Variable domain orientations in antigen receptors

James Dunbar

Department of Statistics

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

TT 2014

For Mum, Dad and Ellie.

Acknowledgements

Over the last four years I have enjoyed the support of a number of people who made this thesis possible. I would most like to thank Charlotte for her dedicated supervision during my DPhil, the advise she has given me and for making OPIG such an enjoyable place to have worked. I feel privileged to have had her as a supervisor. My industrial supervisors, Angelika and Jiye, have provided new and interesting perspectives on my research for which I am very grateful. Additionally I thank Seb, Terry, Alex and Guy for the conversations and collaboration we have had. I am grateful for the funding provided by the ESPRC, UCB and Roche and the training I received during my time at the SABS-IDC DTC.

Writing this thesis has been made much more enjoyable by the encouragement, coffee breaks and crosswords shared with past and present members of OPIG. I would especially like to thank KK for pioneering the "antibody division" and Hannah and Henry who have been my fellow thesis-dungeon inmates for the last three months. My thanks also go to several members of OPIG for their inspiration for interactive elements of this thesis.

Finally I would like to thank my friends and family for their support and giving me perspective of the world away from wobbling proteins.

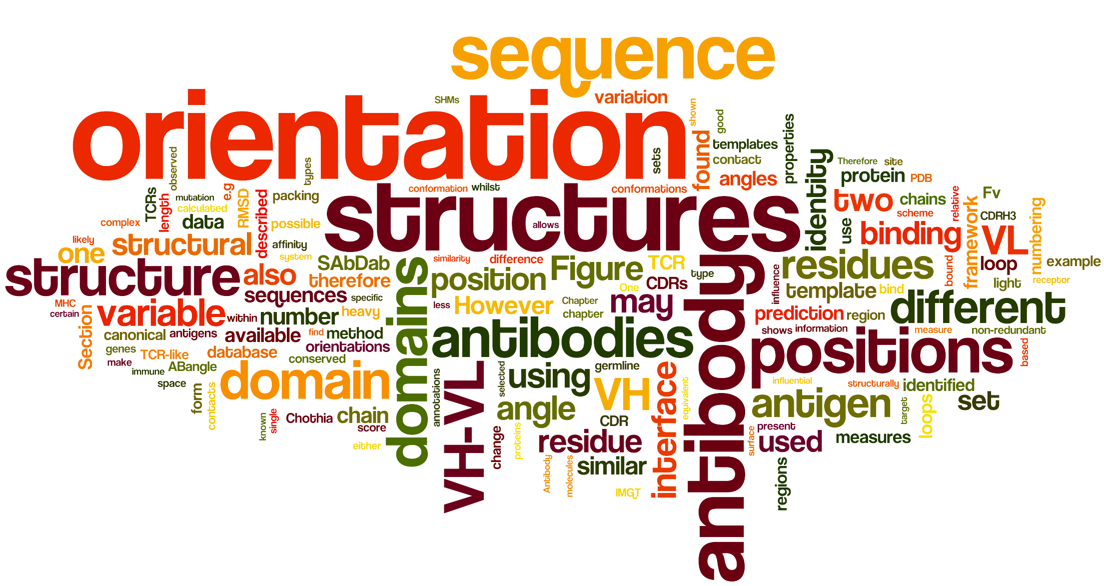
Abstract

Specific recognition of pathogenic molecules by the immune system is mediated by proteins known as antigen receptors. One such component is the antibody. Binding properties of natural and engineered antibodies can be understood by studying the structure of their variable domains, VH and VL. In this thesis we investigate how the two variable domains orientate with respect to one another and therefore influence the geometry of the antigen binding site which is formed between them.

We describe a method which fully characterises the VH-VL orientation in a consistent and absolute sense using five angles and a distance. The ABangle method is used to investigate variable domain orientation in structures collected by our database SAbDab.

Using the ABangle method we compare VH-VL orientation to the corresponding property in a different component of the immune system, the T-cell receptor (TCR). Despite having similar individual domain structures the variable domain orientations of antibodies and TCRs are found to be distinct. This is found to affect an antibody's ability to mimic TCR specificity.

ABangle's characterisation is used to find determinants of the VH-VL orientation. We identify sequence and structural properties that influence the variable domain pose. A feature based method for predicting VH-VL orientation is presented and assessed. Future directions of this research and its application to the development of antibody therapeutics are described.



sequence
orientation
structures
antibody
structure
antibodies
domains
VH-VL
domain
positions
different
residues
antigen
structural
variable
identity
binding
VL
VH
two
position
Figure
interface
residue
using
angle
may
However
also
therefore
SAbDab
one
sequences
number
available
database
template
prediction
framework
loops
set
used
similar
measures
protein
chains
angles
found
templates
contact
properties
variation
chain
Chothia
canonical
Section
data
packing
possible
described
RMSD
length
TCRs
TCR
CDRs
loop
light
numbering
example
identified
region
information
orientations
method
heavy
conserved
ABangle
TCR-like
genes

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

Nomenclature

CDR	Complementarity Determining Region
dc	Separation distance between VH and VL domains
Fab	Antigen binding fragment of an antibody
Fv	Variable fragment of an antibody
HC1	Tilt angle of VH domain
HC2	Twist angle of VH domain
HL	Torsion angle between VH and VL domains
LC1	Tilt angle of VL domain
LC2	Twist angle of VL domain
RMSD	Root mean square deviations. This is calculated using only Cα coordinates unless otherwise stated
TCR	T-Cell receptor
Vα	Alpha chain variable domain of a T-Cell receptor
Vβ	Beta chain variable domain of a T-Cell receptor
VH	Heavy chain variable domain of an antibody
VL	Light chain variable domain of an antibody

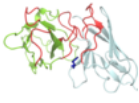
Chapter 1

Introduction and Background

Our immune response relies on being able to recognise those substances that may do us harm. Key components that mediate specific pathogen recognition are the antigen receptors, one type of which is the antibody. These proteins act as the natural immune system's molecular sensors and also form the fastest growing type of pharmaceutical currently in development. Studying their structural properties will allow us to understand how proteins can recognise ligands and inform the rational engineering of therapeutic agents.

This thesis focuses on the structural variation of antibodies. In particular, the relative orientation between two subunits of these molecules, the variable domains. The antibody's binding site is found between its variable domains. We investigate this structural property in antibodies and its analogue in a similar protein, the T-cell receptor. In doing so we aim to provide insight as to how to better engineer therapeutic antibodies and to understand the mechanisms that proteins may use to determine how their domains associate.

In this chapter we introduce the background and context to the thesis. We



briefly describe the sequence, structure and function of proteins. This is followed by a discussion of the antibody's role in the immune system, their structure and how they may be engineered as therapeutics. The chapter is concluded by outlining the contents of the following sections of the thesis.

1.1 Protein sequence, structure and function

Proteins are the machinery of the cell and are the actors of the functional information encoded on the genome. Each protein is formed by a linear sequence of amino-acid residues that, in most cases, fold into three dimensional structures. The chemical and structural properties of protein molecules allow them to perform their numerous biological roles, one of which is the defence of the organism.

1.1.1 Protein sequence

A protein is encoded on the genome by an alphabet of four DNA bases: adenine (A), guanine (G), thymine (T) and cytosine (C). Sequences of DNA nucleotides, genes, are transcribed into RNA, introns are spliced out and the sequence is translated by the ribosome into protein chains consisting of 20 different amino-acid types (Figure 1.1). Each amino-acid residue is covalently bonded to the next through a peptide bond (Figure 1.2a). Protein sequences are conventionally written from N-terminus to C-terminus reflecting the order that the polypeptide chain emerges from the ribosome (N-terminus first).

The side chain, or functional-group (R), of the amino acid gives each residue specific chemico-physical properties. For example, tryptophan has a large aromatic side chain making it hydrophobic. This residue therefore tends to pack into the

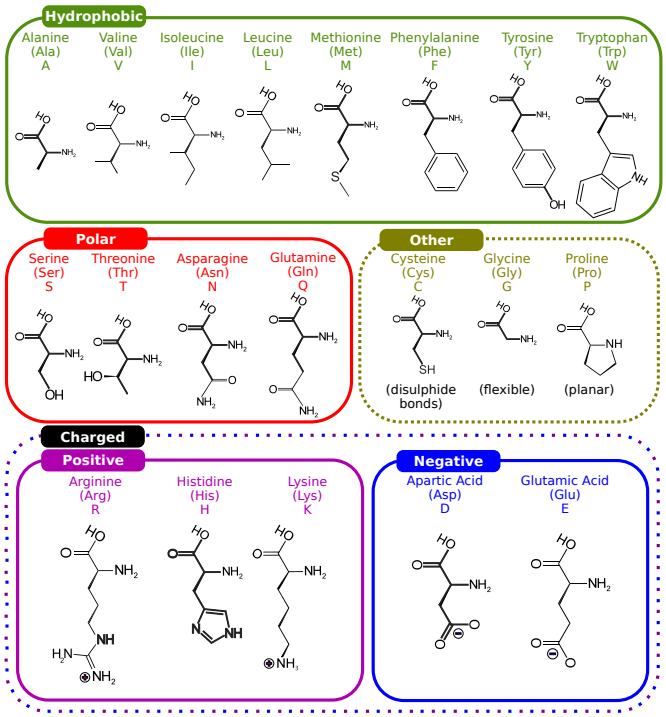
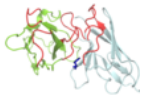


Figure 1.1: The 20 amino-acids that occur naturally as protein residues. Certain residues have similar properties. Here they have been grouped into hydrophobic, polar and charged. Three residues do not fall into these categories and are grouped as "other".



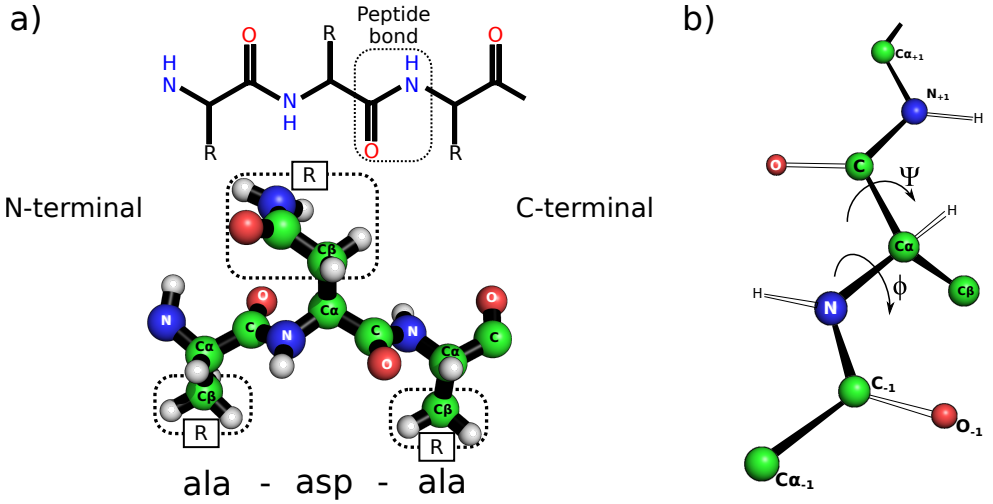


Figure 1.2: a) The peptide bond (top) and the atomic structure of a short polypeptide (bottom). Sequences are conventionally written from the N terminus to the C terminus. When translated, the N terminal end of the protein emerges from the ribosome first. Each of the backbone heavy atoms of the peptide have been labelled. The R group of each residue is indicated. These groups are different for each residue and are shown in Figure 1.1. b) The backbone dihedral angles. These can be used to define the conformation of the main chain of a protein. ϕ is the dihedral angle between the C_{-1}-N-Cα-C atoms. ψ is the dihedral angle between the N-Cα-C-N_{+1} atoms. Different residues have different allowed combinations of ϕ-ψ angles which influence the conformation of the polypeptide chain.

centre of the protein away from the solvent. Other residues are charged and are often involved in interactions with other proteins. Whereas glycine when present can act to allow the protein backbone to achieve different conformations. This is due to a greater allowed space of ϕ-ψ backbone dihedral angles (Figure 1.2b) owing to the residue's lack of a side chain (Figure 1.1).

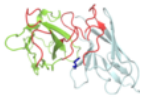
On an evolutionary time scale protein residues may mutate, be inserted or be deleted. Homologous proteins can therefore have different sequences but a generally conserved shape. Related protein sequences can be aligned (e.g. Needleman & Wunsch [1970]; Smith & Waterman [1981]; Notredame *et al.* [2000]; Edgar [2004]). This allows a correspondence between residues in different proteins to be assigned.

1.1.1.1 Sequence identity and sequence similarity

Once aligned, the identity and similarity between a pair of sequences can be calculated. These give measures of how related the two sequences are. Matched sequence identity counts the number of positions in the alignment where both sequences have the same residue type present and neither is a gap. A percentage sequence identity is calculated by dividing this count by the total number of positions at which both sequences have a residue present. Values of sequence identity given in this thesis refer to the matched sequence identity.

Sequence similarity accounts for how related two residues are. To calculate sequence similarity one requires a dissimilarity matrix. This gives a score for each possible residue-residue or residue-gap combination.

A commonly used matrix is the BLOSUM62 matrix [Henikoff & Henikoff, 1992]. BLOSUM62 was derived using alignments of related sequences that shared no more than 62% sequence identity over the aligned segment or block. Its integer log-odds



score represents how likely, compared to random, one amino-acid is to mutate to another at a residue-position, or column, in the alignment. Therefore, a positive score means that the mutation is more likely than random and is large for the value of the amino-acid compared to itself. A negative score means that the mutation is less likely than random. Large negative scores mean that the mutation is unlikely and the evolutionary cost of performing it is high.

As an example, changing a hydrophobic leucine to a negatively charged aspartic-acid has a score of -4. By contrast, changing to an isoleucine scores 2. Leucine and isoleucine are similar (see Figure 1.1). The similarity between two sequences is the summation of the similarity score over each position in the alignment. It is therefore a function of the length of the alignment. However, it may be rescaled by dividing by the length or normalised by dividing by the maximum possible similarity score for the sequence pair.

1.1.2 Secondary structure

A protein's sequence largely determines its structure [Anfinsen, 1973]. Protein secondary structure describes the local shape that the peptide chain forms. Most secondary structure is of three types: α-helices, β-sheets and loops or coils (Figure 1.3). β-sheets consist of multiple β-strands. These strands are connected either in parallel, with both strands running from N to C terminal, or anti-parallel where the strands run in opposite directions. Both have distinctive hydrogen bonding patterns (Figure 1.4).

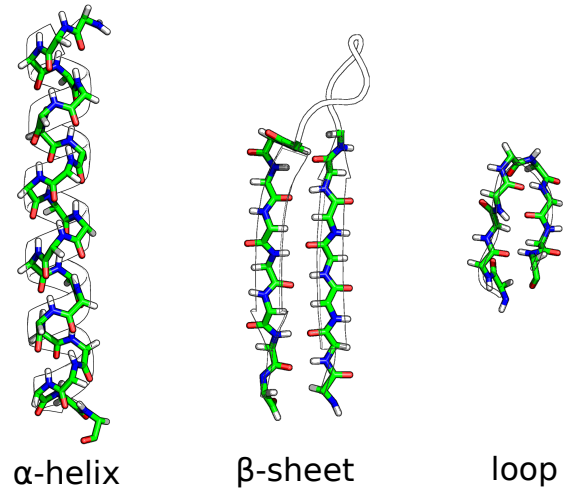


Figure 1.3: Three main types of protein secondary structure: α-helix, β-sheet and loop. The main chain of the protein is shown in stick representation. Nitrogen atoms are coloured blue, oxygen atoms red, carbon atoms green and hydrogen atoms white. The cartoon representation of each secondary structure type is shown as a black outline.

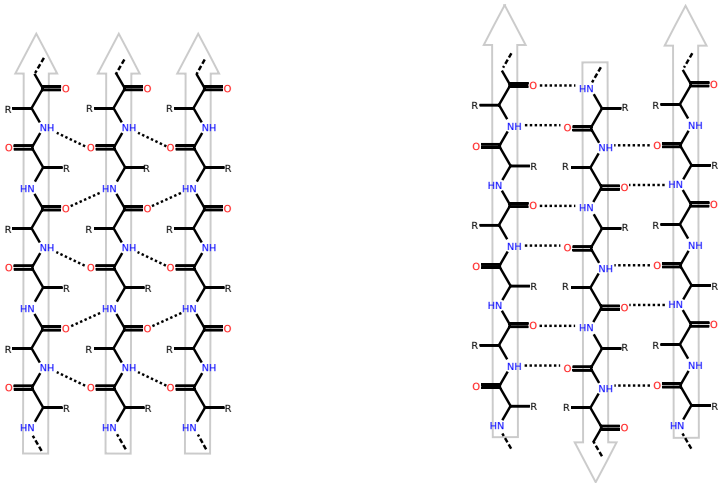
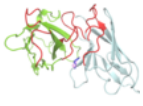


Figure 1.4: Parallel (left) and anti-parallel (right) β-strands. The direction of each strand is indicated by the arrows. Main chain hydrogen bonds are shown as dashed lines black lines. For these bonds the main chain nitrogen acts as the donor whilst the carbonyl oxygen acts as the acceptor. The hydrogen bonding patterns are distinctive for both types of β-sheet. Antibody structures consist of anti-parallel β-sheets.



1.1.3 Domains and tertiary structure

Elements of secondary structure are organised into more complex tertiary structure. Protein domains are semi-independent folding units of compact secondary structure [Richardson, 1981; Koonin *et al.*, 2002]. A wide range of domain sizes have been observed, but typically they contain 100-250 residues. Each has a particular shape, fold, that describes its topology. Domains can be classified by their sequence, fold and function [Lo Conte *et al.*, 2000; Orengo *et al.*, 1999]. Certain domains exist in different combinations and arrangements in the protein structure universe with the same units performing varied biological functions [Koonin, 2003; Lee *et al.*, 2005]. They are therefore thought to be evolutionary building blocks [Ohno, 1970].

1.1.4 Quaternary structure

Many proteins contain multiple domains and can be formed from more than one subunit or chain. The arrangement of these elements is called the quaternary structure of the protein. Advantages of having a quaternary level of structure include multiple functionality of the protein, increased stability, signalling through conformational changes and the ability to form large structures whilst reducing the risk of translational error that increases significantly with chain length [Kessel & Ben-Tal, 2012]. Arrangements of different subunits also allows for the construction of more diverse active sites than would be possible from those built from a single chain. Examples of this mechanism include the active sites of HIV-proteases and, pertinent to this thesis, the immune system's antigen receptors.

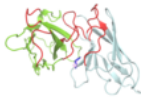
1.1.5 Structure determination

Protein structural data can be obtained using several different techniques. The three most commonly used are Nuclear Magnetic Resonance (NMR), Electron Microscopy and X-Ray Crystallography.

Structure determination using NMR measures the reactions of the nuclei of atoms in the protein to a varying magnetic field. As each atom has a different chemical environment they can be distinguished from one another and set of distance constraints inferred. One advantage of NMR techniques are that they give a measure of uncertainty of distances between atoms [Branden & Tooze, 1991]. This may be interpreted as the flexibility of a protein. However, only small to medium sized proteins can be studied with this technique (less than about 40kDa). Structures solved using NMR are considered to be of relatively low resolution (a measure of the confidence with which the position of each atom is known or the level of detail that can be interpreted from the data).

Electron Microscopy also determines structures at low resolution. However, it is not limited by protein size and is therefore commonly used to investigate the structure of protein complexes [Jonic & Vénien-Bryan, 2009].

X-Ray Crystallography is the most prevalent technique. Obtaining crystals of the protein is key to the process. This is a difficult procedure that often requires testing of many different conditions and modifications to the natural protein [Branden & Tooze, 1991]. Protein crystals are exposed to an X-Ray beam and the diffraction pattern is collected. The patterns can then be used to infer the electron density in the unit cell of the crystal. A model of the protein structure is then fitted into the density iteratively.



The protein data bank (PDB) [Berman *et al.*, 2000] contains structures of proteins that have been determined in the public literature. As of September 2014 it contains over 100,000 entries, many of which are redundant in sequence. Approximately 89% of structures have been solved by X-Ray Crystallography, 10% by NMR and 1% by Electron Microscopy.

1.2 Antibodies and the immune response

The immune response protects an organism against the pathogenic effects that foreign molecules, other organisms or harmful cellular components may have on host tissues. As a first line of defence, the innate immune response can react and defend against pathogens in a non-specific way [Janeway *et al.*, 2001]. For example, bacteria that enter the human body can often be recognised from the common constituents on their cell surface. They are engulfed (phagocytosed) and digested by a type of white blood cell, the macrophage.

However, some pathogens do not have, or disguise, features that allow them to be easily recognised by the innate immune response. In most vertebrates, the adaptive immune response provides mechanisms for protecting against foreign bodies that the organism may not have previously encountered or the species not evolved an innate response against. What distinguishes the adaptive from the innate immune response is the specificity with which it reacts to markers for a particular pathogen, an antigen. Antigens range from cell-surface proteins to nucleic acids to non-biological small molecules. Antigen recognition is mediated by antigen receptors, one type of which is the antibody.

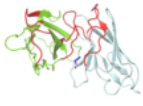
1.2.1 The biological function of antibodies

Antibodies, also known as immunoglobulins, are produced in response to the presence of non-self antigens. Their function is to bind to the antigen and to disable it and the pathogen for which it may be a marker. It does this either by direct inhibition or by recruiting other parts of the immune system. A mature antibody is specific for a particular part of, or patch on, the surface of an antigen, known as the epitope. Here, specific means that an antibody only binds to one epitope and does not bind (or only weakly binds) to other molecules or surfaces. Antibodies bind with high affinity (strongly) to the epitope for which they are specific.

The effector function of an antibody refers to how it inhibits, or helps to dispose of, an antigen. Antibodies have three main mechanisms by which they can do this [Janeway *et al.*, 2001] (Figure 1.5). The first is neutralisation. Here, the antibody binds to an antigen and prevents it from interacting with cellular components of the host organism's tissues. For example, bacteria may release toxins that can bind receptors on the surface of cells. When an antibody binds to a toxin antigen, the toxin can no longer form a complex with the receptor. The toxin's harmful effect is therefore neutralised and they are eventually disposed of by macrophages.

Second is a process called opsonisation. By binding to an antigen, the antibody acts to mark the molecule as pathogenic. Macrophages and neutrophils can recognise part of the antibody and phagocytose both it and the antigen. Bacterial or viral cell surfaces may have many antigens (e.g. surface proteins) to which antibodies can bind. Antibodies, can therefore form a coat (of opsonins) on the foreign organism allowing the whole pathogen to be recognised and ingested by phagocytes.

The third is through the activation of the complement system. Antibodies that



bind to the surface of pathogenic cells can also recruit parts of the innate immune system other than phagocytes. The activated complement can attack the surface of the cell, lysing through the membrane and inducing cell death. Similarly, all components of the antibody, antigen and pathogenic cell are phagocytosed by macrophages.

The effector function of an antibody is determined by its isotype. In humans, there are five different isotypes, IgA, IgD, IgE, IgG and IgM, each with different roles in the immune system [Alberts *et al.*, 2007] (Figure 1.6). For example, IgMs are the first antibody to be produced in response to an antigen and can be soluble or membrane-bound but bind with a relatively weak affinity. However, they can form pentamers and are particularly efficient at activating the complement system [Boes, 2000]. A different isotype, IgD, is usually a membrane-bound antibody on the surface of B-lymphocytes where they act as receptors for the cell. Antibodies in blood serum are mostly IgGs. These have several subtypes and have the widest range of possible effector functions.

1.2.2 In vivo antibody generation through clonal selection

The adaptive immune response involving antibodies is referred to as the humoral response. Antibodies are produced through a mechanism called clonal selection [Burnet, 1957] which is summarised in Figure 1.7 and described below. As mentioned earlier, an antibody will bind specifically to an epitope on an antigen. Antigens can potentially be any shape or size with a range of chemical properties. The humoral immune response must therefore be able to respond with a complementary diverse set of potential antibodies.

The first level of antibody diversity is encoded in the primary B-cell repertoire.

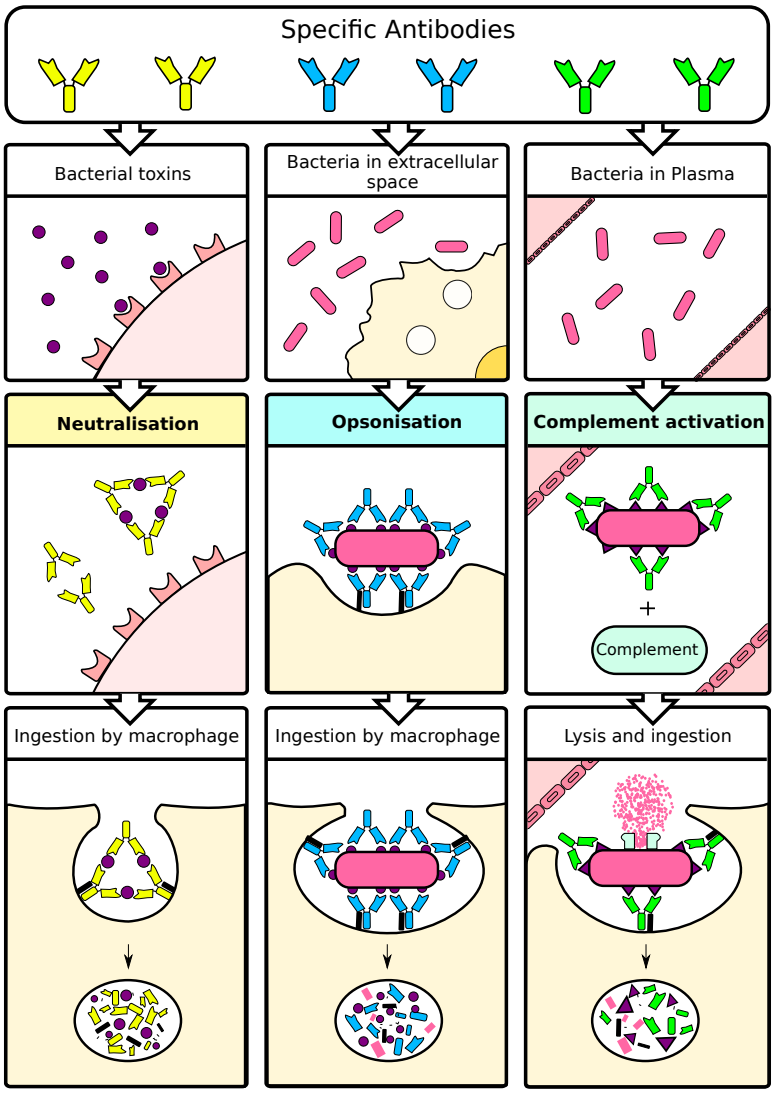
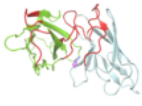


Figure 1.5: Examples of antibody effector functions. Antibodies recognise and bind specifically to particular antigens. Antigens are themselves or are markers for, pathogens. Antibodies defend against pathogens in three main ways. **Neutralisation**: antibodies bind to antigens (e.g. a bacterial toxin) and prevent them from interacting with the host cells by inhibiting binding with receptors. The complex will eventually be engulfed and degraded by macrophages. **Opsonisation**: antibodies coat an antigen making it recognisable as foreign by phagocytes (macrophages and neutrophils). **Complement activation**: once antibodies have bound to an antigen they act as receptors for certain components of the innate immune system (collectively the complement system). These components attack the pathogen (e.g. bacteria) by, for example, lysing the cell membrane, again eventually being ingested by phagocytes. Figure adapted from Janeway *et al.* [2001].



Effector function	IgM	IgD	IgG1	IgG2	IgG3	IgG4	IgA	IgE
Neutralisation	+		++	++	++	++	++	
Opsonisation			+++		++	+	+	
Activation of the complement system	+++		++	+	+++		+	
Sensitisation of mast cells			+		+			+++
Mean serum level (mg ml^{-1})	1.5	0.04	9	3	1	0.5	2.1	$3x10^{-6}$

Figure 1.6: Effector functions of human antibody isotypes. Major functions of each isotype are indicated by $+++$, minor functions by $++$ and very minor functions by $+$. Neutralisation, opsonisation and activation of the component system are described in Figure 1.5. The isotype with the smallest concentration in the blood is IgE. Once produced, receptors on mast cells bind to it with very high affinity. This cell makes the cell sensitive for the antigen that the IgE is specific for. When the antigen is re-encountered, the mast cells will release histamines, and play a key role in allergic reactions. The mean blood serum concentration of each isotype is shown on the last row. Information summarised from Janeway *et al.* [2001] and Schroeder & Cavacini [2010]

Each single B-cell produces antibodies with an identical specificity. Different B-cells in the repertoire produce different antibodies, the genetic mechanism behind which is described in Section 1.2.3.1. At this stage, IgM antibodies are expressed on the surface of a B-cell and act as receptors that will recognise the same antigen as the soluble antibodies the cell can produce. Only those B-cells which do not bind to antigens from the host tissue (self-antigens) are transported from the bone-marrow to the lymph nodes. These B-cells express IgD (and IgM) on their surface. Once in the lymph nodes, B-cells are presented with antigen [Gonzalez *et al.*, 2011]. Those B-cells whose antibody binds to the antigen have the potential of being activated.

For protein antigens and where the antigen is bound or presented by a larger molecule or cell (e.g. a virus), activation can be aided (helped) by other parts of the adaptive immune system. The activation process begins by the B-cell internalising the bound antigen and its connected components. Proteins from the internalised matter

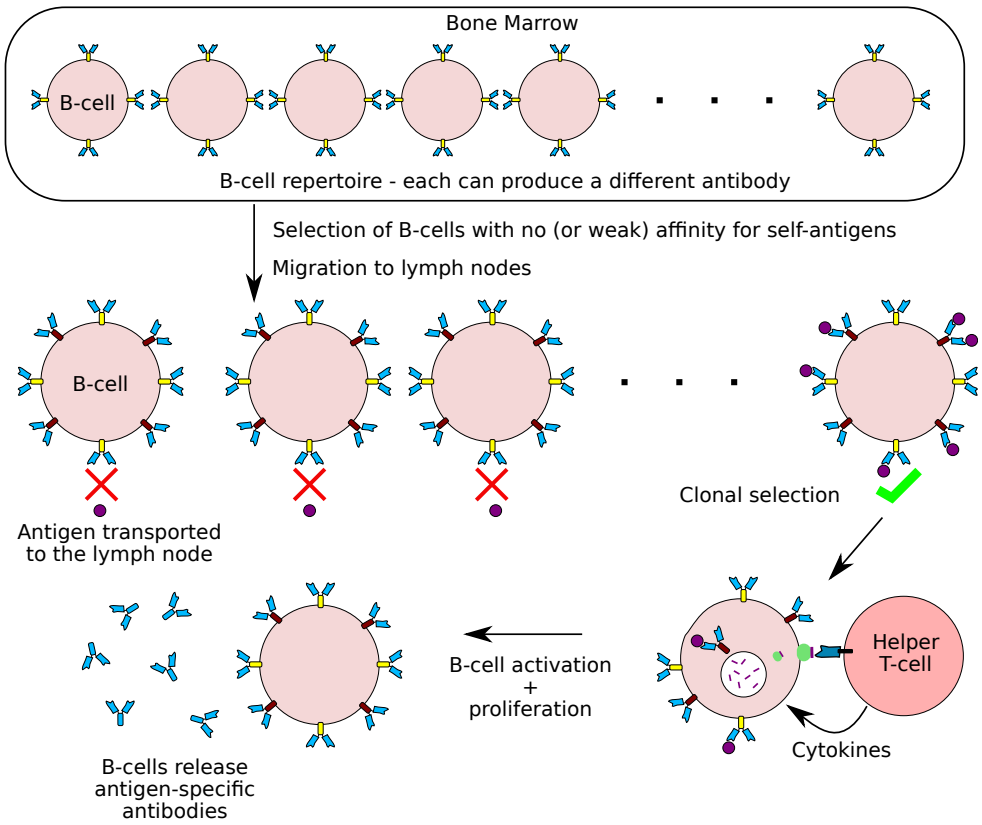
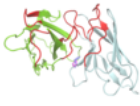


Figure 1.7: Clonal selection and activation of B-Cells. Immature B-cells are produced in the bone marrow. Each one produces antibodies with a different specificity. Diversity at this stage is achieved by combining different genes as described in Section 1.2.3.1 and is known as the germline repertoire. Immature B-cells express IgM antibodies on their surface (yellow-stemmed Y's) which act as receptors for the B-cell. Those B-cells that are able to bind antigens from host tissues (self-antigens) do not progress out of the bone marrow or spleen (negative selection). Others migrate to the lymph nodes as mature B-cells, expressing IgD on their surface (brown-stemmed Y's). Here, they can be exposed to an antigen. Those that the antigen binds to are selected. Certain antigens can activate the B-cell directly. Protein antigens often require activation by helper T-cells. B-cells present peptide fragments of the antigen they have internalised on their surface using MHCII's. T-cells recognise immunogenic peptides and release cytokines to help activate the B-Cells. Once activated, the B-Cell will proliferate and begin to produce soluble antibodies (initially IgM, later predominantly IgG). Further rounds of clonal selection will act to increase the affinity with which the antibodies produced will bind specifically to the antigen. A high rate of somatic hyper-mutation enables this (Section 1.2.3.2).



are digested into peptide fragments, loaded onto Major Histocompatibility Complex Class II (MHCII) and presented back onto the surface of the B-cell. Helper T-cells recognise those peptides that are immunogenic (i.e. non-self) and release cytokines to activate the B-cell. Therefore, the peptide the T-cell receptor recognises and the epitope the B-cell receptor (antibody) recognises do not have to be overlapping. In fact, they can be from different parts of the pathogen e.g. a surface protein recognised by the antibody and a peptide fragment of an intra-cellular protein recognised by the T-cell receptor. The interaction between the peptide, MHCII and T-cell receptor will be discussed in Chapter 4. Other antigen types such as bacterial toxins, can activate a B-cell without the help of other components of the immune system.

Once activated, a B-cell will proliferate and begin to release soluble antibodies (IgM initially). At this stage the antibodies may only have a weak affinity to the antigen they are specific for. As the B-cells proliferate they go through multiple rounds of clonal selection [Rajewsky, 1996]. Each new generation acquires somatic mutations to the genes that encode for the antibody the B-cell produces. Thus, those clones with mutations that improve antibody-antigen binding properties proliferate in a process called affinity maturation (Section 1.2.3.2). This is a further mechanism for achieving the antibody diversity required specifically to complement the vast space of possible antigens.

The final B-cell clones produce antigen-specific, high affinity antibodies. When transported from the lymph node into the blood stream they release predominantly IgG istype antibodies and perform the functions depicted in Figure 1.5. The time taken for a humoral response can be of the order of days if the organism has not previously been exposed to an antigen. However, once a high-affinity-antibody producing B-cell has been generated, a number of these cells are stored by the immune system as memory

B-cells. Re-exposure to the same antigen re-activates the appropriate memory B-cell and thus a much faster humoral response is possible. This is key to the concept of vaccination.

1.2.3 Sources of antibody diversity

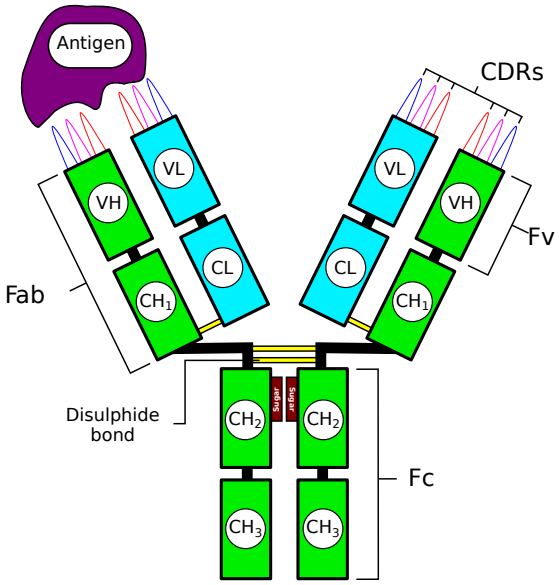


Figure 1.8: A schematic of an IgG antibody. The molecule consists of four polypeptide chains, two heavy (green) and two light (cyan). The variable domains, VH and VL, associate to form variable regions (Fv). The Fv, along with the first constant domains, CH1 and CL1, are known as antigen binding fragments (Fab). The remaining heavy constant domains associate to form the crystallisable or constant fragment (Fc). The complementarity determining regions (CDRs) mediate antigen binding. Three are located on each of the VH domain (H1, H2 and H3) and the VL domain (L1, L2 and L3). They form the majority of the antigen binding site.

The diversity with which different antibodies recognise different antigens, and perform different functions can be understood by examining the protein structure of the antibody (Figure 1.8). Antibodies have a highly conserved Y shaped structure consisting of four polypeptide chains. Two of the chains are identical and called light chains (coloured cyan throughout this thesis). The other two chains are longer and

are also identical to each other. These are called the heavy chains (coloured green throughout this thesis).

Each chain folds independently and has multiple domains, two for the light chains and four or more for the heavy chains. The stem of the Y is known as the crystallisable/constant fragment (Fc). The Fc is responsible for determining the effector function of the antibody. In natural antibodies both arms of the Y are identical. Each arm is known as an antigen-binding fragment (Fab). Thus, natural antibodies can bind to two antigens simultaneously (bivalently).

At the tip of the Fab are the heavy and light variable domains, VH and VL respectively. VH and VL are known collectively as the variable region of the antibody (Fv). As their names suggest, the sequence and parts of the structure vary between different antibodies. The other domains in the antibody have the same sequence between antibodies of the same isotype. These are therefore called constant domains: CL for the light chain, CH1-3 for IgG heavy chains.

The Fv region is responsible for binding to an antigen. The residues the Fv uses to associate with the antigen's epitope is known as the paratope. Within both the VH and VL domains are three linear regions that are "hyper-variable" in sequence. They coincide with residues that are thought to form the paratope (discussed in Section 1.3.2.3). Here, we refer to them as the six complementarity determining regions (CDRs). They are denoted as L1, L2 and L3 on VL and H1, H2 and H3 on VH. CDR H3 is more variable in length, sequence and subsequently structure than the other five CDRs.

Antibody diversity is found in the sequence of the VH and VL domains. The genomic mechanism by which the expressed B-cell repertoire is created provides the first explanation of the achieved variability of the antibody Fv.

1.2.3.1 Genomic mechanisms

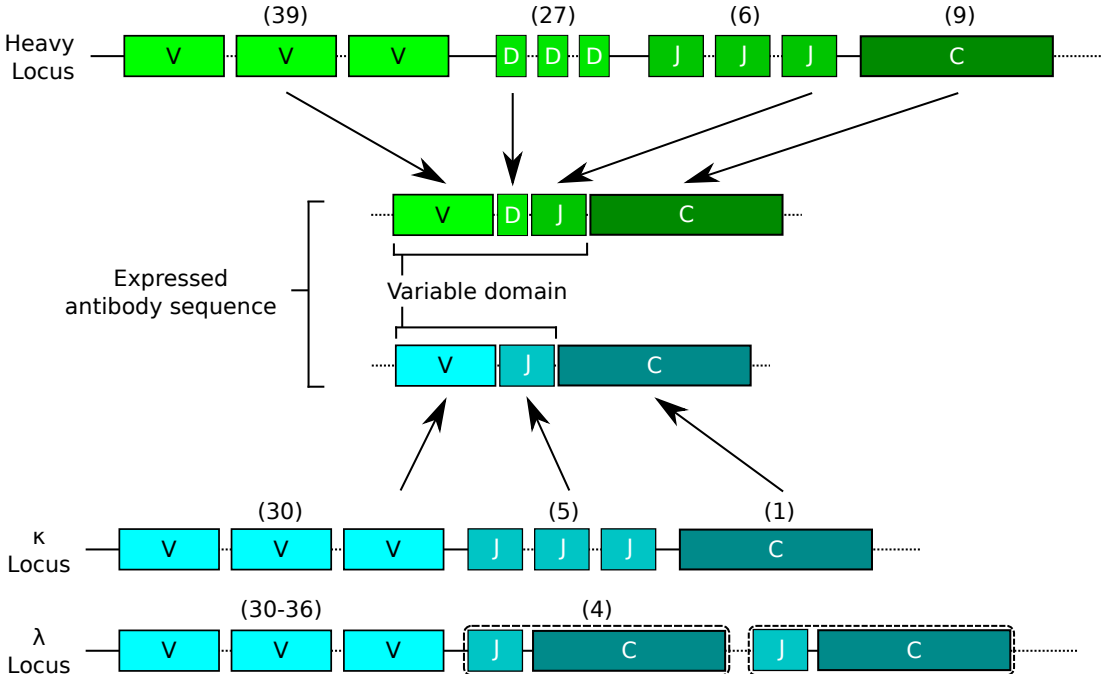


Figure 1.9: A schematic of v(d)j recombination. Antibodies consist of two different chains, heavy and light, generated from multiple genes. In heavy chains these are variable (v), diverse (d), joining (j) and constant (c) genes. Light chains are made up of v, j, and c genes and are translated from either the κ or λ loci. At the heavy loci, there are many different v, j and d genes. The number of functional genes found in humans is shown above each gene-type in brackets. In a B-Cell, all but one of each v, d, and j genes are excised at the heavy locus in a process called v(d)j recombination. Similarly, on the light chain v and j genes will be recombined. κ chains have one c gene encoding for the CL domain. In contrast, the λ chain constant domain can be encoded by a number of c genes that combine in tandem with one of the λ j genes. A B-cell will contain the recombined sequence of heavy and light chains. Either the κ or the λ chain will be translated as the light chain for a given B-cell clone. The heavy constant gene that is transcribed may change and allows the B-cell to switch the isotype of the antibody it produces.

The heavy and light chains are generated at different loci on the genome. In humans, the heavy chain locus is located on chromomsome 14 whilst the light chain can be expressed from either the κ locus on chromomsome 2 or the λ locus on chromomsome 22. At each locus are sets of genes. Multiple different variable (V),

joining (J) and constant (C) genes are present at the heavy, κ and λ loci (Figure 1.9). The heavy loci also has diverse (d) genes. An antibody VH domain is encoded by a V, D and J gene (V and J for a VL domain). Each constant domain on both chains is encoded by a C gene. The heavy constant gene determines the isotype of the antibody (e.g. IgG or IgD).

In a B-Cell, all but one of each of the V, D, and J gene segments are excised at the heavy locus in a process called V(D)J recombination [Nemazee, 2000]. Before each gene segment is a short non-coding recombination signal sequence (RSS). The RSS is recognised by the recombination activating gene proteins, RAG1 and RAG2 [van Gent *et al.*, 1995]. These cleave the DNA. Two genes, are then joined by non-homologous end joining proteins, whilst the DNA between them is excised from the genome. This process is not precise and nucleotides may be incorporated from part of the RSS, inserted by terminal deoxynucleotidyl transferase or deleted. The short sequence of the heavy d genes can also be inverted, deleted and be read in any of the three DNA reading frames. Diversity additional to that encoded by the germline genes is therefore generated at the junctions of the variable domain V-D-J or V-J genes. Junctional diversity corresponds to the CDR3 regions on both the VH and VL domains. Inclusion of the D gene segment on the heavy sequence accounts for the greater diversity of the H3 in both length and sequence than that of L3.

In humans, there are 39 different functional heavy V genes, 27 D genes and 6 J genes [Corbett *et al.*, 1997; Schroeder & Cavacini, 2010]. This gives about 10^4 different possible combinations for the VH domain. Junctional diversity is estimated to increase the possible VH sequences to 10^7 rearrangements. All together, the possible number of different sequences that can be generated by the immune system is of the order 10^9. These sequences are known as germline sequences and form the

primary B-cell repertoire of an organism (Figure 1.7).

Although the genes that encode and generate the germline sequences are different, gene segments can be grouped into similar families [Brodeur & Riblet, 1984; Kirkham & Schroeder, 1994; Kawasaki *et al.*, 1997; Matsuda *et al.*, 1998]. The IMGT database [Lefranc *et al.*, 2009] provides a collection of publicly available antibody sequences. The germline genes of multiple species are available and have an associated ontology [Giudicelli & Lefranc, 1999] (Figure 1.10). Each region of an antibody sequence can be annotated with this ontology. Pairings of heavy and light germline genes have been thought to be essentially random [Brezinschek *et al.*, 1998; de Wildt *et al.*, 1999] although some preferences are observed in the sequences of positively selected antibodies [Jayaram *et al.*, 2012].

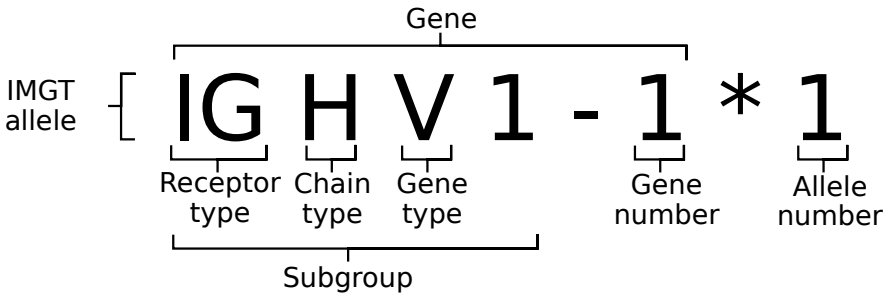
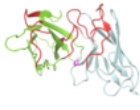


Figure 1.10: The IMGT ontology for germline genes [Giudicelli & Lefranc, 1999]. Here, the annotation of the immunoglobulin, heavy-chain, subgroup-1, gene-1, allele-1 is shown. The receptor type is IG for all antibodies (TR for T-cell receptors). The chain type (or group) is either H (heavy chains), K (κ chains) or L (λ chains). The gene type is either V (variable), J (joining), D (diverse) or C (constant). Each different gene is given a number. Multiple alleles can exist for some genes. These differ from one another by a few point mutations and correspond to sequences identified from different individuals. Each allele is also given a number. Genes are grouped together to form subgroups. Genes within a subgroup each have a sequence identity to one another of 75% or more at the DNA nucleotide level. This example has the IGHV1 subgroup.

Mature antibodies that are produced in response to an antigen rarely have a sequence that could have been achieved by v(d)j recombination alone. Instead, an



additional mechanism of diversification, somatic hyper-mutation (SHM), takes place during affinity maturation of an antibody. This process has been estimated to increase the possible number of different antibody sequences that can be achieved to 10^{16} [Schroeder & Cavacini, 2010].

1.2.3.2 Somatic affinity maturation

Once a B-cell is activated it begins to proliferate. In normal cells, the rate of mutation is of the order of 10^{-8} per base pair per cell cycle [Lynch, 2010]. However, in the variable region encoding nucleotides of a B-cell genome, the rate of somatic mutation is up to 10^{-3} changes per base pair per cell cycle [Schroeder & Cavacini, 2010]. Somatic hyper-mutation (SHM) in B-cells allows further diversification of the possible antibody variable region sequences.

SHMs are thought to be enabled by activation-induced deaminase (AID) [Petersen-Mahrt *et al.*, 2002], a protein specifically expressed in B-cells. AID can change cytosine to uracil, creating a G-U base pair mismatch. This is repaired by DNA polymerases, which are error-prone and may lead to a mutation. Locations of SHMs have been found to be correlated with certain DNA motifs (e.g. AGCT) [Sharpe *et al.*, 1991; Rogozin & Kolchanov, 1992; Betz *et al.*, 1993]. Such motifs are called AID hotspots and have been proposed to play a key role in directing the location of SHMs. AID is also found to be important in a B-Cell's ability to change the isotype of the antibody it produces (class switching). Mice deficient in AID are unable to generate IgG, IgA or IgE antibodies but are able to produce the IgM isotype that are produced early in B-cell development [Muramatsu *et al.*, 2000].

Once a germline B-cell has been activated and proliferates, different lineages of its daughter cells will acquire different SHMs. Those clones that produce antibodies with

an improved antigen-affinity are further selected to proliferate. Those that decrease or lose antigen binding, proliferate at a slower rate. Thus, multiple SHMs are acquired that optimise the antibody-antigen interaction. Positively selected SHMs have a high propensity to be at or near to positions that can make direct contact with the antigen (i.e the CDRs) [Burkovitz *et al.*, 2013].

Once the B-cells enter the blood stream they no longer undergo SHM and are said to be affinity matured. They produce antibodies that are both specific for and have a high affinity with, a particular antigen.

1.3 Antibody structure

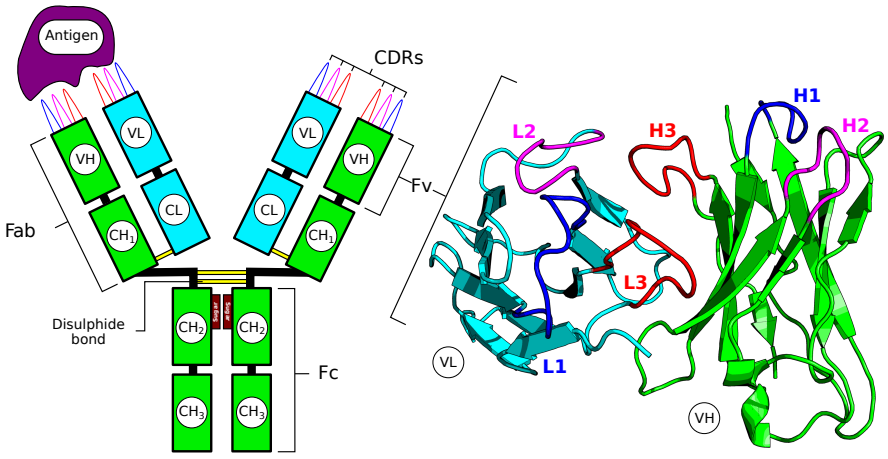


Figure 1.11: A schematic of an IgG antibody and the structure of the Fv region. The schematic on the left is labelled as in Figure 1.8. The structure on the right shows the VH and VL domains coloured green and cyan respectively. VH and VL share similar immunoglobulin fold structures (Section 1.3.1). Their β strands and individual residues can be annotated consistently (Sections 1.3.2.1 and 1.3.2.2 respectively). Each contains three CDR regions, of which there are a number of definitions (Section 1.3.2.3). Here, we show Chothia's characterisation [Chothia & Lesk, 1987; Al-Lazikani *et al.*, 1997]. The way in which VH and VL pack together is described in Section 1.3.3.3. The relative orientation of these two variable domains is the main subject of investigation in this thesis.

As introduced in Section 1.2.3 and shown in Figure 1.11, antibodies are formed

from heavy and light chains, each containing multiple domains. Each domain has a similar highly conserved structure known as the immunoglobulin (Ig) fold

1.3.1 The immunoglobulin domain

The Ig fold consists of two β-sheets that form a sandwich and are linked by a cysteine-cysteine disulphide bond (Figure 1.12). Each of the sheets has a greek-key like β-strand connectivity [Richardson, 1981]. The Ig fold appears in different protein structures, many of which are related to immune functions [Williams & Barclay, 1988]. The Ig folds of the antibody VH and VL domains both have nine β-strands [Chothia *et al.*, 1998].

1.3.2 Annotations of the antibody variable domains

The majority of both VH and VL domains are highly structurally conserved [Padlan, 1994]. Within each domain structural diversity is primarily found in the structures of three loop regions that coincide with the CDRs. The regions between the CDRs are known as the framework and are annotated FR1, FR2, FR3 and FR4 for the VH and the VL domain. Given the structural conservation of VH and VL domains, annotations can be made to consistently label parts of their structure and sequence (Figure 1.13). For example, the tertiary structure can be characterised by the arrangement of two β-sheets as described above. In the following sections we describe the labels given to both VH and VL's β-strands, the numbering used to describe residue positions in the two domains and the different definitions of which positions form the CDRs.

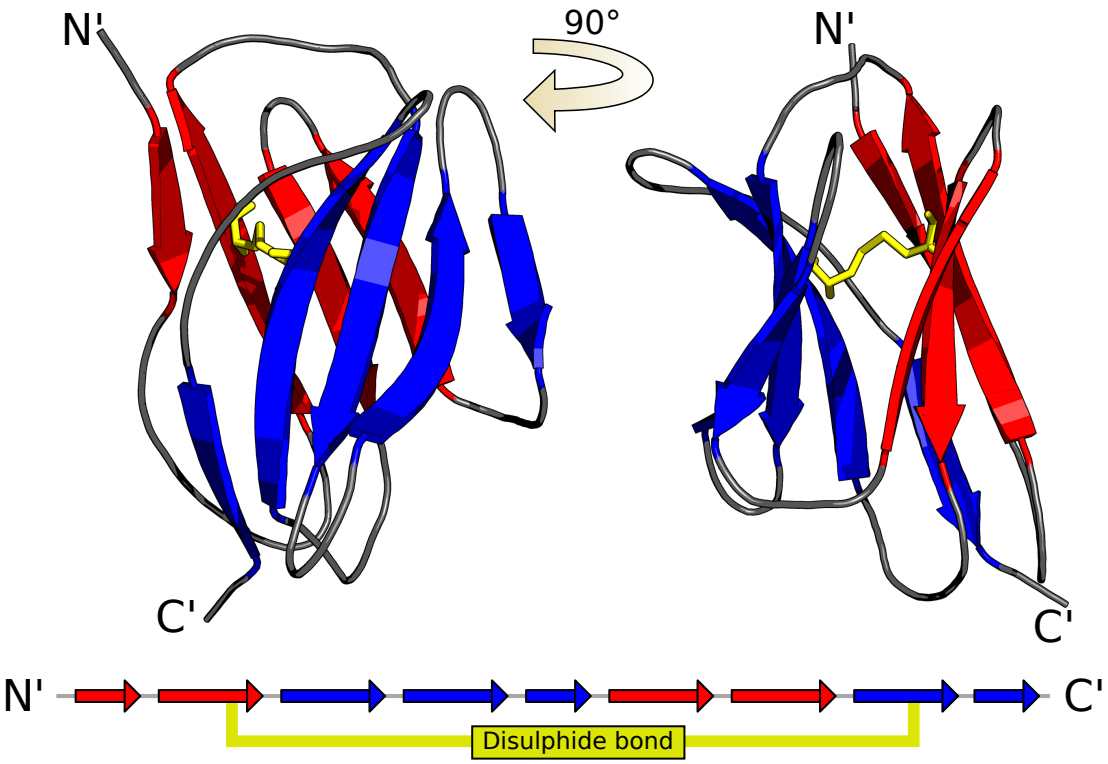


Figure 1.12: Two views of the immunoglobulin (Ig) domain. The Ig domain is formed by two β-sheets (red and blue) with a sandwich architecture. Their association is stabilised by a disulphide bond (yellow). The β-strand connectivity is shown below the views of the structure. Both the antibody VH and VL domains have this architecture. Annotations of variable domain strands are shown in Figure 1.13.

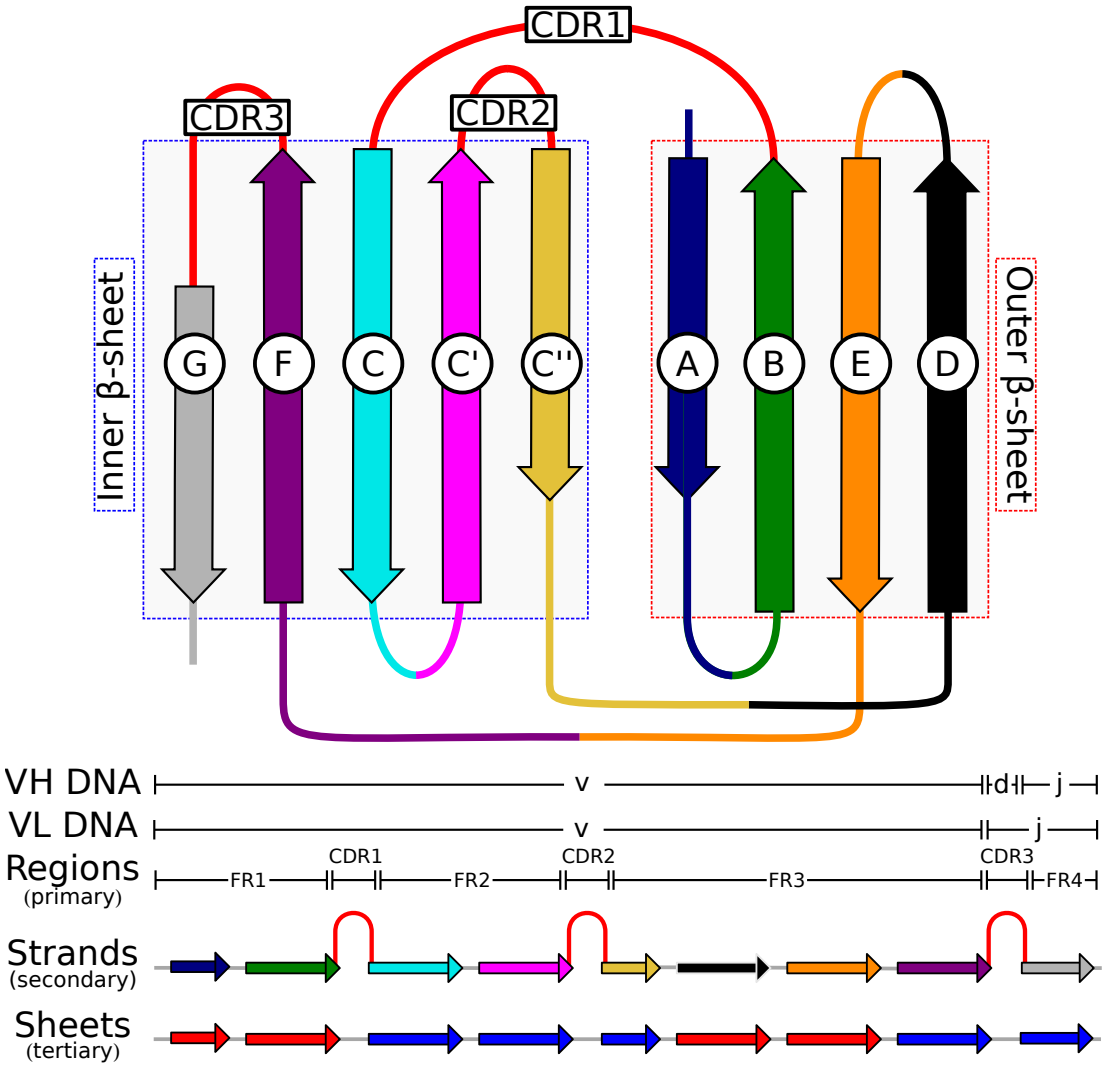


Figure 1.13: Aligned primary, secondary and tertiary annotations to the antibody variable domains. The structure of VH and VL domains are highly similar. The different genes discussed in Section 1.2.3.1 encode for certain regions of the domain. The sequence of the domain is divided into framework (FR) and CDR regions. Positions within each region can annotated with a numbering scheme (Section 1.3.2.2). Each FR region consists of several β-strands connected by loops or turns. The CDRs generally coincide with loop regions. Each strand is annotated separately with a letter (top). Other Ig domains may have fewer strands than the nine in VH and VL. In these cases, the C' and C'' may not be present. Finally, the nine strands are arranged into two β-sheets (inner and outer). The strands in each sheet are bounded by boxes with outlines colour coded as in Figure 1.12.

1.3.2.1 Strand Notation

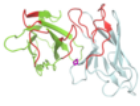
Figure 1.13 shows a schematic of how the β strands are labelled and connected in VH and VL domains. Each strand is labelled with a letter (A, B, C, C', C", D, E, F and G). C' and C" are not present in some Ig-like domains [Harpaz & Chothia, 1994].

1.3.2.2 Antibody numbering

Antibody numbering schemes are useful tools to annotate equivalent residue positions within an antibody sequence. Thus, properties including their amino-acid preferences, influence on other parts of the structure and importance for antigen binding can be analysed consistently. Multiple numbering schemes exist, all with their relative merits based on application. They vary in the nomenclature they use to label positions and the locations along the sequence at which they allow indels. These considerations have been informed by analysing antibody sequence and structural data.

The Kabat scheme [Kabat *et al.*, 1983] was developed before multiple different structures of antibody VH and VL domains were available. Instead, the authors based their scheme on alignments of VH and VL sequences separately. The sequence is numbered sequentially from 1-109 for VL and 1-113 for VH. Position three on the the heavy chain is therefore labelled H3. Those positions at which an insertion can occur are also labelled with a letter e.g. H100A. On VL, insertions are allowed to occur at 27, 95 and 106. On VH, insertions are allowed to occur at 35, 52, 82 and 100.

Later a scheme was devised by Chothia and colleagues [Chothia & Lesk, 1987; Al-Lazikani *et al.*, 1997]. This was very similar to the Kabat scheme, but placed the insertions in CDR loop regions in their "structurally correct positions". Here,



insertions on VL can occur at 30, 95 and 106. VH insertions can occur at 31, 52, 82 and 100. We refer to this set of indel positions as the Chothia numbering scheme.

More recently Abhinandan & Martin [2008] enhanced the Chothia numbering scheme by analysing where deletions and insertions can occur in the framework regions. The Martin scheme allows VL insertions at 30, 40 (with the deletion of 41), 52, 68, 95 and 107. Deletions of VL positions are also possible at 10, 30, 41 (with the insertion at 40), 52, 68 and 95. VH insertions are possible at 8, 31, 52, 72 and 100. VH deletions can occur at 8, 31, 42, 52 and 100. One major difference between this scheme and the original Chothia scheme is the change in location of the VH insertions at position 82 to position 72. Another is the ability to add insertions at H42.

One limitation of these three Kabat-like numbering schemes is that there is no direct equivalence between positions in VH and VL. In contrast, the Aho scheme [Honegger & Plückthun, 2001] uses the same numbering scheme for both VH and VL. IMGT goes further and is a unique numbering scheme for VH and VL domains, T-cell receptor variable domains and other variable-like Ig domains [Lefranc *et al.*, 2003]. Both schemes are similar in construction, giving each possible position a different number. Therefore, positions that are not present in a sequence (e.g. the insertion at L95C in Kabat) appear as a deletion in either the Aho (deletion of L116) or IMGT (deletion of L114) schemes.

Another feature pioneered by the Aho scheme was the ordering with which insertions are annotated in the CDR loop regions of the domains. In Kabat-like schemes, insertions occur unilaterally from a single position e.g. insertions at L30 would be numbered: L30, L30A, L30B, L30C. In the Aho scheme, the insertions are annotated symmetrically about a position e.g the same insertions are centered around L36: L31

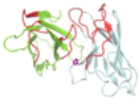
L32 L33 L39 L40. Other schemes have incorporated a similar method for numbering the CDR loop regions of antibodies [Lefranc *et al.*, 2003; Kuroda *et al.*, 2008; North *et al.*, 2011]. The IMGT scheme adds insertions at the CDR-H3 loop symmetrically about position 111. When insertions reach 111 they are added, in turn onto 111 and 112 e.g. 111, 112, 111.1, 112.1, 111.2, 112.2 etc. Kuroda *et al.* [2008] and North *et al.* [2011] defined residue-position numbering for only the CDR loops. Here, they both index residues from the anchor positions e.g. 0, 1, 2, 3, -4, -3, -2, -1. This approach can be particularly helpful when closely examining equivalent positions in the structures of loops.

1.3.2.3 CDR characterisations

In the previous sections of this Chapter we have used the term CDR loosely to describe one of the six regions of VH and VL domains responsible for antigen binding. There are different characterisations that define which specific residues form the CDRs.

The Kabat characterisation of the CDRs was developed by analysing the variation of VH and VL sequences [Wu & Kabat, 1970]. Kabat CDRs are therefore structurally-agnostic and some of the residues at positions within Kabat CDRs are not able to make direct antigen contacts. Chothia's CDRs [Chothia & Lesk, 1987], like the numbering scheme, took into account structural properties of the domains. The boundaries for the Chothia CDRs correspond to the anchors of the loops for the structures available at the time. A compromise between the Chothia and Kabat characterisations is used in the AbM definition [Martin *et al.*, 1989].

Direct consideration of antigen interactions was taken into account by the Contact characterisation [MacCallum *et al.*, 1996]. Here, CDRs were based on the positions at which residues are most frequently involved in antigen contacts. A consensus char-



acterisation for the CDRs of VH, VL and other variable-like Ig-domains was reached by IMGT characterisation [Lefranc *et al.*, 2003].

Different characterisations of CDRs are not refinements of one another and have relative merits for their use based on the application. For example, North *et al.* [2011] studied the conformations of the CDR loops. Therefore, their definition was based on positions that would provide structurally conserved anchors for the loops [Honegger & Plückthun, 2001]. Figure 1.14 shows a comparison between the Kabat, Chothia, Contact and North characterisations of the six CDRs.

1.3.2.4 CDR structure classification

The six CDRs roughly correspond to loop structures on the VH and VL domains. Despite their variability in sequence, a comparatively small set of different structural conformations are observed for five of the loops (L1, L2, L3, H1 and H2) [Chothia & Lesk, 1987; Martin & Thornton, 1996; Al-Lazikani *et al.*, 1997; Oliva *et al.*, 1998; North *et al.*, 2011; Nikoloudis *et al.*, 2014]. These conformations are referred to as the canonical classes. The sixth loop, H3, is much more variable in both sequence and structure. However, common structural properties can also be identified between the stems of different H3 loops [Morea *et al.*, 1998; Shirai *et al.*, 1999; Kuroda *et al.*, 2008; North *et al.*, 2011].

The canonical classes of the non-H3 CDRs have been studied extensively. Chothia & Lesk [1987] visually identified several consistent conformations of loops of the same type and the same length. They proposed that such canonical forms may be recognised by the presence of certain residues at structurally determining positions (SDRs) [Chothia *et al.*, 1989; Tramontano *et al.*, 1989]. However, residues at other positions can change to a number of different amino-acids without influencing the conforma-

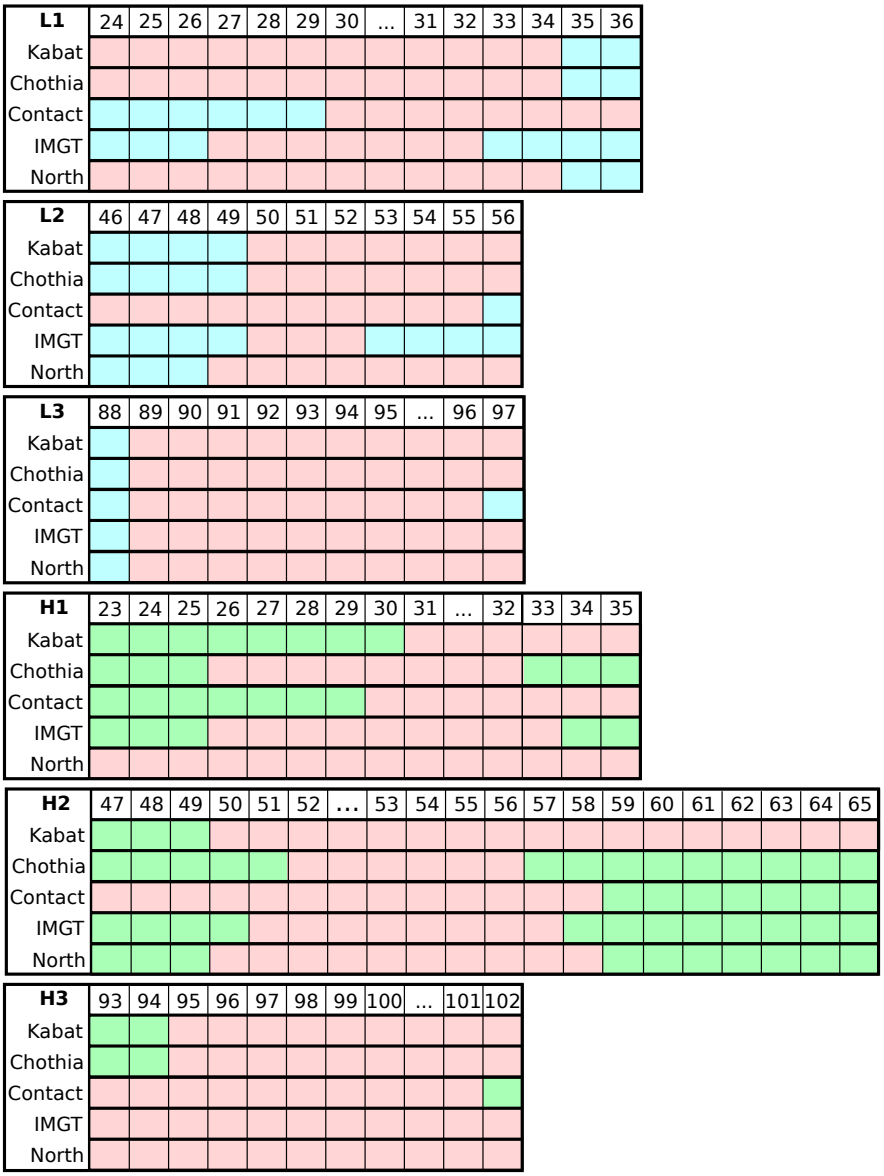
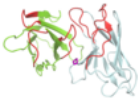


Figure 1.14: Five examples of different characterisations of the six antibody CDRs. Each box shows which Chothia positions are considered to be CDR residues for the Kabat [Jones, 1972], Chothia [Chothia & Lesk, 1987], Contact [MacCallum *et al.*, 1996], IMGT [Lefranc *et al.*, 2003] and North [North *et al.*, 2011] characterisations. Red boxes mean that the position is included in the CDR. Cyan boxes mean that the position is part of the VL framework. Green boxes mean that the position is part of the VH framework.



tion. As the number of structures available has grown, new lengths and conformations have been identified. Martin & Thornton [1996] automated the procedure of identifying distinct conformations by performing a cluster analysis of the torsion angles of the loops and then merging clusters whose backbone coordinates were similar by root mean square deviation. Al-Lazikani *et al.* [1997] analysed the high resolution structures available at the time and identified 25 standard conformations for different canonical structures.

More recently, North *et al.* [2011] revisited the structural classification of antibody CDRs. They found additional conformations, many that could be assigned either as equivalent to one of the original canonical classes or as a new canonical form that was predictable from sequence. Their 72 conformations compared with Al-Lazikani *et al.* [1997]'s 25 suggest that the original canonical classes were not sufficient to capture the potential space of non-H3 structural diversity. It is not clear whether the number of possible loop conformations have reached, or will reach, saturation. The most recent methods for clustering antibody CDRs continuously monitor the redundant set of structures available from the PDB [Dunbar *et al.*, 2014b; Nikoloudis *et al.*, 2014].

1.3.3 Quaternary structure

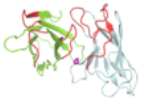
The arrangement of the antibody domains forms the protein's quaternary structure. Here, we focus on the quaternary structure of the antigen binding fragment (Fab). Each Fab is formed by four domains, VH and CH1 on the heavy chain and VL and CL on the light chain. In the folded protein, VH and VL associate non-covalently as do CH1 and CL.

1.3.3.1 The elbow angle

The orientation between the Fv region (VH and VL) and the constant domains (CH1 and CL) is referred to as the elbow angle [Lesk & Chothia, 1988; Padlan, 1994]. It is defined as the angle between the axis that runs between the pseudo-two-fold symmetry axis of the VH and VL domains and the equivalent axis for the two constant domains. This angle is related to whether the light chain is a κ or λ chain [Stanfield *et al.*, 2006]. It has also been found to be able to change upon antigen binding [Sela-Culang *et al.*, 2012]. This mode of antibody dynamics has been proposed to allow additional flexibility when binding surface antigens bivalently [Stanfield *et al.*, 2006] or to possibly act as a form of allosteric signalling through the molecule [Huber *et al.*, 1976; Sela-Culang *et al.*, 2012; Corrada *et al.*, 2013].

1.3.3.2 Binding site shapes

Association of the VH and VL domains brings the six CDRs together to form the binding site. The way that the two domains pack with respect to one another determines the relative positions of the CDRs and therefore influences the shape of the antigen binding site. The shape of the binding site's surface is determined by the relative positioning, conformation and residue content of the CDRs. Antigens can range in size, shape and chemical properties. Those antibodies specific to certain antigen types have been found to share similar binding site shapes [Webster *et al.*, 1994; MacCallum *et al.*, 1996; Lee *et al.*, 2006]. For example, binding sites that are flat or plain-like tend to bind protein antigens whilst those with deep cave-like sites are more likely to bind small molecule antigens (haptens).



1.3.3.3 Variable domain packing

Residues in the inner β-sheet (Figure 1.13) of the VH and VL domains form the majority of the domain-domain interface. Chothia *et al.* [1985] described how the edge C' strand of both domains has conserved a β bulge (Figure 1.15). This is caused by an extra residue compared to a usual anti-parallel β-sheet (Figure 1.4) that induces the strand to twist [Richardson, 1981].

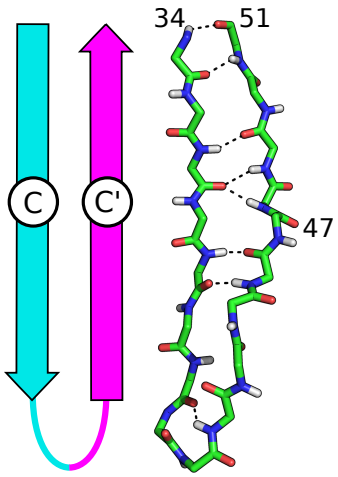


Figure 1.15: A β bulge in an anti-parallel β-sheet. The structure on the right shows the main chain of the C and C' strands of a VH domain. The numbers refer to the Chothia positions of certain residues. Intra-main chain hydrogen bonds are shown as black dashes. An extra residue at position H47 (typically a tryptophan) causes the strand to twist more than would occur in a regular anti-parallel β-sheet. This is a β bulge and is a conserved feature of the interface C' strands of VH and VL domains described by Chothia *et al.* [1985].

As a result, the inner, or interface, sheets form a barrel-like structure when viewed looking down onto the antigen binding site (Figure 1.16 - left). Side chains of residues at two diagonally opposite corners of the interface sheets fold into the centre of the VH-VL interface to form what Chothia called the "inner layer" (Figure 1.16 - right). These core residues from the edge strands in the three structures analysed in Chothia *et al.* [1985]'s study were 44, 96 and 98 on VL and 45, 100 and 103 on VH. Two

layers involving the side chains of central strand residues of the VH and VL domains form more peripheral inter-domain packing.

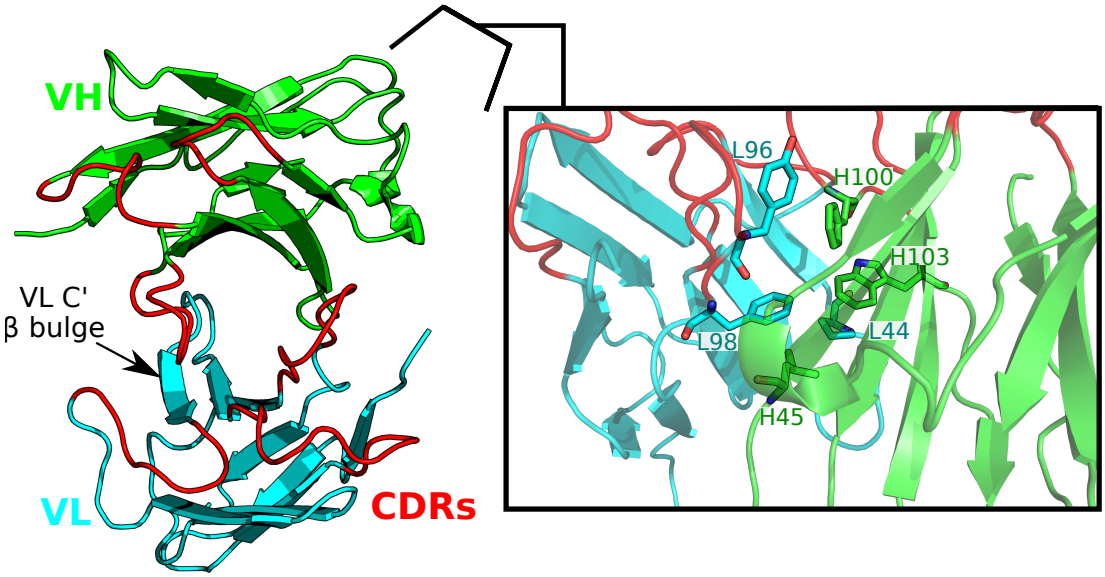
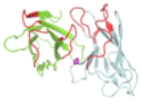


Figure 1.16: Packing of the antibody VH and VL domains. Left: A view looking down onto the antigen binding site. Residues in the interface β sheets of the VH and VL domains are involved in the majority of inter-domain contacts. They form a barrel-like structure with the edge strands making most of the contacts. Both the VH and VL domains have β-bulges in their C' strands. Residues at diagonally opposite edges of the interface sheets contribute to the "inner-layer" as described by Chothia *et al.* [1985]. Right: a view of the VH-VL interface with the "inner-layer" residues shown as stick representation. The Chothia position of the six residues are shown. The approximate view-point of the inset image is shown by the black line.

Chothia's three layer packing model was later refined by Vargas-Madrazo & Paz-García [2003]. Based on an analysis of 23 structures, these authors found more variation in the possible packing of interface residue side chains. Specifically the side chain of Chothia's inner-layer residue L96 in most cases was not involved in core packing and pointed out of the interface. Similarly, the possible insertions of residues in the CDR H3 loop meant that position H100 was also not always involved. In fact, the residue three positions before H103 (position H100 for length 7 Chothia H3s but



H100A for length 8) better described the residue in the inner layer. An analysis of contact frequencies, amino-acid usage and side-chain orientations led Vargas-Madrazo & Paz-García [2003] to propose a zone model for VH-VL domain packing (Figure 1.17). On the surface of both VH and VL domains are three zones: the central zone that contribute inter-domain contacts (VH: 37, 45, 91, 100-3, 103; VL: 36, 44, 87, 89 and 98); the proximal zone that form inter-domain contacts, contacts with the hyper-variable loops and intra-domain contacts (VH: 35, 47 and 95; VL: 34, 46, 91, 96); and the remote zone at the base of the Fv whose residues make inter-domain contacts but do not always pack into the core region (VH: 39, 91; VL 38 and 87).

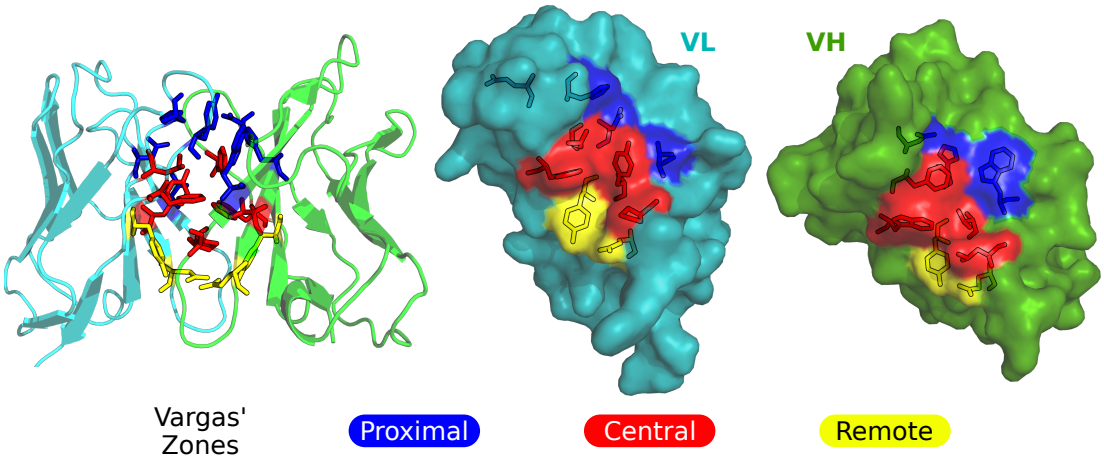


Figure 1.17: The three zones of the VH-VL interface described by Vargas-Madrazo & Paz-García [2003]. The Fv has been "opened up" to display the interacting surface of the VH and VL domains. Each zone is coloured separately. The central zone forms the core of the interface whilst the proximal and remote zones form inter-domain contacts at the top and bottom of the interface. Residues in the proximal zone can also interact with the CDR loops.

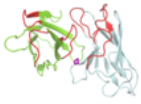
Both Chothia and Vargas' descriptions were based upon a small number of structures compared to the volume of data that is now available. Despite this, many of the general characteristics of variable domain association were captured by their descriptions. However, variation in the packing of the two domains is possible. Changing the

VH-VL packing has been proposed as a mechanism with which antibodies further increase their structural diversity [Colman *et al.*, 1987; Colman, 1988; Foote & Winter, 1992; Davies & Metzger, 1983; Chothia *et al.*, 1985; Stanfield *et al.*, 1993; Khalifa *et al.*, 2000; Vargas-Madrazo & Paz-García, 2003; Abhinandan & Martin, 2010; Chailyan *et al.*, 2011]. By doing so, an antibody can alter the relative orientation of VH and VL and thus the geometry of its antigen binding site.

1.3.3.4 Domain orientation

Although CDR region residues are thought to be the predominant determinants of antigen specificity, mutations to residues in the framework can change antibody-antigen affinity [Riechmann *et al.*, 1988; Foote & Winter, 1992; Chatellier *et al.*, 1996; Banfield *et al.*, 1997; Khalifa *et al.*, 2000; Nakanishi *et al.*, 2008; Fera *et al.*, 2014]. One way a framework residue mutation could influence affinity is to induce a change in conformation of one or more of the CDR loops [Riechmann *et al.*, 1988; Foote & Winter, 1992]. However, certain affinity-improving mutations have been identified at residue positions distant from the binding site [Chatellier *et al.*, 1996; Khalifa *et al.*, 2000; Nakanishi *et al.*, 2008] and are unlikely to make contact with either the CDRs or the antigen directly. Here, a rearrangement of the relative orientation between VH and VL has been proposed as the structural change that influences binding affinity. Such a structural change has been observed to be related to affinity changes during antibody engineering [Banfield *et al.*, 1997] and in natural antibody maturation in response to a changing viral antigen [Fera *et al.*, 2014]. Changes in VH-VL orientation have also been observed as an allosteric effect in response to antigen binding [Colman *et al.*, 1987; Colman, 1988; Stanfield *et al.*, 1993; Teplyakov *et al.*, 2011].

The relation of binding properties with VH-VL orientation has prompted the sys-



tematic study of the inter-domain pose [Narayanan *et al.*, 2009; Sivasubramanian *et al.*, 2009; Abhinandan & Martin, 2010; Chailyan *et al.*, 2011]. In Chapter 3 we will discuss the different methods that have been used to investigate VH-VL orientation. Chapters 3 and 4 will investigate the space of variable domain orientations in antibodies and TCRs, whilst in Chapter 5 we study what factors determine pose.

1.4 Therapeutic and diagnostic applications

An antibody's ability to bind specifically to a particular target makes them attractive as both therapeutics and as diagnostic tools. In the laboratory antibodies are used in numerous assays with applications such as pathology detection, molecular labelling or filtering [Borrebaeck, 2000]. Antibody therapeutics are currently the fastest growing type of pharmaceutical. There are more antibody therapeutics currently undergoing regulatory approval than any other class of drug [Reichert, 2013].

1.4.1 Antibodies as biopharmaceuticals

A major challenge in the development of a pharmaceutical is choosing a compound that will only affect a desired pathway and have minimal off-target interactions. The antibodies of the natural immune response come close to having this "magic bullet" property. Their use as therapeutics is therefore becoming a well established form of treatment for various diseases [Elbakri *et al.*, 2010]. For example, at the time of writing, an outbreak of the Ebola virus is threatening the lives of the west African population. It is an antibody therapeutic, ZMapp, that provides the most promising treatment for this disease and is currently being accelerated through clinical trials [Qiu *et al.*, 2014; McCarthy, 2014].

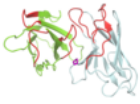
One action of a therapeutic antibody can be to direct the natural immune defence mechanisms (Figure 1.5) towards antigens that it either cannot recognise as pathogenic or is unable to react to quickly enough to prevent disease. Engineered antibodies can also be used to inhibit a particular target that is involved in a disease pathway. Alternatively, the Fc of an antibody can be linked covalently to small molecules [Alley *et al.*, 2010]. The antibody is able to deliver the otherwise toxic conjugated drug specifically to, for example, cancer cells [Zolot *et al.*, 2013; Mullard, 2013].

Antibody formats other than the natural IgG are also in use (Figure A.1 in Appendix A). Their versatile domain structure have allowed the development of innovative single domain antibodies, multi-specific molecules and other non-naturally occurring fragments such as the single chain Fv (scFv) [Holliger & Hudson, 2005]. However, all antibody therapeutics must be selected or engineered to have specific high-affinity binding properties, be non-immunogenic in human patients and have stable biophysical properties that allow them to be manufactured in large quantities.

1.4.2 Antibody Engineering

Once the target antigen has been identified a monoclonal antibody (mAb) can be generated. To date, the majority of therapeutic mAbs have been derived from mice or rats (murines) using hybridoma technology [Köhler & Milstein, 1975]. To prevent an immunogenic response in human patients the murine content of these mAbs must be reduced whilst retaining their original specificity and affinity.

Chimeric antibodies are engineered constructs where the murine constant domains are replaced with their human counterparts. In humanised antibodies only the murine



CDRs are retained [Riechmann *et al.*, 1988]. A challenge here is to choose VH and VL frameworks that will allow the antibody to retain the affinity of the original murine molecule. Often, further engineering is required to regain antigen-affinity by making mutations back to the original murine residues [Adair *et al.*, 1999; Igawa *et al.*, 2011]. As discussed in Section 1.3.3.4 one possible reason this is required is to retain the variable domain orientation of the murine antibody.

Fully human therapeutic mAbs can be generated *in vitro* using phage display [Weinblatt *et al.*, 2003]. Here, a human antibody repertoire is expressed by bacteriophages. Those phages producing antibodies able to bind to the target antigen are selected and amplified using an iterative procedure [McCafferty *et al.*, 1990]. A degree of rational engineering is possible by introducing mutations to residues in certain regions of the antibody [Groves *et al.*, 2013].

The engineering process of mAbs using phage display technology can benefit from analysis of existing antibody data to make rational decisions. For example choosing the size, diversity and composition of a repertoire can influence the effectiveness at which specific high affinity antibodies are chosen [Finlay & Almagro, 2012]. Directing the location of additional mutations can be informed by computational predictions of the likely effect of their introduction on binding and biophysical properties such as stability [Honegger, 2008; Kuroda *et al.*, 2012].

1.4.3 Immunoinformatics for therapeutic antibody development

Computational analyses and tools are increasingly being employed to aid the antibody engineering process [Kuroda *et al.*, 2012; Shirai *et al.*, 2014]. The antibody

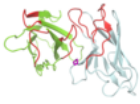
numbering discussed earlier (Section 1.3.2.2) is one informatics concept that allows the analysis and comparison of different antibodies. These schemes can be applied computationally [Abhinandan & Martin, 2008; Lefranc *et al.*, 2003] and form one type of annotation available from several antibody databases.

1.4.3.1 Databases

Several databases that handle antibody data currently exist [Johnson & Wu, 2001; Retter *et al.*, 2005; Lefranc *et al.*, 2009; Martin, 2010; Ansari *et al.*, 2010; Ponomarenko *et al.*, 2011; Chailyan *et al.*, 2012]. Of these, most are sequence-based or are antibody discovery tools. DIGIT [Chailyan *et al.*, 2012], provides sequence information for immunoglobulins and, like KabatMan [Martin, 1996] and Abysis [Martin, 2010], has the advantage over earlier sequence databases (Kabat [Johnson & Wu, 2001], IMGT [Lefranc *et al.*, 2009], Vbase2 [Retter *et al.*, 2005]) of providing heavy and light chain sequence pairings. However, it does not incorporate structural data. AntigenDB [Ansari *et al.*, 2010] and IEDB-3D [Ponomarenko *et al.*, 2011] do include structural data. However, both focus on collecting epitope data and do not include unbound antibody structures. In comparison, both IMGT [Lefranc *et al.*, 2009] and the Abysis portal [Martin, 2010] provide the ability to inspect and download individual bound and unbound antibody structures.

1.4.3.2 Binding site prediction

Computational methods exist that attempt to predict which residues form either the antibody paratope (e.g. Paratome [Kunik *et al.*, 2012b,a], ProABC [Olimpieri *et al.*, 2013] and Antibody i-patch [Krawczyk *et al.*, 2013]) or the antigen epitope (reviewed in El-Manzalawy & Honavar [2010]). Used together this information can improve the



performance of complex structure prediction by docking algorithms [Krawczyk *et al.*, 2014]. Such tools aim to give insight about the possible influence of making certain mutations in the engineering process.

Some limited success has also been achieved in predicting potential epitope-specific antibody sequences computationally [Lippow *et al.*, 2007; Pantazes & Maranas, 2010]. *In silico* prediction of the sequences of potential lead antibodies is a major challenge but is hoped in the future to complement existing laboratory antibody discovery techniques.

1.4.3.3 Modelling antibody structures

When it is unfeasible to determine an antibody's structure experimentally it is necessary to turn to structural modelling to yield atomic three dimensional information. A large volume of structural data is available for antibodies making homology or comparative modelling the general procedure of choice [Almagro *et al.*, 2014]. As the most diverse region of the molecule most effort is focussed on predicting the Fv. Generally, the framework of both VH and VL domains can be predicted to a good resolution. The principle challenges are the prediction of the CDR loops and determining the VH-VL orientation [Kuroda *et al.*, 2012].

As discussed in Section 1.3.2.4 five of the CDR loops have shapes that can often be recognised by their sequence. In most cases, empirical rules for predicting the canonical form [Martin & Thornton, 1996; Al-Lazikani *et al.*, 1997; North *et al.*, 2011] or database search methods [Choi & Deane, 2011] give acceptable predictions. For CDR-H3, the problem is more challenging and the predictions are less accurate [Almagro *et al.*, 2011, 2014]. Here, prediction methods are often template-free and must generate and score different possible conformations (e.g. with the ROSETTA

energy function Sircar *et al.* [2009]). A conformation can score well only if it fits between the loop's anchor coordinates and is compatible with the loop's environment. This environment is, in part, determined by the relative orientation of VH and VL.

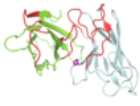
Prediction of VH-VL orientation in antibody modelling protocols have often taken a conservative approach. For example, the WAM algorithm [Whitelegg & Rees, 2000] used the same, average, orientation to assemble models of VH and VL. The more modern PIGS prediction server [Marcatili *et al.*, 2008] chooses a particular template to use for VH-VL orientation based on the sequence similarity of positions in the VH-VL interface. Other methods use energy functions to select the predicted pose from potential conformations [Narayanan *et al.*, 2009; Sivasubramanian *et al.*, 2009]. In the recent antibody modelling assessments it was unclear which type of protocol performs best with respect to the VH-VL orientation especially in the general case [Almagro *et al.*, 2014].

1.5 Overview

This introductory chapter provides the background about antibodies and how analysis of their structures can assist the engineering of such molecules as therapeutics. The rest of this thesis is organised into the following chapters:

1.5.1 Chapter 2

In Chapter 2 we describe the Structural Antibody Database (SAbDab) [Dunbar *et al.*, 2014b]. SAbDab collects, curates and presents the publicly available structural data for antibodies. We discuss the pipeline used to collect the data, the annotations made to each structure and the different ways in which the information can be retrieved



and analysed. The database statistics as of September 2014 are presented. Chapter 2 is concluded with an analysis of the structural variation in the antibody variable domains. We demonstrate that the structural diversity of the Fv cannot be accounted for by variation within the individual VH and VL domains. Instead the variable domain orientation in this type of antigen receptor is an additional form of structural diversity that should be investigated.

1.5.2 Chapter 3

In Chapter 3 we describe a method to characterise the variable domain orientation in antibodies. The computational tool ABangle [Dunbar *et al.*, 2013] that implements our method, is used to calculate the six absolute measures of orientation required to fully characterise VH-VL orientation. We compare previous studies and show how the modes of orientation being identified relate to movements of different angles. Thus, we are able to explain why different studies identify different structural clusters and different residues as important. Given this result, we then identify those positions and their residue identities which influence each of the angular measures of orientation. Further, we describe how ABangle can be applied to find sequence determinants of the orientation, compare different bound states of an antibody and assess the level of orientation conservation that is related to the type of antigen an antibody is specific for.

1.5.3 Chapter 4

In Chapter 4 we apply the ABangle methodology to a different antigen receptor, the T-cell receptor. We compare the analogous Vα and Vβ domains to the antibody

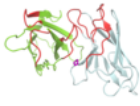
VH and VL domains [Dunbar *et al.*, 2014a]. The sets of orientations observed in the two proteins are found to be distinct from one another. As a result, the likely functional reasons for this are investigated by examining the MHC-TCR complex. The therapeutic importance of TCR-like antibodies is discussed and a possible explanation for the difficulty in their development proposed. Finally, a sequence based approach to determine residue positions that give rise to the difference in domain orientation is presented. From this, those framework positions which may be best suited as candidates to improve antibody specificity for MHCs are suggested.

1.5.4 Chapter 5

In Chapter 5 we investigate what determines the VH-VL orientation. We analyse the accuracy with which modelling protocols should aim to predict the VH-VL orientation. We assess at what level of target-template sequence identity does using the most sequence similar template become effective at predicting orientation. Therefore, we propose the maximum target-template sequence identity that should be used when benchmarking prediction protocols. The influence of sequence and structural features on VH-VL orientation are investigated. Chapter 5 is concluded by describing a feature based predictor for orientation and show its performance as a modelling protocol with different levels of template-target sequence identity.

1.5.5 Chapter 6

In Chapter 6 we conclude the thesis and summarise the findings of this DPhil research project. We present the possible future research directions and the extensions of the work presented here.



Chapter 2

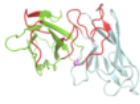
SAbDab: the Structural Antibody Database

The majority of the work presented in this chapter is contained within the following publication and is my own contribution unless otherwise stated.

J Dunbar, K Krawczyk, J Leem, T Baker, A Fuchs, G Georges, J Shi and CM Deane, 2014. SAbDab: the Structural Antibody Database. *Nucleic Acids Res.*, **42**, D1140-D1146.

2.1 Introduction

In the previous chapter we described the function and structure of antibodies. The amount of structural data available for antibodies is growing rapidly. This is due to increased research interest in the molecules themselves and to their use as experimental tools in the laboratory [Walsh, 2010]. Whether as a concerted effort or as a bi-product, the wealth of available data allows for a better understanding of the



structural determinants of these molecules and the influences of structure on the antibody-antigen interaction. In this chapter we will describe our automatically updating database to collect, curate and present available antibody structural data. The implementation of SAbDab (Structural Antibody Database) [Dunbar *et al.*, 2014b] will be detailed and the annotations that have been collected discussed. The data collected by SAbDab provides a basis for the analysis in the rest of the work in this thesis. In this chapter we investigate the coverage of antibody sequence space by structural data and the variability that is observed between structures of the variable domains.

Part of the work described in this chapter was carried out in collaboration with other members of the research group. Specifically, Konrad Krawczyk developed the majority of the front end to the SAbDab website and Jinwoo Leem curated the affinity data for structures in the database.

2.1.1 Protein Structure and the Protein Data Bank

To understand the function of a protein one can investigate properties such as its interactions [Barabási & Oltvai, 2004], expression levels in various cells [Greenbaum *et al.*, 2003] or evolutionary conservation of sequence [Altschul *et al.*, 1990]. However, arguably the most informative data to elucidate how a molecule works is its three-dimensional structure [Teichmann *et al.*, 2001]. For instance, protein structure can be used to estimate the likely binding sites for cofactors and ligands or to view the possible mechanisms that allow a process to take place. As a result, much effort goes into determining the structure of proteins (see Section 1.1.5).

The Protein Data Bank (PDB) [Berman *et al.*, 2000] contains structures of pro-

teins that have been determined in the public literature. The database strives to standardise the data contents and provides various annotations associated with each structure. At the time of writing the database updates with approximately 300 new structures each week. These are provided both through a web-based interface and a file transfer protocol service (ftp).

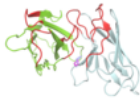
As the interest in, and use of, antibodies has increased, the number of available structures of the molecules has come to represent a significant proportion of the data in the PDB. Whilst the PDB does annotate general structures in detail with improving accuracy and consistency, antibody specific annotations are not available.

2.1.2 Antibody Structure Databases

Several databases exist that deal with antibody sequence and structure [Shirai *et al.*, 2014]. Here, we describe those that have had updates in the last two years and have structural components to them.

2.1.2.1 SACS and Abysis

SACS is a summary of antibody structures in the protein data bank [Allcorn & Martin, 2002]. It updates on a regular basis and details those structures that contain an antibody chain in the PDB. The database comprises of a summary page which contains a list of antibody chains with each entry linking to a page detailing some annotations of the structure. Whilst some annotations are provided there are no details about the heavy and light chain pairings or the antigen chains that are bound to variable domains. Structures and flat file annotations are not available for download directly from the database. SACS is incorporated into a more extensive database,



Abysis [Martin, 2010]. Abysis allows sequence and structural queries and has several integrated tools that allow for antibody numbering [Abhinandan & Martin, 2008] and assessing the humanness of a particular sequence [Abhinandan & Martin, 2007]. The system requires a licence to be used locally.

2.1.2.2 IMGT/3Dstructure-DB

The international ImMunoGeneTics information system (IMGT) is a database that collects DNA and protein sequence data for immunoglobulins, T-cell receptors and other immuno-proteins. It includes some three dimensional structure data for these molecules in the 3Dstructure-DB sub-database [Lefranc *et al.*, 2009]. Extensive annotations are available for each structure focussing on sequence features. One such feature is the V, J and C germline gene segments associated with each antibody (Section 1.2.3.1). An associated ontology to these assignments is used over the whole of the database (Giudicelli & Lefranc [1999] and described in Section 1.2.3.1). Structures are available for download from the database but only on an individual basis. No structural annotations are made and other annotations are not available as flat files. Although the database is regularly updated (about every six months) the number of antibody structures in the PDB far exceeds the number in IMGT.

2.1.3 Motivation for a new Antibody Structure Database

The volume of new antibody structures that are released each week warrants an automated approach to be taken to collect them. Curating each of the antibody structures so that they are annotated in a consistent manner informs structural analysis and allows for an extensive collection of data for use in homology modelling. We set out to

create a database of structures with the initial aim of having consistent data for use within the research group.

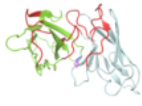
The resulting database, SAbDab, and the back-end functionality has been made available and is in use at the two sponsoring companies of this thesis, UCB and Roche. In addition to the work included in this thesis, the API is being used within the research group to aid antibody-antigen docking and CDR loop modelling, at UCB in a antibody modelling pipeline and at Roche where it is integrated with the company's internal structure database. The SAbDab API is written in a modular structure such that it can be easily extended and distributed.

In the following sections we will describe the pipeline followed to collect the data, the algorithms used and the annotations made that are presented by SAbDab.

2.2 Methods

2.2.1 The SAbDab Pipeline

The process through which structures are added to the database is summarised in Figure 2.1. Each week the newly available structures are filtered to find sequences that antibody numbering can be applied to. Those that are successful have their chains annotated with type (heavy, light or non-antibody), paired and associated with the correct antigen molecule. Annotations are then made that include experimental (e.g. determination method and affinity), sequence (e.g. numbering and gene ontology) and structural (e.g. CDR clusters and VH-VL orientation) properties. Structures are stored in the database and can be filtered and inspected in number of ways using an application programming interface (API) or a web-based interface.



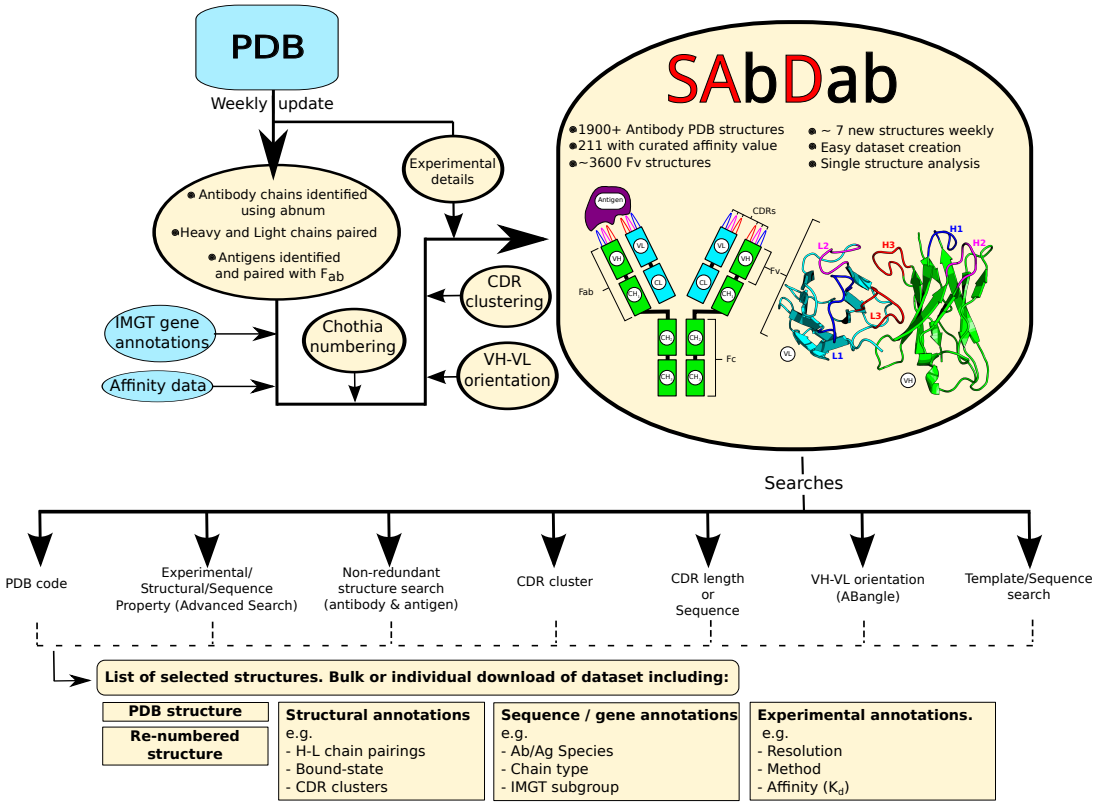


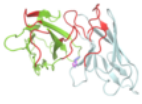
Figure 2.1: SAbDab's workflow. Each week new structures from the PDB are analysed to find antibody chains. These structures are then annotated with a number of properties and stored in SAbDab. Users may access and select this data using a number of different criteria. Structures and annotations can be downloaded individually or as a dataset. Inset, a schematic of the IGg antibody structure and the Fv fragment formed by the heavy and light variable domains, VH and VL.

2.2.2 Antibody numbering

The primary algorithm used to number sequences and identify antibody chains (Figure 2.1) in SAbDab is ABnum [Abhinandan & Martin, 2008]. If it is successful, the annotations are retained and the process applied recursively to the rest of the sequence in order to identify each variable region of the chain. This enables the identification of single-chain Fvs (scFvs) that have not been split into separate chains. If the ABnum web-service is unavailable, our own numbering algorithm ARNACI (Antigen Receptor Numbering and ClasIffication) is used to perform the same task. When a PDB entry contains an unequal number of heavy and light chains, an alignment algorithm is used to test that any non-antibody chains are so. The three algorithms are detailed below.

2.2.2.1 ABnum

ABnum is an antibody numbering program that is able to apply structural position annotations to immunoglobulin sequences [Abhinandan & Martin, 2008]. It is able to apply the Chothia [Chothia & Lesk, 1987], Martin [Abhinandan & Martin, 2008] and Kabat [Wu & Kabat, 1970] numbering schemes (Section 1.3.2.2). The algorithm uses the relative conservation of the sequences in the framework of antibodies to identify different regions. A profile can be created for any position in a multiple sequence alignment that notes the frequency of each amino acid observed. A residue in a given sequence can be assigned a score that measures how likely it is to correspond to that position using the profile. For instance, a position that is a conserved cysteine will score 1 for a cysteine residue and 0 otherwise. A position that is 50% proline and 50% leucine will score 0.5 for these two amino acids and 0 otherwise. Sets of positions can be selected and the sum of their profile scores calculated over a window of sequence



to give a profile-set score.

At the edge of each framework region, sets of 6 positions are taken and profiles created using a multiple sequence alignment of known and numbered antibody domains. To number a sequence each profile-set is scored against each window in the sequence. The profile-set score is the sum of the profile scores in the set. The maximum score for a profile-set gives the best placement for that set of 6 positions. If the ordering of the profile-sets is correct then a numbering scheme can be applied. Depending on the scheme used, the numbering can be applied backwards and forwards from these anchor points.

The profile sets used can be generated using specific sets of sequences. Most simple is to use profile sets from VH domains and VL domains. However, these can be differentiated into VH, Vκ and Vλ domains or by species.

2.2.2.2 ARNACI

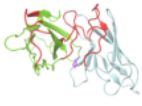
Antigen Receptor Numbering and ClasIffication (ARNACI) is my own algorithm for numbering antibodies built in order to extend the range of numbering schemes and possible domain types that can be handled. ARNACI is able to number sequences of heavy and light domains but also number the T-cell receptor Vα and Vβ domains using the IMGT numbering scheme [Lefranc *et al.*, 2003]. The annotations obtained are similar to that of ABnum and it is also able to differentiate between a larger number of domain types. This program does not require a licence for local use and is used in the SAbDab pipeline when the ABnum web-service is unavailable. However, ARNACI is currently less computationally efficient than ABnum and continues to be developed.

2.2.2.3 Numbering through sequence alignment

Although the ABnum algorithm is able accurately to number the majority of sequences, it is sensitive to unusual sequences. For instance it is unable apply numbering to the heavy sequences of the structures 4k3d and 4k3e as they have CDR-H3 loops that are 61 residues in length. An alternative method to apply numbering is to use a multiple sequence alignment of the candidate chain with known antibody sequences. To do this non-redundant sets of sequences for VH domains and VL domains were created. A multiple alignment is created using MUSCLE [Edgar, 2004] for both sets and the candidate chain sequence. The alignment where the candidate chain has the highest sequence identity to the known sequences taken forward to apply numbering. Any chain with a sequence identity to both profiles lower than 35% is considered non-antibody and discarded. The alignment gives a mapping of known framework positions onto the candidate chain. Numbering of the CDRs can then be extended from anchor points of the loops. This algorithm is less sensitive to unusual sequence features but it not as accurate or as fast as the ABnum or ARNACI algorithms.

2.2.3 Pairing Heavy and Light Chains

To pair heavy and light chains the constraint is applied that the conserved cysteine at Chothia position 92 on a heavy chain must be within 22Å of the conserved cysteine at position 88 on a light chain. Each pair of chains is referred to in this thesis by its PDB code, heavy chain identifier and light chain identifier. For example, PDB 12e8 contains two pairs of heavy and light chains, H-L and P-M. We label them 12e8_HL and 12e8_PM respectively.



2.2.4 Identifying antigen molecules

Potential antigens are identified from the non-antibody chains and the non-polymer, nucleic-acid or carbohydrate molecules. Those small molecules that are recognised as common solvents [Weichenberger *et al.*, 2013] (e.g. glycerol) are discarded. Antibody chains are then paired with their antigen molecules by calculating the number of CDR residues that are within 7.5Å of each candidate.

We classify antigens into five types: protein, peptide, carbohydrate, nucleic acid or hapten (non-polymeric ligands). Each antigen identified in the structures in SAbDab has been annotated, if available, with its sequence and the species of origin. Polypeptide antigens are deemed to be peptides if they contain less than 50 residues and proteins otherwise. This is an arbitrary threshold that ensures all antigens with "peptide" in their header names were categorised as such. However, the sequence length is recorded so that structures can be filtered based on a user selected threshold. For hapten molecules, annotations are taken from the ligand-expo database [Berman *et al.*, 2000] which is updated in-step with the PDB. The chemical structure, name and formula is extracted for these antigens.

2.2.5 Structure annotations

In addition to the chain pairing, antigen pairing and sequence numbering detailed above, each structure is annotated with a number of other features. These include annotations from a number of external sources and our own assigned properties.

2.2.5.1 PDB

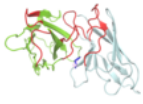
From the PDB summary files and structure headers the experimental method, resolution, R-factors and publication details such as authors and Pubmed identifiers are extracted. In addition we parse species level annotations and compound names for each of the protein chains in the structure. This allows annotation of both the antibody and antigens with information about their origins. Unfortunately this information is not available for every structure.

2.2.5.2 IMGT

For all structures in our database where an IMGT entry is available we parse the gene and species information for the antibody chains in the structure. For instance, the light chain from the structure 12e8 is species *Mus musculus* has V, J and C alleles of IGKV6-13*01, IGKJ5*01 and IGKC*01 respectively. It is therefore a κ light chain with a germline IGKV6 variable gene subgroup (see Figure 1.10). When, as in all new cases, the structure is unavailable in IMGT we assign these annotations ourselves using the known germline sequences for each species. Here, we only annotate down to the subgroup level for variable, joining and constant genes.

2.2.5.3 CDRs

There are multiple characterisations of antibody CDRs e.g. [Wu & Kabat, 1970; Chothia & Lesk, 1987; MacCallum *et al.*, 1996; Lefranc *et al.*, 2003]. In SAbDab the Kabat [Wu & Kabat, 1970], Contact [MacCallum *et al.*, 1996] and Chothia [Chothia & Lesk, 1987] CDRs are given (Section 1.3.2.3). The length and sequence of the CDRs according to these three definitions is extracted for each structure and recorded in



SAbDab. Additionally, the structure of each of these loops are extracted and populate a continuously expanding library of CDR structures for use in loop prediction methods.

The canonical conformations of a given CDR type and length were originally created with the aim of linking sequence with structure (Section 1.3.2.4). These groupings have been studied extensively [Chothia & Lesk, 1987; Al-Lazikani *et al.*, 1997; Chothia *et al.*, 1989; North *et al.*, 2011; Lara-Ochoa *et al.*, 1996; Martin & Thornton, 1996]. Given the exponential growth of the number of antibody structures in the PDB (Figure 2.3), we provide a standardised tool for studying the structural classes of CDRs. SAbDab regularly clusters the latest set of Chothia CDRs for each type (H1, H2, H3, L1, L2 and L3) and length using a method implemented by another member of the research group, Konrad Krawczyk. The clustering is performed by calculating the pairwise root mean square deviation between the CDRs and using a UPGMA clustering algorithm [Maechler *et al.*, 2013] at a number of cut-offs. Correspondence to previously defined canonical classes are noted for each cluster. This feature automatically monitors the conformational space of the CDRs as the amount of antibody structural data continues to increase.

2.2.5.4 VH-VL orientation

The antigen binding site is formed between the variable domains, VH and VL, of an antibody. The topography of the site is therefore influenced by how the domains are orientated with respect to one another. Optimising the VH-VL orientation has been proposed as a mechanism to fine tune antibody-antigen affinity. Indeed, in humanisation experiments, affinity is found to be regained after making mutations that are distant from the antigen-binding site and therefore indicative of a structural change: modifying the VH-VL orientation [Riechmann *et al.*, 1988; Foote & Winter,

1992; Banfield *et al.*, 1997]. In SAbDab we use the ABangle methodology [Dunbar *et al.*, 2013] that characterises the orientation in an absolute sense using 6 measures: 5 angles and a distance. These measures allow for the orientation space of antibodies to be characterised and will be described in the next chapter of this thesis. In SAbDab we automatically calculate these measures for each Fv region in the database.

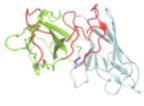
2.2.5.5 Affinity Data

Antibody binding affinity data was collected by another member of the research group, Jinwoo Leem. Data was primarily obtained from two databases: PDB-Bind [Wang *et al.*, 2005] and the structure-based benchmark [Kastritis *et al.*, 2011]. All the antibody entries were selected and only those with K_A or K_D data were kept. Where available, meta-data that is pertinent to affinity data (e.g. experimental conditions) was also collected.

Currently SAbDab contains 211 structures with an associated affinity value. One hundred and forty six are bound to proteins, 41 to peptides and 24 to hapten antigens. This curated dataset should serve as a useful benchmarking resource for the antibody-antigen docking prediction community and the antibody engineering community. It will also continue to be updated as more data becomes available.

2.2.5.6 Other structure annotations

The structures in SAbDab are each annotated with a number of other properties that are either stored or derived on-the-fly from other stored information. Such properties include whether the constant domains are present for a given receptor, if an Fv is a single chain Fv or whether there are residues that have not been resolved in the structure.



2.2.5.7 Manual flags

Although in most cases antibody identification and annotation can occur automatically, there are cases where manual inspection is required. For example, differentiating crystal contacts from real protein-protein contacts is a known problem [Carugo & Argos, 1997]. In most cases it is trivial to find which antibody chains bind to an antigen chain. However, if more than one chain is close to the antibody binding site, it is difficult to distinguish between a true multi-chain antigen and additional contacts due to crystal packing. In these cases when antigen assignment is ambiguous, the structure is highlighted for manual inspection before inclusion in the database.

Other manual flags occur when the header details of a structure contains words similar to "T-Cell" or "MHC". This to ensure that any T-cell receptors, that may have been incorrectly recognised as antibodies, are not included in the database.

2.3 Accessing the database

SAbDab was primarily created to integrate the expanding available antibody structural data easily into other pipelines and tools. It is therefore implemented in Python with an application programming interface (API) that allows a user to inspect properties of each structure and use a suite of antibody analysis tools. The API has been used to create a web-based interface developed together with Konrad Krawczyk. This website is for public use and acts as a platform for the database as well as antibody applications that have and are being developed by the Oxford Protein Informatics group.

2.3.1 Python API

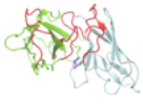
The SAbDab API is implemented as a Python package allowing the user to import the database. Once imported, an entry object corresponding to a PDB structure may be fetched from SAbDab. This object has various methods that allow information to be retrieved simply. For instance, a user may "get" the resolution or details about the paired heavy and light chains. Each set of paired chains, Fab, has information such as the numbered sequence, missing or unresolved residues in the structure or germline annotations.

Structure objects may also be retrieved that contain the atomic coordinates. Here, the Biopython PDB parser [Hamelryck & Manderick, 2003; Cock *et al.*, 2009] has been modified to put the structure into an antibody context. For example, paired antibody heavy and light chains are combined into a single object allowing them to be treated as a Fv or Fab region. Residues in these chains are automatically renumbered with the Chothia numbering scheme and antigen molecules are recognised and classified. Each element of the structure or sets of structures can be visualised directly from the API in the molecular viewer PyMol [Schrödinger, 2010].

The API also contains a number of functions that can be used separately from the database. Such utilities include interfaces for each of the antibody numbering programs described earlier, protocols to compare structures such as RMSD and functions to extract information from online databases.

2.3.2 Web interface

SAbDab is available as a web interface at http://opig.stats.ox.ac.uk/webapps/sabdab. The website was developed together with Konrad Krawczyk who imple-



mented the majority of the HTML, javascript and PHP. This author's contribution was to provide support for the back-end API and design the presentation of some of the tools and data. The front-end of the database allows users to select, inspect and download antibody structure data. Structures and their associated annotations are available both for individual and bulk download. Several selection methods are available that allow a user to create datasets or analyse particular structures.

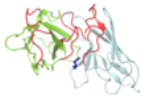
2.3.2.1 Individual structure information

An individual structure can be accessed using its PDB accession code (e.g. 1ahw). When a structure is accessed the user is brought to its summary page as shown in Figure 2.2a. Here, the structure can be visualised with the heavy chain, light chain, antigen and CDRs annotated in different colours. Clicking on the structure information tab shows details including: experimental method used to acquire the structure, species information, the number of paired heavy and light chains and, if available, the associated K_D and ΔG values for antibody-antigen binding.

Under the paired chains information tab, further details about each paired heavy and light chain (Fab) can be found. These include: H and L chain identifiers, the bound state of the Fab, the IMGT subgroup gene annotations, the Chothia numbered sequence of each chain, information about each CDR and the orientation measures between the VH and VL domains. If present, details of the antigen and its sequence are provided.

The summary page also allows the user the full set of download options. Links are also provided to the original PDB entry and if available, to the entry for the structure in IMGT.

Figure 2.2: Selected screen-shots of the SAbDab website a) The structure summary page for an entry in SAbDab. Detailed information about the structure and a visualisation of the antibody and antigen is available. b) The advanced search form. Structures may be selected using a number of methods. Here, the advanced search selects the required attributes of each structure in the selection c) The alignment between a query sequence and a template identified by the template search function. d) The ABangle orientation search tool. Users may select Fv structures by choosing specific regions of the VH-VL orientation space.



2.3.2.2 Advanced search tool

The advanced search tool (Figure 2.2b) allows the user to select structures based on a number of attributes. Attributes include: experimental method, resolution cut-off (for x-ray structures), r-factor, bound-state (bound or unbound), antigen type, antibody species and antibody light chain type (κ or λ). Users can also specify amino-acid types that must be present at Chothia positions. Similarly, structures can be limited to those that have an associated affinity value or those that have the constant domains of the Fab region present.

After clicking on the "get structures" button, the user will be presented with a list of structures that satisfy their selection criteria. Basic information is shown for each structure with a link to each entry's summary page. The "downloads" section of the results page provides options to download the selected structures.

2.3.2.3 Non-redundant dataset creation

The antibody and antigen structures in the PDB are highly redundant in terms of sequence. For instance, 5% of the bound antigens in SAbDab are lysozymes. The over representation of certain types of antigens in analysis datasets may bias results, especially in the antibody-antigen docking field where algorithms are often trained using paratope-epitope contacts. To overcome this problem we provide a non-redundant dataset creation tool. Structures are clustered using CD-hit [Li & Godzik, 2006] based on their sequence identity with respect to both antibody and antigen sequences. Users may select sequence identity levels for the antibody and antigen separately and specify other constraints for the structures returned.

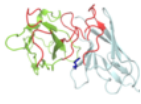
2.3.2.4 CDR search tools

SAbDab offers a CDR-specific search functionality. A user may select CDRs using similar criteria as in the advanced search tool (Section 2.3.2.2). In addition, CDR structures can be searched with respect to their CDR type and length in accordance with different CDR definitions and their membership of structural clusters or canonical classes SAbDab will return a list of the selected CDR structures. These can be inspected individually or downloaded as a dataset. The CDR search tool also allows a non-redundant set of CDR structures to be selected. In this case, only non-identical structures, with respect to type, length and sequence are returned. For identical sequences, the structure with the best resolution is returned.

2.3.2.5 Template search tool

The template search tool allows users to identify those structures in SAbDab with the highest sequence identity to a given antibody sequence. The returned entries may act as good templates for use in a modelling protocol. Structures can be searched according to their sequence identity over either the heavy or light chain or over both chains at once. Users may specify whether they wish to calculate sequence identity over the full variable region, only the framework regions, only the CDRs or only a particular CDR. An option is also provided that requires each template to have the same structurally equivalent positions as the query sequence i.e. that there are no insertions or deletions between template and query.

On submission, the top N templates (as specified by the user) are returned, ranked by their matched sequence identity to the query. Each structure may be inspected individually and the Chothia-numbered alignment between the template and the query



sequence visualised (Figure 2.2c). An option is given to download all returned structures individually or *en masse* along with a multiple sequence alignment of the template sequences to the query sequence.

2.3.2.6 ABangle search tool

As described in Section 2.2.5.4, the orientation between the variable domains can be characterised using six absolute measures. Users can explore the VH-VL orientation space using our ABangle search tool (Figure 2.2d). The distribution of each measure has been divided into discrete bins. To select structures with a particular orientation, a user may click on one or multiple (or no) bins for each of the distributions. On submission, each Fv region with a VH-VL orientation that falls within the selected orientation range will be returned. Alternatively, the same criteria as in the Section 2.3.2.2 can be used to select structures and visualise where they lie in orientation space. For instance, if a user selected structures with a proline (P) at Chothia position L44 these would show a different orientation preference to those with a tryptophan (W) at the same position [Dunbar *et al.*, 2013; Chailyan *et al.*, 2011; Abhinandan & Martin, 2010].

The "select by pdb code" function allows for the selection of a number of individual structures for comparison of their VH-VL orientation. One application of this tool is that it allows comparison of the VH-VL orientation of antibodies in their bound and unbound form. For example the HIV-1 neutralising antibody 50.1 has been crystallised both without and in complex with its peptide antigen [Stanfield *et al.*, 1993]. These structures have been cited as evidence for conformational changes in the antibody upon antigen-binding. Interestingly, it is the unbound form (1GGC and 1GGB) that has an unusual orientation whilst the bound from (1GGI) has an orientation typical

of known antibody structures.

2.4 Contents of the database

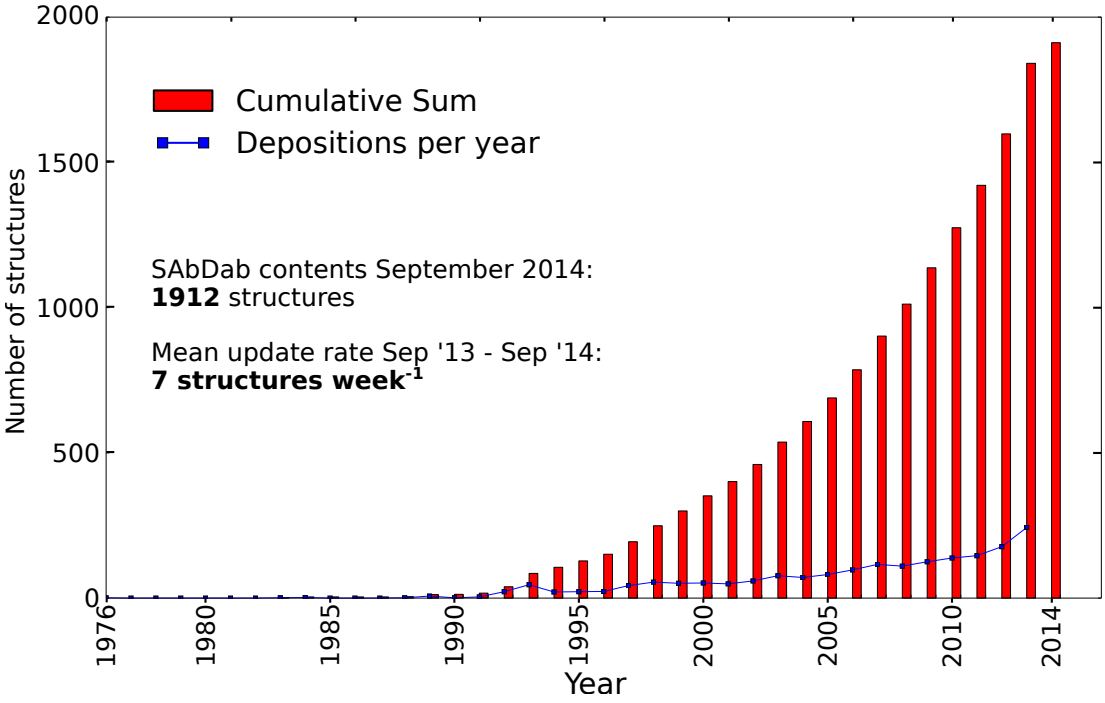


Figure 2.3: The number of antibody structures in the PDB is rising rapidly. The red bars show the total number of antibody structures. The blue line shows the number of structures deposited per year. In the past year approximately 350 structures have been added to the SAbDab database, about the same total number of structures that was available in the year 2000. Here we use the deposition date of a structure to the PDB not its public release date.

SAbDab is automatically updated each week from the PDB. The rate at which antibody structures are deposited to the database is steadily increasing (Figure 2.3). In the following sections we will investigate how these data are partitioned, how much of antibody sequence space is covered by structural data and the variation observed in structures.

As of September 10th 2014 the database contains 1912 antibody structures. A single PDB entry may include more than one paired heavy and light chain in its unit cell. In total there are 3610 Fv regions in the full redundant set.

2.4.1 Database statistics

SAbDab contains antibodies from a range of species, in different bound states and of varying formats. The data also spans a number of experimental techniques (Table 2.1).

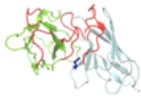
Method	Frequency
X-Ray Diffraction	1865
Electron Microscopy	26
Solution NMR	14
Solution NMR / Homology Model	5
Solid-State NMR	1
Electron Microscopy / Homology Model	1

Table 2.1: The experimental techniques used to collect the PDB structures contained in SAbDab. The frequency is the number of PDB structures in the current database solved using each technique.

Table 2.2 shows the species content of the structures in SAbDab. The majority of structures are from mice reflecting the methods by which antibodies are generated and studied in the laboratory. Next most populated are structures of human origin. The database is also well populated by single domain antibodies. Referred to as VHH's these antibodies are found naturally in Camelid species (e.g. Camels, Llamas and Alpacas) and do not have light chains. Although not naturally occurring in humans, their simpler structure gives desirable physio-chemical properties that are attractive for pharmaceutical use [van der Linden *et al.*, 1999]. Indeed, some antibody therapeutics

Species	No. Fvs	No. PDBs
Mouse	1589	979
Human	1477	657
Chimeric Human/Mouse	117	71
Llama	165	79
Camel	59	31
Synthetic construct	51	28
Unclassified/unreported	34	11
Black Rat	29	10
Norwegian Rat	15	10
Rabbit	13	8
Alpaca	10	7
Armenian Hamster	10	6
Channel Catfish	10	3
Rhesus Monkey	8	4
Chimeric Human/Rabbit	7	3
Other Camelid	4	3
Cattle	4	2
Chimpanzee	4	2
Harbor Seal	2	1
Chimeric Human/Norwegian Rat	1	1
Domestic Dog	1	1

Table 2.2: The species source of all antibodies in SAbDab ordered by frequency. The number of Fvs refers to the number of paired heavy and light chains or, in the case of single domain antibodies (Llama, Camel, Alpaca and other Camelid), the number of VH domains. The number of PDBs is the number of structures that contain an antibody of each species (a single PDB may contain multiple antibodies from different species). The species of humanised antibodies are not reported consistently in the PDB. Various authors have annotated them with chimeric human/murine, human or synthetic construct. Further work is required for SAbDab to consistently annotate different types of engineered antibodies such as chimeric and humanised formats.



have been engineered to mimic the VHH format using human-like sequences [Holt *et al.*, 2003].

Every antibody variable domain in SAbDab is annotated with the v and j genes from which it was most likely expressed from. Both human and mouse variable domain genes are well characterised and form sequence-similar sub-groups (Section 1.2.3.1). The number of genes in each mouse (*Mus musculus*) and human (*Homo sapiens*) v-subgroup according to the IMGT database is shown in Figure 2.4. The current structural coverage that SAbDab has of these subgroups is also shown. There is some relationship with the number of genes and the structural coverage of the subgroup. Whilst some of the subgroups are well represented (e.g. KV1 in both mouse and human) others have far less structural information available even though they have a large number of characterised germline genes (e.g. mouse KV13). This may be due to their relative natural usage in antibodies or a bias in the types of antibodies that have been structurally examined.

2.4.2 Data redundancy

The structures in the PDB are highly redundant in terms of the sequences they represent. We investigated the level of redundancy observed in the antibody structures in SAbDab. The sequences of all the variable regions in the database were extracted. The program cd-hit [Li & Godzik, 2006] was used to cluster the sequences at decreasing identity thresholds. At each threshold the number of clusters was noted. Six different selections were used to perform the analysis: the full Fv, the Fv framework region, VH only, VL only and their respective framework regions. Figure 2.5 shows the redundancy of each of these regions.

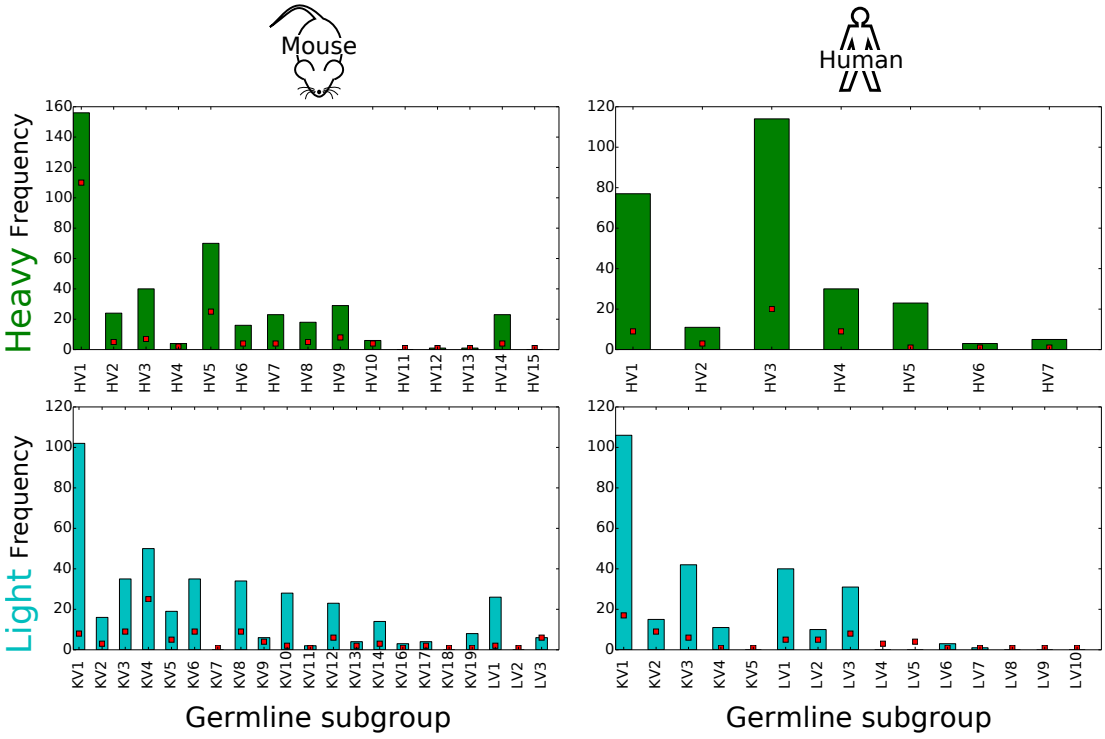


Figure 2.4: The number of non-redundant SAbDab structures for each human and mouse IMGT functional v-gene subgroup (bars) and the number of genes in each of the subgroups (red squares). Subgroups starting with K are from the κ locus whilst those starting with L are from the λ locus. In general, the number of structures available is associated with the number of functional genes in the subgroup. The structural coverage of potential sequence space is not evenly spread. We have a high volume of structural data for certain subsets of sequence-similar antibodies e.g. the mouse light KV1 subgroup. Other areas of the potential sequence space are less well structurally characterised e.g. the mouse light KV2 subgroup.

The CDRs of an antibody are more variable than the rest of the variable domain, the framework. Therefore, when removed from the analysis a higher level of redundancy in sequence is seen. VH domains are found to be more variable in sequence than VL domains over the whole variable region and in just the framework. This is in agreement with other studies that find heavy germline sequences to be more variable than light [Janeway *et al.*, 2001] and for somatic hyper-mutation sites to be more prevalent on VH domains [Burkovitz *et al.*, 2013]. The number of non-identical Fv structures is approximately 800.

The structural data in SAbDab can be used for several different applications. One area is to study how an antibody interacts with its antigen. Each antigen is annotated with its type, sequence and other information. The number of antibody-antigen complexes in SAbDab is presented in Table 2.3. In addition to redundancy in the antibody sequences there is also redundancy in the antigens of the complexes. Studies that focus on antibody-antigen docking often create non-redundant datasets based on the antibody sequence. However, this does not exclude redundant antigens and in several cases different antibodies can bind to identical antigens even at the same epitope. Table 2.3 therefore also presents the number of non-redundant complexes present in the database where both antibody *and* antigen have been considered. In the case of polypeptide antigens a 95% sequence identity threshold is used to create non-redundant sets. For hapten antigens, the HETATM name is used to filter identical molecules. For both carbohydrate and nucleic-acid antigens, the name of the antigen is used. A large proportion of complexes are removed from the antibody non-redundant set when the antigen is also considered.

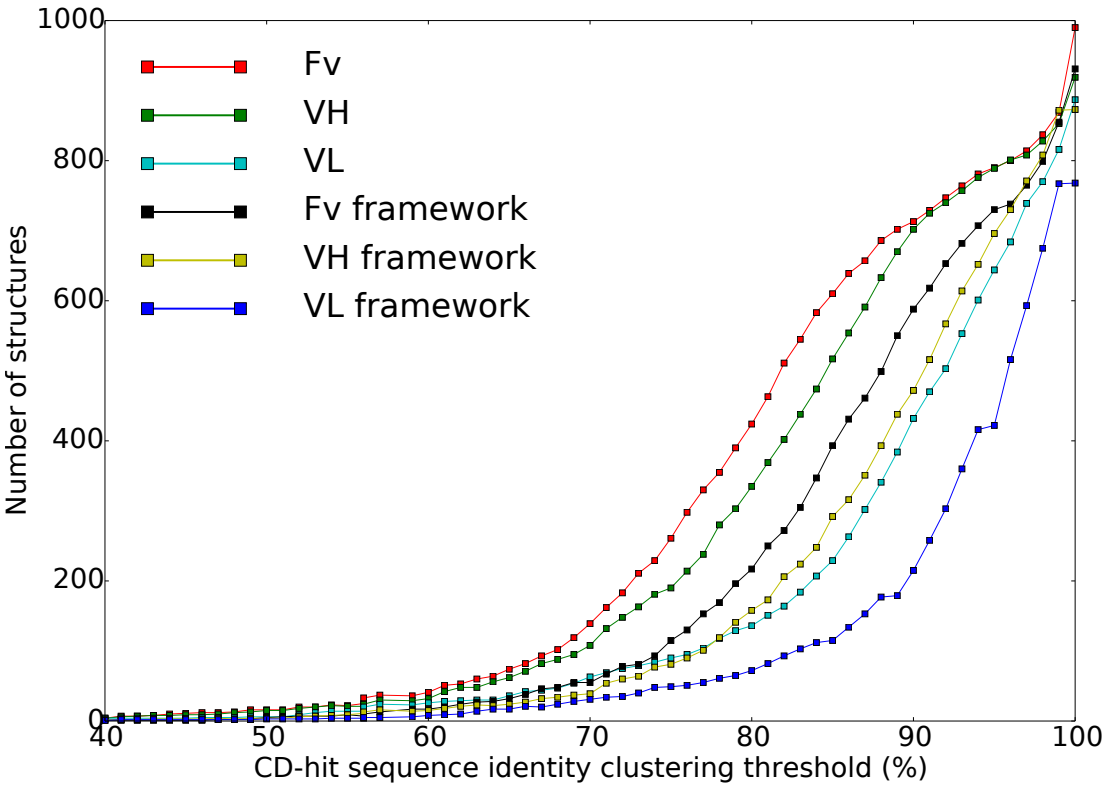
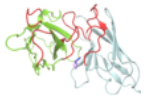


Figure 2.5: A total of 3610 structures of Fv regions are currently available in SAbDab. Of these 2746 are formed from a VH and a VL domain (i.e. not a single domain or a homodimer) which are both represented by two separate chains in the PDB structure file. This figure shows that there are just under one thousand non-identical Fv regions in the current database (100% CD-hit clustering threshold). The redundancy of the latter set is demonstrated by clustering them using different thresholds of sequence identity. Each line corresponds to the number of clusters found by CD-hit if a sequence identity threshold is applied over the regions specified in the legend. Each threshold corresponds to the minimum sequence identity between any two sequences in each cluster. Therefore at 100% all structures in a cluster must have identical sequences.



Antigen Type	Redundant complexes (PDBs)	Non-redundant ab-only	Non-redundant ab and ag
protein	1202 (676)	396	259
peptide	312 (227)	135	76
hapten	216 (147)	101	88
carbohydrate	94 (68)	27	13
nucleic-acid	16 (13)	7	7

Table 2.3: SAbDab's antibody-antigen complex data contents. The redundant complexes column refers to all Fv-Antigen pairings excluding single domain antibodies. In brackets is the number of unique PDB entries that these complexes arise from. The non-redundant ab-only column refers to the number of Fv-Antigen complexes when a sequence identity threshold of 99% is applied to the antibody variable domains. The non-redundant ab and ag column shows the number when an additional filter for the antigen is applied. This is a sequence identity of 95% for protein and peptide antigens. We call non-polypeptide antigens non-redundant if their antigen names are different.

2.4.3 Structural variation in antibody Fv structures

One application of SAbDab is to curate the structural space of antibodies. As discussed in Section 1.3 the most structurally diverse region between different antibodies is the Fv. Modelling protocols tend to focus on predicting this part of the structure as it is most influential in determining the antibody-antigen interaction (Section 1.4.3.3).

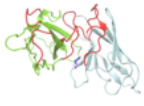
The structural variation of the Fv's VH and VL domains is demonstrated in Figure 2.6. Here, two representative domains from each variable subgroup have been selected from SAbDab for both mice and human antibodies. There are 10 mouse heavy subgroups, 14 mouse light subgroups, 6 human heavy subgroups and 7 human light subgroups with at least two structures available. The structures with the best resolution have been used in each case. The domains within each of the four selections (mouse-VH, mouse-VL, human-VH and human-VL) have been structurally aligned using mammoth-mult [Lupyan *et al.*, 2005]. Within each alignment we calcu-

late the root mean square deviation (RMSD) of the Cα coordinates of each Chothia residue position. The aligned structures in Figure 2.6 are coloured according to this position specific RMSD.

The most structurally diverse regions coincide with the CDR loops, in particular CDR-H3 (warmest colours in the human and mouse VH panels in Figure 2.6). However, some framework regions of both domains are also shown to vary between subgroups.

One such region is the C-terminal of the VL domain that, in a Fab, connects it to the CL domain. As this variation appears dependent on whether the chain is κ or λ, it may be related to Stanfield *et al.* [2006]'s observation that the elbow angle corresponds to the light chain type (Section 1.3.3.1). Chothia residue positions 8 and 9 on the VH domain are also found to be structurally variable, with multiple different conformations possible in both mouse and human structures. Although not present in any of our selected structures, a residue insertion is also possible in this region [Honegger & Plückthun, 2001; Abhinandan & Martin, 2008]. Another region found to be more structurally diverse than the rest of the framework is the loop connecting the C and C' β-strands in both domains. We will return to investigating the different possible conformations of these loops in Chapter 5.

A single measure of how different two structures are can be obtained by structurally aligning them and calculating the RMSD over chosen residues. In Figure 2.7, we show the distributions of pairwise RMSDs using different sets of residues of the antibody Fv. The variation in the framework regions is less than when the CDRs residues are included for either the whole Fv or in each of the individual variable domains. Sequence identical structures are also more structurally similar over each region than structures with different sequences. The majority of sequence identical VH and VL



domains have an RMSD of less than 1Å. However, the distribution over the whole Fv is higher. Therefore, another form of structural variation that cannot be accounted for by variation within each domain is possible.

To assess the contribution to the full Fv RMSD that cannot be explained by intra-domain variations (e.g. different conformations of CDR loops) we calculated the unexplained RMSD fraction:

$$U_{RMSD} = 1 - \frac{R_{VH} \times N_{VH} + R_{VL} \times N_{VL}}{R_{Fv} \times (N_{VH} + N_{VL})} \tag{2.1}$$

where for each pair of structures, R_{VH} is the RMSD of equivalent VH residues after superimposing the VH domain structures. R_{VL} is the RMSD of equivalent VL residues after superimposing the VL domain structures. R_{Fv} is the RMSD of all Fv residues after superimposing both VH and VL domains together. N_{VH} and N_{VL} are the number of equivalent positions in the VH and VL domains respectively. Figure 2.8 shows the distribution of U_{RMSD} for the pairs of structures in SAbDab with non-identical sequences. A median value of 17.4% of the full RMSD cannot be explained by secondary or tertiary structure variation. Instead, a quaternary difference in the Fv must account for the discrepancy. We refer to this structural variation as the VH-VL orientation.

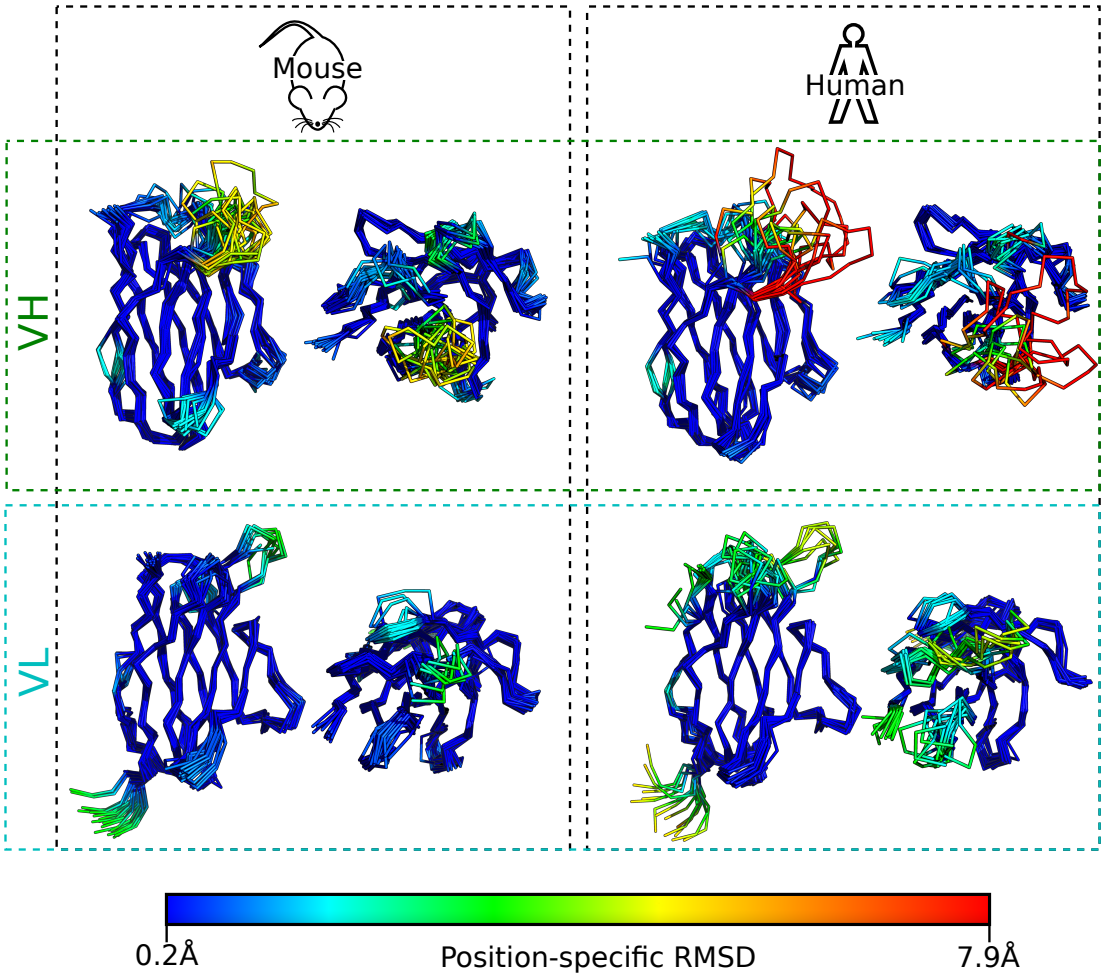
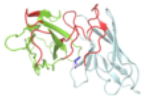


Figure 2.6: Regions of structural variability in the VH and VL domains of mice and human antibodies. Each panel shows an alignment of the best resolution structures from each of the subgroups (two structures per subgroup). Twenty structures form the mouse-VH alignment, 28 the mouse mouse-VL alignment, 12 the human-VH alignment and 14 the human-VL alignment. The two views are rotated at 90° to one another. The root mean square deviation of the Cα coordinate at each position within each of the four alignments is coloured according to the colourbar. Warmer colours indicate higher structural variation. Some of the most variable regions correspond well to the CDRs of each domain. For example, the red region in the human VH alignment is CDR-H3. However, some regions of the framework are also structurally variable.



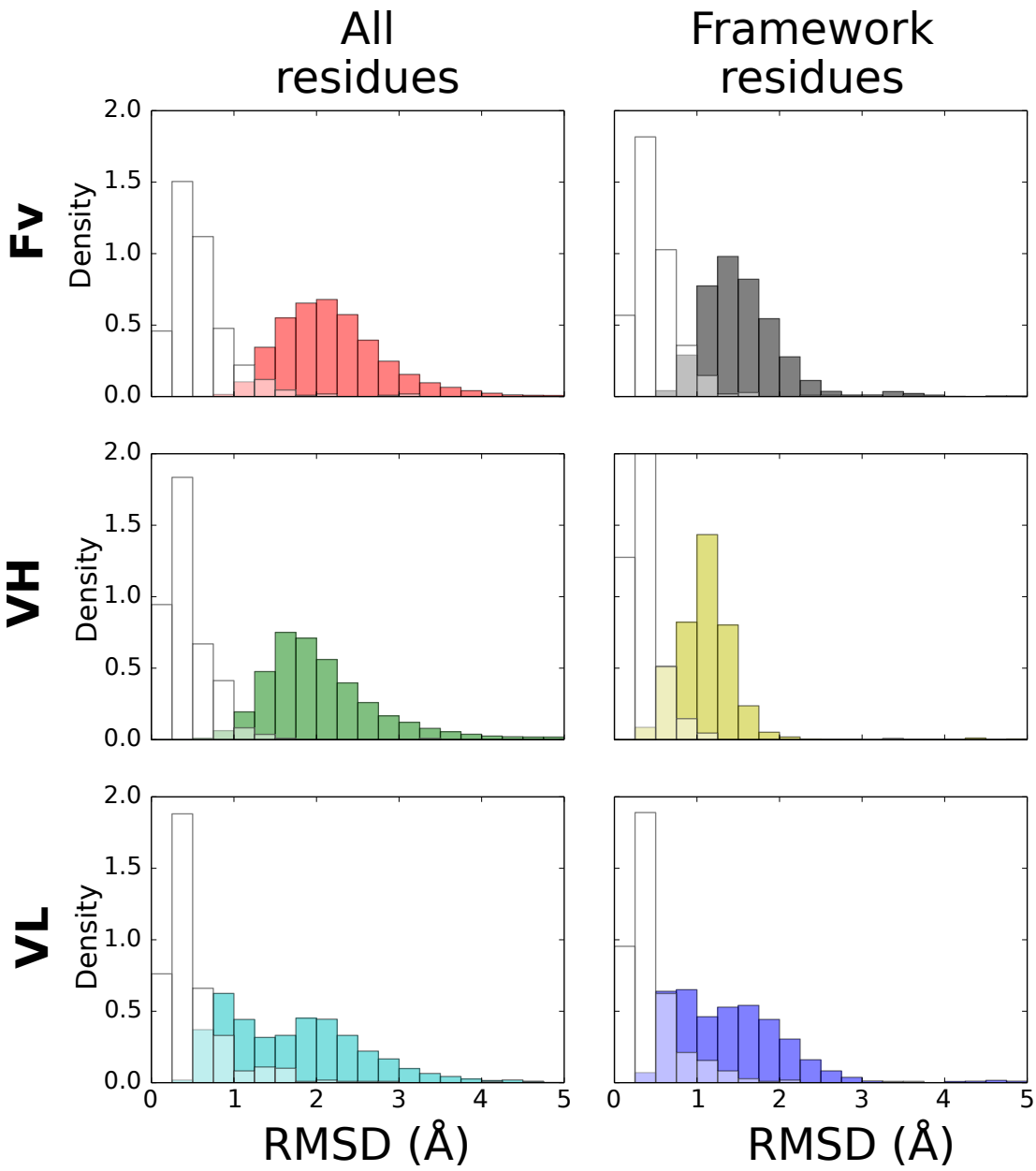


Figure 2.7: Distributions of pairwise RMSDs in a non-redundant set between all residues and only framework residues over the Fv, VH domain and VL domain (coloured histograms). The white histograms are the same measure between antibodies that are sequence identical. Comparing only residues in the framework regions gives a lower RMSD than when the CDRs are included, especially for the VH domain. Those antibodies that are sequence identical are more structurally similar than those that have different sequences.

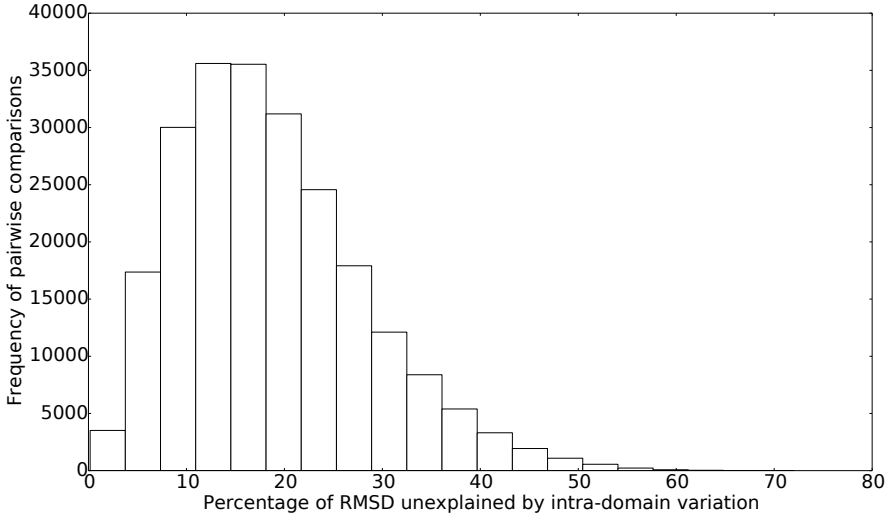
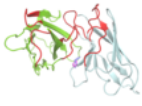


Figure 2.8: The distribution of the percentage of RMSD in pairwise comparisons of Fv structures that is not explained by variation within the VH or the VL domain. The fraction of the Fv-RMSD not accounted for by the VH-RMSD and VL-RMSD (U_{RMSD}) is calculated as in Equation 2.1 and reported here as a percentage. The median value for all pairwise comparisons in this non-redundant set is 17.4%. This discrepancy can be accounted for by differences in the VH-VL orientation.

2.5 Conclusion

In this chapter we have described an automatically updating database, SAbDab. SAbDab contains all the antibody structures in the PDB and annotations of a range of properties. The structures are highly redundant in their sequences and certain regions of antibody sequence space are better structurally covered than others. The accuracy and confidence with which one can model the structure of a given sequence is dependent on the extent of this coverage.

Variation in the structures of antibodies was investigated. As might be expected, sequence identical structures vary much less than the non-identical structures. Although some differences are measured, the mean RMSD is less than 1Å between individual VH or VL domains with identical sequences. This sets a gold standard for



the accuracy by which antibodies should be hoped to be modelled.

The RMSDs in the full framework cannot be fully explained by internal variation in each of the two domains. Another possible source of structural diversity is the way that one variable domain sits in space relative to the other.

In the next chapter we will assess how the VH-VL orientation affects the structure of the antibody. We will characterise how the orientation varies in an absolute sense and use the resulting method to help identify its sequence determinants, analyse its conservation and find its relation of antigen size.

Chapter 3

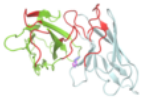
Characterising the VH-VL orientation in antibodies

The majority of the work presented in this chapter is contained within the following publication and is my own contribution unless otherwise stated.

J Dunbar, A Fuchs, J Shi, and CM Deane, 2013. ABangle: characterising the VH-VL orientation in antibodies. *Protein Eng. Des. Sel.*, **24**(10), 611-620.

3.1 Introduction

In the previous chapter we described a database (SAbDab) that collects and curates the structural information currently available for antibodies. As described in Chapter 1 and seen in Chapter 2 antibodies have a very conserved structure with most of their diversity in the antigen binding site. The antigen binding site is primarily comprised of residues in the six CDR loops that are located on the VH and VL domains. However, the geometry of the binding site is further modulated by how the VH and VL domains



orientate with respect to one another [Colman *et al.*, 1987; Colman, 1988; Foote & Winter, 1992].

This variation in domain orientation has been proposed as an additional mechanism to increase the repertoire of antibody specificity [Davies & Metzger, 1983; Chothia *et al.*, 1985; Stanfield *et al.*, 1993; Khalifa *et al.*, 2000; Vargas-Madrazo & Paz-García, 2003]. Mutations to residues at framework positions (i.e. those not in the CDRs of the Fv) in the VH-VL interface have been shown to change antigen affinity [Riechmann *et al.*, 1988; Foote & Winter, 1992; Banfield *et al.*, 1997; Fera *et al.*, 2014]. These positions are distant from the binding site and unable to make direct contact with the antigen. Therefore, their effect on antigen affinity are likely to be due to a structural change of the binding site geometry for example, modifying the VH-VL orientation.

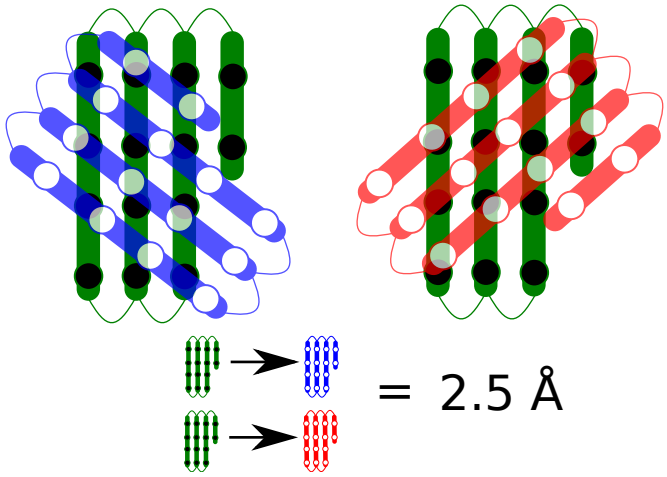


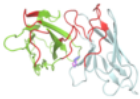
Figure 3.1: Pairs of structures can be compared using a relative measure such as the root mean square deviation (RMSD) of equivalent atoms (circles). The RMSD calculated between the green structure and the blue structure has the same relative value (e.g. 2.5Å) as the RMSD between the green structure and the red structure. However, the structures are dissimilar in *different* ways. An absolute measure puts the green, red and blue structures on a scale of structural space, giving a description for *how* the structures are different.

When making a comparison between any two protein structures, it is common

to use distance-based metrics such as the root mean square deviation (RMSD) of equivalent atoms. This measure has been used by several studies to quantify the relative changes in the VH-VL orientation for specific pairs of structures e.g. [Li *et al.*, 2000; Narayanan *et al.*, 2009; Sela-Culang *et al.*, 2012]. In other cases, the angle required to rotate between the two conformations has been reported [Colman *et al.*, 1987; Stanfield *et al.*, 1993; Banfield *et al.*, 1997; Teplyakov *et al.*, 2011]. Whilst these relative measures do quantify changes in domain orientation, they do not consistently report on how the pose changes between antibody structures. For example, whilst two pairs of structures may both differ by 2.5Å RMSD, one is unable to tell whether the pairs differ in the same way (Figure 3.1). Additionally, there is no direct way to tell how the pairs relate to one another. Similarly, a rotation angle may be reported for each pair [Stanfield *et al.*, 1993; Banfield *et al.*, 1997; Teplyakov *et al.*, 2011]. However, the angle can be reported along an arbitrary axis such that its direction is unclear.

Although the CDRs of the antibody Fv are highly variable, the structure of the framework regions of the VH and the VL domains are relatively conserved [Chothia *et al.*, 1989]. In both VH and VL the framework has a β-sandwich architecture (Section 1.3.1). This conservation can be utilised to define the VH-VL orientation in an absolute sense. In doing so, the structural space of an antibody Fv can be quantified.

Abhinandan & Martin [2010] defined such an absolute measure with their VH-VL packing angle. This was a torsion angle measured between a vector fitted through the Cα coordinates of conserved positions in the interface β-sheet of the VH domain and a similar vector on the VL domain (Figure 3.2). Their packing angle varied from -60.8° to -31.0° in known molecules and allowed each antibody to be placed



on an absolute scale of structural space. With this definition of VH-VL orientation Abhinandan and Martin identified positions that are influential in determining pose: L38, L40, L41, L44, L46, L87, H33, H42, H45, H60, H62, H91 and H105 under the Chothia numbering scheme (see Section 1.3.2.2) [Chothia & Lesk, 1987].

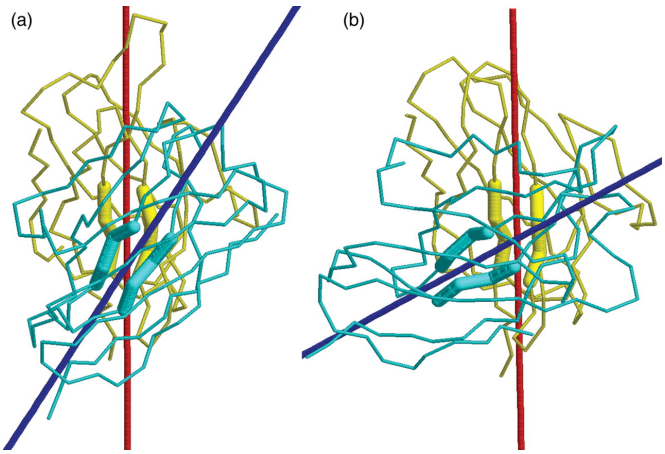


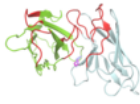
Figure 3.2: Abhinandan & Martin [2010]'s packing angle was defined as the torsion angle between vectors (red and blue lines) fitted through interface positions of the VH (cyan) and VL (yellow) domains. This was the first method to define the orientation in an absolute sense. The structures in a and b have different orientations as described by packing angles of -31.0° and -68.0° respectively. However, a single angle cannot fully capture all modes of variation. Figure reproduced from Abhinandan & Martin [2010].

A different approach was taken by Chailyan *et al.* [2011]. They focused on identifying different types of VH-VL interface within antibody structures by using a relative measure, GDT-HA [Zemla, 2003; Read & Chavali, 2007]. This measure calculates the structural similarity between two structures by assessing the fraction of corresponding positions that can be superimposed to within different distance thresholds. Analysis of the VH-VL interface structural similarity of 101 Fv regions identified two main clusters of antibody structures (A and B). A similar clustering was also identified by Sivasubramanian *et al.* [2009] using an RMSD measure. Given this clustering, a difference in the orientation of the variable domains and binding site geometry would be

expected to be observed. Indeed, Chailyan *et al* calculated that structures in cluster B had a significantly smaller binding site area than those in A. Structures in B were also found to be specific for smaller antigens.

However, no equivalent clustering is observed in Abhinandan and Martin's packing angle measure of VH-VL orientation. Furthermore, the eight positions that Chailyan *et al* found best discriminated between clusters (L8, L28, L36, L41, L42, L43, L44 and L66) are in agreement with only two of Abhinandan and Martin's positions, L41 and L44. As highlighted in a recent review [Kuroda *et al.*, 2012], this inconsistency may be due to the inability of a single torsion angle to capture all modes of orientation variation between the domains.

In this Chapter, we describe a method for fully characterising the VH-VL orientation in an absolute sense. This allows us to compare the VH-VL pose of a single Fv region, to that of all other known structures. This method has been implemented in the computational tool, ABangle. To demonstrate its use, we compare how Chailyan *et al*'s clusters differ in orientation. We resolve the apparent discrepancy in the positions assigned to be important for Abhinandan and Martin's packing angle and Chailyan *et al*'s clusters. Additionally, we use ABangle to find those positions and their residue identities that are most influential for determining VH-VL orientation. Finally a case study analysis of the conservation in orientation in sequence-identical Fv structures, suggests that those antibodies that bind to hapten antigens are more rigid than those that bind to larger protein antigens.



3.2 Methods

To characterise the orientation between any two three dimensional objects, it is necessary to define:

- a frame of reference on each object.
- axes to measure orientation parameters about.
- terminology to describe and quantify these parameters.

In the case of two human heads (Figure 3.3) the frame of reference might be the frontal plane defined relative to the conserved features of the face (e.g. mouth, ears, eyes). A line drawn consistently between the noses provides one choice of an axis about which to measure angles and distances.

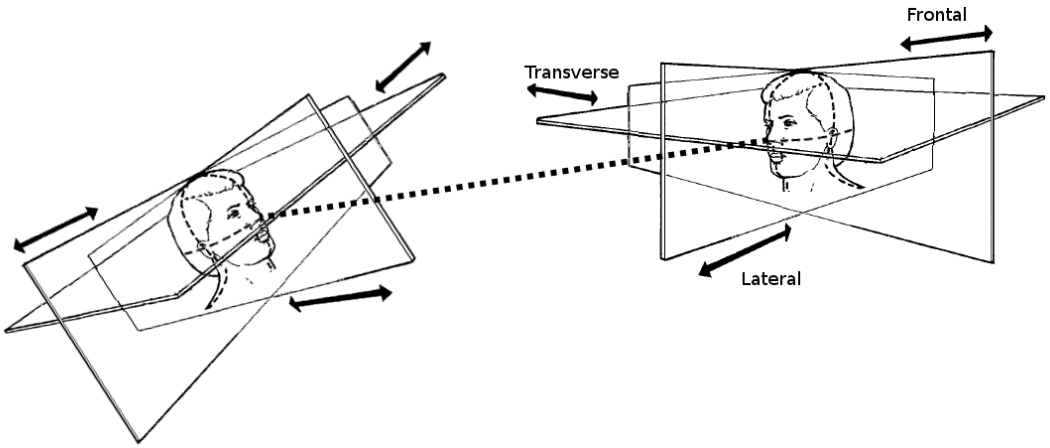


Figure 3.3: Describing the orientation between two human heads requires consistent frames of reference (e.g. the frontal plane), an axis to make measurements about (the dashed nose vector) and terminology to describe the orientation (angles and distance). Figure modified from Hall [2003]

Equivalent decisions can be made to describe the orientations between two domains of a protein. Our method for characterising the orientation between VH and

VL domains in antibodies is outlined in Figure 3.4 and described in the following sections.

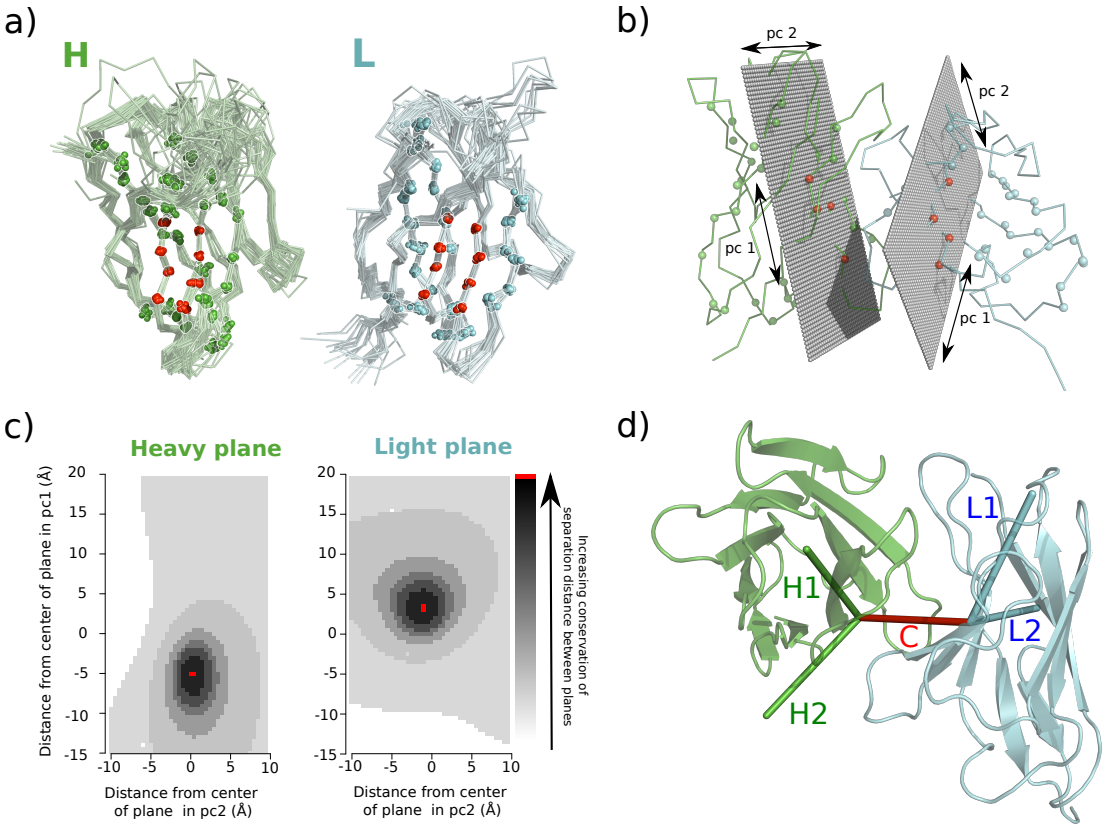
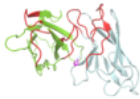


Figure 3.4: a) Superposition of 30 representative VH (green) domains showing the core-set positions (spheres) and the eight positions (red), 240 coordinates sets, used to generate the VH plane. In cyan is the corresponding image for VL. b) The average core-set positions (consensus structures) and VH and VL reference planes aligned to the antibody Fv 1B4J_HL. c) Calculation of vector C which runs through the points on the VH and VL reference planes that have the most conserved distance over the 352 Fv structures in the non-redundant set. d) Our coordinate system mapped onto 1B4J_HL. H1 and H2 are vectors that are parallel to the principal components used to create the VH reference plane in b). L1 and L2 are similarly defined for VL.



3.2.1 Dataset

One thousand and sixty one antibody structures were extracted from the protein data bank (PDB) [Berman *et al.*, 2000] in January 2012. Ninety seven were discarded as they were either single chain Fvs, single domain antibodies or Bence Jones proteins (Light chain dimers). Chothia antibody numbering [Chothia & Lesk, 1987] was applied to each of the antibody chains in the remaining 964 files using Abnum [Abhinandan & Martin, 2008]. Chains that were successfully numbered were paired to form Fv regions. This was done by applying the constraint that the H37 position Cα coordinate of the heavy chain must be within 20Å of the L87 position Cα coordinate of the light chain. One thousand two hundred and ninety six Fv regions were identified. Of these, only the 1265 X-ray crystallography structures were taken to form the full redundant dataset.

A non-redundant set of antibodies was created using CD-hit [Li & Godzik, 2006], applying a sequence identity cut-off over the framework of the Fv region of 99%. This resulted in a set of 352 structures with a resolution of 3Å or better. A high sequence identity threshold was chosen in order that we could investigate the effect that making only a few amino acid changes has on VH-VL orientation. The set comprises of 248 mouse, 87 human, 7 Rat and 9 chimeric antibody structures.

3.2.2 Identifying the core-set positions of the VH and VL domains

The most structurally conserved residue positions in the heavy and light domains were used to define domain location. We refer to these positions as the VH and VL core-sets. In order to identify these core-sets, the analysis described below was performed

separately for the VH and VL domains.

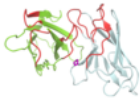
To identify the VH core-set we first selected all non-CDR positions that were present in all structures in our-non redundant set. We define a structural variation score, S_i for each of these positions:

$$\begin{aligned} \bar{d}_{ij} &= \sum_{n=1}^{N} \frac{d_{n,ij}}{N} \\ RMSD_{ij} &= \sqrt{\frac{1}{N}\sum_{n=1}^{N}(d_{n,ij} - \bar{d}_{ij})^2} \\ S_i &= \sum_{j=1;j\neq i}^{R} \frac{RMSD_{ij}}{\bar{d}_{ij}} \end{aligned} \tag{3.1}$$

Where R is the number of selected positions in the domain and N is the number of structures in the non-redundant set. The Euclidean distance between Cα coordinates of the ith and jth positions in the nth structure is denoted as $d_{n,ij}$.

Positions that are less conserved have a higher structural variation score (S_i). In order to give an estimate of how many positions should be included in the VH core-set, the position with the highest S_i was removed and the scores recomputed. The process was repeated until all positions were removed, noting the score at removal and reducing R by 1 upon each iteration. The point where the score decreased approximately linearly with R, was used to choose a cut-off for the number of positions to include in the core-set (Figure 3.5, Table 3.1).

Thirty five positions were chosen for the VH core-set. An analogous procedure was performed for the light variable domain to form the VL core-set, again containing 35 positions (Table 3.1, Figure 3.5). With this number of positions remaining, the



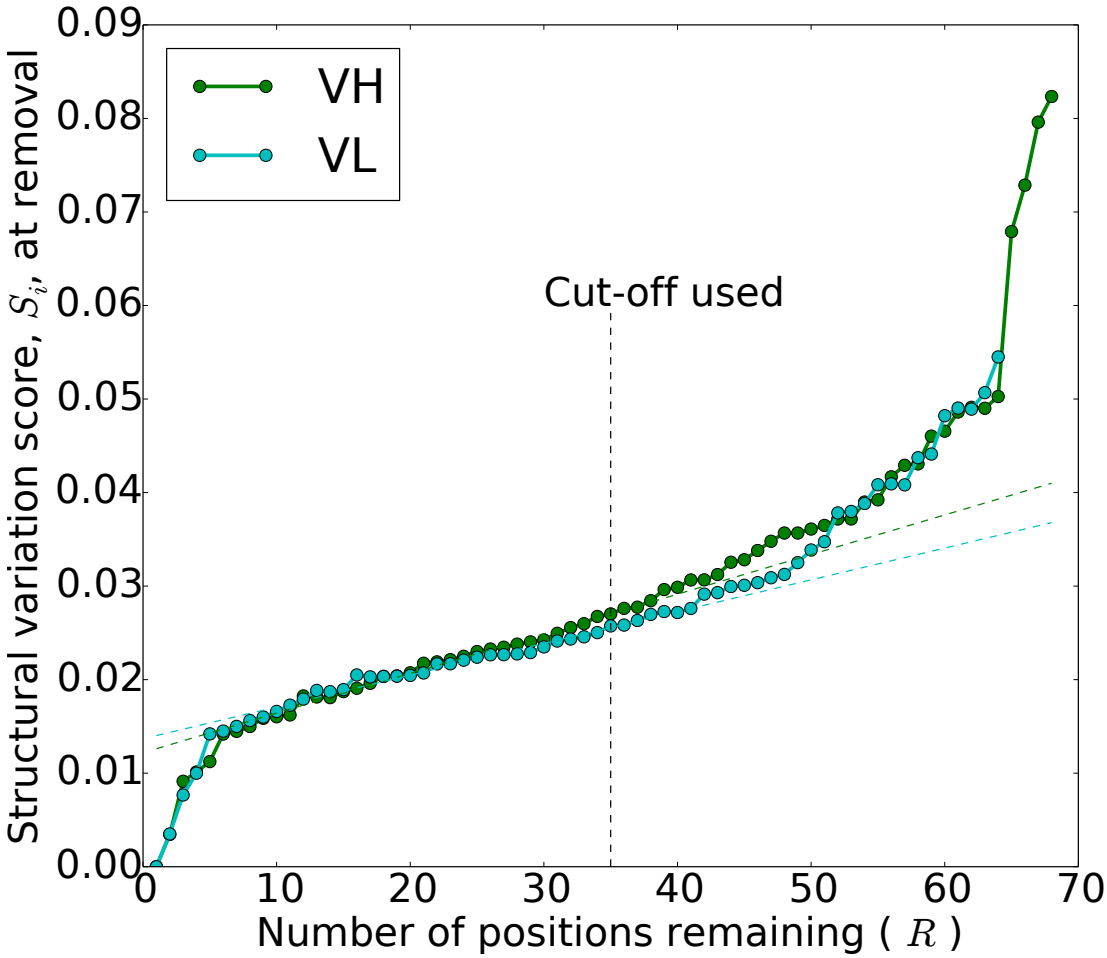


Figure 3.5: Selection of the core-set positions is detailed in Section 3.2.2. The structural variation score, S_i, of each framework position as it is removed from the set is plotted against the number of positions left in the set (R). The choice of the number of positions to include in the VH and VL core-sets is made by taking the point where the score decreases approximately linearly with R (35 positions for both VH and VL). The dashed green and cyan lines show the regions at which S_i decays with R linearly for VH and VL respectively.

value of S_i for the least conserved among the 35 positions was approximately the same for both VH and VL.

Figure 3.4a shows the location of the 35 positions used for each core-set. As might be expected, these positions are predominantly located on the β strands of the framework and form the core of each domain.

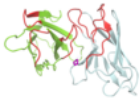
3.2.3 Defining frames of reference and consensus structures

To measure the VH-VL orientation a definition of the location of each domain must be made. These locations are described using frames of reference. One method to define frames of reference is to use conserved features on each individual domain. Abhinandan and Martin achieved this by fitting vectors through conserved positions in the β-sheets of the VH-VL interface. This description of the domain position is sensitive to the relative spacing of a small number of points. The effect of this when measuring a torsion angle is small. However, measurement of other modes of rigid-body orientation would be affected to a greater extent.

To minimise the effect of local deformations we used the core-set positions to register frames of reference onto the domains. By doing so, we capture the rigid-body locations of the whole of each domain and not just those positions at the interface.

The VH frame of reference was created as follows. The VH domains in the non-redundant dataset were clustered using CDHIT, applying a sequence identity cut-off of 80% over framework positions in the domain. One structure was randomly chosen from each of the 30 largest clusters.

This set of domains was aligned over the VH core-set positions using Mammoth-



VL domain	VH domain
L44	H35
L19	H12
L69	H38
L14	H36
L75	H83
L82	H19
L15	H94
L21	H37
L47	H11
L20	H47
L48	H39
L49	H93
L22	H46
L81	H45
L79	H68
L80	H69
L23	H71
L36	H70
L35	H17
L37	H72
L74	H92
L88	H84
L38	H91
L18	H90
L87	H20
L17	H21
L86	H85
L85	H25
L46	H24
L70	H86
L45	H89
L16	H88
L71	H87
L72	H22
L73	H23

Table 3.1: The Chothia positions that form the core-sets for the heavy and light variable domains. These are the 35 most structurally conserved position within the VH and VL domains as calculated using the method described in Section 3.2.2. The most conserved position for each domain is listed at the bottom of the table.

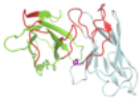
mult [Lupyan *et al.*, 2005]. From this alignment we extracted the Cα coordinates corresponding to the same eight structurally conserved positions in the β-sheet interface as identified by Abhinandan and Martin (H36, H37, H38, H39, H89, H90, H91 and H92). We fit a plane through the resulting 240 coordinates. This was done by taking the first two components of a principal components analysis. The plane is the frame of reference for the VH domain. The coordinates used to define the frame of reference were chosen to allow direct comparison with Abhinandan and Martin's orientation measure. A consensus VH structure was also created by taking the mean Cα coordinate for each of the VH core-set positions in the 30 aligned structures.

To register the reference frame plane onto an individual VH structure, we perform a superposition of the consensus structure to the core-set positions of the real VH domain, using TM-align [Zhang & Skolnick, 2005]. The resulting transformation matrix can then be used to map the plane onto the structure of the individual VH domain.

An analogous procedure was performed to create the reference frame plane and consensus structure for the VL domain. Here, positions L35, L36, L37, L38, L85, L86, L87 and L88 were used to fit the plane. Figure 3.4b shows the VH and VL reference frame planes and consensus structures when they have been mapped to the structure 1B4J_HL.

3.2.4 Choosing an axis to measure VH-VL orientation about

The procedure described above allows us to map the two reference frame planes onto any Fv structure. We can therefore think of measuring VH-VL orientation as equivalent to measuring the orientation between the two planes. To do this fully and in



an absolute sense requires a minimum of six parameters: a distance, a torsion angle and four bend angles. These parameters must be measured about a consistently defined vector that connects the planes. We used the vector that had the most conserved length in the structures in our non-redundant set. This choice was made in order to maximise the variation in orientation which may be described using angular measures. Thus, it effectively acts as the pivot axis of VH-VL orientation. We call this vector C.

To identify C, the reference frame planes were registered onto each of the structures in our non-redundant set and a mesh placed on each plane (Figure 3.4b). Each structure therefore had equivalent mesh points and thus equivalent VH-VL mesh point pairs. The Euclidean distance was measured for each pair of mesh points in each structure. The pair of points with the minimum variance in their separation distance was identified. Figure 3.4c shows where these two points are located on the planes. The vector which joins these points is defined as C. As we can identify where these points are located on each individual domain, we can also map C onto every Fv structure and define a coordinate system about it in a consistent manner.

3.2.5 Defining a coordinate system and measures for VH-VL orientation

The coordinate system is fully defined using vectors which lie in each plane and are centred on the points corresponding to C. H1 is the vector running parallel to the first principal component of the VH plane, whilst H2 runs parallel to the second principal component. They therefore lie approximately parallel and perpendicular to the strands in the VH β-sheet interface, respectively. L1 and L2 are similarly defined on the VL

domain. Figure 3.4d shows the coordinate system defined on an Fv region.

To describe the VH-VL orientation we use six measures, a distance and five angles. These are defined in the coordinate system as follows:

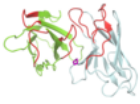
- The length of C, **dc**
- The torsion angle, **HL**, from H1 to L1 measured about C.
- The bend angle, **HC1**, between H1 and C.
- The bend angle, **HC2**, between H2 and C.
- The bend angle, **LC1**, between L1 and C.
- The bend angle, **LC2**, between L2 and C.

The HL angle is a torsion angle between the two domains and it is similar to Abhinandan and Martin's packing angle. The HC1 and LC1 bend angles are equivalent to tilting like variations of one domain with respect to the other. The HC2 and LC2 bend angles describe twisting like variations of one domain to the other.

These measures provide a method of placing the orientation of individual structures onto an absolute scale. By describing the pose of the VH and VL domains for all known Fv structures, one can compare both individual and groups of structures in a consistent manner.

3.2.6 Identification of Chailyan et al's interface types within the non-redundant set

Chailyan *et al* have previously identified two types of VH-VL interface by clustering a curated set of antibody structures. They denoted these as clusters A and B. To



investigate if and how these sets of structures differ in our orientation measures, we assigned each of the structures in our non-redundant set to one of these two clusters.

The residue identity at position L44 was found to be most informative of cluster membership by Chailyan *et al*. Structures in cluster A all have proline at L44, 302 of our 351 structures have this residue identity.

Structures in cluster B can have either phenylalanine, valine or isoleucine at L44. In Chailyan *et al*'s data set these residues occurred in the ratio 24:5:2 and therefore cluster B was largely characterised as having phenylalanine at L44. In our non-redundant dataset we identified structures in cluster B with the residues phenylalanine, valine and isoleucine in the ratio 19:16:10. This difference in proportion of residues led us to stratify the structures in cluster B by the amino-acid type at L44 (phe-L44, val-L44 and ile-L44). Four structures in our data set were not placed in either cluster. Three of these had asparagine at L44 and one had leucine at this position.

3.2.7 Random Forest Regression

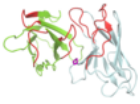
Our method allows for each structure to be placed on an absolute scale of orientation. Previous studies have identified positions that are thought to be influential for determining VH-VL orientation e.g. Abhinandan & Martin [2010] and Chailyan *et al.* [2011]. Here, we wish to find both the positions that are important and their residue identity for which they become so. For instance, position L44 can be either proline, phenylalanine, valine, leucine, asparagine or isoleucine. Although the presence of phenylalanine at this position may be influential for the HC1 and LC1 angles, it may not influence the HL torsion angle. Or conversely, the presence of valine at L44 may be informative for HL but not for the HC1 or LC1 angles.

We performed a regression on the distributions of our angles using a random forest algorithm [Breiman, 2001]. This algorithm builds an ensemble, or forest, of decision trees based on input variables in order to predict a response variable. Each decision tree is built on a subset of the data to prevent over-fitting. Those input variables that are most informative of the response variable are identified by randomly permuting each of them in turn and assessing the reduction in prediction performance. Here, we use conditional inference trees as they have been shown to allow for input variable importance assignment that is unbiased by their entropy [Strobl *et al.*, 2007].

In order to find both positions and residues that are important for the orientation angles, we created binary input variables for the regression algorithm. For instance, the residue at position L87 can be either phenylalanine, tyrosine, isoleucine or histidine. We therefore create four variables L87F, L87Y, L87I and L87H. These are 1 for a structure when the corresponding residue is present at L87 and 0 otherwise. The L87F binary variable therefore differentiates between structures that have phenylalanine at position L87 or any other residue instead.

Binary variables were created for each position that is in the VH-VL interface and therefore able to directly mediate inter-domain orientation. These positions were identified by examining the change in solvent accessible surface area (SASA) when the domains are taken individually and in complex with one another. SASA was calculated using JOY [Mizuguchi *et al.*, 1998]. Any position that had a change in SASA greater than 15% in more than 5% of the structures in the non-redundant set was defined as an interface position. We found 64 such positions, 30 from the light chain and 34 from the heavy chain.

As positions that are highly conserved in their identity are unlikely to be informative about variation in VH-VL orientation, those variables that were 1 for 80% or more



structures were discarded. Similarly, residues which appear at positions infrequently are unable to provide statistically significant information about the pose. Therefore, those variables which accounted for less than 2% of structures were also discarded. We also combined those variables which were deemed to be highly correlated. The Jaccard distance [Jaccard, 1908] was used as a measure of variable dissimilarity, with those variables of a score of less 0.4 being combined. For instance, the variables L42G and L43T were combined to form the variable L42G/L43T. This has the value of 1 when a structure has either a glycine at L42 or a threonine at position L43. Combinations only occur between variables relating to different positions.

This resulted in a total of 349 input variables. For each of the angular measures (HL, HC1, HC2, LC1 and LC1), 50 conditional inference forests were built using the R package "party" [Strobl *et al.*, 2009]. From these, the mean importance measure was extracted and the variables ranked in accordance with it.

3.2.8 Orientation RMSD

A relative measure of the difference in orientation between two antibody Fv structures is the orientation root mean square deviation (RMSD) [Narayanan *et al.*, 2009]. As an example, the orientation RMSD between two Fv structures (Tx and Ty) is calculated as follows. Tx and Ty are structurally fitted using the Cα coordinates of their shared VL domain framework positions. Tx's VH domain is then independently fitted to Ty's VH domain. The RMSD is calculated between Tx's VH domain in its native orientation and the transformed position. The same procedure is also performed using the VH domains for the initial fitting and calculating RMSD between VL domains. The mean of these two values describes the difference in the relative orientation.

3.3 Results

3.3.1 Distributions of the measures

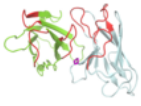
Our VH-VL orientation measures were calculated for all of the Fv regions in our dataset. The distribution of each measure is shown for the non-redundant set in Figure 3.6a.

As described in the methods section, the vector C was chosen to have the most conserved length over the non-redundant set of structures. The distance, dc, is this length. It has a mean value of 16.2 Å and a standard deviation of only 0.3 Å.

The HL torsion angle has the largest range. It varies from -72.2° to -45.14°. The angle with the smallest range is HC1 which varies from 64.8° to 77.4°. However, variation in each angle cannot be compared on the same scale. For instance, a one degree change in the HC1 angle is not equivalent to a one degree change in the HC2 angle. They describe different directions of movement and therefore affect the physical coordinates of the domains by different amounts.

We compared our absolute measures to the relative measure of orientation RMSD (Section 3.2.8). This was carried out by calculating the change in our absolute measures and the orientation difference as measured by RMSD, between every pair of structures in the non-redundant set. The measure that was most correlated with RMSD was the HC2 angle (Spearman's $\rho = 0.54$). However, for a given RMSD there can be a range of angle changes. This range of angle differences increases with RMSD. For instance, pairs of structures with an RMSD of 0.5Å $\pm$ 0.1Å have HL angle differences in the range 0.0° to 4.8°, whilst those with an RMSD of 3Å $\pm$ 0.1Å have a range from 0.0° to 19.6°.

When plotting the angles against each other no direct correlation is observed



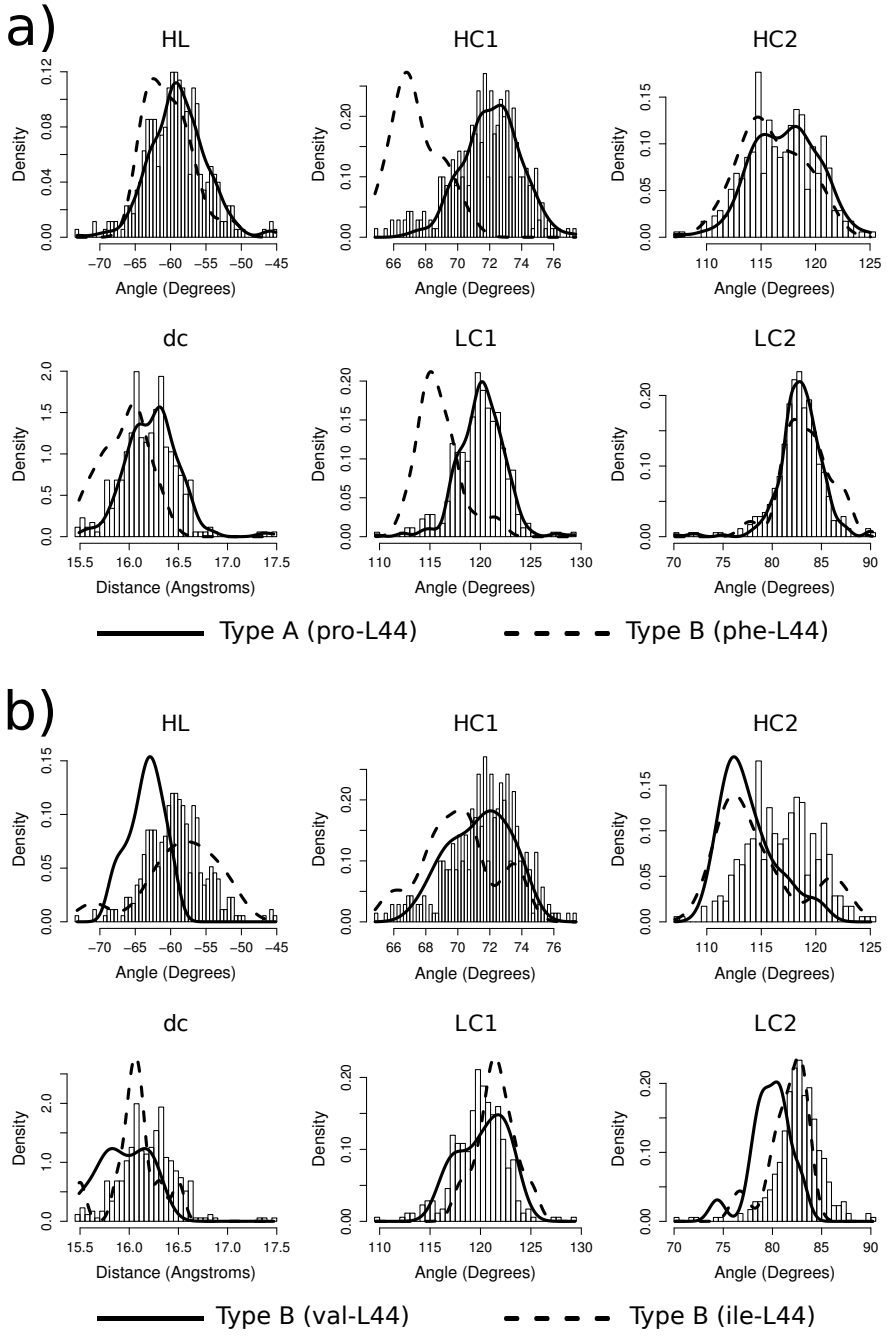


Figure 3.6: Histograms showing the distribution of each of our VH-VL orientation measures for the non-redundant set of structures. Each antibody variable domain can be placed at a position in this structural space. The location on each distribution of structures with position L44 occupied by a) proline or phenyalanine or b) valine or isoleucine are shown in each measure. Each line represents the Gaussian density estimation for the relevant distribution.

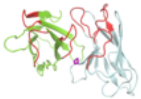
between any of the measures. This suggests that the orientation of the variable domains does not vary from one structure to another about a single axis, i.e. one cannot define a single torsion angle to adequately define VH-VL orientation.

3.3.2 Comparing the orientation of interface type clusters

In order to compare with Chailyan *et al*'s work, structures in the non-redundant set were stratified into four subsets in accordance with the residue present at position L44 (Section 3.2.6). Figure 3.6 shows how these subsets of structures differ in their orientation measures.

Most striking is the location of the phe-L44 subset of structures on the LC1 and HC1 bend angle distributions. The structures in Chailyan *et al*'s cluster B were predominantly from this subset. Those structures in the pro-L44 subset best represent Chailyan *et al*'s cluster A. Comparing the distribution of HC1 angles for the phe-L44 subset to that for the pro-L44 subset using a Kolmogorov-Smirnov (K-S) test [Massey Jr, 1951] showed that the former is significantly more acute (p-value $= 4.0\times10^{-12}$). The same is true for the LC1 angle (p-value $= 3.6\times10^{-10}$). These differences correspond to a tilting of the variable domains towards each other at the binding site in cluster B structures relative to those in cluster A. It is also indicative of the significantly smaller binding site area in cluster B structures than in cluster A structures observed by Chailyan *et al*.

We therefore propose that the difference in orientation that Chailyan *et al* describe with their clusters is in the HC1 and LC1 angles. This is not the same mode of orientation variation described by Abhinandan and Martin's packing angle. In fact, a change in Abhinandan and Martin's torsion angle is approximately perpendicular to a



change described by the HC1 or LC1 angles. Therefore, the apparently inconsistent sets of positions that these studies propose as influential for VH-VL orientation, may be due to identifying positions that are important for different directions of pose.

However, Chailyan *et al*'s cluster B also contained a small number of structures that did not have phenylalanine at L44. Instead, valine or isoleucine was present. We find that structures in our ile-L44 subset do not have significantly different preferences for orientation in any measure than structures with a residue other than isoleucine at position L44. In contrast, those structures in the val-L44 subset have significantly different HL, HC2 and LC2 angles to structures with a residue other than valine at position L44 (K-S test p-values 5.7×10^{-5}, 9.1×10^{-4} and 2.0×10^{-6} respectively). This is not the same mode of orientation differentiation that is found in phe-L44 structures and is more similar to the mode of orientation variation described by Abhinandan and Martin's packing angle.

These results indicate that different residues at the same position may influence the VH-VL orientation in different directions. Residues at position L44 can discriminate between structures that have preferences for either the HL torsion angle or the HC1 and LC1 angles. This may explain why L44 was one of only two positions that both Chailyan *et al* and Abhinandan and Martin assigned high importance in their descriptions of orientation.

Those structures with phenylalanine at L44 have light chains predominantly from the mouse IGLV1 subgroup. Those with valine at the same position instead are predominantly from the mouse IGKV10 subgroup. As these subsets were found to have distinct orientations, we investigated the effect of heavy and light subgroup pairings on the orientation measures. Further dependence on the subgroup type was not found. Similarly, a recent study compared Abhinandan and Martin's packing

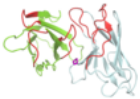
angle to VH-VL subgroup pairing and found no particular preference for subgroup pairs (Jayaram *et al.*, 2012). Therefore, we moved to consider the residue identity at individual positions for their influence on the orientation measures.

3.3.3 Important positions and residues for determining VH-VL orientation

Table 3.2 lists the top 10 positions and residues identified by the random forest algorithm as being important in determining each of our angular measures of VH-VL orientation. For instance, L87F is the highest scoring position and residue for the HL angle. Therefore, to model the structure of an antibody that has phenylalanine at position L87, one should only use template structures of antibodies that share this property in order to better predict the VH-VL orientation with respect to the HL angle. However, as this position does not score highly for the other measures, not using this information will not affect the prediction of the VH-VL orientation in the other angles.

Three of the positions that Abhinandan and Martin found to be influential for their packing angle (L44, L87 and H62) are also identified by our method to be influential for our similar torsion angle, HL. A further three positions, L38, L41 and L46, that Abhinandan and Martin found to be influential are included in our sets for at least one other of our measures.

Five of the eight positions that Chailyan *et al* proposed as influential for VH-VL orientation also score highly for the HC1 or LC1 measures. As shown in the previous section, these angles best discriminate between the authors' two clusters of structures. The remaining three positions were not interface positions and therefore



Angle	Top ten important input variables
HL	L87F* L42G†/L43T† L44V*† H61D L89L H43Q H43N/H44K H62K*/H89V L55H L53R
HC1	X*† L56P L41D*† L89A L97V L94N L34H L34N L96W L100A
HC2	H62S* H62K*/H89V H43K H50W H46K/H62D* H35S H61Q H43Q H33W H58T
LC1	L91W L89A X*† L97V L94N L50G H43Q L56P H62S* L55A
LC2	L50Y L42G†/L43T† L44V*† L42Q† L55H H99Y L93T L94L L53R L85T

Table 3.2: X represents the variable L36V†/ L38E*/ L42H†/ L43L†/ L44F*†/ L45T/ L46G*/ L49G/ L95H. * denotes positions also found to be influential by Abhinandan & Martin [2010]. † denotes those positions also found to be influential by Chailyan *et al.* [2011].

were not included in our analysis.

The HC2 measure is found to have a strong dependence on heavy chain positions and especially on the position H62. Examination of those structures which have lysine at H62 and those with serine at H62 finds that they have different preferences for the HC2 angle. Those with lysine have significantly smaller HC2 angles than those with serine (K-S test, p-value $= 2.0 \times 10^{-14}$). Further investigation of the relationship between HC2 and the residue at position H62 revealed that the size of the residue present is generally inversely related to the size of the angle. For instance, those structures with aspartic acid at H62 have small angles whilst those with alanine have large ones. The size of the amino-acid affects the packing at the domain interface and therefore the VH-VL orientation as measured by HC2.

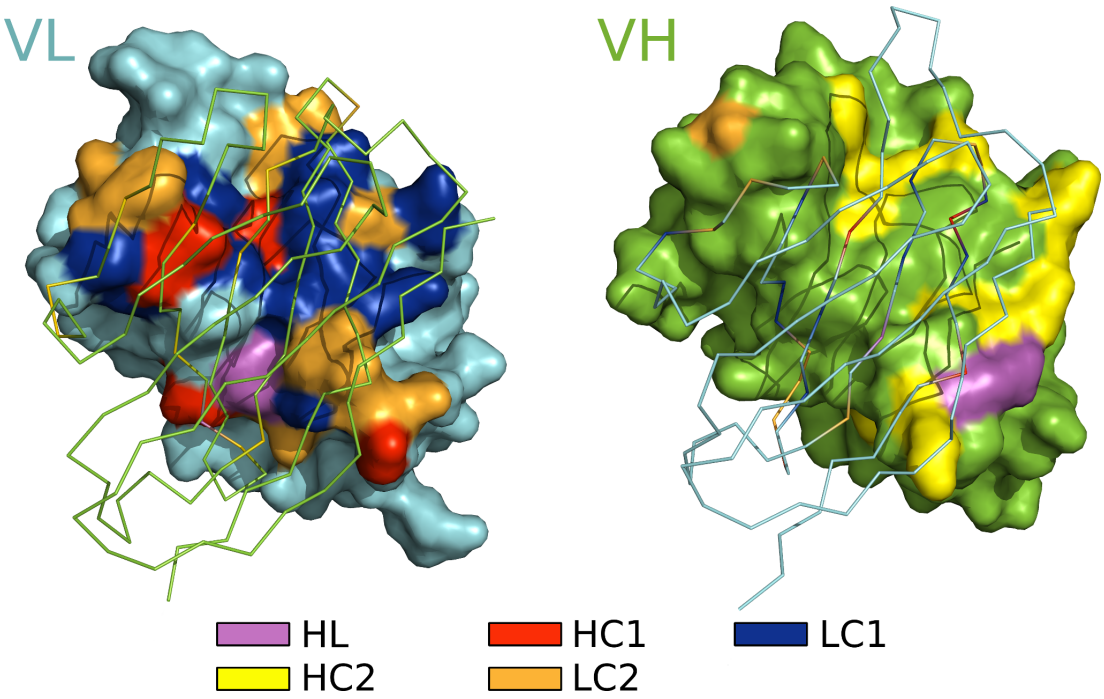


Figure 3.7: The location of positions that are found to be influential for the angular measures. Those positions which are influential for more than one measure are coloured in order of priority: LC1, HC1, LC2, HC2, HL. Positions that are deemed to be influential for the HC2 and LC2 measures are located on the periphery of the VH-VL interface, whilst those for the LC1 and HC1 measures pack into the centre of the interface.

3.3.4 Location of important positions on the VH-VL interface

Figure 3.7 shows the location on the variable domains of the positions we have identified to be influential for each angle. The HC1 and LC1 measures describe a tilting like motion of one domain towards the other. The positions we identify as important for these angles tend to be in the core of the interface and predominantly on the VL domain. Whereas LC2 and HC2 describe a twisting like motion of one domain with respect to the other. In this case, the positions tend to be on the periphery of the inter domain interface (predominantly on the VL domain for LC2 and VH for HC2). The positions that are important for the HL torsion angle tend to also be important for either the HC2 or the LC2 measures (i.e. sites on the periphery of the interface).

3.3.5 Variation in orientation between sequence identical structures is dependent on antigen type

In this section we consider if our VH-VL orientation measures are informative with respect to antibody-antigen binding. No significant relationship was found between the absolute VH-VL orientation and the antigen type the antibody is specific for. To test the conservation of the VH-VL orientation in antibodies, sequence-identical structures with the same bound-state were identified in the full dataset. In the set of Fvs with antigens bound, 205 sequences were identified that had two or more structures. The difference in the VH-VL orientation angle was calculated for each pair of sequence-identical structures and the mean difference calculated for each sequence case. Similarly, in the set of Fvs with no antigen bound, 45 sequences were identified that had two or more structures and the mean difference in angles was calculated for each.

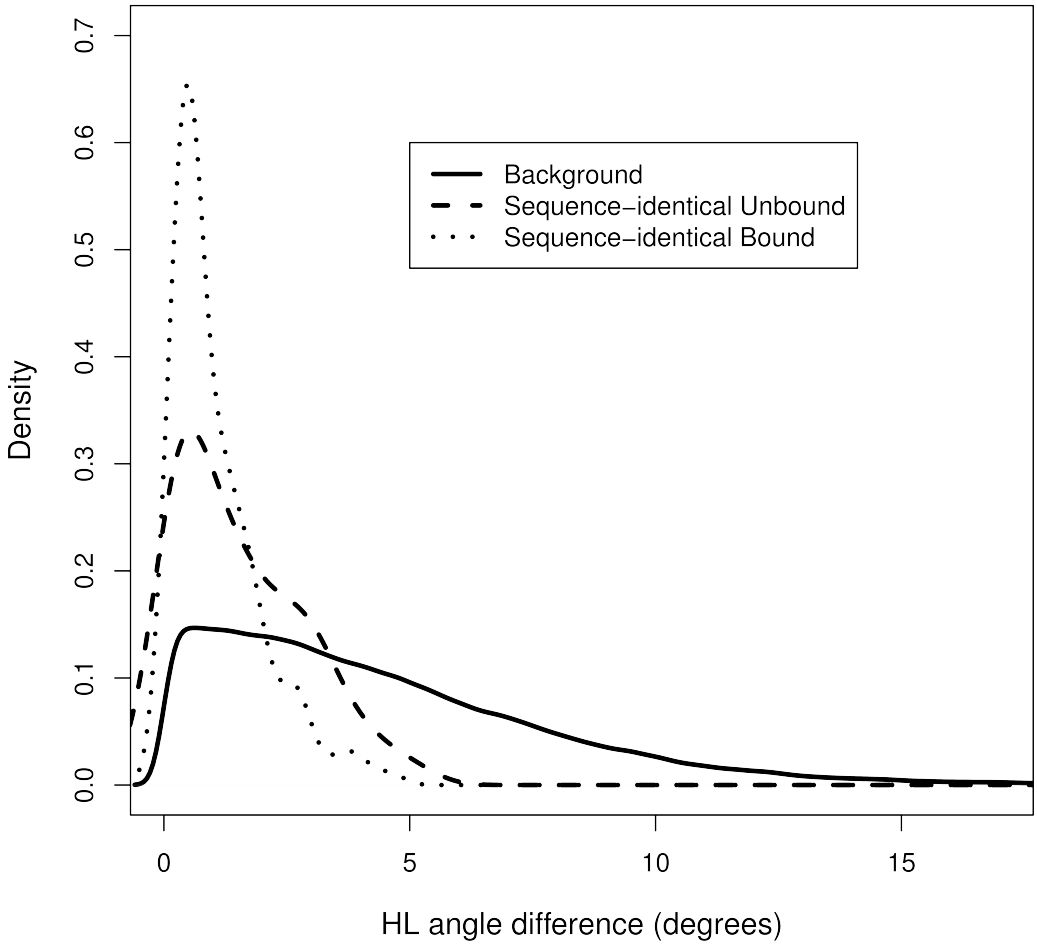


Figure 3.8: Distributions for the variation in the HL angle for sequence identical bound structures, sequence identical unbound structures and structures with sequence identity of less than 90% (background). The structures of sequence identical antibodies have a more conserved angle than the variation observed between non-identical antibodies. However, the HL angle of bound sequence identical antibodies is more conserved than that for unbound sequence identical antibodies.

The variation was found to be different between the bound and unbound sets of structures in only the HL angle. The bend angles are more likely to be conserved in sequence-identical structures as large differences would imply a loss of contacts at the VH-VL interface. However, larger changes may occur in a torsion angle motion while still maintaining the VH-VL contact surface.

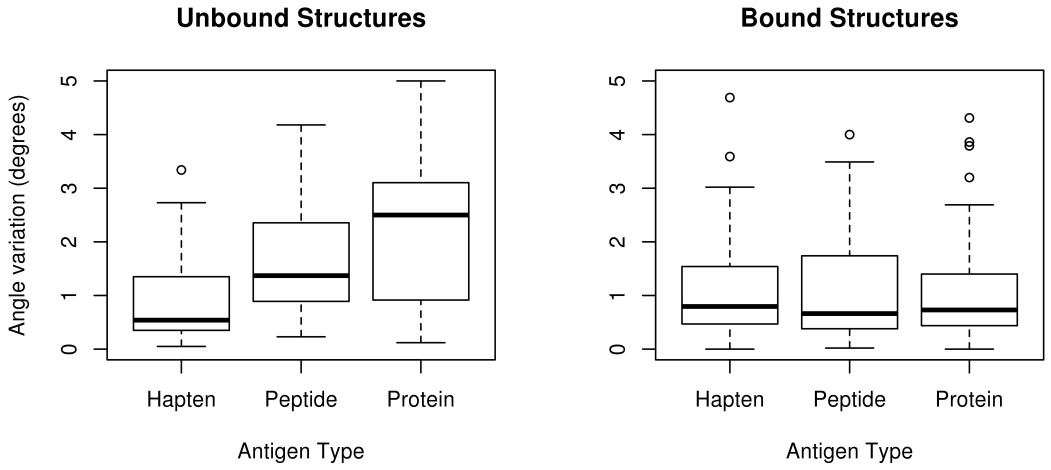


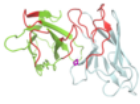
Figure 3.9: Distributions of the variation in HL angle of sequence-identical antibodies stratified by antigen type for a) bound structures and b) unbound structures. Polypeptide antigens of 25 or more residues in length are classified as proteins, those with less than 25 residues as peptides and those antigens that are small molecules as haptens. The variation in HL angle for unbound structures of sequence-identical protein-binding antibodies is significantly more than for unbound structures of sequence-identical hapten-binding antibodies. However, the variation in VH-VL orientation for bound structures of sequence-identical antibodies is independent of the antigen for which they are specific for. This suggests that the VH-VL orientation for protein-binding antibodies is more flexible than it is for hapten-binding antibodies in the free-state. However, when bound, protein-binding antibodies rigidify and share the same degree of VH-VL orientation conservation as hapten-binding antibodies.

Figure 3.8 shows the distributions of the HL angle variation for unbound structures of sequence identical antibodies and bound structures of sequence identical structures. The distribution of HL angle differences for those structures which are less than 90% sequence identical is also shown in order to demonstrate the background variation. We

find that the bound structures of sequence identical antibodies have a more conserved HL angle than unbound structures. Although not a direct indication of the dynamics of the molecule, this result may reflect the effect of complex formation reducing the structural space available to the antibody. If this is true, then the degree to which the structure is stabilised is likely to be dependent on the size of the antigen it binds.

To test whether there is any dependence of orientation variation on antigen size, we stratified the sets of sequence-identical structures into three types of antibodies: hapten binding; peptide (or carbohydrate) binding; and protein binding. There were 99 protein binding, 38 peptide binding and 68 hapten binding sets of bound structures and 15 protein binding, 11 peptide binding and 19 hapten binding sets of unbound structures. Figure 3.9 shows the distributions in HL angle variation for the bound and unbound sets stratified in this way. For protein binders, we found that the variation in unbound structures was significantly larger than the variation in bound structures (p-value $= 0.0014$). However, neither the peptide nor the hapten distributions were significantly different between unbound and bound forms. The variation in angle in unbound structures is also significantly larger for protein binders than for hapten binders (p-value $= 0.0076$), whilst the variation in bound structures is not significantly different.

These results suggest that the VH-VL orientation for protein binding antibodies is more flexible than for hapten binding antibodies. However, upon binding, both types of antibodies are found to have a similar degree of conservation in variable domain pose. The flexibility of a protein leads to a higher entropic cost for binding. Therefore, this result may be due to the fact that larger antigens are able to overcome this cost, whilst smaller hapten antigens require a more rigid binding partner. This result is physically intuitive and statistically significant. However, the number of data



points is small and the analysis would benefit from the availability of more structures, especially in the unbound form.

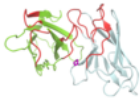
3.3.6 ABangle

We have implemented our method of calculating the VH-VL orientation in antibody structures in the computational tool, ABangle. Given a structure in PDB format, the software fully automates the procedure of recognising all Fv regions present and will calculate and report the orientation measures for each VH-VL pair. This includes calculation for multiple NMR models and an option for automation of orientation calculation for single chain Fvs. The distribution of angles for single chain Fvs is generally similar to that of the standard VH-VL pairs.

ABangle also allows for the analysis of VH-VL orientation. Individual Fv regions found in the PDB can be retrieved using their PDB code and its chain identifiers. Sets of structures can also be selected by a number of properties including residue identity at a Chothia position, species, heavy or light chain subgroup and CDR loop length. The orientation of these structures can then be visualised in two ways: as plot of the distribution of the orientation measures against the non-redundant set background e.g. Figure 3.6b; or using PyMol [Schrödinger, 2010] by aligning all the structures to either the VH or VL consensus structure.

In addition to the work presented in this thesis, ABangle is in active use at both UCB and Roche, has been used to investigate variable domain conformations in molecular dynamics simulations [Knapp *et al.*, 2014] and is used as part of the Antibody Modelling Assessment competition [Teplyakov *et al.*, 2014]. We provide ABangle at `http://opig.stats.ox.ac.uk/webapps/abangle` and as part of SAbDab as

described in Section 2.3.2.6.



3.4 Conclusion

In this chapter we have presented a method to fully characterise the VH-VL orientation of antibody structures in an absolute sense. This allows us to investigate not just relative changes between Fv regions, but how this change relates to variation observed in all structures.

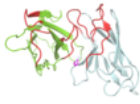
We use our method to explain why two previous studies [Abhinandan & Martin, 2010; Chailyan *et al.*, 2011] identify different framework positions as important for the orientation of the VH and VL domains. We find that the difference between the two clusters identified by Chailyan *et al* is related to a change in our LC1 and HC1 angles, whilst the difference described by Abhinandan and Martin, the VH-VL packing angle, relates instead to a change in our HL torsion angle. Thus, the apparent inconsistency in the positions that these studies find influential for pose is because they have described approximately perpendicular modes of variation.

Our orientation measures have allowed us to investigate which positions and their residue identity affect pose in different directions. We find similar positions to both Abhinandan and Martin and Chailyan *et al* in the analogous modes of orientation and identify others that may have a significant influence for different modes of variation. Our measures also offer insight into structural variation between bound and unbound forms of antibodies. We find that the variation in VH-VL orientation in antibodies in their unbound form is dependent on the size of antigen they bind. However, in the bound form, no such dependence is found suggesting a reduction in conformational space available to the antibody.

Our method has been implemented in the computational tool, ABangle. This allows researchers to investigate the structural space of antibodies. We have demon-

strated its use for the applications of comparing sets of structures, finding influential positions and investigating the variation of orientation in homologues. It could also be used to compare the orientation of specific antibodies, especially in their unbound and bound forms. The orientation measures allow absolute scales of variation to be quantified. They can therefore be incorporated into Fv modelling protocols as a framework for modulating the VH-VL pose or for model assessment [Teplyakov *et al.*, 2014]. ABangle's ability to automatically and rapidly calculate the VH-VL orientation of a number of structures also lends itself to the investigation of the conformational space observed in NMR models of the Fv or for molecular dynamics studies of antigen receptors [Knapp *et al.*, 2014].

Although ABangle has been developed to investigate domain orientations in antibodies, the methodology can be applied generally to pairs of domains with high numbers of examples in the PDB. In the next chapter we demonstrate this by investigating domain orientations of a different yet related protein, the T-cell receptor.



Chapter 4

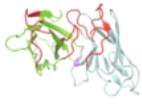
Comparing variable domain orientations in different antigen receptors

The majority of the work presented in this chapter is contained within the following publication and is my own contribution unless otherwise stated.

J. Dunbar, B. Knapp, A. Fuchs, J. Shi, and CM. Deane, 2014. Examining variable domain orientations in antigen receptors gives insight into TCR-like antibody design. *PLoS Comp. Bio.*, **10**, e1003852

4.1 Introduction

In the previous chapter we showed how the orientation of the variable domains in antibodies could be characterised in a consistent manner. We developed the ABangle method and used it to show that different types of antibodies had different orien-



tations. One result showed that the conservation of antibody VH-VL orientation is related to the class of antigen the antibody binds.

Another component of the immune system, the T-cell receptor (TCR), only binds to peptide antigens and only when they are presented on the surface of a cell by the major histocompatibility complex (MHC). However, like the antibody, the TCR binds using its variable region that consists of two domains, Vα and Vβ, which are analogous to the antibody VL and VH domains.

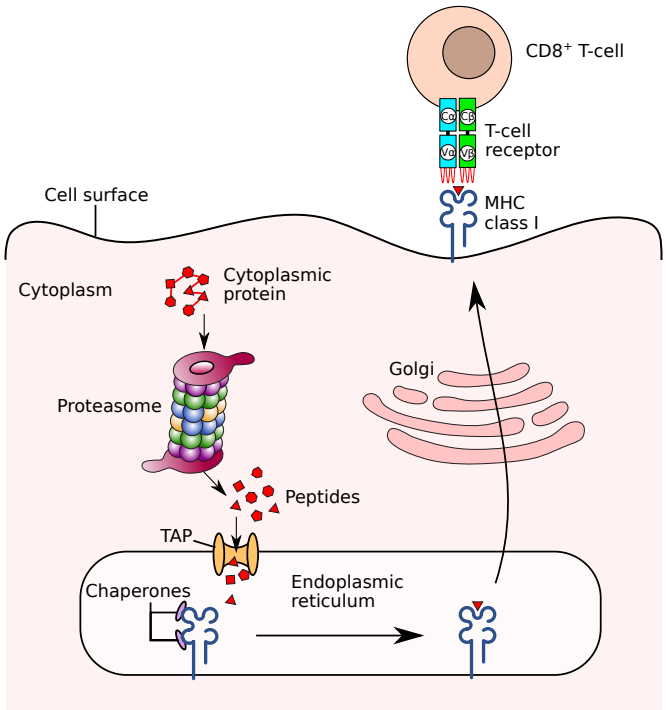


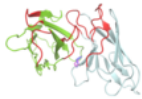
Figure 4.1: Peptide loading, presentation and recognition. In a cell, proteins are degraded by the proteasome into peptide fragments. These are loaded onto MHC class I in the endoplasmic reticulum and transported via the Golgi to the cell surface. Here, the MHC-peptide epitope is presented to the immune system. Immunogenic peptides are recognised as being such by receptors on T-cells that signal for apoptosis to be induced in the presenting cell. Shown here is a CD8+ T-cell whose receptors recognise immunogenic peptides presented by MHC class I. The process is similar for a CD4+ T-cell. However, in this case peptides are derived from extracellular proteins and presented by MHC class II. Figure adapted from Yewdell *et al.* [2003]

The humoral immune system can only directly interact with antigens either in extracellular space or on the surface of cells (Section 1.2.1). However, the markers of a pathology can be within a cell (e.g. a mutated protein expressed by a cancerous cell). Therefore a different pathway, largely separate from the action of antibodies, is used by the immune system to defend the organism in these cases. This cell-based immunity is predominantly mediated by T-cells [Janeway *et al.*, 2001].

In most human cells proteins are continuously broken down into short peptide fragments by the proteasome [Rock *et al.*, 1994] (Figure 4.1). They are transported to the endoplasmic reticulum by the transporter for antigen processing (TAP) [Neefjes *et al.*, 1993]. Here, molecular chaperones help load peptides onto MHC class I (MHCI). The MHCI-peptide complex is transported to the cell surface by the Golgi. At the surface, the MHCI presents the peptide to the surroundings [Yewdell *et al.*, 2003]. T-cells that express the CD8 co-receptor on their surface, scan the peptides for those that are immunogenic and indicative of a pathology [Janeway, 1992]. When such peptides are recognised, the T-cell will induce apoptosis in the antigen presenting cell.

Another function of T-cells is in the activation of B-cells (Section 1.2.1). B-cells are able to internalise antigens that they bind. These molecules are broken down into peptides and processed in the endoplasmic reticulum. In B-cells, these peptides are loaded onto MHC class II (MHCII) and presented on the surface of the cell. T-cells with the CD4 co-receptor on their surface (helper T-cells) recognise immunogenic peptides and release cytokines to activate the B-cell [Janeway, 1992].

Whilst the immunological synapse is formed by several components, including either the CD4 or CD8 co-receptor, in both cases of MHC-peptide recognition TCRs are the molecular binders to the T-cell epitope (i.e the MHC-peptide complex) [Bromley



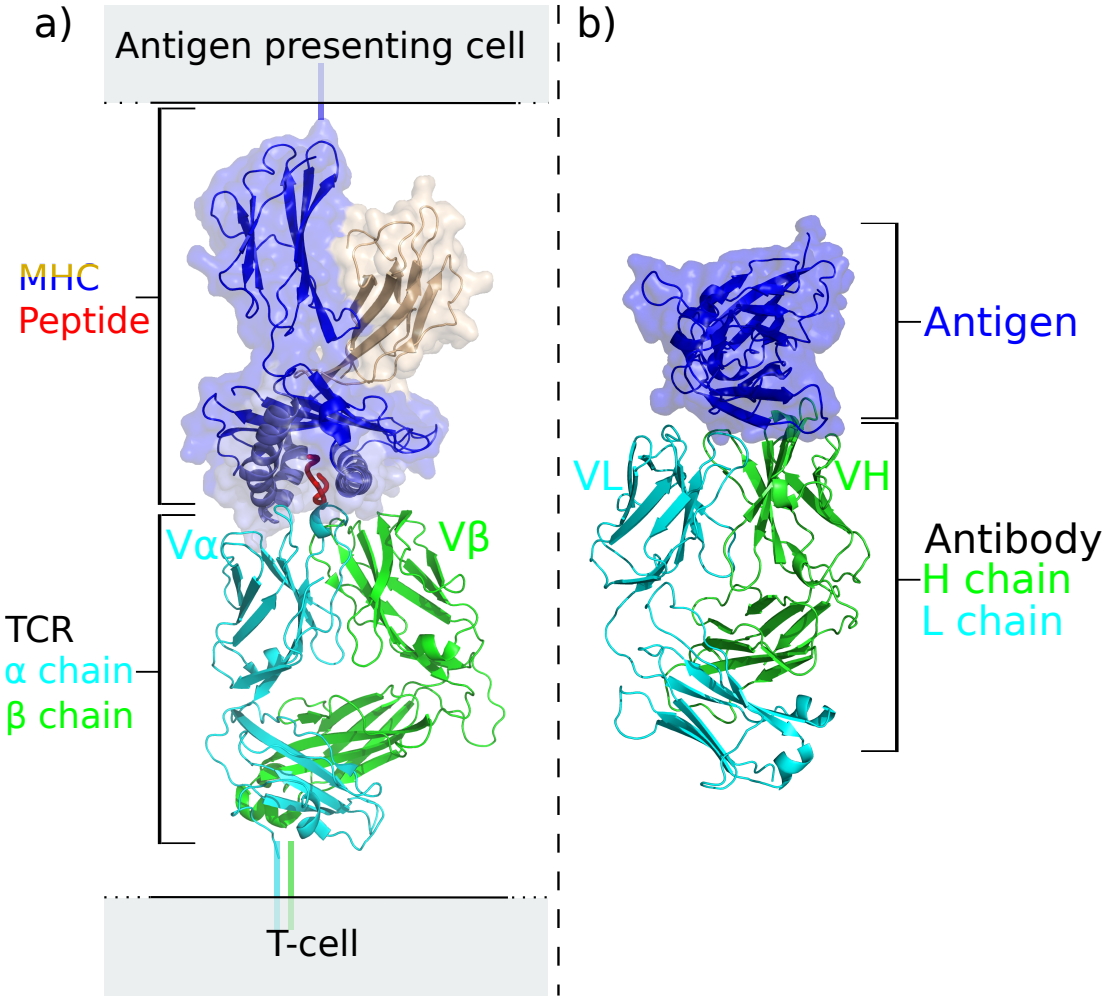
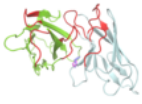


Figure 4.2: a) The TCR-MHC-peptide complex (PDB 1mi5 [Kjer-Nielsen *et al.*, 2003]). TCRs are expressed on the surface of T-cells and specifically recognise the complex between MHC and immunogenic peptides. Here, the immunogenic peptide is in complex with an MHC I. The TCR consists of two chains each with a variable domain (Vα and Vβ). These domains are analogous to the antibody VH and VL domains shown in b) (PDB 1ahw [Huang *et al.*, 1998])

et al., 2001]. Like the antibody Fab, the TCR consists of two chains, α and β. A TCR also uses its variable domains, Vα and Vβ, to directly interact with the antigen (Figure 4.2). The domain structures are similar to VH and VL and have similar framework and CDR regions [Garcia *et al.*, 1999]. Vβ domains, like the antibody VH, are built from v, j and d genes. Vα domains, like the antibody VL, are built from v and j genes, The process of v(d)j recombination (Section 1.2.3.1) allows the diversity in the sequence and structure of the TCR. However, unlike B-cells, T-cells do not undergo somatic hyper mutation (Section 1.2.3.2). Instead, they are generated in the thymus where they undergo thymic education [Janeway *et al.*, 2001]. Only T-cells with TCRs that can bind at least weakly to an MHC are allowed to proliferate (positive selection). However, those whose TCRs bind strongly to MHCs presenting peptides of proteins produced by the host's cell are deleted (negative selection).

Whilst antibodies can bind to a highly diverse set of antigenic shapes and sequences, TCR epitopes are relatively similar to one another. Therefore, although similar in function, the binding sites of the two molecule types are under different evolutionary pressures. Antibodies must be versatile in binding to many different targets whilst the TCR antigen binding site can optimise to associate specifically with the shape of MHCs and the peptides they present.

Although the diversity in potential T-Cell epitopes is far smaller than for antibodies, variation is present in the MHC class, allele, and the bound peptide [Rudolph *et al.*, 2006]. MHC class I molecules (MHCI) are present on the surface of most cells. The binding groove consists of a single chain and is closed at both ends. This enables MHCI to present peptides of typically nine residues in length. MHCI peptides are of intracellular origin. The MHCI is attached by one trans-membrane region to the cell surface. In contrast, MHC class II molecules (MHCII) form their binding



groove with two chains. The binding groove of MHCII is open at both ends allowing for peptides of lengths of 20 or more residues to be presented [Rammensee, 1995]. MHCII peptides are derived from extracellular proteins. The MHCII is attached with two transmembrane regions to the cell surface. Even within the same MHC class, sequence differences allow for a specific binding repertoire of peptides and TCRs [Rammensee *et al.*, 1999]. Currently 8,124 human MHCI and 2,409 human MHCII alleles are known [Robinson *et al.*, 2013]. Despite these differences MHCs share the same overall fold with the peptide presented above an anti-parallel beta sheet floor and flanked by two kinked alpha helices. This means the TCR binding site is highly structurally conserved.

The relative similarity between T-Cell epitopes is also reflected in the structure of the complex formed between the TCR and MHC/peptide (pMHC). Typically, the TCR is found to bind diagonally with respect to the MHCs peptide groove (Figure 4.3). This geometry of interaction is referred to as the canonical binding mode. Those cases where the TCR is positioned in a different orientation are known as non-canonical [Rudolph *et al.*, 2002]. Although no strict definition for canonical association has been made, a number of TCRs that bind unusually have been found to be related to autoimmune responses [Wucherpfennig & Call, 2009; Yin *et al.*, 2012]. Ensuring a canonical binding mode may therefore be important for preventing the recognition of self-antigens.

Naturally, T-Cell and B-Cell receptors perform different functions in the immune system. An example of a non-natural interface between the two receptor types are TCR-like antibodies. These are antibodies that are specific for T-cell epitopes. In diseases where T-cell functionality is inhibited, engineered TCR-like antibodies provide a method for delivering cytotoxic drugs and induction of infected cell apoptosis [Cohen

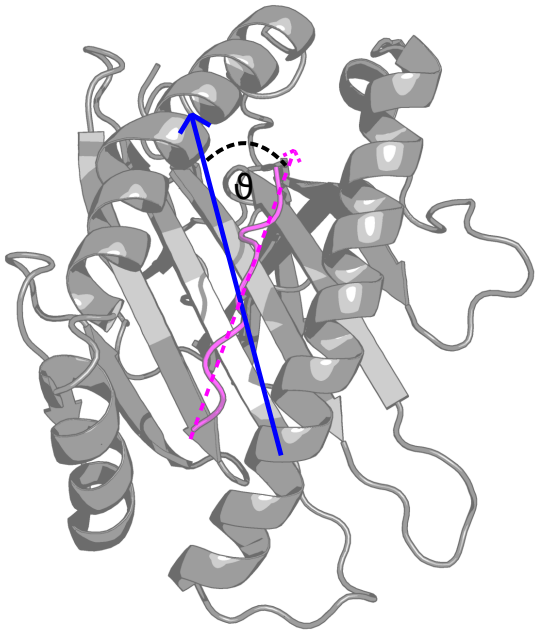


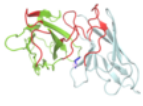
Figure 4.3: The structure of an MHC class I (grey) with a bound peptide (magenta). The peptide binding groove is formed between two α helices and has a β sheet floor. The arrow shows a schematic of the TCR-MHC docking angle. The docking angle (θ) is the angle made between the principal axis of the peptide (magenta, dashed line) and the vector between the interfacial cysteine Cα's (IMGT position 104) of the Vα and Vβ domains (blue). The majority of TCRs bind diagonally with respect to the MHC helices. This is known as the canonical binding mode.

& Reiter, 2013; Dahan & Reiter, 2012]. An advantage of using a TCR-like antibody as a therapeutic over a soluble recombinant TCR [Molloy *et al.*, 2005] is their higher natural affinity to an antigen [Neumann *et al.*, 2009]. Subsequently, a lower level of affinity enhancement is required. In addition, the serum half-life of an antibody is measured in days to weeks, whilst that of a TCR is of the order of hours [Dostalek *et al.*, 2013].

Antibodies able to recognise T-cell epitopes must mimic a TCR and replicate their binding site properties to enable pMHC specificity. Producing TCR-like antibodies is a difficult procedure inhibited by the process of isolating the desired MHC/peptide epitope [Cohen & Reiter, 2013]. However, advances in phage-display technology have allowed the generation of such molecules for diagnostic and potential therapeutic purposes [Andersen *et al.*, 1996; Biddison *et al.*, 2003; Cohen *et al.*, 2003]. Successfully generated TCR-like antibodies share some similar pMHC-binding properties with TCRs that have the same specificity [Stewart-Jones *et al.*, 2009]. For example, mutations to certain positions on the MHC disrupt both TCR and antibody binding [Biddison *et al.*, 2003]. However, previous comparisons between individual antibodies and TCRs have also reported that the receptors use different features to achieve pMHC specificity [Mareeva *et al.*, 2004; Hülsmeyer *et al.*, 2005]. This suggests that antibodies do not necessarily mimic TCR canonical binding. The therapeutic importance of TCR-like antibodies motivates the investigation of how they may be engineered to exhibit improved MHC specificity. Whilst the TCRs CDR residues primarily mediate MHC recognition, the orientation between the TCR Vα and Vβ domains has previously been proposed to contribute in determining T-Cell epitope specificity [McBeth & Seamons, 2008].

In this chapter we investigate the Vβ-Vα domain orientations in TCR variable

regions and compare them to the analogous structural space in antibodies. Functional reasons for differences in the two receptor types are proposed by analysing the influence of domain orientation on the structural properties of a TCR-MHC complex. Given that the orientation space is different we investigate which sequence positions may be contributing factors and suggest residues on the domain interface of antibodies that may induce a more TCR-like orientation.



4.2 Methods

4.2.1 Dataset

All X-Ray crystal structures containing a paired TCR α and β chain were extracted from the protein data bank (PDB) [Berman *et al.*, 2000] using the international ImMunoGeneTics information system (IMGT) [Lefranc *et al.*, 2009]. IMGT numbering was retained for each of the TCR variable domains. Those 92 structures with a resolution of better than 3Å formed the full redundant dataset. This set contained 49 TCRs bound to MHCI, 15 bound to MHCII and 28 unbound. A sequence identity filter of 90% was applied over the TCR variable domains using CD-hit [Li & Godzik, 2006] to form a non-redundant set of 39 structures. Nineteen were bound to MHCI, 11 to an MHCII and 9 were unbound.

A non-redundant antibody dataset was created using the structural antibody database (SAbDab) [Dunbar *et al.*, 2014b]. Again, a sequence identity filter of 90% was applied to the full sequence of the variable domains using CD-hit. The non-redundant antibody set consisted of 441 structures. The IMGT DomainGapAlign tool [Ehrenmann *et al.*, 2010] was used to apply IMGT numbering to the variable domains in each of the structures.

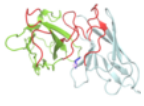
The structures in SAbDab were filtered to identify TCR-like antibodies bound to MHCs. A BLAST search [Altschul *et al.*, 1990] for the sequence of the MHCI and MHCII was performed using a database created from the antigens in SAbDab. Manual inspection of these top hits and a keyword search for phrases similar to "TCR like antibody" identified a total of 4 structures. One of the structures (3hae) is higher affinity mutant of another (3gjf) from the same experiment [Stewart-Jones *et al.*, 2009]. Therefore 3 cases were identified for TCR-like antibodies in complex with an

MHC.

4.2.2 Rationale for domain equivalence

To compare the variable domain orientations of antibodies and TCRs using the ABangle tool it is necessary to define which domains are equivalent in the two receptor types. Is VH equivalent to Vβ or Vα? Comparing the mean sequence identity between the domain types showed that Vκ and Vλ (collectively VL) domains were more similar to both Vα and Vβ than VH was to either. Therefore, sequence identity alone provides no clear indication.

To compare VH-Vα/VL-Vβ or VH-Vβ/VL-Vα equivalence we examined the residue conservation at individual IMGT positions in the sets of domains. To check the VL/Vα and VH/Vβ equivalence we identified IMGT positions where the VH and Vα domains have the same conserved residue whilst in VL domains a different conserved amino-acid is present (Figure 4.4). Positions such as these suggest that the alternative equivalence should be used. These positions were identified for both equivalences i.e. VH-Vβ/VL-Vα and VH-Vα/VL-Vβ. A position was deemed to be conserved if it had the same amino-acid in 50% or more sequences. Five positions support making VH-Vα/VL-Vβ equivalent whilst 10 positions support making VH-Vβ/VL-Vα equivalent. This equivalence is also supported by the similarity in the combination of genes that make up corresponding domains. Both Vβ and VH are generated from v, d and j genes whilst Vα and VL are built from just v and j genes. We therefore we use the VH-Vβ/VL-Vα equivalence herein.



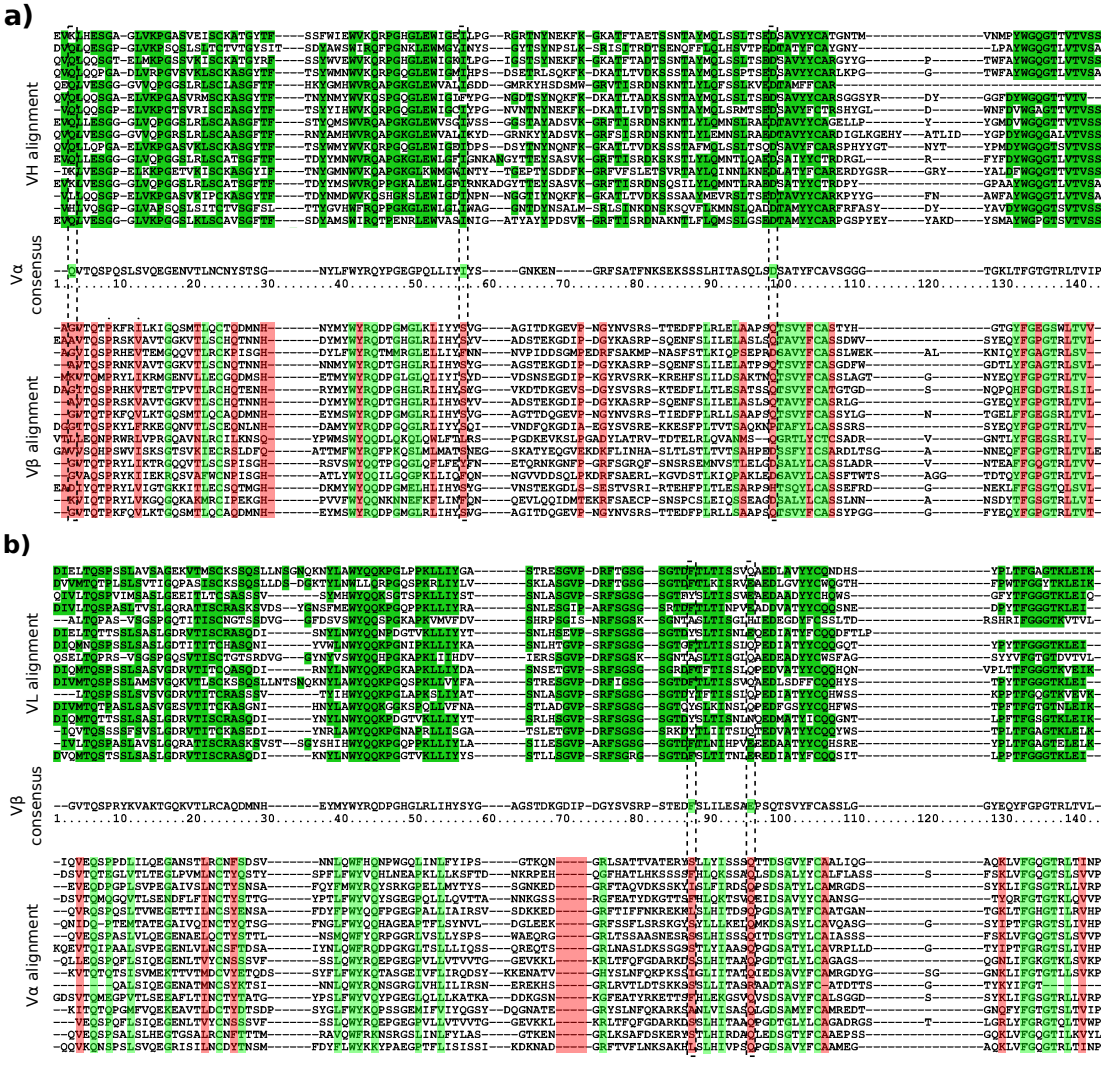


Figure 4.4: Comparing the sequence profiles of the four domain types. a) Part of the VH domain alignment is shown with conserved (50%) columns coloured dark green. In the Vβ domain alignment, the columns are highlighted if the corresponding position in both VH and Vβ domains is conserved. These positions are coloured green if the conserved residue is the same and red if it is different in the two domains. Where the conserved residue is different, the position is highlighted with a dashed box if the most popular amino-acid in Vα domains (Vα consensus) is the same as the conserved VH residue. These positions are those that suggest that the alternative pairing should be used i.e. VH-Vα and VL-Vβ. b) The corresponding image comparing the VL domain alignment to the Vα domain alignment.

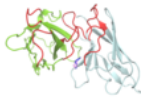
4.2.3 Variable domain orientation root mean square deviation

The variable domain orientation root mean square deviation (RMSD) is a measure of the difference in orientation between two structures. As an example, the orientation RMSD between two TCR structures (Tx and Ty) is calculated as follows. Tx and Ty are structurally aligned using the Cα coordinates of their shared Vα domain framework positions. Tx's Vβ domain is then independently aligned to Ty's Vβ domain. The RMSD is calculated between Tx's Vβ domain in its native orientation and the transformed position. The same procedure is also performed using the Vβ domains for the initial alignment and calculating RMSD between Vα domains. The mean of these two values describes the difference in the relative orientation. All structural alignments were performed using the Biopython SVD superimposer [Cock *et al.*, 2009].

The orientation RMSD was calculated for each pair of structures in the non-redundant antibody set, the non-redundant TCR set and between the sets. These measurements were used as distances between the structures allowing them to be clustered using a complete-linkage hierarchical clustering method [Maechler *et al.*, 2013].

4.2.4 Applying the ABangle methodology to TCRs

Whilst relative measures of orientation differences such as the orientation RMSD give a magnitude to changes in pose they do not reveal how structures vary. The ABangle methodology was developed to characterise the orientation between the VH and VL domains in antibodies. It gives six absolute measures, five angles and a distance, to fully describe the orientation between the two domains. The same method can be applied to any set of structurally defined homologous domain pairs. The process is



presented in detail in Chapter 3 and can be summarised as:

1. Define the most structurally conserved positions for both domains (core-sets).
2. Fit frames of reference through interface positions using coordinates from multiple structural alignment
3. Define consensus (mean) structures for both domains.
4. Compute or choose the pivot axis about which to measure orientation.
5. Define six measures about the resulting coordinate system to describe orientation.

This process was followed for the TCR structures in our dataset. The core-set positions are listed in Table C.1 in Appendix C. To allow for direct comparisons to be made with antibodies, the pivot axis, C, was chosen to be the same as in the antibody ABangle work. However, the pivot axis was also calculated for TCRs and has been used in a separate study [Knapp *et al.*, 2014] to investigate variation in structural properties of the TCR-MHC/peptide complex using molecular dynamics simulations. Whilst the two axes are similar, the TCR-derived one lies closer to the centre of the domain-domain interface than the antibody derived one.

The resulting coordinate system allows for the 6 absolute measures of orientation described in the previous chapter to be defined. To recap, the distance dc is the length of C. HL describes a torsion angle of one domain with respect to the other. HC1 and LC1 are angles that describe the tilt of one domain towards the other. HC2 and LC2 describe twisting-like differences between the orientations of variable domain structures. In each of these cases the ABangle antibody nomenclature has been retained.

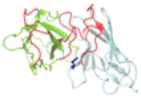
4.2.5 The MHC-TCR docking angle

TCRs are often observed to bind to the MHC in a diagonal mode with respect to the peptide binding groove. As described in Section 4.1, qualitative observations have labelled this association geometry as the canonical binding mode. The binding geometry can also be described quantitatively using the MHC-TCR docking angle [Mareeva *et al.*, 2008]. This is defined as the angle between the major axis of the peptide and the vector between the Cα atoms of the interface cysteines (IMGT position 104) on the Vα and Vβ domains.

4.2.6 Measuring the effect of variable domain orientation in a TCR-MHC complex

The effect that changing the Vβ-Vα orientation has on the structure of a TCR-MHC complex was investigated. A single TCR-MHC-peptide complex was chosen (PDB 1mi5 [Kjer-Nielsen *et al.*, 2003]) as the native structure. The MHC is of class I and presents a peptide 9 residues in length from the Epstein Barr Virus. Both the MHC and TCR are human in origin and the structure is solved at a resolution of 2.5Å. The docking angle of the TCR-MHC is 51.4° and is therefore considered to be in the canonical range. A set of 20 non-redundant structures were chosen from across the antibody orientation space (Table C.2 in Appendix C). Similarly, a set of 20 structures was sampled from the non-redundant TCR dataset (Table C.3 in Appendix C). These two sets are referred to as the antibody and TCR decoy sets respectively.

The native complex was compared to structures made when the Vβ-Vα orientation was changed to assume the poses in the TCR and antibody decoy sets. The following protocol was used to change the Vβ-Vα orientation.



4. Comparing variable domain orientations in different antigen receptors

Given a decoy structure, the VL domain (or Vα domain for TCR decoys) was aligned to the Vα domain of the native complex (Figure 4.5a). This structural alignment was performed using the shared framework positions of the two domains. The native Vβ domain was then transformed independently to assume the resulting pose of the decoy VH domain (or Vβ domain for TCR decoys). Thus, the Vα domain is in a native position relative to the MHC-peptide complex whilst the Vβ domain inherits the orientation of the decoy (Figure 4.5b,c). After remodelling the residue side chains with Scwrl4 [Krivov *et al.*, 2009], clashes were identified between the Vβ domain in its decoy position and the MHC/peptide chains using the program Molprobity [Davis *et al.*, 2007]. Any pair of residues with a Cβ-Cβ atomic distance of less than 7Å were also recorded. In addition, the DOPE score [Shen & Sali, 2006] of the entire complex was calculated to assess the energetic change in modifying the domain orientation. No further optimisation of the structure was performed other than the repacking of side chains. It may be possible for a different score to be obtained after applying conformational optimisation. The process was then repeated in an analogous way to change the Vα position to assume the decoy orientation whilst maintaining a native Vβ pose.

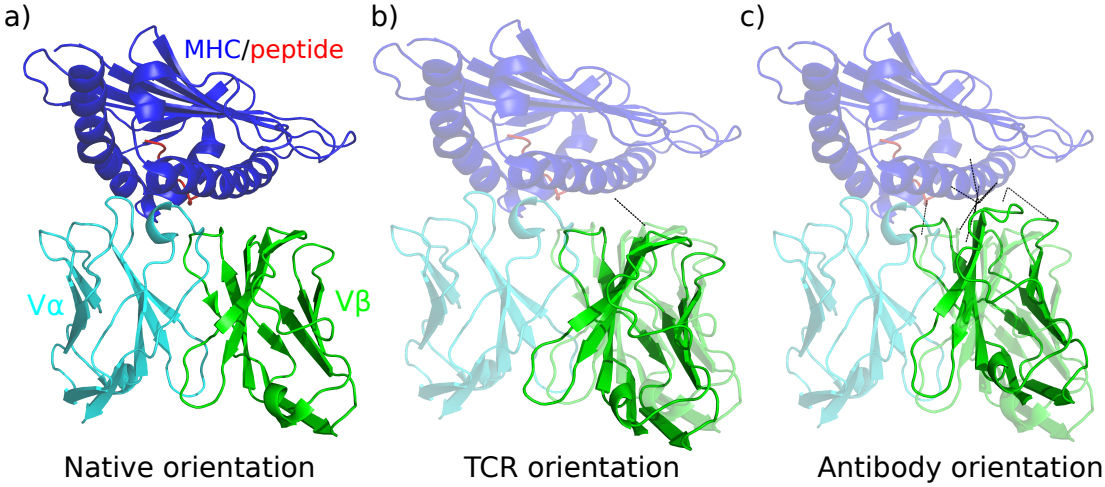


Figure 4.5: Changing the Vβ-Vα orientation of a TCR/MHC-peptide complex. a) The native complex 1mi5. No residue clashes are observed between the TCR and the MHC-peptide. b) The complex with the TCR placed in the Vβ-Vα orientation of another TCR structure. A single clash (black line) is identified between the Vβ domain and the MHC. c) The complex with the TCR placed in the VH-VL orientation of an antibody. Multiple residue clashes, shown by black lines, are found between the Vβ domain and the MHC.

4.3 Results

4.3.1 Vβ-Vα orientations are different from VH-VL orientations

Antibody and TCR structures were clustered based on their orientation RMSD (Figure 4.6). The TCR and antibody structures fall into separate clusters with very little mixing of the two types. This demonstrates a difference in the variable domain orientations of TCRs and antibodies. The orientations of the TCRs alone fall into two distinct clusters. One might suspect that this may be related to MHC type or species. However, neither cluster has a strong bias for either property. Furthermore, no clear general characteristic can be identified to separate TCRs with these different conformations. In general, TCR binding orientation appears to show little relationship

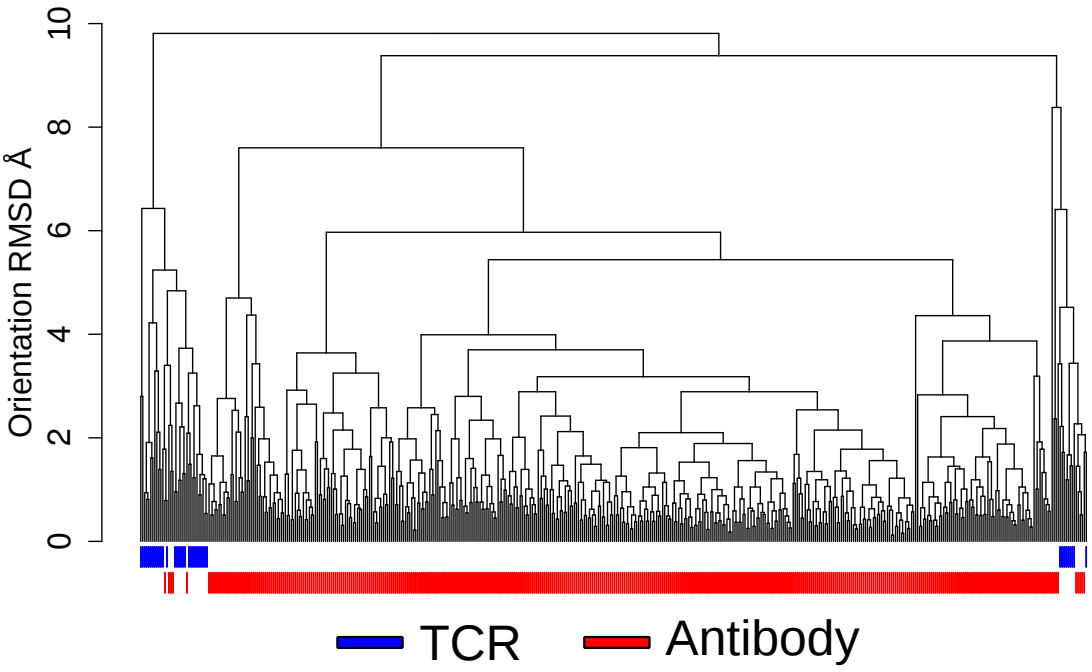


Figure 4.6: Clustering of the antibody and TCR structures by their relative orientation RMSD. The two types of structure fall largely into different clusters.

to Vβ-Vα orientations, however the scarcity of non-canonical binding examples makes it hard to draw any specific conclusions.

Whilst the RMSD measure shows that the orientations of antibodies and TCRs have distinct conformations it does not show us how the structures differ. We therefore used the ABangle methodology to identify structural differences.

4.3.2 ABangle measures reveal how antigen receptors differ

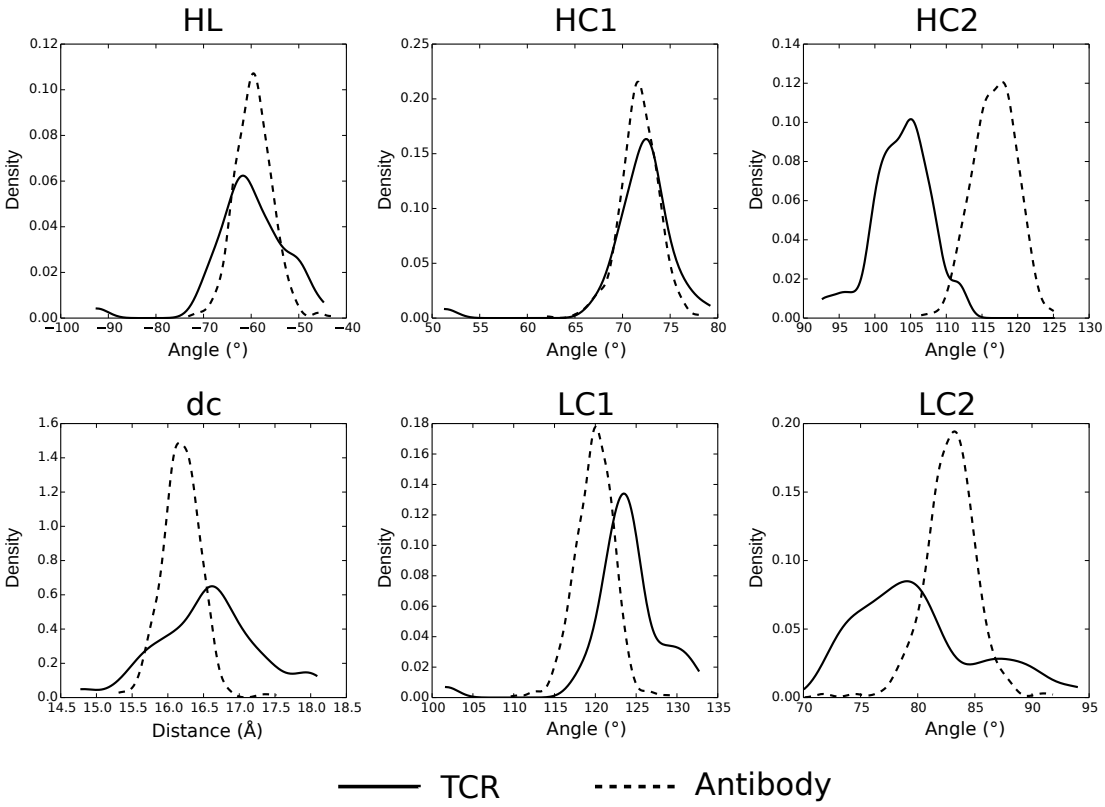


Figure 4.7: The ABangle orientation measures for antibodies and TCRs. The receptor types are found to have different orientations. The difference is best characterised by the HC2 angle. This corresponds to a twist of the VH or Vβ domain with respect to the VL or Vα domain respectively.

ABangle's six orientation measures were calculated for each structure in the

dataset. Figure 4.7 shows the distributions for both antibodies and TCRs. A similar magnitude of variation is observed in the orientations of the two sets of structures. However, as with the RMSD measure of orientation, a clear difference in orientation preference is seen between the two structure sets. This is best characterised in the HC2 angle where the distributions are significantly different (Kolmogorov-Smirnov test p-value 2.2×10^{-16}). In fact, antibodies are almost never observed to reach the extreme HC2 orientation seen in TCRs. Two of the other orientation angles, LC1 and LC2 are also significantly different (p-values 7.1×10^{-12} and 8.6×10^{-10} respectively) whilst the two receptor types do not have significantly different HL torsion angles or HC1 bend angles. The distributions of the dc length are significantly different with a p-value of 3.46×10^{-8}. The difference in orientation between antibodies and TCRs can be best described as a twisting-like change of the variable domains with respect to one another. Figure 4.8 demonstrates such a structural difference in orientation between the antigen receptor types.

4.3.3 Antibody orientations are incompatible with binding in a TCR-like mode

For such similar molecules in terms of their function and domain structure, it is somewhat surprising that TCRs and antibodies have such different orientations. As discussed previously, TCRs bind specifically to peptide antigens only when they are presented by MHCs on the surface of a cell. In comparison, antibodies are not restricted in the same way and bind to a far more varied set of antigenic shapes.

To investigate whether the variable domain orientation is likely to affect function we calculated the influence that changing the Vβ-Vα orientation had on the structure

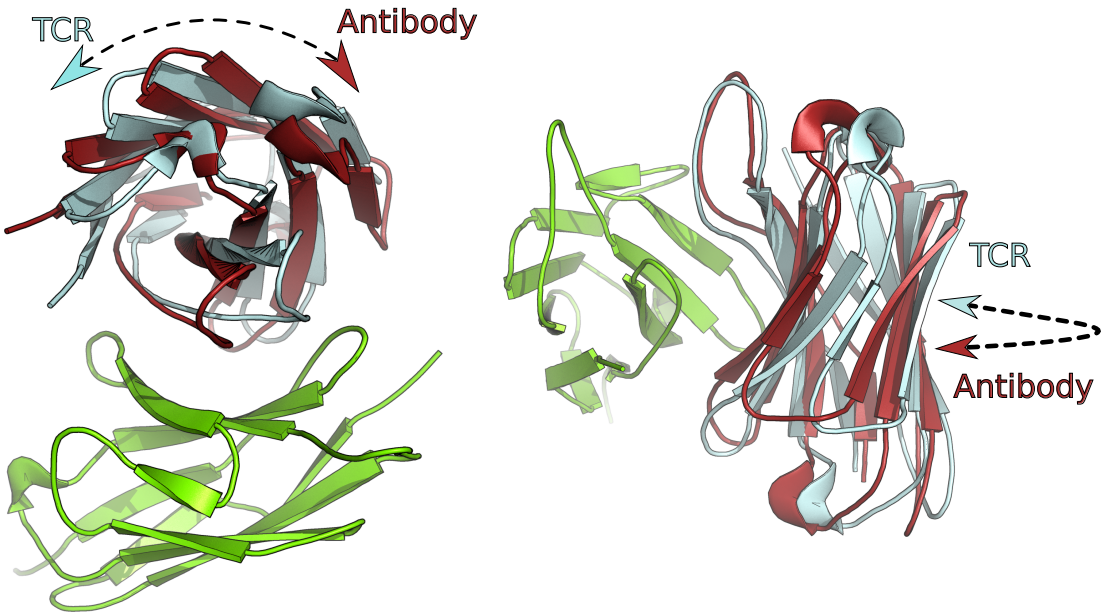


Figure 4.8: Two views showing the difference in variable domain orientation between antibodies and TCRs. The VH domain (red) and VL domain (green) of the antibody structure 3hzk is shown in its native orientation. The same antibody is also shown in the orientation assumed from the TCR structure 3qiu. In this case the VH domain is shown in cyan. The difference in variable domain orientation can be described as a twist of one domain with respect to the other and is best characterised using our HC2 twist angle.

of a single TCR-MHC complex PDB 1mi5). Two sets of orientation decoys were used, one from antibody structures and the other from other TCR structures (for full details see the methods section). For both of the decoy sets, the number of Vα-MHC/peptide residue clashes and number of Vβ-MHC/peptide residue clashes were counted using Molprobity. No clashes were observed between either domain and the MHC/peptide chains in the native complex. For the TCR decoy set medians of 3.5 and 5 clashes were found for the Vα and Vβ domains respectively (4 and 8.5 before side chain rearrangement). In comparison, medians of 5 and 8 clashes were induced for the equivalent domains using the antibody decoys (10 and 14 before side chain rearrangement). A Mann-Whitney U test found the increase in the total number of clashes to be statistically significant (p-value 0.005). The increased number of clashes between the chains suggests that antibody orientations are incompatible with binding in a canonical TCR-like mode. For the native complex and the re-orientated complex sets, the mean Cβ-Cβ contact distances were calculated. The contact matrices for each set are shown in Figure 4.9.

The difference between the antibody-orientated complexes and the TCR-orientated complexes can also be measured using the energetics of the structures. Figure 4.10 shows the distributions of the DOPE score relative to the native state. Making the assumption that the native state approximates an energy minimum, we find that both sets of complexes are transformed to less favourable, higher energy states represented by higher DOPE scores. However, the set of complexes with orientations assumed from other TCRs are found to be significantly closer to the native state score than those with antibody orientations.

These results suggest that binding in a TCR-like mode requires the receptor to have a TCR-like variable domain orientation. Although the Vβ-Vα orientation is at

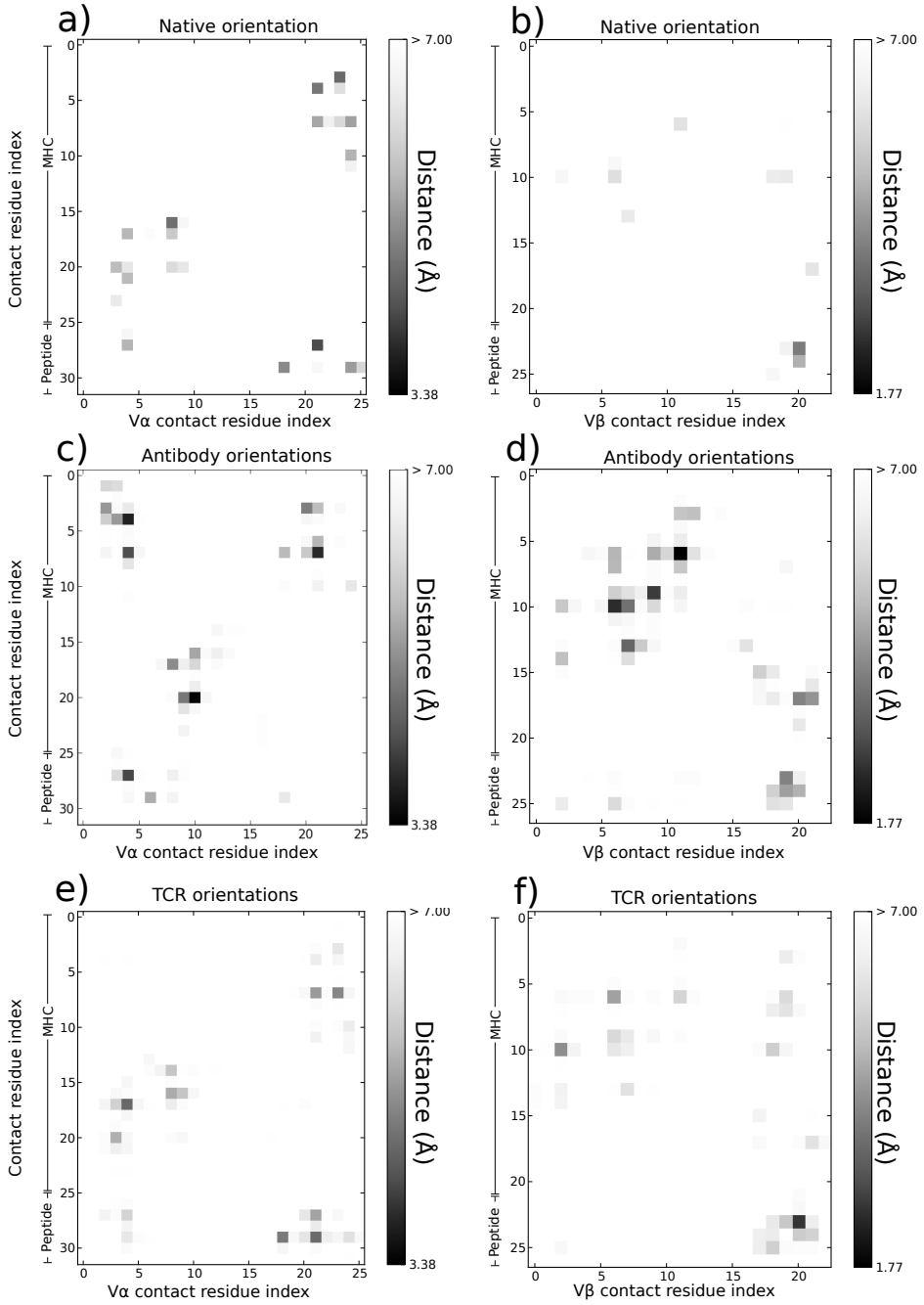
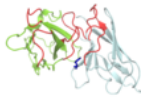


Figure 4.9: a) The Cβ-Cβ contact distances between the Vα domain and the MHC/peptide in the native complex. b) The corresponding image for Vβ and the MHC/peptide contacts. The TCR is made to assume variable orientations taken from the antibody decoy set. The mean contact distance over all of these decoys for c) the Vα domain and d) the Vβ domain is calculated. Similarly, the mean contact distances for e) the Vα domain and f) the Vβ domain are calculated when the structure assumes orientations from other TCR. When antibody orientations are assumed, unfavourable clashes are induced between the TCR and MHC. When other TCR orientations are assumed, the complex is disrupted to a lesser extent.



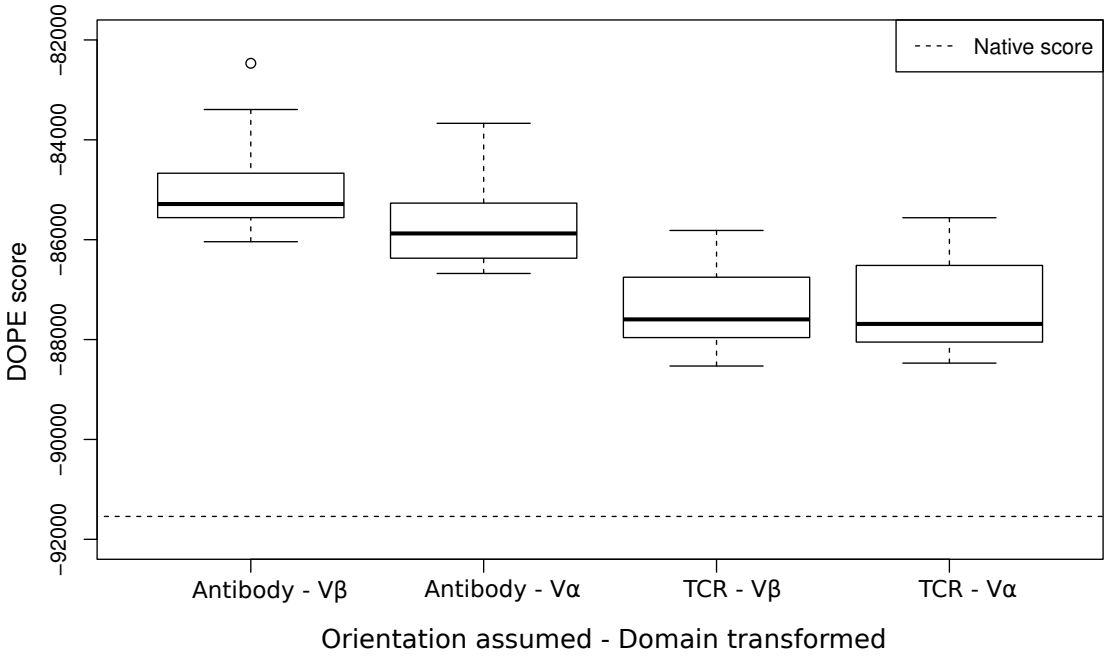


Figure 4.10: The DOPE scores of the TCR/MHC-peptide complex structure when placed in Vβ-Vα orientations assumed from the antibody decoy set and the TCR decoy set. The first boxplot corresponds to keeping the Vα domain in its native position and transforming the Vβ domain to positions observed in antibodies. The second boxplot is analogous to the first except now the Vβ domain is fixed and the Vα domain is repositioned. The third and fourth distributions are the respective boxplots for the TCR decoy set. The DOPE score is significantly higher for the antibody decoys than the TCR decoys representing less favourable interactions for the antibody orientations.

least as variable as the VH-VL orientation, TCR conformations are generally compatible with binding to the MHC in a canonical mode. In contrast, the space of antibody orientations is such that the canonical mode becomes more difficult to obtain. Despite this observation, antibodies may be engineered such that they bind to MHC-peptide epitopes. These molecules are known as TCR-like antibodies.

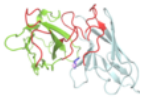
4.3.4 Orientations of TCR-like antibodies

The orientations of each of the TCR-like antibodies found in SAbDab were calculated using the ABangle methodology. Figure 4.11 shows where each of these lie compared to the background antibody distribution. There appears to be no preference for their variable domain orientation. However, one structure, PDB 3cvh, has an orientation that is considered extreme for an antibody. It has an HC2 angle that leans towards the domain orientations of TCRs. The remaining TCR-like antibodies have HC2 angles that are towards the other extreme of the antibody distribution.

We calculated the docking angle for both TCRs and antibodies bound to an MHC (Section 4.2.5). A canonical angle for TCRs is defined as being between 40° and 85° (Figure 4.12), a range similar to that found by previous studies [Hennecke & Wiley, 2001; Rudolph *et al.*, 2002].

Most of the antibodies in complex with an MHC do not bind in a canonical manner (Table 4.1). Instead, they bind diagonally but appear to use residues in the VL domain in the same way as a TCR uses the Vβ residues. Likewise the VH domains sit in a similar position to Vα domains in the canonical TCR-MHC complexes.

Only one of the TCR-like antibodies, 3cvh, binds in a similar mode to the TCR structures and with the expected equivalence. As discussed above, 3cvh, is the only



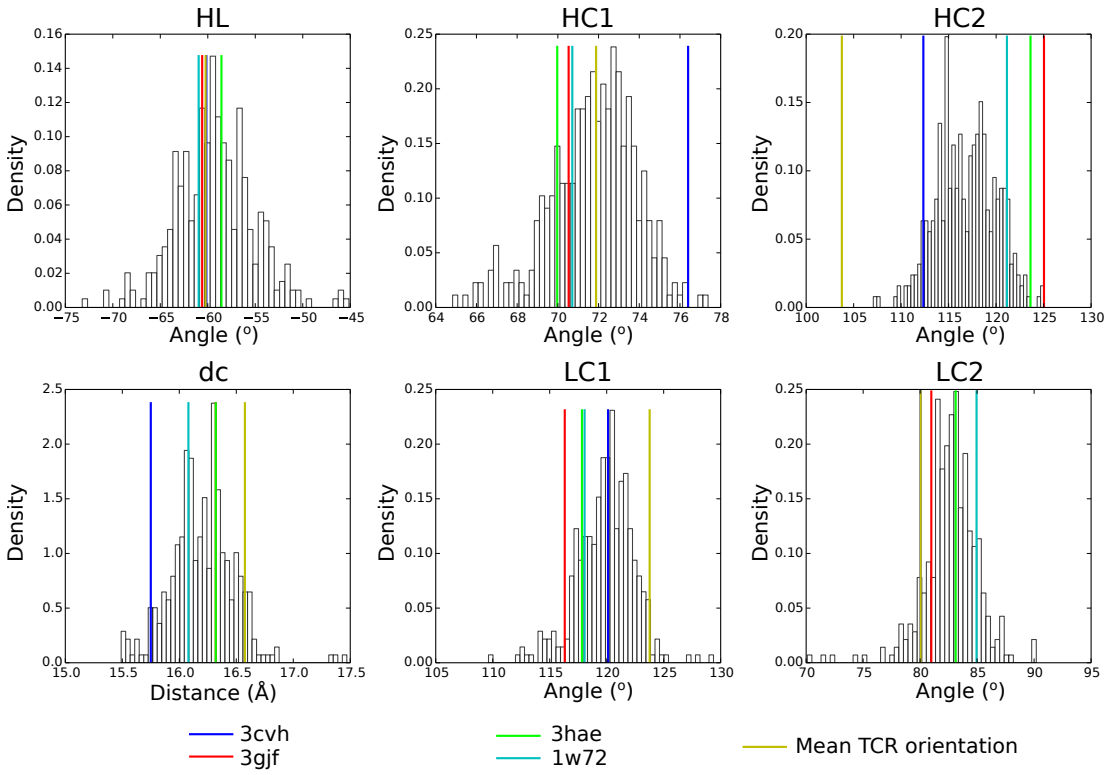


Figure 4.11: Orientation measures of TCR-like antibodies compared to the mean orientation of TCRs. Each vertical line corresponds the orientation measures of a single pair of variable domains in the unit cell of a crystal structure. The difference in orientation between TCRs and antibodies is best described using the HC2 twist angle. Only one antibody, 3cvh, has a TCR-like orientation in this measure.

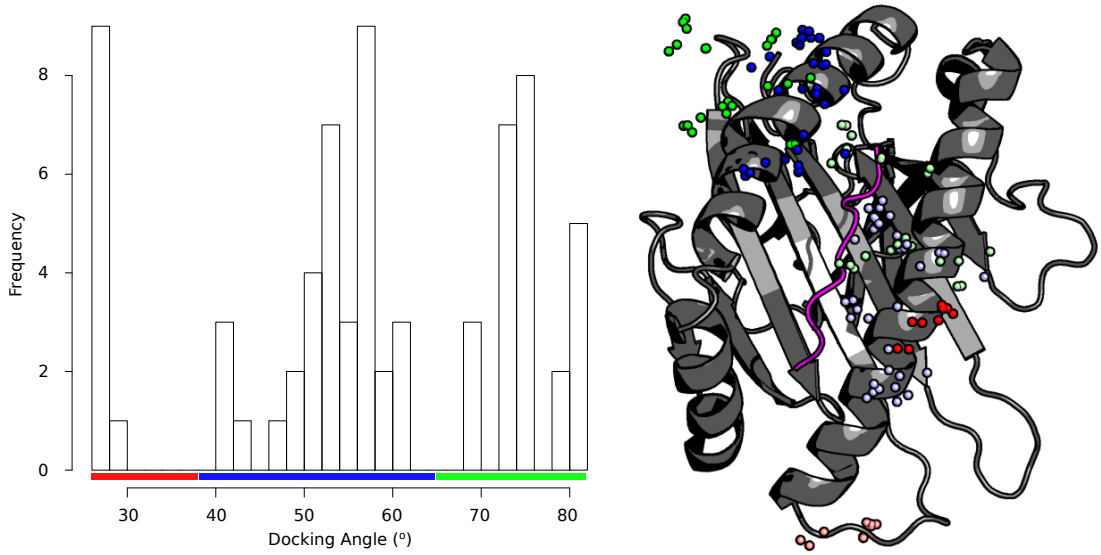
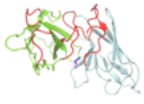


Figure 4.12: The TCR-MHC docking angle. Left, the distribution of docking angles found in the TCR dataset. We observe three qualitative clusters of angles. The blue and green clusters correspond to the canonical binding mode. The red cluster corresponds to the non-canonical binding mode found in a minority of complexes. Right, a representation of the docking poses mapped onto the MHC structure 1mi5. Each cluster member is represented by two spheres coloured according to its docking angle on the left. The darker sphere corresponds to the cysteine at IMGT position 104 on the Vα domain and the lighter sphere to the same position on the Vβ domain.

PDB	Heavy chain	Light chain	Docking Angle
3cvh	H	L	48.9°
3cvh	Q	R	49.3°
3gjf	H	L	122.3°
3gjf	M	K	125.1°
3hae	H	L	124.7°
3hae	I	G	124.9°
3hae	O	N	126.2°
3hae	T	S	123.6°
1w72	H	L	139.6°
1w72	H	L	140.3°

Table 4.1: The docking angle for each TCR-like antibody/MHC complex. For comparison, the distribution of TCR/MHC complex docking angles are shown in Figure 4.12.



antibody that had an variable domain orientation that approached that seen in the TCR space. Although only a single example, this structure suggests that in order to be able bind in a TCR-like mode to the MHC, a TCR-like variable domain orientation should be promoted.

4.3.5 Factors for promoting a TCR-like VH-VL orientation

Given that variable domain orientation may be related to the functions of antibodies and TCRs, we investigated which factors give rise to the structural differences between them. The type of antigen an antibody binds does not determine its absolute VH-VL orientation (Section 3.3.5 and Figure C.1 in Appendix C). Consequently, peptide-binding antibodies are no more similar in orientation to TCRs than general antibodies. The largest sequence variation in both receptor types occurs in the CDRs. Insertions also occur in these regions and cause variation in CDR loop structures length. The CDR3 in each domain type is likely to be the most influential loop as it makes the most inter-chain contacts.

Figures 4.13a and 4.13b show the length distributions of CDR3 in VL/Vα domains and VH/Vβ domains respectively. The CDR3s in both VH and Vβ have a wide range of lengths. Their length distributions are similar and are both centred on a length of 12 residues. The CDR3s of VL and Vα have very different length distributions. The antibody loop (CDR-L3) is almost always 9 residues long whilst CDR3 in the TCR domain (Vα) is longer with a modal length of 13 residues. The VL/Vα CDR3 loop is partially packed in the domain-domain interface. The fact that CDR-Vα3 is generally longer than CDR-L3 can be related to the observed difference in the HC2 twist angle (Figure 4.14) between TCRs and antibodies. The location of the loop

at the interface means that in order to compensate for a larger number of residues TCRs tend to reduce the HC2 angle as the interface twists open.

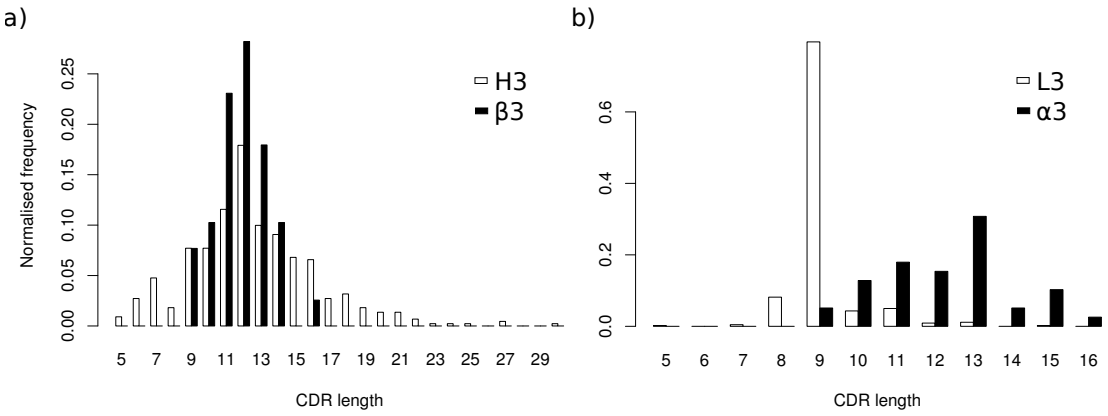


Figure 4.13: Length distributions of the CDR3 in a) the equivalent VH and Vβ domains and b) VL and equivalent Vα domains. In each case the IMGT definition of CDR has been used.

To test the effect of the CDR-L3 length on orientation in antibodies we identified all non-redundant structures in SAbDab with CDR-L3 loops of 13 residues. Nine such antibodies exist (Table C.4 in Appendix C). Eight of them have orientations that are typical of antibodies. This would suggest that having a long CDR-L3 loop alone does not promote a TCR-like orientation. One antibody, 3B5H10 Fab, (PDB structures 3s96 and 4dcq [Peters-Libeu *et al.*, 2012]) has an orientation similar to that of the TCRs. However, four of the other antibodies in this set have the same light variable germline subgroup (mouse IGLV3), are at least 95% identical sequences, and have identical L3 loops to 3B5H10. They are paired with different heavy chains suggesting that it is VH interface residues that give rise to the difference in orientation.

To isolate these positions, we compared the VH framework residues involved in contacts in the five IGLV3 structures (Table 4.2). Fifteen VH positions are involved in contacts in all the structures. In one position, IMGT 50, the antibody with the

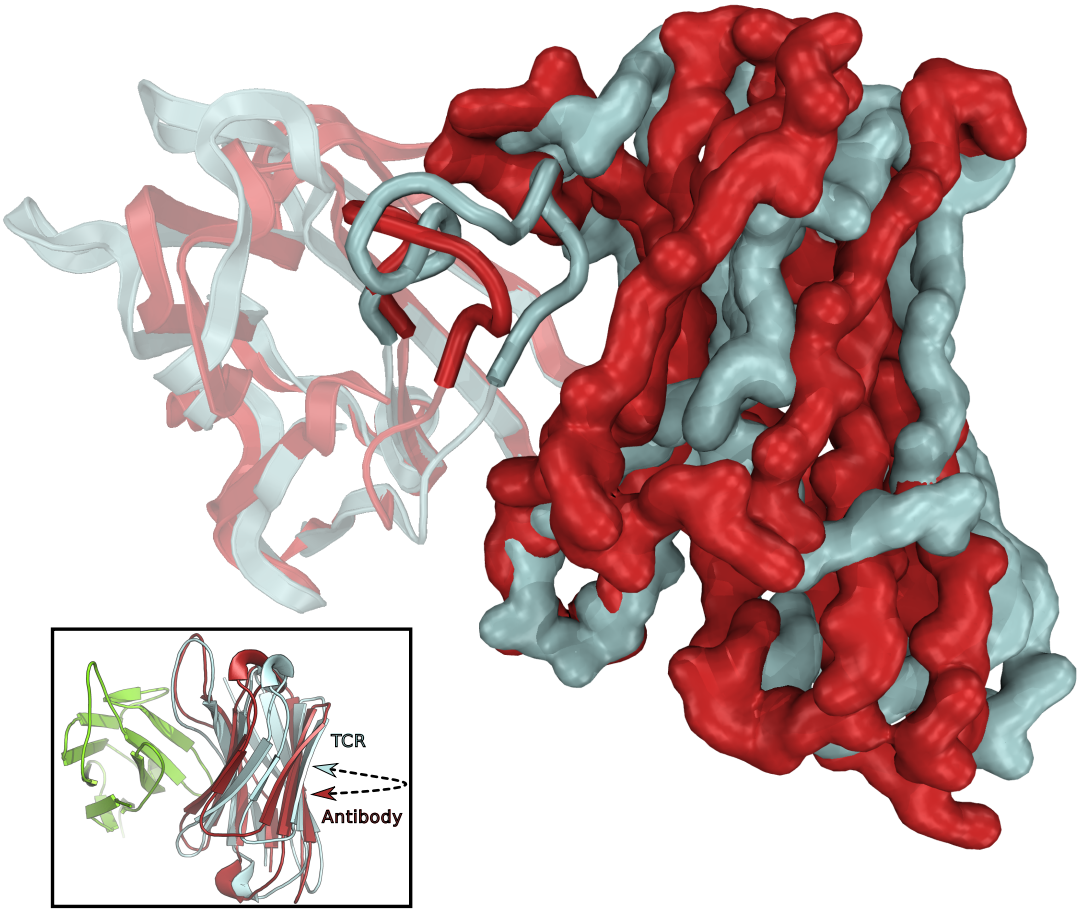


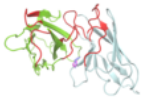
Figure 4.14: The influence of the VL/Vα CDR3 on variable domain orientation. Inset is part of Figure 4.8 where a single antibody structure was made to assume its native (antibody) orientation and a typical TCR orientation. The main panel shows a representative TCR structure in cyan and an antibody structure in red. The structures are aligned using the shared framework positions of the VL and Vα domains (left - ribbon). On the right are the VH and Vβ domains. The Vβ domain twists relative to the position that the VH domain sits in. The difference in orientation can be explained by the different packing of the CDR-L3 and the longer CDR-α3 loops (shown in ribbon representation). The TCR loop packs into the domain interface and opens up the orientation whilst the shorter antibody loop makes fewer inter-chain contacts. The difference in orientation is captured in our HC2 angle measure.

TCR-like orientation is a phenylalanine whilst the other four have leucine at the same position. This large aromatic residue makes contact with the CDR-L3 loop and is packed in the periphery of the interface. There are nine other structures in SAbDab with non-identical sequences that have phenylalanine at the same position. In each case, the residue has been mutated from the germline leucine. Seven of these have small HC2 angles ($<115°$) similar to that observed in TCRs. This suggests that making a mutation from leucine to phenylalanine at IMGT position H50 pushes the VH-VL orientation towards a more TCR-like conformation.

Structure	42	44	48	49	50	51	52	55	66	98	103	118	119	120	121
2qhr H L	V	Q	K	R	L	E	W	Y	Y	P	Y	W	G	Q	G
4isv B A	V	Q	K	G	L	K	W	W	T	A	F	W	G	Q	G
3ffd A B	I	Q	K	R	L	E	W	T	Y	P	Y	W	G	Q	G
2otu D C	V	Q	K	R	L	E	W	F	Y	P	Y	W	G	Q	G
3s96 A B	V	Q	K	G	**F**	E	W	W	T	V	F	W	G	Q	G
TCRs (Modal %)	Y (100)	Q (97)	H (25)	G (77)	L (82)	R (49)	L (74)	Y (64)	T (56)	Q (59)	F (64)	F (100)	G (100)	P (51)	G (100)

Table 4.2: Heavy contact positions in L3 length 13 structures with mouse IGLV3 subgroup.

IMGT position 50 in Vβ domains is also predominantly leucine. In TCRs the orientation therefore appears to be achieved using other interface positions. To examine how a VH-VL interface could be made more TCR-like, we examined the relative amino-acid frequencies at interface positions. Framework positions were identified that are conserved in both TCRs and in antibodies ($>50\%$) but with a different amino acid in each set. Ten positions were identified and are listed in Table 4.3. In order for a position to provide a feasible target for engineering it is preferable for a potential mutation to have been observed naturally. We therefore consulted the Abysis database [Martin, 2010] to find those positions where the conserved TCR amino-acid is also observed at the same position in 2% or more of non-identical antibody sequences. We find five such positions and refer to them as antibody compatible mutations.



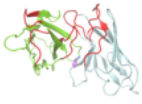
Position	Antibody residue (%)	TCR residue (%)	Antibody compatible mutation?
H/β 103	Y (80)	F (62)	Yes
H/β 100	A (93)	S (56)	No
H/β 120	Q (76)	P (51)	Yes
H/β 52	W (95)	L (72)	No
H/β 72	K (78)	P (69)	No
H/β 118	W (90)	F (97)	No
H/β 6	E (53)	Q (97)	Yes
H/β 4	L (98)	V (79)	Yes
H/β 42	V (75)	Y (97)	No
L/α 4	M (55)	V (72)	Yes

Table 4.3: Interface positions that are conserved in both TCRs and antibodies but with a different amino acid. Those positions where the TCR residue is also present in more than 2% of sequences are labelled as antibody compatible mutations.

Each of the antibody compatible mutations is located on the periphery of the VH-VL interface. Positions from a similar region of the domain interface were previously found to be influential for variation in the HC2 measure of VH-VL orientation (Section 3.3.4). Examining the sequences of the TCR-like antibodies finds that only the antibody that binds in the TCR canonical binding mode and has the most TCR-like variable domain orientation (3cvh) has the conserved TCR amino acid at any of these five positions (glutamine at H/β 6). When the ABangle measures for antibodies with the TCR amino-acid at these positions are calculated we find that two of the four positions, H6 and H4, select for structures with small HC2 angles indicative of the TCR-like orientations.

In summary, the packing of the CDR3 of Vα and VL domains is likely to be an influential factor for domain orientation. However, having a long antibody CDR-L3

does not necessarily replicate the TCR orientation. Instead, the VH-VL orientation can be made more TCR-like by adopting specific amino acids at certain interface positions: phe-H50, gln-H6 and val-H4.



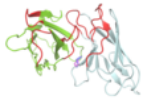
4.4 Discussion

In Chapter 3 we discussed and showed that the orientation between the antibody variable domains, VH and VL, is influential in determining the geometry of the antigen binding site. Here, we have compared the VH-VL orientation to the analogous property in the T-Cell receptor, the Vβ-Vα orientation. The two receptor types have distinct sets of orientations. Using the ABangle methodology, we characterised how the orientation differs. The best descriptor of orientation difference is the change in the HC2 bend angle. This corresponds to a twisting-like change of the VH or Vβ domain with respect to the VL and Vα domains respectively.

The functional implications of variable domain orientation were investigated by analysing its effect on the structure of a TCR-MHC complex. The Vβ-Vα orientation was changed to assume the orientation of a set of antibody structures and a set of other TCR structures. The contacts, clashes and energetics of the transformed complex structures were analysed. Antibody orientations are found to be incompatible with binding to the MHC in a canonical manner. In contrast, assuming orientations of other TCRs disrupted the complex structure far less. The TCR and MHC are thought to have co-evolved over millions of years allowing for their interaction to be optimised [Garcia & Adams, 2005]. Variable domain orientation may be one example of a structural property that the TCR has evolved to enhance its ability to recognise the MHC. Alternatively, such properties may not be encoded on the genome but instead selected for during thymic education of T-cells [Collins & Riddle, 2008]. Therefore, wider ranges of Vβ-Vα orientations may be possible but only particular conformations are chosen as they are structurally compatible with MHC and co-receptor interactions [Rangarajan & Mariuzza, 2014].

Despite the apparent steric restriction imposed by their variable domain orientations, it is possible to engineer antibodies that bind to the MHC. The small number of structures of TCR-like antibodies bound to an MHC were examined and their docking angle calculated. Only one of the three cases, 3cvh, binds in the canonical TCR binding mode. This antibody has a VH-VL orientation that is similar to the TCRs. The other TCR-like antibodies also bind in a diagonal mode but use their VH domains in the same way as a TCR uses its Vα domain. Reversing the equivalence of the domains allows them to overcome the steric restrictions imposed by the variable domain orientation. The current number of available structures is small so it is not possible to make statistically robust conclusions. As more structures become available how TCR-like antibodies bind the MHC might be better understood. To understand how an antibody may be engineered to bind specifically to the MHC in a canonical manner, we examined the factors that cause antibodies and TCRs to be distinct in the HC2 angle.

TCRs tend to have longer CDR-α3 loops than antibodies have CDR-L3 loops. The CDR-α3 packs into the domain-domain interface and acts to open up the Vβ-Vα orientation. However, antibody structures with similar longer CDR-L3 loops are generally not observed to share the TCR orientation. The length of the loop alone is therefore not predictive of the variable domain orientation. However, a TCR-like orientation is found in antibodies with particular interface residues. The most influential position we identified was at IMGT position 50 on the VH domain. Here, antibodies with a phenylalanine instead of the germline leucine tend to have a TCR-like orientation. The TCR itself predominantly has a leucine at this position. Thus in TCRs the interface is twisted open due to packing a longer CDR-α3 in the interface but in antibodies the same effect is achieved by the incorporation of a bulky residue,



phenylalanine, in the interface. Along with IMGT 50 on the VH domain, additional candidate positions were also identified that could be mutated to increase the similarity of the VH-VL interface to the Vβ-Vα interface. Together, these positions provide promising targets for the rational engineering of antibodies specific to MHCs.

The analysis in both this and the previous chapter suggests that variable domain orientation is influenced by certain recognisable features. In the next chapter we extend our investigation into what determines VH-VL orientation and how this information might be used in a predictive capacity.

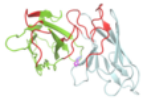
Chapter 5

Determinants of VH-VL orientation

5.1 Introduction

In the previous chapters we have shown how the orientation between the VH and VL domains of an antibody can be characterised in an absolute sense. The ABangle method has then been used to describe the differences between sets of antibody structures and between antigen receptor types. In this chapter we turn our attention to investigating what determines VH-VL orientation and how we can build a set of rules for its prediction at "low" sequence identity.

As discussed in Chapter 1, determining the VH-VL orientation is an important step when modelling an antibody structure [Kuroda *et al.*, 2012; Almagro *et al.*, 2011, 2014]. Not only does it affect the accuracy of the structure directly, but may also influence how well other parts of the structure are predicted. For example, the structure of the CDR-H3 is often influenced by contacts with the rest of the antibody [Choi & Deane, 2010; Almagro *et al.*, 2014; Messih *et al.*, 2014]. These contacts are in part determined by the relative orientation of the two domains.



5. Determinants of VH-VL orientation

Early modelling protocols did not explicitly take into account the VH-VL orientation as an independent feature and simply used the same template for VH, VL and domain orientation [Whitelegg & Rees, 2000]. One of the first methods to include orientation prediction was the Prediction of ImmunoGlobulin Structure (PIGS) webserver [Marcatili *et al.*, 2008]. Here they allow a user to choose a different template for the VH-VL orientation than for the individual domains. This template is the one with the highest sequence similarity to the target over a set of interface residues that are thought to be the most influential for determining the orientation.

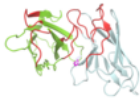
In the recent antibody modelling assessments [Almagro *et al.*, 2011, 2014; Teplyakov *et al.*, 2014], PIGS has performed well compared to other methods with respect to variable domain orientation. A different approach was taken by Abhinandan and Martin [Abhinandan & Martin, 2010]. As discussed in Chapter 3, the authors defined a VH-VL orientation packing angle and predict this single measure parameter to propose likely templates. Their packing angle predictor is also sequence-based but uses a neural network to convert interface-residue types into a single angle. One advantage of an approach that directly gives the geometry of the system is that the orientation can, in principle, be predicted in a template-free manner. However, Abhinandan and Martin make the assumption that VH-VL orientation can be described by a single angle. We showed in Chapter 3 that this is not the case.

In contrast to these sequence-based methods, energetics has been used to select variable domain orientation templates. Narayanan *et al.* [2009] found that the native domain orientation corresponded to a local energy minima. The authors used a physics-based energy function to choose the most suitable template for domain orientation. They found that the protocol worked significantly better if the CDR-H3 was in its native conformation suggesting the orientation is not simply determined by

framework residues. Whilst the orientation is predicted to reasonable accuracy, this protocol relies on having good models for both domains including the structure of the most challenging part, CDR-H3. However, it does demonstrate the principle that the VH-VL orientation observed in crystal structures is related to the energetics of the biological molecule.

Arguably the most complex prediction method currently in use is Rosetta Antibody [Sivasubramanian *et al.*, 2009]. Here, a hybrid method using sequence and energetics predicts the packing in tandem with the rest of the structure. An initial pose is assumed taken from a template with high sequence identity. During a refinement process, different orientations are sampled in a Monte-Carlo fashion using the Rosetta Energy [Simons *et al.*, 1999] as an objective function. A potential advantage of this repacking method is the ability to sample better VH-VL orientations and CDR-H3 conformations that may not have been scored highly using the initial predictions of each [Almagro *et al.*, 2014]. A considerable difference between the energy methods and pure sequence-based methods are their relative computational running times. For instance PIGs runs in minutes or less, whilst Rosetta Antibody has a running time of hours.

In the previous chapters we found that variable domain orientation is a conserved feature between sequence-identical antibody structures, thus as with most homology modelling situations, the higher the sequence similarity between a known structure and the target sequence, the more likely that the structure will provide a good template. As the sequence identity between two proteins reduces below a certain threshold, it becomes uncertain whether they will share similar structures. This threshold is commonly referred to as the twilight zone [Dolittle, 1989; Rost, 1999]. Above this threshold, predicting the structure of a protein becomes relatively simple. However,



the level at which two proteins are deemed to be structurally similar and a prediction accurate is not uniquely defined. For example one may call proteins structurally similar if they share the same major secondary structure arrangements and topological connections (fold) [Lo Conte *et al.*, 2000]. In contrast, for structures of proteins of the same type, structural similarity may be defined to atomic resolution. Thus, the threshold is dependent on application.

One aim of this Chapter is to define where the twilight zone is for predicting the VH-VL orientation in antibodies. That is, at what template-target sequence identity threshold does prediction become trivial and at what level of sequence similarity should prediction methods focus their efforts?

In this Chapter we assess how accurately a modelling protocol should aim to predict the VH-VL orientation and how likely a structure will be a good template given it has a certain level of sequence identity to the target. We investigate different properties that may influence orientation and therefore help in its prediction. Finally, we discuss the use of the ABangles in a feature-based predictor for VH-VL orientation.

5.2 Methods

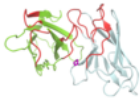
5.2.1 Dataset

All structures from SAbDab [Dunbar *et al.*, 2014b] were extracted in January 2014 and filtered for those that had paired VH and VL domains and a resolution of less than 3Å. This resulted in a redundant set of 1895 Fv structures. The redundant dataset was then filtered at a sequence identity cut-off of 99% using CD-hit [Li & Godzik, 2006] to form a non-redundant set of 678 structures. This cut-off removed identical structures and reduced the median highest sequence identity of an Fv to another structure to 85%. This set was then randomly divided into a training set of 509 structures and a testing set of 169 structures (75% and 25% of structures respectively).

5.2.2 Residue and strand numbering

The VH and VL domains have similar β-sandwich structures. Each β-strand in the both domains is annotated according to the nomenclature discussed in Section 1.3.2.1 and shown in Figure 1.13.

As with the rest of this thesis, the Chothia numbering scheme was use to annotate structurally equivalent positions in the antibody. However, as has been noted by other authors [Honegger & Plückthun, 2001; Lefranc *et al.*, 2003; North *et al.*, 2011] the Chothia numbering has limitations in the way it numbers the CDR-H3 loop. The scheme labels insertions in the loop sequentially instead of symmetrically about the anchors. Figure 5.1 demonstrates a problem with using the Chothia numbering scheme for annotating the CDR-H3 loop. Numbering from the anchor residues means



that more structurally equivalent positions are annotated identically. For instance, when Chothia numbering is used, the third to last residue in the loop is given a different annotation depending on the length of the CDR (e.g. 100A, 100B, 100C...). If a loop is numbered from the anchors, this "-3rd" position is consistently annotated. Such a scheme appears to be most effective for aligning positions in the C terminus region of the loop. For this chapter we follow an anchor-based numbering scheme similar to that of North *et al.* [2011] and apply it to the third hypervariable loop on both domains. The Chothia-Anchor numbering equivalence for CDR-H3 is shown in Table 5.1.

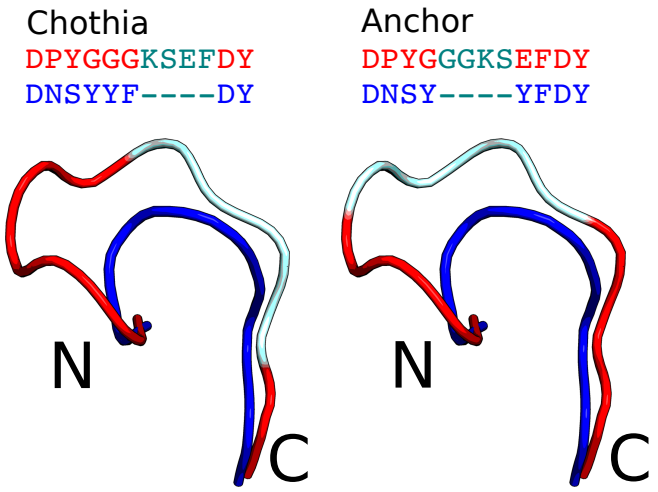
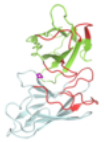


Figure 5.1: Comparing Chothia numbering of CDR-H3 with numbering from anchor positions. Two CDR-H3 loops of different lengths are shown. In red are the aligned positions according to the Chothia numbering scheme (left) and the anchor numbering scheme (right). With the anchor numbering scheme we see that the aligned positions are in closer proximity than the Chothia scheme. This is particularly the case towards the C terminus and it is this region that makes more contacts with the VL domain. Performing the anchor renumbering helps to ensure equivalent structural positions are considered in the orientation prediction protocol.

Chothia	95	96	97	98	99	100	A	B	C	D	E	F	G	H	I	J	K	L	101	102
Length																				
1	0																			
2	0																			-1
3	0	1																		-1
4	0	1																	-2	-1
5	0	1	2																-2	-1
6	0	1	2	-3															-2	-1
7	0	1	2	-4	-3														-2	-1
8	0	1	2	3	-4	-3													-2	-1
9	0	1	2	3	-5	-4	-3												-2	-1
10	0	1	2	3	4	-5	-4	-3											-2	-1
11	0	1	2	3	4	-6	-5	-4	-3										-2	-1
12	0	1	2	3	4	5	-6	-5	-4	-3									-2	-1
13	0	1	2	3	4	5	-7	-6	-5	-4	-3								-2	-1
14	0	1	2	3	4	5	6	-7	-6	-5	-4	-3							-2	-1
15	0	1	2	3	4	5	6	-8	-7	-6	-5	-4	-3						-2	-1
16	0	1	2	3	4	5	6	7	-8	-7	-6	-5	-4	-3					-2	-1
17	0	1	2	3	4	5	6	7	-9	-8	-7	-6	-5	-4	-3				-2	-1
18	0	1	2	3	4	5	6	7	8	-9	-8	-7	-6	-5	-4	-3			-2	-1
19	0	1	2	3	4	5	6	7	8	9	-9	-8	-7	-6	-5	-4	-3		-2	-1
20	0	1	2	3	4	5	6	7	8	9	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1

Table 5.1: The equivalence between Chothia numbering and numbering from the anchor residues for CDR-H3. insertions in the Chothia scheme are represented as A, B, C etc. This table demonstrates how the Chothia annotation of, for example, the third from last loop position (-3), changes with loop length.



5.2.3 Contact positions

Residues found in the interface between VH and VL are likely to constrain how the two domains can associate and determine their relative orientation. Using the non-redundant training set we calculated the frequency at which each residue position makes a contact with positions in the opposing domain. Here, two residues were defined as being in contact if any of their atoms were within 5Å of each other. The frequency at which each contact pair is observed is presented in Appendix D Table D.1.

All contacts that occur in 10% or more of the training set structures were retained and the positions involved referred to as contact residue positions herein. Positions in the framework regions that make more than one contact with residues in the other domain (framework and CDRs), are listed in Table 5.2. These are referred to prolific contact residue positions.

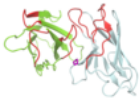
The framework positions that make more than one contact with either framework or CDR residue positions on the other domain, are listed in Table 5.2. These are referred to as prolific contact residue positions.

5.2.4 Germline pairing and the effect of somatic hyper-mutations

As described in Section 1.2.3.1 antibody variable domains have a germline sequence formed from the combination of v, d and j genes. During affinity maturation, the sequence of the receptor will undergo somatic hyper-mutation, changing it away from that of the original germline sequence. Regions of the variable domains that have an enrichment of positively selected SHMs correlate well with CDR characterisations [Burkovitz *et al.*, 2013]. However, SHMs are not limited to these regions and can

VH residue	VL residue contacts
47	$l3_{-4}$ $l3_{-3}$ $l3_{-2}$ $l3_{-1}$ $l3_{0}$ $l3_{2}$ 98
103	36 43 44 45 46 98
45	36 38 44 87 98 99
91	38 42 43 44
44	87 99 100
39	38 44 87
61	$l3_{-4}$ $l3_{-3}$ 1
43	85 87 100
105	41 42 43
60	$l3_{-3}$ $l3_{-2}$ 1
59	$l3_{-4}$ $l3_{-3}$
58	$l3_{-4}$ $l3_{-3}$
50	$l3_{-4}$ $l3_{-2}$
VL residue	**VH residue contacts**
98	$h3_{-3}$ 37 45 46 47 103
46	$h3_{-5}$ $h3_{-4}$ $h3_{-3}$ $h3_{-2}$ $h3_{3}$ 103
36	$h3_{-5}$ $h3_{-4}$ $h3_{-3}$ $h3_{-2}$ 45 103
49	$h3_{-6}$ $h3_{-5}$ $h3_{-4}$ $h3_{3}$ $h3_{5}$
43	91 103 104 105 106
50	$h3_{-7}$ $h3_{-6}$ $h3_{-5}$ $h3_{-4}$ $h3_{5}$
87	39 43 44 45
34	$h3_{-6}$ $h3_{-5}$ $h3_{-4}$ $h3_{-3}$
44	39 45 91 103
32	$h3_{-7}$ $h3_{-6}$ $h3_{-5}$
38	39 45 91
55	$h3_{-4}$ $h3_{-2}$ $h3_{-1}$
1	60 61
100	43 44
42	91 105
99	44 45

Table 5.2: Residue positions in the framework (left column) of the VH and VL domains that make more than one contact with the other domain in over 10% of structures in the training set. The framework or CDR positions that they make contact with are listed in the right hand column. Positions $h3_n$ or $l3_n$ refer to the n^{th} positions in the Chothia CDR-H3 and CDR-L3 loop respectively.



occur away from the binding site.

In an evolutionary process a mutation may be selected for if it increases the fitness of a system. However, it may also remain if it happens to occur, with neutral effect, in parallel with another fitness-improving mutation. SHMs at non-antigen-interacting positions may therefore be neutral, background mutations or make modifications to the molecule that improve its antigen binding properties in more subtle ways. One example of such a change could be modification of the VH-VL orientation [Fera *et al.*, 2014]. Here, we investigate whether SHMs at certain non-antigen-interacting positions are more influential than others for modifying the angles from what might be expected for the germline sequence. If so, the reasoning for positively selecting such mutations during affinity maturation may be deconvoluted from the effect of SHMs at residues that make direct antigen contacts.

5.2.4.1 Identifying somatic hyper-mutations

The germline v and j genes were assigned to each structure using the method described in Section 2.2.5.2. Any position where the residue was different in the structure's sequence to the germline sequence was determined to be an SHM. As the germline sequence of CDR-H3 is ambiguous to determine, we do not attempt to assign SHMs in the "diverse" encoded region of this loop (typically between Chothia positions 96-101 exclusive). For each structure the position, germline residue and mutated residue was recorded.

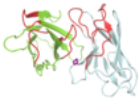
5.2.4.2 Assessment of mutation magnitude

Substitutions from one amino-acid to another in proteins are not equally likely. In general, the more dissimilar two residues are the less likely a mutation between them

is likely to occur. We classify the magnitude of a mutation based on two metrics for measuring amino-acid dissimilarity, BLOSUM62 [Henikoff & Henikoff, 1992] and Miyata [Miyata *et al.*, 1979]. The BLOSUM62 matrix [Henikoff & Henikoff, 1992] gives a log-odds score that represents how likely it is that one residue mutates to another in an alignment of protein sequences. It therefore characterises how related each of the residues are in the context of protein evolution. In contrast, the Miyata dissimilarity matrix [Miyata *et al.*, 1979] is primarily based on the physico-chemical relationship of two residues. However, it should be noted that the physical properties used to parameterise the Miyata matrix were chosen based on their agreement with mutation propensities observed in protein evolution[Yampolsky & Stoltzfus, 2005]. In both cases, a large positive score is equivalent to two residues being similar whilst a large negative score means two residues are dissimilar. We use these two scores to assess the magnitude of a change of one residue to another in an SHM.

5.2.4.3 Determining angle changes

Given the small number of examples, studying structural changes between true germline antibodies and those that have undergone somatic hyper-mutation is not currently possible. Instead, we compared the ABangles within sets of structures with the same germline variable gene subgroup. Sets of structures with three or more representatives were identified. Only one chain was considered at a time (i.e. the other domain may have a different germline). Within each set we calculated the difference in ABangles between all those with the same mutation (e.g. leu to phe at position H45) and all those that retain the germline residue (leu). A list of these angle differences and the magnitude of the mutation (see Section 5.2.4.2) was recorded for each position in the variable domains.



5.2.4.4 τ_b statistic

The Kendall τ_b statistic measures the rank correlation between two variables [Kendall, 1938]. It is defined as:

$$\tau_b = \frac{N_c - N_d}{\sqrt{N_1} \times \sqrt{N_2}} \tag{5.1}$$

where N_c is the number of concordant pairs and N_d is the number of discordant pairs. In the τ_b statistic a weighting is given for tied ranks. N_1 is the number of pairs in the first variable that are not tied. N_2 is the number of pairs in the second variable that are not tied. It is bound between -1 and 1. A value of 0 means that no correlation is found between the ranks of the two variables.

We calculated the τ_b value for the correlation between the magnitude of the angle changes and the magnitude of the SHM mutations at each position in the variable domains. This was done for each of the six ABangles. If mutations at a position are influential for changing orientation we might expect that the more different the residues are the more the angle will change. In this case we would expect the τ_b to have a negative value.

5.2.5 Structural variation of the Heavy and Light framework interface loops

In Chothia *et al.* [1985]'s original description of the packing of the VH and VL domains (see Section 1.3.3.3), the authors reported that the strands at the edge (C' strands in Figure 1.3.2.1) of the interface β-sheets twist and pack residues into the domain-domain interface. More recently Vargas-Madrazo & Paz-García [2003] analysed 23

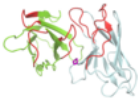
crystal structures and found that several residues in this region were consistently involved in inter-domain contacts (39, 45 and 47 on the VH domain and positions 34, 36, 38 and 44 on the VL domain). Our analysis of the 509 structures in our training set finds that many of the prolific contact positions (Table 5.2) are either in or with the framework loop that connects the C and C' beta strands in both VH and VL (Figure 5.2). As discussed in Chapter 3, the identity of certain residues in these regions have been found to influence VH-VL orientation [Dunbar *et al.*, 2014b; Abhinandan & Martin, 2010; Chailyan *et al.*, 2011]. However, to our knowledge, the explicit structural properties of these interface framework loops have not been investigated previously. We therefore decided to examine their structural conformations in more detail.

5.2.5.1 Definition of the Heavy and Light framework interface loops

The Heavy interface framework loop (Hifw-loop) is defined as Chothia positions H39-H47. The Light interface framework loop (Lifw-loop) is defined as Chothia positions L38-L46. Both loop locations are shown in Figure 5.2.

5.2.5.2 Loop clustering

The Hifw-loop from each of the structures in the training set were excised and aligned to the Fv structure 2ai0_KO. For the Hifw-loop alignments the Cα, Cβ and carbonyl oxygen coordinates of the residues at positions 37, 38, 47 and 48 were used. These positions act as anchors for the loops and form parts of structurally conserved β strands. Pairwise RMSDs were calculated using the Cα, Cβ and carbonyl oxygen coordinates over the loop length (positions 39-46). Using the carbonyl oxygen in the calculation allows the direction of the side-chain in each of the eight residues to



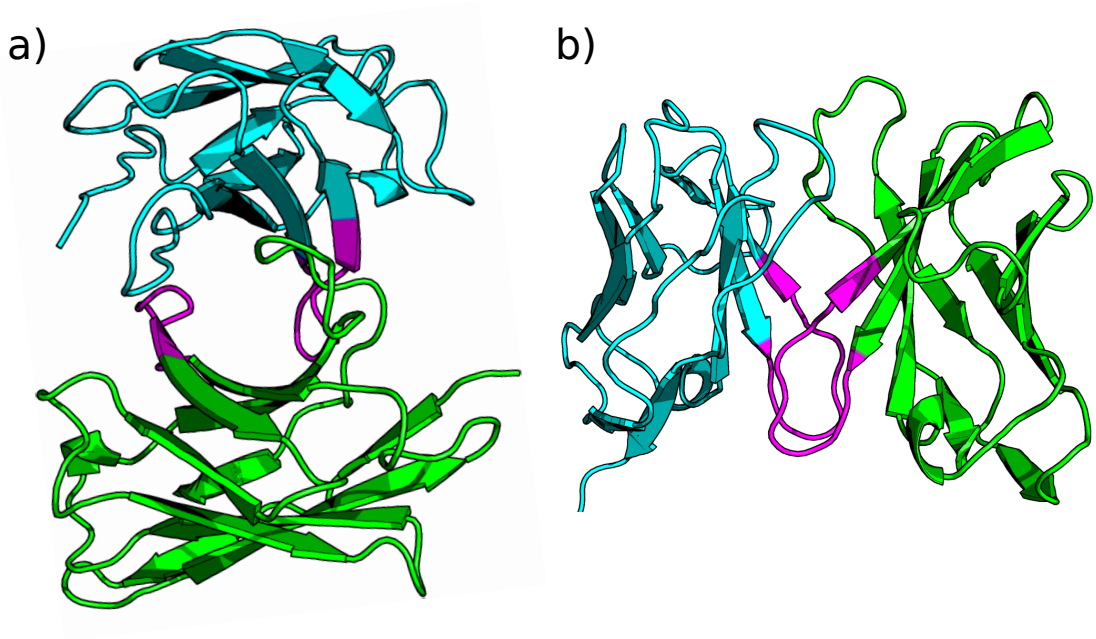


Figure 5.2: The location of the interface framework loops (magenta) on the VH (green) and VL (cyan) domains. a) Looking down from the antigen-binding site b) Side view. Chothia *et al.* [1985]'s original description of the packing of the VH and VL domains (see Section 1.3.3.3) described how the edge strands of the interface β-sheets of both domains twist, causing residues to pack into the interface. We find many of the prolific contact positions (Table 5.2) are located in these regions.

be represented. Loops were then clustered using the UPGMA algorithm [Sokal & Michener, 1958] and a cut-off of 1.5Å. Therefore, any two structures in the same cluster will not have an RMSD of more than 1.5Å. An analogous procedure was performed for the Lifw-loop using positions 36, 37, 46 and 47 for the anchors and positions 38-45 as the loop.

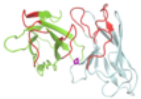
5.2.5.3 Canonical assignment

To identify distinct canonical structures from the clustering we examined the distributions of Phi-Psi angles for each of the loop residues. Only positions 41-44 of the Hifw-loop had dihedral angles that discriminated between RMSD-clusters. Similarly, the structurally equivalent positions 40-43 of the Lifw-loop were also the only ones to differentiate between RMSD-clusters.

For the Hifw-loop, the RMSD-clusters were each manually classified into groups according to the dihedral angles of residues 41-44. To do this a dihedral string was assigned to each cluster. As an example, Figure 5.3 shows the distribution of Phi and Psi angles for the two largest clusters of Hifw-loop structures. We annotate cluster one with the dihedral string A-B-A-A and cluster two with the dihedral string B-A-C-A. Each letter of the string corresponds to a region of the dihedral space of residues H41, H42, H43 and H44 respectively.

The distributions of dihedral angles for all RMSD cluster and the annotations of regions in the Phi-Psi space can be found in Appendix D, Figure D.1. All RMSD-clusters that shared the same dihedral string were joined to form a canonical class for the loop. Table 5.3 shows the mapping between the canonical class, dihedral string and corresponding RMSD-cluster with the largest number of representatives.

The same procedure was applied to the Lifw-loop. Similarly, the distributions of



dihedral angles for each RMSD cluster and the annotations of regions of the Phi-Psi space for the Lifw-loop structures are shown in Appendix D, Figure D.2. The mapping between Lifw-loop canonical class, dihedral string and corresponding RMSD-cluster with the largest number of representatives is shown in Table 5.4.

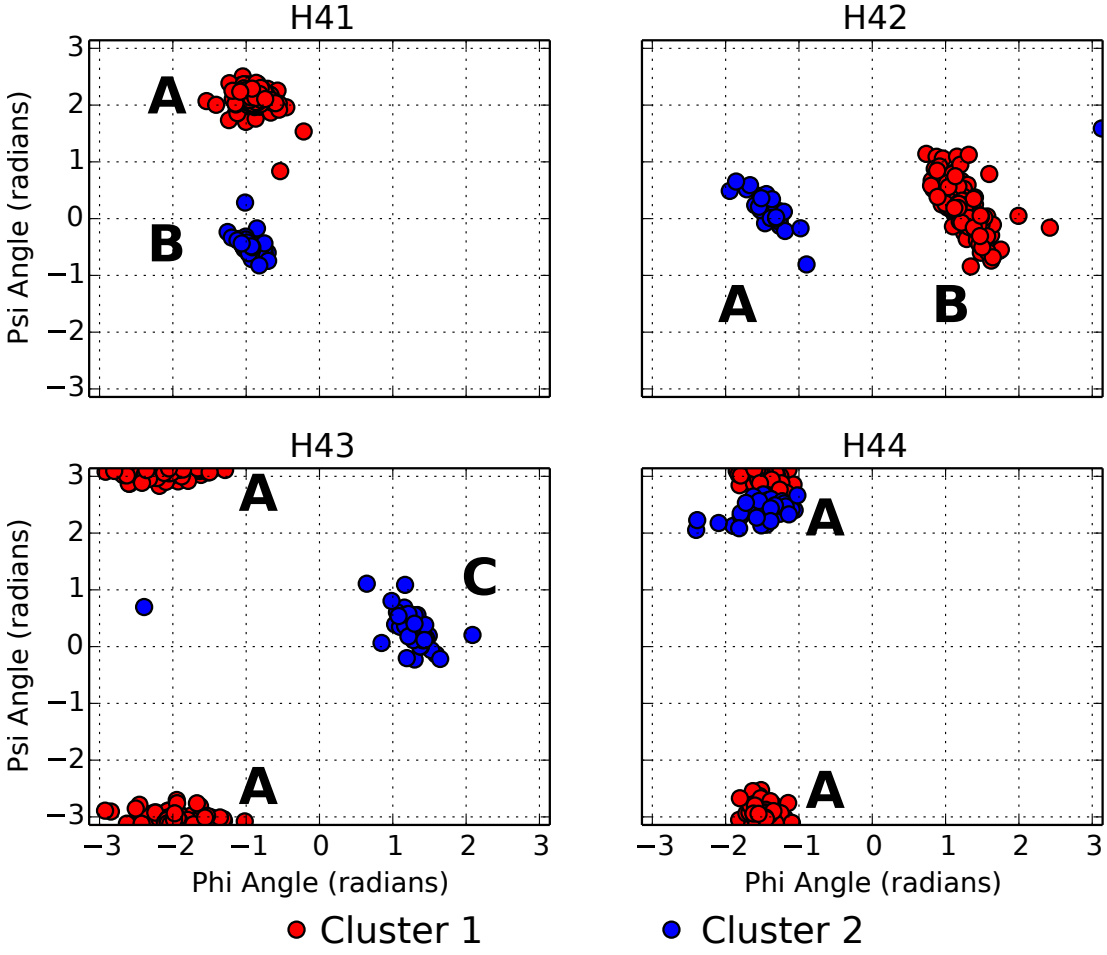


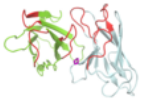
Figure 5.3: Hifw-loop structures were clustered according to their RMSD using the method described in Section 5.2.5. Here we show the dihedral angles for the two largest clusters of loops over the four residues where the clusters show different conformations. For each residue we manually assign a dihedral string to each cluster based on where the majority of its structures are positioned on the Ramachandran plot. For example cluster 1 has a string A-B-A-A whilst cluster 2 has a string B-A-C-A. The Ramachandran plots for all of the Hifw-loop clusters are shown in Appendix D Figure D.1, whilst the corresponding images for the Lifw-loop are shown in Appendix D Figure D.2. Those clusters with the same dihedral string are joined to form the canonical shapes described in Tables 5.3 and 5.4.

Canonical ID	H41	H42	H43	H44	Largest RMSD cluster
1	A	B	A	A	1
2	B	A	C	A	2
3	A	B	C	A	6
4	B	A	B	B	7
5	B	A	A	A	11
6	C	B	D	A	12
7	A	B	B	B	13

Table 5.3: The unique dihedral strings identified for the Hifw-loop. Each letter corresponds to a region of the Ramachandran plot shown in Appendix D Figure D.1. Each string corresponds to a different shape of loop. These shapes are the Hifw-loop canonical structures. The corresponding RMSD-cluster with the largest number of representatives is listed on the right.

Canonical ID	L40	L41	L42	L43	Largest RMSD cluster
1	A	B	A	A	1
1 (sub)	B	A	A	A	1
2	B	A	C	A	2
3	A	B	B	A	4

Table 5.4: The unique dihedral strings identified for the Lifw-loop. Each letter corresponds to a region of the Ramachandran plot shown in Appendix D Figure D.2. Each string corresponds to a different shape of loop. These shapes are given the Lifw-loop canonical structures. The corresponding RMSD-cluster with the largest number of representatives is listed on the right.



5.2.5.4 Classification by sequence

For CDR canonical classes, the shape of the loop can often be predicted from sequence e.g. North *et al.* [2011]. We attempted to develop an equivalent classification procedure for the classes of the framework loops we identified.

To identify sequence traits of the Hifw-loop shapes, we built a decision tree using the eight residue loop sequence using the R package rpart [Therneau *et al.*, 2014]. We also included the amino-acid identity of the positions on the light chain that make the most contacts with the loop (i.e. 36, 38, 44, 85, 87, 98, 99 and 100). Similarly a classification tree was built for the Lifw-loop shapes.

5.2.6 A feature based predictor

Previous explicit prediction methods of the VH-VL orientation from sequence have relied on machine-learning methods [Abhinandan & Martin, 2010; Chailyan *et al.*, 2011]. Indeed, in Chapter 3 we turned to a random forest approach to extract sequence patterns that are influential for each of the ABangles. Here, we attempt to predict the orientation by selecting known structures in SAbDab that will provide good templates for a given antibody sequence. The problem becomes significantly easier when a high sequence identity template is available. Therefore, we focus on developing a method for prediction in "low" sequence identity cases. The method for selecting a template for orientation is outlined in Figure 5.4 and in detail below.

Initially, the 20 most sequence similar structures in the training set to the target sequence are selected as potential templates. Only those structures with a sequence identity of less than a given threshold are available for selection for training. All positions in the variable region are used in this calculation and the CDR-3 regions

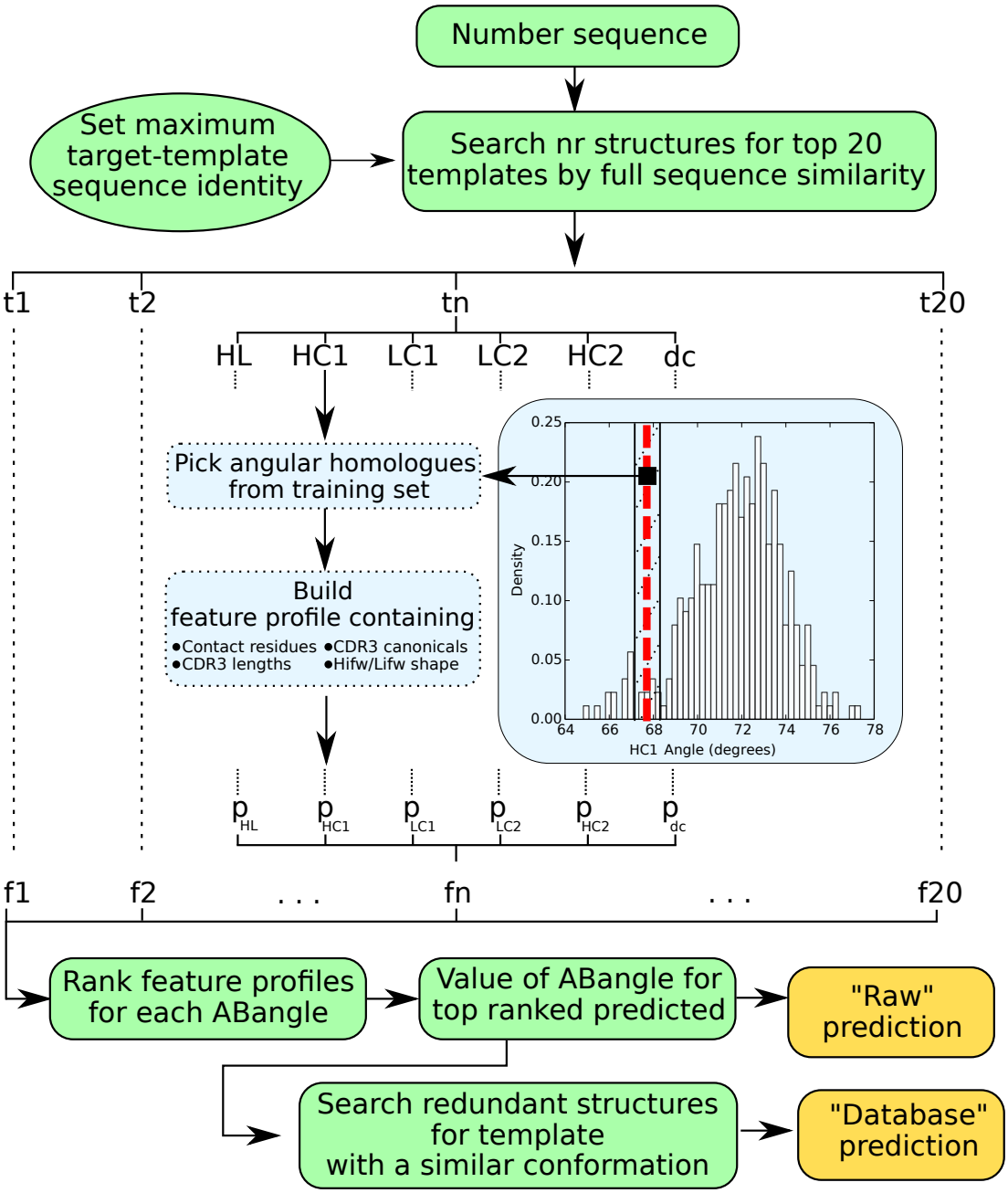


Figure 5.4: The feature based prediction method as described in section 5.2.6.

renumbered according to the scheme described in Section 5.2.2. For each potential template, a feature profile is built as described in Section 5.2.6.1 for each of the ABangles. A pairwise comparison is made between the templates using a profile scoring scheme described in Section 5.2.6.2.

5.2.6.1 Feature profiles

A feature profile records the relative frequencies of values of a given property. A commonly used example in protein informatics are sequence profiles (e.g. for sequence alignment [Edgar, 2004], protein classification [Sonnhammer *et al.*, 1998] or secondary structure prediction [Jones, 1999]). Here, the frequency of occurrence of each amino acid is recorded for a column in a multiple sequence alignment. Some columns, or positions, are informative about the sequences used to build the multiple sequence alignment. For example, sequences from mice may be enriched for one residue type at a position whilst sequences from humans have a preference for another.

Similarly, we can build a profile using structures that have a similar orientation angle to a given template. Thus, we can incorporate information about structurally similar antibodies in order to assess whether a particular template would be suitable to model a target antibody. In other words, if a target sequence shares a similar property to a template *and* to those structures to which the template is structurally similar, the template is more likely to be "good" (i.e. be within 1.5Å orientation RMSD to the target structure). Features do not need to be limited to the residue identities at positions as in a sequence profile. Our profiles contain each contact position (5.2.3). We also include the lengths of CDR-H3 and CDR-L3, the North canonical classes of CDR-H3 and CDR-L3 [North *et al.*, 2011] and the canonical classes of the Hifw-loop and Lifw-loop (5.2.5.3).

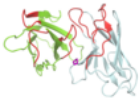
For one of the potential templates (t_n) we build a set of six feature profiles, one for each ABangle. To build, for example, the HC1 profile (p_{HC1}), we choose the 20 other structures from the training set with the most similar HC1 angle to t_n (angle similar structures). The same procedure is performed for each ABangle and for each potential template.

5.2.6.2 Comparing template feature profiles

We test which of two templates is a better match for the target sequence using only features for which the template profiles are considered different. For example, assume we have two templates T_1 and T_2 and target sequence S. At position P, S has residue α. If α is the most frequent residue at position P in both the T_1 and T_2 profiles, both templates score 0. If α is present in only one template, that template scores 1 and the other scores 0. Otherwise, the templates score the frequency at which α occurs in their respective profiles. The final score is taken as the sum over all features. The template with the higher score is taken as a better match for the target antibody. Pairwise comparison of the initial templates as described above gives a re-ranking of the templates using features that are relevant for the angle being considered.

5.2.6.3 Selection of predicted template

For each ABangle we have a top predicted template, but this may not be the same one for all. We choose the template orientation in two ways. The first is to reconstruct the orientation geometry from the angles in each of the top predicted templates. We call this the "raw" prediction. A model of the VH and VL domains is made to assume this orientation even if no real structure exists with these precise angles. The second is to use the redundant structures in SAbDab to identify a template with



the closest orientation to this set of angles. We call this the "database" prediction. This template, whether or not it was in the original set of templates, is used as the prediction orientation. The second protocol aims to ensure that the combination of angles are physically compatible.

5.2.6.4 Assessment methods

As discussed earlier in this chapter, predicting orientation of the domains becomes a relatively easy problem when a template with high sequence identity is available. New methods to predict orientation should therefore demonstrate their performance in a scenario analogous to the "twilight zone" in general homology modelling cases. We ran our prediction method for a range of thresholds of maximum sequence identity allowed to build any part of the system. A leave-one-out cross validation method was used to benchmark the predictor performance at each cutoff.

To assess how well the predictor performs we use a metric to assess how many times the orientation is predicted within an acceptable orientation RMSD (see Section 5.3.1). We compare the protocols' performance to the equivalent performance achieved if the maximum full sequence identity template to the target is used.

5.3 Results

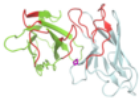
5.3.1 How well should we aim to predict?

When assessing the accuracy of a model, the crystal structure is used as the ground truth. However, the structure itself is also a model with an associated uncertainty. Such errors can be attributed to experimental artefacts and the limitation of representing a dynamic molecule as a static snapshot obtained from a non-biological environment. We therefore examined the available crystal structure data to set a threshold at which to call the orientations of two structures identical.

Figure 5.5 shows the distribution of mean orientation RMSD (see Section 3.2.8) in sets of sequence-identical sets of structures. We find that 85% of sets have an orientation RMSD of below 1 Å, whilst 92% are more similar than 1.5 Å. Given this, we aim to predict the orientation of a given antibody to within 1.5 Å orientation RMSD. A prediction is considered to be good if it is below this threshold to the known crystal structure of the target.

5.3.2 Prevalence of good templates in antibody structure space

In most antibody prediction protocols, the VH-VL orientation is predicted using the template with the highest sequence identity or similarity to the target sequence. We investigated how well this was likely to predict the orientation at different template-target sequence identities. The orientation RMSD was calculated for each pair of structures in the training set and binned by their full sequence identity. Within each bin, the fraction of pairs that were good orientation templates for each other was calculated. Figure 5.6 shows the change of prevalence of good orientation templates



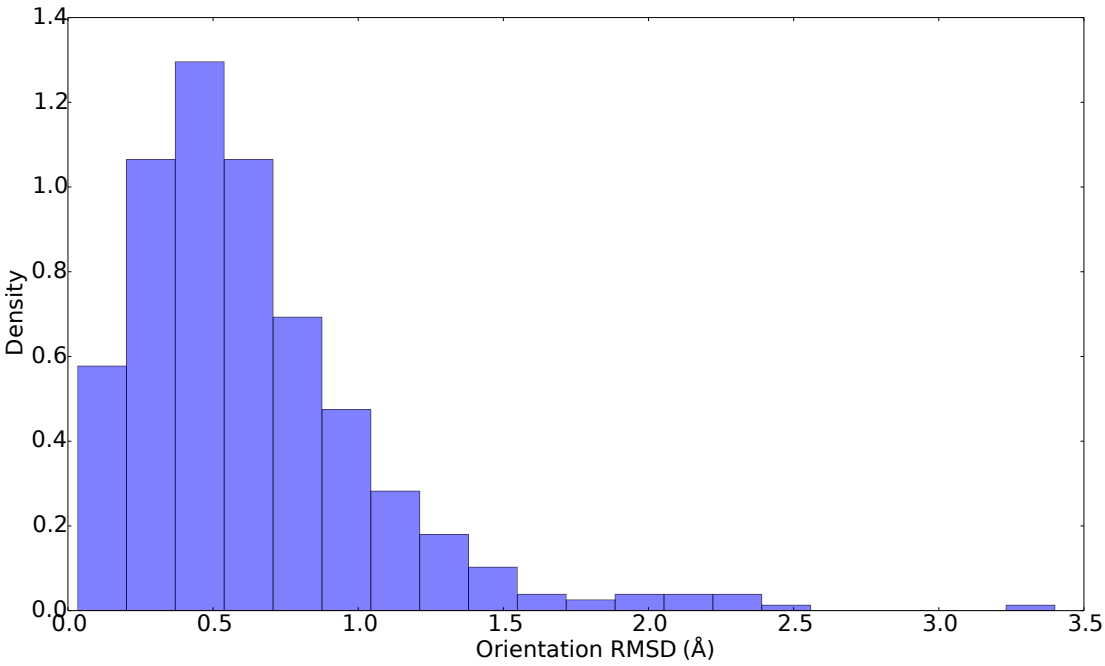


Figure 5.5: The difference in VH-VL orientation between sets of sequence identical structures. The distribution of the mean orientation RMSD for the sets is shown. Sets of structures with less thn 1.5Å account for 92% of the data. We use 1.5Å as a threshold to call the orientation between any two structures identical.

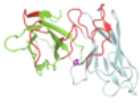
throughout antibody structure space. As might be expected, the more similar the structures, the more likely they are to have the same orientation. However, there is a drop of prevalence of good templates below approximately 85% sequence identity. Below this threshold, if one was to simply use the highest sequence identity template, the orientation would be acceptable less than 50% of the time.

The following sections of this chapter investigate factors that may allow us to enrich the likelihood of selecting a good template given that the maximum template-target sequence identity is below at least 85%.

5.3.3 Does using interface residues enrich the selection of good templates?

Interactions between the VH and VL domains are mediated by residues that make contact in the interface. One might expect that these residues have the most influence on determining the VH-VL orientation. Therefore, choosing templates based on their similarity to the target over interface residue positions may provide a more effective way of picking good orientation templates. For reference, the relationship between the probability of choosing a good template and the interface sequence identity is shown in Appendix D Figure D.3.

To directly compare the expected ability of interface residues to select good templates to when all residues are used, we performed the following procedure. For each full sequence identity bin as described in Section 5.3.2 we counted the number (N) of pairs that involved the same target structure. For this target, the N template structures with the highest interface sequence similarity (using BLOSUM62) to the target structure, were identified. A maximum full sequence identity equal to the upper limit



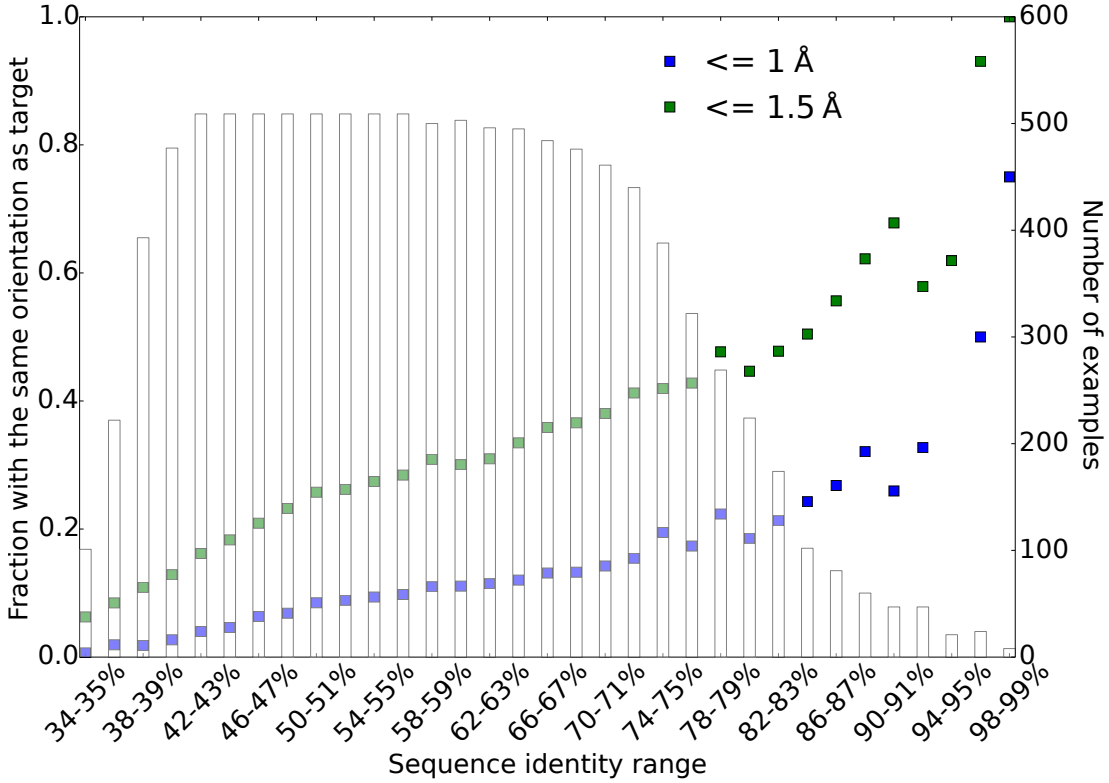


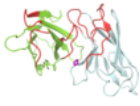
Figure 5.6: The prevalence of good templates for VH-VL orientation through antibody structure space. Pairs of structures were binned into 1% ranges of sequence identity. Orientation RMSDs were also calculated for each pair. A pair is referred to as being good orientation templates for each other if their RMSD is below a threshold value of either 1.5Å. For each sequence identity bin we calculate the mean value of the fraction of good pairings within the bin (left hand axis, shown by dots). This is equivalent to the probability of a structure being a good orientation template given it has a certain sequence identity to the target. The bars (right hand axis) refer to the number of structures that have templates in each range of sequence identity. Above approximately 85% sequence identity, selecting a good template becomes "easy" (greater than 50% chance it will be a good template). However, only 156 out of the 509 structures in our training set have any template above this threshold. The majority of structures only have templates below this threshold where the probability of being a good template drops quickly with sequence identity.

of the bin was imposed when selecting the templates. We calculated the fraction of the N templates that were good (less than 1.5Å orientation RMSD to the target). The mean value for all the structures with a pair in the bin gave the final number for the full sequence identity range.

We expected that this procedure would enrich the probability of picking a good template. Figure 5.7 shows that this is not the case. Using the similarity of interface residues alone is worse at selecting good templates compared with the full sequence identity. The same is seen when the sequence identity of interface residues is used instead of sequence similarity. Further, when we enforce that N can have a maximum value of 20, using interface similarity is even worse at selecting good templates than full sequence identity.

One explanation for this result is that the identity of residues within each domain are not independent. For example, the majority of the variable domain is expressed from a single v gene. Therefore, an interface residue that is important for determining a particular orientation may be correlated with residues that are not on the interface .

Residues on the interface may not be the only influential factors for determining orientation. Other positions, where residues do not make direct contact with the other domain, could influence the shape of the interface and therefore the VH-VL orientation. If these positions are excluded from the set of residues we used to select templates, their contribution will be ignored.



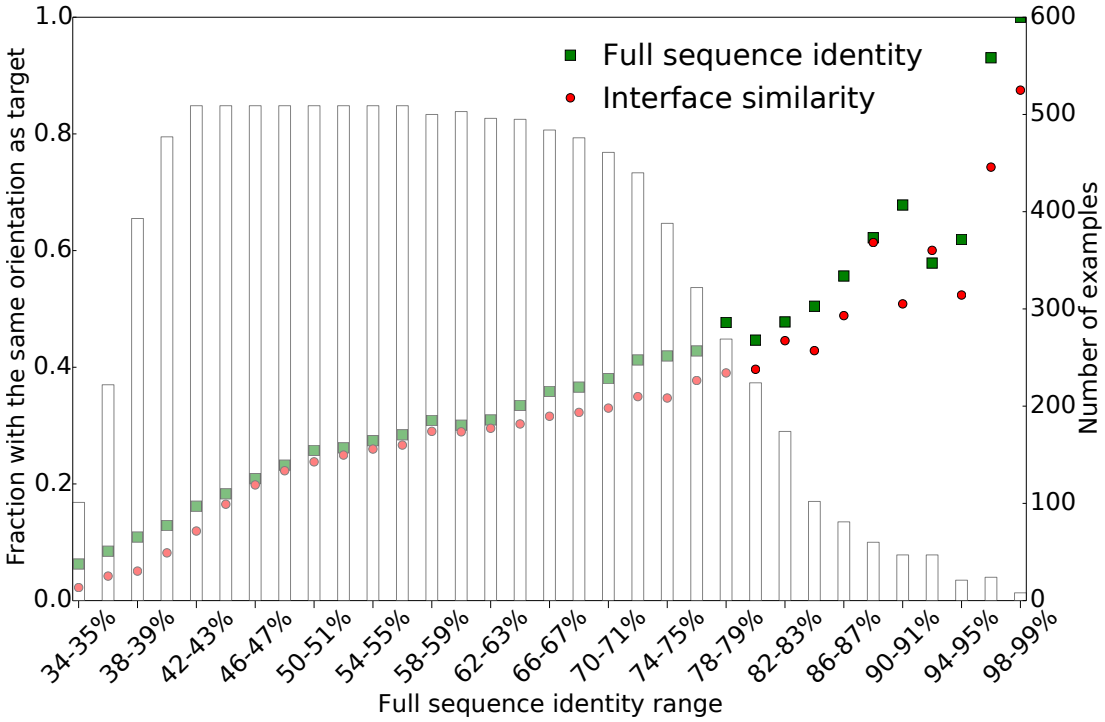


Figure 5.7: The prevalence of good orientation templates (<1.5Å orientation RMSD) when selected by full sequence identity and when selected by interface residue sequence similarity using the procedure described in Section 5.3.3 (left hand axis, shown by dots). The bars (right hand axis) refer to the number of structures that have templates in each range of sequence identity. Using interface residues alone performs worse at selecting good orientation templates than using the full sequence identity. This suggests that examining only interface residues is not a sufficient method of enriching the chance of selecting a good template for antibody modelling.

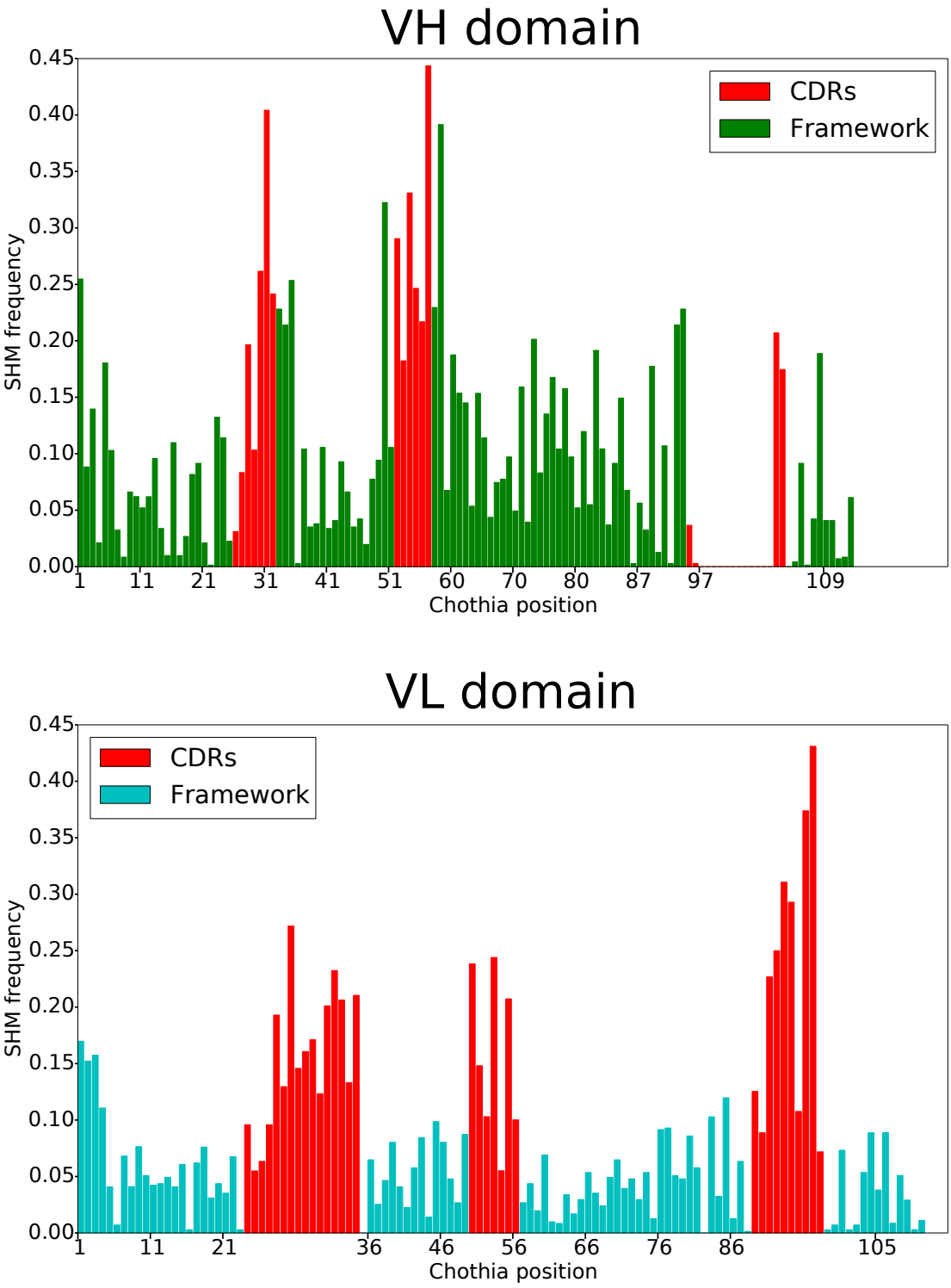


Figure 5.8: The frequency at which any mutation from the germline sequence occurs at variable domain positions. SHMs were not calculated for Chothia positions VH 96-101 exclusive as germline sequence in this region is ambiguous. The CDRs overlap with regions that have a high frequency of SHMs.

5.3.4 The role of SHMs in non-antigen binding regions for changing orientation

SHMs are thought to act to improve an antibody's interaction with its antigen in the maturation process [Tonegawa, 1983]. Figure 5.8 shows the frequency at which SHMs occur at each Chothia position in the variable domains of our non-redundant dataset. As might be expected, the regions that normally make the most contacts with the antigen have a high mutation frequency. SHMs at these positions may have improved binding properties by making direct physico-chemical changes to antigen contacts and therefore be selected for during B-Cell maturation. However, the Chothia CDRs do not completely encompass the region of high SHM frequency. We therefore take a more relaxed definition of an antigen binding region (ABR) and use the union of the Chothia [Chothia & Lesk, 1987; Al-Lazikani *et al.*, 1997], AbM [Abhinandan & Martin, 2008], IMGT [Lefranc *et al.*, 2003] and Contact [MacCallum *et al.*, 1996] characterisation of CDRs (see Section 1.3.2.2 for a description of each) as summarised in Table 5.5. The Kabat characterisation of CDRs [Wu & Kabat, 1970] is not used in the union. It was derived by analysing sequences alone without consideration of each position's ability to form antigen contacts. If included it would extend the end position of H2 to 65.

Although not as frequent, SHMs also occur in regions that cannot make direct contact with the antigen [Ramirez-benitez & Almagro, 2001; Raghunathan *et al.*, 2012]. These could have been selected for by mutating in parallel with another position that improved affinity through a direct antigen contact. Alternatively, they could have independently improved binding properties by making indirect structural changes [Ramirez-benitez & Almagro, 2001; Raghunathan *et al.*, 2012; Burkovitz

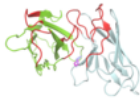
ABR	Start	End
H1	26	35
H2	47	58
H3	95	102
L1	24	36
L2	46	56
L3	89	97

Table 5.5: Antigen binding regions (ABRs) are defined as the union of the Chothia [Chothia & Lesk, 1987; Al-Lazikani *et al.*, 1997], AbM [Abhinandan & Martin, 2008], IMGT [Lefranc *et al.*, 2003] and Contact [MacCallum *et al.*, 1996] characterisation of CDRs. Here, the starting and ending Chothia positions are shown for each of the six ABRs.

et al., 2013]. We hypothesise that modifying the VH-VL orientation may be one such change and ask whether the presence of certain SHMs should influence the choice of particular templates in a modelling protocol. For example, if an antibody has a particular SHM should a template that also has this mutation be chosen despite another with a higher overall sequence identity being available?

We therefore, examined the SHM frequency in non-ABRs of the variable domains. Mutations to residues that make inter-domain contacts might be expected to affect variable domain orientation more than residues that do not make inter-domain contacts. For 20% of the inter-domain contact positions no SHM is found. In contrast, only 9% of the non-contact positions are free of SHMs. Therefore, within the germline, a higher proportion of the VH-VL interface region must remain conserved compared to the rest of the non-ABRs of the domains.

Figure 5.9 shows the frequency distributions of non-ABR positions that *do* allow SHMs to occur. We find that the mean SHM frequency to be slightly higher for inter-domain contact positions (8.3%) than non-inter-domain contact positions (7.0%).



This suggests that where mutations do occur, they are more likely to be selected for at interface positions than non-interface positions. However, the difference is not found to be statistically significant by a KS-test. A similar result was found by Burkovitz *et al.* [2013] who compared the propensity of SHMs in different parts of the Fv region of antibody structures. The authors calculated that the VH-VL interface had a higher propensity for SHMs than other residues in the structure that were not in ABRs. Additionally, they found that predicted antibody-antigen binding energies were affected by mutations at the VH-VL interface more than those in other regions of the antibody not in the ABRs.

Given these results, we investigated the influence that SHMs have on the VH-VL orientation using the procedure described in Section 5.2.4. The aim was to identify positions where SHMs changed the orientation away from other structures that had the germline residue present. We assumed that the larger the "size" of the mutation, the more an SHM would affect the orientation if a position is influential for an ABangle measure. The τ_b value defined in Section 5.2.4.4 quantifies this relationship.

We find no region to be enriched with mutations that correlate with orientation differences in any of the angles as measured by the τ_b statistic. The five positions in non antibody binding regions with the best correlation between mutation magnitude and orientation change are shown in Table 5.6. The positions returned when BLOSUM62 or Miyata dissimilarities are used to quantify mutation magnitude are largely similar. There appears to be no preference for interface or non-interface residues. In addition, these value are not significantly more correlated than others in the domains. We therefore do not find any conclusive evidence for SHMs at particular positions to effect orientation. However, given the limitations in the available data this does not mean that there is none.

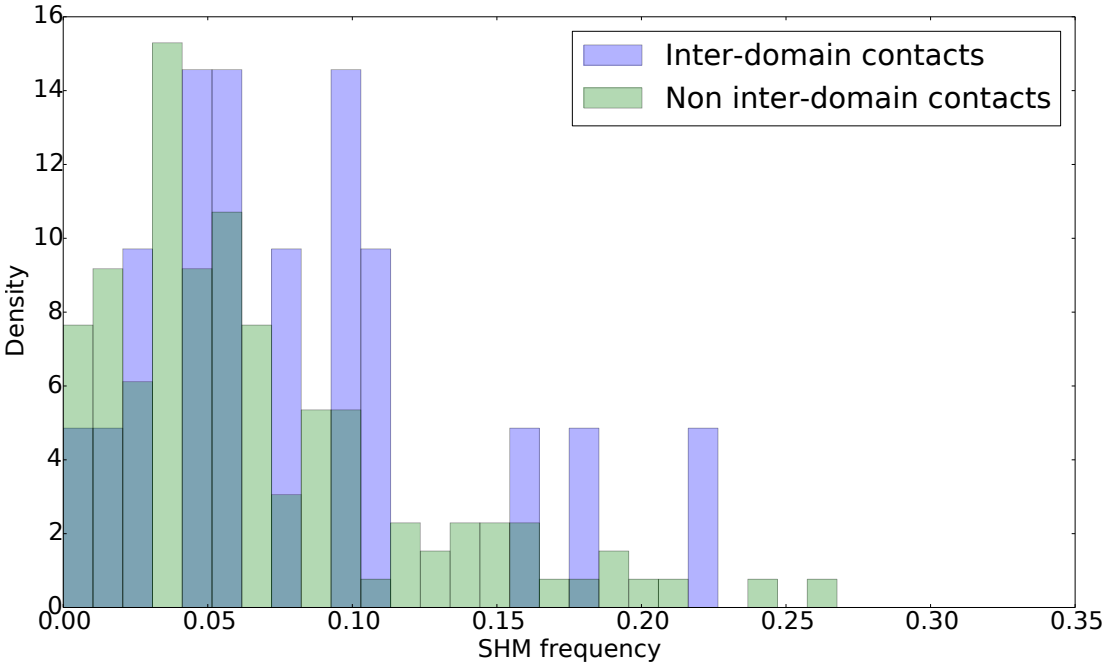


Figure 5.9: The frequency of SHMs at positions in non-antigen-binding regions. Only positions where the SHM frequency is non-zero are shown. Positions that make inter-domain contacts on average have a slightly higher SHM frequency than position that do not make inter-domain contacts.

BLOSUM62	
Angle	**Position (τ_b)**
HL	L109 (-0.33) H21 (-0.27) L59 (-0.22) L68 (-0.20) L42 (-0.17)
HC1	L61 (-0.33) L109 (-0.31) L102 (-0.28) L59 (-0.27) H80 (-0.19)
HC2	L44 (-0.35) H17 (-0.26) L65 (-0.25) L38 (-0.22) L101 (-0.20)
LC1	L101 (-0.58) L109 (-0.44) H111 (-0.36) L59 (-0.33) H21 (-0.32)
LC2	L101 (-0.60) H21 (-0.38) L42 (-0.27) L65 (-0.23) L60 (-0.22)
Miyata	
Angle	**Position (τ_b)**
HL	L109 (-0.33) L38 (-0.31) H4 (-0.20) L100 (-0.17) L68 (-0.17)
HC1	L109 (-0.31) L102 (-0.28) L61 (-0.28) L21 (-0.27) L8 (-0.21)
HC2	L44 (-0.35) L64 (-0.23) H87 (-0.22) L65 (-0.21) L101 (-0.20)
LC1	L101 (-0.58) L109 (-0.44) H111 (-0.36) L38 (-0.31) H86 (-0.29)
LC2	L101 (-0.60) L105 (-0.36) H15 (-0.26) L42 (-0.22) L65 (-0.20)

Table 5.6: The five positions for each angle where the mutation magnitude of SHMs have the strongest correlation with angle change (see Section 5.3.4). Mutation magnitude is defined as either the BLOSUM62 or Miyata dissimilarity scores. In brackets after each position is the value of the τ_b correlation (Section 5.2.4.4).

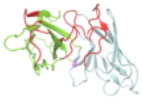
5.3.5 Structural differences in framework secondary structure and its relationship with VH-VL orientation

The structures of five of the six CDRs have long been thought to be represented by a small set of discrete shapes known as the canonical classes (see Section 1.3.2.4). Similarly, the shapes that CDR-H3 can adopt have also been found to be constrained at least at the basal regions of the loop [Kuroda *et al.*, 2008; North *et al.*, 2011]. Such shapes can be successfully determined from sequence. Some positions in the loops are key for maintaining the loop structure. Changing the residue at these positions can, in some cases, change the CDR from one canonical form to another. At other positions the residue can vary between a subset of amino-acids whilst retaining the backbone-conformation.

Given this sequence dependence for antibody CDR conformations, we investigated whether a similar property is seen for the Hifw-loop and the Lifw-loop. Packing of residues in these regions is thought to play a key role in variable domain association [Chothia *et al.*, 1985; Vargas-Madrazo & Paz-García, 2003]. The identity of key residues in these loops has also been previously found to influence VH-VL orientation [Abhinandan & Martin, 2010; Chailyan *et al.*, 2011]. Here we show a possible mechanism for how residues in these regions are able to vary the VH-VL orientation by changing the conformation of the Hifw-loop and the Lifw-loop.

5.3.5.1 The shapes of the Heavy and Light interface framework loops

Distinct conformations for the Hifw-loop and Lifw-loop were identified as described in Section 5.2.5. Seven discrete shapes were found for the Hifw-loop and 4 for the Lifw-loop. Each canonical shape is shown in Figure 5.10. The largest RMSD cluster



for the Lifw-loop is represented by two different shapes, referred to as canonical 1 and 1-sub in Table 5.4. Examining representative structures from the two populations shows that the difference in conformation is due to a rotation of glycine at position L41.

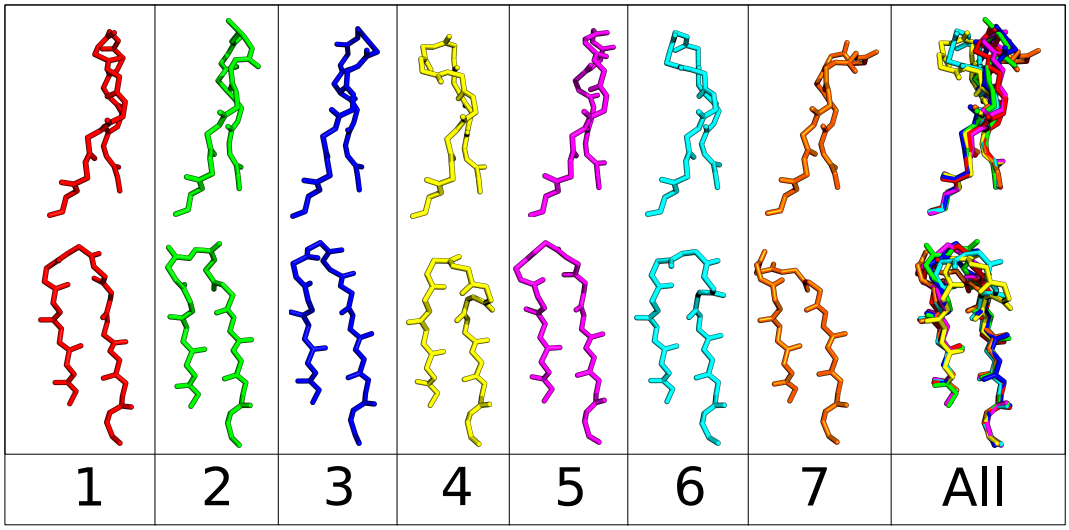
Using the method described in Section 5.2.5.4 sequence traits of the Hifw-loop and Lifw-loop were identified. Figures 5.11 and 5.12 show the classification trees for the heavy and light loops respectively. For the Hifw-loop we are able to classify 5 of the 7 classes. Two of the shapes, 5 and 6, are indistinguishable from larger populations of shapes using sequence alone. Shape 5 is very similar to shape 1 and differs by a rotation of the residue at H42 that is a glycine for most cases.

5.3.5.2 Packing of the residue at H43

Position H43 in the Hifw-loop is almost always a positively charged lysine or polar glutamine/asparagine. Given this, it might be expected for the side-chain to point into solvent. This is observed for the majority of cases. We refer to it as the "out" conformation of H43 as shown in Figure 5.13. However, in a significant minority of structures the residue instead packs deeply into the VH-VL interface. We refer to this conformation as the "in" conformation of H43 also shown in Figure 5.13. Such a difference in packing may influence (or be caused by) a difference in VH-VL orientation. We classified conformation of the H43 residue in each structure by calculating the direction of its $C\alpha$-$C\beta$ bond. Ambiguous structures were manually inspected in Pymol.

Figure 5.14 shows a decision tree for classifying the two conformations built using the Hifw-loop sequence and H43 contact positions L85, L87 and L100. It appears that the conformation is "in" if the "out" state is restricted in some way. For example, if

a)



b)

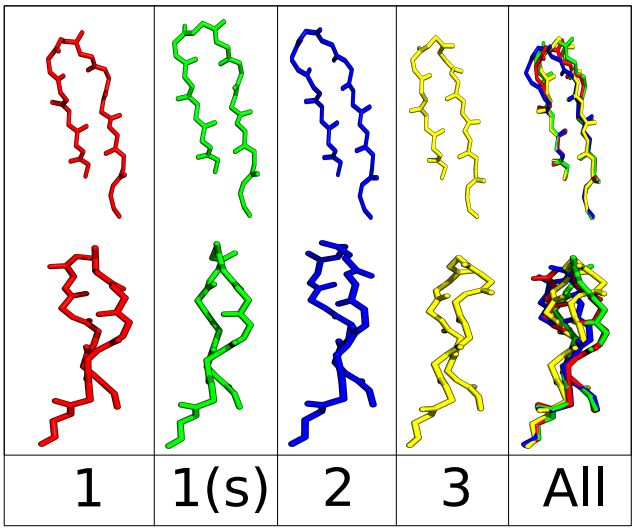
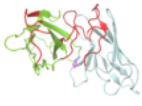


Figure 5.10: The canonical interface framework loop shapes for a) the VH domain and b) the VL domain. A front and side view is shown for each shape. Only backbone atoms are shown.Seven shapes are identified for the Hifw-loop. Four shapes are found for the Lifw-loop. The first two shapes (1 and 1(s)) fall into the same RMSD cluster. However, they have distinct dihedral strings due to a flip of a glycine residue at L41.



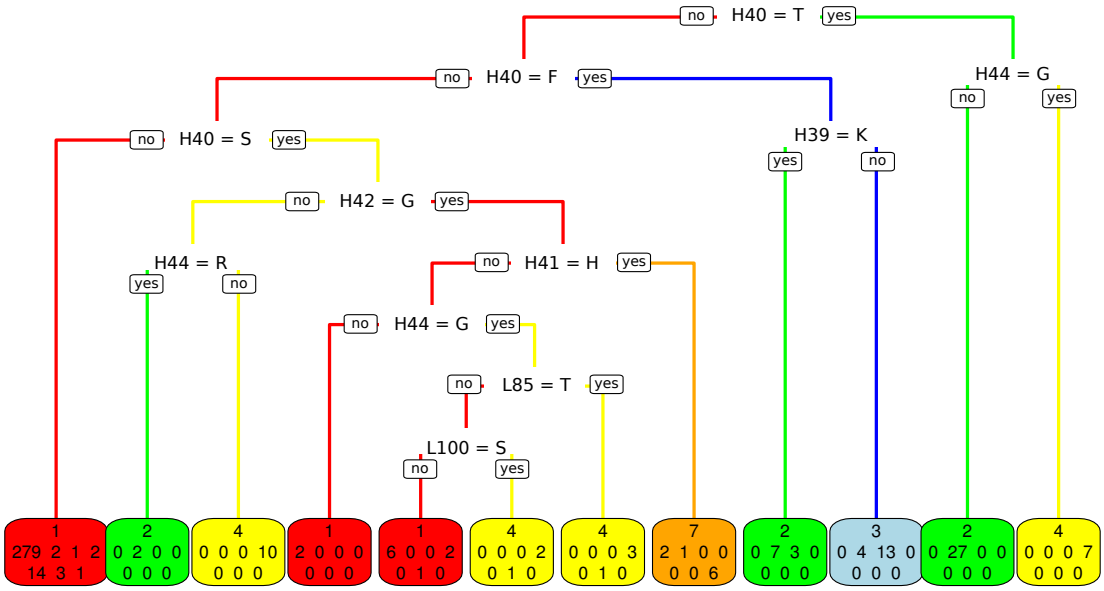


Figure 5.11: Classification by sequence of the seven Hifw-loop canonical shapes shown in Figure 5.10a. Each node of the tree describes whether a particular residue is present at a given position. All colours correspond to the shape that has the most number of representives in the branch or node and are equivalent to Figure 5.10a. Each leaf is labelled with how many of the shapes are classified into it.

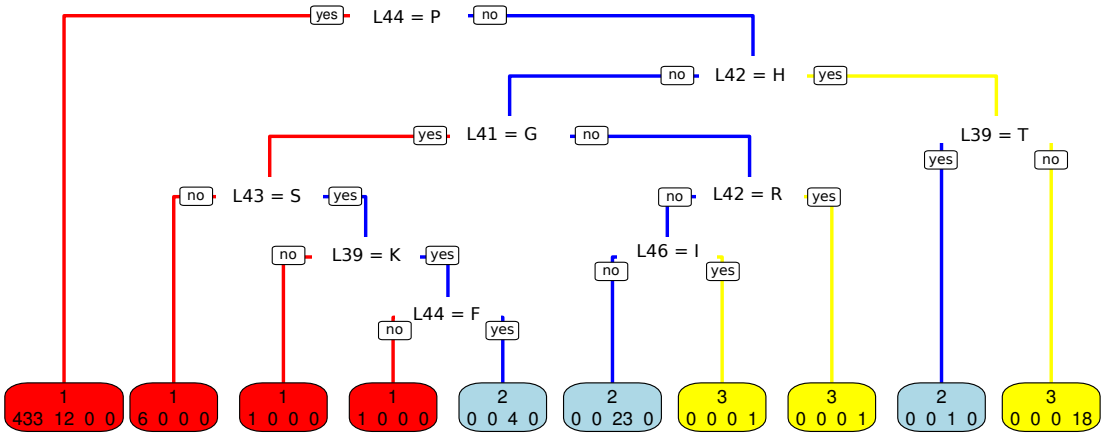


Figure 5.12: Classification by sequence of the four Lifw-loop canonical shapes (1, 1(s),2,3) shown in Figure 5.10b. See legend of Figure 5.11 details of annotation. Colours correspond to those in Figure 5.10b.

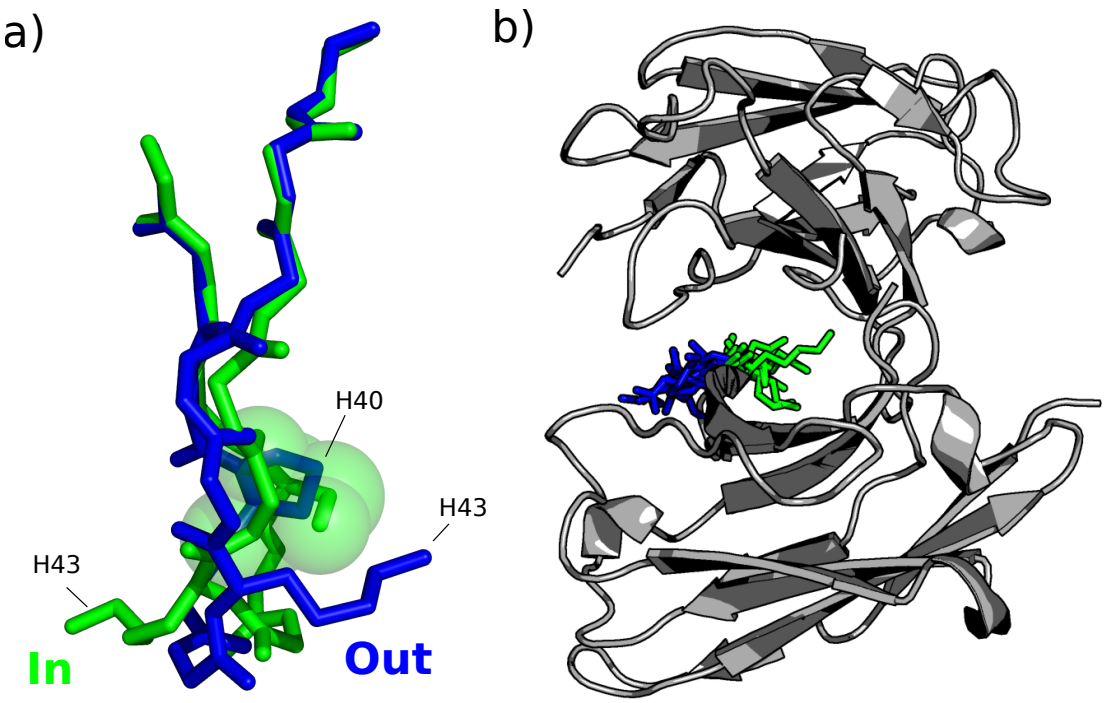
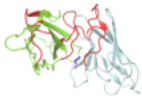


Figure 5.13: Examples of the "in" and "out" conformations. a) When a residue larger than proline or alanine is present at H40, the side-chain of H43 cannot occupy the "out" conformation. In this example, the "in" structure has a threonine present at H40 (spheres). The "out" structure has a proline at the same position. b) Position H43 is located on the Hifw-loop distal from the antigen binding site. The "in" conformation means that the residue packs into the interface between the VH and VL domains. For the "out" conformation the side-chain points out into the solvent.



H44 is a glycine, the loop can rotate at this position and almost all structures have the "out" conformation. Otherwise, there is a restriction on the dihedral angles allowed at this position. The state is then largely determined by the residue at H40. In general, if a small amino-acid is present (alanine or proline), the "out" conformation can still be achieved. If a larger residue is present, the side-chain occupies the same space that the H43 amino acid would assume and therefore an "in" conformation is found. This difference is shown in Figure 5.13.

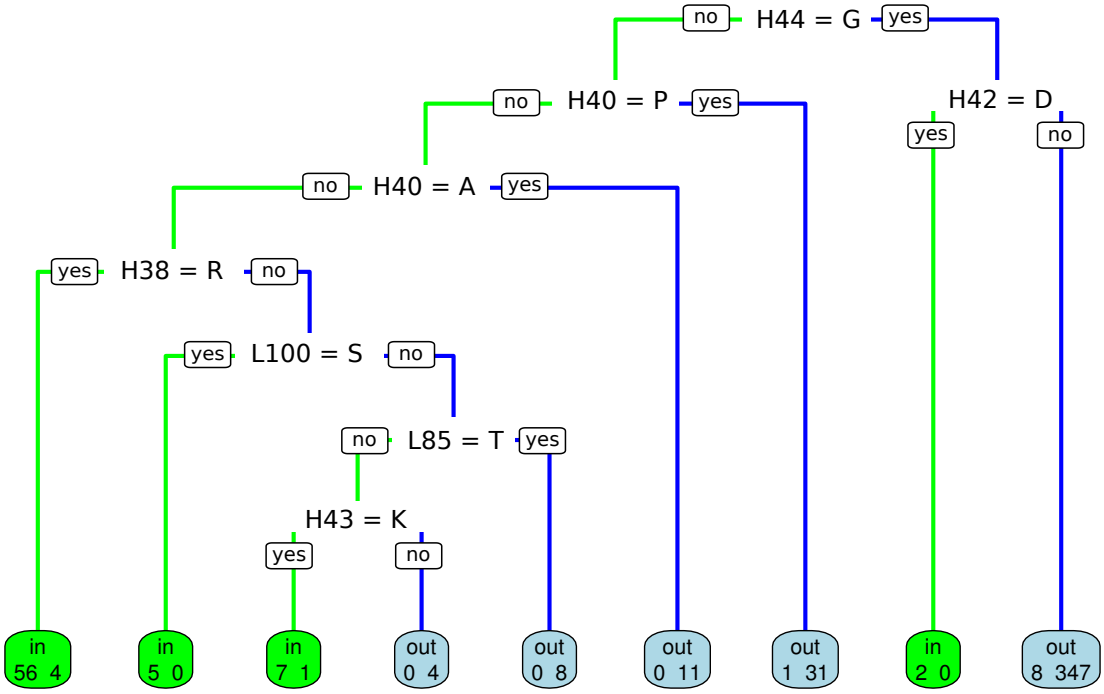


Figure 5.14: Classification by sequence of the "in" and "out" conformations as shown in Figure 5.13. See legend of Figure 5.11 details of annotation. Colours correspond to those in Figure 5.13

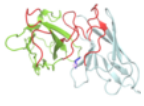
5.3.5.3 Structural differences influence orientation

We assessed the relationship between the structural classes of the framework loops and the VH-VL orientation. On average structures with an "in" conformation of the

H43 residue have a significantly wider HL torsion angle than those with an "out" conformation (p-value 3×10^{-8} Figure 5.15). In addition, the populations have significantly different LC1 and HC1 tilt angles (p-values 1.0×10^{-5} and 1.5×10^{-5} respectively). The orientation in the three other measures are not significantly different. The difference in packing at the interface appears to relate to a change in angle in the tilt and torsion angles. Interestingly, in Chapter 3 we identified that the residue at contact position L44 could also change the packing and influence the tilt angles. Here a similar effect (changing the tilt angles) is achieved using a different mechanism. Instead of a contact residue changing, a different framework loop conformation changes the direction of the H43 side chain and therefore inter-domain packing. The "in" conformation appears to be mainly determined by a non-contact position, H40. This effect highlights how residues that do not make direct contact with the opposing domain may influence the VH-VL orientation.

We also assessed the relationship between combinations of the in/out conformation, Hifw-loop canonical and Lifw-loop canonical. Structures in the training set were grouped by their combination of the three properties. The orientation distributions of those groups with five or more representatives is shown in Figure 5.16. A degree of specificity is observed for particular in certain orientation angles. For example, the combinations of in-2-1 and in-7-1 are specific for different regions of the HC2 angle. This may be due to a different level of interface side-chain packing that the conformation of the Hifw-allows.

We propose that the structural variation accounts, in part, for the difference in orientation between certain antibodies. Varying these properties, either on the germline or during B-cell maturation may be an *in vivo* mechanism to mediate variation in VH-VL orientation.



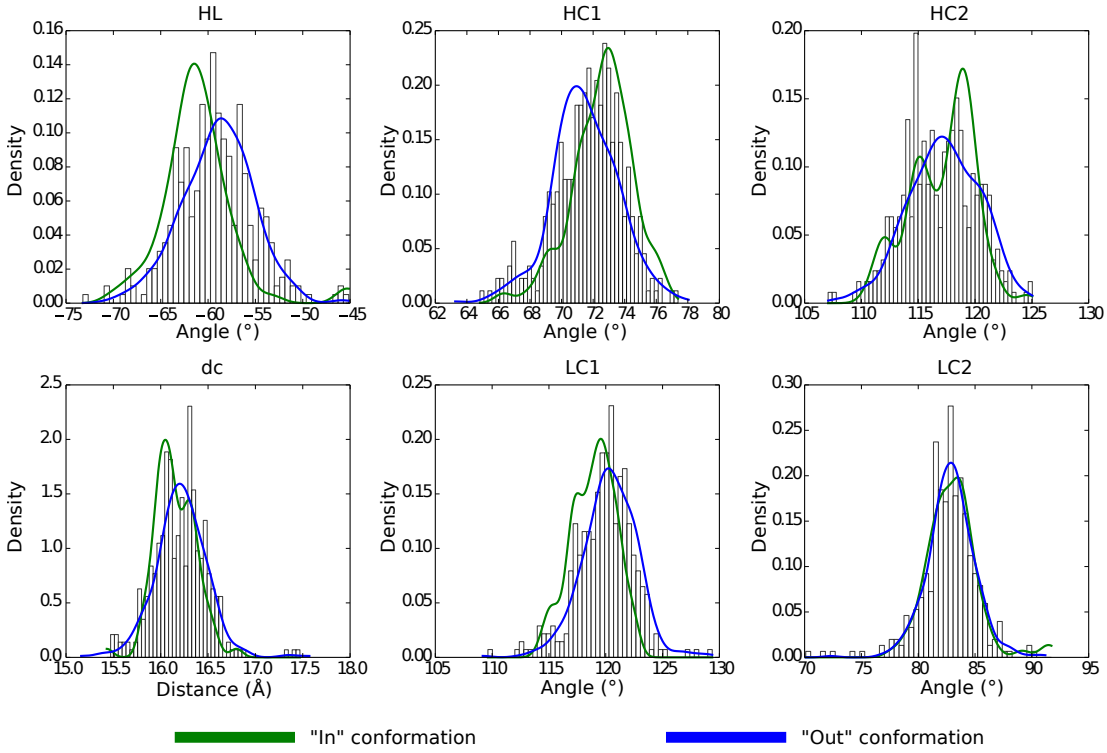


Figure 5.15: A comparison of the VH-VL orientations of structures with the "in" and "out" conformations of H43 as shown in Figure 5.13. When the H43 has the "in" conformation, its side-chain packs into the VH-VL interface whilst the "out" conformation points into solvent. Small yet significant differences in these two sets of structures are found in the torsion and tilt angles. The "in" conformation appears to widen the torsion angle compared and tilt the VH domain towards the VL domain.

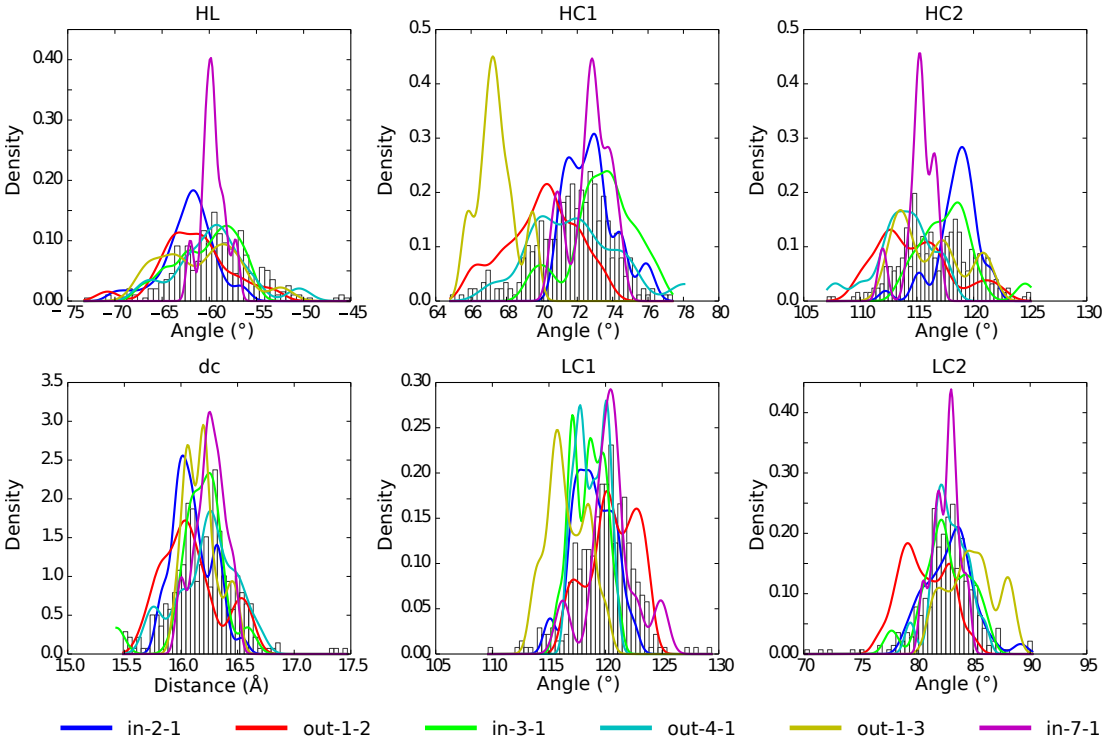


Figure 5.16: The relationship between combinations of structural classes of the interface framework loop and the orientation angles. The legend corresponds to the combination of in/out conformation (Figure 5.13), Hifw-loop canonical and Lifw-loop canonical (Figure 5.10). Only those combinations with 5 or more representative are shown. The majority of structures have the out-1-1 combination. This distribution is not shown.

5.3.6 The end of CDR-H3 is more important than the beginning

We investigated whether the North *et al.* [2011] canonical forms of the CDRs influenced the VH-VL orientation. No relationship is observed for individual canonical forms of the loops. Similarly, combinations of the shapes also showed no relationship to the orientation.

However, CDR-H3 residues are involved in many contacts with the VL domain [Chothia *et al.*, 1985; Morea *et al.*, 1998; Vargas-Madrazo & Paz-García, 2003; Kuroda *et al.*, 2008; Messih *et al.*, 2014]. Indeed, several authors include information about VL contact positions in their prediction methods and discussions of CDR-H3 loop conformations [Morea *et al.*, 1998; Kuroda *et al.*, 2008; Choi & Deane, 2011; Messih *et al.*, 2014]. Figure 5.17 shows the frequency at which the residue positions in CDR-H3 make contact with the VL domain. Here, we quantify an observation made by Chothia *et al.* [1985] based on a small number of available structures, that residues towards the end of CDR-H3 are involved in inter-domain contacts more often than those towards the beginning. We find that the last five residues of CDR-H3 have a mean inter-domain contact frequency of 99% whilst the first five residues have a mean contact frequency of 50%. It is not possible to see this effect using the Chothia numbering scheme because the annotation of, for example, the third last position depends on the length of the loop (Section 5.2.2 and Table 5.1). This highlights the importance of using an anchor-based numbering scheme for the CDR-H3 loop especially for analysis, like ours, that focuses on inter-domain contacts.

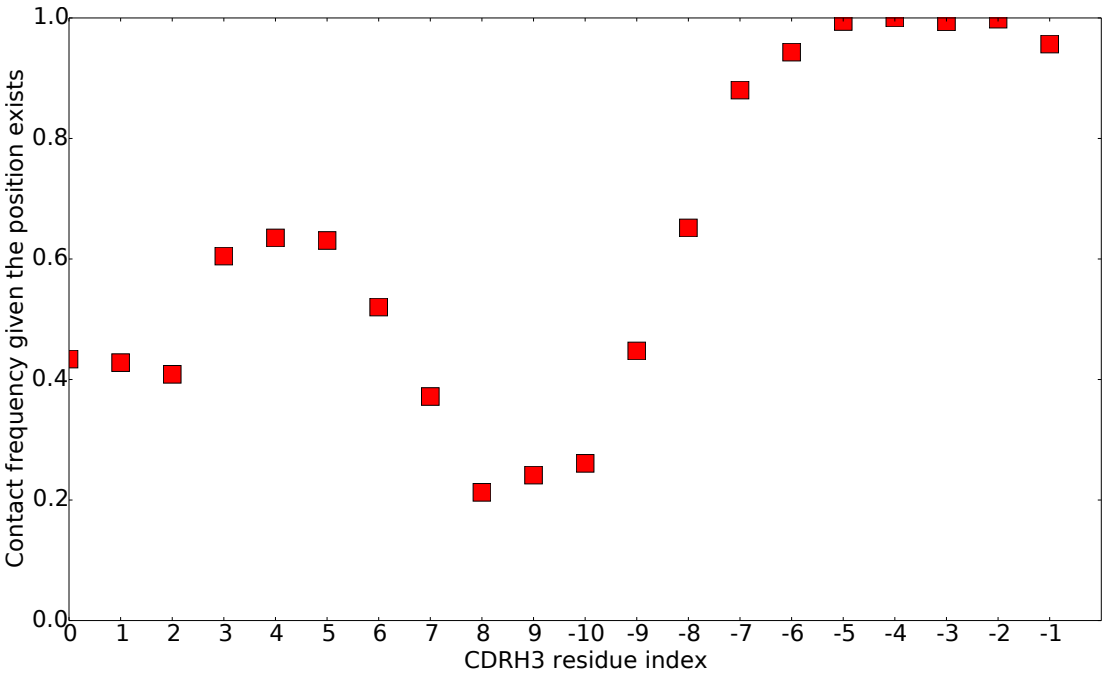
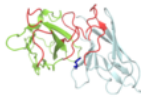


Figure 5.17: The frequency at which CDR-H3 residue positions are found to be in contact with any residue in the VL domain. Here, the anchor numbering scheme as described in section 5.2.2 is used so that, for example, -3 refers to the third from last residue in CDR-H3. The contact frequency at each position is calculated using only structures for which the CDR-H3 position exists (see Table 5.1 for a correspondance with loop length). The mean contact frequency of the last five positions is 99% whilst the mean for the first five is only 50%. Residues near the end of the loop are therefore more likely to be influential for VH-VL orientation than those near the beginning. For longer CDR-H3s, the residues near the middle of the loop (e.g. positions 8, 9 and -10) are less likely to be in contact the VL domain and therefore less likely to affect the VH-VL orientation.

s



5.3.6.1 Missing CDR-H3 contacts cause the domains to tilt

Given that when present the last five positions make contact with the VL domain, we investigated if a common difference in orientation is found when some of the positions are missing i.e. when the CDR-H3 loop is seven or less residues in length (see Table 5.1). These we call short CDR-H3 loops and others normal CDR-H3 loops. All structures with short CDR-H3s for which there exists a structure with a sequence identity higher than 85% and a normal length CDR-H3 were identified. For these 19 pairs, the differences in ABangles were calculated.

Figure 5.18a shows the distributions in angle difference for the short-normal pairs. For comparison we calculated the angle difference when the short-looped structure was replaced with another structure with a normal length loop and a sequence identity higher than 85%. The normal-normal distributions are shown in Figure 5.18b. We find a positive skew in the HC1 and LC1 tilt angles in the short-normal distributions that is not present in the normal-normal distributions. This corresponds to more acute HC1 and LC1 angles in the short looped structures than their normal counterparts. Therefore, if contact positions at the end of the CDR-H3 are not present, as with short looped structures, one should expect VH and VL to tilt towards each other relative to similar antibodies with normal length CDR-H3 loops. This should be considered during humanisation protocols if murine CDR-H3 loops longer than 7 residues are replaced with short human CDR-H3 loops or *visa versa*.

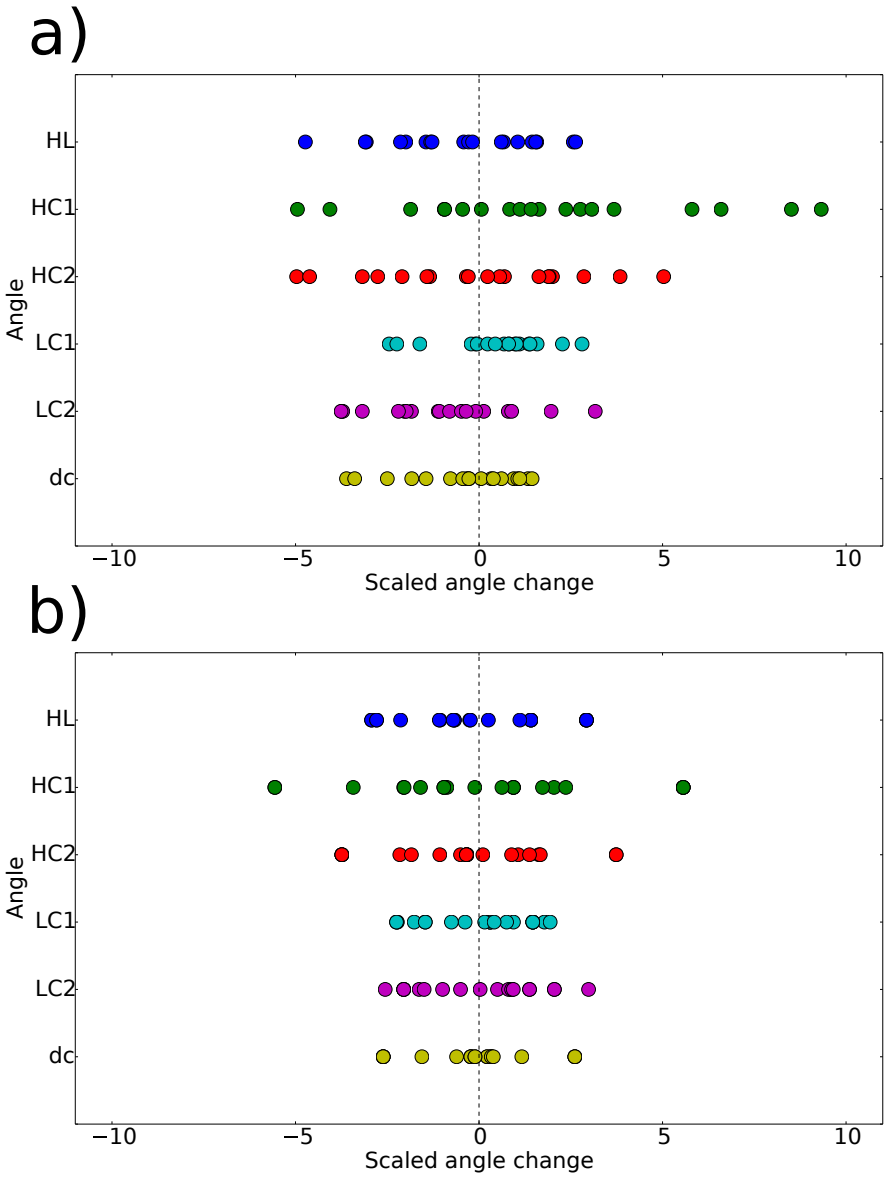
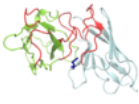


Figure 5.18: Assessing the influence of short CDR-H3s on orientation. a) The distributions of scaled angle differences between structures with short loops (seven or less residues) and similar (>85% sequence identity) structures with normal length loops (8 or more residues). The HC1 and LC1 tilt angles are skewed to the right suggesting that on average the structure with a short loop will have more acute tilt angles. b) The equivalent plot for when the short looped structure is replaced with another similar structure (>85% sequence identity) with a normal length loop. Here, we find no skew in the angles. The angle differences have been scaled by the difference in angle one can expect from two sequence-identical structures.



5.3.7 Performance of feature based predictor using non-similar antibodies

We have shown in Section 5.3.2 that selecting a good template for orientation becomes more difficult when the full sequence identity falls below approximately 85%. It is in this region in which modelling protocols should be benchmarked. Our prediction protocol described in Section 5.2.6 ranks templates based on their suitability for each of the ABangles. To do this, feature profiles are built for each template using other structures that are similar in angle to the template. Given the target sequence, these profiles are compared and the templates ranked. A final template is predicted by taking the value of the top predicted template for each angle and reconstructing the geometry (raw prediction). Additionally, the database of redundant structures is search to find a real template with a similar orientation to the raw prediction (database prediction).

Figure 5.19 shows how the two prediction methods perform in comparison to taking the maximum sequence-identity template. Predictions are binned by the maximum sequence identity of the available templates to the target. Our raw method predictions slightly enrich the probability of picking a good orientation template compared to using the top sequence identity. This method can return a prediction for all targets. Our database method takes the raw prediction and checks that an antibody structure with a similar orientation exists for any structure in the redundant set with a sequence identity lower than the maximum template-target sequence identity. This means that it has lower prediction coverage. However, when it does make a prediction it is significantly better than using the top sequence identity template, especially as the sequence identity reduces.

The performance of the methods were assessed on the test set of structures described in Section 5.2.1. Using the top full-sequence identity template in the training set, a good prediction was made for 54.8% of the targets. In comparison, the database method made a good prediction for 61.6% of targets whilst the raw method achieved 56.0%.

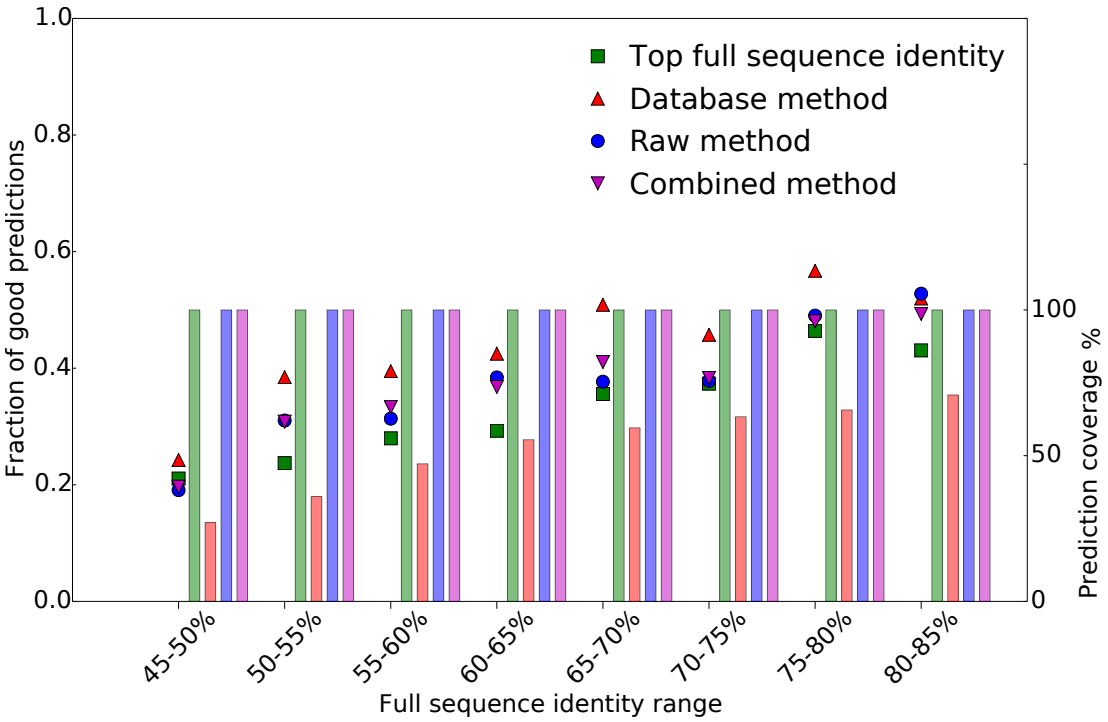
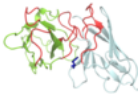


Figure 5.19: The performance of the feature based predictor. The predictor was run at a range of allowed maximum template-target full sequence identity thresholds. Predictions are binned by the sequence identity of the maximum available template-target sequence identity. The points show the fraction of predictions made that gave good orientation templates when each of the methods described in Section 5.2.6 were used. The combination method makes the database method prediction if it returns an answer and the raw method prediction otherwise. The bars show the prediction coverage for each method within each maximum sequence identity range. The method with the highest prediction accuracy is the database method. However, this also has the lowest coverage. All three of our methods outperform the prediction accuracy of simply using the top sequence identity template. However, the prediction accuracy still remains low.

Although we are able to improve over a simple protocol that improvement is minor.



One reason for this could be the lack of any good templates in the initial 20 most sequence similar structures. However, we find that for all the targets where we make a bad prediction, there is at least one good template in this set. Another, is that our definition of what is a good template is too strict. Relaxing the condition naturally raises the perceived performance of any predictor. We do not find that our prediction method improves relative to using simple sequence identity measures at more relaxed thresholds.

Most likely is that the orientation of a particular antibody is determined by multiple different factors (e.g. residues at positions), the specific combination of which determines the pose. Unless such a combination exists in the dataset, the reliable high-resolution prediction of such an orientation is challenging. We see some evidence of this in the drop in orientation-similar antibodies as sequence identity drops below 90% (Figure 5.6).

However, certain trends do exist and different residues at certain positions are found to be influential for determining orientation. These properties are useful to rationalise what sequence changes may affect orientation on a case-by-case basis. When selecting a template for prediction, many of these properties are captured by using templates with high sequence similarity. Properties such as the correct framework loop conformation or conservation of the majority of interface residues will be shared by all potential templates.

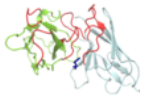
The limited success of our method suggests that more accurate VH-VL orientation prediction may benefit from moving away from template-based prediction. This may be done through explicit prediction of the ABangles using a more complex approach or through an energy-based selection of possible orientations.

5.4 Conclusion

Variable domain orientation is a property that antibodies can use to diversify the shape of their antigen binding site above and beyond CDR shape and residue types. In this chapter we have investigated properties that may influence the orientation.

Sequence identical structures were found to have a VH-VL orientation that can vary up to 1.5Å orientation RMSD. This value provides a threshold at which a prediction can be considered correct. Models built using homology techniques generally improve as the sequence identity of the highest available template increases. At above 85% target-template sequence identity, the most similar structure available is likely to provide a good orientation template. Below this threshold orientation prediction is non-trivial. Methods that predict this property should be benchmarked when the maximum target-template sequence identity is lower than 85%.

Changing the variable domain orientation is a possible mechanism that could be used to adjust antigen binding properties during affinity maturation. We investigated this biological mechanism by analysing the relationship between the magnitude of SHMs at each residue position and how they changed the six ABangle measures. The results were not consistently interpretable. A limitation to the approach taken is decoupling the effect of one particular SHM and other differences between structures. Ideally we would wish to have structures of germline antibodies and corresponding molecules with a small number of point mutations introduced. In the absence of these data, we compared structures with a particular SHM to similar structures that have the germline residue at the same position. A compromise had to be made between the level of similarity used and the amount of data available to perform statistical tests. We chose to call structures similar if they shared the same v gene subgroup for the



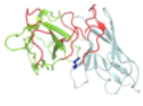
domain on which the SHM was located. Given the number of examples available, the level of signal to noise is currently too low to detect any consistent effects. Our results are therefore inconclusive as to the role of SHMs in influencing VH-VL orientation.

Residues in the framework loops between the C and C' β-strands in the VH and the VL domains form many of the inter-domain contacts. A structural property that may be linked to VH-VL orientation are the conformations of these interface framework loops. We identified several distinct shapes of the Hifw-loop and the Lifw-loop. Analogous to the CDR loops, these shapes are recognisable by certain sequence properties and can be categorised into canonical forms.

Certain residues within the two interface framework loops are found to be important for determining orientation (Section 3.3.3). We found that the residue at position H43 can adopt a conformation that points into or out of the VH-VL interface, thus affecting inter-domain residue packing. Changing conformation of this residue appears to affect the VH-VL orientation. However, its conformation is largely determined by the amino-acid at H40, a residue that does not make contact with the opposing domain. The identity of non inter-domain contact residues should therefore be considered when determining orientation. Mutating the residue at H40 may act as an orientation "switch" and should be considered when engineering therapeutic antibodies.

No significant relationship was identified between the VH-VL orientation and the structures of CDR-H3 and CDR-L3. However, the VH and VL domains of antibodies with short CDR-H3 loops tilt towards each other more than compared to similar structures with longer loops. As the short loop cannot contribute as many residues to inter-domain packing, the variable domains collapse towards each other. This is captured by our HC1 and LC1 tilt angles.

We used both sequence and structural properties to build a prediction method for the VH-VL orientation based on the six ABangle measures. Although prediction performance was better than a more simple method, the improvements made by our feature based protocol were minor in our benchmark over maximum target-template sequence identity space. A major limitation is the use of "real" templates to model a target sequence. Further work is required to improve the prediction of VH-VL orientation. One approach may investigate the energetics of putting a pair of domains in given orientations. In the next chapter we discuss the future directions for this research and conclude the thesis.



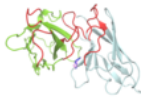
Chapter 6

Conclusions and future directions

Antibodies play a key role both in the natural immune response and increasingly as therapeutic agents. Understanding their structural properties can provide insight into how they, and other antigen receptors, are able to specifically sense epitopes and give clues as to how to better engineer these molecules in the laboratory. In this thesis I have investigated one such structural property, the VH-VL orientation. In doing so I have created a database to automatically collect the rapidly growing amount of antibody structural data; developed methodology to characterise the absolute VH-VL orientation; used this methodology to compare conformations in different receptor types; and identified sequence and structural properties that influence the inter-domain pose.

6.1 SAbDab: the Structural Antibody Database

In Chapter 2 I described the Structural Antibody Database (SAbDab). My motivation for developing SAbDab was to have, within the research group, the ability to automatically obtain all publicly available antibody structure data. These data and their annotations were then available for analysis and integrated into antibody



computational tools. SAbDab's Python API allows the desired data retrieval and includes other functionality such as an interface to antibody numbering programs, identification of missing residue coordinates and the recognition of SHM positions. This API has been installed and is in active use at Roche where it extends the company's existing structure database, at UCB where it is part of a modelling protocol and antibody-antigen docking pipeline and within my research group as the basis for the SAbDab webserver.

Within SAbDab and throughout the majority of this thesis, the Chothia numbering scheme has been used. This decision was made due to the scheme's popularity in the literature and that it can be applied rigorously and simply using the ABnum tool. More modern schemes can provide annotations that better reflect equivalent residue positions (e.g. IMGT). However, there are no stand-alone computational tools that are freely available and suitable to numbering many antibody sequences with these schemes. The ARNACI algorithm described in Section 2.2.2.2 is able to provide IMGT numbering but further work is required to improve the algorithm and to extend the different schemes it can apply.

One application of SAbDab is to search the database for good templates to model an antibody given a its sequence. The template search capability ranks potential templates based simply on their matched sequence identity to the sequence. A better way to perform the ranking would be to calculate a BLAST like E-value score. Such a score would not treat amino-acid differences at each position equally. Those positions at which the residue is variable may influence the overall structure less than those that are typically more conserved. Weightings based on structural information could also be incorporated. For example, the positions that determine the conformation of the interface framework loops (Section 5.3.5.1) could be up-weighted for selecting an

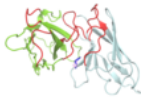
appropriate template.

SAbDab continues to develop within the group. Areas that should be improved include more detailed antigen annotation, highlighting parts of antibody structures that may be unusual and better descriptions of the level of engineering undertaken to produce the molecule (e.g. annotations of humanised antibodies).

6.2 Characterising the variable domain orientation in antibodies

In Chapter 3 I described a method to characterise, in an absolute sense, the VH-VL orientation in antibodies. The ABangle method extends on Abhinandan & Martin [2010]'s single torsion angle characterisation of VH-VL packing to the six measures required to fully describe the orientation between three dimensional objects. These measures allow one to compare both the magnitude and the direction of how structures differ within a consistently defined framework. Therefore the influence that an individual residue has on orientation in a particular direction can be studied. This ability was demonstrated by showing that influential residues identified by different authors were describing changes in different modes of orientation.

ABangle can be used to measure VH-VL orientation and its response to different factors. The conservation of orientation was found to be dependent on the type of antigen that the antibody is specific for. Having a flexible binding site may impose more of an entropic penalty for binding to a small molecule hapten antigen than a larger protein antigen. Indeed, unbound hapten-binding antibody structures were found to have a more conserved VH-VL orientation that protein-binding antibodies.



These results suggest that antigen specificity may be related to the dynamics of the molecule. However, variation in crystal structures does not directly correspond to protein flexibility. Future work should study flexibility of domains either experimentally using NMR or computationally using molecular dynamics simulations and other conformational sampling techniques [Sim *et al.*, 2012].

Given the relationship between antibody specificity and orientation conservation in the crystal data one might expect to find sequence and structural features that are influential for determining flexibility. One structural feature that should be explored is the relationship between flexibility and packing of residues in the CDRs. Different densities of binding site packing may stabilise the variable domain orientation to different extents. Classifying different packings could be achieved by recognising different antibody binding site shapes [Lee *et al.*, 2005].

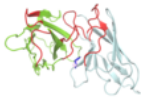
Whilst ABangle has thus far been used as an analysis tool it may in the future be used in the computational design of antibodies. For example, loop structures and sequences complementary to a patch on an antigen could be predicted or taken from known binding partners. In a similar manner to humanisation, one could graft these loops onto an antibody framework. Here, ABangle could provide a mapping between the putative geometry of a predicted paratope and sequence that would allow the required VH-VL orientation to be achieved.

6.3 Comparing variable domain orientations in different antigen receptors

Engineering of the VH-VL orientation to allow antibodies to bind specifically to a target was investigated in Chapter 4. A different type of antigen receptor, the TCR, was found to have distinct variable domain orientations to antibodies. Although antibodies and TCRs perform similar roles, they are under different selective pressures. To be able to recognise immunogenic peptides, TCRs must be able to bind with weak affinity to an MHC. Any antibody with such specificity would be deleted by the immune system to prevent auto-immune disease.

The difference between the receptor's variable domain orientations makes it difficult for an antibody to bind the pMHC in the same way a TCR does. This structural difference may be an evolved mechanism to differentiate the receptors. It would therefore be encoded by the respective variable region genes. However, a TCR-like variable domain orientation could be a feature selected against in antibody clonal selection or selected for during TCR thymic education. If so, some common features may be found in either the non-functional antibody and TCR variable genes or the unselected sequences of each receptor. For example, as described in Section 5.3.6.1, the packing of the long TCR CDR-α3 appears to influence the orientation. One could investigate whether unselected T-cells produce TCRs with CDR-α3 loops that have significantly different inter-domain packing properties to functional T-cells. In antibodies a TCR-like orientation can be achieved with a bulky residue at IMGT position H50. This mimics the CDR-α3 packing in TCRs. One could examine the non-functional variable genes to see whether this feature might be responsible for their negative selection.

The results of this chapter suggest that variable domain orientation should be



considered when developing therapeutic TCR-like antibodies. I identified candidate positions that could be mutated to increase the similarity of the VH-VL interface to the Vβ-Vα interface. The next stage would be to experimentally verify whether these mutations are able to promote a TCR-like VH-VL orientation and thus improve pMHC specificity. Influences on VH-VL orientation could be investigated by examining the crystal structures of the same antibody with and without the mutation (e.g. leucine to phenylalanine and IMGT H50). The effect on pMHC specificity could be assessed by comparing the performance of phage libraries with and without the mutation to generate high affinity antibodies.

In this chapter I demonstrated how the ABangle methodology can be used to characterise the orientations between pairs of domains other than VH and VL. The protocol outlined in Section 3.2 can be followed to set up a framework to investigate the absolute conformations between any two domains with multiple examples available in the PDB. One such example is the HIV protease homo-dimer. Here, the methodology could be used to study dynamics of protease upon ligand binding and inform the subsequent development of anti-viral drugs that target this protein.

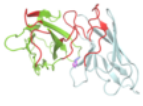
6.4 The determinants of VH-VL orientation

In Chapter 5 I investigated what determines VH-VL orientation with a view to improve its prediction from sequence. Homology modelling protocols tend to perform better as the available target-template sequence identity increases. Above a certain point, structure prediction becomes trivial to within a level of accuracy. Therefore, I attempted to define both the level of accuracy one should aim to predict VH-VL orientation and the sequence identity at which the most sequence similar template

should always be used. These values were within 1.5Å orientation RMSD, reflecting the variation within sequence-identical structures, and above approximately 85%, corresponding to the point where one would expect to reach the accuracy at least 50% of the time. Methods to predict orientation should therefore report performance benchmarks when their underlying structure databases exclude structures with template-target sequence identities higher than this threshold.

The VH-VL orientation may be changed during antibody maturation. I therefore attempted to identify whether SHMs at certain positions have more or less effect on VH-VL orientation. No significant correlation was found between mutation magnitude and the change in angle from other antibodies sharing the same germline subgroup. However, such an analysis is currently unlikely to give a fair representation of the biological mechanism due to a lack of data. For example, there are very few available examples of germline antibody structures making it difficult to assess the "true" germline angles. Further work is required to estimate the angles for each germline sequence. In addition, SHMs rarely occur on their own. Without additional data, the ability to assess an individual mutation's effect is limited.

Many of the residues that have been found to influence VH-VL orientation are located in, or make contacts with, the framework loops at the base of the interface. In this chapter I identified different conformations of these loops analogous to the canonical forms of the CDRs. Changing the shape of the loop may be a structural mechanism that allows variation in domain orientation. Indeed, certain loop conformations are preferable for certain regions of the ABangle distributions. Different conformations may change the packing of the side chains at the interface. For example H43 either packs into or out of the interface. Its conformation is largely determined by the identity of a residue that does not make contact with the VL do-



main, H40. A recent study by Fera *et al.* [2014] has shown how VH-VL orientation can change in response to a changing viral antigen. One of the mutations made in the lineage is at H40. Indeed both the framework loop conformation and the VH-VL orientation changes in presence of this mutation. This highlights the importance of non inter-domain contact residues in determining the VH-VL orientation.

I used both sequence and structural features to build a prediction method for the VH-VL orientation. Although some improvement was gained compared to selecting the template based on highest sequence similarity alone, the benefit was minor. Further work on VH-VL prediction should consider several factors.

When building the training set I chose a single structure from each cluster to make a non-redundant set of templates. Each template therefore has a single orientation thus ignoring the possible variation in angle at this stage. Instead, the flexibility is accounted for at the assessment stage where a prediction deemed to be accurate if the model it produces has an orientation RMSD of less than 1.5Å to the target. Predictions may be improved by considering variation in angle at an earlier stage of the protocol.

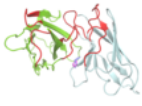
In this research I have focussed on determining the orientation based on sequence. It appears that, although some common sequence traits can be found, the precise angle is determined by different residues in different antibodies. Thus, unless a highly sequence similar template is already available, there is little guarantee that the orientation can be predicted by sequence alone. An alternative approach could be to use energetics. This has been largely avoided in this thesis due to the computational cost associated with energy-based predictions. However, preliminary results do show that more simple, computationally fast, statistical potential based functions (e.g. DFIRE [Zhou & Zhou, 2002]) can discriminate between good and bad orientation templates.

This ability is significantly reduced when the residue side chains are remodelled.

Further work should be performed to improve the side chain modelling for antibodies, particularly at the VH-VL interface. Current methods select rotamer angles from a library generated from known proteins. Given the level of data available for antibodies it may be possible to generate an antibody, or even position, specific rotamer library. Constraints for each rotamer's selection may be incorporated based on, for example, the conformation of the interface framework loops or the residue identity of their potential contacts. This may improve the ability to build accurate models of VH and VL and subsequently make rapid prediction of VH-VL orientation using an energy function possible.

6.5 Closing remarks

In this thesis I have investigated structural variation of the molecular sensors for the immune system, antibodies and TCRs. By developing SAbDab I have contributed a method by which to monitor the rapidly growing structural space of antibodies. This system enables further research to be performed using data curated in a consistent manner. The ABangle methodology addresses the problem of fully characterising variable domain orientation. I have used it to investigate the VH-VL orientation in antibodies and the Vβ-Vα orientation in TCRs. In doing so I have identified sequence and structural properties that influence variable domain orientation biologically and may be harnessed in the development of biotherapeutic agents. However, further challenges remain in the accurate prediction of variable domain orientation in antigen receptors.



References

Abhinandan, K.R. & Martin, A.C.R. (2007). Analyzing the degree of humanness of antibody sequences. *J. Mol. Biol.*, **369**, 852–862. 50

Abhinandan, K.R. & Martin, A.C.R. (2008). Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol. Immunol.*, **45**, 3832–3839. 28, 41, 50, 53, 75, 88, 180, 181

Abhinandan, K.R. & Martin, A.C.R. (2010). Analysis and prediction of VH/VL packing in antibodies. *Protein Eng. Des. Sel.*, **23**, 689–697. 37, 38, 66, 83, 84, 96, 104, 112, 152, 163, 168, 185, 207

Adair, J.R., Athwal, D.S. & Emtage, J.S. (1999). Humanised antibodies. *US Pat. 5,859,205*. 40

Al-Lazikani, B., Lesk, A.M. & Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.*, **273**, 927–948. 23, 27, 30, 32, 42, 58, 180, 181

Alberts, B., Alexander, J., Julian, L., Martin, R., Keith, R. & Walter, P. (2007). The Adaptive Immune System. In *Cell*, chap. 25, Garland Science, 5th edn. 12

Allcorn, L.C. & Martin, A.C.R. (2002). SACS: self-maintaining database of antibody crystal structure information. *Bioinformatics*, **18**, 175–181. 49

Alley, S.C., Okeley, N.M. & Senter, P.D. (2010). Antibody-drug conjugates: targeted drug delivery for cancer. *Curr. Opin. Chem. Biol.*, **14**, 529–37. 39

Almagro, J.C., Hernandez-Guzman, F., Maier, J., Shaulsky, J., Butenhof, K., Labute, P., Thorsteinson, N., Teplyakov, A., Luo, J., Sweet, R. & Gilliland, G.L. (2011). Antibody modeling assessment. *Proteins*, **79**, 3050–3066. 42, 151, 152

REFERENCES

Almagro, J.C., Teplyakov, A., Luo, J., Sweet, R.W., Kodangattil, S., Hernandez-Guzman, F. & Gilliland, G.L. (2014). Second antibody modeling assessment (AMA-II). *Proteins*, **82**, 1553–1562. 42, 43, 151, 152, 153

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410. 48, 124

Andersen, P.S., Stryhn, A., Hansen, B.E., Fugger, L., Engberg, J. & Buus, S. (1996). A recombinant antibody with the antigen-specific, major histocompatibility complex-restricted specificity of T cells. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 1820–1824. 122

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science (80-.).*, **181**, 223–30. 6

Ansari, H.R., Flower, D.R. & Raghava, G.P.S. (2010). AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res.*, **38**, D847–853. 41

Banfield, M.J., King, D.J., Mountain, A. & Brady, R.L. (1997). VL:VH domain rotations in engineered antibodies: crystal structures of the Fab fragments from two murine antitumor antibodies and their engineered human constructs. *Proteins*, **29**, 161–171. 37, 59, 82, 83

Barabási, A.L. & Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113. 48

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. 10, 48, 56, 88, 124

Betz, A.G., Rada, C., Pannell, R., Milstein, C. & Neuberger, M.S. (1993). Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proc. Natl. Acad. Sci. U. S. A.*, **90**, 2385–2388. 22

Biddison, W.E., Turner, R.V., Gagnon, S.J., Lev, A., Cohen, C.J. & Reiter, Y. (2003). Tax and M1 Peptide/HLA-A2-Specific Fabs and T Cell Receptors Recognize Nonidentical Structural Features on Peptide/HLA-A2 Complexes. *J. Immunol.*, **171**, 3064–3074. 122

Boes, M. (2000). Role of natural and immune IgM antibodies in immune responses. *Mol. Immunol.*, **37**, 1141–1149. 12

Borrebaeck, C. (2000). Antibodies in diagnostics from immunoassays to protein chips. *Immunol. Today*, **21**, 379–382. 38

Branden, C. & Tooze, J. (1991). *Introduction to protein structure*. Garland Publishing Company, New York. 9

Breiman, L. (2001). Random forests. *Mach. Learn.*, **45**, 5–32. 97

Brezinschek, H.P., Foster, S.J., Dörner, T., Brezinschek, R.I. & Lipsky, P.E. (1998). Pairing of variable heavy and variable kappa chains in individual naive and memory B cells. *J. Immunol.*, **160**, 4762–4767. 21

Brodeur, P. & Riblet, R. (1984). The immunoglobulin heavy chain variable region (IghV) locus in the mouse. I. One hundred IghV genes comprise seven families of homologous genes. *Eur. J. Immunol.*, **14**, 922–930. 21

Bromley, S., Burack, W., Johnson, K.G., Somersalo, K., Sims, T.N., Sumen, C., Davis, M.M., Shaw, A.S., Allen, P.M. & Dustin, M.L. (2001). The immunological synapse. *Annu. Rev. Immunol.*, **19**, 375–396. 117

Burkovitz, A., Sela-Culang, I. & Ofran, Y. (2013). Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. *FEBS J.*, **281**, 306–319. 23, 72, 158, 180, 182

Burnet, F. (1957). A modification of Jerne's theory of antibody production using the concept of clonal selection. *Aust. J. Sci.*, **20**, 67–69. 12

Carugo, O. & Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Sci.*, **6**, 2261–2263. 60

Chailyan, A., Marcatili, P. & Tramontano, A. (2011). The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *FEBS J.*, **278**, 2858–2866. 37, 38, 66, 84, 96, 104, 112, 163, 168, 185

Chailyan, A., Tramontano, A. & Marcatili, P. (2012). A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res.*, **40**, D1230–1234. 41

Chatellier, J., Van Regenmortel, M.H., Vernet, T. & Altschuh, D. (1996). Functional mapping of conserved residues located at the VL and VH domain interface of a Fab. *J. Mol. Biol.*, **264**, 1–6. 37

Choi, Y. & Deane, C.M. (2010). FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins*, **78**, 1431–1440. 151

REFERENCES

Choi, Y. & Deane, C.M. (2011). Predicting antibody complementarity determining region structures without classification. *Mol. BioSyst.*, **7**, 3327–3334. 42, 194

Chothia, C. & Lesk, A.M. (1987). Canonical Structures for the Hypervariable Regions of Immunoglobulins. *J. Mol. Biol.*, **196**, 901–917. 23, 27, 29, 30, 31, 53, 57, 58, 84, 88, 180, 181

Chothia, C., Novotný, J., Bruccoleri, R. & Karplus, M. (1985). Domain Association in Immunoglobulin: The Packing of Variable Domains. *J. Mol. Biol.*, **186**, 651–663. 34, 35, 37, 82, 162, 164, 185, 194

Chothia, C., Lesk, A., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., Colman, P.M., Spinelli, S., Alzari, P.M. & Poljak, R.J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877–883. 30, 58, 83

Chothia, C., Gelfand, I. & Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.*, **278**, 457–479. 24

Cock, P.J.a., Antao, T., Chang, J.T., Chapman, B.a., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M.J.L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423. 61, 127

Cohen, C.J., Denkberg, G., Lev, A., Epel, M. & Reiter, Y. (2003). Recombinant antibodies with MHC-restricted, peptide-specific, T-cell receptor-like specificity: new tools to study antigen presentation and TCR-peptide-MHC interactions. *J. Mol. Recognit.*, **16**, 324–332. 122

Cohen, M. & Reiter, Y. (2013). T-Cell Receptor-Like Antibodies: Targeting the Intracellular Proteome Therapeutic Potential and Clinical Applications. *Antibodies*, **2**, 517–534. 120, 122

Collins, E.J. & Riddle, D.S. (2008). TCR-MHC docking orientation: natural selection, or thymic selection? *Immunol. Res.*, **41**, 267–294. 148

Colman, P., Laver, W. & Varghese, J. (1987). Three-dimensional structure of a complex of antibody with influenza virus neuraminidase. *Nature*, **326**, 358–363. 37, 82, 83

Colman, P.M. (1988). Structure of antibody-antigen complexes: Implications for immune recognition. *Adv. Immunol*, **43**, 99–132. 37, 82

Corbett, S.J., Tomlinson, I.M., Sonnhammer, E.L., Buck, D. & Winter, G. (1997). Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. *J. Mol. Biol.*, **270**, 587–597. 20

Corrada, D., Morra, G. & Colombo, G. (2013). Investigating Allostery in Molecular Recognition: Insights from a Computational Study of Multiple Antibody-Antigen Complexes. *J. Phys. Chem. B*. 33

Dahan, R. & Reiter, Y. (2012). T-cell-receptor-like antibodies - generation, function and applications. *Expert Rev. Mol. Med.*, **14**. 122

Davies, D.R. & Metzger, H. (1983). Structural basis of antibody function. *Annu. Rev. Immunol.*, **1**, 87–117. 37, 82

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., Snoeyink, J., Richardson, J.S. & Richardson, D.C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, **35**, W375–383. 130

de Wildt, R.M., Hoet, R.M., van Venrooij, W.J., Tomlinson, I.M. & Winter, G. (1999). Analysis of heavy and light chain pairings indicates that receptor editing shapes the human antibody repertoire. *J. Mol. Biol.*, **285**, 895–901. 21

Dolittle, R.F. (1989). Of URFS and ORFS a Primer on How to Analyze Derived Amino Acid Sequences. *J. Basic Microbiol.*, **29**, 246. 153

Dostalek, M., Gardner, I., Gurbaxani, B.M., Rose, R.H. & Chetty, M. (2013). Pharmacokinetics, pharmacodynamics and physiologically-based pharmacokinetic modelling of monoclonal antibodies. *Clin Pharmacokinet*, **52**, 83–124. 122

Dunbar, J., Fuchs, A., Shi, J. & Deane, C.M. (2013). ABangle: characterising the VH-VL orientation in antibodies. *Protein Eng. Des. Sel.*, **24**, 611–620. 44, 59, 66

Dunbar, J., Knapp, B., Fuchs, A., Shi, J. & Deane, C.M. (2014a). Examining Variable Domain Orientations in Antigen Receptors Gives Insight into TCR-Like Antibody Design. *PLoS Comput. Biol.*, **10**, e1003852. 45

REFERENCES

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J. & Deane, C.M. (2014b). SAbDab: the structural antibody database. *Nucleic Acids Res.*, **42**, D1140–D1146. 32, 43, 48, 124, 155, 163

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797. 5, 55, 170

Ehrenmann, F., Kaas, Q. & Lefranc, M.P. (2010). IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.*, **38**, D301–307. 124

El-Manzalawy, Y. & Honavar, V. (2010). Recent advances in B-cell epitope prediction methods. *Immunome Res.*, **6**. 41

Elbakri, A., Nelson, P.N. & Abu Odeh, R.O. (2010). The state of antibody therapy. *Hum. Immunol.*, **71**, 1243–1250. 38

Fera, D., Schmidt, A.G., Haynes, B.F., Gao, F., Liao, H.X., Kepler, T.B. & Harrison, S.C. (2014). Affinity maturation in an HIV broadly neutralizing B-cell lineage through reorientation of variable domains. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 10275–10280. 37, 82, 160, 212

Finlay, W.J.J. & Almagro, J.C. (2012). Natural and man-made V-gene repertoires for antibody discovery. *Front. Immunol.*, **3**. 40

Foote, J. & Winter, G. (1992). Antibody framework residues affecting the conformation of the hypervariable loops. *J. Mol. Biol.*, **224**, 487–499. 37, 58, 82

Garcia, K., Teyton, L. & Wilson, I. (1999). Structural basis of T cell recognition. *Annu. Rev. Immunol.*, **17**, 369–97. 119

Garcia, K.C. & Adams, E.J. (2005). How the T cell receptor sees antigen-a structural view. *Cell*, **122**, 333–336. 148

Giudicelli, V. & Lefranc, M.p. (1999). Ontology for immunogenetics : the IMGT-ONTOLOGY. *Bioinformatics*, **15**, 1047–1054. 21, 50

Gonzalez, S.F., Degn, S.r.E., Pitcher, L.a., Woodruff, M., Heesters, B.a. & Carroll, M.C. (2011). Trafficking of B cell antigen in lymph nodes. *Annu. Rev. Immunol.*, **29**, 215–233. 14

Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.*, **4**. 48

Groves, M.A., Amanuel, L., Campbell, J.I., Rees, D.G., Sridharan, S., Finch, D.K., Lowe, D.C. & Vaughan, T.J. (2013). Antibody VH and VL recombination using phage and ribosome display technologies reveals distinct structural routes to affinity improvements with VH-VL interface residues providing important structural diversity. *MAbs*, **6**, 236–245. 40

Hall, S.J. (2003). *Basic Biomechanics*. McGraw Hill, 4th edn. 86

Hamelryck, T. & Manderick, B. (2003). PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310. 61

Harpaz, Y. & Chothia, C. (1994). Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing. *J. Mol. Biol.*, **238**, 528–539. 27

Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 10915–10919. 5, 161

Hennecke, J. & Wiley, D.C. (2001). T cell receptor-MHC interactions up close. *Cell*, **104**, 1–4. 139

Holliger, P. & Hudson, P.J. (2005). Engineered antibody fragments and the rise of single domains. *Nat. Biotechnol.*, **23**, 1126–1136. 39, 235

Holt, L.J., Herring, C., Jespers, L.S., Woolven, B.P. & Tomlinson, I.M. (2003). Domain antibodies: proteins for therapy. *Trends Biotechnol.*, **21**, 484–490. 70

Honegger, A. (2008). Engineering antibodies for stability and efficient folding. *Handb. Exp. Pharmacol.*, 47–68. 40

Honegger, A. & Plückthun, A. (2001). Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J. Mol. Biol.*, **309**, 657–670. 28, 30, 75, 155

Huang, M., Syed, R., Stura, E.a., Stone, M.J., Stefanko, R.S., Ruf, W., Edgington, T.S. & Wilson, I.a. (1998). The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor, Fab 5G9 and TF.G9 complex. *J. Mol. Biol.*, **275**, 873–894. 118

Huber, R., Deisenhofer, J., Colman, P., Masaaki, M. & Palm, W. (1976). Crystallographic structure studies of an IgG molecule and an Fc fragment. *Nature*, **264**, 415–420. 33

REFERENCES

Hülsmeyer, M., Chames, P., Hillig, R.C., Stanfield, R.L., Held, G., Coulie, P.G., Alings, C., Wille, G., Saenger, W., Uchanska-Ziegler, B., Hoogenboom, H.R. & Ziegler, A. (2005). A major histocompatibility complex-peptide-restricted antibody and t cell receptor molecules recognize their target by distinct binding modes: crystal structure of human leukocyte antigen (HLA)-A1-MAGE-A1 in complex with FAB-HYB3. *J. Biol. Chem.*, **280**, 2972–2980. 122

Igawa, T., Tsunoda, H., Kuramochi, T., Sampei, Z., Ishii, S. & Hattori, K. (2011). Engineering the variable region of therapeutic IgG antibodies. *MAbs*, **3**, 243–252. 40

Jaccard, P. (1908). *Nouvelles recherches sur la distribution florale*, vol. 44. Bull. Soc. vaud. sci. nat. 98

Janeway, C. (1992). The T cell receptor as a multicomponent signalling machine: CD4/CD8 coreceptors and CD45 in T cell activation. *Annu. Rev. Immunol.*, **10**, 645–674. 117

Janeway, C., Travers, P., Walport, M. & Shlomchik, M. (2001). *Immunobiology*. Garland Science, New York, 5th edn. 10, 11, 13, 14, 72, 117, 119

Jayaram, N., Bhowmick, P. & Martin, A.C.R. (2012). Germline VH/VL pairing in antibodies. *Protein Eng. Des. Sel.*, **25**, 523–530. 21

Johnson, G. & Wu, T.T. (2001). Kabat Database and its applications: future directions. *Nucleic Acids Res.*, **29**, 205–206. 41

Jones, B. (1972). Construction of a Three-Dimensional Model of the Polypeptide Backbone of the Variable Region of Kappa Immunoglobulin Light Chains Variable Regions of Kappa Chains. *Proc. Natl. Acad. Sci. U. S. A.*, **69**, 960–964. 31

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202. 170

Jonic, S. & Vénien-Bryan, C. (2009). Protein structure determination by electron cryo-microscopy. *Curr. Opin. Pharmacol.*, **9**, 636–642. 9

Kabat, E., Wu, T., Bilofsky, H., Reid-Miller, M. & Perry, H. (1983). Sequence of Proteins of Immunological Interest. National Institutes of Health, Bethesda. 27

Kastritis, P.L., Moal, I.H., Hwang, H., Weng, Z., Bates, P.A., Bonvin, A.M.J.J. & Janin, J. (2011). A structure-based benchmark for protein-protein binding affinity. *Protein Sci.*, **20**, 482–491. 59

Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, a., Schmeits, J.L., Wang, J. & Shimizu, N. (1997). One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res.*, **7**, 250–261. 21

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81–93. 162

Kessel, A. & Ben-Tal, N. (2012). *Introduction to proteins: structure, function, and motion*. CRC Press, New York. 8

Khalifa, M.B., Weidenhaupt, M., Choulier, L., Chatellier, J., Rauffer-Bruyère, N., Altschuh, D. & Vernet, T. (2000). Effects on interaction kinetics of mutations at the VH-VL interface of Fabs depend on the structural context. *J. Mol. Recognit.*, **13**, 127–139. 37, 82

Kirkham, P.M. & Schroeder, H.W. (1994). Antibody structure and the evolution of immunoglobulin V gene segments. *Semin. Immunol.*, **6**, 347–360. 21

Kjer-Nielsen, L., Clements, C.S., Purcell, A.W., Brooks, A.G., Whisstock, J.C., Burrows, S.R., McCluskey, J. & Rossjohn, J. (2003). A structural basis for the selection of dominant alphabeta T cell receptors in antiviral immunity. *Immunity*, **18**, 53–64. 118, 129

Knapp, B., Dunbar, J. & Deane, C.M. (2014). Large Scale Characterization of the LC13 TCR and HLA-B8 Structural Landscape in Reaction to 172 Altered Peptide Ligands: A Molecular Dynamics Simulation Study. *PLoS Comput. Biol.*, **10**, e1003748. 110, 113, 128

Köhler, G. & Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, **256**, 495–497. 39

Koonin, E., Wolf, Y. & Karev, G. (2002). The structure of the protein universe and genome evolution. *Nature*, **420**, 218–223. 8

Koonin, E.V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.*, **1**, 127–36. 8

Krawczyk, K., Baker, T., Shi, J. & Deane, C.M. (2013). Antibody i-Patch Prediction of the Antibody Binding Site Improves Rigid Local Antibody-Antigen Docking. *Protein Eng. Des. Sel.*, **26**, 621–629. 41

Krawczyk, K., Liu, X., Baker, T., Shi, J. & Deane, C.M. (2014). Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, **30**, 2288–2294. 42

REFERENCES

Krivov, G.G., Shapovalov, M.V. & Dunbrack, R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795. 130

Kunik, V., Ashkenazi, S. & Ofran, Y. (2012a). Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Res.*, **40**, W521–524. 41

Kunik, V., Peters, B. & Ofran, Y. (2012b). Structural consensus among antibodies defines the antigen binding site. *PLoS Comput. Biol.*, **8**, e1002388. 41

Kuroda, D., Shirai, H., Kobori, M. & Nakamura, H. (2008). Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins*, **73**, 608–620. 29, 30, 185, 194

Kuroda, D., Shirai, H., Jacobson, M.P. & Nakamura, H. (2012). Computer-aided antibody design. *Protein Eng. Des. Sel.*, **25**, 507–522. 40, 42, 85, 151

Lara-Ochoa, F., Almagro, J.C., Vargas-Madrazo, E. & Conrad, M. (1996). Antibody-antigen recognition: a canonical structure paradigm. *J. Mol. Evol.*, **43**, 678–684. 58

Lee, D., Grant, A., Marsden, R.L. & Orengo, C. (2005). Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins*, **59**, 603–615. 8, 208

Lee, M., Lloyd, P., Zhang, X., Schallhorn, J.M., Sugimoto, K., Leach, A.G., Sapiro, G. & Houk, K.N. (2006). Shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. *J. Org. Chem.*, **71**, 5082–5092. 33

Lefranc, M.P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V. & Lefranc, G. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77. 28, 29, 30, 31, 41, 54, 57, 155, 180, 181

Lefranc, M.P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. & Duroux, P. (2009). IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.*, **37**, D1006–1012. 21, 41, 50, 124

Lesk, A.M. & Chothia, C. (1988). Elbow motion in the immunoglobulins involves a molecular ball-and-socket joint. *Nature*, **335**, 188–190. 33

Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659. 64, 70, 88, 124, 155

Li, Y., Li, H., Smith-Gill, S. & Mariuzza, R. (2000). Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63. *Biochemistry*, **39**, 6296–6309. 83

Lippow, S.M., Wittrup, K.D. & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, **25**, 1171–1176. 42

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259. 8, 154

Lupyan, D., Leo-Macias, A. & Ortiz, A.R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**, 3255–3263. 74, 93

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 961–968. 22

MacCallum, R.M., Martin, A.C.R. & Thornton, J.M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.*, **262**, 732–745. 29, 31, 33, 57, 180, 181

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. (2013). cluster: Cluster Analysis Basics and Extensions. 58, 127

Marcatili, P., Rosi, A. & Tramontano, A. (2008). PIGS: automatic prediction of antibody structures. *Bioinformatics*, **24**, 1953–1954. 43, 152

Mareeva, T., Lebedeva, T., Anikeeva, N., Manser, T. & Sykulev, Y. (2004). Antibody specific for the peptide.major histocompatibility complex. Is it T cell receptor-like? *J. Biol. Chem.*, **279**, 44243–44249. 122

Mareeva, T., Martinez-Hackert, E. & Sykulev, Y. (2008). How a T cell receptor-like antibody recognizes major histocompatibility complex-bound peptide. *J. Biol. Chem.*, **283**, 29053–29059. 129

REFERENCES

Martin, A.C.R. (1996). Accessing the Kabat antibody sequence database by computer. *Proteins*, **25**, 130–131. 41

Martin, A.C.R. (2010). *Antibody Engineering Vol. 2*, vol. 2. Springer Berlin Heidelberg, Berlin, Heidelberg, 2nd edn. 41, 50, 145

Martin, A.C.R. & Thornton, J.M. (1996). Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J. Mol. Biol.*, **263**, 800–815. 30, 32, 42, 58

Martin, A.C.R., Cheetham, J.C. & Rees, A.R. (1989). Modeling antibody hypervariable loops: a combined algorithm. *Proc. Natl. Acad. Sci. U. S. A.*, **86**, 9268–72. 29

Massey Jr, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc.*, **46**, 68–78. 101

Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K.I., Hayashida, H., Miyata, T. & Honjo, T. (1998). The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J. Exp. Med.*, **188**, 2151–2162. 21

McBeth, C. & Seamons, A. (2008). A New Twist In TCR Diversity Revealed By A Forbidden $\alpha\beta$ TCR. *J. Mol. Biol.*, **375**, 1306–1319. 122

McCafferty, J., Griffiths, A.D., Winter, G. & Chiswell, D.J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. **348**, 552–554. 40

McCarthy, M. (2014). US signs contract with ZMapp maker to accelerate development of the Ebola drug. *BMJ*. 38

Messih, M.A., Lepore, R., Marcatili, P. & Tramontano, A. (2014). Improving the accuracy of the structure prediction of the third hypervariable loop of the heavy chains of antibodies. *Bioinformatics*, **30**, 2733–2740. 151, 194

Miyata, T., Miyazawa, S. & Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, **12**, 219–36. 161

Mizuguchi, K., Deane, C., Blundell, T., Johnson, M. & Overington, J. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623. 97

Molloy, P.E., Sewell, A.K. & Jakobsen, B.K. (2005). Soluble T cell receptors: novel immunotherapies. *Curr. Opin. Pharmacol.*, **5**, 438–443. 122

Morea, V., Tramontano, A., Rustici, M., Chothia, C., Lesk, A.M., Angeletti, M.P. & Roma, P. (1998). Conformations of the Third Hypervariable Region in the VH Domain of Immunoglobulins. *J. Mol. Biol.*, **275**, 269–294. 30, 194

Mullard, A. (2013). Maturing antibody-drug conjugate pipeline hits 30. *Nat. Rev. Drug Discov.*, **12**. 39

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y. & Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*, **102**, 553–563. 22

Nakanishi, T., Tsumoto, K., Yokota, A., Kondo, H. & Kumagai, I. (2008). Critical contribution of VHVL interaction to reshaping of an antibody: The case of humanization of anti-lysozyme antibody, HyHEL-10. *Protein Sci.*, **17**, 261–270. 37

Narayanan, A., Sellers, B.D. & Jacobson, M.P. (2009). Energy-based analysis and prediction of the orientation between light-chain and heavy-chain antibody variable domains. *J. Mol. Biol.*, **388**, 941–953. 38, 43, 83, 98, 152

Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453. 5

Neefjes, J.J., Momburg, F. & Hämmerling, G.J. (1993). Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter. *Science (80-.).*, **261**, 769–771. 117

Nemazee, D. (2000). Receptor selection in B and T lymphocytes. *Annu. Rev. Immunol.*, **18**, 1–28. 20

Neumann, F., Sturm, C., Hülsmeyer, M., Dauth, N., Guillaume, P., Luescher, I.F., Pfreundschuh, M. & Held, G. (2009). Fab antibodies capable of blocking T cells by competitive binding have the identical specificity but a higher affinity to the MHC-peptide-complex than the T cell receptor. *Immunol. Lett.*, **125**, 86–92. 122

Nikoloudis, D., Pitts, J. & Saldanha, J. (2014). A complete, multi-level conformational clustering of antibody complementarity-determining regions. *PeerJ*, **2**, e456. 30, 32

REFERENCES

North, B., Lehmann, A. & Dunbrack, R.L. (2011). A new clustering of antibody CDR loop conformations. *J. Mol. Biol.*, **406**, 228–256. 29, 30, 31, 32, 42, 58, 155, 156, 168, 170, 185, 194

Notredame, C., Higgins, D.G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217. 5

Ohno, S. (1970). *Evolution by gene duplication.*. Berlin, Heidelberg and New York: Springer-Verlag. 8

Olimpieri, P., Chailyan, A., Tramontano, A. & Marcatili, P. (2013). Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. *Bioinformatics*, **29**, 2285–2291. 41

Oliva, B., Bates, P.A., Querol, E., Avilés, F.X. & Sternberg, M.J. (1998). Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J. Mol. Biol.*, **279**, 1193–1210. 30

Orengo, C.A., Pearl, F.M., Bray, J.E., Todd, A.E., Martin, A.C.R., Lo Conte, L. & Thornton, J.M. (1999). The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279. 8

Padlan, E. (1994). Anatomy of the antibody molecule. *Mol. Immunol.*, **31**, 169–217. 24, 33

Pantazes, R.J. & Maranas, C.D. (2010). OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng. Des. Sel.*, **23**, 849–858. 42

Peters-Libeu, C., Miller, J., Rutenber, E., Newhouse, Y., Krishnan, P., Cheung, K., Hatters, D., Brooks, E., Widjaja, K., Tran, T., Mitra, S., Arrasate, M., Mosquera, L.a., Taylor, D., Weisgraber, K.H. & Finkbeiner, S. (2012). Disease-associated polyglutamine stretches in monomeric huntingtin adopt a compact structure. *J. Mol. Biol.*, **421**, 587–600. 143

Petersen-Mahrt, S., Harris, R. & Neuberger, M. (2002). AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature*, **418**, 99–103. 22

Ponomarenko, J., Papangelopoulos, N., Zajonc, D.M., Peters, B., Sette, A. & Bourne, P.E. (2011). IEDB-3D: structural data within the immune epitope database. *Nucleic Acids Res.*, **39**, D1164–1170. 41

Qiu, X., Wong, G., Audet, J., Bello, A., Fernando, L., Alimonti, J.B., Fausther-Bovendo, H., Wei, H., Aviles, J., Hiatt, E., Johnson, A., Morton, J., Swope, K., Bohorov, O., Bohorova, N., Goodman, C., Kim, D., Pauly, M.H., Velasco, J., Pettitt, J., Olinger, G.G., Whaley, K., Xu, B., Strong, J.E., Zeitlin, L. & Kobinger, G.P. (2014). Reversion of advanced Ebola virus disease in nonhuman primates with ZMapp. *Nature*, **514**, 47–53. 38

Raghunathan, G., Smart, J., Williams, J. & Almagro, J.C. (2012). Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *J. Mol. Recognit.*, **25**, 103–113. 180

Rajewsky, K. (1996). Clonal selection and learning in the antibody system. *Nature*, **381**, 751–758. 16

Ramirez-benitez, M.C. & Almagro, J.C. (2001). Analysis of antibodies of known structure suggests a lack of correspondence between the residues in contact with the antigen and those modified by somatic hypermutation. *Proteins Struct. ...*, **206**, 199–206. 180

Rammensee, H., Bachmann, J., Emmerich, N.P., Bachor, O.a. & Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219. 120

Rammensee, H.G. (1995). Chemistry of peptides associated with MHC class I and class II molecules. *Curr. Opin. Immunol.*, **7**, 85–96. 120

Rangarajan, S. & Mariuzza, R.a. (2014). T cell receptor bias for MHC: co-evolution or co-receptors? *Cell. Mol. Life Sci.*, **71**, 3059–3068. 148

Read, R. & Chavali, G. (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins*, **69**, 27–37. 84

Reichert, J.M. (2013). Antibodies to watch in 2014. *MAbs*, **6**, 5–14. 38

Retter, I., Althaus, H.H., Münch, R. & Müller, W. (2005). VBASE2, an integrative V gene database. *Nucleic Acids Res.*, **33**, D671–674. 41

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339. 8, 24, 34

REFERENCES

Riechmann, L., Clark, M., Waldmann, H. & Winter, G. (1988). Reshaping human antibodies for therapy. *Nature*, **332**, 323–327. 37, 40, 58, 82

Robinson, J., Halliwell, J.a., McWilliam, H., Lopez, R., Parham, P. & Marsh, S.G.E. (2013). The IMGT/HLA database. *Nucleic Acids Res.*, **41**, D1222–1227. 120

Rock, K.L., Gramm, C., Rothstein, L., Clark, K., Stein, R., Dick, L., Hwang, D. & Goldberg, A.L. (1994). Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. *Cell*, **78**, 761–71. 117

Rogozin, I.B. & Kolchanov, N.a. (1992). Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim. Biophys. Acta*, **1171**, 11–18. 22

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94. 153

Rudolph, M.G., Luz, J.G. & Wilson, I.a. (2002). Structural and thermodynamic correlates of T cell signaling. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 121–149. 120, 139

Rudolph, M.G., Stanfield, R.L. & Wilson, I.a. (2006). How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.*, **24**, 419–466. 119

Schrödinger, L. (2010). The PyMOL Molecular Graphics System, Version 1.3r1. 61, 110

Schroeder, H.W. & Cavacini, L. (2010). Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.*, **125**, 41–52. 14, 20, 22

Sela-Culang, I., Alon, S. & Ofran, Y. (2012). A systematic comparison of free and bound antibodies reveals binding-related conformational changes. *J. Immunol.*, **189**, 4890–4899. 33, 83

Sharpe, M.J., Milstein, C., Jarvis, J.M. & Neuberger, M.S. (1991). Somatic hypermutation of immunoglobulin kappa may depend on sequences 3' of C-kappa and occurs on passenger transgenes. *EMBO J.*, **10**, 2139–2145. 22

Shen, M.y. & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, **15**, 2507–2524. 130

Shirai, H., Kidera, A. & Nakamura, H. (1999). H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Lett.*, **455**, 188–197. 30

Shirai, H., Prades, C., Vita, R., Marcatili, P., Popovic, B., Xu, J., Overington, J.P., Hirayama, K., Soga, S., Tsunoyama, K., Clark, D., Lefranc, M.P. & Ikeda, K. (2014). Antibody informatics for drug discovery. *Biochim. Biophys. Acta*, **14**, S1570–9639. 40, 49

Sim, A.Y.L., Levitt, M. & Minary, P. (2012). Modeling and design by hierarchical natural moves. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 2890–2895. 208

Simons, K., Bonneau, R., Ruczinski, I. & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins Struct. Funct. Bioinforma.*, **176**, 171–176. 153

Sircar, A., Kim, E.T. & Gray, J.J. (2009). RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.*, **37**, W474–W479. 43

Sivasubramanian, A., Sircar, A., Chaudhury, S. & Gray, J.J. (2009). Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins*, **74**, 497–514. 38, 43, 84, 153

Smith, T. & Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197. 5

Sokal, R.R. & Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, **28**, 1409–1438. 165

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322. 170

Stanfield, R., Takimoto-Kamimura, M., Rini, J., Profy, A. & Wilson, I. (1993). Major antigen-induced domain rearrangements in an antibody. *Structure*, **1**, 83–93. 37, 66, 82, 83

Stanfield, R.L., Zemla, A., Wilson, I.a. & Rupp, B. (2006). Antibody elbow angles are influenced by their light chain class. *J. Mol. Biol.*, **357**, 1566–1574. 33, 75

Stewart-Jones, G., Wadle, A., Hombach, A., Shenderov, E., Held, G., Fischer, E., Kleber, S., Nuber, N., Stenner-Liewen, F., Bauer, S., McMichael, A., Knuth, A., Abken, H., Hombach, A.A., Cerundolo, V., Jones, E.Y. & Renner, C. (2009). Rational development of high-affinity T-cell receptor-like antibodies. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 5784–5788. 122, 124

REFERENCES

Strobl, C., Boulesteix, A.L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 1–21. 97

Strobl, C., Hothorn, T. & Zeileis, A. (2009). Party on! *R J.*, **1**, 14–17. 98

Teichmann, S.a., Murzin, A.G. & Chothia, C. (2001). Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.*, **11**, 354–363. 48

Teplyakov, A., Obmolova, G., Malia, T. & Gilliland, G. (2011). Antigen recognition by antibody C836 through adjustment of VL/VH packing. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **67**, 1165–1167. 37, 83

Teplyakov, A., Luo, J., Obmolova, G., Malia, T.J., Sweet, R., Stanfield, R.L., Kodangattil, S., Almagro, J.C. & Gilliland, G.L. (2014). Antibody modeling assessment II. Structures and models. *Proteins*, 1–20. 110, 113, 152

Therneau, T., Atkinson, B. & Ripley, B. (2014). *rpart: Recursive Partitioning and Regression Trees*. 168

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, **302**, 575–581. 180

Tramontano, A., Chothia, C. & Lesk, A. (1989). Structural determinants of the conformations of mediumsized loops in proteins. *Proteins Struct. Funct. Bioinforma.*, **6**, 382–394. 30

van der Linden, R.H., Frenken, L.G., de Geus, B., Harmsen, M.M., Ruuls, R.C., Stok, W., de Ron, L., Wilson, S., Davis, P. & Verrips, C.T. (1999). Comparison of physical chemical properties of llama VHH antibody fragments and mouse monoclonal antibodies. *Biochim. Biophys. Acta*, **1431**, 37–46. 68

van Gent, D.C., McBlane, J.F., Ramsden, D.a., Sadofsky, M.J., Hesse, J.E. & Gellert, M. (1995). Initiation of V(D)J recombination in a cell-free system. *Cell*, **81**, 925–934. 20

Vargas-Madrazo, E. & Paz-García, E. (2003). An improved model of association for VH-VL immunoglobulin domains: asymmetries between VH and VL in the packing of some interface residues. *J. Mol. Recognit.*, **16**, 113–120. 35, 36, 37, 82, 162, 185, 194

Walsh, G. (2010). Biopharmaceutical benchmarks 2010. *Nat. Biotechnol.*, **28**, 917–924. 47

Wang, R., Fang, X., Lu, Y., Yang, C.Y. & Wang, S. (2005). The PDBbind Database: Methodologies and Updates. *J. Med. Chem.*, **48**, 4111–4119. 59

Webster, D.M., Henry, A.H. & Rees, A.R. (1994). Antibody-antigen interactions. *Curr. Opin. Struct. Biol.*, **4**, 123–129. 33

Weichenberger, C.X., Pozharski, E. & Rupp, B. (2013). Visualizing ligand molecules in twilight electron density. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.*, **69**, 195–200. 56

Weinblatt, M.E., Keystone, E.C., Furst, D.E., Moreland, L.W., Weisman, M.H., Birbara, C.a., Teoh, L.a., Fischkoff, S.a. & Chartash, E.K. (2003). Adalimumab, a fully human anti-tumor necrosis factor alpha monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial. *Arthritis Rheum.*, **48**, 35–45. 40

Whitelegg, N.R.J. & Rees, A.R. (2000). WAM : an improved algorithm for modelling antibodies on the WEB. *Protein Eng.*, **13**, 819–824. 43, 152

Williams, A. & Barclay, A. (1988). The immunoglobulin superfamily-domains for cell surface recognition. *Annu. Rev. Immunol.*, **6**, 381–405. 24

Wu, T. & Kabat, E. (1970). An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.*, **132**, 211–250. 29, 53, 57, 180

Wucherpfennig, K. & Call, M. (2009). Structural alterations in peptide-MHC recognition by self-reactive T cell receptors. *Curr. Opin. Immunol.*, **21**, 590–595. 120

Yampolsky, L.Y. & Stoltzfus, A. (2005). The exchangeability of amino acids in proteins. *Genetics*, **170**, 1459–1472. 161

Yewdell, J.W., Reits, E. & Neefjes, J. (2003). Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat. Rev. Immunol.*, **3**, 952–961. 116, 117

Yin, Y., Li, Y. & Mariuzza, R.a. (2012). Structural basis for self-recognition by autoimmune T-cell receptors. *Immunol. Rev.*, **250**, 32–48. 120

Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374. 84

Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309. 93

Zhou, H. & Zhou, Y. (2002). Distancescaled, finite idealgas reference state improves structurederived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726. 212

Zolot, R.S., Basu, S. & Million, R.P. (2013). Antibody-drug conjugates. *Nat. Rev. Drug Discov.*, **12**, 259–260. 39

Appendix A

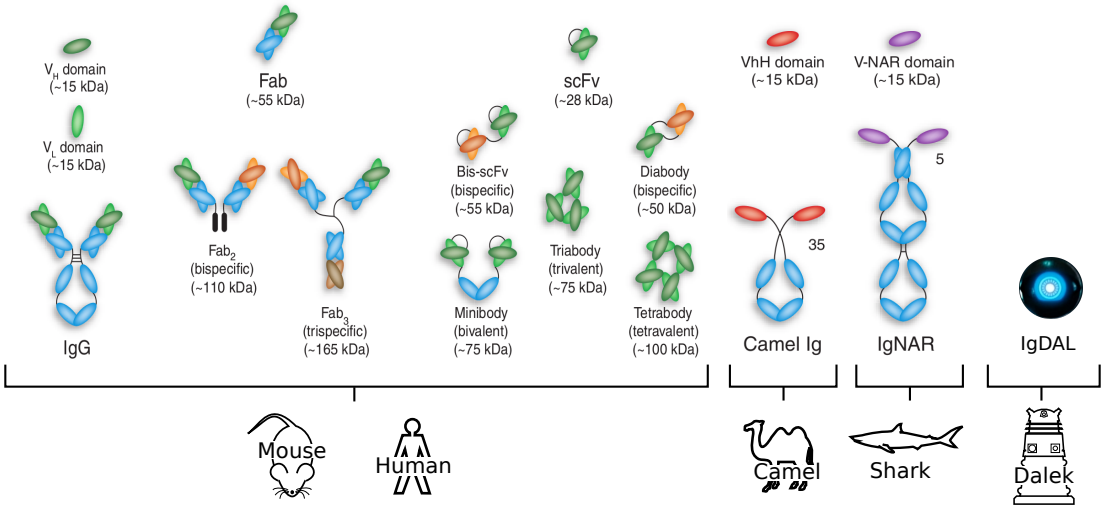


Figure A.1: Schematic representations of different antibody formats. The species from which each natural or engineered antibody is derived from is shown below the representations. IgG's described in Chapter 1 are the most abundant natural antibody in the blood serum of humans and mice. The versatile nature of the immunoglobulin domain allows other non-naturally occurring formats to be engineered. For example the Fab_2 format has non-identical variable regions and can therefore bind to two different antigens. The VH and VL domains of the scFv format are covalently linked and can be expressed as a single chain. This has advantages in the industrial production of therapeutic molecules owing to better stability properties than a similarly sized multi-chain antibody fragment. Camelids and sharks have antibodies that lack light chains. Due to their ability to bind antigens without the additional diversity brought by a VL domain, the camel heavy variable domain (VHH) is of therapeutic interest. Antibodies are thought by some to exist in extraterrestrial organisms[2]. The IgDAL of a Dalek is a particularly ferocious molecule and should be avoided at all costs. Figure adapted from Holliger & Hudson [2005]

[2]Doctor Who: Series 34, Episode two - Into the Dalek

Appendix B

Species	Subgroup	Functional genes	Source
Mouse	IGHV1	110	1
Mouse	IGHV2	5	1
Mouse	IGHV3	7	1
Mouse	IGHV4	2	1
Mouse	IGHV5	25	1
Mouse	IGHV6	4	1
Mouse	IGHV7	4	1
Mouse	IGHV8	5	1
Mouse	IGHV9	8	1
Mouse	IGHV10	4	1
Mouse	IGHV11	1	1
Mouse	IGHV12	1	1
Mouse	IGHV13	1	1
Mouse	IGHV14	4	1
Mouse	IGHV15	1	1
Mouse	IGKV1	8	2
Mouse	IGKV2	3	2
Mouse	IGKV3	9	2
Mouse	IGKV4	25	2
Mouse	IGKV5	5	2
Mouse	IGKV6	9	2
Mouse	IGKV7	1	2
Mouse	IGKV8	9	2
Mouse	IGKV9	4	2
Mouse	IGKV10	2	2
Mouse	IGKV11	1	2
Mouse	IGKV12	6	2

Mouse	IGKV13	2	2
Mouse	IGKV14	3	2
Mouse	IGKV16	1	2
Mouse	IGKV17	2	2
Mouse	IGKV18	1	2
Mouse	IGKV19	1	2
Mouse	IGLV1	2	3
Mouse	IGLV2	1	3
Mouse	IGLV3	6	3
Human	IGHV1	9	4
Human	IGHV2	3	4
Human	IGHV3	20	4
Human	IGHV4	9	4
Human	IGHV5	1	4
Human	IGHV6	1	4
Human	IGHV7	1	4
Human	IGKV1	17	5
Human	IGKV2	9	5
Human	IGKV3	6	5
Human	IGKV4	1	5
Human	IGKV5	1	5
Human	IGLV1	5	6
Human	IGLV2	5	6
Human	IGLV3	8	6
Human	IGLV4	3	6
Human	IGLV5	4	6
Human	IGLV6	1	6
Human	IGLV7	1	6
Human	IGLV8	1	6
Human	IGLV9	1	6
Human	IGLV10	1	6

Table B.1: The number of functional genes in each IMGT subgroup. The source numbers correspond to the following URLs.
1. www.imgt.org/IMGTrepertoire/LocusGenes/repertoires/mouse/IGH/IGHV/Mu_IGHVrep.html
2. www.imgt.org/IMGTrepertoire/LocusGenes/repertoires/mouse/IGK/IGKV/Mu_IGKVrep.html
3. www.imgt.org/IMGTrepertoire/LocusGenes/repertoires/mouse/IGL/IGLV/Mu_IGLVrep.html
4. www.imgt.org/IMGTrepertoire/LocusGenes/repertoires/human/IGH/IGHV/Hu_IGHVrep.html
5. www.imgt.org/IMGTrepertoire/LocusGenes/repertoires/mouse/IGK/IGKV/Mu_IGKVrep.html
6. http://www.imgt.org/IMGTrepertoire/LocusGenes/repertoires/mouse/IGL/IGLV/Mu_IGLVrep.html

Appendix C

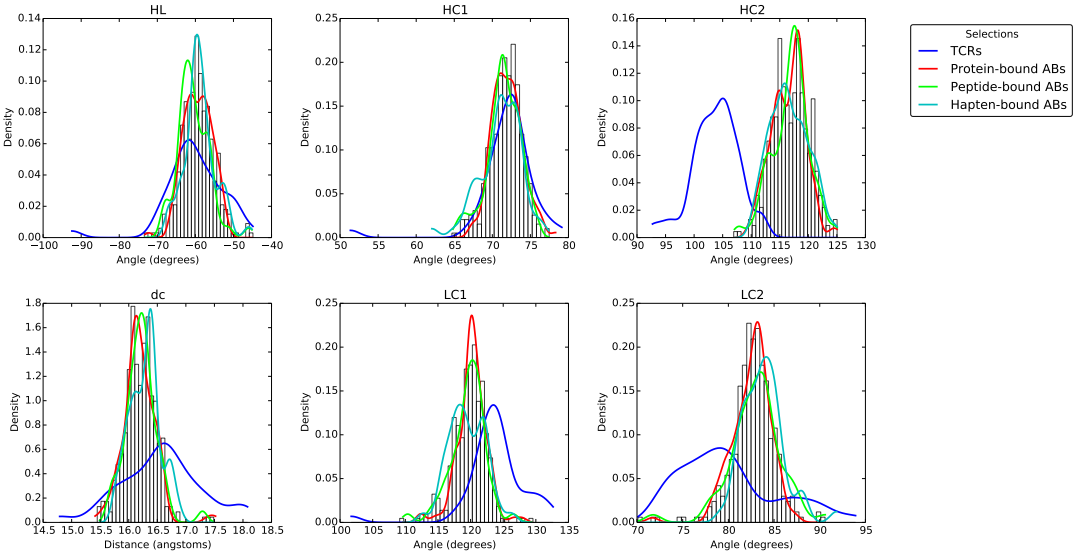


Figure C.1: The distributions of ABangle orientation measures for TCRs and antibodies stratified by the type of antigen they are bound to (Protein, Peptide or Hapten). The histograms show the background distribution for general antibodies. No preference in orientation is found for any of the three antibody groups. Thus, the antigen type is found not to determine the absolute VH-VL orientation. No group is any more similar to the TCR distribution than to the general antibody distribution.

Appendix C

Vα coreset	Vβ coreset
16	5
17	6
18	8
20	9
22	10
38	11
39	12
40	14
41	15
42	16
43	22
44	23
45	24
52	38
85	39
86	40
88	41
89	42
90	43
94	44
95	51
96	52
97	53
98	86
99	99
100	100
101	101
102	102
103	103
104	104

Table C.1: The Vα and Vβ coreset IMGT positions used for the TCR ABangle procedure.

PDB	Heavy chain	Light chain
3o11	H	L
1kb9	J	K
3qeg	H	L
3qos	B	A
43c9	F	E
3o2d	H	L
1iqw	H	L
1fpt	H	L
1ind	H	L
3eyq	D	C
1rjl	B	A
3h42	H	L
2e27	H	L
1sy6	H	L
1rur	H	L
3c08	H	L
1yee	H	L
3sob	H	L
1kel	H	L
3nh7	I	M

Table C.2: The Fv structures used as decoys to change the orientation of the native complex.

PDB	α chain	β chain
3utt	I	J
3ffc	D	E
3kpr	I	J
1ymm	D	E
1nfd	A	B
3dx9	A	B
2ial	A	B
3qiw	C	D
3pl6	C	D
3mff	A	B
2esv	D	E
3rgv	A	B
3qib	C	D
1fo0	A	B
2uwe	L	M
2xna	A	B
3he6	C	D
3he7	C	D
3o6f	C	D
3rev	A	B

Table C.3: The TCR structures used as decoys to change the orientation of the native complex.

Structure
3mlw
2qhr
4isv
2b1h
3ffd
3s96
2otu
4ht1
4jo2

Table C.4: Antibodies with length 13 CDR L3 loops

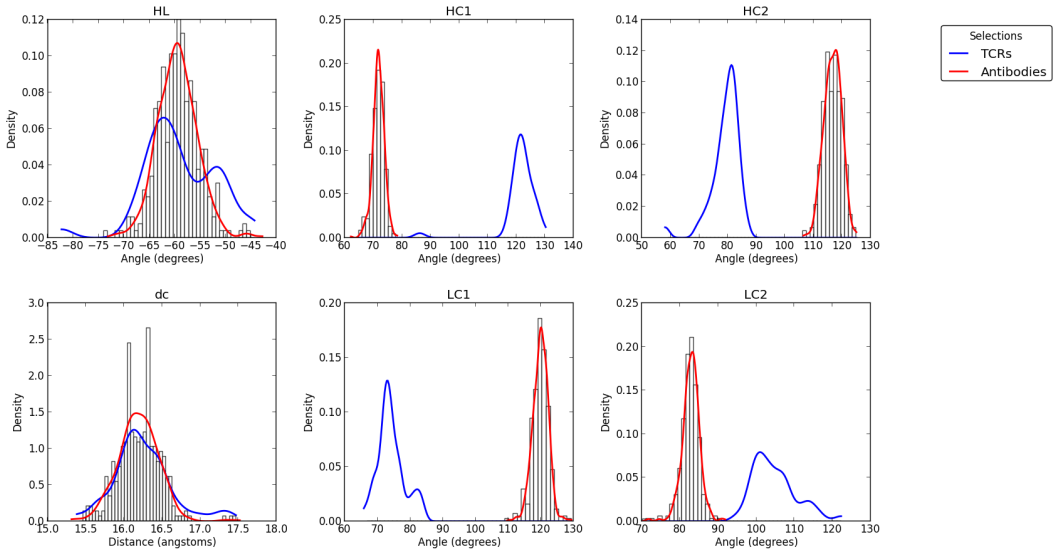


Figure C.2: The absolute measures of the Vα-Vβ orientation (TCRs) compared with the VH-VL orientation space (antibodies) when the Vα-VH, Vβ-VL domain equivalence is used. Here, we observe that none of the bend angles (HC1, HC2, LC1 and LC2) are in the range of antibody orientations.

Appendix D

H index	H insertion code	L index	L insertion code	Frequency (N=509)
-1	cdrh3	45		11
2	cdrh3	-4	cdrl3	11
3	cdrh3	56		11
-8	cdrh3	30		11
-4	cdrh3	33		11
3	cdrh3	50		11
5	cdrh3	53		11
-7	cdrh3	46		11
-7	cdrh3	31		11
4	cdrh3	55		11
1	cdrh3	55		11
-4	cdrh3	53		11
-7	cdrh3	-4	cdrl3	11
-5	cdrh3	31		12
-5	cdrh3	4	cdrl3	12
56		-4	cdrl3	12
45		100		12
-5	cdrh3	56		12
44		4		12
5	cdrh3	46		12
7	cdrh3	50		12
-3	cdrh3	56		13
4	cdrh3	-2	cdrl3	13
-2	cdrh3	49		13
43		101		13
37		0	cdrl3	13
103		0	cdrl3	14

-8	cdrh3	3	cdrl3	14
5	cdrh3	55		14
-8	cdrh3	-2	cdrl3	14
-5	cdrh3	53		14
-7	cdrh3	34		14
-7	cdrh3	3	cdrl3	15
62		-3	cdrl3	15
-6	cdrh3	4	cdrl3	15
-3	cdrh3	49		16
2	cdrh3	32		16
7	cdrh3	32		16
-7	cdrh3	53		16
37		-2	cdrl3	17
3	cdrh3	34		17
60		-1	cdrl3	17
61		-5	cdrl3	17
-4	cdrh3	54		17
35		0	cdrl3	18
-8	cdrh3	31		18
6	cdrh3	32		18
2	cdrh3	46		18
6	cdrh3	50		19
-6	cdrh3	-4	cdrl3	20
5	cdrh3	2	cdrl3	20
44		101		21
50		-3	cdrl3	21
-8	cdrh3	49		21
43		103		22
89		42		22
-1	cdrh3	46		22
-8	cdrh3	50		23
6	cdrh3	49		23
4	cdrh3	32		23
-2	cdrh3	44		23
35		-4	cdrl3	24
-2	cdrh3	56		24
-6	cdrh3	53		25
-5	cdrh3	3	cdrl3	25

-3	cdrh3	45		25
-6	cdrh3	31		26
33		-2	cdrl3	26
-5	cdrh3	1	cdrl3	26
3	cdrh3	-2	cdrl3	28
-6	cdrh3	55		28
52		-4	cdrl3	28
104		44		28
39		42		28
-6	cdrh3	3	cdrl3	28
52		-2	cdrl3	29
33		-4	cdrl3	29
-4	cdrh3	48		29
-7	cdrh3	-2	cdrl3	29
3	cdrh3	2	cdrl3	29
-5	cdrh3	-4	cdrl3	30
35		2	cdrl3	30
4	cdrh3	50		31
-4	cdrh3	56		32
58		2	cdrl3	33
2	cdrh3	36		33
58		-5	cdrl3	33
-1	cdrh3	56		33
4	cdrh3	2	cdrl3	33
-8	cdrh3	32		34
91		41		34
2	cdrh3	34		35
-4	cdrh3	32		36
44		98		37
4	cdrh3	49		37
-8	cdrh3	2	cdrl3	38
-6	cdrh3	46		38
5	cdrh3	32		38
-3	cdrh3	55		38
50		2	cdrl3	42
43		38		43
2	cdrh3	0	cdrl3	45
-5	cdrh3	55		45

-2	cdrh3	45		47
3	cdrh3	55		48
62		1		48
58		-2	cdrl3	48
-7	cdrh3	49		49
60		-4	cdrl3	50
43		85		53
5	cdrh3	50		57
47		2	cdrl3	57
-6	cdrh3	-2	cdrl3	60
45		36		61
3	cdrh3	49		62
61		-4	cdrl3	65
-7	cdrh3	32		68
3	cdrh3	46		69
-4	cdrh3	-2	cdrl3	70
-7	cdrh3	2	cdrl3	73
5	cdrh3	49		73
-7	cdrh3	50		76
-3	cdrh3	2	cdrl3	78
60		1		79
43		100		85
39		44		88
-5	cdrh3	46		90
105		41		90
58		-3	cdrl3	94
-4	cdrh3	50		98
-1	cdrh3	55		98
2	cdrh3	2	cdrl3	99
-4	cdrh3	55		99
47		0	cdrl3	100
-2	cdrh3	36		104
105		42		105
-5	cdrh3	36		105
61		1		107
60		-2	cdrl3	108
-6	cdrh3	34		111
106		43		112

47		-1	cdrl3	113
-3	cdrh3	34		114
44		99		122
-5	cdrh3	32		138
59		-3	cdrl3	139
-5	cdrh3	50		141
59		-4	cdrl3	144
-6	cdrh3	32		154
-6	cdrh3	49		172
2	cdrh3	-2	cdrl3	172
50		-4	cdrl3	178
-5	cdrh3	0	cdrl3	180
-6	cdrh3	2	cdrl3	183
-5	cdrh3	-2	cdrl3	185
61		-3	cdrl3	187
-5	cdrh3	49		188
-6	cdrh3	50		191
45		99		205
-4	cdrh3	2	cdrl3	224
-4	cdrh3	0	cdrl3	229
50		-2	cdrl3	229
-5	cdrh3	34		263
-2	cdrh3	55		279
103		45		284
103		46		291
-5	cdrh3	2	cdrl3	297
35		-2	cdrl3	304
-3	cdrh3	-2	cdrl3	307
60		-3	cdrl3	342
-3	cdrh3	98		353
45		38		356
44		100		356
-3	cdrh3	0	cdrl3	360
-4	cdrh3	36		370
58		-4	cdrl3	379
43		87		388
-4	cdrh3	49		390
47		-4	cdrl3	394

-4	cdrh3	46		394
-4	cdrh3	34		400
44		87		401
105		43		431
-3	cdrh3	36		433
103		43		458
91		43		459
104		43		461
47		-3	cdrl3	465
39		87		465
-3	cdrh3	46		468
103		98		472
45		44		473
91		42		478
91		38		481
46		98		490
-2	cdrh3	46		496
47		98		496
37		98		502
91		44		502
103		36		503
45		87		504
39		38		506
47		-2	cdrl3	506
103		44		507
45		98		508

Table D.1: VH-VL contact position pairs and there frequencies in the non-redundant set of structures. Positions are labelled using the Chothia numbering scheme except for those in CDRH3 and CDRL3. Instead these positions are labelled from their respecitive anchor positions.

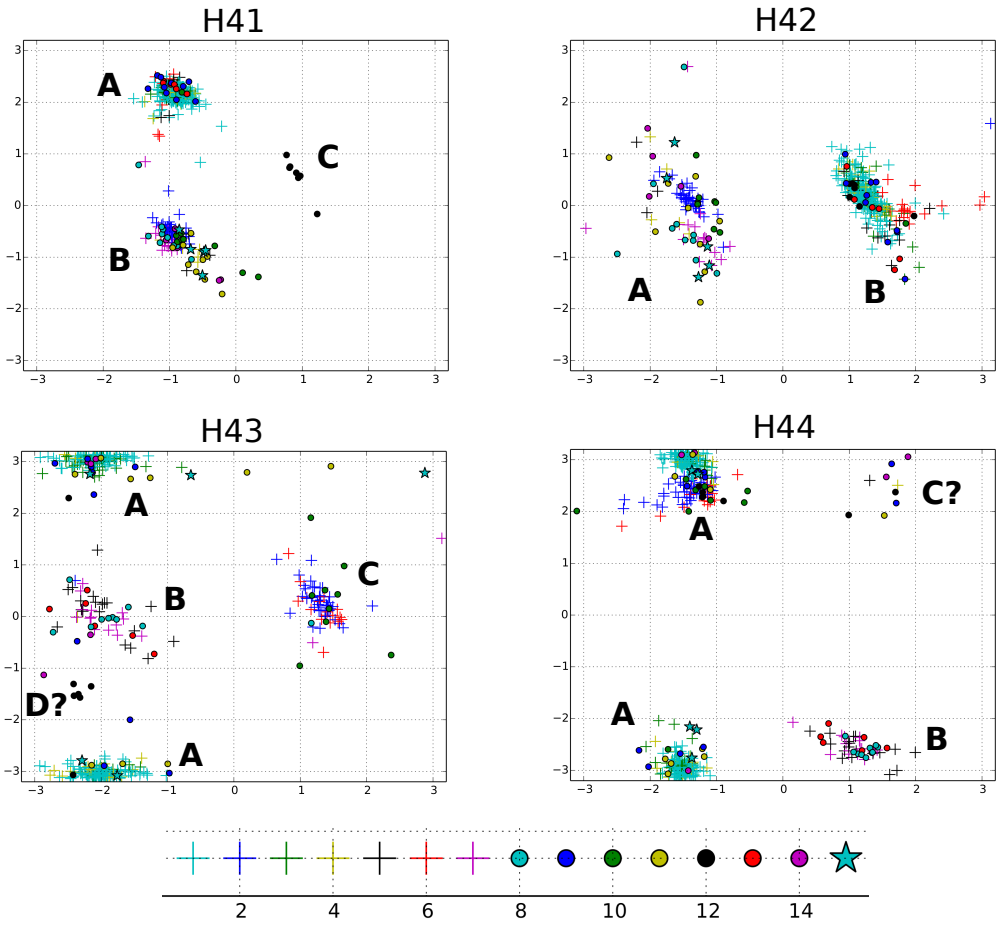


Figure D.1: Ramachandran plots for the Hifw-loop residues that discriminate between RMSD-clusters. The dihedral angles of each structure are shown according to the RMSD-cluster that they belong to. The most populated is cluster 1 and the least cluster 15. Only those clusters with 5 or more members are shown. Manual classifcation of the diherdral conformations are performed and each RMSD cluster given a dihedral string. For example cluster 1 has a string A-B-A-A whilst cluster 2 has a string B-A-C-A.

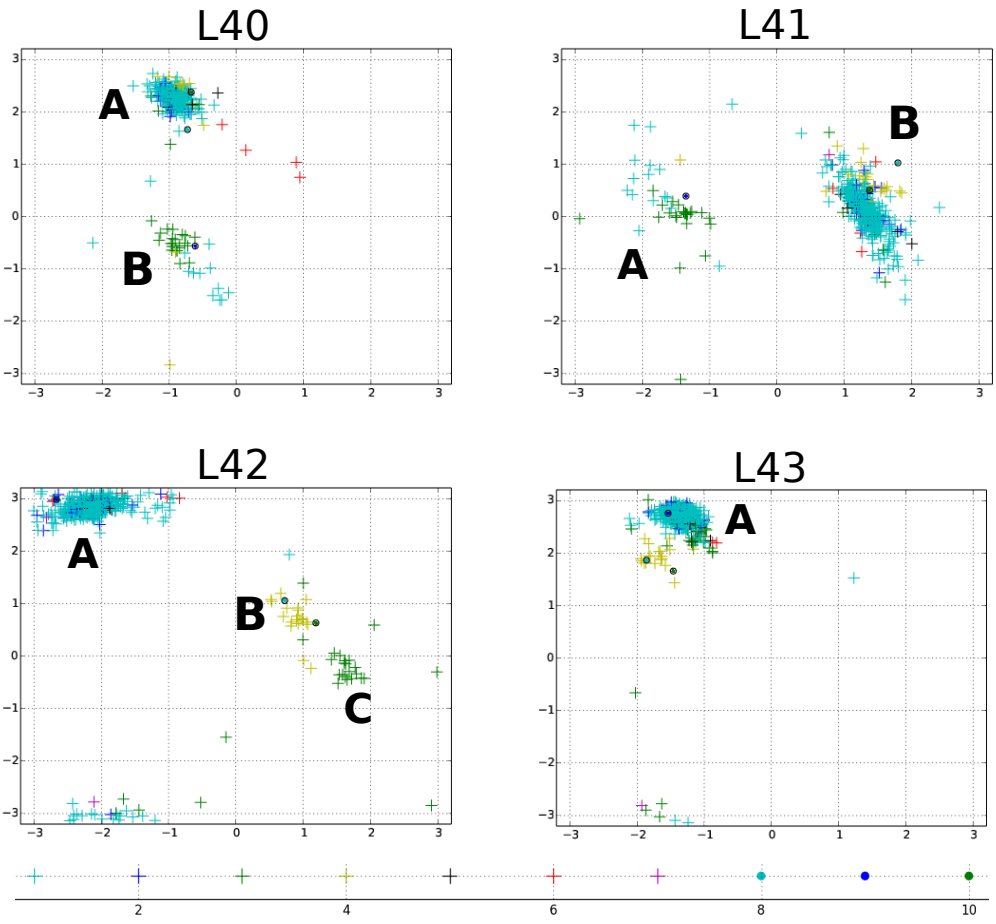


Figure D.2: The corresponding plot to Figure D.1 for the Lifw-loop.

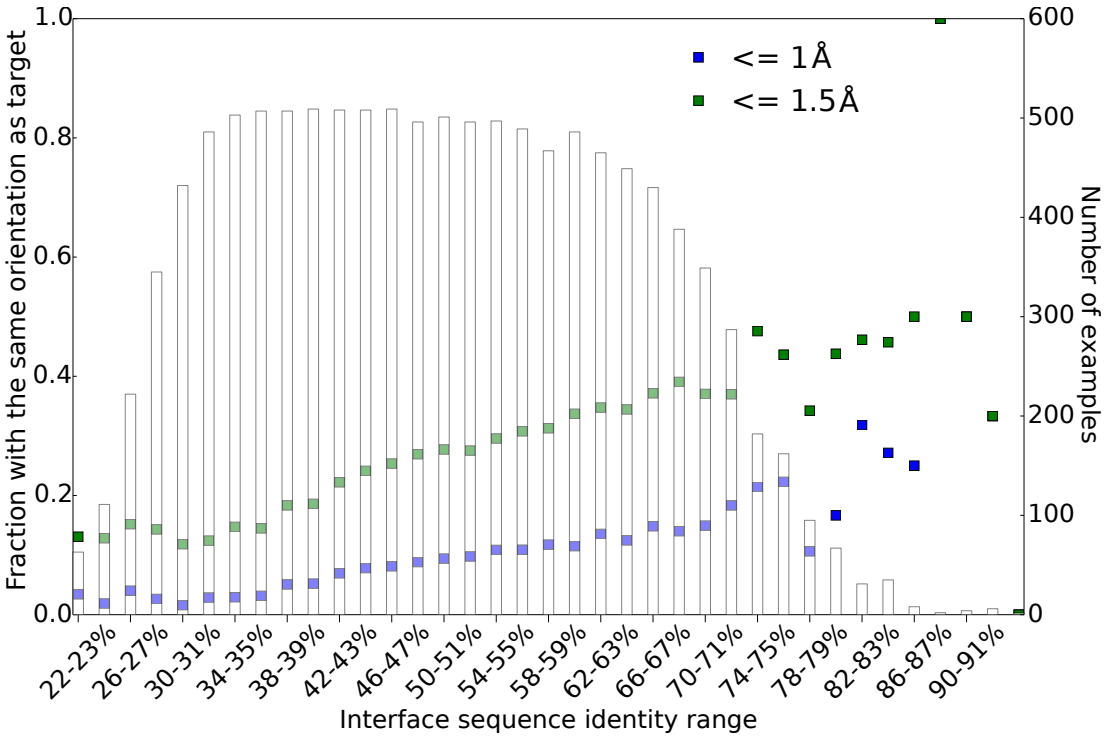


Figure D.3: The relationship between sequence identity over interface positions and the prevalence of good templates. This figure is equivalent to Figure 5.6 but only interface positions, as defined in Section 5.2.3, are used in the calculation of pairwise sequence identity. All pairs with a full sequence identity of over 85% are removed from the analysis. We find little enrichment for detecting good templates when only using the interface residues when this maximum sequence identity threshold is imposed.