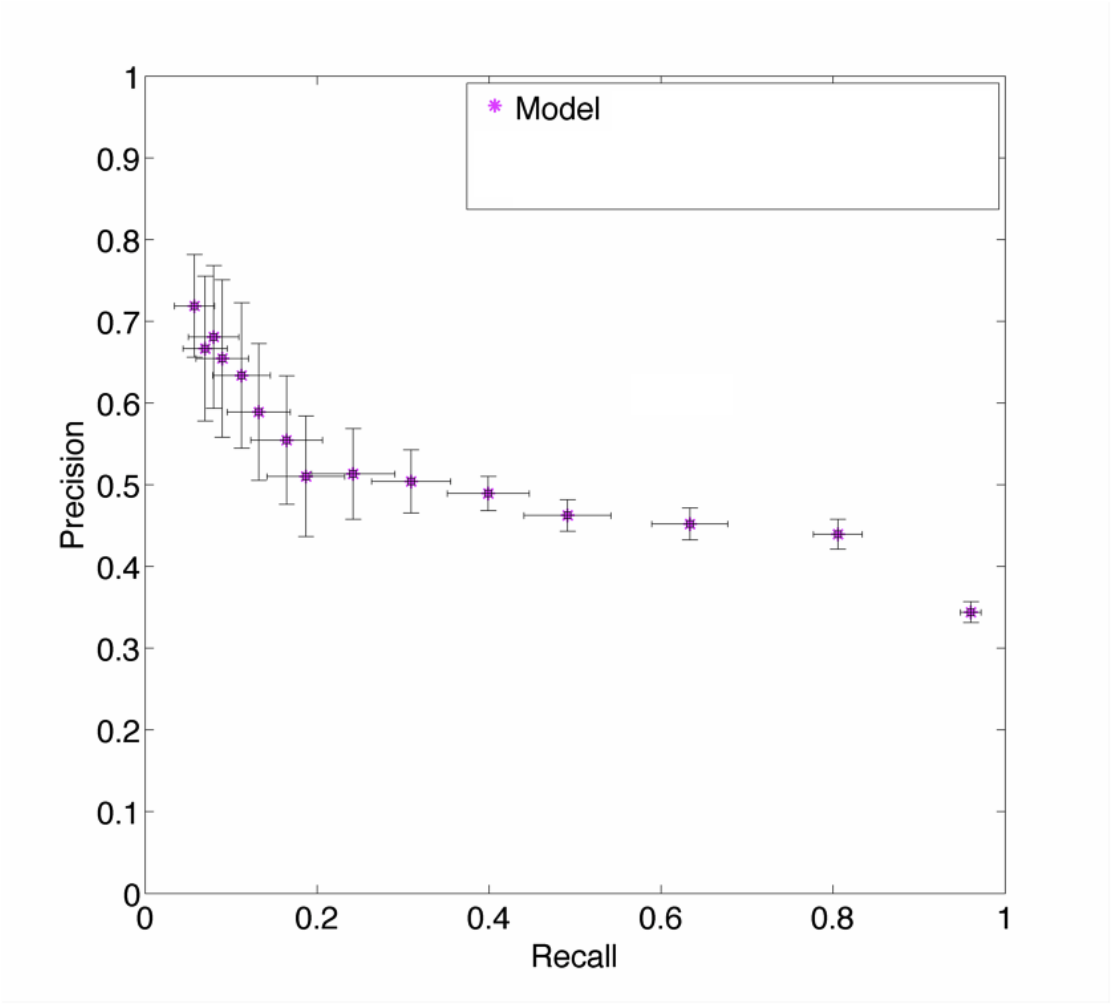


FIGURE A.5: **Performance of i-Patch on homology model dataset RA-h.**

Appendix B

Supplementary information for local Ab-Ag docking

B.1 NR-subet dataset

NR-subset constituted the test set for docking in the original manuscript. The 30 PDB codes chosen at random from NR-full are presented in Table B.1:

Table B.1: NR-subset

PDB	Ab Heavy Chain	Ab Light Chain	Ag
1ahw	E	D	F
1dee	D	C	G
1fns	H	L	A
1jrh	H	L	I
1kb5	H	L	B
1lk3	H	L	A
1nfd	F	E	B
1nsn	H	L	S
1rjl	B	A	C
1xiw	H	G	E
2adf	H	L	A
2nr6	F	E	B
2r56	I	M	B
2vh5	H	L	R
2vxs	J	N	D
2xwt	A	B	C
2yc1	D	E	F
3ab0	B	C	D
3gbm	I	M	D
3hb3	C	D	B
3hi1	B	A	J
3iu3	H	L	I
3k2u	H	L	A

3lev	H	L	A
3lzf	H	L	A
3mxw	H	L	A
3r1g	H	L	B
3ru8	H	L	X
3s35	H	L	X
3skj	H	L	E

B.2 PDB codes for the docking dataset SnugDock-H

PDBs for the homology dataset SnugDock-H are given in Table B.2.

B.3 Supplementary docking results

B.3.1 NR-subset

Here we present the docking results for other extended epitope cut-offs and for PatchDock. ZDOCK results are presented with their respective standard deviations (given in parentheses, next to the mean). The ZDOCK results are given in Figures B.1, B.2 and B.3. The corresponding results for PatchDock are given in Figures B.4, B.5 and B.6.

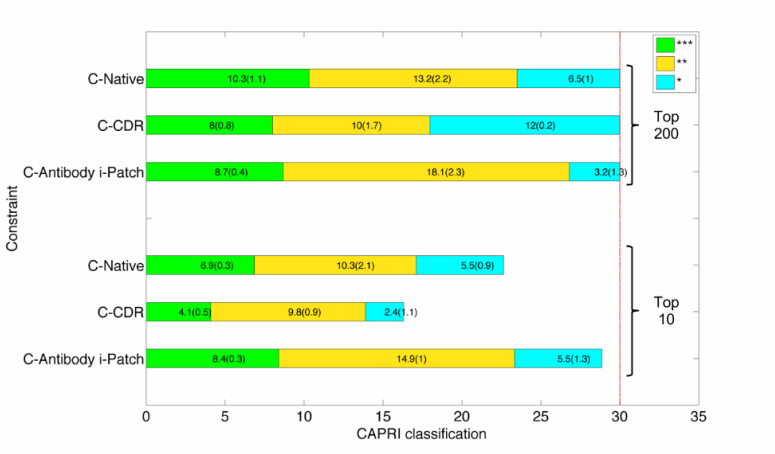


FIGURE B.1: **Docking results for ZDOCK at 4Å cut-off extended epitope and dataset NR-test.** The values in parentheses are the standard deviations.

Bound complex PDB	Bound Ab chain (H,L)	Bound Ag chain	Unbound Ab PDB	Unbound Ab chain (H,L)	Unbound Ag PDB	Unbound Ag chain	Ag comment
1mlc	B A	E	1mlb	B A	1lza	A	Hen egg white lysozyme
1ahw	B A	C	1fgn	H L	1boy	A	Human tissue factor
1jps	H L	T	1jpt	H L	1tfh	A	Human tissue factor
1wej	H L	F	1qbl	H L	1hrc	A	Horse heart cytochrome C
1vfb	B A	C	1vfa	B A	8lyz	A	Hen egg white lysozyme
1bql	H L	Y	—	- -	1dkj	A	Bobwhite quail lysozyme
1k4c	A B	C	—	- -	1jvm	A	Potassium channel KcsA, in high concentration of K+
2jel	H L	P	—	- -	1poh	A	Hpr, a phosphocarrier protein of the phosphoenolpyruvate:sugar phosphotransferase system of Escherichia coli
1jhl	H L	A	—	- -	1ghl	A	Pheasant egg white lysozyme
1nca	H L	N	—	- -	7nn9	A	Influenza A subtype N9 neuraminidase
2bdn	H L	A	—	- -	—	-	Small inducible cytokine A2
1ynt	B A	F	—	- -	—	-	Toxoplasma gondii surface antigen 1 (SAG1) p30
2aep	H L	A	—	- -	—	-	Neuraminidase (NA) of influenza virus A/Memphis/31/98 (H3N2)
2b2x	H L	A	—	- -	—	-	Integrin VLA1 RdeltaH I-domain
1ztx	H L	E	—	- -	—	-	Envelope protein of the West Nile virus

Table B.2: Summary of the homology data SnugDock-H.

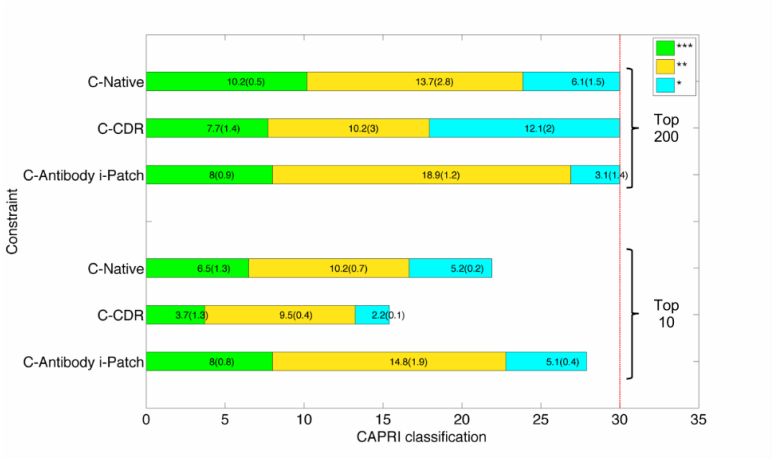


FIGURE B.2: **Docking results for ZDOCK at 5Å cut-off extended epitope and dataset NR-subset.** The values in parentheses are the standard deviations.

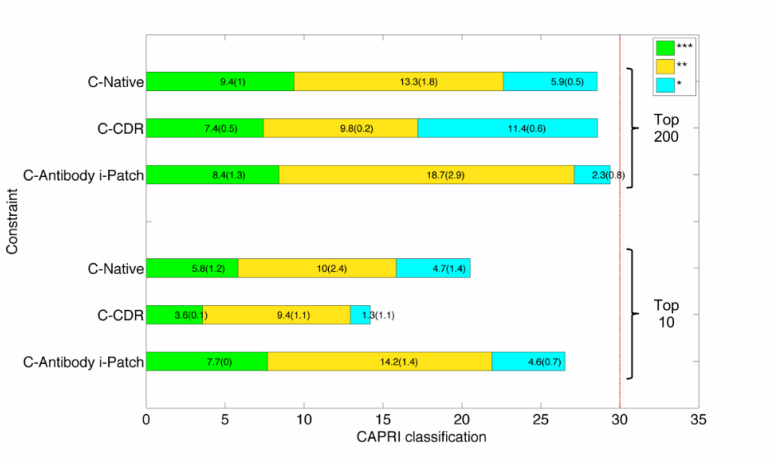


FIGURE B.3: **Docking results for ZDOCK at 6Å cut-off extended epitope and dataset NR-subset.** The values in parentheses are the standard deviations.

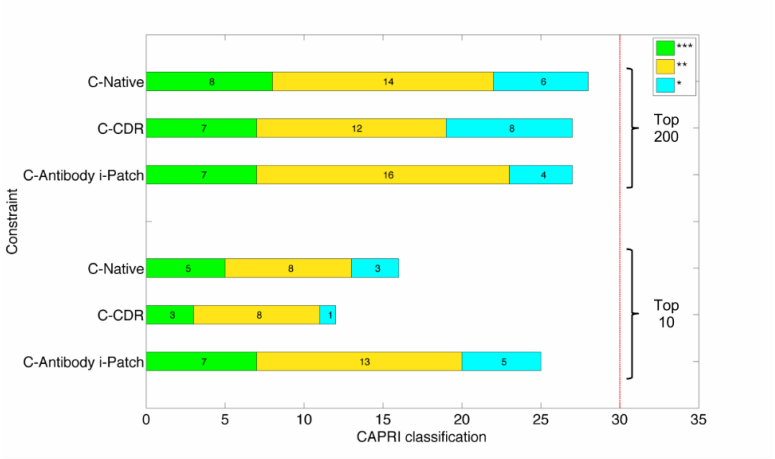


FIGURE B.4: **Docking results for PatchDock at 4Å cut-off extended epitope and dataset NR-subset.** The values in parentheses are the standard deviations.

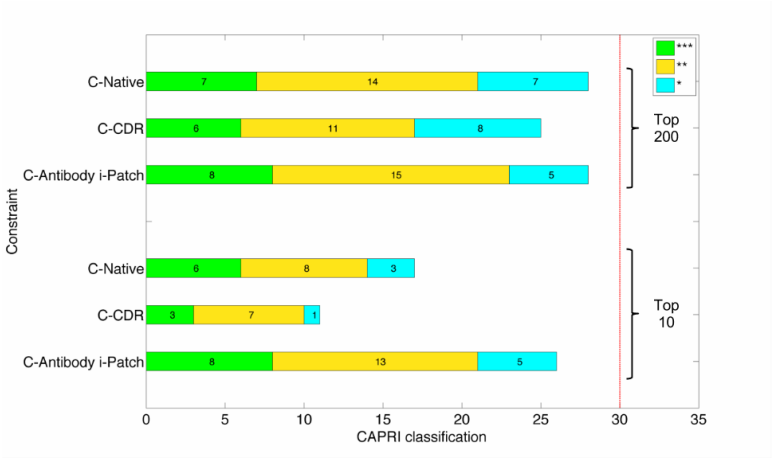


FIGURE B.5: **Docking results for PatchDock at 5Å cut-off extended epitope and dataset NR-subset.** The values in parentheses are the standard deviations.

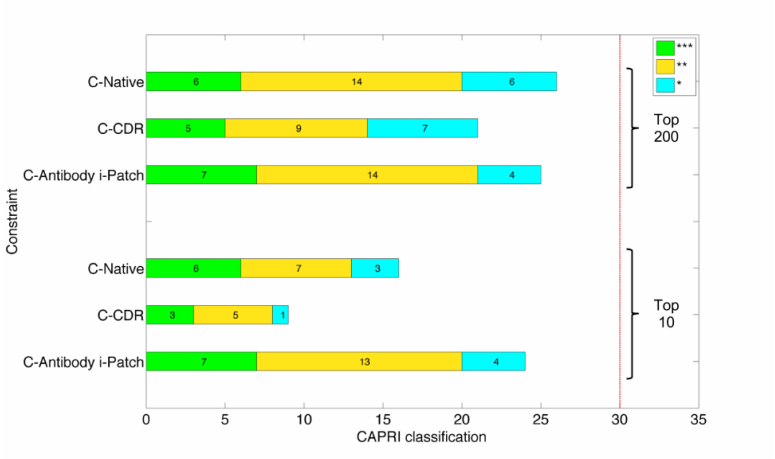


FIGURE B.6: **Docking results for PatchDock at 6Å cut-off extended epitope and dataset NR-subset.** The values in parentheses are the standard deviations.

B.3.2 SnugDock-H

Here we present the docking results for other extended epitope cut-offs on dataset SnugDock-H. ZDOCK results are presented with their respective standard deviations (given in parentheses, next to the mean). The SnugDock-H results are given in Figures B.7, B.8 and B.9.

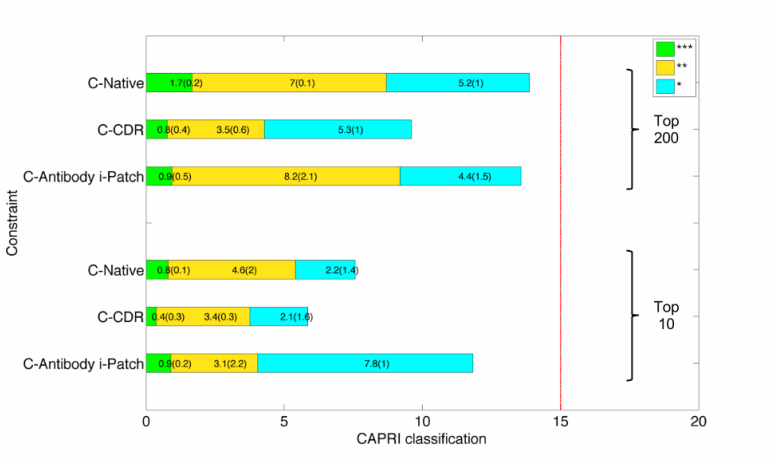


FIGURE B.7: **Docking results for ZDOCK at 4Å cut-off extended epitope and dataset SnugDock-H.** The values in parentheses are the standard deviations.

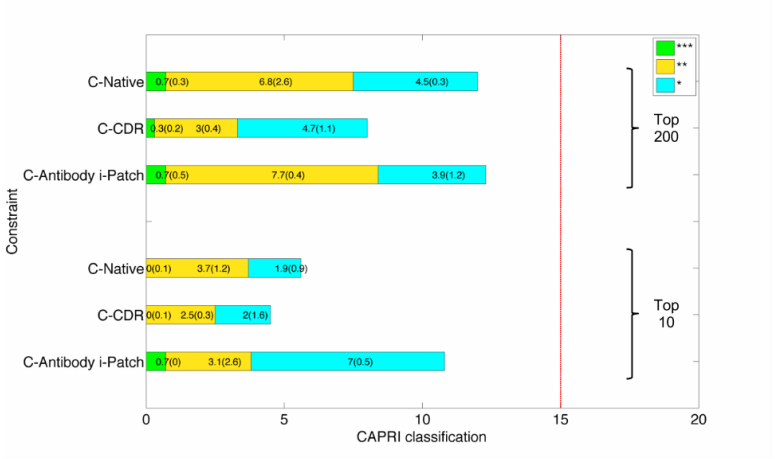


FIGURE B.8: **Docking results for ZDOCK at 5Å cut-off extended epitope and dataset SnugDock-H.** The values in parentheses are the standard deviations.

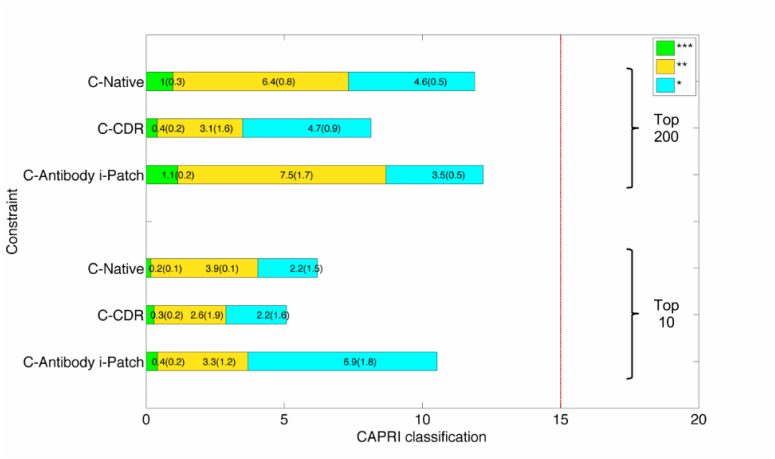


FIGURE B.9: **Docking results for ZDOCK at 6Å cut-off extended epitope and dataset SnugDock-H.** The values in parentheses are the standard deviations.

Appendix C

Supplementary information for global antibody-antigen docking

C.1 Data

The datasets used in this study are presented in the sections below.

C.1.1 X-dataset

The data used in dataset X-dataset is presented in Table C.1.

Table C.1: Summary of the data constituting dataset X-dataset.

PDB	Heavy chain	Light chain	Antigen Chain(s)	Antigen name
3d85	B	A	C	interleukin-23 subunit p19
1p2c	B	A	C	lysozyme c
3t3p	E	F	C	integrin alpha-iib
3zkm	C	D	B	beta-secretase 2
3gjf	H	L	A	hla class i histocompatibility antigen, a-2alpha chain
2uzi	H	L	R	gtpase hras
4i3s	H	L	G	outer domain of hiv-1 gp120 (ker2018 od4.2.2)
2vxs	J	N	D	interleukin-17a
1nsn	H	L	S	staphylococcal nuclease
3sqo	H	L	A	proprotein convertase subtilisin/kexin type 9
2jel	H	L	P	histidine-containing protein
1e6j	H	L	P	capsid protein p24
3lzf	H	L	A	hemagglutinin, ha1 subunit
3o2d	H	L	A	t-cell surface glycoprotein cd4
3eoa	H	L	I	integrin alpha-l
1nmb	H	L	N	n9 neuraminidase

3so3	C	B	A	suppressor of tumorigenicity 14 protein
3rkd	H	L	A	capsid protein
3lh2	I	M	T	4e10 1vi7a s0 002 n (t88)
3ngb	H	L	G	envelope glycoprotein gp160
4hwb	H	L	A	interleukin-13 receptor subunit alpha-1
3uc0	I	M	B	envelope protein
4hj0	P	Q	A	gastric inhibitory polypeptide receptor
1hez	B	A	E	protein l
3r1g	H	L	B	beta-secretase 1
3b9k	D	C	F	t-cell surface glycoprotein cd8 beta chain
4dkf	H	L	A	interleukin-34
1fj1	B	A	F	outer surface protein a
4g6m	H	L	A	interleukin-1 beta
3ma9	H	L	A	transmembrane glycoprotein
2xqy	G	L	A	envelope glycoprotein h
3hb3	C	D	B	cytochrome c oxidase subunit 2
1rjl	B	A	C	outer surface protein b
3i50	H	L	E	envelope glycoprotein
1n8z	B	A	C	receptor protein-tyrosine kinase erbb-2
1eo8	H	L	A	hemagglutinin (ha1 chain)
2xwt	A	B	C	thyrotropin receptor
3rvv	D	C	A	peptidase 1
4ag4	H	L	A	epithelial discoidin domain-containing receptor 1
4dtg	H	L	K	tissue factor pathway inhibitor
4ene	E	F	B	h(+)/cl(-) exchange transporter clca
4ffy	H	L	A	envelope glycoprotein
1ahw	E	D	F	tissue factor
2j88	H	L	A	hyaluronon glucosaminidase
1fsk	C	B	A	major pollen allergen bet v 1-a
3qwo	H	L	P	motavizumab epitope scaffold
1lk3	H	L	A	interleukin-10
3raj	H	L	A	adp-ribosyl cyclase 1
2ih3	A	B	C	voltage-gated potassium channel
2xqb	H	L	A	interleukin 15
3q1s	H	L	I	interleukin-22
1egj	H	L	A	cytokine receptor common beta chain precursor

1pkq	G	F	J	myelin oligodendrocyte glycoprotein
4ers	H	L	A	glucagon receptor
3sdy	H	L	B	hemagglutinin ha2 chain
2qqn	H	L	A	neuropilin-1
3lev	H	L	A	rna polymerase sigma factor
2j6e	H	L	A	ig gamma-1 chain c region
2h9g	B	A	R	tumor necrosis factor receptor superfamily member 10bprecursor
4hcr	M	N	B	mucosal addressin cell adhesion molecule 1
1tqb	B	C	A	prion protein
3hmx	H	L	A	interleukin-12 subunit beta
3p0y	H	L	A	epidermal growth factor receptor
2adf	H	L	A	von willebrand factor
3mxw	H	L	A	sonic hedgehog protein
1xiw	D	C	A	t-cell surface glycoprotein cd3 epsilon chain
3v6o	C	E	A	leptin receptor
2r0l	H	L	A	hepatocyte growth factor activator
1h0d	B	A	C	angiogenin
3nfp	H	L	I	interleukin-2 receptor subunit alpha
4dn4	H	L	M	c-c motif chemokine 2
4d9q	E	D	B	factor d
2zch	H	L	P	prostate-specific antigen
3tt1	H	L	B	leucine transporter leut
2fd6	H	L	U	urokinase plasminogen activator surface receptor
4leo	A	B	C	receptor tyrosine-protein kinase erbb-3
4f3f	B	A	C	mesothelin
3mj9	H	L	A	junctional adhesion molecule-like
3kr3	H	L	D	insulin-like growth factor ii
3u9p	K	M	C	neutrophil gelatinase-associated lipocalin
3q3g	H	F	I	integrin alpha-m
2ypv	H	L	A	lipoprotein
1jrh	H	L	I	interferon-gamma receptor alpha chain
4hkx	A	B	E	hemagglutinin ha1
2qqk	H	L	A	neuropilin-2
3ru8	H	L	X	epitope scaffold 2bodx43
1iqd	B	A	C	human factor viii

4dw2	H	L	U	urokinase-type plasminogen activator
3cx5	J	K	E	cytochrome b-c1 complex subunit rieske,mitochondrial
2q8b	H	L	A	apical membrane antigen 1
4aei	I	M	B	alpha-mammal toxin aah2
3sob	H	L	B	low-density lipoprotein receptor-related protein 6
2yc1	D	E	F	beta-mammal toxin cn2
1v7m	H	L	V	thrombopoietin
1bgx	H	L	T	taq dna polymerase
3pnw	H	G	I	tudor domain-containing protein 3
4jr9	H	L	A	nitrite extrusion protein 1
2arj	B	A	R	t-cell surface glycoprotein cd8 alpha chain
3bgf	B	C	A	spike protein s1
1oaz	J	N	B	thioredoxin 1
4k2u	I	M	B	erythrocyte binding antigen 175
2r29	H	L	A	envelope protein e
4i9w	E	D	A	potassium channel subfamily k member 4
1kb5	H	L	A	kb5-c20 t-cell antigen receptor
3grw	H	L	A	fibroblast growth factor receptor 3
4f2m	C	D	F	spike protein
3ab0	B	C	A	bcla protein
4dqo	H	L	C	1fd6-v1v2 scaffold zm109 hiv-1 strain
1ors	B	A	C	potassium channel
1mhp	H	L	A	integrin alpha 1, (residues 169-360)
1fns	H	L	A	von willebrand factor
3vg9	C	B	A	adenosine receptor a2a
4etq	H	L	C	imv membrane protein
3liz	H	L	A	aspartic protease bla g 2
4i77	H	L	Z	interleukin-13
3s37	H	L	X	vascular endothelial growth factor receptor 2
3nh7	J	N	C	bone morphogenetic protein receptor type-1a
4am0	H	L	R	envelope protein,
3hi1	B	A	J	glycoprotein 120
2hmi	D	C	B	hisubunit of v-1 reverse transcriptase
4ffv	D	C	B	dipeptidyl peptidase 4
4hf5	H	L	A	hemagglutinin ha1

3ld8	C	B	A	bifunctional arginine demethylase and lysyl-hydroxylasejmjd6
4ht1	H	L	T	tumor necrosis factor ligand superfamily member 12
3o0r	H	L	B	nitric oxide reductase subunit b
3skj	H	L	E	ephrin type-a receptor 2
3dvg	B	A	Y	ubiquitin
1wej	H	L	F	cytochrome c
2r56	I	M	B	beta-lactoglobulin
1tzh	B	A	W	vascular endothelial growth factor a
4jpk	H	L	A	germline-targeting hiv-1 gp120 engineered outer domain,eod-gt6
3pgf	H	L	A	maltose-binding periplasmic protein
1yjd	H	L	C	t-cell-specific surface glycoprotein cd28
3vi3	H	L	D	integrin beta-1
2oz4	H	L	A	intercellular adhesion molecule 1
3ks0	H	L	B	cytochrome b2, mitochondrial
4g3y	H	L	C	tumor necrosis factor
3l95	H	L	Y	neurogenic locus notch homolog protein 1
4kuc	F	E	I	ricin
4fqj	H	L	A	hemagglutinin
1nfd	F	E	B	n15 alpha-beta t-cell receptor
4f37	H	L	A	colicin-e7 immunity protein
2vxq	H	L	A	pollen allergen phl p 2
1ob1	B	A	C	major merozoite surface protein
4jqi	H	L	A	beta-arrestin-1
2aep	H	L	A	neuraminidase
2vxt	H	L	I	interleukin-18
2nyy	D	C	A	botulinum neurotoxin type a
3gi9	H	L	C	uncharacterized protein mj0609

C.1.2 X-test

The data used in dataset X-test is presented in Table C.2.

TABLE C.2: Summary of the data constituting dataset X-test.

PDB	Heavy chain	Light chain	Antigen Chain(s)	Antigen name
1p2c	B	A	C	lysozyme c
3t3p	E	F	C	integrin alpha-iib
3zkm	C	D	B	beta-secretase 2
3gjf	H	L	A	hla class i histocompatibility antigen, a-2alpha chain
3o2d	H	L	A	t-cell surface glycoprotein cd4
4hj0	P	Q	A	gastric inhibitory polypeptide receptor
1hez	B	A	E	protein l
3r1g	H	L	B	beta-secretase 1
3ma9	H	L	A	transmembrane glycoprotein
3i50	H	L	E	envelope glycoprotein
1n8z	B	A	C	receptor protein-tyrosine kinase erbb-2
3rvv	D	C	A	peptidase 1
4ene	E	F	B	h(+)/cl(-) exchange transporter clca
3raj	H	L	A	adp-ribosyl cyclase 1
2ih3	A	B	C	voltage-gated potassium channel
3q1s	H	L	I	interleukin-22
3u9p	K	M	C	neutrophil gelatinase-associated lipocalin
1v7m	H	L	V	thrombopoietin
4jr9	H	L	A	nitrite extrusion protein 1
3ab0	B	C	A	bcla protein
1fns	H	L	A	von willebrand factor
3liz	H	L	A	aspartic protease bla g 2
4i77	H	L	Z	interleukin-13
4am0	H	L	R	envelope protein,
4ht1	H	L	T	tumor necrosis factor ligand superfamily member 12
3o0r	H	L	B	nitric oxide reductase subunit b
1tzh	B	A	W	vascular endothelial growth factor a
3pgf	H	L	A	maltose-binding periplasmic protein
4g3y	H	L	C	tumor necrosis factor
1nfd	F	E	B	n15 alpha-beta t-cell receptor
2vxt	H	L	I	interleukin-18

Bibliography

C Abergel and J Claverie. A strong propensity toward loop formation characterizes the expressed reading frames of the d segments at the ig h and t cell receptor loci. *European journal of immunology*, 21(12): 3021–3025, 1991.

K R Abhinandan and A C R Martin. Analysis and prediction of vh/vl packing in antibodies. *Protein Engineering Design and Selection*, 23(9):689–697, 2010.

Martin A C. R Abhinandan K R. Analysis and improvements to kabat and structurally correct numbering of antibody variable domains. *Mol Immunol.*, 14:3832–9., 2008.

E. E. Abola, F. C. Bernstein, and T. F. Koetzle. Protein data bank. Technical report, Brookhaven National Lab., Upton, NY (USA), 1984.

B Al-Lazikani, A M Lesk, and C Chothia. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol.*, 4:927–948, 1997.

L C Allcorn and A C R Martin. Sacsself-maintaining database of antibody crystal structure information. *Bioinformatics*, 18(1):175–181, 2002.

P Aloy and R B Russell. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences*, 99(9):5896–5901, 2002.

H R Ansari, D R Flower, and G P S Raghava. AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res.*, 38(Database issue):D847–53, 2010.

R Barderas, J Desmet, P Timmerman, R Meloen, and J I Casal. Affinity maturation of antibodies assisted by in silico modeling. *Proceedings of the National Academy of Sciences*, 105(26):9029–9034, 2008.

S J Bell and M A Kamm. The clinical role of anti-tnfa antibody treatment in in crohn's disease. *Aliment Pharmacol Ther*, 14:501–514, 2000.

E Besmer, P Gourzi, and F N Papavasiliou. The regulation of somatic hypermutation. *Current opinion in immunology*, 16(2):241–245, 2004.

S. Birtalan, Y. Zhang, F. A. Fellouse, L. Shao, G. Schaefer, and S. S. Sidhu. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *Journal of molecular biology*, 377(5):1518–1528, 2008.

R Bonneau, J Tsai, I Ruczinski, D Chivian, C Rohl, C E M Strauss, and D Baker. Rosetta in casp4: progress in ab initio protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 45(S5):119–126, 2001.

A M J J Bonvin. Flexible protein–protein docking. *Current opinion in structural biology*, 16(2):194–200, 2006.

G L Boulianne, N Hozumi, and M J Shulman. Production of functional chimaeric mouse/human antibody. 1984.

R. Brenke, D. R. Hall, G. Y. Chuang, S. R. Comeau, T. Bohnuud, D. Beglov, O. Schueler-Furman, S. Vajda, and D. Kozakov. Application of asymmetric statistical potentials to antibody–protein docking. *Bioinformatics*, 28(20):2608–2614, 2012.

C J Camacho and C Zhang. Fastcontact: rapid estimate of contact and binding free energies. *Bioinformatics*, 21(10):2534–2536, 2005.

P J Carter. Potent antibody therapeutics by design. *Nature Reviews Immunology*, 6(5):343–357, 2006.

A Chailyan, P Marcatili, and A Tramontano. The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *FEBS Journal*, 278(16):2858–2866, 2011.

A Chailyan, A Tramontano, and P Marcatili. A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res.*, 40(Database issue):D1230–1234, 2012.

S Chaudhury and J J Gray. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and nmr ensembles. *J Mol Biol.*, 381:10681087, 2008.

P Y Chen, C M Deane, and G Reinert. Predicting and validating protein interactions using network structure. *PLoS Comput Biol.*, 4:e1000118, 2008.

R Chen, L Li, and Z Weng. Zdock: An initial-stage protein docking algorithm. *Proteins*, 1:80–87, 2003.

S Cherian et al. Antigen-antibody docking reveals the molecular basis for cross-reactivity of the 1918 and 2009 influenza a/h1n1 pandemic viruses. *Bioinformation*, 6(1):35, 2011.

Y Choi and C M Deane. Fread revisited: Accurate loop structure prediction using a database search algorithm. *Proteins*, 78:1431–40, 2010.

Y. Choi and C. M. Deane. Predicting antibody complementarity determining region structures without classification. *Molecular BioSystems*, 7(12):3327–3334, 2011.

C Chothia and A M Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol.*, 4:901–17, 1987.

C Chothia, J Novotnỳ, R Bruccoleri, and M Karplus. Domain association in immunoglobulin molecules: the packing of variable domains. *Journal of molecular biology*, 186(3):651–663, 1985.

C Chothia, A M Lesk, A Tramontano, M Levitt, S J Smith-Gill, G Air, S Sheriff, E A Padlan, D Davies, W R Tulip, P M Colman, S Spinelli, P M Alzari, and R J Poljak. Conformations of immunoglobulin hypervariable regions. *Nature*, 342:877 – 883, 1989.

G Chuang, D Kozakov, R Brenke, S R Comeau, and S Vajda. Dars (decoys as the reference state) potentials for protein-protein docking. *Biophysical journal*, 95(9):4217–4227, 2008.

L A Clark, P Boriack-Sjodin, J Eldredge, C Fitch, B Friedman, K JM Hanf, M Jarpe, S F Liparoto, Y Li, A Lugovskoy, et al. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein science*, 15(5):949–960, 2006.

A V J Collis, A P Brouwer, and A C R Martin. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *Journal of molecular biology*, 325(2):337–354, 2003.

L L Conte, C Chothia, J Janin, et al. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5):2177–2198, 1999.

S Covaceuszach, A Cassetta, P V Konarev, S Gonfloni, R Rudolph, D Svergun, D I nad Lamba, and A Cattaneo. Dissecting ngf interactions with trka and p75 receptors by structural and functional studies of an anti-ngf neutralizing antibody. *J Mol Biol.*, 4:881–96, 2008.

P I W de Bakker, M A DePristo, D F Burke, and T L Blundell. Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the amber force field with the generalized born solvation model. *Proteins: Structure, Function, and Bioinformatics*, 51(1):21–40, 2003.

S J de Vries and A M J J Bonvin. Cport: a consensus interface predictor and its performance in prediction-driven docking with haddock. *PLoS One*, 6(3):e17695, 2011.

K A Dill, S B Ozkan, M S Shell, and T R Weikl. The protein folding problem. *Annual review of biophysics*, 37:289, 2008.

C Dominguez, R Boelens, and A M J J Bonvin. Haddock: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737, 2003.

D Duhovny, R Nussinov, and H J Wolfson. Efficient unbound docking of rigid molecules. *In Gusfield et al., Ed. Proceedings of the 2'nd Workshop on Algorithms in Bioinformatics(WABI) Rome, Italy, Lecture Notes in Computer Science*, 2452:185–200, 2002.

J Dunbar, A Fuchs, J Shi, and C M Deane. ABangle: characterising the VH-VL orientation in antibodies. *Protein Eng., Des. Sel.*, 26(10):611–620, 2013a.

J Dunbar, K Krawczyk, J Leem, T Baker, A Fuchs, G Georges, J Shi, and C M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, doi:gkt1043, 2013b.

R C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.

F Ehrenmann, Q Kaas, and M Lefranc. Imgt/3dstructure-db and imgt/domaingapalign: a database and a tool for immunoglobulins or antibodies, t cell receptors, mhc, igsf and mhcsf. *Nucleic acids research*, 38(suppl 1):D301–D307, 2010.

L P Ehrlich and R C Wadey. Protein–protein docking. *Reviews in Computational Chemistry, Reviews in Computational Chemistry*, page 61, 2003.

T Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

M Feldmann. Development of anti-tnf therapy for rheumatoid arthritis. *Nature Reviews Immunology*, 2 (5):364–371, 2002.

F. A. Fellouse, B. Li, D. M. Compaan, A. A. Peden, S. G. Hymowitz, and S. S. Sidhu. Molecular recognition by a binary code. *Journal of molecular biology*, 348(5):1153–1162, 2005.

N Fernandez-Fuentes, B Oliva, and A Fiser. A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic acids research*, 34(7):2085–2097, 2006.

A Fiser and A Šali. Modeller: generation and refinement of homology-based protein structure models. *Methods in enzymology*, 374:461–491, 2003.

J M Gershoni, A Roitburd-Berman, D D Siman-Tov, N T Freund, and Y Weiss. Epitope mapping. *BioDrugs*, 21(3):145–156, 2007.

V Giudicelli, P Duroux, C Ginestoux, G Folch, J Jabado-Michaloud, D Chaume, and M Lefranc. Imgt/ligm-db, the imgt® comprehensive database of immunoglobulin and t cell receptor nucleotide sequences. *Nucleic Acids Research*, 34(suppl 1):D781–D784, 2006.

R. Golub, J. S. Fellah, and J. Charlemagne. Structure and diversity of the heavy chain vdj junctions in the developing mexican axolotl. *Immunogenetics*, 46(5):402–409, 1997.

J J Gray, S E Moughan, C Wang, O Schueler-Furman, B Kuhlman, C A Rohl, and D Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol.*, 1:281–299, 2003.

I Halperin, B Ma, H Wolfson, and R Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 47(4):409–443, 2002.

R Hamer, Q Luo, J P Armitage, G Reinert, and C M Deane. i-patch: interprotein contact prediction using local network information. *Proteins*, 78:2781–97, 2010.

S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

A Hildebrand, M Remmert, A Biegert, and J Söding. Fast and accurate automatic structure prediction with hhpred. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):128–132, 2009.

J Huang and W Honda. Ced: a conformational epitope database. *BMC immunology*, 7(1):7, 2006.

H Hwang, T Vreven, J Janin, and Z Weng. Protein–protein docking benchmark version 4.0. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3111–3114, 2010.

S Idrees and U A Ashfaq. Structural analysis and epitope prediction of hcv e1 protein isolated in pakistan: an in-silico approach. *Virol J*, 10(1):113, 2013.

M B Irving, O Pan, and J K Scott. Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics. *Current opinion in chemical biology*, 5(3): 314–324, 2001.

L C James, P Roversi, and D S Tawfik. Antibody multispecificity mediated by conformational diversity. *Science*, 299(5611):1362–1367, 2003.

J Janin. Docking predictions of protein-protein interactions and their assessment: The capri experiment. In *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, pages 87–104. Springer, 2013.

N Jiang, J A Weinstein, L Penland, R A White III, D S Fisher, and S R Quake. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences*, 108(13):5348–5353, 2011.

R Jimenez, G Salazar, K K Baldridge, and F E Romesberg. Flexibility and molecular recognition in the immune system. *Proceedings of the National Academy of Sciences*, 100(1):92–97, 2003.

G Johnson and T T Wu. Kabat Database and its applications: future directions. *Nucleic Acids Res.*, 29(1):205–206, 2001.

P T Jones, P H Dear, J Foote, M S Neuberger, and G Winter. Replacing the complementarity-determining regions in a human antibody with those from a mouse. 1986.

L C U Junqueira, J Carneiro, R O Kelley, et al. Basic histology. 1998.

E A Kabat, T Te Wu, H M Perry, K S Gottesman, and C Foeller. *Sequences of proteins of immunological interest*. Diane Books Publishing Company, 1992.

W Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

P L Kastritis and A M J J Bonvin. Are scoring functions in protein- protein docking ready to predict interactomes? clues from a novel binding affinity benchmark. *Journal of proteome research*, 9(5): 2216–2225, 2010.

M B Khazaeli, R M Conry, and A F LoBuglio. Human immune response to monoclonal antibodies. *Journal of immunotherapy*, 15(1):42–52, 1994.

T J Kindt, R A Goldsby, B A Osborne, and J Kuby. *Kuby immunology*. WH Freeman & Company, 2007.

P M Kirkham, D Neri, and G Winter. Towards the design of an antibody that recognises a given protein epitope. *Journal of molecular biology*, 285(3):909–915, 1999.

G Köhler and C Milstein. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256(5517):495–497, 1975.

D Kozakov, R Brenke, S R Comeau, and S Vajda. Piper: An fft-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics*, 65(2):392–406, 2006.

K Krawczyk, T Baker, J Shi, and C M Deane. Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Protein Eng., Des. Sel.*, 26(10):621–629, 2013.

J V Kringelum, C Lundegaard, O Lund, and M Nielsen. Reliable b cell epitope predictions: impacts of method development and improved benchmarking. *PLoS computational biology*, 8(12):e1002829, 2012.

J V Kringelum, M Nielsen, S B Padkjær, and O Lund. Structural analysis of b-cell epitopes in antibody: protein complexes. *Molecular Immunology*, 53(1):24–34, 2013.

V Kunik and Y Ofran. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Engineering Design and Selection*, 2013.

V Kunik, B Peters, and Y Ofran. Structural consensus among antibodies defines the antigen binding site. *PLoS Comput Biol.*, 8:e1002388, 2012.

D Kuroda, H Shirai, M P Jacobson, and H Nakamura. Computer-aided antibody design. *Protein Engineering Design and Selection*, 25(10):507–522, 2012.

F Lara-Ochoa, J C Almagro, E Vargas-Madrazo, and M Conrad. Antibody-antigen recognition: a canonical structure paradigm. *J Mol Evol.*, 6:678–84, 1996.

J W Larrick and K E Fry. Recombinant antibodies. *Human Antibodies*, 2(4):172–189, 1991.

M C Lawrence and P M Colman. Shape complementarity at protein/protein interfaces. *Journal of molecular biology*, 234(4):946–950, 1993.

M Lee, P Lloyd, X Zhang, J M Schallhorn, K Sugimoto, A G Leach, G Sapiro, and K N Houk. Shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. *The Journal of organic chemistry*, 71(14):5082–5092, 2006.

M Lefranc et al. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.*, 37 (Database issue):D1006–12, 2009.

M P Lefranc. Imgt unique numbering for the variable (v), constant (c), and groove (g) domains of ig, tr, mh, igsf, and mhsf. *Cold Spring Harb Protoc.*, 6:633–42., 2011.

A Li, M Rue, J Zhou, H Wang, M A Goldwasser, D Neuberg, V Dalton, D Zuckerman, C Lyons, L B Silverman, et al. Utilization of ig heavy chain variable, diversity, and joining gene segments in children with b-lineage acute lymphoblastic leukemia: implications for the mechanisms of vdj recombination and for pathogenesis. *Blood*, 103(12):4602–4609, 2004.

B Li and D Kihara. Protein docking prediction using predicted protein-protein interface. *BMC Bioinformatics*, 13:7, 2012.

W Li and A Godzik. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

S M Lippow, K D Wittrup, and B Tidor. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol.*, 10:1171–6, 2007.

R M MacCallum, A C R Martin, and J M Thornton. Antibody-antigen interactions: contact analysis and binding site topography. *Journal of molecular biology*, 262(5):732–745, 1996.

A D MacKerell, D Bashford, M Bellott, R L Dunbrack, J D Evanseck, M J Field, S Fischer, J Gao, H Guo, S Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

P. Marcatili, A. Rosi, and A. Tramontano. Pigs: automatic prediction of antibody structures. *Bioinformatics*, 24(17):1953–1954, 2008.

A C R Martin. Accessing the kabat antibody sequence database by computer. *Proteins: Structure, Function, and Bioinformatics*, 25(1):130–133, 1996.

A C R Martin. Antibody Engineering Vol. 2. In *Antibody Engineering*, volume 2, chapter Protein Se, pages 33–51. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

A C R Martin and J M Thornton. Structural families of loops in homologous proteins: Automatic classification, modelling and application to antibodies. *J Mol Biol.*, 263:800–815, 1996.

F Matsuda, K Ishii, P Bourvagnet, K Kuma, H Hayashida, T Miyata, and T Honjo. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *The Journal of experimental medicine*, 188(11):2151–2162, 1998.

A J McCoy, V Chandana Epa, and P M Colman. Electrostatic complementarity at protein/protein interfaces. *Journal of molecular biology*, 268(2):570–584, 1997.

B A McKinney, N L Kallewaard, J E Crowe Jr, and J Meiler. Using the natural evolution of a rotavirus-specific human monoclonal antibody to predict the complex topography of a viral antigenic site. *Immunome Res*, 3:8, 2007.

R Mendez, R Leplae, M F Lensink, and S J Wodak. Assessment of capri predictions in rounds 35 shows progress in docking procedures. *BMC Bioinformatics*, 60:150–169, 2005.

I. S. Mian, A. R. Bradwell, and A. J. Olson. Structure, function and properties of antibody binding sites. *Journal of molecular biology*, 217(1):133–151, 1991.

J C Mitchell, R Kerr, and L F Ten Eyck. Rapid atomic density methods for molecular shape characterization. *Journal of Molecular Graphics and Modelling*, 19(3):325–330, 2001.

S Mohan, N Sinha, and S J Smith-Gill. Modeling the binding sites of anti-hen egg white lysozyme antibodies hyhel-8 and hyhel-26: an insight into the molecular basis of antibody cross-reactivity and specificity. *Biophysical journal*, 85(5):3221, 2003.

V Morea, A Tramontano, M Rustici, C Chothia, and A M Lesk. Antibody structure, prediction and redesign. *Biophysical chemistry*, 68(1):9–16, 1997.

V Morea, A Tramontano, M Rustici, C Chothia, and A M Lesk. Conformations of the third hypervariable region in the vh domain of immunoglobulins. *Journal of molecular biology*, 275(2):269–294, 1998.

S L Morrison, M J Johnson, L A Herzenberg, and V T Oi. Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *Proceedings of the National Academy of Sciences*, 81(21):6851–6855, 1984.

R Mosca, C Pons, J Fernandez-Recio, and P Aloy. Pushing structural information into the yeast interactome by high-throughput protein docking experiments. *PLoS Comput Biol.*, 5:e1000490, 2009.

J Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, 15(3):285–289, 2005.

J P Murad, O A Lin, E. E. V. Paez, and F T Khasawneh. Current and experimental antibody-based therapeutics: Insights, breakthroughs, setbacks and future directions. *Current molecular medicine*, 2012a.

M H Murad, M T Drake, R J Mullan, K F Mauck, L M Stuart, M A Lane, N O Abu Elnour, P J Erwin, A Hazem, M A Puhan, et al. Comparative effectiveness of drug treatments to prevent fragility fractures: a systematic review and network meta-analysis. *Journal of Clinical Endocrinology & Metabolism*, 97(6):1871–1880, 2012b.

M S Neuberger. Antibody diversification by somatic mutation: from burnet onwards. *Immunology and cell biology*, 86(2):124–132, 2008.

A Nicholls and B Honig. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the poisson–boltzmann equation. *Journal of computational chemistry*, 12(4):435–445, 1991.

B North, A Lehmann, and R L Dunbrack Jr. A new clustering of antibody cdr loop conformations. *J Mol Biol.*, 2:228–56, 2011.

B Oliva, P A Bates, E Querol, F X Avilés, M J Sternberg, et al. Automated classification of antibody complementarity determining region 3 of the heavy chain (h3) loops into canonical forms and its application to protein structure prediction. *Journal of molecular biology*, 279(5):1193, 1998.

R J Pantazes and C D Maranas. Optcdr: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng Des Sel.*, 11:849–58, 2010.

P A Patten, N S Gray, P L Yang, C B Marks, G J Wedemayer, J J Boniface, R C Stevens, and P G Schultz. The immunological evolution of catalysis. *Science*, 271(5252):1086–1091, 1996.

B. G. Pierce, Y. Hourai, and Z. Weng. Accelerating protein docking in zdock using an advanced 3d convolution library. *PloS one*, 6(9):e24657, 2011.

J Ponomarenko, N Papangelopoulos, D M Zajonc, B Peters, A Sette, and P E Bourne. IEDB-3D: structural data within the immune epitope database. *Nucleic acids research*, 39(Database issue): D1164–70, 2011.

J V Ponomarenko and P E. Bourne. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct Biol.*, 7:64, 2007.

G. Raghunathan, J. Smart, J. Williams, and J. C. Almagro. Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *Journal of Molecular Recognition*, 25(3):103–113, 2012.

R Rapberger, A Lukas, and B Mayer. Identification of discontinuous antigenic determinants on proteins based on shape complementarities. *Journal of Molecular Recognition*, 20(2):113–121, 2007.

J B Reece et al. *Campbell biology*. Benjamin Cummings/Pearson, 2011.

U Reimer. Prediction of linear b-cell epitopes. In *Epitope Mapping Protocols*, pages 335–344. Springer, 2009.

Ida Retter, H H Althaus, R Münch, and W Müller. VBASE2, an integrative V gene database. *Nucleic Acids Res.*, 33(Database issue):D671–674, 2005.

L Riechmann, M Clark, H Waldmann, G Winter, et al. Reshaping human antibodies for therapy. *Nature*, 332(6162):323–327, 1988.

J M Rini, U Schulze-Gahmen, and I A Wilson. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science*, 255(5047):959–965, 1992.

H Sakano, R Maki, Y Kurosawa, W Roeder, and S Tonegawa. Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. 1980.

M C Saraf, G L Moore, Nina M Goodey, Vania Y Cao, Stephen J Benkovic, and Costas D Maranas. Ipro: an iterative computational protein library redesign and optimization procedure. *Biophysical journal*, 90(11):4167–4180, 2006.

J F Schildbach, R I Near, R E Bruccoleri, E Haber, P D Jeffrey, J Novotny, S Sheriff, and M N Margolies. Modulation of antibody affinity by a non-contact residue. *Protein Science*, 2(2):206–214, 1993.

D Schneidman-Duhovny, Y Inbar, R Nussinov, and H J Wolfson. Patchdock and symmdock: servers for rigid and symmetric docking. *Nucl. Acids. Res.*, 33:W363–367, 2005.

H W Schroeder Jr. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Developmental & Comparative Immunology*, 30(1):119–135, 2006.

H W Schroeder Jr, G C Ippolito, and S Shiokawa. Regulation of the antibody repertoire through control of hcdr3 diversity. *Vaccine*, 16(14):1383–1390, 1998.

T Schwede, J Kopp, N Guex, and M C Peitsch. Swiss-model: an automated protein homology-modeling server. *Nucleic acids research*, 31(13):3381–3385, 2003.

J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl 2):W382–W388, 2005.

I Sela-Culang, V Kunik, and Y Ofran. The structural basis of antibody-antigen recognition. *Frontiers in Immunology*, 4:302, 2013.

H Shirai, A Kidera, and H Nakamura. Structural classification of cdr-h3 in antibodies. *FEBS letters*, 399(1):1–8, 1996.

B A Shoemaker, D Zhang, R R Thangudu, M Tyagi, J H Fong, A Marchler-Bauer, S H Bryant, T Madej, and A R Panchenko. Inferred biomolecular interaction servera web server to analyze and predict protein interacting partners and binding sites. *Nucleic acids research*, 38(suppl 1):D518–D524, 2010.

L Simonelli, M Pedotti, M Beltramello, E Livoti, L Calzolai, F Sallusto, A Lanzavecchia, and L Varani. Rational engineering of a human anti-dengue antibody through experimentally validated computational docking. *PloS one*, 8(2):e55561, 2013.

N Sinha, S Mohan, C A Lipschultz, and S J Smith-Gill. Differences in electrostatic properties at antibody–antigen binding sites: implications for specificity and cross-reactivity. *Biophysical journal*, 83(6):2946–2968, 2002.

A Sircar and J J Gray. Snugdock: Paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol.*, 6:e1000644, 2010.

A Sivasubramanian, A Sircar, S Chaudhury, and J J Gray. Toward high-resolution homology modeling of antibody fv regions and application to antibody-antigen docking. *Proteins*, 74:497–514, 2009.

G P Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317, 1985.

G R Smith and M J Sternberg. Evaluation of the 3d-dock protein docking suite in rounds 1 and 2 of the capri blind trial. *Proteins*, 52:7479, 2003.

S Soga, D Kuroda, H Shirai, M Kobori, and N Hirayama. Use of amino acid composition to predict epitope residues of individual antibodies. *Protein Engineering Design and Selection*, 23(6):441–448, 2010.

R R Sokal. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–1438, 1958.

R L Stanfield, M Takimoto-Kamimura, J M Rini, A T Profy, and I A Wilson. Major antigen-induced domain rearrangements in an antibody. *Structure*, 1(2):83–93, 1993.

R L Stanfield, H Dooley, P Verdino, M F Flajnik, and I A Wilson. Maturation of shark single-domain (ignar) antibodies: evidence for induced-fit binding. *Journal of molecular biology*, 367(2):358–372, 2007.

J Sun, T Xu, S Wang, G Li, D Wu, and Z Cao. Does difference exist between epitope and non-epitope residues? analysis of the physicochemical and structural properties on conformational epitopes from b-cell protein antigens. *Immunome Research*, 7(3), 2011.

P Sun, H Ju, Z Liu, J Zhang, X Zhao, Y Huang, Z Ma, and Y Li. Bioinformatics resources and tools for conformational b-cell epitope prediction. *Comput Math Methods Med.*, page 2013: 943636, 2013.

M C Thielges, J Zimmermann, W Yu, M Oda, and F E Romesberg. Exploring the energy landscape of antibody- antigen complexes: protein dynamics, flexibility, and molecular recognition. *Biochemistry*, 47(27):7237–7247, 2008.

A Tovchigrechko, C A Wells, and I A Vakser. Docking of protein models. *Protein Sci.*, 11:18881896, 2002.

S Vajda. Classification of protein complexes based on docking difficulty. *Proteins: Structure, Function, and Bioinformatics*, 60(2):176–180, 2005.

R Vita, L Zarebski, J A Greenbaum, H Emami, I Hoof, N Salimi, R Damle, A Sette, and B Peters. The immune epitope database 2.0. *Nucleic acids research*, 38(suppl 1):D854–D862, 2010.

K L Wark and P J Hudson. Latest technologies for the enhancement of antibody affinity. *Adv Drug Deliv Rev.*, 5:657–70, 2003.

G J Wedemayer, L H Wang, P A Patten, P G Schultz, and R C Stevens. Crystal structures of the free and liganded form of an esterolytic catalytic antibody. *Journal of molecular biology*, 268(2):390–400, 1997.

C X Weichenberger, E Pozharski, and B Rupp. Visualizing ligand molecules in twilight electron density. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.*, 69(Pt 2):195–200, 2013.

J A Weinstein, N Jiang, R A White, D S Fisher, and S R Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810, 2009.

N R Whitelegg and A R Rees. Wam: an improved algorithm for modelling antibodies on the web. *Protein Eng.*, 12:819–24., 2000.

I. A. Wilson and R. L. Stanfield. Antibody-antigen interactions. *Current opinion in structural biology*, 3(1):113–118, 1993.

T T Wu and E A. Kabat. 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.*, 2:211–50, 1972.

J L Xu and M M Davis. Diversity in the cdr3 region of vh is sufficient for most antibody specificities. *Immunity*, 13(1):37–45, 2000.

B Yao, D Zheng, S Liang, and C Zhang. Conformational b-cell epitope prediction on antigen protein structures: A review of current algorithms and comparison with common binding site prediction methods. *PloS one*, 8(4):e62249, 2013.

C Yu, H Peng, C Chen, Y Lee, J Chen, K Tsai, C Chen, J Chang, E Yang, P Hsu, et al. Rationalization and design of the complementarity determining region sequences in an antibody-antigen recognition interface. *PLoS One*, 7(3):e33340, 2012.

M. Zemlin, M. Klinger, J. Link, C. Zemlin, K. Bauer, J. A. Engler, H. W. Schroeder, and P. M. Kirkham. Expressed murine and human cdr-h3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *Journal of molecular biology*, 334 (4):733–749, 2003.

Q C Zhang, L Deng, M Fisher, J Guan, B Honig, and D Petrey. Predus: a web server for predicting protein interfaces using structural neighbors. *Nucleic acids research*, 39(suppl 2):W283–W287, 2011.

X W Zhang. A combination of epitope prediction and molecular docking allows for good identification of mhc class i restricted t-cell epitopes. *Computational biology and chemistry*, 2013.

L Zhao and J Li. Mining for the antibody-antigen interacting associations that predict the b cell epitopes. *BMC structural biology*, 10(Suppl 1):S6, 2010.

L Zhao, L Wong, and J Li. Antibody-specified b-cell epitope prediction in line with the principle of context-awareness. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(6): 1483–1494, 2011.

J Zimmermann, F E Romesberg, Charles L Brooks I, and I F Thorpe. Molecular description of flexibility in an antibody combining site. *The journal of physical chemistry B*, 114(21):7359–7370, 2010.