Using structural fragments to design antibody binding sites

Cristian Regep

Jesus College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2017

Abstract

Antibodies are an essential part of the immune system. They are able to attain high specificity and affinity to almost any antigen, and through their modularity make a robust framework for protein engineering. As a result, development of therapeutic antibodies has grown rapidly and they now account for the majority of revenue in the sales of new bio-therapeutics.

The established design methods are heavily based on experimental simulation of the antibody maturation process outside of the human organism, and make little use of rational computational design methods. In this thesis we analysed the applicability and issues surrounding an existing modelling paradigm and developed a novel approach towards de novo antibody design.

The majority of the affinity and specificity of antibodies is modulated by a set of six binding loops called the Complementarity Determining Region (CDR). We analysed the proportion of antibody CDRs that can be modelled in a set of 15 million antibody sequences. Out of all the CDRs H3 is the only one where we observed a large number of sequences that can not be modelled accurately. We then further explored why CDR H3 is so hard to model and found that the reason is not method dependent, but is a result of a lack of suitable structures. The H3 CDR shows unique structural characteristics at full loop, four residue and single residue levels of granularity.

On the topic of de-novo antibody design we developed SAbDesigner, an automated computational pipeline for designing antibodies by mimicking the binding interface of a receptor of the target protein on the CDR. SAbDesigner automatically identifies important loops for binding on the receptor and then identifies an antibody framework that is able to accommodate it through loop grafting. The designs are then computationally validated, further refined using both point mutations, and optimisations of other CDRs. Through this pipeline we proposed a set of 15 novel antibodies that target Interleukin-5, an important therapeutic target, and also performed an applicability study across other major important therapeutic targets.

Using structural fragments to design antibody binding sites

Cristian Regep

Jesus College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2017

To *purpose*. I hope one day you will find your way to me.

Acknowledgements

Jesus, where to start!? I have many people to thank for such a small thesis. But they all thoroughly deserve it.

First of all I would like to thank Charlotte who had the idea for this project. She has been incredibly helpful, patient, and understanding with me as I was progressing in what was an entirely new field. I would also like to thank UCB, GSK, Roche and Medimmune who sponsored this project and who have provided valuable input and discussion, especially GSK which is kindly running complex validation experiments for our suggested antibody designs.

Next, I would like to acknowledge and thank my family. Without them and their support I would not be here today, or the person I am now. It is remarkable how I can trace many of the traits which make me who I am to my parents and family. I discover more and more day by day. Special mentions go to my godfather Dorin Vingan, the only academic in my extended family, who at the time of the original submission of this thesis was battling cancer (unfortunately a week later he passed away). Because of the timing I like to believe that my future lined-up job in Cancer Research is not coincidental. I will use this memory to help me refocus whenever I feel that things are hard and not working out.

I would also like to acknowledge all the current and past members of OPIG that I have had the pleasure to meet and work with. Thank you for being such great company and also good scientific advisors. I would like to mention Eleanor, Jaro, and Jin with who I've shared many profound conversations ranging from the meaning of life to the latest advances in running power meter technology. I will really miss these.

Throughout my time in Oxford I have gone through quite a transition, when I started I was the proverbial couch potato and by the end I somehow picked up long distance triathlon and cycling time trials. There have been many people from who I have drawn inspiration through this transition, or who have pulled me out of particularly difficult and bleak moments relating to injury or apathy. For this I would like to thank Eleanor, Jin, Martin, Jenny, Ben, members of JCBC (the number of the first 4 chapters are coloured green #bleedgreen) and members of OUCC. Jin gets a special mention here for making me cycle to Cambridge three weeks post-surgery, hands down the stupidest and best thing I have ever done (aside maybe from this DPhil).

Aside from Jesus College during my time in Oxford I have also been attached to Merton College as Deputy Principal of the Postmasters (the coolest title I have ever had). This is acknowledged through the maroon colour for the last chapter number. I would like to thank all the members of staff for being such good and helpful colleagues, with special mentions to Jenny, Mark and Jonathan.

Last but not least, the people outside the Oxford circle. I would like to thank Tudor and Iulia (or Alina as she is known otherwise) for showing great stoicism in listening to my senseless rants. Tudor should get a medal for what he had to endure for more than a year (chapeau!). I would also like to acknowledge two school teachers, Tache Marian and Petrean Diana, who on a number of occasions have pushed me to get out of the apathy and idleness of school and do extra work. Looking back this has given me invaluable skills at an early age, which have served me well in adult life.

Finally, I would like to leave a historical note for future generations in this thesis about Brexit through the quotes at the top of the first page of each chapter. As I am an EU student who intends to settle in the UK in the past year I have been oscillating between feelings of anxiety and rage. The quotes were carefully selected to reflect how superficially the major politicians have treated the entire Brexit debate, the most important decision for the UK in recent times.

Abstract

Antibodies are an essential part of the immune system. They are able to attain high specificity and affinity to almost any antigen, and through their modularity make a robust framework for protein engineering. As a result, development of therapeutic antibodies has grown rapidly and they now account for the majority of revenue in the sales of new bio-therapeutics.

The established design methods are heavily based on experimental simulation of the antibody maturation process outside of the human organism, and make little use of rational computational design methods. In this thesis we analysed the applicability and issues surrounding an existing modelling paradigm and developed a novel approach towards de novo antibody design.

The majority of the affinity and specificity of antibodies is modulated by a set of six binding loops called the Complementarity Determining Region (CDR). We analysed the proportion of antibody CDRs that can be modelled in a set of 15 million antibody sequences. Out of all the CDRs H3 is the only one where we observed a large number of sequences that can not be modelled accurately. We then further explored why CDR H3 is so hard to model and found that the reason is not method dependent, but is a result of a lack of suitable structures. The H3 CDR shows unique structural characteristics at full loop, four residue and single residue levels of granularity.

On the topic of de-novo antibody design we developed SAbDesigner, an automated computational pipeline for designing antibodies by mimicking the binding interface of a receptor of the target protein on the CDR. SAbDesigner automatically identifies important loops for binding on the receptor and then identifies an antibody framework that is able to accommodate it through loop grafting. The designs are then computationally validated, further refined using both point mutations, and optimisations of other CDRs. Through this pipeline we proposed a set of 15 novel antibodies that target Interleukin-5, an important therapeutic target, and also performed an applicability study across other major important therapeutic targets.

Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in the text.

Cristian Regep
4th October 2017

Contents

List of Figures

List of Tables

List of Abbreviations

AA-constrained . NGS dataset comprising unique full chain sequences
ANARCI . . . . . Antigen receptor numbering and receptor classification
ASA . . . . . Accessible surface area
BSA . . . . . Buried surface area
CAS . . . . . Computational alanine scanning
CDR . . . . . Complementarity–determining region
CH, CL . . . . . Constant heavy, constant light
cryo-EM . . . . . Electron microscopy
ΔG, ΔΔG . . . . . Change in free energy, or change in free energy changes
DNA . . . . . Deoxy–ribonucleic acid
ESS . . . . . Environment–specific substitution
ETD . . . . . Extended target database
Fab . . . . . Antigen–binding fragment
Fc . . . . . CH2 and CH3 constant domains
FFT . . . . . Fast Fourier Transform
FPR . . . . . False positive rate
Fv . . . . . Variable fragment
IC50 . . . . . Half maximal inhibitory concentration
Ig . . . . . Immunoglobulin, Antibody
IL-5 . . . . . Interleukin-5
IMGT . . . . . International Immunogenetics Information System
K_D . . . . . Dissociation constant
MD . . . . . Molecular dynamics
MHC . . . . . Major Histocompatibility Complex
NGS . . . . . Next–generation sequencing

NMR . . . . . Nuclear magnetic resonance

Non-Ig . . . . . Non–Antibody

PDB . . . . . Protein Data Bank

RMSD . . . . . Root–mean square deviation

RNA . . . . . Messenger ribonucelic acid

SAbDab . . . . . Structural antibody database;

TPR . . . . . True positive rate

VH,VL . . . . . Variable heavy, variable light

Brexit means brexit

— Theresa May MP

Context: I am an EU citizen living in the UK

1 Introduction

Contents

In the 20th century the most important drug therapies came in the form of small compounds and vaccines. Their development was of incredible importance, helping eradicate or control diseases like polio, smallpox and many bacterial infections. In the current century we face new challenges, with many of the diseases at the front line of research not being caused by a foreign organism, but instead by over-expression or under-expression of human proteins (e.g. PKD, forms of asthma), autoimmune response (e.g. rheumatoid arthritis, MS) or by uncontrolled growth of human cells (i.e. cancer). These diseases are not easily tackled by the classic approach of small molecule compounds.

In the last few decades interest has gradually shifted focus towards our own immune system, and how it can be tweaked to help tackle disease. One area that has proved successful is the development of therapeutic antibodies. With their high modularity, low off target effects and ability to recruit other parts of the immune system, antibodies are a versatile source of bio-therapeutics. However, antibody design methods that have produced the majority of antibody bio-therapeutics to date have relied on simulating the random process of affinity maturation either in animal model or *in vitro* through a phage display library. Over the last few years computational methods have become more widely used in antibody therapy development. In this thesis we analyse one of these methods and also propose a fully automated computational pipeline for designing antibodies. Our approach to designing is novel, mimicking the binding mode of a known receptor of the target on the antibody scaffold.

The main topic of this thesis is therefore antibodies, but they are just one of many types of proteins. Therefore, this chapter will firstly introduce generic background information about all proteins and their structure. Next, antibodies and their specific characteristics are presented, along with the immune system and their role and generation during the process of immune response. Then, as the specific focus of the thesis is the topic of computational antibody design, the

existing methods are detailed, both experimental and computational. Finally, the structure of the thesis is summarised.

1.1 Proteins

1.1.1 Introduction

Proteins are macromolecules found within cells and the blood stream, and the majority of functions of an organism are performed by them. In a single cell more than half the dry weight is taken up by proteins (Milo et al., 2009). Human proteins generally vary in length between 50 and 2000 amino acids, with the median human protein containing 346 amino acids (Milo et al., 2009). An antibody is slightly longer than the median protein, with the longest protein chain in an antibody having between 450-550 amino acids (Milo et al., 2009).

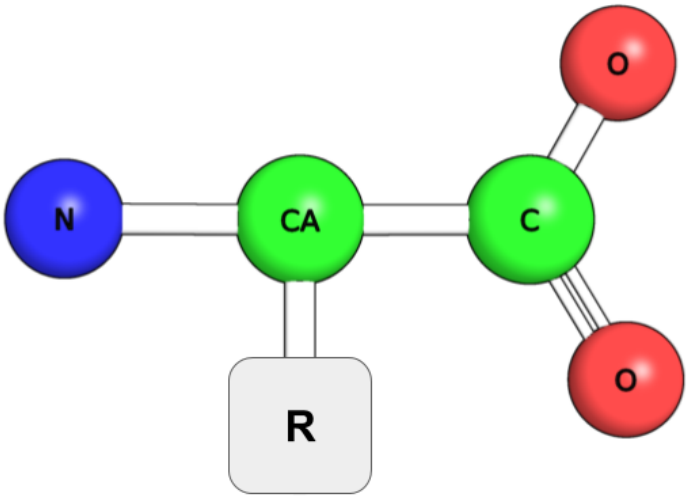

Figure 1.1: Ball and stick model of an amino acid. Carbon atoms are coloured green, oxygen atoms are coloured red, and nitrogen atoms are coloured blue. For increased clarity the hydrogen atoms have been omitted.
An amino acid (A) has a Cα atom (CA) bonded to an amine group and a carboxyl group. R denotes the side-chain, which varies between the different amino acid types (see Figure 1.3)

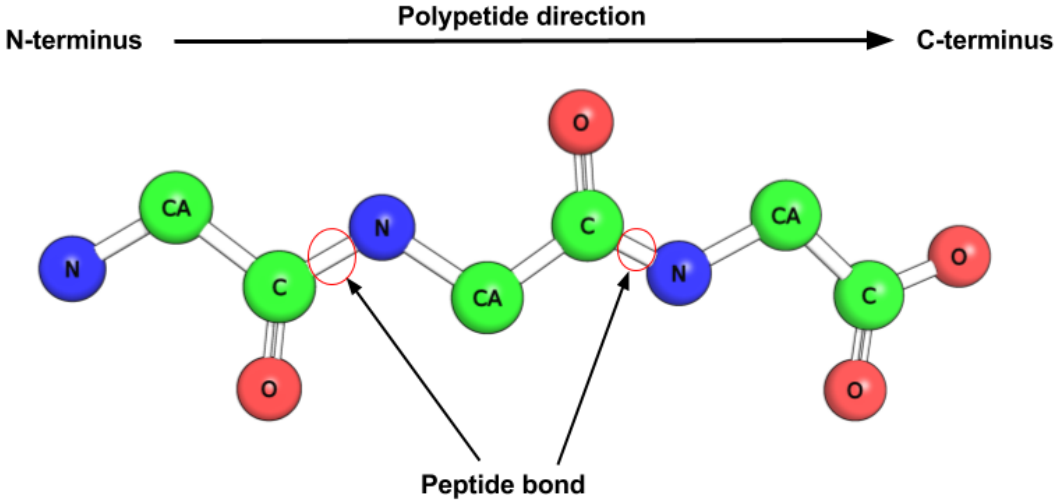

Figure 1.2: A three amino acid polypetide connected by two peptide bonds. Side-chains have been omitted for clarity. The peptide bond is found between the CO group of an amino acid and the NH group of the subsequent amino acid. During the formation of the peptide bond the single bonded oxygen is lost. Polypetides have directionality, which is shown with the black arrow, from the terminus with the amine (N-terminus) to the terminus with the carboxyl (C-terminus)

1.1.2 Amino acids

Amino acids are the building blocks of proteins, a protein being formed of a sequence of amino acids. An amino acid is an organic compound composed of an amine, a carboxyl, a side-chain and a hydrogen connected by a central carbon called the Cα (see Figure 1.1). The connected N-Cα-C-O region is called the backbone of the amino acid. The sequence of amino acids of a protein are linked to each other through a peptide bond (see Figure 1.2)

The side-chain varies between the different amino acid types (see Figure 1.3). The vast repertoire of protein functionality is generated generated through the variation in side-chains of the amino acids. These affect the shape a protein takes and the type of interactions they make with other proteins. Side-chains can be classified as hydrophobic, polar, positively charged and negatively charged (see Figure 1.3).

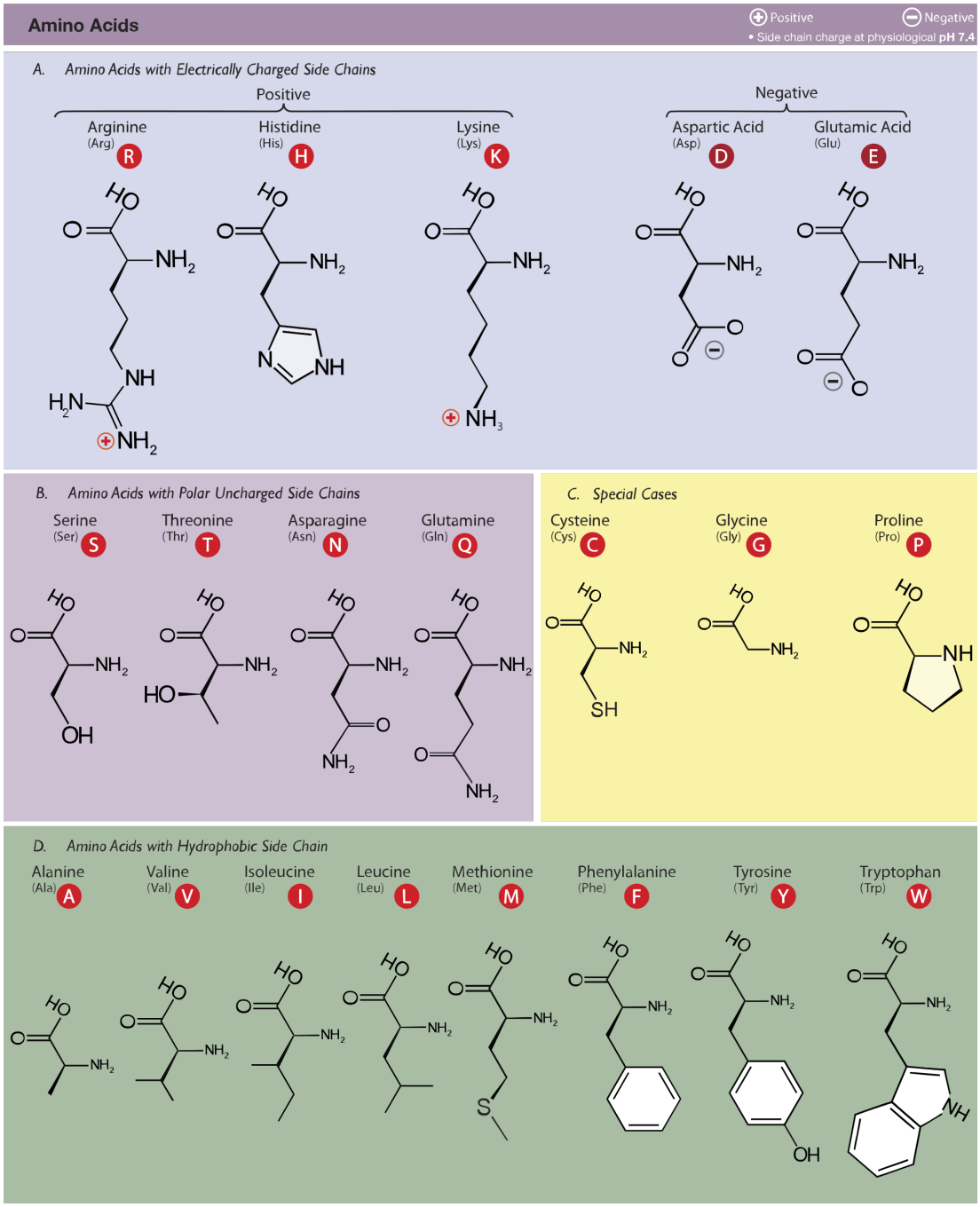

Figure 1.3: The 20 standard amino acids types, grouped by the chemical properties of their side-chains. For each of the side-chains the skeletal formula is presented. They are split into charged (positively and negatively), polar uncharged and neutral (hydrophobic). Cysteine, proline and glycine have characteristics unique from the other groups and have been categorised separately in 'Special cases'. (The image is adapted from an original image created by Dan Cojocari of University of Toronto released under the Creative commons licence 3.0)

1.1.3 Folding

As the amino acids of a protein are translated by a ribosome the protein starts to fold into a three-dimensional structure (Fedorov and Baldwin, 1997). The amino acids that are hydrophobic will generally tend to pack against each other in the core of the protein avoiding solvent, while polar residues will generally prefer the surface of the protein being more exposed to solvent (Munson et al., 1996).

1.1.3.1 Dihedral angles and Ramachandran plots

The structure of a protein can be defined as the combination of all of the degrees of freedom of its amino acids. These degrees of freedom are found in the polypeptide backbone, and are generated by rotatable bonds. The only rotatable bonds are the N-Cα and the Cα-C single bonds. C-O is a double bond, and the polypeptide bond resonates with the C-O bond switching between a single and a double bond, which prevents it from rotating (Pauling et al., 1951). Therefore, for an individual amino acid the dihedral angles of the N-Cα single bond (the Phi angle) and the Cα-C single bonds (the Psi angle) modulate all its possible backbone conformations (see Figure 1.4A).

The combinations of Phi and Psi angles for the residues of a protein can be scattered on a two dimensional grid. This type of plot is called a Ramachandran plot (see Figure 1.4B). From the available structures Ramachandran et al. (1963) observed that amino acids have preferred dihedral combinations which are energetically more favourable.

1.1.3.2 Secondary structure

Through folding proteins create regular shapes called *secondary structure*. The creation of secondary structure is defined by hydrogen patterns between backbone NH and CO groups which stabilise the shapes. The two main types of secondary structure are **α-helices** and **β-sheets** (Pauling et al., 1951).

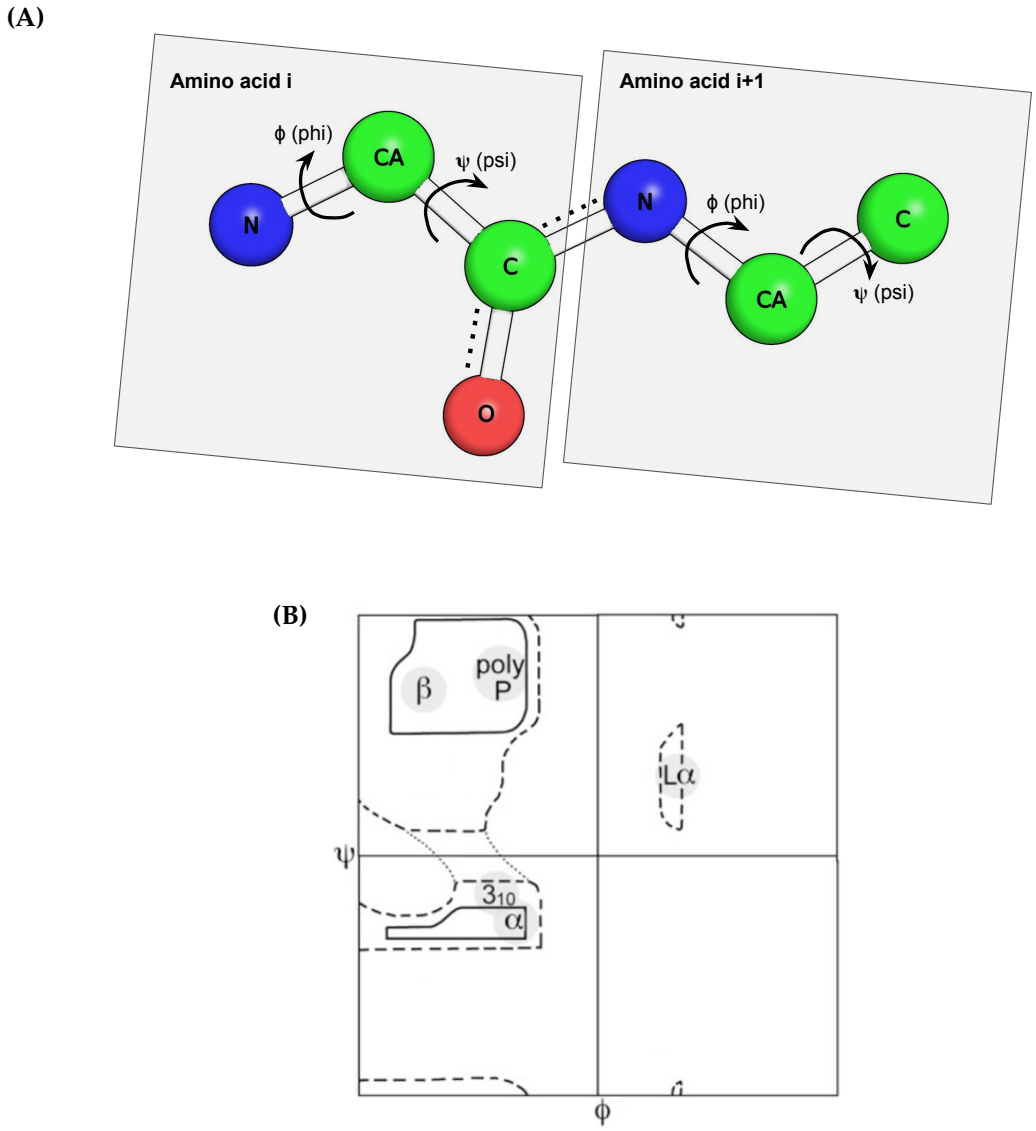

Figure 1.4: (A) Two amino acids with their phi and psi dihedral angles highlighted. The resonance between the C-O double bond and the peptide bond is highlighted with black dots.
(B) A Ramachadran plot showing the ϕ-ψ (phi-psi) dihedral angle combinations preferred by proteins. The higher density areas are outlined with dotted lines, and the areas which are associated with types of secondary structure are outlined by solid lines. (Image author: Jane Shelby Richardson. Released under Creative Commons 3.0 licence)

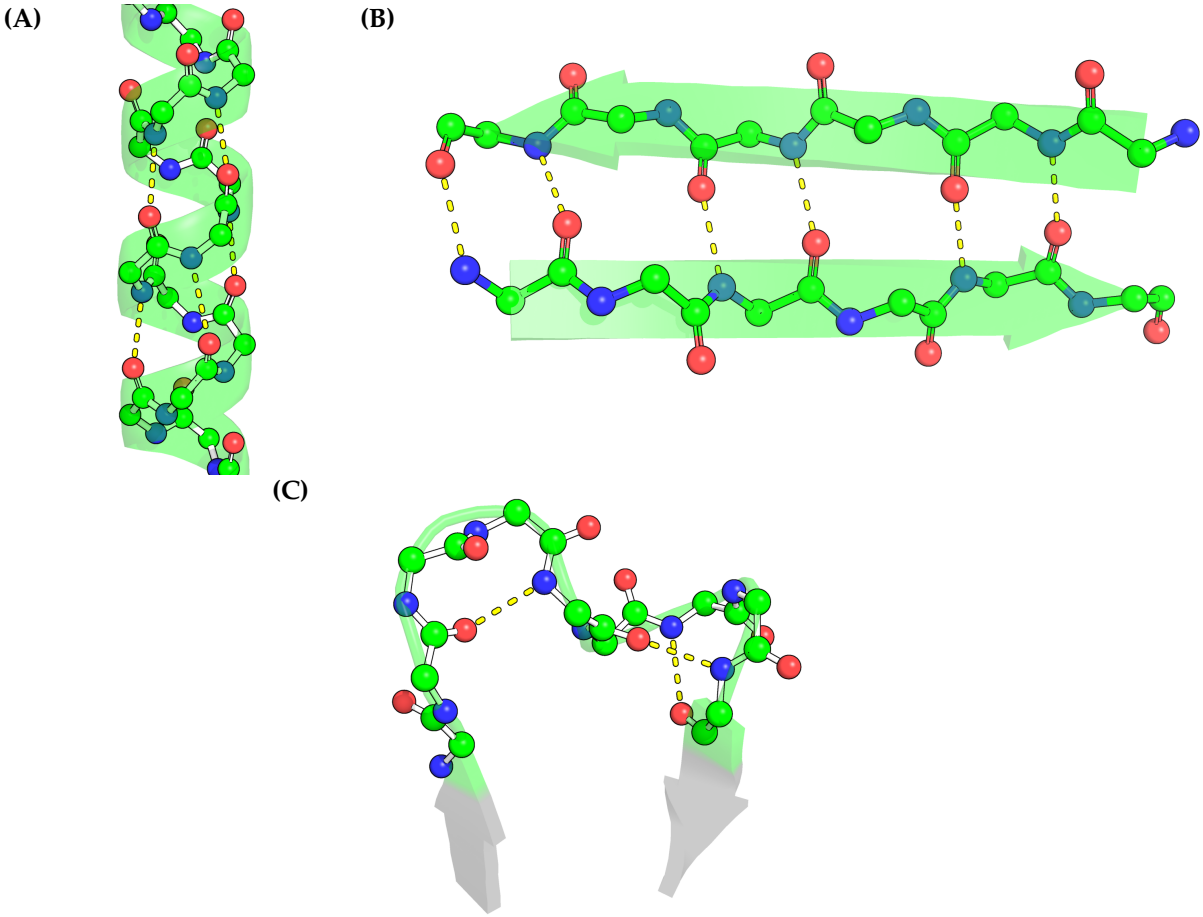

Figure 1.5: Protein secondary structure shown with the backbone of their amino acids and a translucent illustration of the shape of the secondary structure. The hydrogen bond pattern is shown with yellow dashed lines.
(A) An alpha-helix is stabilised by the regular hydrogen bond pattern between the CO group of one amino acid and the NH group of the fourth following amino acid.
(B) A beta-sheet is stabilised by the regular sideways hydrogen bond pattern between the NH and CO groups of amino acids in different strands. The direction of the strands is shown by the direction of the arrows. As the strands have opposing directions, this particular beta sheet is considered anti-parallel.
(C) A loop between two anti-parallel β strands (show in grey). The structure is stabilised by an irregular hydrogen pattern.

An **α-helix** is a coil-like structure which is stabilised by hydrogen bonding between the CO group of one residue and the NH group of the 4^{th} following residue in the direction of the polypeptide (see Figure 1.5A). These helices can be right handed or left handed (the rotation direction is in the polypeptide direction), although the ones naturally observed are right handed.

A **β-sheet** is a structure composed of two or more elongated peptide sequences (called strands) that stabilise each other through lateral hydrogen bonds (see Figure 1.5B). These can either be parallel if the matching strands have the same directionality, or anti-parallel if they have opposing directionality.

The other areas of the proteins that do not form α-helices or β-sheets form either structured turns or loops. Loops are areas of irregular structure which can either be stabilised by hydrogen bonds, or can be unstabilised. Areas that contain loops are usually the areas through which a protein interacts with other proteins. This is especially true of antibodies, with their binding site being formed of six loops (see Section 1.2.2.1).

1.1.3.3 Tertiary and quaternary structure

Long range interactions between residues, or interactions between different elements of secondary structure, create tertiary structure in proteins (Merz and LeGrand, 2012). For example in α-helices one side is usually hydrophobic and the other side is non-hydrophobic. Multiple such α-helices will pack against each other burying their hydrophobic sides, while their polar sides are exposed to solvent (see Figure 1.6A). Another long range interaction that gives rise to tertiary structure is the strong covalent disulphide bond between the thiol groups of two cysteines. They are important for maintaining the stability and structure of some proteins, and are a key feature of the structure of antibody domains (see Figure 1.11B).

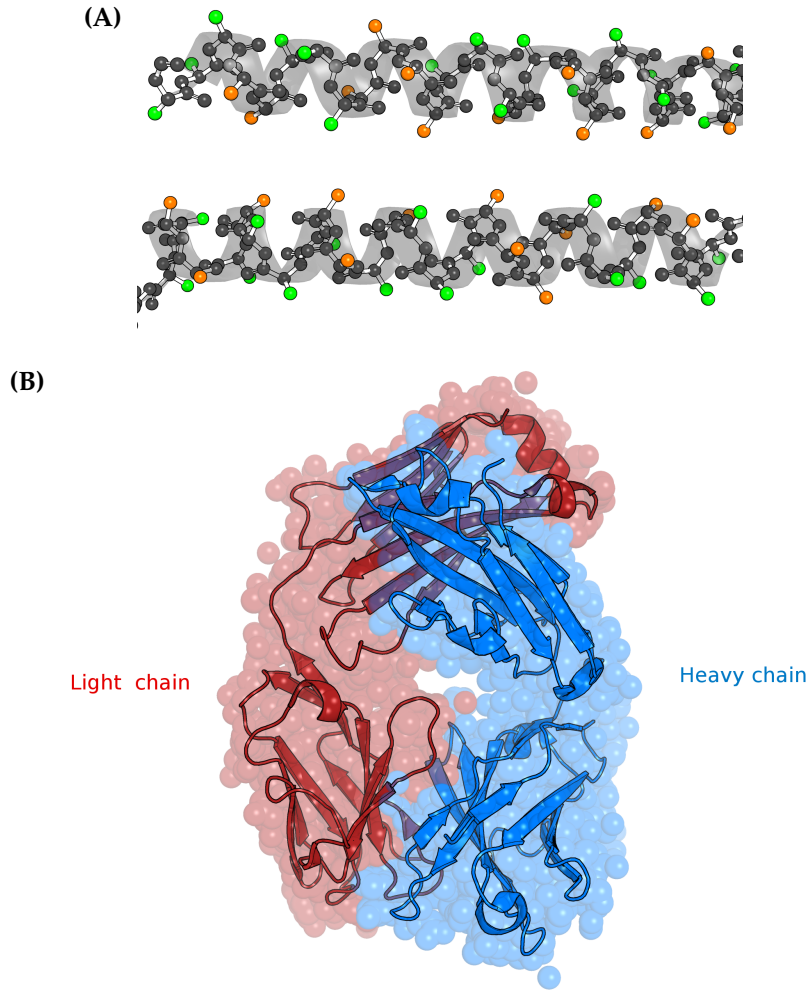

Figure 1.6: (A) Tertiary structure created by two alpha helices packing against each other. For each amino acid the side-chain is shown as one atom, the Cβ. The atom is coloured to orange if the side-chain is hydrophobic or green if the side-chain is not hydrophobic. The backbone atoms of the amino acids have been all coloured grey to aid visualisation. The majority of hydrophobic side-chains from the alpha helices point towards each other, while the majority of non-hydrophobic side-chains point outwards to solvent.
(B) The two chains of an antibody (light chain in red and heavy chain in blue) forming quaternary structure by packing against each other.

Some proteins are formed from more than one polypeptide chain. These chains pack against each other and form what is called the quaternary structure (Branden et al., 1999). This is also a feature of antibodies, as the standard antibody is formed from two chains (see Figure 1.6B and Section 1.2.2.3).

1.1.3.4 Structure classification - SCOP

Proteins can be classified according to the patterns of secondary and tertiary structures they contain. Murzin et al. (1995) developed a hierarchical classification system based on a perceived structural relationship between proteins. The classification is stratified on four levels, with the lowest corresponding to the greatest degree of evolutionary relationship, and the top one the least. The lowest level is the *family*, where proteins with high degree of sequence similarity and structure similarity are placed together. The next level is the *superfamily*, where proteins with a high degree of structure similarity and lower degree of sequence similarity are placed. The 2^{nd} level of the pyramid is the *fold* where superfamilies which have structural similarities are placed together. The top level is the class which categorises together folds that have similar patterns of secondary structure (see Figure 1.7).

The SCOP database has been created through manual annotation, and as such there are many proteins for which an annotation does not exist yet. In these cases the Superfamily package (Gough and Chothia, 2002) can classify automatically a protein based on its sequence using Hidden Markov models that describe the sequence patterns inside each family.

The chains that make up an antibody are classified as an All β protein, with a β sandwich fold, and in the immunoglobulin superfamily.

A similar system of classifying proteins hierarchically based on their secondary structure is CATH (Orengo et al., 1997). The CATH database has a similar four level hierarchy, but there are differences to SCOP as the classification

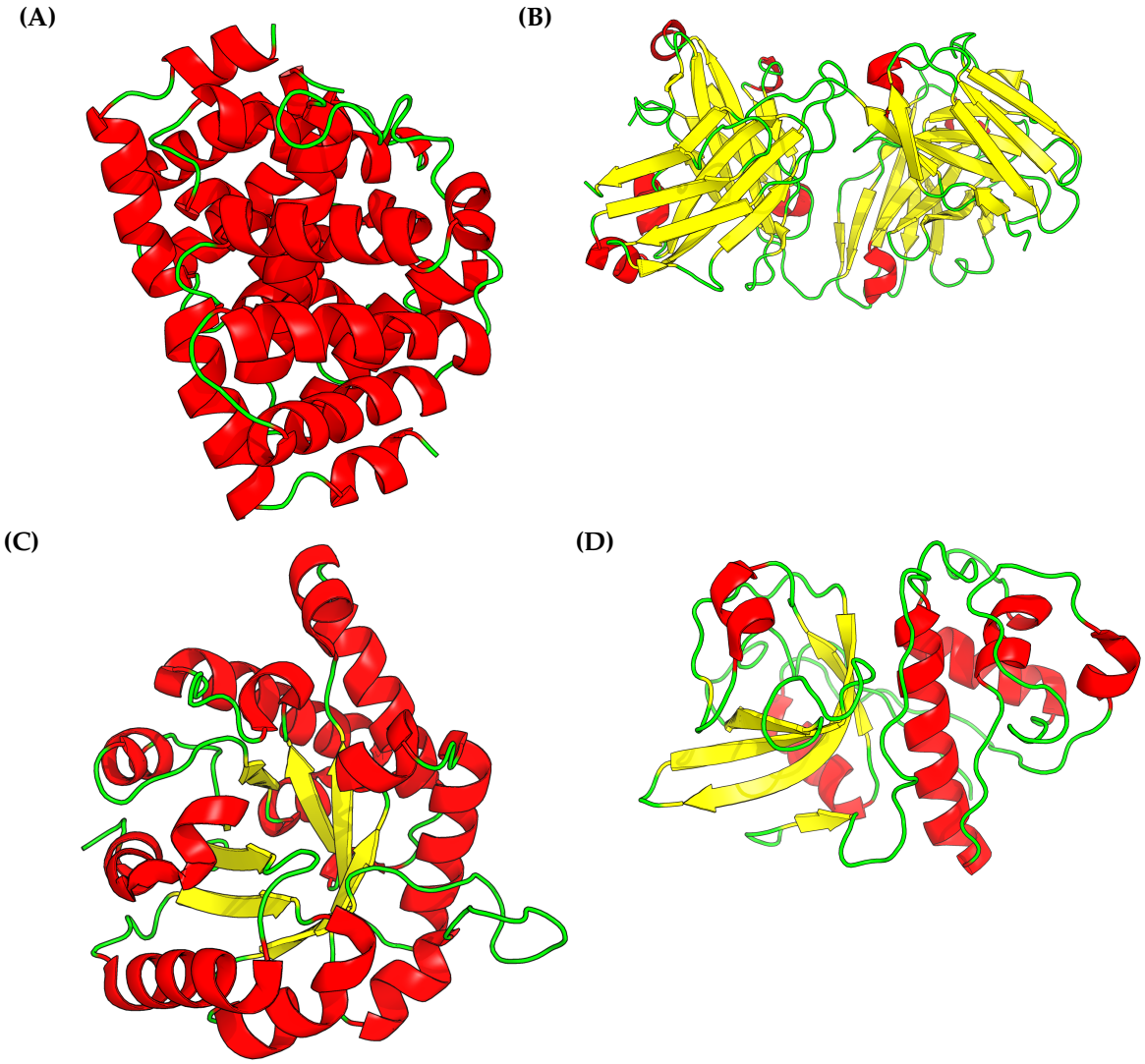

Figure 1.7: Example members for the major SCOP classes. In each figure α-helices are coloured red, β strands are coloured yellow and loops are coloured green.
(A) All α (containing only alpha helices)
(B) All β (containing only beta sheets)
(C) α/β (secondary structure continuously oscillates between beta strands and alpha helices)
(D) α+β (there are separate areas that contain either beta sheets or α helices)
There are other scop classes which are not represented here, but they have relatively few number of members in comparison to the ones presented.

methodology is different and contains more automated steps. In our work these differences are not important and we chose to use SCOP.

1.1.4 Structure determination

1.1.4.1 Importance and challenges

Solving the structure of proteins is an important topic in structural biology. A solved structure can elucidate how a protein performs its function, it can show its binding site and can inform on how its interactions are modulated. For protein engineering purposes a structure can also guide the process of making affinity or stability enhancements (Matthews et al., 1987). Also, in the area of drug discovery, a solved structure of a therapeutic protein target is useful to identify what are druggable areas, and can guide the process of developing and refining potential drug candidates (Anderson, 2003).

1.1.4.2 X-ray crystallography

X-ray crystallography is currently the best available method for providing atomic level data of a protein's native structure (Wlodawer et al., 2013). The locations of individual atoms are identified by measuring the diffraction pattern when an x-ray beam is directed towards the protein. However, before this step the protein needs to be purified and prepared in a crystal form. In order for the diffraction data to be transformed to a reliable structure the instances of the protein that form the crystal need to be arranged in a regular pattern (i.e. the crystal needs to be ordered).

Purified soluble protein in solution is typically coaxed into a crystal structure by adding a precipitant (e.g salt, organic solvents) (Smyth and Martin, 2000). The exact precipitant solution is different for every protein, and the process of identifying the correct precipitant requires multiple trial and error rounds, with no guarantee of success. When a protein does not form a regular crystal

structure it is also common for mutations to be made to the surface residues which are predicted to have impeded the process (Derewenda, 2004).

If a crystal is obtained its structure is solved using an x-ray beam, which can be generated either in a lab tube or in a synchrotron. The x-rays scattered from the atoms of the protein are measured with either an x-ray film or a detector, and are then used to produce the electron density map using a Fourier transformation (Smyth and Martin, 2000). The structure of the individual amino acids of the protein can then be modelled on the resulting electron density. The x-ray film and the detector measure the amplitude of the diffracted wave, but do not measure phase, which is an added complication to the method. This problem is solved by fitting data to a model with a similar structure to identify phase, or by soaking the structure with heavy atoms and comparing the diffraction to the native (Smyth and Martin, 2000).

The protein structure data used in this thesis is solely from structures solved using x-ray crystallography.

1.1.4.3 Other methods

A method which does not require a crystal for solving the structure is Nuclear magnetic resonance spectroscopy (NMR). In NMR highly concentrated purified protein is suspended in an magnetic field and then probed using radio waves. This allows for the identification of distances between conserved pairs of atoms and when sufficient constraints are identified they are projected to a model that satisfies those constraints (Wüthrich, 2001). This method is particularly well suited when the target protein is small (having around or under 100 amino acids) and rigid, and when a crystal can not be obtained (Wüthrich, 2001). It also has the added benefit of identifying structural characteristics while the protein is in a non-crystal environment. It, however, suffers from the fact that if not sufficient constraints can be identified it will produce more than one potential structure.

Cryo-Electron Microscopy (cryo-EM) is a method for structure determination using electrons. With this method the protein solution is frozen in a cryogenic liquid, not requiring a crystal structure, and then the structure is mapped in a similar way to protein crystallography by beaming electrons instead of x-rays (Bai et al., 2015). The detectors for electrons can capture phase data (as opposed to x-ray detectors), and can also capture multiple two-dimensional snapshots per second (Bai et al., 2015). These images are then combined to form a three-dimensional structure. The drawbacks of this method are that it is currently only well suited to large complexes that can withstand electron beams, and the data produced is at low resolution (Bai et al., 2015).

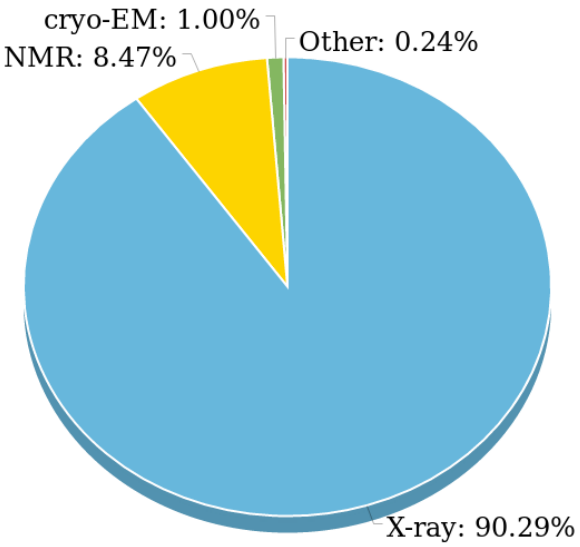

Figure 1.8: Pie chart of the structures deposited in the PDB according to the method used to solve the structure. The numbers that generated the pie chart were taken on the 5^{th} of September 2017.

1.1.4.4 Protein Data Bank

The data obtained from the above methods are stored in the Protein Data Bank (PDB), a publicly available data archive for protein structure data (Berman et al.,

2000). The most common methods for structure determination by number of structures deposited are in reverse order: x-ray crystallography, NMR and cryo-EM (see Figure 1.8). Due to the different nature of the methods some provide one snapshot (e.g. x-ray crystallography) and others multiple snapshots (e.g. NMR). Each individual snapshot is reported inside a PDB file, and is called a model (Berman et al., 2000). In each model the identified atoms from the residues are listed with their coordinates. An example of a line describing the position of an individual atom can be seen in Figure 1.9. For structures solved using

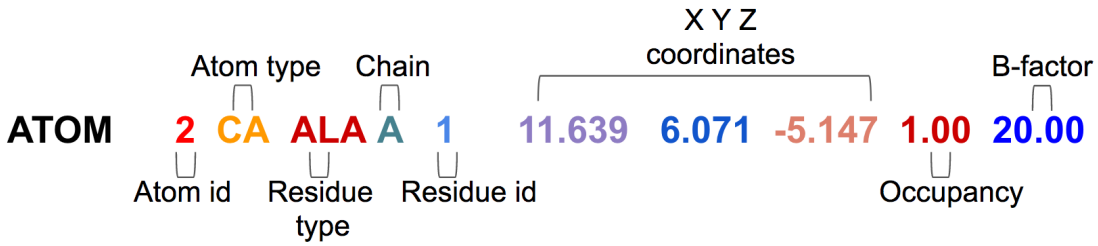

Figure 1.9: Line describing the Cα atom of an alanine residue inside a PDB file. The atom, chain and residue are presented along with the geometric coordinates. The structure from which this line came from was solved using X-ray crystallography and as such the B-factor and the occupancy are also reported.

x-ray crystallography aside from the coordinates two other values are usually reported, the B-factor and the occupancy. The B-factor (or temperature factor) for an individual atom (see Figure 1.9) shows the average divergence of the atoms in the individual cells of the crystal lattice from the coordinates reported on the line. This value is affected by the movement of an atom when they are struck by the X-ray wave, atoms in rigid areas having less movement than ones in flexible areas (Smyth and Martin, 2000). There are, however, cases where this value can include modelling uncertainty resulting from low resolution which is further described in Section 2.2.2.6. The occupancy is reported because a residue may have multiple conformations (e.g. the backbone, or rotamer of the side-chain, may switch from one conformation to another between individual

cells in the lattice). When this is observed individual conformations are reported, and the proportion of cells in the lattice that have a specific conformation are indicated through the occupancy value (Callaway et al., 1996). The sum of the occupancy values for a residue equates to 1.0.

There are areas of a protein, or residues, which do not generate enough electron density to be able to reliably create a model structure. For these residues either the side-chain, or the entire residue is omitted.

The header of a PDB file also contains meta information about the structure. This ranges from a description of the protein that was solved, to its original sequence (which can be different from the solved sequence (see Section 1.1.4.2)), to secondary structure annotations, and to crystallisation conditions.

1.1.5 Structure alignment

A key method used in this thesis is measuring the structural difference between two protein structures. To do this we used the Kabsch superposition algorithm (Kabsch, 1978). This algorithm computes the optimal superposition (i.e. the superposition with the least divergence) of two sets of points using singular value decomposition. The divergence is then quantified using the root mean square deviation (RMSD) value of the two sets of points:

$$RMSD(a_i, a_j) = \sqrt{\frac{1}{n}\sum_{k=1}^{n} |a_{ik} - a_{jk}|^2} \tag{1.1}$$

where a_i and a_j are the two sets of points. n is the total number of points of each of the sets. This algorithm is only suited for two structures with the same number of amino acids.

For a loop there are two ways in which RMSD can be measured: local and global. Local RMSD involves firstly superpositioning the backbone atoms of the loops (C, Cα, N, O) and the computing the RMSD. Global RMSD involves using the anchor backbone atoms (two amino acids upstream and

downstream from the loop) to firstly perform the superposition, and the RMSD is afterwards calculated for the backbone atoms of the loop. In this thesis we preferentially use local RMSD measurements and where global was used is it specifically mentioned

1.2 Antibodies and the immune system

1.2.1 Introduction

The topic of this thesis is antibodies, which are an important part of the immune system. Through the immune system a mammalian organism neutralizes and eliminates pathogens. Pathogens are foreign organisms that enter the body and perform actions which are detrimental to its normal function. Firstly, we will present the antibody and its different functional areas and then we present how an immune response is generated, how antibodies are produced and the specific role they play.

1.2.2 Antibodies

1.2.2.1 Domains and functional areas

The standard structure of a mammalian antibody is a symmetric Y shape, each part of the symmetric unit has two chains: a heavy chain (H) and a light (L) chain (see Figure 1.10). The H chain is formed of four domains, three constant and one variable (VH), while the L chain has one constant and one variable domain (VL). The VH and VL are referred together as the Fv area of an antibody, while the Fv area in combination with the first constant domain on both of the chains is called the F_{AB} area (see Figure 1.10).

The region of an antibody which determines the majority of specificity towards antigens is found on the variable domains and is named the Complementarity-determining region (CDR) (Wu and Kabat, 1970). The CDR is formed of three loops for each of the VH and VL chains. They are named L1, L2, L3, H1, H2, and H3. The specificity towards antigens primarily arises from variation in the sequence and structure of these CDR loops (Wu and Kabat, 1970).

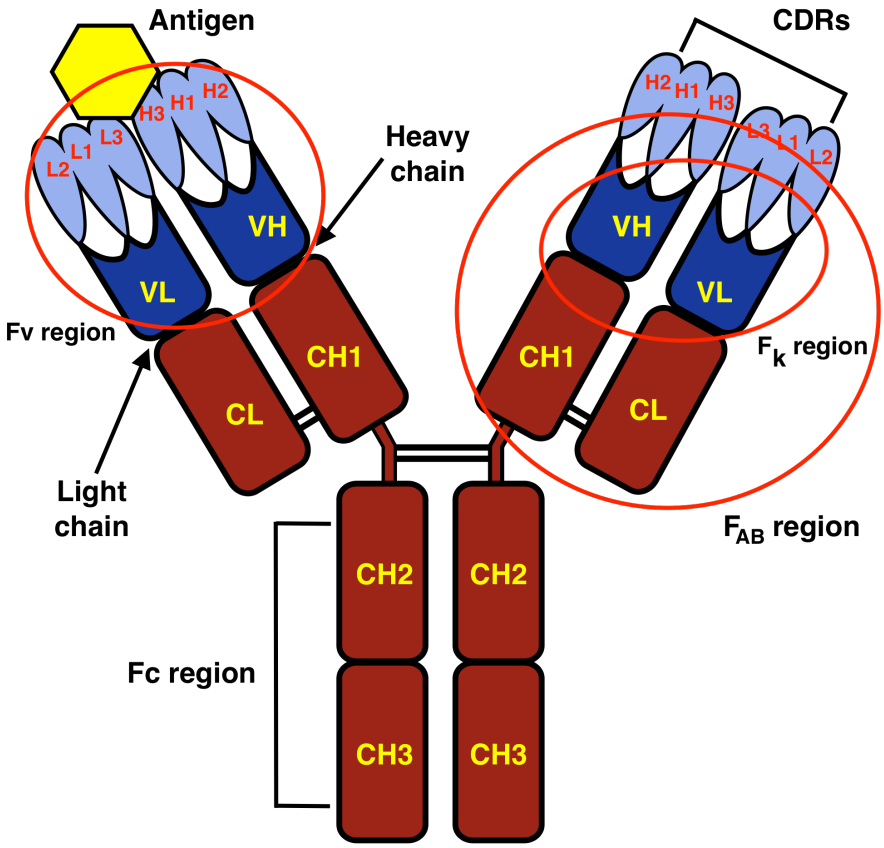

Figure 1.10: The key areas of an antibody protein. The three constant domains of the heavy and the light chain are shown in maroon. The variable domains are shown in the dark blue, with the CDRs being depicted in light blue. The antigen is shown as a hexagon with a yellow background. (The image has been adapted from an image created by Anypodetos released under the Creative commons licence 3.0)

The CH2 and CH3 domains of the heavy chain area called the Fc region, which is responsible for antibody-antigen complex recognition by other cells of the immune system.

1.2.2.2 Numbering and CDR definitions

The issue of how to define which residues form part of the CDR was first addressed by Wu and Kabat (1970). With no structural information, they used sequence information to define the CDR of an antibody as the most variable region in terms of sequence. A second definition of the CDR loops

CDR	Chothia Numbers
L1	L24-L34
L2	L50-L56
L3	L89-L97
H1	H26-H32
H2	H52-H56
H3	H95-H102

Table 1.1: The Chothia CDR definition based on the Chothia numbering. For each of the CDRs the start and end residue number is reported (Source (Chothia and Lesk, 1987)).

was proposed by Chothia and Lesk (Chothia and Lesk, 1987). They generated their definition by looking at the structural differences between loops from the same area on different antibodies.

More recently the International Immunogenetics Information System (IMGT) conducted an automated sequence analysis of all variable regions of proteins that interact with antigens (e.g. antibodies, B Cell Receptors and T Cell Receptors) and produced a more general definition of the CDR independent of the type of receptor (Lefranc et al., 2003). Honegger and Plückthun (2001) have also developed the aHo numbering scheme which focuses on assigning the same number to structurally similar positions inside the CDRs.

In our work we are interested in identifying what are the structurally variable areas of an antibody and therefore make use of the Chothia structural definition (Chothia and Lesk, 1987). We used the ANARCI (Dunbar and Deane, 2015) antibody numbering software package to number residues in an antibody Fv according to the latest Chothia numbering (Al-Lazikani et al., 1997). This numbering differs from the original numbering (Chothia and Lesk, 1987) by having the insertion point at residue L30 instead of L31. The CDR regions are delimited using the numbers in Table 1.1.

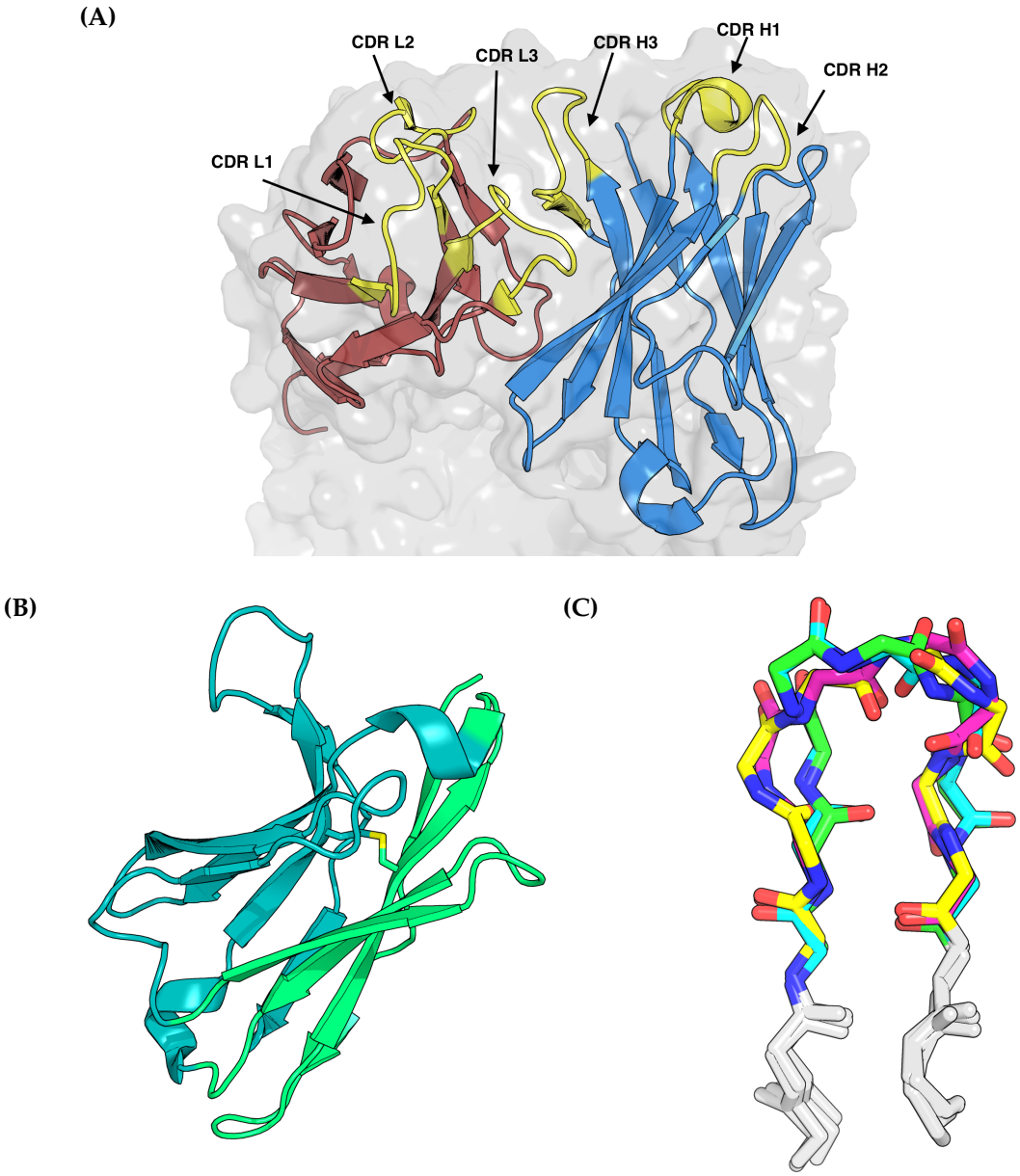

Figure 1.11: Structure of an antibody.
(A) The structure of an Fv domain with the light chain coloured in red, and the heavy chain coloured in blue. The CDRs are highlighted in yellow.
(B) The two beta sheets making up a heavy chain of an antibody (one green, one teal) stabilised by a disulphide bond (yellow).
(C) Four loops from two different canonical classes aligned based on their anchor residues (two residues upstream and downstream from the loop coloured in gray).

1.2.2.3 Structure

The standard structure of the variable region of an antibody chain is that of a β sandwich. This is formed of two anti-parallel beta sheets which pack against each other and are stabilised by a disulphide bond (see Figure 1.11). The CDRs are composed of six loops all on the same side of the β sandwich (see Figure 1.11A). Based on structural information CDRs apart from H3 can be clustered into *canonical classes* (Chothia and Lesk (1987), North et al. (2011)). A canonical class is a cluster of loops of the same length from one CDR with similar backbone structures (see Figure 1.11C). Nowak et al. (2016) later proposed a method to cluster loops into length-independent canonical classes.

1.2.3 Immune response

1.2.3.1 Antigens

An antigen is a molecule or cell that can induce an immune response. In the normal functioning of the immune system these are of foreign origin, however, there are cases where the body detects its own cells or proteins as antigens which is called auto-immunity (Alberts, 2017).

1.2.3.2 Innate immunity

The innate immune system is the first line of defence for all living organisms (Alberts, 2017). A typical innate immune response is generated by Phagocytes that recognise features common among antigens to ingest and destroy them.

1.2.3.3 Adaptive immunity

The second line of defence of the immune system is the *adaptive* immune response (Flajnik and Kasahara, 2010). It is called adaptive because its main components evolve over time to counteract the antigen. There are two separate

adaptive responses, the humoral and the cellular. They differ in the ways in which they remove their targets, but they are also interconnected.

Antigen presenting cells express on their surface Major Histocompatibility Complex (MHC) proteins with a bound peptide from proteins digested inside the cells. These proteins can be of foreign origin (i.e. pathogen), or endogenous. During their circulation they will enter lymphoid organs (spleen, lymph nodes and patches inside the gut). Inside those organs they are detected by a type of immune cell, called a naïve T-cell (a T-cell that has never encountered an antigen). When a naïve T-cell encounters an antigen presenting cell it activates proliferation and differentiation which will create effector T-cells that can exert other functions on the antigen and its peptides (see Figure 1.12A).

There are two main types of T-cells, CD4 and CD8 T-cells. A naïve CD8 T-cell binding an antigen presenting cell will differentiate to a cytotoxic effector T-cell (Andersen et al., 2006) (see Figure 1.12B). A cytotoxic T-cell kills cells presenting the same peptide on its surface through an MHC class I presenting protein. The process presented so far is the cellular response system.

A CD4 T-cell binding an antigen presenting cell will become a helper effector T-cell (Reiner, 2007) (see Figure 1.12C). It is called a helper T-cell because it assists B-cells recognising the same antigen to further proliferate and differentiate into plasma cells which produce antibodies against that antigen.

Naïve B-cells recognise the same antigen through their immunoglobulin receptor. When the receptor binds the antigen it sends a signal inside the cell which promotes endocytosis of the receptor-antigen complex (Malhotra et al., 2009). Inside the cell, through the same process used by phagocytes, the antigen is broken down and its peptides are presented on the surface, but this time using MHC class II proteins (Wollenberg and Bieber, 2002). During circulation a naïve B-cell presenting the antigen will enter a lymph organ, and its MHC class II presenting complex will be detected by a helper T-cell that recognizes the same

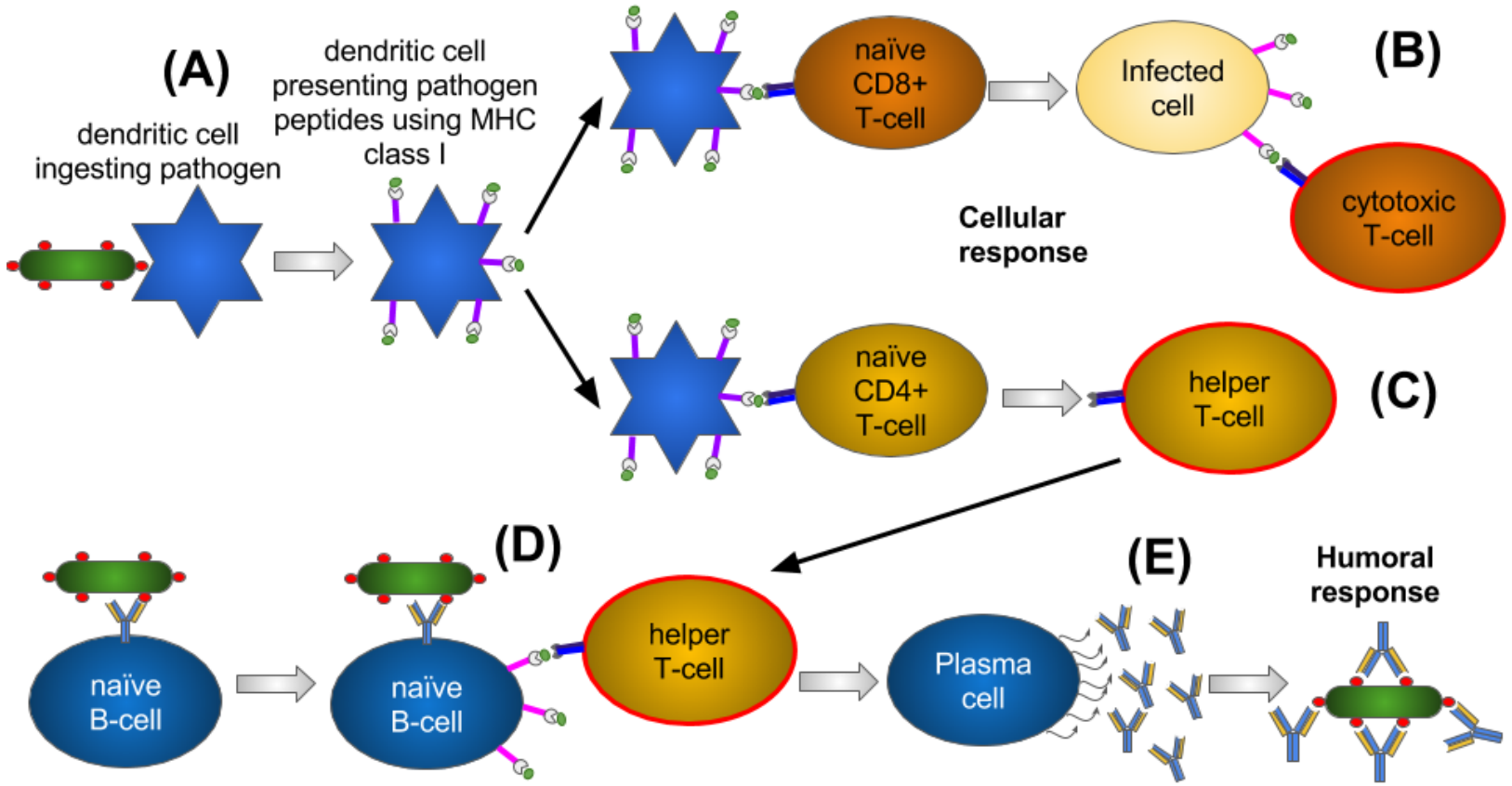

Figure 1.12: The process of immune response from antigen recognition to generation of antibodies.
(A) Phagocytes ingest and digest antigen (in this case a pathogen), presenting their peptides using MHC class I proteins.(B) CD8 T-cells binding to the phagocytes differentiate to cytotoxic T-cells. (C) CD4 T-cells binding to the phagocytes differentiate to helper T-cells. (D) Antibodies bind to the same pathogens using their immunoglobulin receptor, ingesting and digesting the pathogen, and then presenting peptides using MHC class II proteins. The helper T-cells from (C) will bind to the B-cell MHC with the peptide and activate differentiation to a plasma cell. (E) When the B-cell matures to a plasma cell it releases large volumes of antibodies which stop the pathogen through the methods described in Figure 1.13

peptide (see Figure 1.12D) . When the T-cell binds the B-cell MHC complex it releases cytokines which are detected by the B-cell through its cytokine receptor. The simultaneous signals of the B-cell receptor bound to antigen, the T-cell bound to the MHC receptor and the detection of cytokines will start the proliferation and differentiation of the B-cell clone into a plasma cell. A plasma cell will produce high volumes of high affinity antibodies which are identical in the variable region to the immunoglobulin receptor on its surface (see Figure 1.12E). During this period somatic hypermutation is also activated, which is further described in Section 1.2.3.5. Antibodies will then interact with antigens in three major ways: neutralization, opsonisation and complement activation (Hofmeyr, 2001). Neutralization is performed by coating the antigen with antibodies, which will inhibit the antigen from binding to cells and then entering/infecting them (see Figure 1.13A). Opsonisation is also performed by coating of an antigen, but in this case the Fc region of the antibodies is detected by phagocytes which promotes faster ingestion uptake (see Figure 1.13B). Complement activation is performed in similar ways to opsonisation, with the Fc region binding other proteins of the complement system. These are also detected by phagocytes, increasing even more the rate of antibody-antigen ingestion (see Figure 1.13C).

1.2.3.4 Antibody isotypes

There are several types of antibodies called isotypes. The main isotypes are IgA, IgG, IgM, IgD and IgE. IgMs are naïve low affinity antibodies normally found in the circulatory system. These are the antibodies first created upon detection of an antigen. Their Fc region type favours the formation of pentamers. A pentameric antibody is useful because IgMs have naturally low affinity, but as a pentamer they can bind to multiple instances of the target protein on the antigen at the same time creating avidity (Rudnick and Adams, 2009).

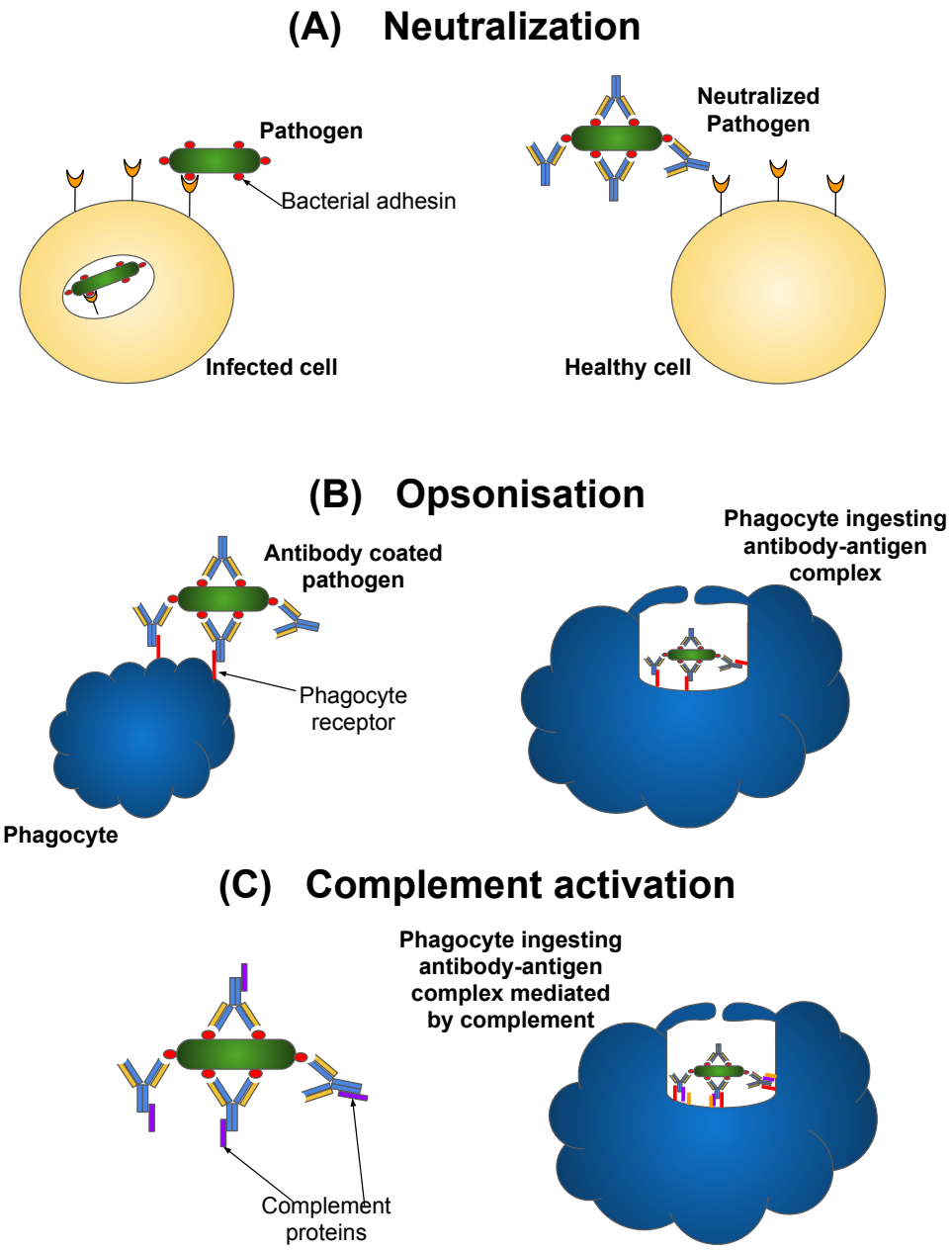

Figure 1.13: An illustration of the main methods in which antibodies mediate immunity. In (A) an antigen in the form of a pathogen is neutralized by blocking the adhesin proteins which help the pathogen infect a cell by crossing its membrane. In (B) antibodies coat an antigen, and their Fc areas are recognised and ingested by phagocytes. In (C) complement proteins bind the Fc region of the antibody which enhances opsonisation by the fact that both the complement and the Fc are recognised by the phagocyte.

Segment	κ	λ	heavy
Variable	34-38	29-33	38-46
Diversity	0	0	23
Joining	5	4-5	6
Constant	1	4-5	9

Table 1.2: The number of variants for each type of gene segment (Source Murphy (2008))

During the maturation process of a B-cell the switching of the isotype will be favoured. This is achieved by the selection of a different constant region for the antibody, which will then be used to exert a different mechanism of action. In the case of IgA antibodies their Fc favours a dimer, instead of a pentamer, while in the case of IgG antibodies their Fc favours a monomeric state. IgGs are the most prevalent type of antibody and have high affinity to their target (Vidarsson et al., 2014).

1.2.3.5 Generating diversity and affinity maturation

One individual B-cell has many instances of the same receptor expressed on its surface. The sequence and the structure of its CDR area determines the antigen to which it binds. These cells, and therefore the receptors, are required to bind to a diverse palette of antigens, and for this there are two processes that sample through the sequence space in order to produce ones which will bind more strongly to the target. The first process randomly rearranges genes segments from a pool of genes in order to generate diversity and is called somatic recombination (see Figure 1.14A) . The second one involves introducing random point mutations and is called somatic hypermutation. (see Figure 1.14B). Through these processes it is hypothesized that the immune system is capable of producing 10^{10} different antibody sequences (Glanville et al., 2009).

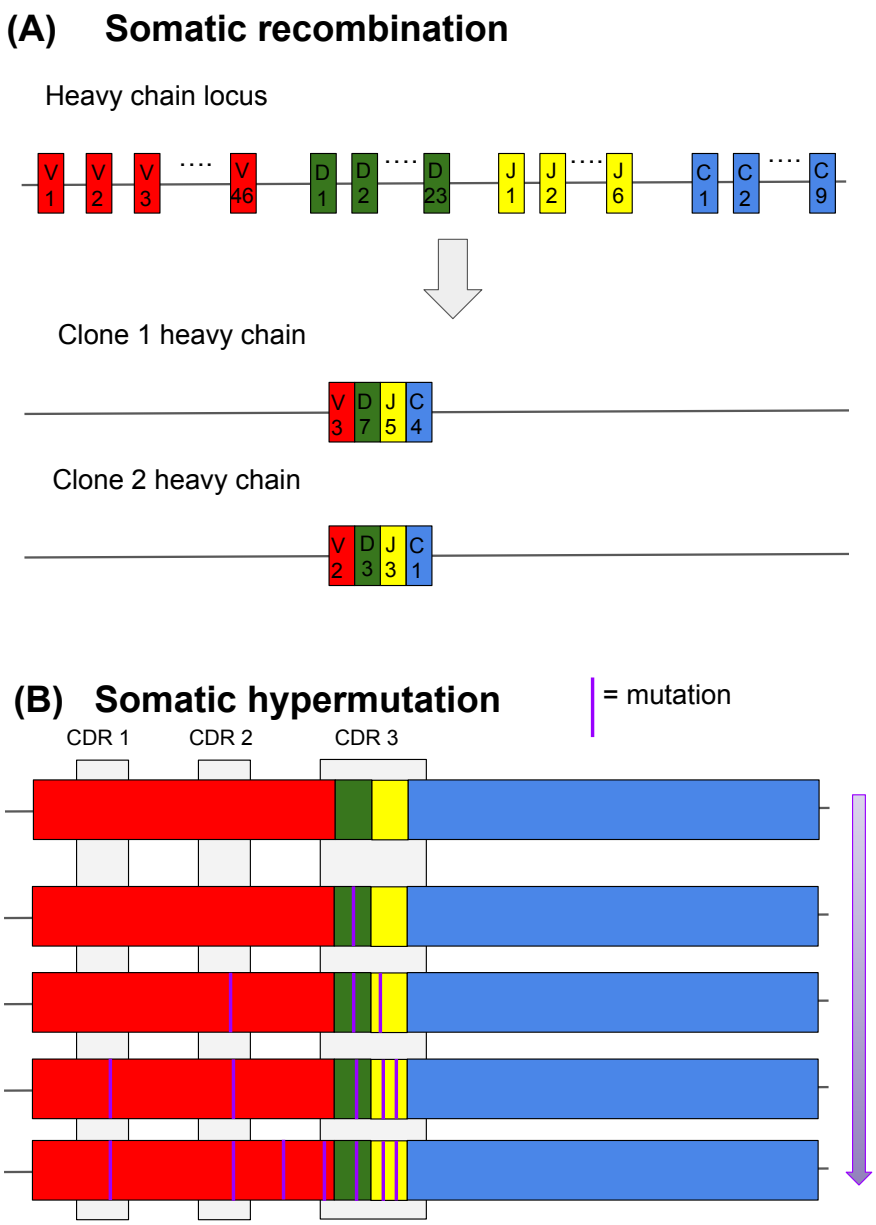

Figure 1.14: The two processes through which antibodies generate diversity in their variable regions. In somatic recombination (A) multiple variants for each type of segment (Variable, Diversity, Joining and Constant see Table 1.2) are randomly brought together through somatic recombination. The segments are represented using coloured rectangles (red for variable, green for diversity, yellow for joining and blue for constant) and the chromosome is represented by a black line. In somatic hypermutation (B) random point mutations are introduced in the Fv region, with a preference for the CDRs.
This figure represents heavy chains, light chains not having a diversity gene segment

1.3 Antibody design

1.3.1 Introduction

As described above antibodies are part of the adaptive immune system and a complex interplay between different cells of the immune system produces them in order to tag antigens for destruction or inhibit their natural function. There are, however, diseases for which antibodies will not be produced, because the targets are not pathogens, but normal cells or proteins (e.g. cancer cells, over-expressed human proteins). In these cases antibodies for these targets can be produced by simulating the process *in vitro* or *in vivo* inside an animal model.

Antibodies are able to attain specificity and affinity to their antigen. They are also modular, the antibody framework region being stable across billions of potential sequences for the CDRs. Finally, the presence of the Fc region is known not to be necessary to neutralize a target (Carter, 2006). These characteristics have made the antibody scaffold attractive for design.

1.3.2 Wet-lab methods

1.3.2.1 Animal model immunisation

Monoclonal antibodies can be generated in high volumes and isolated *in vitro* through a hybridoma (Köhler and Milstein, 1975). A hybridoma is created by taking an antibody producing B-cell and fusing it to an immortal myeloma cell culture. This can produce monoclonal antibodies for an extended period of time.

Initially, an animal model is immunised with the targeted antigen with the hope that it will generate an immune response against the target (Nissim and Chernajovsky, 2008). This step might be repeated multiple times and is not guaranteed to succeed because it depends on the capability of the animal model to generate an appropriate immune response. If an antibody is generated against

the target, the B-cell producing the antibody is isolated and used to create the hybridoma (Nissim and Chernajovsky, 2008).

Through this method Muromonab-CD3 was produced, the first antibody therapeutic approved for clinical use in humans (Group et al., 1985). It was used to promote apoptosis in T-cells that recognise and generate an immumogenic response against cells from transplanted organs (Group et al., 1985). At the time it was a breakthrough in reducing organ rejection after transplant surgery.

One of the counter-indications of Muromonab-CD3 is that it can not be used for people allergic to mouse proteins (Nissim and Chernajovsky, 2008). The reason for this is that the antibody is a mouse protein, having been generated by the murine immune system. When being administered to a human this creates the potential of eliciting a counter immunogenic response.

1.3.2.2 Chimerisation and Humanisation

To solve the issue of immunogenicity one option is the chimerisation of the antibody (Chames et al., 2009). Chimerisation was firstly achieved by replacing the Fc region of the animal model with an Fc from a human antibody, creating the chimeric antibody (Morrison et al., 1984). This lead to the creation of Abciximab, an anti-coagulant used after angioplasty surgery (Nissim and Chernajovsky, 2008). The -xi- substem in the name of the antibody indicates that it is a chimeric antibody in the Nomenclature of monoclonal antibodies (Organization et al., 2009).

The Fv domain, however, is still of murine origin and still capable of eliciting an immunogenic response. Jones et al. (1986) proposed a different method where only the CDRs were transferred from the non-human antibody to the human antibody. It was shown to transfer binding specificities and since then there have been numerous experimental studies that confirm this principle (Nicaise et al. (2004), Nakano et al. (2010), Nishimoto and Kishimoto (2008), Hwang

et al. (2005), Asano et al. (2006)). The first antibody therapeutic to be developed with this method was daclizumab, initially developed for prevention of kidney transplant rejection, but is now used for relapsing multiple sclerosis (Vincenti et al. (1997), Bielekova et al. (2004)). It works by inhibiting T-cells involved in the autoimmune response. The -zu- substem in the name of the antibody indicates that it is a humanized antibody.

1.3.2.3 Transgenic mouse

Transgenic mice can be a solution to humanisation by generating a human antibody inside the animal model. This is achieved by inserting the human antibody germline genes inside the mouse embryo cells, while silencing the mouse antibody germline (Lonberg, 2008). When the mouse is presented with the antigen it produces human antibodies. The antibodies are then produced via the same hybridoma process. One drawback of this method is that the transgenic immune system requires more rounds of immunisation on average to obtain an antibody than the native mouse immune system (Lonberg, 2008).

1.3.2.4 Phage display

Phage display is different from the previous methods because it can select an antibody binder without the use of a mammalian host organism. The heavy and light chain gene segments of the antibody are fused together into one gene segment, and then ligated to a gene that contains a coating protein of a bacteriophage virus (i.e. a virus that infects bacteria) (Smith and Petrenko, 1997). This will eventually express the antibody on the surface of the virus. The virus with the attached antibody is called a phage antibody (McCafferty et al., 1990). This process is done for a library of antibody genes and the resulting phage antibodies are then inserted into bacteria giving rise to millions of different surface antibodies. The phage antibodies are then presented to their intended

target, which is immobilized on a plate well. The well is then washed, leaving only phage antibodies which bind to the required target (McCafferty et al., 1990). The phage antibodies that bind can then be sequenced because their DNA is contained within the phage. This process can go through multiple iterations where only the genes of the binding antibodies from the previous iteration are used to create the library. This promotes the identification of more specific and higher affinity antibodies for the required target in every round, essentially acting as a *panning* method for identifying the highest affinity binders (Frenzel et al., 2016).

Phage display can be used in three different ways. The first use is to pan high affinity binders from an immune library of antibodies (i.e. from the bloodstream of an immunised organism). This is especially useful to identify the antibodies naturally developed in an organism against a target, an example being identifying neutralizing antibodies in non-progressing HIV patients (i.e. patients who do not naturally progress to AIDS long-term) (Trott et al., 2014). The second use is on a universal library, or a blind antibody repertoire (i.e. antibodies that haven not been shown previously to bind to the target) to identify if there is a binding antibody (Winter et al., 1994). For this case a naïve human IgM repertoire composed of rearranged gene segments is used as a library to be panned. Finally, the third use case is as a directed mutagenesis technique by using a synthetic library generated by randomly mutating a reduced set of initial antibodies (Pini et al., 1998). At each step the high affinity binders can be identified, and in the next rounds they can be further mutated, effectively simulating somatic hypermutation (see Section 1.2.3.5) *in vitro*.

Phage display was the method of development for Adalimumab (Humira), one of the highest selling bio-therapeutics of all time (King, 2013). Adalimumab is used against rheumatoid arthritis by suppressing the Tumor Necrosis Factor

(TNF) proteins which create the inflammatory response. The -u- substem in the name of the antibody indicates that it is a human antibody.

1.3.3 Computational methods

Computational methods for rational design of antibodies are currently less well established and there is no antibody therapeutic that can trace its origin from a computational antibody design pipeline. Computational tools for designing antibodies exist, and will be presented in the following, but the majority of methods tackle one individual topic of antibody design.

Computational antibody design methods can be split into three major themes: modelling the structure of the antibody from sequence, affinity maturation on an existing antibody structure, and *de novo* antibody design. In the following modelling and *de novo* design are detailed as they are the main focus of our work.

1.3.3.1 Protein modelling

In this subsection we introduce the basic principles behind protein structure modelling, and in the next subsection focus on their application to antibodies.
Template-based modelling, also known as homology modelling, works on the empirical observation that proteins, or fragments, with high sequence similarity also have high structure similarity (Chothia and Lesk, 1986). For a given sequence a template is predicted from the structures available, and then the template is refined according to its unique amino acid sequence. The drawback of this method is that for more than half of protein sequence families a solved structure does not exist (Kamisetty et al., 2013).
De-novo prediction If a template does not exist the structure can be predicted from first principles. In the same way as the protein folds to a more energetically favourable and stable structure the problem of identifying the combinations of conformations of its individual amino acids is treated as a global optimisation

problem. For this a sampling method iterates through conformations of amino acids in a gradient descent fashion, with the aid of an energy function that associates higher scores to native conformations (Dorn et al., 2014). The problem with this method is that the search space is very large (Levinthal, 1969) and the evaluation of the energy function is computationally intensive. To this date only small proteins of less than 100 amino acids can be solved accurately with this method, running on dedicated supercomputers for extended periods of times (Maximova et al., 2016).

Loop modelling The modelling of a loops is a special case of protein modelling, and is particularly important for antibodies. Loop modelling can be performed either using a template or ab-initio approach. In the case of template-based modelling a suitable loop model is identified from a database of loops using a scoring function based on the sequence of the loop. The scoring of loops poses a harder problem because a mutation has a higher weight in a short peptide than in an entire protein. Loops can also be flexible and one sequence can be found in multiple conformations, or similar sequences can have high degree of structure difference (Babor and Kortemme, 2009). The predicted structure of the loop also has to be melded on the structure its native protein, and differences in the orientations of the anchors can have a significant impact on the global RMSD of the loop (Deane and Blundell, 2001). Some of the methods that perform template-based loop modelling are FREAD (Deane and Blundell, 2001), LoopWeaver (Holtby et al., 2013) and LoopIng (Messih et al., 2015).

Ab initio methods sample Phi-Psi backbone conformations for a loop and use a general or dedicated energy function to score each conformation. The significant challenge with these methods is the ranking of models, as the energy functions are not able to systematically rank first the best conformation (Marks et al., 2017). Examples of ab-initio loop structure modelling algorithms include

Modeller (Fiser et al., 2000), Loopy (Xiang et al., 2002) and Rosetta loop modelling (Stein and Kortemme, 2013).

1.3.3.2 Antibody modelling

The important topic in antibody modelling is the mapping of the sequence of a variable domain chain (i.e. a heavy or a light chain) to its structure. According to the latest assessment of antibody modelling the existing tools can predict the majority of the areas of a variable domain chain with an average accuracy close to 1.0Å (Almagro et al., 2014). There is one area, however, the CDR H3, which is consistently modelled poorly, with an average RMSD > 1.5Å . In the following the major modelling challenges for an antibody are detailed.

Framework

When predicting the structure of a chain the template-based modelling paradigm can be used. This works on the empirical observation that proteins, or fragments, with high sequence identity also have high structure similarity (Chothia and Lesk, 1986). In the case of antibodies, Leem et al. (2016) have shown that for the framework region (the variable region without the CDRs) a template structure with a sequence identity greater than 80% has on average an RMSD < 1.0Å . Through a further modelling analysis of an NGS dataset of 15 million sequences our group has observed that for 99% of the antibodies a template can be identified in the PDB with more than 80% sequence identity (unpublished data). This suggests that at the current time there are enough antibody structures solved to model the majority of existing antibody frameworks.

non-H3 CDRs

In the case of CDR regions template based modelling is more challenging because of the increased sequence variability, which decreases the probability of identifying a model with a high sequence identity. However, it is known that for CDRs L1, L2, L3, H1 and H2 modelling is aided by the fact that the

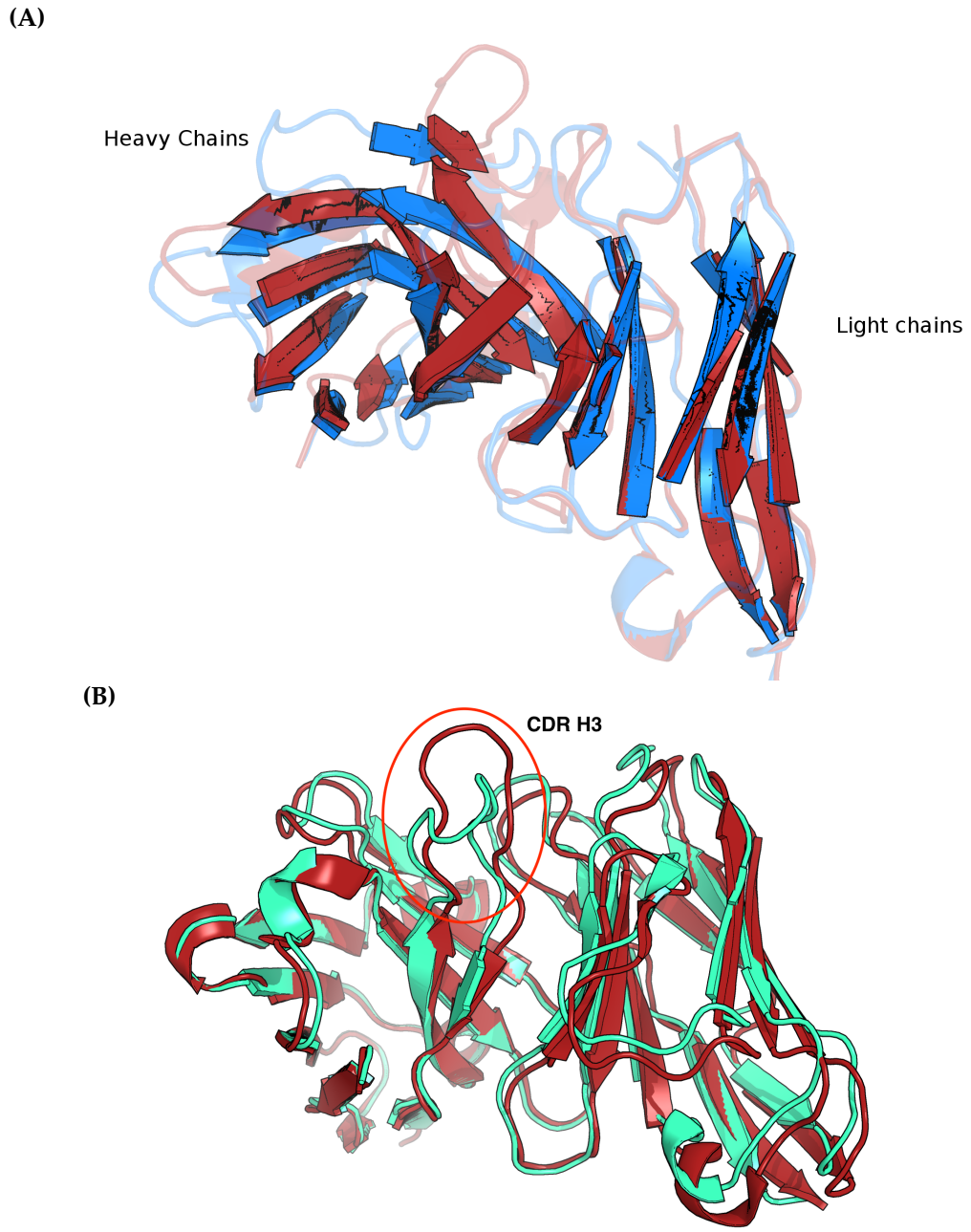

Figure 1.15: (A) Angle between the heavy and the light chain. The picture shows the alignment of two antibodies (one red, one blue) by their light chains. The blue heavy chain has a different angle to the light chains than the red heavy chain.
(B) Alignment of the structure of an average target (in terms of prediction quality) from the second antibody modelling assessment (Almagro et al., 2014) against the best prediction from ABodyBuilder (Leem et al., 2016). In most areas the local conformation is predicted correctly (although there are some displacements due to angle difference), but the CDR H3 conformation is completely different in the model.

backbone structures can be clustered into *canonical forms* (see Section 1.2.2.1). These can be predicted based on length and a few key residues (e.g. Chothia and Lesk (1987), North et al. (2011)). Recently Nowak et al. (2016) have also shown that the canonical class model can be extended to be length independent, based on the observation that loops with similar sequences but different lengths also share common structural motifs. An alternative method of predicting the structure of non-H3 CDRs is by using FREAD (Choi and Deane, 2010), which identifies suitable structural models for loops based on a sequence similarity score, rather than a sequence identity. This score was created based on dihedral angle patterns of the template loops and the probability of that changing by mutation to each residue type. Leem et al. (2016) have shown that for the non-H3 CDRs the top FREAD model is accurate to an RMSD < 1.0Å .

H3 CDR

The H3 loop is the only area of the antibody for which current template based methods systematically fail to produce reliable models (see Figure 1.15B). In Chapter 2 we provide an analysis that shows that this inability extends to a much broader dataset of antibodies, and it is due to the lack of suitable templates in the currently available solved structures. An alternative modelling paradigm, *ab initio*, can be used for this case. In *ab initio* a sampling method iterates through conformations of amino acids in a gradient descent fashion, with the aid of an energy function that associates higher scores to more native-like conformations (Dorn et al., 2014). Marks and Deane (2017) developed a hybrid method from which the sampling starts from the best available template, and then the conformations are sampled based on Ramachadran distributions for each amino acid. Weitzner and Gray (2017) use a similar approach, but restrict the conformation of the C-terminal region of the loop to a kinked conformation based on the observation that 80% of H3 loops share this conformation. There are two caveats to these methods. Firstly they are computationally intensive,

for a target the runtime is on the scale of hours, as opposed to template-based methods which can make a prediction on the timescale of seconds. Secondly these methods produce hundreds of potential models, but are unable to rank them correctly, the top ranked prediction in the majority of the cases not being the best prediction.

VH/VL angle

In order to complete a full antibody variable area after both chains have been predicted, the angle between them has also to be predicted. It has been observed that the angle between the two chains varies between different antibody structure (Dunbar et al., 2013a) (see Figure 1.15A). In some cases this angle can be predicted from a number of key residues (Dunbar et al. (2013a), Bujotzek et al. (2015)). An alternative method for predicting the angle is by using one template that encompasses both chains where the angle can be transferred from the template to the target.

1.3.3.3 De novo design

In *de novo* antibody design the structure of the target is provided and the task is to design an antibody that will bind to the antigen. A number of methods have been developed that are able to produce *de novo* antibody designs. These are still at the computational level, with very little comprehensive experimental validation being provided to date. These will be detailed in the following.

OptCDR and OptMaven

One of the first attempts at *de novo* design of antibodies was OptCDR (Pantazes and Maranas, 2010), which identifies the best CDR loop combination for a specific target. OptCDR can be set to scan the entire antigen for potential epitopes (and design for each), or can be restricted to a specific epitope. For an individual epitope in the first step the backbone of the CDR loops is picked from pre-compiled libraries. These libraries were generated for each CDR from

existing antibody structures. The combinations of six CDRs are ranked according to their geometric complementarity to the epitope, which is scored based on a proprietary inter-atomic distance heuristic. In the next steps the amino acid sequence is initialised for the backbone based on favourable electrostatic interactions with the epitope, and then both the backbone structure and sequence are further refined. Finally, a short-list of potential designs is produced and ranked by an energy function they developed and trained on an antibody dataset consisting of mutants of one antibody with associated binding affinities.

OptCDR does not haven any input outside of the CDR area, leaving the framework to be added by the user. In order to design the whole antibody variable area OptCDR was extended to OptMaven to use the modular antibody parts which were generated from fragments mapping gene segments of an antibody, and not the CDRs (Pantazes and Maranas, 2013). An antibody is therefore constructed from a pre-compiled database of structures for each gene segment in a similar fashion to how diversity is generated by somatic recombination (see Figure 1.14A). Recently, (Poosarla et al., 2017) have shown experimentally that three (out of 5) designs created using OptMaven against a peptide antigen bind with high affinity.

AbDesign

Lapidoth et al. (2015) developed AbDesign, using a similar approach to OptMaven where they construct an antibody from modular parts which are not CDRs. Their modular parts were generated by identifying area of high structure conservation on the antibody for segmentation (e.g. the Cysteines forming the disulphide bridge). Their refining sampling methodology is also geared towards sequences of antibodies instead of random point mutations.

Liu et. al (2017)

In this case an automated method was not proposed, but rather a principle for a *de novo* antibody method was demonstrated using a case study on the Keap1

protein. Their approach is different from the methods above because in the first instance they design only one CDR, and they do this by grafting a known antigen binding motif from a non-antibody protein. The motif is, however, not grafted in the classic sense, they instead identify antibodies that contain this motif on CDR H2. These antibodies are then further refined by grafting other H3 structures that might increase binding affinity, using an energy function for selection. Finally, they validate experimentally that some of the refined designs bind with nM affinity to their target and also obtain crystal structures to show that their designed antibody indeed binds to the expected epitope, in the correct pose.

The methods presented so far share a common drawback, the conformations used to design the binding loops of the antibody are either from existing antibody structures (OptCDR, OptMaven, AbDesign) or need to have been previously observed in an antibody (in Liu et al. (2017)). In this thesis we present a method that transcends this drawback by using fragments from non-antibody proteins to design an antibody.

1.4 Thesis Overview

1.4.1 Chapter 2. Antibody CDR loops structural diversity

Chapter 2 contains an analysis of the proportion of the antibody sequence space that can be modelled using CDR structures deposited in the PDB. The analysis was done using an Next generations sequencing dataset of 15 million sequences of naïve IgM antibodies. The results show that for all but CDR H3 the majority of sequences have a suitable structural model. We then analysed why CDR H3 is hard to model computationally and found that even if the perfect template selection algorithm existed the majority of structures would not have a suitable template. We then analysed what makes the H3 hard to model computationally and found it is more diverse than loops in the rest of

the protein world, with unique four residue fragments and unusual dihedral angle combinations for Tyrosine and Glycine.

The work on analysing the problems with modelling H3 in this chapter has been published in *Regep, Cristian, Guy Georges, Jiye Shi, Bojana Popovic, and Charlotte M. Deane. "The H3 loop of antibodies shows unique structural characteristics." Proteins: Structure, Function, and Bioinformatics (2017)*. The work on modelling the sequences in the NGS data set has been written up as part of a broader article which is submitted and currently under review.

1.4.2 Chapter 3. SAbDesigner: Designing antibodies using non-antibody protein loops or fragments

Chapter 3 presents SAbDesigner, an automated pipeline for *in-silico de novo* antibody design against a specific target. The antibody is designed by mimicking the binding interface of a receptor of the target. This chapter describes the process of selecting the best loops for grafting and identifying a framework on which they can be grafted. Interleukin-5 (IL-5) is used as a target to guide the explanation of the methodology, but designs against other important therapeutic targets are also presented.

1.4.3 Chapter 4. SAbDesigner: Validation and Refinement

Chapter 4 presents the *in-silico* validation pipeline used for SAbDesigner. Methods include variation in buried surface area between native and engineered environment and structure displacement after relaxation. Chapter 4 also includes methods for refining the designs by suggesting favourable point mutations, and optimisations to the other CDRs to increase the buried surface area of the designed antibody. The methodology is presented by applying it to the initial designs for IL-5 from the previous chapter, and we created a further 10 designs that resulted after applying the methodology.

1.4.4 Chapter 5. Conclusion and future work

Chapter 5 includes the further directions that the research can take. It briefly summarises the results of each chapter and at each point will propose further research venues that can explored. This section will also present the current progress in experimental validation, and comment on the results.

I was asked if a group of Romanian men moved in next to you, would you be concerned? And if you lived in London, I think you would be

— Nigel Farage MEP

Context: I am a Romanian national

2 Antibody CDR loops structural diversity

Contents

2.1 Introduction

The Complementarity-determining region (CDR) loops of an antibody play an important role for its specificity and affinity, and as such modelling their structure from sequence is one of the prevailing topics in antibody modelling and design. For all but one of the CDR loops, the H3, modelling is aided by the fact that their backbone structures can be clustered into a number of

canonical forms (e.g. Chothia and Lesk (1987), North et al. (2011), Nowak et al. (2016)). Using just a few residues the canonical form and thus the structure of a CDR can be modelled. This has been validated, however, on only the approximately 2000 structures available in the Protein Data Bank (PDB), which is a tiny subset of the antibody space.

Humans are able to produce from germline through somatic recombination and affinity maturation an estimated 10^{10} distinct antibody sequences (Glanville et al. (2009), Fanning et al. (1996), Perelson and Oster (1979), Tonegawa (1983)). We, therefore, tested a more representative sample to see how much of the antibody sequence space can be modelled with the existing structures in the PDB. For this we used an Next generation sequenceing (NGS) dataset of 15 million naïve IgM antibody variable domain chains. To establish if a CDR sequence can be modelled we used a modified version of the loop modelling software FREAD (Choi and Deane, 2010), using the existing CDR loop structures from the PDB used as a template library. The ability to model a sequence was determined if the CDR sequence passes our established modelling thresholds for FREAD.

For CDRs L1, L2, L3 and H1 we found that the known structures can accurately model the majority of loops. For the H2 and the H3 we observed that the majority can not be modelled. In the case of H2 we believe that this is a result of an artefact of our loop modelling algorithm. In the case of H3 we believe this is an accurate picture of the ability to model H3 CDRs, although the fact that more than a third of sequences can be modelled was surprising.

In previous structural studies the H3 was also the only CDR for which computational methods consistently fail to produce sub-angstrom models (Almagro et al. (2014)). H3 is considered to be the most important CDR for antigen binding. When bound to an antigen the H3 is located in the centre of the binding site (Zemlin et al. (2003)). It also gains the most mutations through affinity maturation (Clark et al. (2006)) and has on average the largest

number of contacts with the antigen (MacCallum et al. (1996)). This makes accurate modelling of H3 vital.

We then analysed what makes the H3 loop so hard to model. Previously, a number of theories for the difficulty in H3 modelling have been proposed. It is known that H3 loops sample a large number of conformations through the process of somatic recombination and somatic hypermutation (Tonegawa (1983)). A kink in the C-terminal end of CDR H3 has been previously hypothesized to be enable this high H3 structural diversity (Weitzner et al. (2015), Teplyakov et al. (2016)). It could be this larger diversity that prevents accurate modelling. A computational study has also suggested that H3 loops are highly flexible, owing to their longer residue sequences and reduced number of stabilising bonds (Babor and Kortemme (2009)). This could make modelling highly challenging. The length distribution of H3 is much broader than for other CDRs and the number of solved crystal structures could be too low to effectively allow for the clustering of shapes (Kuroda et al. (2012)).

In this chapter we have analysed H3 loop flexibility through a systematic study of the normalized temperature factor and show that H3 structures in the PDB are if anything less flexible than general protein loops. Given that H3 is not more flexible than other loops we explored in detail what differentiates it. We compared the structures of H3, the other five CDRs and 18 other loop sets from well populated superfamilies to a non-redundant set of structures from the PDB. We found that H3 contains by far the largest percentage of unique conformations ($\sim$30%), on average 10 times more than the other loops. Next, we analysed the regions within the H3 loop which cause these differences. We found over 1,000 four residue fragments which adopt conformations not seen in any other structure. These fragments are consistently found in the area around the tip of the H3 loop and show a high propensity for Tyrosine and Glycine in unfavourable conformations. These results suggest that H3 loops present

structural characteristics which are unique in the protein world and it is this uniqueness that allows antibodies to target the highly diverse space of antigen structures, but also makes them difficult to model computationally.

The work in this chapter can be found in two publications. The analysis of the NGS data is part of a broader article in collaboration with other members of the Oxford Protein Informatics group, which is submitted and under review. The work on the structural uniqueness of the H3 is published in *Regep, Cristian, Guy Georges, Jiye Shi, Bojana Popovic, and Charlotte M. Deane. "The H3 loop of antibodies shows unique structural characteristics." Proteins: Structure, Function, and Bioinformatics (2017).*

2.2 Methods

2.2.1 CDR structural variability

2.2.1.1 Next generation sequencing dataset

UCB provided an NGS dataset of 15 million full sequences of antibody variable domains collected from approximately 500 people. An NGS dataset for proteins is created by generating the reverse transcription of RNA molecules from a cell into cDNA. To ensure that only the cDNA of antibodies is reverse transcribed from the sample (which can contain RNA from other proteins) reverse oligonucleotides that are specific for antibodies are used (Frese et al., 2013). The resulting cDNA is then amplified through PCR and sequenced through high throughput sequencing by fluorescent labelling.

The sequences in this dataset comprise naïve antibodies (antibodies that have not gone through affinity maturation) and memory IgM antibody molecules sourced from peripheral blood, bone marrow and spleen. The make-up in terms of descent, organs and age is further detailed in Table 2.1. The sequences have been provided as single variable domain chains, but without heavy-light chain

Number of donors	Immune cell harvest organ	Descent	Age
426	peripheral blood	Asian	18-52
56	bone marrow	Asian	22-85
12	spleen	Caucasian	18-54

Table 2.1: The make up of the NGS data set of 15 million sequences. The antibodies have been harvested from different areas of the body, and for each organ the number of donors, age range and the racial background is reported.

pairing. In 2015, when this dataset was provided, it was one order of magnitude higher in the number of human donors and two orders of magnitude higher in the number of chains sequenced than any other publicly available dataset (DeKosky et al. (2016), DeWitt et al. (2016)). This combination made it the best representative snapshot of the antibody sequence space available to us.

2.2.1.2 Dataset processing

The original files containing the DNA reads from high throughput sequencing were processed with IgBlast 1.4.0 (Ye et al., 2013) using the Human V, D & J germline references. The output from IgBlast was then processed, and the sequences that satisfied the following criteria were retained:

- the sequence has to contain at least the human germline V & J genes
- the sequence needs to cover the full length of an antibody variable domain, with a maximum of two base pairs missing at 5′ or 3′ ends
- the sequence may not contain ambiguous nucleotide calls or stop codons

The NGS dataset consists of approximately 13.5 million unique complete V-gene sequences divided into κ (approximately 5 million chains), λ (approximately 3.5 million chains) and heavy (approximately 5 million chains), covering 47 out of 52 functional heavy genes, 39 out of 39 functional lambda genes and 31 out of 33 functional kappa genes annotated in the IMGT database (Lefranc et al., 2003).

There are cases where two exact copies of the exact same antibody sequence are found in the dataset. In order to remove bias induced by this chain redundancy the unique full chain sequences were extracted from these into a dataset which will be referred to as "AA-constrained" from now on. The resulting amino acid sequences were then numbered according to the Chothia numbering using the ANARCI antibody numbering program (Dunbar and Deane, 2015). The framework regions and CDRs have then been delimited and extracted according to the Chothia structural definition (see Section 1.2.2.2).

The work on harvesting the antibodies and sequencing was performed by UCB Pharma and the work on parsing and cleaning the reads into amino acid format with the Chothia numbering attached was performed by Dr. Konrad Krawczyk from the Oxford Protein Informatics Group.

2.2.1.3 FREAD loop modelling

To model the structure of the CDR loops in the NGS dataset we used a modified version of the FREAD loop modelling protocol (Choi and Deane (2010), Choi and Deane (2011)). The algorithm was modified because the task performed here is different as the structure of the rest of the protein can not be provided as input. To understand better the changes the initial steps of the algorithm are described below.

FREAD was developed to model missing segments in protein structures. These segments are usually loop regions. FREAD models the loop using an existing fragment of protein structure from a pre-compiled library. The library of fragments contains the actual structure, its amino acid sequence, an encoding of the dihedral angle combination of the backbone of the loop, and the Cα distances between amino acids of the anchors of the fragment. The anchors are two residues upstream and two residues downstream from the fragment. The

dihedral angles are encoded using a discretization of the dihedral angle space Choi and Deane (2010). The standard inputs for the algorithm are the following:

- the amino acid sequence of the loop to be modelled
- the structure of the rest of the protein

When trying to identify a suitable model for a loop FREAD first calculates the Environment Specific Substitution (ESS) score for each fragment of identical length in the library. The amino acids from the target loop and the library fragment are paired according to their position in the loop, and for each pairing a score is attached. The score is extracted from an Environment Specific Substitution Table (ESST) which takes into account for a given position amino acid change (from library fragment to target fragment), and the dihedral angle bin of the library residue. The ESS score is the sum of the scores for each position in the sequence. Choi and Deane (2010) showed that a total ESS score greater or equal than 25 for a library fragment can yield a prediction with a reasonable degree of accuracy. These potential hits are then further filtered by only keeping those where the superposition RMSD of the backbone anchor structures (between the target and library fragment) is less than 1.0Å. If these criteria are satisfied the library fragment is considered to be a suitable structural model (hit).

When modelling the sequences from the NGS dataset the structure of the rest of the antibody is not provided. How we overcome this limitation is described in Section 2.2.1.5. The original algorithm also checks for clashes between the model loop and the rest of the structure, and also ranks the loops when multiple possible hits are found. In this study we are only interested in identifying if a model exists, and not selecting the best possible one, therefore these were not relevant.

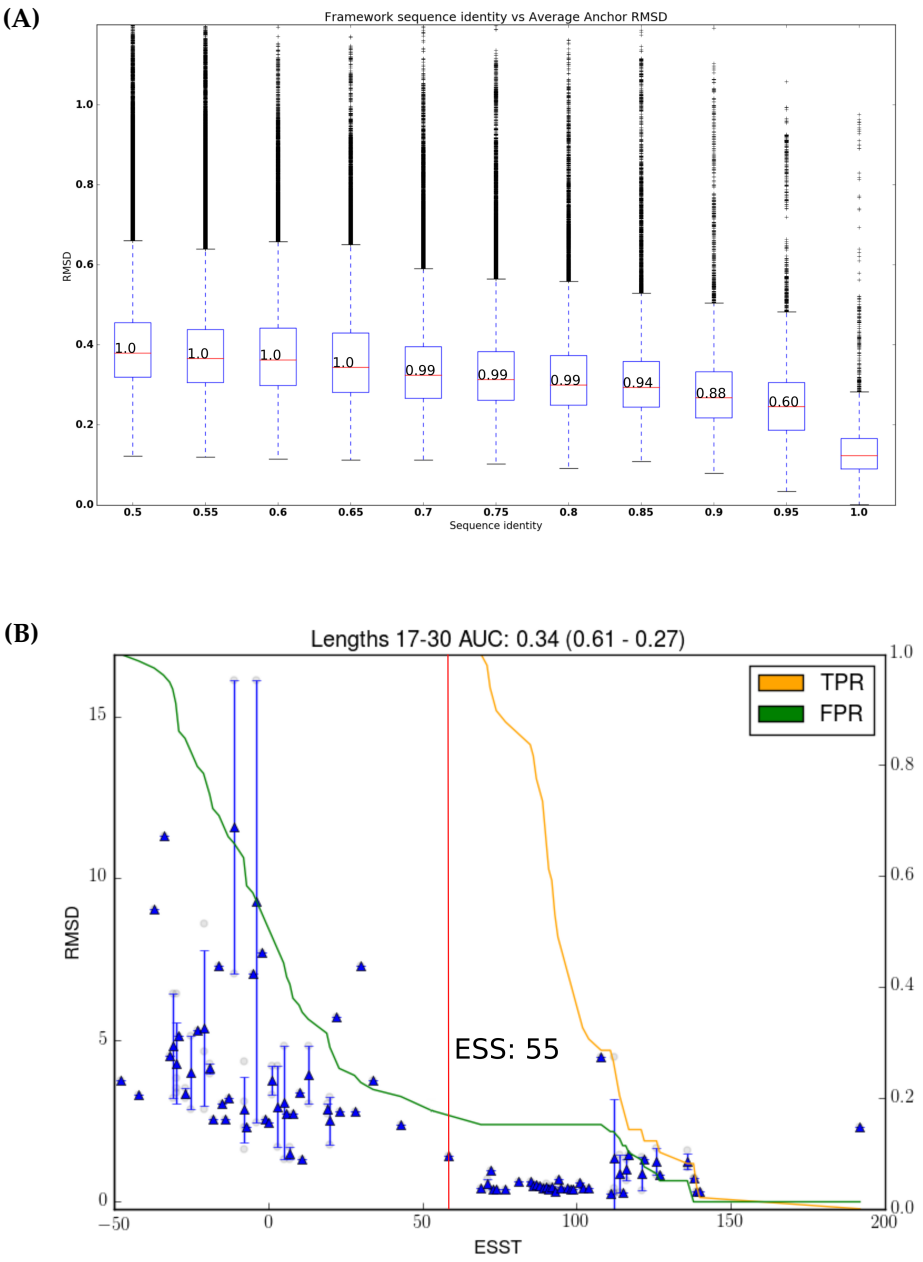

Figure 2.1: (A) Average anchor RMSD for all CDRs at different values of framework identity. Inside of each boxplot the coverage (the percentage of frameworks that have a model at that sequence identity) is reported. The 1.0 bin was calculated separately from frameworks with identical sequence to show the natural variation in anchor RMSD.
(B) The one vs all cross-validation for the ESS scoring system for lengths 17-30. All of the prediction instances are represented as lightly shaded circles with the horizontal axis showing the ESS score and the left hand vertical axis showing RMSD. At each ESS position where there is a prediction the False positive rate (FPR) and the True positive rate (TPR) curve is shown, with the right hand vertical axis being for rate values. The threshold is selected at the point where the FPR drops below 0.2

2.2.1.4 Loop structure library

The first stage in our protocol is to build a sensible fragment library. Choi and Deane (2011) found that when predicting loop structures for antibodies the fragment library should be composed of CDR loops alone. We therefore created a database for each CDR, composed of the existing loop structures of the particular CDR. The structures were extracted from SAbDab (Dunbar et al., 2014) on the 27 of November 2015. At that time there were approximately 2000 structures of antibodies. These structures are highly redundant with only 685 unique sequences for L1, 435 for L2, 778 for L3, 538 for H1, 736 for H2 and 1004 for H3 from the 2000 antibodies. The unique sequences were obtained by removing all the sequence duplicates (i.e. sequences with a hamming distance greater than 0).

When introducing a loop to the library FREAD will also split the loops in fragments of lower lengths. For example for a loop of length seven aside from the original loop the two fragments of length six, and the three fragments of length five, and so on up to length three will be introduced in the library. Our interest for this analysis was to see the amount of loops that can be modelled by exact loop matches, and therefore we do not include fragments of CDR loops.

2.2.1.5 Framework model

When using the NGS dataset we do not have access to the structure of the antibody framework, and as a result the anchor residues. It is therefore necessary to model the framework. For each chain in the NGS dataset we identified its closest sequence match to known PDB frameworks. If the highest scoring match was at over 90% sequence identity it was used as a template for the framework.

The sequence identity was calculated using the Chothia numbering for alignment. The 90% sequence identity cut-off was selected as it is the closest threshold in terms of anchor RMSD to the best possible anchor RMSD where the majority of frameworks can be modelled (see Figure 2.1A). This analysis was

Loop length	FREAD threshold
<13	25
13-16	40
>16	55

Table 2.2: ESS score thresholds used in modelling. The original ESS score had a threshold of 25 for a loop to be considered a potential hit. For our purposes this algorithm has been modified to work in the absence of the framework structure, and the thresholds were recalculated. The new thresholds are dependent on the amino acid length.

performed on a non-redundant set of antibody frameworks (both light chains and heavy chains without CDRs) from SAbDab.

2.2.1.6 New thresholds

Introducing the change may have an impact on the quality of the predictions, so we then tested whether the threshold of 25 for the FREAD cut-off score was still valid. The way we tested this was by performing one vs all cross-validation. Each loop in turn from the library was predicted using the rest of the library, removing loops with an identical sequence. We then calculated the ESS score where the False positive rate (FPR) drops below 20%.

The FPR is defined as the proportion of false positive cases at a given ESS threshold out of all the negative cases. A negative case is defined as an instance where two loops have an RMSD>1.0Å , while a positive cases is where their RMSD<1.0Å. A false positive case is where a loop is predicted to be positive, but is in fact negative. We found that for loop lengths below 13 the cut-offs should remain the same, but for lengths greater than that they should be increased (see Figure 2.1B). The final length dependent cut-offs are detailed in Table 2.2.

It is important to note that when calculating the FPR and TPR at a given ESS threshold we use all predictions from all targets. For example, if for one target loop there are three predictions over the ESS threshold with the top one being

a false positive, and the other true positives all three are used in calculating the TPR and FPR, and not just the top one. The reason we are doing this is because we are using FREAD as a method of estimating if a suitable model can be predicted, and not as an absolute predictor of the exact structure.

2.2.2 CDR H3 analysis

2.2.2.1 Antibody CDRs

We took all of the Fv chains found in the SAbDab database (Dunbar et al. (2014)) on 8^{th} of October 2015. The ones in PDB files with resolution >3.0Å were removed. This resulted in 1,779 structures with 4,989 chains. From these chains the CDR loops were extracted according to the *Chothia* definition using the ANARCI numbering software (Dunbar and Deane (2015)). In order to avoid loops with atoms that have high positional uncertainity we discarded the ones that have backbone atoms with a temperature factor higher than 80.0.

2.2.2.2 Loops from other superfamilies

The control loops for our analysis were compiled from other large structural superfamilies. Eighteen superfamilies were selected by randomly picking from those superfamilies that have more than 500 loops with unique sequences. We used the SCOP superfamily assignments (Murzin et al. (1995)) and the Superfamily package (Gough et al. (2001)) to predict the superfamily for the chains in the PDB that do not already have a manual assignment. Loops were then extracted from these chains as a region of more than three residues between two secondary structures as annotated by DSSP (Kabsch and Sander (1983a)). The superfamilies and number of loops are detailed in Table 2.3.

Scop ID	Superfamily	Class	Loops
50939	Sialidases	All β	979
51604	Enolase C-terminal domain-like	α/β	1976
50249	Nucleic acid-binding proteins	All β	896
51569	Aldolase	α/β	2829
49785	Galactose-binding domain-like	All β	1125
51556	Metallo-dependent hydrolases	α/β	2323
50494	Trypsin-like serine proteases	All β	2060
48264	Cytochrome P450	All α	1746
48557	L-aspartase-like	All α	850
49503	Cupredoxins	All β	1303
48208	Six-hairpin glycosidases	All α	1044
53187	Zn-dependent exopeptidases	α/β	1475
55486	Metalloproteases ("zincins"), catalytic domain	$\alpha+\beta$	1380
56672	DNA/RNA polymerases	Multi	1971
56235	N-terminal nucleophile aminohydrolases	$\alpha+\beta$	1539
56601	beta-lactamase/transpeptidase-like	Multi	1809
81296	E set domains	All β	1022
52518	Thiamin diphosphate-binding fold	α/β	1396

Table 2.3: The Scop IDs of the superfamilies from which the control loops are extracted, along with the a description and the number of loops. These are used in the study of H3 structural divergence.

2.2.2.3 Non antibody like protein loops

For the comparison to general protein loops we used all the loops from every chain in the PDB that has a resolution better than 3.0Å and is not antibody-like (Ig-like). We used DSSP (Kabsch and Sander (1983a)) as described above to define loops. Loops which have backbone atoms with a temperature factor higher than 80.0 were removed. We defined a chain as being Ig-like if it is either in an antibody chain included in SabDab (Dunbar et al. (2014)) or contains in the PDB description field terms related to MHCs or T-Cell Receptors.

2.2.2.4 Bound loop definition

In some tests we split CDRs into different categories depending on whether the antibody in the PDB file is bound or not to an antigen. A CDR is considered to be bound if it is part of an antibody-antigen complex as indicated by SabDab (Dunbar et al. (2014)). For non-Ig proteins a loop is considered to be bound if any of its atoms are within 5.0Å of any atom from a residue found on a different chain in the same PDB structure.

2.2.2.5 Non-redundant set of protein structures

A non-redundant set of protein structures was created by culling the chains in the PDB with resolution better than 3.0Å , at 90% sequence identity. To perform this we used PISCES (Wang and Dunbrack Jr (2003)). This resulted in 31028 chains with an average number of 260 residues per chain. From these chains we extracted all overlapping fragments between three and 30 residues.

2.2.2.6 Temperature factor normalization and flexibility

The temperature factors were used as a measure to compare flexibility between structures. This comparison is, however, difficult because the uncertainty of an atom position increases with a decrease in resolution (see Figure 2.2).

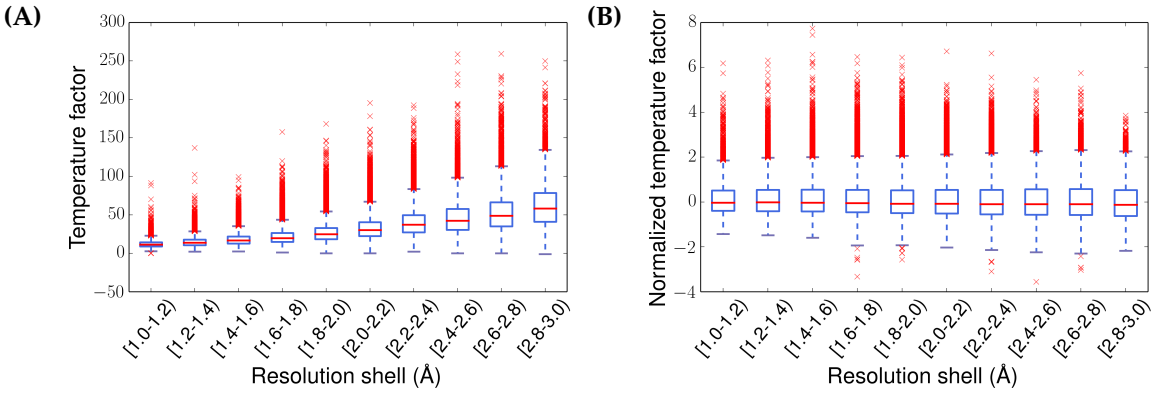

Figure 2.2: Variation of temperature factor (A) and normalized temperature factor (B) with increase of resolution in a non-redundant set of protein loops. The temperature factor is seen to increase with the resolution of the structure while the normalized temperature factor averages around zero for all resolution shells.

Hence, using an average temperature factor for comparing flexibility would be biased by resolution. We therefore normalized the value of each temperature factor to the Z-score (the mean and the variance are calculated from all the temperature factors of the backbone atoms in the specific PDB file) as suggested by Parthasarathy and Murthy (1997). Using this method we observe that the median of the normalized distribution does not increase with resolution.

Alternate conformations could potentially offer a more accurate picture of the flexibility of H3 loops, but there are very few structures that have backbone atoms for the H3 loop with multiple occupancies (one example is PDB structure with accession code 2VXU, chain H, residues 95 and 96), meaning it cannot be used at this time.

2.2.2.7 Length matched sets

When comparing two sets of loops the result might be biased by the fact that their length distributions are different. For example longer loops have more degrees of freedom and may be more flexible. To remove this potential bias we generated length matched sets (LMS). If set B is compared to set A, and B

has a different length distribution to A, a sample from B is randomly extracted without replacement such that at each length it matches the proportion of loops of that length in set A. For example if in set A 5% of loops have length 6, 3% length 9 and 2% length 12, then LMS(B) will be a sample of B which has 5% of loops at length 6, 3% at length 9 and 2% at length 12.

2.2.2.8 Unique loop fragments

We analysed whether the H3 contains unique fragments in comparison to non-Ig protein loops. We define a fragment as a continuous chain of four amino acids. The set of fragments of a loop consists of all its overlapping four residue fragments (e.g. for a loop of length five there are two overlapping fragments of length four). We also added an anchor of two residues (upstream and downstream) to the loop to account for the difference between the DSSP loop definition and the CDR definition. Two fragments are considered to be structurally different if the superposition of their backbone atoms (see Section 1.1.5) has an RMSD greater than 1.0Å .

The naïve method of identifying if a fragment from an H3 loop is unique requires comparison to all the non-Ig protein fragments. There are approximately $1.5*10^4$ H3 fragments and $12*10^6$ non-Ig protein loop fragments, which would result in a total of $18*10^{10}$ comparisons. This is computationally infeasible and we therefore decided to cluster the 12 million fragments into unique shapes to reduce the number of comparisons.

The clustering of 12 million four residue fragments is also a computational intensive task. A standard clustering algorithm requires as input a distance matrix between all the elements to be clustered, which in this case would amount to around 10^{14} initial comparisons. We therefore developed a sequential clustering method that resembles the sorting mechanism behind *Insert sort*

(Friedman, 2000). The algorithm is summarised in the following and described in detail in appendix section A.1:

- the algorithm starts from a list of fragments, and no clusters.
- each fragment in turn is compared to the representative of each existing cluster (the first fragment is made automatically a cluster). if it is unique it is added to the list of clusters, if it is not it is discarded.
- the algorithm continues until the list of fragments is exhausted.

Through this method the four residue fragments were reduced to a list 64,830 unique shapes. We considered an H3 fragment to be unique when its closest structural neighbour from the 64,830 clusters of non-Ig shapes has an RMSD > 1.0Å.

2.2.2.9 Dihedral angles

To define the expected dihedral angles in loops we took non-redundant set of non-Ig loops and meshed their backbone atom Phi-Psi dihedral angle space into bins of 3.0 x 3.0 degrees. The frequency for each bin was computed and a 90% contour plot was generated. The algorithm for the contour plot used a greedy 'highest-frequency first' approach up to 90% of the density. If an angle falls out of the generated contour it is considered to be energetically unfavourable.

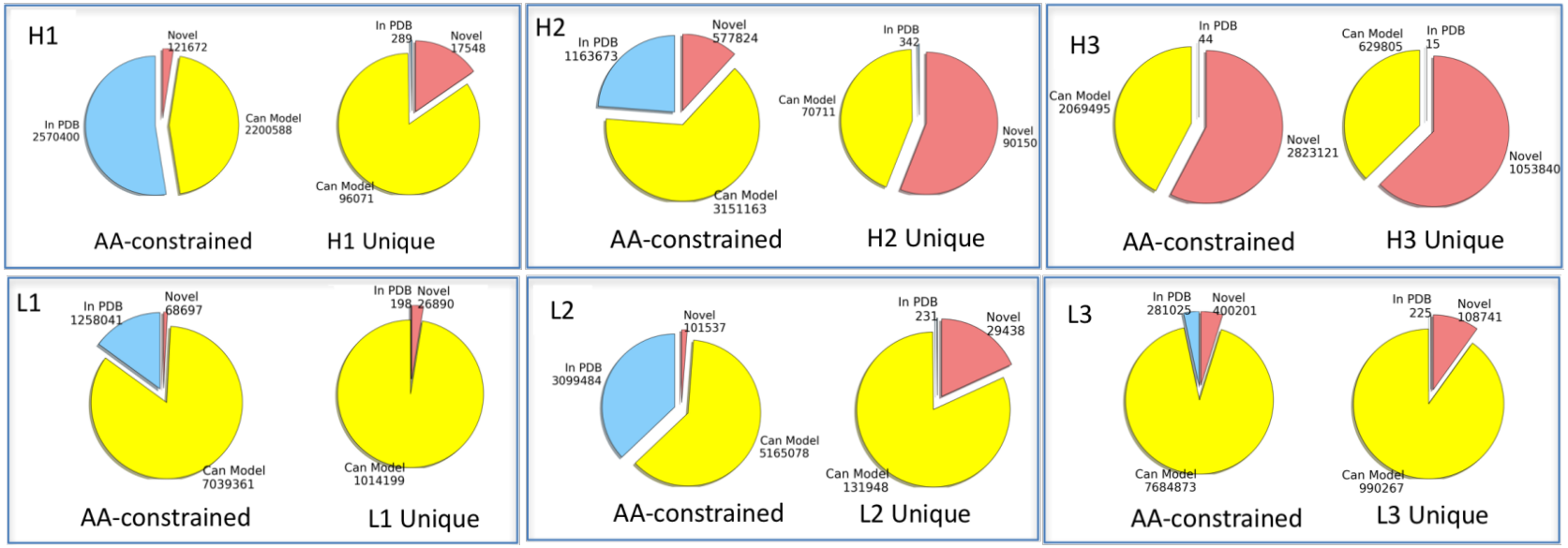

Figure 2.3: Structure prediction of the Chothia CDR loops. For each CDR, statistics are given for both the AA-constrained and CDR-unique sequence entries. CDR-unique sequences are a subset of AA-constrained. For example there are 4,892,660 H1 sequences in AA-constrained set of which only 113,908 are H1-unique, 2% of the AA-constrained set. The blue portion of the chart indicates the proportion of sequences for which we find an identical match in the PDB. The yellow section denotes the sequences for which we do not find an identical match in the PDB, but for which we could produce a reliable structural model, using the modified version of FREAD. The remaining sequences, which are potentially novel structures, are shown in red.

2.3 Results

2.3.1 CDR structural variability

Using the modified version of the loop modelling protocol FREAD we have predicted structures for the CDR sequences in the AA-constrained NGS data set. The results of this analysis are summarised in Figure 2.3, split by AA-constrained and unique data sets. Analysis on the AA-constrained set of loops reflects the preferred use of the sequence space by antibodies, while the unique set reflects the breadth of the space explored. There is a large redundancy in the NGS CDR loops, with the number of unique loop sequences being much lower than AA-constrained sequences. The H3 loop uses the most of its structural diversity (approximately 3 times less redundancy than the closest CDR loop, L3). This means that although there is a high breadth of sequences that are sampled, only a reduced set is preferred for use.

It can also be observed that the potential antibody structure space is not much more diverse than the structures already known for the non-H3 CDRs, with a considerable percentage of sequences in the AA-constrained set having identical structures available in the Protein Data Bank. In the AA-constrained sets of L1, L2 and L3 there are direct PDB matches for 15%, 37% and 3.3% sequences respectively. In the case of the heavy chain H1 and H2 there are identical PDB structures in 52.5% and 23.7% of cases. These do not need a predicted structural model as it is expected that a structure with the same sequence suffices. On top of the identical matches there is also a high percentage of sequences for which a suitable structural model can be predicted. For all but H3 the majority of the AA-constrained space is covered by suitable model predictions. In the case of H3 only 44 of the AA-constrained sequences have a sequence-identical match in the PDB, and with the ones that can be modelled it goes up to 40%.

In terms of the unique sequence space the combination of identical matches and structural models are able to provide structural data for the majority of the repertoire for all the CDRs except two, H2 and H3. In the case of the H2 loop the high number of unmodellable loops is unexpected, as it is known to present canonical forms. We believe that the results are not an accurate picture of the actual structural diversity, but are in fact an artefact of the ESST scoring system used by FREAD. 65% of the unique H2 sequences in the NGS dataset are of length 6, and they explore the sequence space at least 3000 times more than the ones at length 8 for example, which is closest in terms of sequence breadth. If a loop contains only six amino acids FREAD can to be too sensitive. One amino acid substitution weighs 66% more in terms of the total FREAD score for a loop of length six versus a loop of length 10. Structurally the H2 loop is a short β-turn, with a reduced amount of flexibility and freedom to explore the conformational space, and it is therefore likely that in light of this structural restrictions even with the many different amino acid combinations they preserve to the known canonical classes.

For the H3 CDR which is already known to have the greatest structural diversity and importance for antibody binding we observe a lower level of redundancy than for the other CDRs. There are only 15 unique sequences that have an exact match in the PDB, out of the over 1000 unique H3 sequences with structures in the PDB. This could be due to the fact that the sequences in the NGS dataset are from a naïve repertoire, and the ones for which a crystal structure is usually solved are generally high affinity binders which have gone through rounds of somatic hypermutation. The H3 is known for gaining most mutations out of all CDR lops through this stage (Clark et al., 2006). Nevertheless, it is still possible to model over a third of the H3 sequences just with the structures currently available in the PDB.

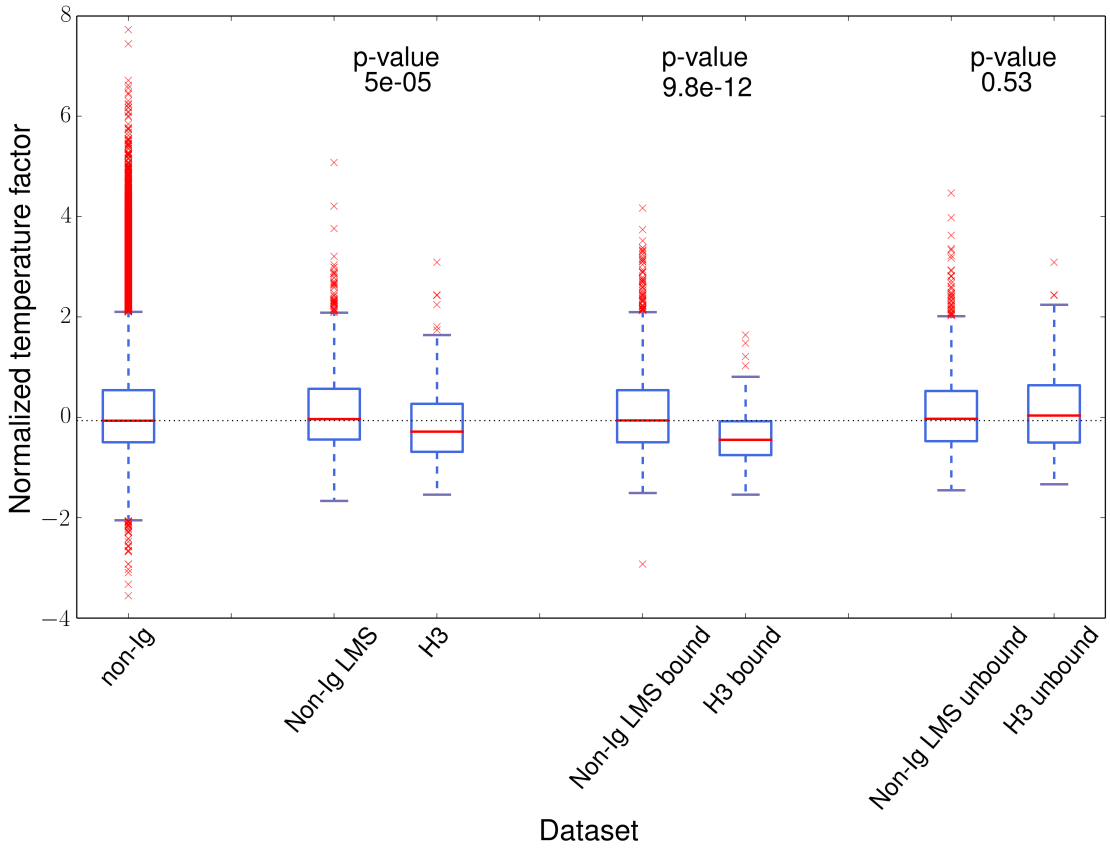

Figure 2.4: Flexibility comparison between H3 loops and non-Ig protein loops using the distribution of normalized temperature factors, one value per loop. For each of the H3, H3 bound, and H3 unbound datasets 10 length matches samples were generated from the non-Ig set and amassed to produce their associated length matched set (LMS) distribution: non-Ig LMS, non-Ig LMS bound and non-Ig LMS unbound respectively. Between each H3 loop set and its associated non-Ig LMS the p-value from a two tailed Welch t-test (Welch (1947)) is reported

2.3.2 CDR H3 analysis

In the previous section we have seen that the inability to model H3 CDRs (see Section 1.3.3.2) extends to the NGS dataset, where for close to two thirds of the unique CDR H3 sequences a suitable structure from the PDB can not be proposed. We therefore analysed the characteristics that make H3 hard to model.

2.3.2.1 Flexibility

We first tested using normalized temperature factors whether H3 loops are more flexible than other loops as a source of structural diversity. Figure 2.4 shows how the distribution of normalized temperature factors of H3 loops compares to that of general protein loops (a length matched set (LMS) is also shown to correct for possible bias from the differences in length distribution). We find that the H3 loop does not show an increased flexibility. We also considered the potential bias induced by the fact that H3 loops are found in two states: bound and unbound. It has previously been suggested that loops involved in binding are less flexible. We therefore examined the bound and unbound H3 loops separately. We observe the expected increase of normalized temperature factor in the unbound H3 loops, however there is no significant difference to the behaviour of unbound general protein loops (p-value 0.53).

2.3.2.2 Residue propensity and length distribution

We then analysed the length distribution and residue propensity distributions of all H3 loops. Residue propensity is defined as:

$$Propensity(AminoAcid) = \frac{\sum_{i=1}^{n} Occurences(AminoAcid, loop_i)}{\sum_{i=1}^{n} Length(loop_i)} \tag{2.1}$$

where Occurences(AminoAcid,$loop_i$) is the number of occurences of the amino acid in $loop_i$ and Length($loop_i$) is the number of residues in $loop_i$.

We compared these distributions to over 200,000 loops from the non-redundant set of structures (see Figure 2.5). H3 loops tend to be longer, peaking at length 10 as opposed to non-Ig loops which peak at length four. They also have a higher propensity for Tyrosine, Glycine, Aspartic Acid and Phenylalanine. These differences have been previously reported in other studies (e.g. Zemlin et al. (2003) and Birtalan et al. (2008)). However, if we carry out the same test for other CDRs (e.g. H2 or H1), H2 loops peak at length six and they have a higher

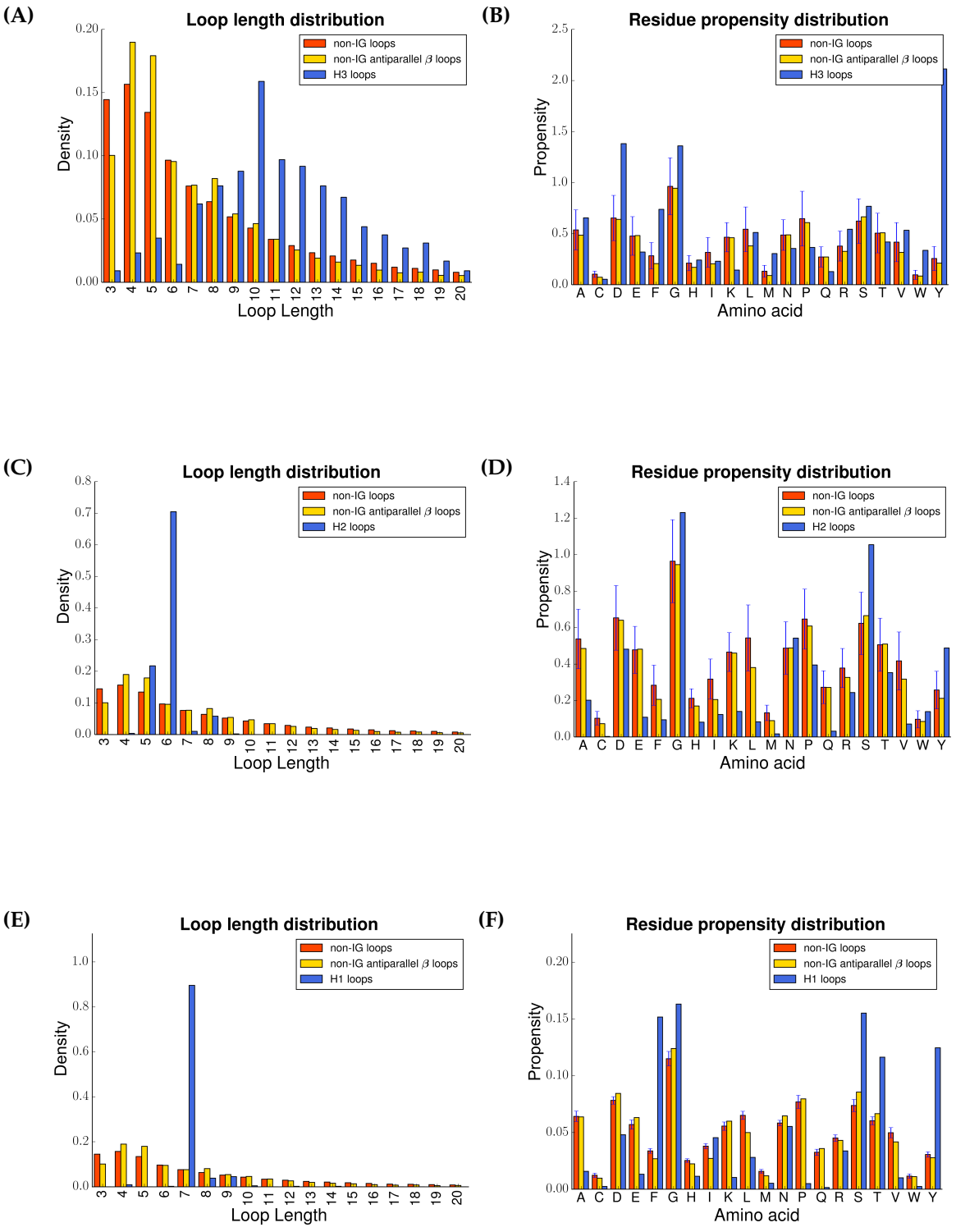

Figure 2.5: Length distribution and residue propensity of H3 (A,B), H2 (C,D) and H1 (E,F) compared to non-Ig loops (see Methods). For residue propensity distributions the error bars are obtained by generating length matched samples.

propensity for Serine and Glycine than the general set. As all these sets are just subsets of the whole this result is perhaps not surprising but it suggests that it is not just length differences or particular amino acid preferences that are the reason for the difficulties in predicting H3 loops.

2.3.2.3 Full loop structure

Given that H3 loops have a unique length and residue distribution we next looked at its structural divergence. For each of the H3 loops we computed the superposition and RMSD to every loop from all non-Ig chains in all crystal structures in the PDB with <3.0Å resolution (2,281,826 loops). We did not cull the list of chains based on sequence because loops with the same sequence in different crystal structures can have different conformations (e.g. H3 loop in structure with PDB id 3v6f chain H and H3 loop in 3v6z chain C share the same sequence but have an RMSD of 2.69Å). To represent how H3 loops and the other CDRs compare in terms of structural similarity to the rest of protein world, we plot distributions of minimum RMSD. For every loop in the query set we retained the value of the closest structural neighbour in all other proteins, excluding the query set.

As all CDRs apart from H3 adopt canonical forms (e.g. Chothia and Lesk (1987), North et al. (2011), Nowak et al. (2016)). We checked whether our results are biased by removing shape duplicates. Shape duplicates are sets of loops which have a superposition RMSD of less than 1.0Å, and for each set we retain only one loop. There are many definitions of canonical forms (these have been compared in several papers e.g. Nowak et al. (2016)). We chose to use a very simple 1.0Å RMSD cut-off as this is a standard definition of structural equivalence (e.g. Fidelis et al. (1994), Irving et al. (2001), Baeten et al. (2008), de Oliveira et al. (2015)) and one which also provides a framework for including H3 in the analysis. Figure 2.6A shows this distribution for H3 loops

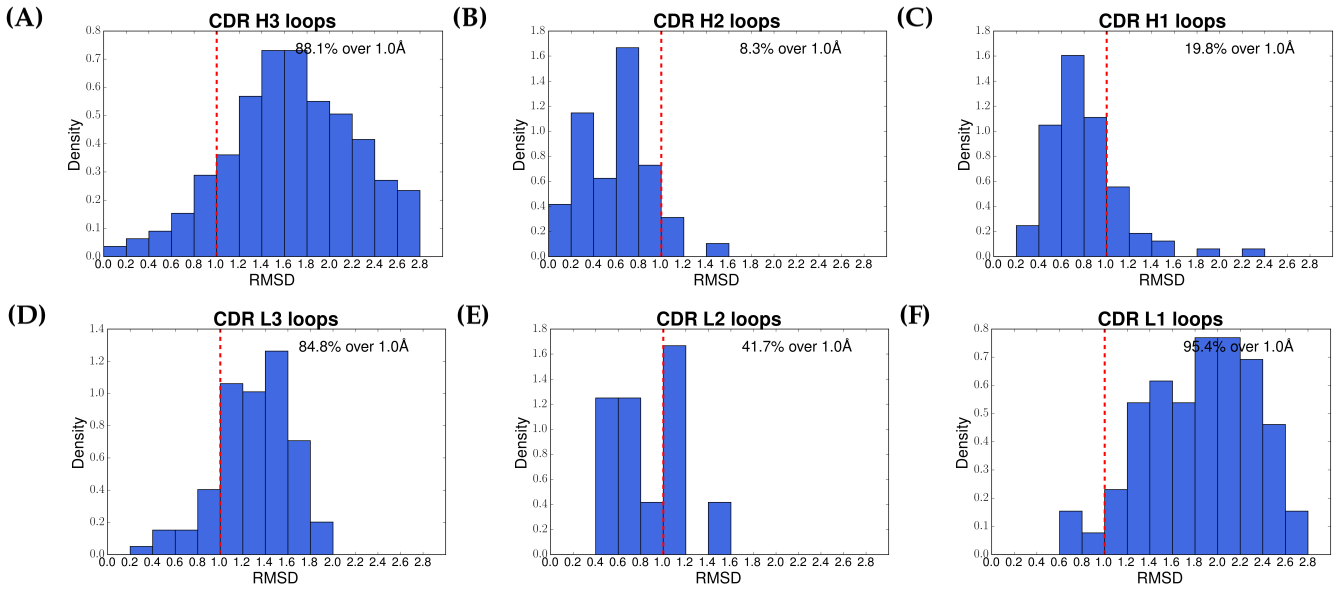

Figure 2.6: For each CDR loop the closest structural neighbour in the rest of the protein world has been identified. The distribution of RMSD between the loops and their closest structural neighbour has been summarised as a histogram to show the structural similarity between the respective set of loops and general proteins. For each CDR the percentage of loops that have their closest structural neighbour at over 1.0Å RMSD (the unique threshold) is reported. Shape duplicates have been removed in each data set. We define as shape duplicates sets of loops which have a superposition RMSD of less than 1.0Å to another loop in the data set. For each set of such duplicates we retain only one loop. In the case of CDRs H1, H2, L1, L2, L3 this is approximately equivalent to retaining only one loop for each canonical class.

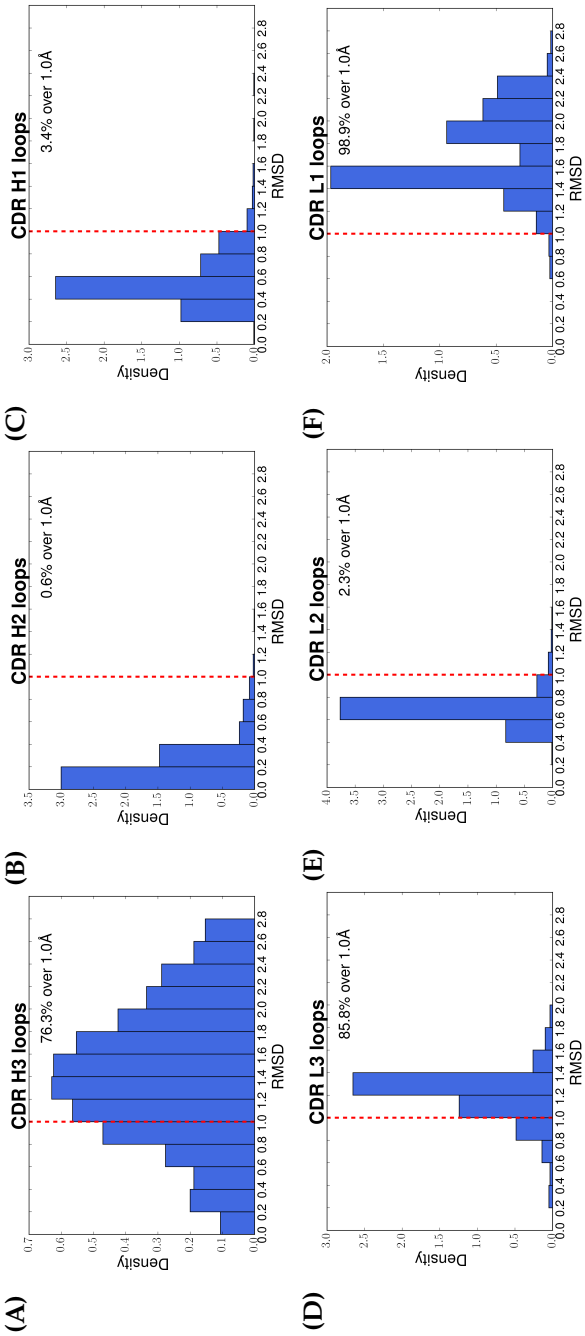

Figure 2.7: Results of the same analysis performed for Figure 2.6 without removing shape duplicates.

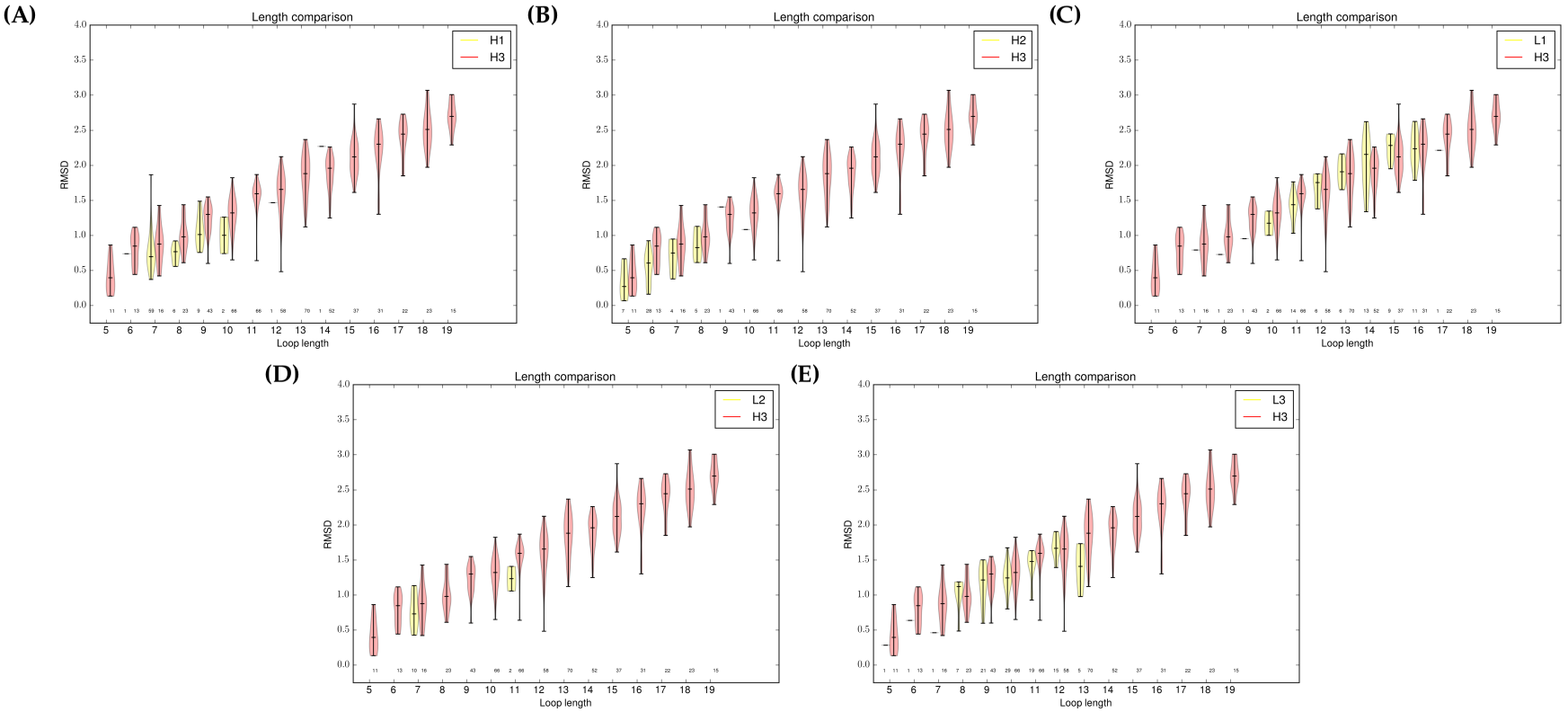

Figure 2.8: The results of the closest structural neighbour analysis from Figure 2.6 split by loop length. The H3 loop is compared to the other CDR loops. For each of the non-H3 CDRs the loops have been split in loop lengths bins, and the distribution of closest structural neighbour RMSD in each length is compared to the respective H3 loop length through violin plots. The datasets shown are H1 (A), H2 (B), L1 (C), L2(D), L3 (E).

is approximately normal, peaks around 1.5Å, and 88% of the conformations are not found in the rest of the protein world. The other five panels in Figure 2.6 (B-F) show the same data for the other CDRs. CDRs L1 and L3 also have most of their conformations over 1.0Å, while H1, H2 and L2 have most of their conformations under 1.0Å . We also stratified this analysis by length to check for length bias (see Figure 2.8), and this shows the same overall results. This was necessary to show that the results in the original analysis are not biased by the fact that H3 loops tend to be longer than the other CDRs (see Figure 2.5). We also include in Figure 2.7 the results without shape duplicates removed, where for all but CDR L2 we observe similar overall results. The difference for CDR L2 is caused by the fact that there are only seven unique shapes.

As L1 and L3 are known to take take on canonical shapes it is likely if we allowed structures from the same superfamily (in this case the Ig fold) to be included we would expect L1 and L3 to have close structural neighbours whereas H3 may well still not.

To show this we compared the CDRs to a non-redundant set of protein structures which include antibodies. This dataset consists of all overlapping fragments from 31028 protein chains (includes secondary structure as well as loop - see Section 2.2.2.5). Between 5.3 to 6.8 million fragments were compared to each loop (dependent on length). We found that H3 loops are structurally unique (have a closest structural neighbour with an RMSD >1.0Å) at least 10 times more frequently than the other CDRs (see Figure 2.9).

To show that this diversity is not only unique for H3 in comparison to the other CDRs, but also in the general protein world we also selected 18 sets of loops from highly populated SCOP superfamilies (Table 2.3) and carried out the same test (see Figure 2.9). These loop sets also have only a small number of unique structures. The largest percentage of unique structures seen for anything other than H3 is 5.6%, and the average is approximately 3%. As H3 tends

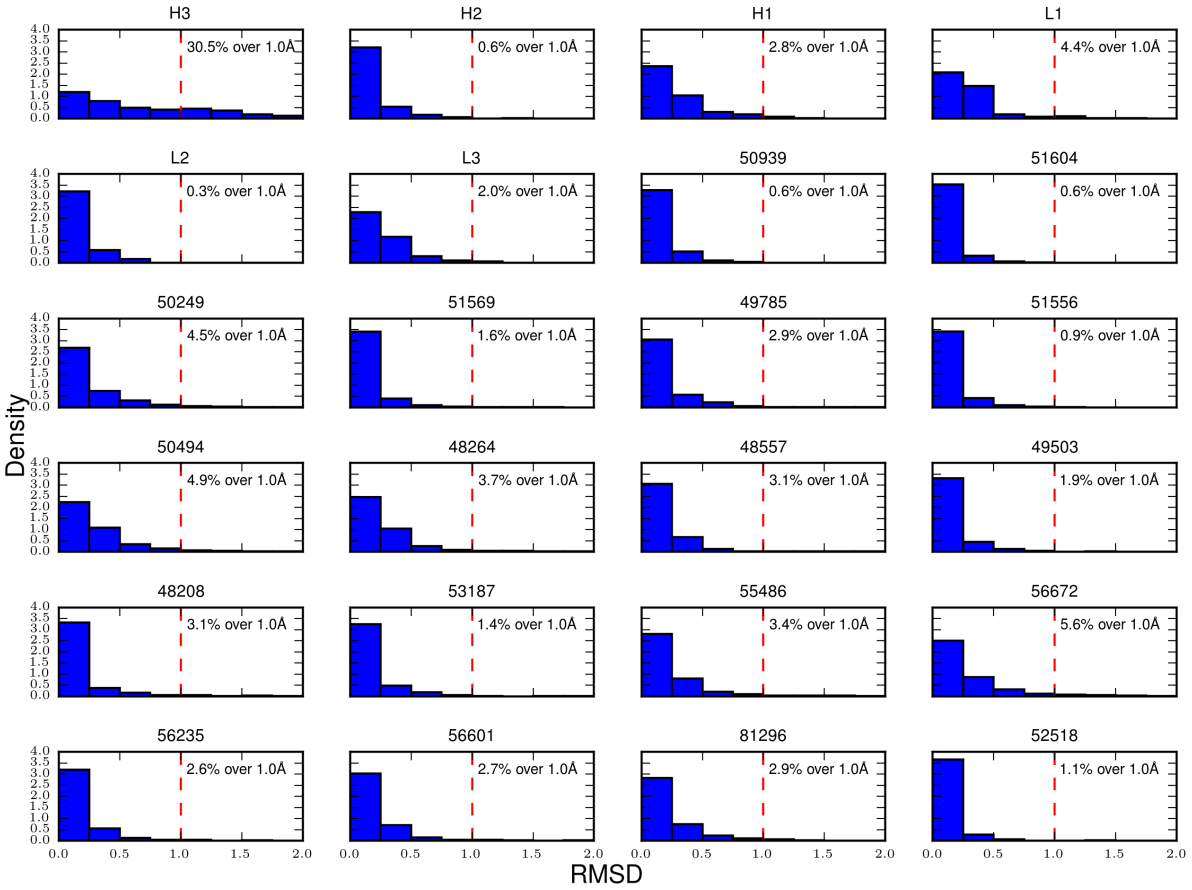

Figure 2.9: For every loop in the sets of CDR loops and the 18 sets from other superfamilies a histogram of the RMSD of their closest structural neighbour from our non-redundant set of all protein structures is shown. The 18 control loop sets are from SCOP with the ID of the superfamily being provided as a title (details can be found in Table 2.3 of the SI). The percentage of loops with no close structural neighbour ($>$ 1.0 RMSD) is given

to be longer on average than other loops, we checked whether the observed structural difference was due to this length difference. Figure 2.10 shows that for all the lengths between five and 19 the closest structural neighbour to an H3 is on average further away than for other loops.

We also checked whether our results might be affected by the fact that in each control set the loops are homologous. A key characteristic of the loops in the H3 dataset is that they are mutated by the unique process of somatic recombination and somatic hypermutation, and therefore do not share the same amount of homology that the loops from the same superfamily usually share.

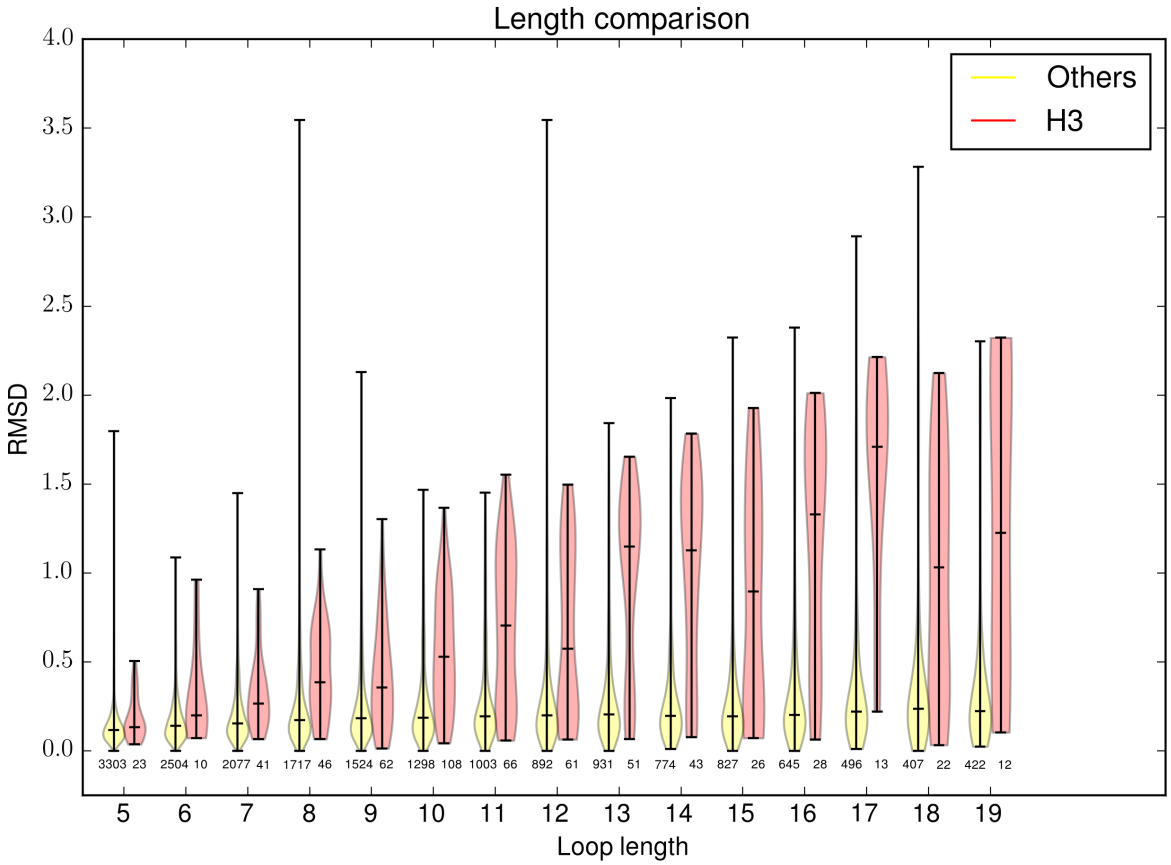

Figure 2.10: A violin plot comparing the difference in closest structural neighbour RMSD of H3 loops to the loops from the 18 control datasets at different lengths (see Figure 2.9). At all lengths the H3 sets have on average a higher RMSD to their closest structural neighbours in the non-redundant set of protein structures.

We performed an analysis where H3 is compared to five random samples of loops from all the superfamilies, the random selection removing the potential for homology. We find that the initial result holds under these new conditions as well (see Figure 2.11).

The challenge of modelling H3 appears to arise from its structural novelty. These results show that even if a perfect scoring system existed such that we could always select the closest structural neighbour as a prediction we would fail to achieve sub-Angstrom accuracy at least 88 % of the time if we used only non-Ig loops as the prediction library, and at least 30 % of the time if

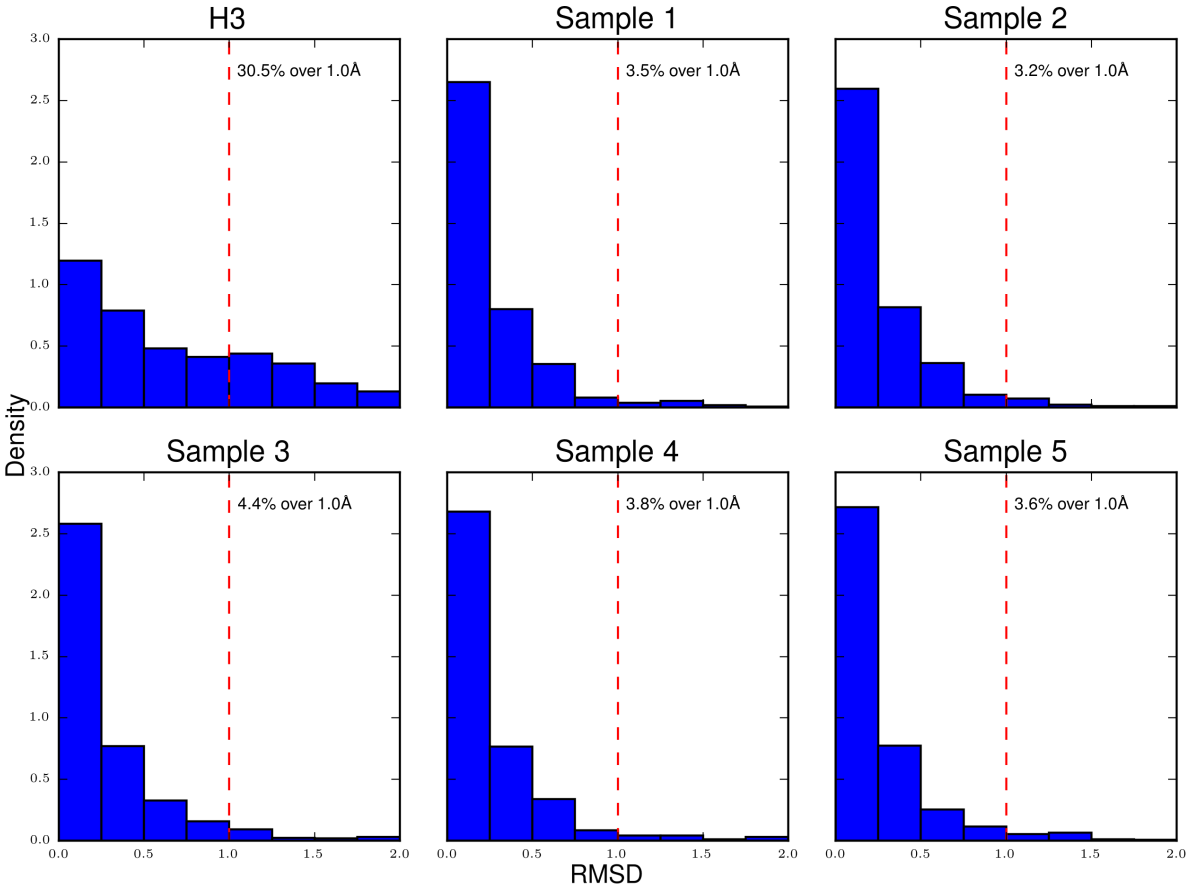

Figure 2.11: The same analysis performed in Figure 2.9, but with the Sample data sets being generated randomly from across all the 18 control data sets in table 2.3 and matching the length distribution of the H3.

antibody loops were included.

2.3.2.4 Unique fragment conformations

Next we tested whether the entire H3 or only segments of the loop are structurally unique. We extracted all the four residue overlapping fragments from every H3 loop and compared it to the set of 64,830 structurally unique four residue segments found in the rest of the PDB (see Section 2.2.2.8). We identified a list of over 1,000 fragments that are unique to H3, with over 30% of H3 loops containing at least one unique fragment. Figure 2.13 shows the characteristics of these fragments. The fragments tend to occur close to the tip of the H3 loop.

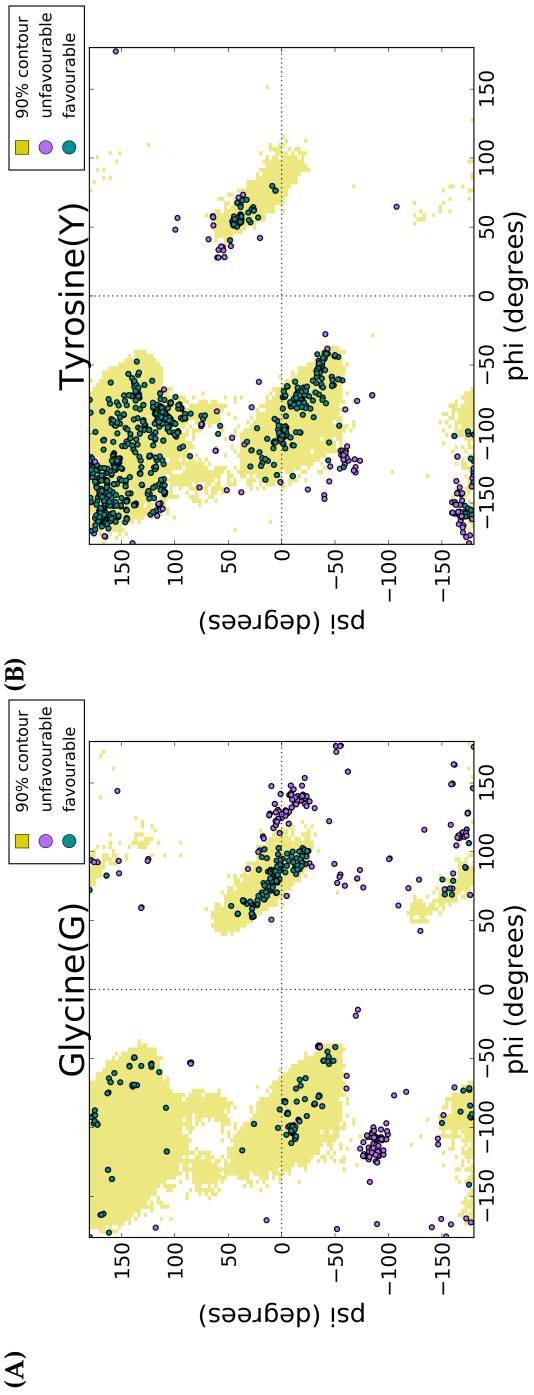

Figure 2.12: Ramachandran plots for Glycine (A) and Tyrosine (B). The background distribution in yellow is a 90% contour of the distribution of dihedral angles from a non-redundant set of protein loops. Each dot on the plot is a residue from a unique H3 fragment. A magenta dot indicates a residue with a conformation which is outside the 90% contour and therefore considered potentially energetically unfavourable. A green dot indicates a residue which is inside the 90% contour.

We define the tip as the residue in the loop that contains the C_α at the greatest distance from the C_α of the residues at the start and end of H3. In order to identify whether these unique H3 fragments have a sequence preference we calculated their amino acid propensities. We observed that the unique fragments have a high propensity for Tyrosine and Glycine, even when compared to the rest of the H3 fragments (Figure 2.13A). Tyrosine and Glycine are known to have a high propensity throughout H3 (Figure 2.5 B), but our result suggests that they are even more concentrated within the unique fragments. Examining these residues we found that the unique fragments contain large numbers of Tyrosine and Glycine residues adopting energetically unfavourable Phi - Psi angle combinations (Figure 2.12).

We initially checked if the full loop structure of H3 loops is more flexible, and we found this is not the case. The results for fragments, however, can be affected if only the four residue fragment is more flexible. We, therefore, analysed if the flexibility of these fragments is significant when compared to an average H3 fragment (see Figure 2.14). The results show that this is not the case. It, therefore, appears that the unique fragments and thus unique H3 conformations may arise from Glycine and Tyrosine residues with unusual dihedral patterns.

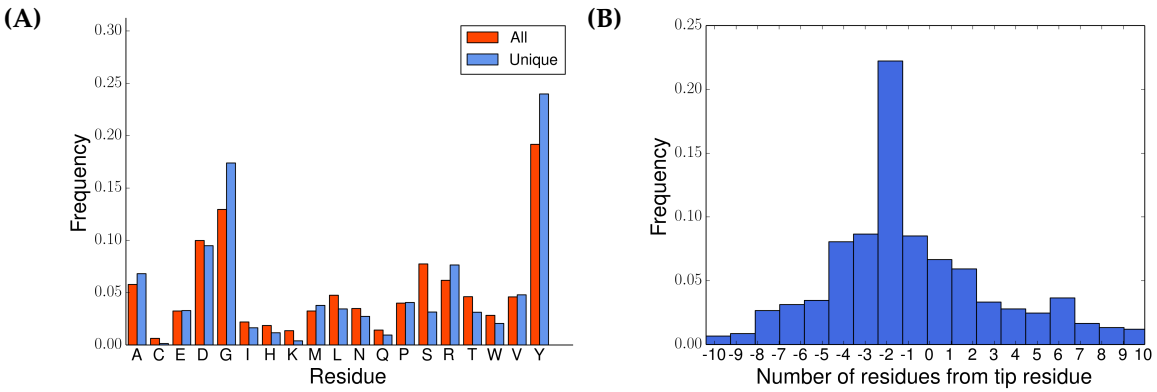

Figure 2.13: (A) Residue propensity distribution of the four residue H3 fragments in unique conformations (blue) and all of the H3 (red). (B) Distribution of the distance of the first C_α in a unique fragment from the tip of the loop.

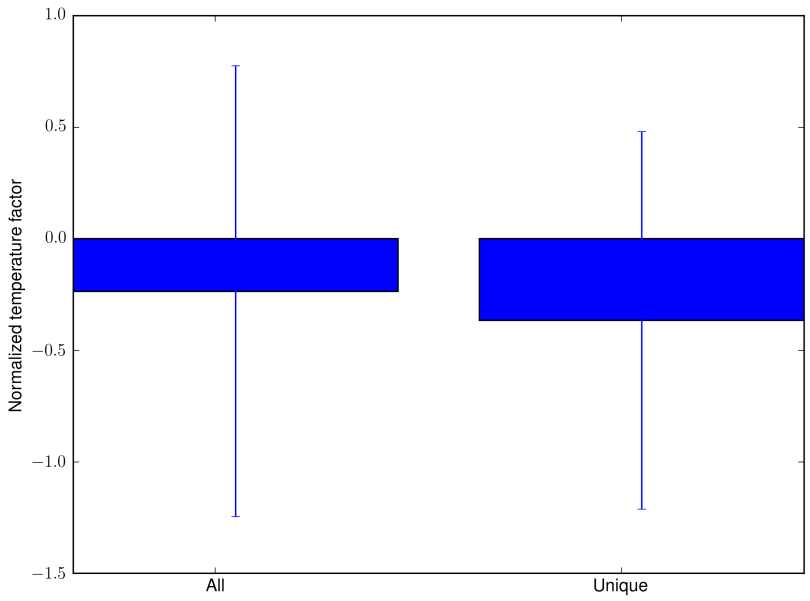

Figure 2.14: Normalized temperature factor for unique fragments in comparison with all the H3 fragments

2.4 Discussion

We analysed the CDR sequences found in an NGS data set of 15 million antibody variable domains, from the naïve repertoire of 500 humans. The results show that in the case of the non-H3 CDRs with the currently available antibody structures we are able to provide a structural model for the majority of sequences. For the H2 loop the initial results showed that over a half of the sequences can not be modelled. However, we believe that this is the result of an algorithm artefact, the sequences in the NGS dataset exploring in detail length six which is not reflected by the loop library, even though there are limited conformations available for that loop to explore and sample.

For the H3 loop only a third of the sequences in the NGS dataset can be modelled. The fact that the majority of sequences are un-modeallable is not surprising, given that the H3 loop is known to be the most structurally diverse. What is surprising is that a structure library of loops formed of

approximately 1000 unique sequences from the PDB can model one third of the approximately 1.7 million unique H3 sequences in the NGS data set. This means that the immune system in its ground state may be using only a limited set of structures to generate diversity.

We next analysed why H3 is so difficult to model.Through the process, which is unique to antibodies, of somatic recombination and somatic hypermutation the CDR loops (including the H3) are refined to achieve high affinity and specificity to target antigens. The H3 CDR loop in antibodies is often the most important loop for antigen binding. To be able to modulate binding to a very large palette of potential antigens the H3 is known to have very high structural variability. It has been previously suggested that a source of its structural variability is an increased flexibility because of its longer length and lack of stabilising bonds. However, the same study suggested that affinity matured antibodies present rigid backbone conformations. What we observe is that the antibodies present in the PDB do not show an increased flexibility when compared to general protein loops. This could be because most crystallised antibodies are matured high affinity binders. Nevertheless, high flexibility is not present and can not explain the difficulty in modelling the H3 loops of the structures in the PDB.

What we did identify is that H3 loops are distinctive in their structural characteristics and diversity from other loops. Thirty percent of H3 loops are unique compared to a non-redundant set of PDB structures, on average 10 times more than our control datasets. Also, 88% of these H3 loops do not have a sub-Angstrom structural neighbour in non-Ig proteins. This result is mirrored by the fact that some of the best predictions in the Antibody Modelling Assessment (Almagro et al. (2014)) relied on physics-based approaches. To try and understand the origin of these unique H3 structures we examined all four residue fragments from H3s and found over 1,000 unique four residue fragments. These fragments have conformations which are not seen in the rest of the PDB.

A high proportion of these fragments are found in close proximity to the tip of the H3 loop. We also observed that these fragments have increased levels of Tyrosine and Glycine compared to other H3 fragments which already have high levels of these amino acids. The uniqueness is further cemented by the fact that these residues are seen to adopt energetically unfavourable dihedral angles, which could be the reason for the structural diversity we observe. These results are a strong indication that the use of fragments of known structure from non-Ig proteins will not be effective in attempts to model the H3 loop to sub-Angstrom accuracy. There is therefore a necessity to develop modelling methods which focus specifically on the characteristics of these unique loops.

Boris, he's the life and soul of the party. But he's not the man you want to drive you home at the end of the evening

— Amber Rudd MP, referring to Boris Johnson MP during the Brexit debate

3

SAbDesigner: Designing antibodies using non-antibody protein loops or fragments

Contents

3.1 Introduction

In the previous chapter we looked at the diversity of the natural antibody sequence and structural space. In this chapter we will switch focus to expanding this diversity through SAbDesigner, a tool which computationally designs antibodies by grafting novel conformations from non-antibody proteins onto antibody scaffolds.

3.1.1 Antibody design

Antibodies have a modularity which is favourable for engineering, their binding mode being mostly determined by the residues from just six loops, the CDRs. Owing to these favourable characteristics there is great interest in the engineering and design of antibodies against therapeutic protein targets. The two most established methods for designing antibodies, animal model immunisation and phage display, involve simulating the *in vivo* evolution and selection mechanism of antibodies (Smith, 1985). These methods are not guaranteed to produce a usable hit antibody, require multiple rounds over long periods of time, and in the case of antibodies developed in animals require an extra humanisation step (Miersch and Sidhu (2012), Carter (2006)). Furthermore, they are not epitope specific which means that a lead might bind the target but not to the appropriate site and therefore may not be successful in altering its activity.

3.1.2 Computational techniques

Computational methods for rational design of antibodies are currently less well established. Existing methods range from modelling the structure of the antibody (Weitzner et al. (2017), Dunbar et al. (2016)) or individual parts (e.g

the H3 loop (Marks and Deane, 2017)) from sequence, to improving affinity and stability by point mutations when provided with an existing structure of the antibody (Lippow et al., 2007), to *de novo* antibody design.

In this chapter we focus on *de novo* antibody design. In this scenario only the structure of the target is provided and the task is to design an antibody that will bind to the antigen. OptCDR (Pantazes and Maranas, 2010) performed *de novo* antibody design against a specific epitope by assembling many potential antibody CDR regions from a database of canonical forms, ranking combinations based on computed interaction energies between the potential antibody and the antigen. The designs were then further refined through iterative random point mutations. AbDesign Lapidoth et al. (2015) used a similar approach but put an emphasis on molecularly matching the fringes of the canonical forms to the fringes of the framework in order to produce stable CDRs, and also to only sample known sequences of antibodies instead of random point mutations for their affinity maturation. Liu et al. (2017) take a slightly different approach by grafting the residues of a non-antibody motif which is known to bind to the target, but only if the backbone structure already exists in an known antibody CDR. These methods share a common drawback, the conformations used to design the binding loops of the antibody have to be from known antibodies. A more detailed description of these methods is available in section 1.3.3.3.

3.1.3 Loop grafting and SAbDesigner

Loop grafting is an established method for humanising antibodies. By grafting a CDR loop from one antibody Fv to another in many cases both native structure and function can be transferred and preserved (e.g. Jones et al. (1986), Nicaise et al. (2004) and Nakano et al. (2010)). Transferring binding specificities by grafting a loop from antibodies to non-antibody proteins has also been successful (e.g. Smith et al. (1995), Kohli et al. (2009)).

Grafting of loop fragments from non-antibody proteins to antibodies in order to transfer specificities has also proven successful, with the antibody Fv showing great versatility in accepting a variety of structures. Liu et al. (2015) grafted an acyclical Trypsin inhibitor peptide into the H3 loop of an antibody, achieving similar or higher levels of inhibition and a longer half life than the native peptide on its own. On a larger scale Peng et al. (2015) inserted into CDRs an entire binding domain by using highly flexible junctions, with the binding mode of the domain being preserved. The previous examples have in common the use of structures that are not of antibody origin, but each is an individual case study using human intervention to redesign the antibody. Sormanni et al. (2015) succesfully grafted a non-atibody motif on the H3 loop that forms β-strand bonds with the epitope on a disordered protein. This method was further extended by Aprile et al. (2017) to succesfully bind and disrupt aggregation of Aβ42 on multiple epitopes.

SAbDesigner further explores this principle, instead of grafting an existing CDR loop conformation onto an antibody framework we graft loops of novel conformation from non-antibody proteins that are known binders of the target. We show that for a large palette of therapeutic targets there exists a publicly available structure in the PDB of it in complex with a cognate protein (which will be referred to as the receptor). For these structures SAbDesigner proposes ways to mimic the binding site of the receptor on an antibody by selecting the loops which participate in binding and then automatically identifying ways in which they can be grafted on an antibody scaffold. These loops may well have conformations which have not been seen in known antibody structures and may not be easily sampled by antibodies from the germline in an immunisation scenario.

3.2 Methods

3.2.1 Non-antibody loops

3.2.1.1 Loop definition

The first step of SAbDesigner is to identify the residues and loops of the receptor that participate in binding to the target of interest. The residues which form part of the binding interface are defined as residues which have an atom within 5.0Å of an atom on the target. Receptor loops that contain any identified interface residues are retained. Loops are characterised using DSSP (Kabsch and Sander, 1983b). A loop is defined as any non-interrupted sequence of three loop residues or more between two areas of secondary structure, each of which must be at least three residues long.

3.2.1.2 Buried surface area ranking criteria

Our primary criteria for ranking the importance of a receptor loop is the amount of loop surface that is buried by the target. This is obtained by running NACCESS (Hubbard et al., 1992) on the complex and on the receptor protein in isolation. The total buried surface area (BSA) is the difference between the accessible surface area of the loop from the protein in isolation and in complex. The accessible surface area (ASA) of a loop is defined as the sum of the ASA of all its residues from the output of NACCESS. SAbDesigner considers a loop to have potential for transferring binding specificity if its BSA is over 175Å^2. This value was suggested by previous analyses of protein-protein complex formation (Hamer et al., 2010). Other studies (e.g Chen et al. (2013), Chung et al. (2006)) use higher protein complex BSA thresholds. The choice of a higher threshold potentially reflects the fact they take into consideration the complete binding area of the receptor. This is usually composed from more than a single loop.

SAbDesigner selects one loop at a time, and therefore the threshold is set to reflect what is reasonable for a single loop.

3.2.1.3 Computational Alanine scanning criteria

A second criteria that can be used to rank the importance of a loop is computational alanine scanning (CAS). In SAbDesigner the importance of a loop is defined as the average of the importance of all its binding residues. The importance of a residue is the ΔG incurred if the residue is mutated to alanine. The mutation is performed using Rosseta with repacking of the neighbouring side-chains (within 5.0Å). The Rosetta energy is calculated for the original sequence and for the one with the mutated residue to alanine (for both cases the structure is relaxed before the calculation (Chaudhury et al., 2010)). As the relaxation algorithm is stochastic we repeat the process 10 times and retain the median energy as the ΔG. The difference between the median value of the original sequence and the one with the residue mutated to alanine is taken as the importance of the residue.

3.2.2 Computational loop grafting

Once potential loops have been selected SAbDesigner attempts to graft the loops onto all of the CDRs from the database of antibody frameworks (see section 3.2.3). SAbDesigner tests whether a loop can be grafted in a specific structural location on an antibody using the anchors, which are two residues upstream and downstream from the start and end point of the loop. As a rapid first step we measure the average distance between the Cαs of the anchors in the loop to be grafted and the anchors of the grafting location. If this value between the anchors is less than 1.0Å we go on to perform the structural superposition. The backbone structure of the anchors of the loop to be grafted and the anchors of the grafting location are superimposed (Kabsch, 1978) (see Figure 3.1D-E).

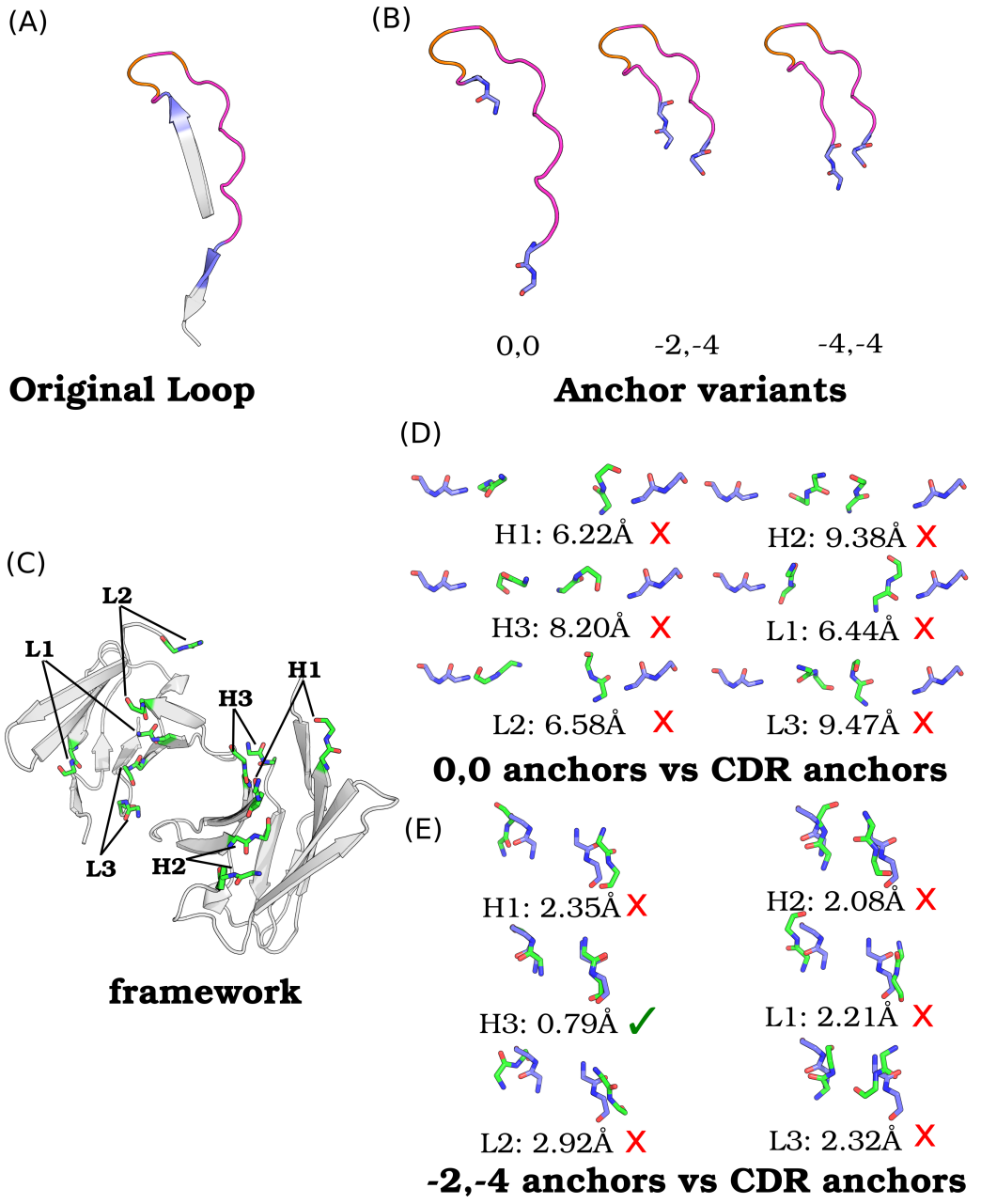

Figure 3.1: (A) A protein loop which is put forward for grafting on an antibody. The loop characterised by DSSP is coloured in magenta, except for the residues participating in binding which are coloured in orange. The anchors are coloured in blue and the adjacent secondary structure is coloured in grey.
(B) The "0,0", "-2,-4" and "-4,-4" anchor variants of loop (A).
(C) A potential antibody framework for loop (A). The anchors of the CDRs are highlighted with green sticks.
(D) Superposition of the "0,0" variant anchors on the all the CDRs of the framework (C). The CDR anchors are coloured in green while the variant anchors are coloured in blue. No CDR is found to have anchors that superpose under the 1.0Å threshold
(E) Superposition of the "-2,-3" variants anchors vs the all the CDRs of framework (C), with the same colouring scheme as in (D). This anchor variant was found to satisfy the anchor criteria for CDR H3.

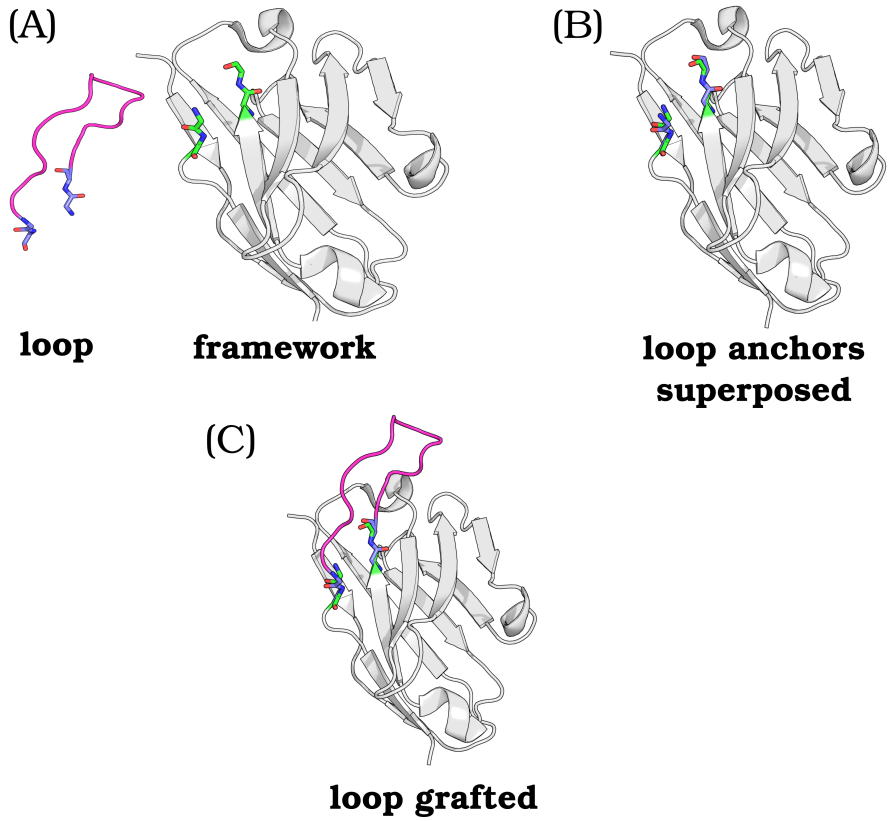

Figure 3.2: The process of grafting a loop on an antibody framework
(A) shows a loop in magenta, with its anchors coloured in blue. The antibody framework is coloured in grey with the anchors of the target CDR coloured in green.
(B) The anchors of the loop are superposed on the anchors of the CDR
(C) The atoms of the rest of the loop are transferred using the rotation and translation transformations from the superposition in (B)

SAbDesigner considers that a loop can be grafted if the RMSD of the anchor superposition is less than 1.0Å(Choi and Deane, 2010). All the atoms of the residues of the loop are then grafted on the antibody framework using the same rotation and translation transformation (see Figure 3.2).

The definition of an antibody CDR does not have to start and end on secondary structure, as opposed to our loop definition. From the existing antibodies deposited in SAbDab (Dunbar et al., 2013b) we found that 40% of H3 loops contain at least the first residue of the proceeding beta strand. Therefore, to increase the chances of identifying a potential design SAbDesigner expands the search by including or excluding up to four residues from the surrounding

structure of the loop, with the anchors being adjusted accordingly (see Figure 3.1B). We refer to these as anchor variants, and label them with two numbers, e.g. anchor variant "-1, -2". The numbers denote the distance in amino acids and direction as opposed to the start and end of the the loop respectively. A positive distance indicates a change towards the C-terminus of the protein, while a negative distance indicates a change towards the N-terminus. Therefore, anchor variant "-1, -2" means a loop which includes an extra residue from secondary structure at the start and removes two residues from the end of the loop, while anchor variant "0,0" is the original loop as characterised by DSSP. The search can be expanded further if no results are found. Anchor variants that exclude residues that form contacts in the native complex are not considered.

3.2.3 Antibody frameworks

3.2.3.1 Numbering and CDR definition

The locations on antibodies that are used for grafting loops by SAbDesigner are the CDR loops. For a given Fv the location of the CDR is identified using the Chothia structural definition (Al-Lazikani et al., 1997), based on Chothia numbering provided by ANARCI (Dunbar and Deane, 2015).

3.2.3.2 Initial Database of antibody scaffolds

There were 1049 human antibody frameworks listed by SAbDab (Dunbar et al., 2013b) on the 12th of January 2017. The structural shape of the anchors of the six CDRs from these antibodies was clustered. The clustering algorithm starts with the list of all possible frameworks and then randomly picks and compares to the existing clusters (the first picked will become the first cluster as there are no existing clusters). If the current framework is unique in comparison to all the existing clusters it is made a cluster, otherwise it was discarded. Uniqueness is

defined as having an RMSD greater than 0.7Å. This resulted in 76 human unique frameworks that SAbDesigner uses as its initial database of antibody scaffolds.

3.2.3.3 Docking by matched molecular pairs

After a loop is grafted on the antibody the design is then docked as a rigid body on the target antigen using the matched molecular pairs of the native loop. This is achieved by computing the Kabsch superposition (Kabsch, 1978) of the grafted loops from the antibody and the native loops on the receptor in the docked pose. This returns the rotation and translation transformations, which are then used to transfer the atoms of the rest of the antibody.

3.2.4 $C\beta$ thresholds for clash detection

3.2.4.1 Intra-protein

The intra-protein threshold for the minimum distance between two $C\beta$ atoms was calculated from a list of chains of protein structures with resolution greater than 2.0Å culled at 40% sequence identity using PISCES (Wang and Dunbrack Jr, 2003). For each chain all the pairwise $C\beta$ euclidean distances were calculated and the ones lower than 8.0Å were retained. For each residue type the density distribution is discretized in 0.25Å bins. The threshold is set at the point where 99% of the density is included. This is achieved by sequentially traversing the bins from 8.0 to 0.0, stopping at the point where 99% of the density is on the right of the current bin. The thresholds computed and used by SAbDesigner are listed in Appendix Table A.1. As Glycine does not have a Cβ atom the Cα atom was used for calculating its threshold distance.

3.2.4.2 Inter-protein

The inter-protein $C\beta$ distances were established by the same methodology as for the intra-protein but for residues on different chains. The dataset used

was formed of known protein complexes. We used the *bona fide* set of protein complexes compiled to benchmark Zdock (Hwang et al., 2010). The thresholds computed are listed in Appendix Table A.2.

3.2.5 Canonical class structure database

In some cases in order to avoid clashes it may be necessary to replace the CDR of an antibody with one of its other canonical forms. Such cases include situations when changing the canonical form of a CDR will remove clashes between a grafted loop and that CDR. When replacing the structure of a CDR to another canonical form of the same type SAbDesigner uses the structures available in the PDB on the 12th of January 2017 and indicated as being antibody CDRs by SAbDab (Dunbar et al., 2013b). SAbDesginer iterates through all the available canonical forms in order of length starting from shortest. This order is preferred because it is more likely that a shorter loop will not clash with the grafted loop than a longer one, and therefore reduce the computational time of solving the clash. SAbDesigner in its automatic form restricts canonical form changes to allowable pairwise combinations (e.g. if changing the canonical form of L1 only one from the subset of canonical forms that have been seen to exist in combination with the L3 and L2 present on the antibody are chosen). These allowable combinations were identified from the modellable loops in the NGS dataset from Chapter 2.

3.2.6 Viable design and termination criteria

A viable design is one in which a binding loop with more than 175Å^2 BSA can be grafted on an antibody scaffold satisfying the anchor criteria and the Cβ thresholds. A viable design can be achieved even if initially there are clashes, if these can be resolved through CDR replacement.

In its default mode SAbDesigner terminates when it reaches the first viable design. It can also be set to continue the search and to show all the distinct viable designs. For a CDR and anchor variant (e.g. anchor "-2, 0" on CDR H1) it will, however, always stop at the first viable design. Grafting in the same position with the same anchor variant is not likely to yield a distinctive design because of the high conservation in antibody frameworks. There could be differences due to the make-up of the other CDRs. This issue is addressed in the next chapter in refinement through CDR optimization where the other CDRs are optimized for the target antigen.

3.2.7 Therapeutic Targets

3.2.7.1 Extended target database

To identify protein structures which are important therapeutic targets we downloaded the Therapeutic Target Database 5.1.02, released on the 16th of November 2016 (Zhu et al., 2011). This database provides for known protein targets their respective ID inside UNIPROT (Consortium et al., 2014). The ID can then be used to identify the gene and further on to identify the homologous protein in different species using the UNIPROT API. The algorithm we used is described in Appendix section A.3. This was performed for cases where the human protein does not have a structure deposited in the PDB, or a receptor-target complex, and it was necessary to use the homologue from another species as a substitute. The resulting list of human and homologous proteins was then intersected with the entries in PDBe (Velankar et al., 2015). This allowed the matching of the proteins deposited in the PDB with their UNIPROT IDs. The PDB entries that contain at least two proteins with different UNIPROT IDs (i.e. a complex), and where at least one of those proteins was in our list, are retained. This database of potential targets with co-crystal structures in the PDB will be referred to as the *Extended Target Database* (ETD).

3.2.7.2 PDBBind

PDBBind (Wang et al., 2005) is a database that matches PDB entries with experimentally determined affinity data. We downloaded the list of protein-protein complexes available on the 22nd of June 2017 at:

http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp

The database, however, does not distinguish within a PDB entry which chains form each member of the complex. PDBe was used to separate the chains within a PDB file into different proteins according to their UNIPROT ID.

3.3 Results

3.3.1 Overview

Inspired by the success of loop grafting experiments we have built a computational tool that automatically identifies the best antibody scaffolds on which to graft loops that are known to bind a given target. The input requirement for SAbDesigner is the structure of a protein complex between the target and a receptor protein. SAbDesigner automatically identifies the loops modulating binding in the complex. The algorithm then proceeds to exhaustively search through known human antibody scaffolds for ones on which the selected loops can be grafted. The tests performed validate that the loop will structurally fit, that the loop will not denature in its new environment, that it will not destabilise in its new environment, and that other areas of the antibody will not impede complex formation. In certain cases this can require the solving of steric or electrostatic clashes by either individual point mutations or replacement of the other CDRs with other available canonical structures. All of these steps are performed automatically. After a candidate antibody is developed it can be further optimised by identifying the best combinations of CDRs that will maximize the amount of BSA between the antibody and the target. SAbDesigner

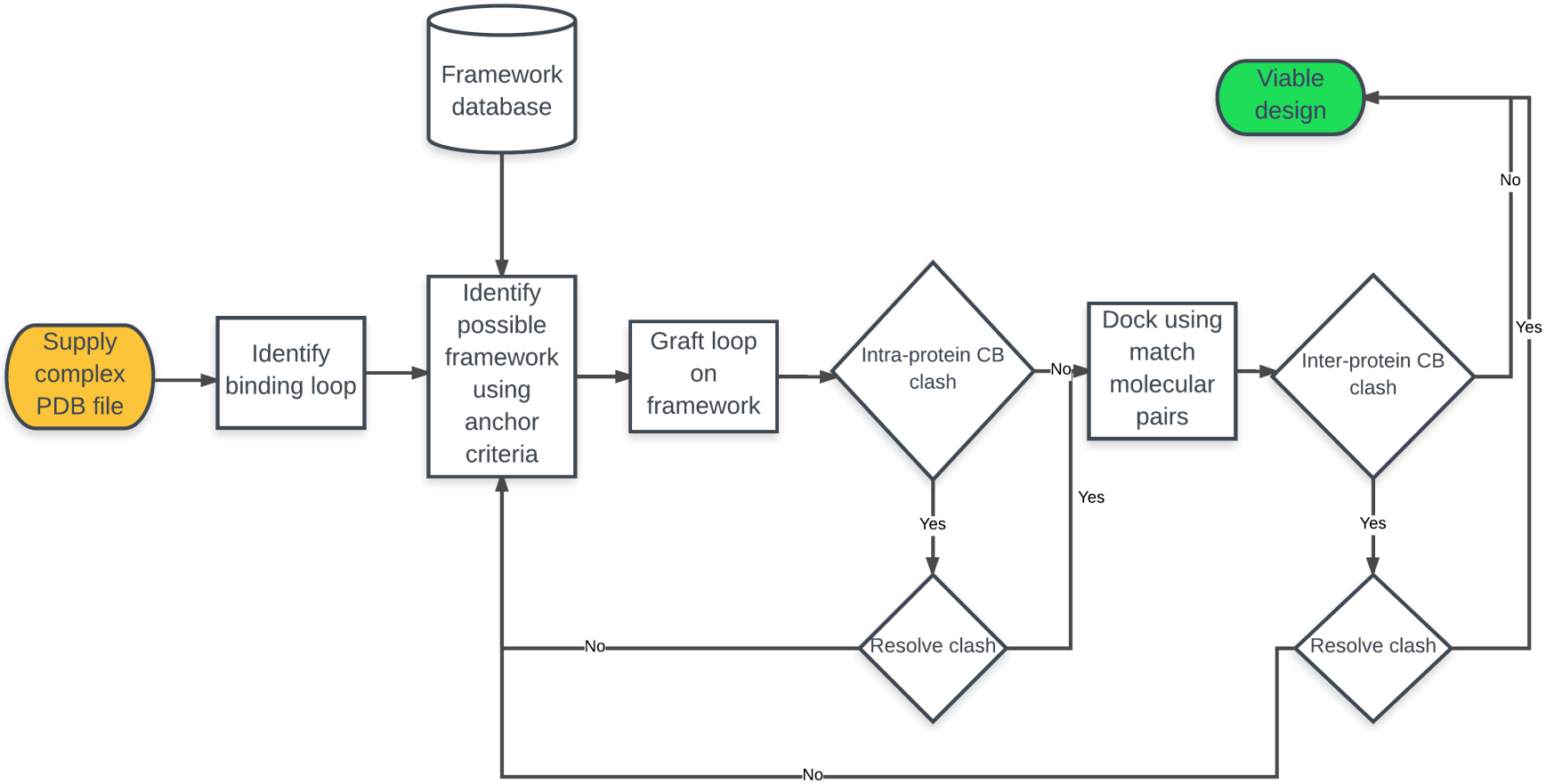

Figure 3.3: Flowchart of the steps in generating in a viable design (see Methods).

Loop ID	BSA	CAS weight
Loop 0	240Å^2	-3.51
Loop 1	122Å^2	-3.08
Loop 2	85Å^2	-3.27
Loop 3	217Å^2	-2.34
Loop 4	13Å^2	-3.27
Loop 5	25Å^2	-2.25

Table 3.1: Comparison between the BSA and the CAS weight, the two ranking criterias, for the binding loops of IL-5.

allows fully automated generation of candidate antibodies that can then be further matured with other experimental or computational techniques. In the following paragraphs we detail the steps of SAbDesigner using the Interleukin 5 (IL-5) target in complex with its alpha receptor as a guiding example (PDB accession code 3va2). IL-5 is a cytokine secreted by T-cells and its overexpression has been shown to be linked to allergic inflammatory diseases (Kusano et al., 2012). It is therefore an important therapeutic target.

3.3.2 Binding loop identification

The input to SAbDesigner is an X-ray crystal structure file of the target in complex with its receptor in the PDB format. For example in the case of IL-5 this is the PDB structure with accession code 3va2, in which IL-5 is in complex with its alpha receptor (see Figure 3.4A).

SAbDesigner firstly identifies the interface of the receptor (see Figure 3.4B). The loops from the receptor which are identified as part of the interface are then put forward for potential grafting. In the case of the alpha receptor of IL-5 this resulted in a total of six loops with BSA ranging from 13Å^2 to 240Å^2. Loop 3 and Loop 0 were chosen as they pass the 175Å^2 threshold, meaning they present a realistic possibility for transferring binding specificity (see Figure

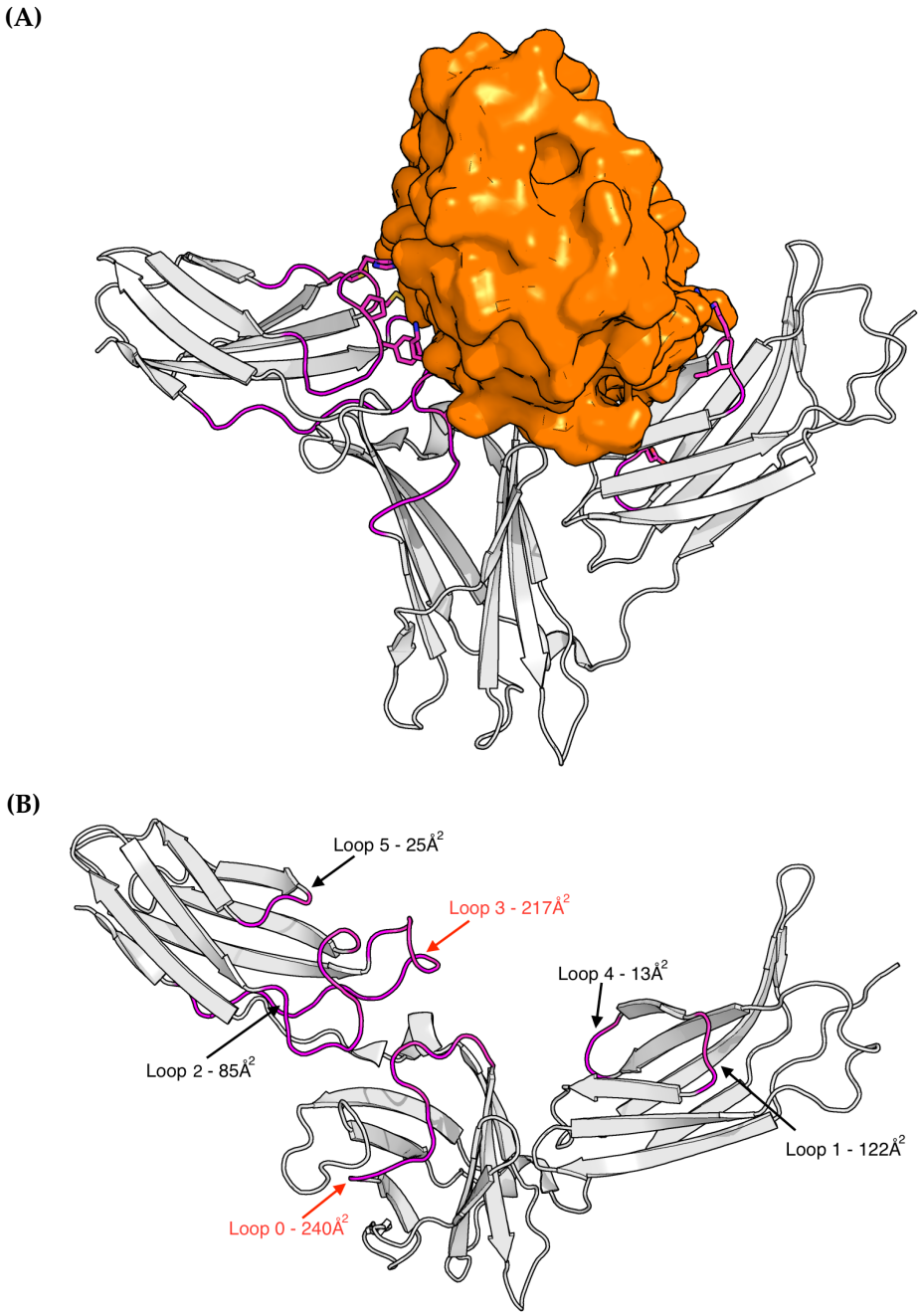

Figure 3.4:
A) IL5 (orange surface) in complex with its receptor (grey). The residues which were identified as forming part of the receptor interface have been colored in magenta. B) The receptor with the binding loops highlighted and their associated BSA. The BSA values that pass the threshold are coloured in red.

3.4). An alternative ranking criteria that can be used is based on the importance in binding of the residues of the loop, and is computed by CAS (see Section 3.2.1.3). A comparison between the results of these two methods can be seen in Table 3.1. The results show that Loop 3 and Loop 5 have the highest importance in terms of CAS weight, as opposed to Loop 3 and Loop 0 for BSA. Loop 5, however, has only one binding residue and its mutation to alanine is seen as favourable (i.e. not important for binding). The other loops each have at least one residue with positive contribution, with the others being negative. We have taken forward the loops identified by the BSA criteria, as we believe that this criteria is more accurate in measuring importance for binding.

3.3.3 Antibody framework identification

For an individual loop the first step is to identify computationally a framework on which it can be grafted. All the CDRs from every framework from our database of 76 frameworks are tested for the anchor loop grafting criteria, including anchor variants (see Figure 3.1). All the anchor variants - CDR combinations that pass the threshold are ranked by anchor RMSD and then tested until a viable design is reached. If no viable designs are identified then the search is expanded to all the other existing human frameworks, and all possible variable anchors outside the -4 to +4 residue range. After identifying the locations where the loops can be grafted the residues of the loop are then transferred using the transformation from the anchor superposition (see Figure 3.2).

If two or more loops from a receptor can be grafted on the same antibody framework the designs are merged together to see if the loops fit within topologically similar locations as in their native environment. The loops from the native receptor are superpositioned on the loops grafted on the antibody and if the RMSD < 1.0Å the new design is retained.

3.3.4 Identifying Clashes

The next step for each grafted loop-framework combination (design) is to identify if the grafted loop clashes with its new environment. We run under the assumption that the loop will fold in its native conformation, and from this point try to identify if this will be impeded, or the presence of the loop will impede the correct folding of the rest of the antibody. The method searches for steric clashes, which it attempts to resolve (see section 3.3.5). If they cannot be resolved the design is discarded. Longer range effects and their potential to destabilise are also taken into consideration. For this threshold distances between every pair of amino acid types were established based on the $C\beta$ atom of the residue (see Section 3.2.4). A clash is declared if any residue from the grafted loop is at a distance lower than the threshold to any residue of the antibody. Clashes are also possible between the designed antibody and the target, when the designed antibody is docked on the target based on the matched molecular pairs of the grafted loop (see Section 3.2.3.3). We apply the same clashing tests, but in this scenario the $C\beta$ thresholds are established from a data set of protein complexes (see Methods inter-protein $C\beta$).

3.3.5 Removing clashes

If a clash occurs between the grafted loop and one of the other CDRs the possibility of replacing the other CDR with one of its other allowable canonical forms is computed automatically. SAbDesigner grafts the other canonical forms using the same methodology described for non-protein loops, and they must satisfy all the same $C\beta$ thresholds described in section 3.2.4 . If an alternative canonical form satisfies all the criteria and removes the clash with the grafted binding loop then the design is retained and is classified as viable design.

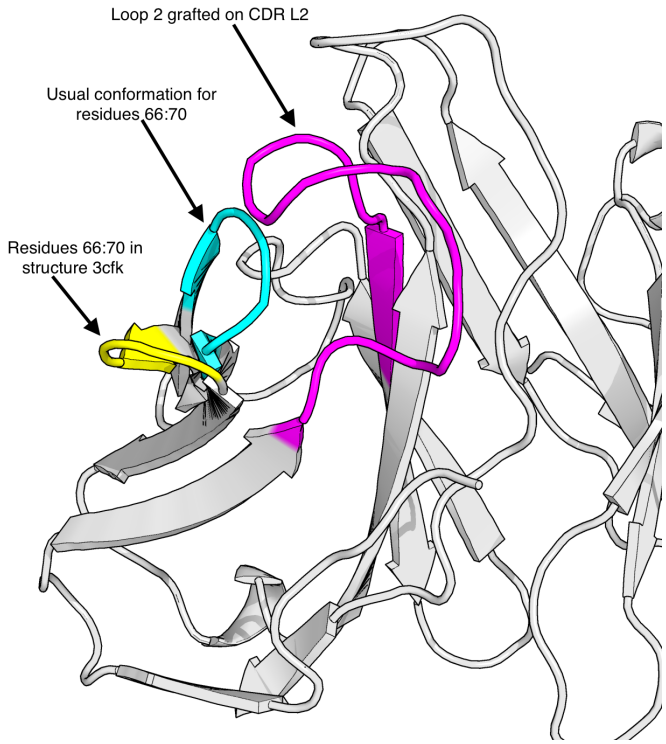

Figure 3.5: Example of variation in framework residues that allow a conformation to be grafted. Loop 2 is shown coloured in magenta, with the usual conformation of residues 66:70 of the light chain show in cyan. This conformation clashes with the grafted loop. The light chain in PDB file 3cfk chain G, however, has an alternative conformation that allows for the grafting of the loop (coloured in yellow).

The other possibility is for the clash to occur between the grafted loop and the framework. SAbDesigner's framework database was generated by clustering all the human antibody frameworks in a reduced set of 76, using only anchor information. We have identified, however, that there are instances where slight variations in frameworks can allow some loop conformations to be grafted while still satisfying the Cβ thresholds (see Figure 3.5). Therefore the SAbDesigner algorithm was extended to test all human antibody frameworks if a loop satisfies the anchor criteria for a CDR but fails because of framework clashes on all of the 76 initial frameworks. In this way the entire set of human loops is used only when it has been identified that this is necessary, and not every time. This results in a significant reduction of computational time to

ID	Scaffold	Chains	Grafted Loops	Other changes
1	3cfk	I,G	L3: Loop 3 (-4, -4) L1: Loop 2 (-1, 4)	None
2	4r7d	O,P	L3: Loop 3 (-4, -4)	None
3	1zlu	M,K	H1: Loop 0 (-1, 2)	H3: 3u0t chain B H2: 2ojz chain H
4	5cil	H,L	H1: Loop 0 (-1, 4)	H3: 1mjj chain B
5	4lss	H,L	H3: Loop 3 (-2, -4)	None

Table 3.2: The details of the initial anti IL-5 designs. The table lists on the first column the PDB ID of the original antibody scaffold, the chains (heavy chain and light chain respectively) in the 2nd column, the grafted loop and the CDR 3rd column in the format *CDR: grafted loop anchor variants*, and other changes (i.e. canonical structure replacements) in the 4th column

generate designs for one target.

3.3.6 IL-5 designs

For the IL-5 target SAbDesigner identified five possible designs. These are described in Table 3.2, with their full sequences being listed in Appendix section A.4. The design incorporate either Loop 3 or Loop 0 as the main loops, the ones with BSA greater than 175Å^2 in the native complex (see Figure 3.4). Design 1 (See Figure 3.6B) is an example of SAbDesigner grafting two loops (Loop 3 and Loop 2) from the receptor in the same topological position. Although Loop 2 has a low BSA it is important because in the native complex it forms a disulfide bridge through its cysteine residue to the cysteine on Loop 3. We believe this is an important feature for stabilising the two loops, and it also has the added benefit of not leaving an unmatched cysteine residue on Loop 3 which might cause the antibody to misfold or aggregate.

Design 2 (see Figure 3.6A) is the equivalent of Design 1 without Loop 2. Design 5 (see Figure 3.7B) is also an example of grafting only Loop 3, but this time on CDR H3 with a different anchor variant. This design has the added

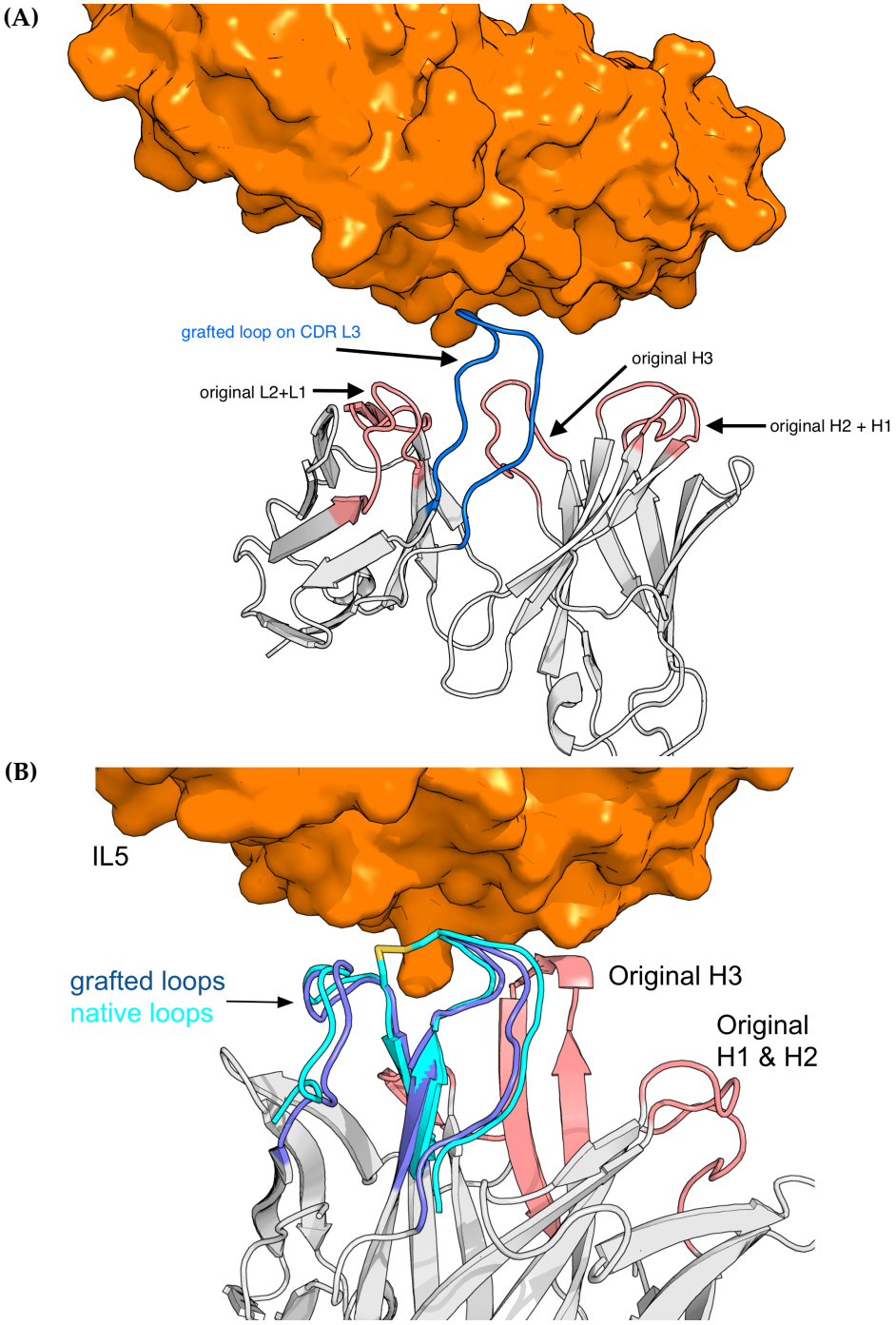

Figure 3.6:
Anti IL-5 designed antibody antigen complexes. IL-5 is show in orange, with the antibody framework show in gray. Grafted loops are shown in dark blue, and original CDRs from the scaffolds are show in salmon.
A) Design 2 - Loop 3 grafted on CDR H3.
B) Design 1 - Loop 3 + Loop 2 replicated together on an antibody scaffold 3cfk, on CDRs L1 and L3 respectively. They are in the same topological position as in their native environment.

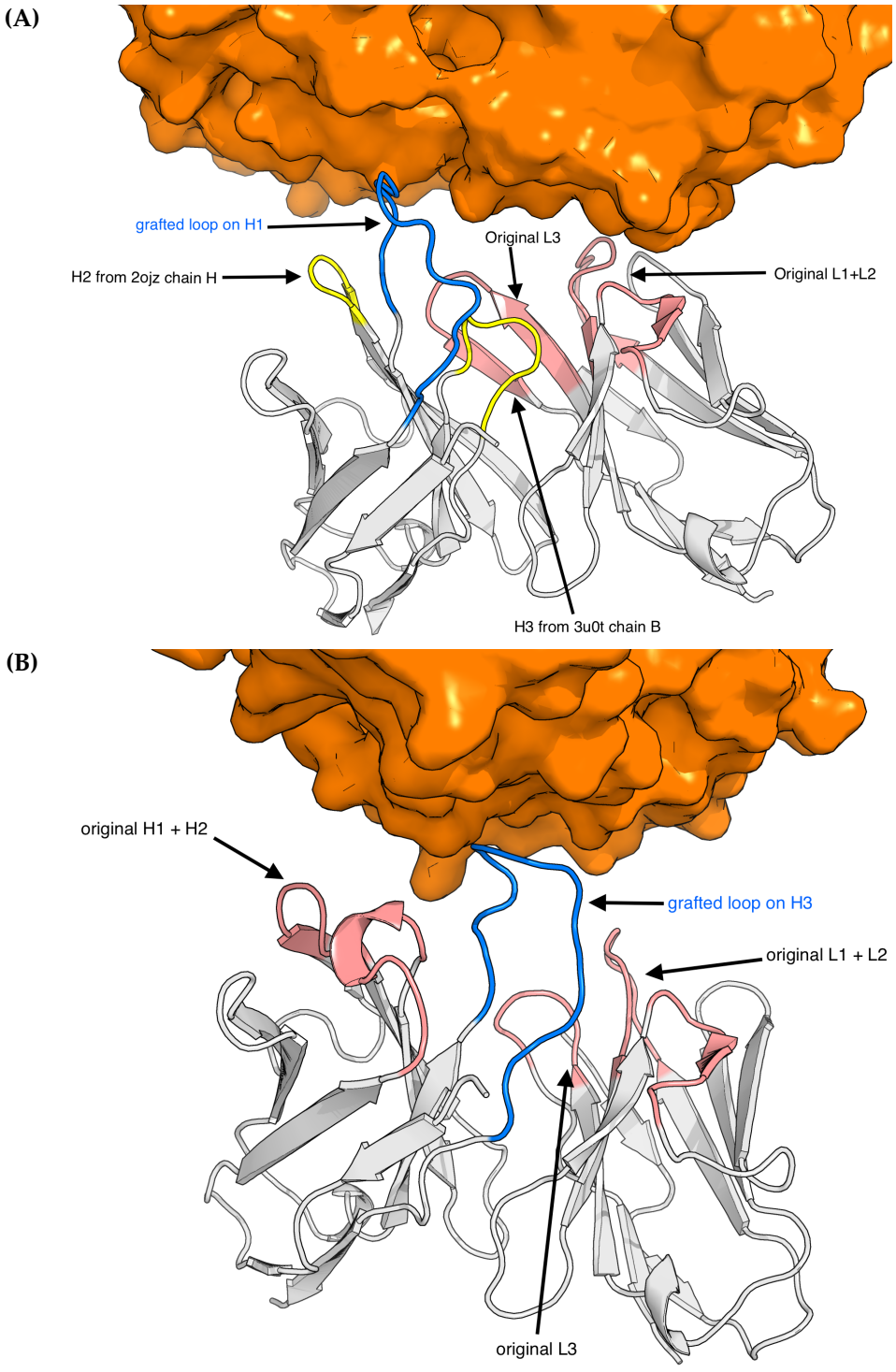

Figure 3.7:
Anti IL-5 designed antibody antigen complexes. IL-5 is shown in orange, with the antibody framework show in gray. Grafted loops are shown in blue, original CDRs from the scaffolds are show in salmon, and replaced CDR are shown in yellow.
A) Design 3 - Loop 0 grafted on CDR H1.
B) Design 5 - Loop 3 grafted on CDR H3.

PDB ID	Target name	Target Chain	Affinity data	Figure
1xxd	Coagulation factor XI	A	IC50=600nM	3.8A
1shy	Hepatocyte growth factor	A	Kd=90nM	3.8B
1eer	Erythropoietin	A	Kd=1nM	3.8C

Table 3.3: A list of some of the important targets for protein therapeutics that have a co-crystal structure with a receptor which SAbDesigner can replicate on an antibody. The affinity data is extracted from PDBBind (see Methods)

benefit over Design 2 of grafting the novel loop in the structurally most diverse area of the antibody, CDR H3. Both designs, however, have the drawback of containing an unmatched cysteine inside the CDR in the absence of Loop 2. This issue will be further addressed in the next chapter.

Design 3 (see Figure 3.7A) and Design 4 contain Loop 0 grafted on the H1 CDR, the only difference between them being the anchor variant.

3.3.7 Therapeutic Targets

To test the large scale applicability of SAbDesigner we compiled the ETD database (see Section 3.2.7.1), a list of usable therapeutic targets from the Protein Data Bank. From the 2925 protein targets in the Therapeutic Target Database 1949 were found to have a complex structure in the PDB of either the human version or an animal homologue. As the designs will probably mimic a subset of the interface of the receptor they will also probably achieve only a fraction of the affinity of the original receptor. It is therefore beneficial in experimental studies to start from a high affinity complex. To identify the ones that have high affinity we intersected the 1949 complexes with the list of complexes for which the affinity has been experimentally and some of the resulting designs are detailed in Table 3.3 and Figure 3.8.

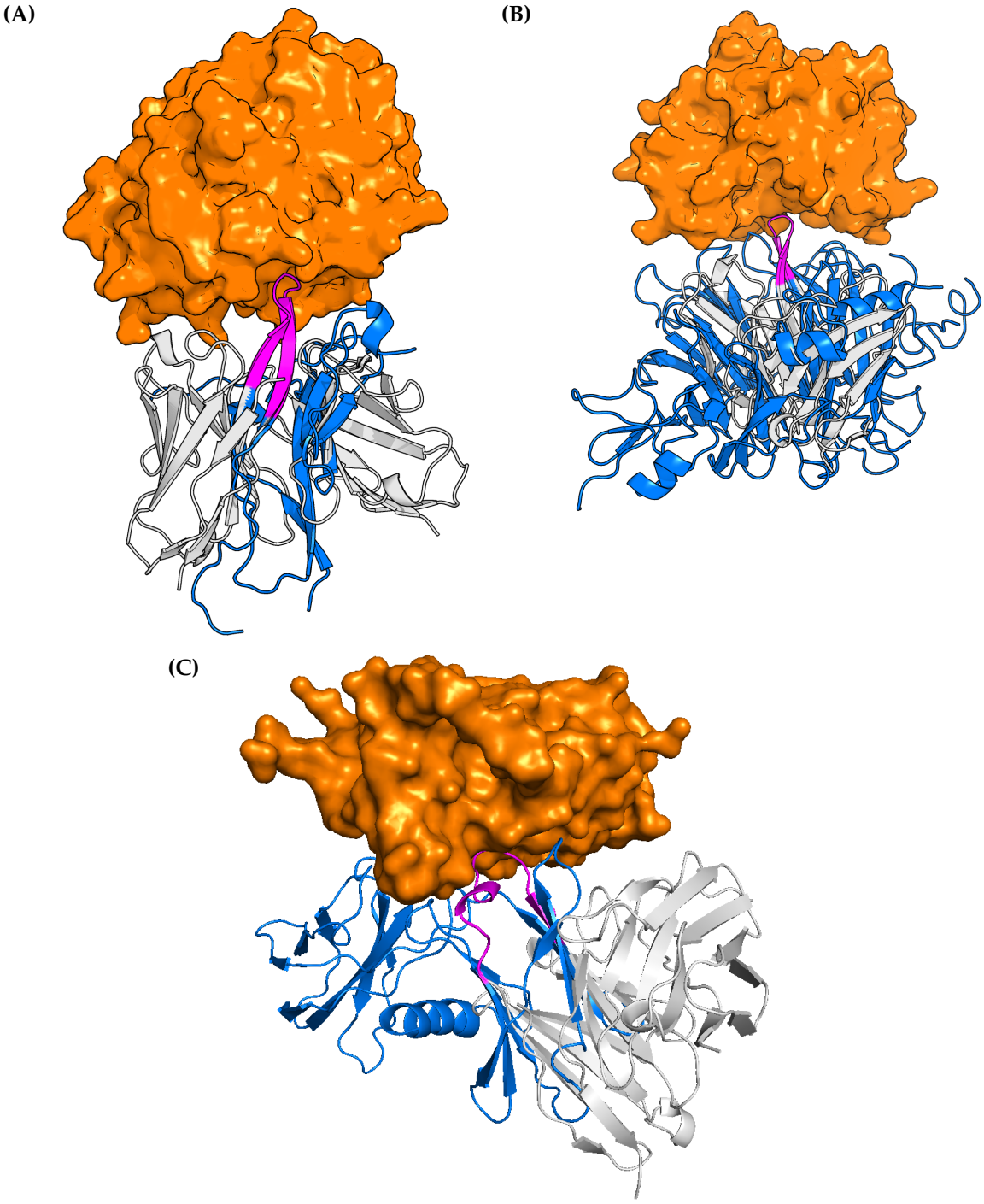

Figure 3.8:
Example designs for therapeutic targets from complexes of known affinity. In each case the antigen is shown in orange as a surface, with the antibody in gray, and the receptor in blue. The grafted loop which is common for both the antibody and the receptor is shown in magenta.
A) Designed antibody against "Coagulation factor XI" derived from 1xxd complex
B) Designed antibody against "Hepatocyte growth factor" derived from 1shy complex
C) Designed antibody against "Erythropoietin" derived from 1eer complex

3.4 Discussion

Rational computational antibody design methods have the potential to alter the established pipeline for developing antibody therapeutics, one which is currently focused on methods that replicate the evolution and selection mechanism of antibodies. In this chapter we present SAbDesigner, a tool that uses computational techniques to automatically design antibodies against a specified target by using loop structures which do not originate in the antibody germline. SAbDesigner uses the principle of loop grafting to transfer binding specificities, a method that has been shown to be effective in the transfer of binding specificity in both antibody and non-antibody proteins. SAbDesigner builds an antibody structure that mimics the interface of an existing non-antibody receptor of the target. It automatically determines the loops and fragments which participate in the formation of a complex and an antibody scaffold in the PDB on which the loop can be grafted.

SAbDesigner is the first fully automated procedure for loop grafting in antibodies that does not have to use antibody CDRs, but rather loops from any protein structure. It alos tests if the loop as a rigid body fits sterically within the constraints of the antibody. In cases where clashes occur SAbDesigner attempts to automatically solve them by either replacing the other CDRs with other known canonical forms or structures, or by identifying if other frameworks have small structure differences which remove the clash.

For the IL-5 target SAbDesigner proposed five designs that each incorporate one of the two main binding loops. There were three designs proposed for Loop 3 and two for Loop 0. The designs use several of the CDRs, with the loops being grafted on all but CDR L2 or H2. Design 1 is an example of SAbDesigner's capability to graft two loops from a receptor in the same topological position as in the native complex,identifying the only known antibody framework that

can accept this conformation. The search performed in this case was exhaustive, however, the algorithm does not jeopardize computational runtime in cases where this is not required.

The main limitation of SAbDesigner is that it requires a structure of the target with a binding partner. This may not always be readily available. However, proteins that are known to be important therapeutic targets have often been crystallised. We found that in a dataset of 2952 potential targets, 1949 have a publicly available crystal structure along with their receptor.

Another limitation of the method is that as we only mimic a subset of the known binding interface. Therefore, the affinity of the designed antibody is likely to be lower than that of the cognate receptor. This in turn can mean that the designed antibody may not be well suited for a competition scenario with the original receptor, or if starting from low affinity will have an affinity which can not be reliably detected. We have therefore intersected our list of important therapeutic targets with the list of protein-protein complexes with high affinity from PDBBind. We list three (see Table 3.3) high affinity antibodies that would be good candidates for proof-of-concept experimental validation.

However, in antibody therapeutic development low affinity designs can also be used as starting points for a phage display library. The advantage of designing a phage library using these antibodies instead of using a random array of antibodies is that they have a higher probability of selecting for the desired epitope, and also attaining a detectable affinity faster as they contain a motif which is known to participate in binding to the target. There is also the benefit of using a non-antibody conformation which is unlikely to be reached by sampling from the germline.

The designs proposed by SAbDesigner to this point are only validated in terms of their rigid bodies, using the $C\beta$ thresholds. This method, however, does

not show if the modified loop or the rest of the antibody will preserve their structure. In the next chapter we present a full computational validation pipeline.

Experimental validation is also currently being performed. Our industrial partner GSK is currently expressing the IL-5 antibody clones with the aim of running a competition assay against the native IL-5 receptor. This will indicate the range of affinity the designed antibodies have. They will also run a suite of biophysical characterisation methods to ascertain the solubility and stability of the designed antibodies.

I think the people of this country have had enough of experts with organisations from acronyms saying that they know what is best and getting it consistently wrong

— Michael Gove MP

4 SAbDesigner: Validation and Refinement

Contents

4.1 Introduction

In the previous chapter we showed how loops from non-antibody proteins can be used to design novel antibodies. The validation provided for the

designs was limited, it only took into account the molecular shape of the grafted loop. The loop was tested to see if it would fit within the geometric constraints of the CDR area, and within the constraints of the antigen, using pair-wise residue dependent Cβ thresholds. Most existing computational *de novo* antibody design methods contain more robust validation methods that range from molecular docking, to energy based calculations and native structure recapitulation. In the case of AbDesign (Lapidoth et al., 2015) their validation arises from comparing their designed antibodies to known antibody structures that bind the antigen in the same pose, and computational docking. They show that their designs are able to recapitulate CDR structures similar to ones seen in native antibodies, with low sequence identity equivalents but similar hydrogen patterns. In the case of OptCDR (Pantazes and Maranas, 2010) they rely on their energy function which was trained on a set of over 100 mutations of an HIV antibody with experimentally determined affinities. It can distinguish favourable versus unfavourable mutations with an accuracy of 78%. They then used it to compare the energy values of their designs to the known antibody binders of the target obtaining similar values. Liu et al. (2017) rely on the fact that the grafted structural fragment is known to bind the the target, and they use known natural antibody scaffolds that contain that motif. They therefore do not use further *in silico* validation as both the binding fragment and the designs have been validated experimentally.

SAbDesigner is different from the first two methods described above, and more similar to Liu et al. (2017), as it relies on the assumption that at least the grafted loop on the antibody is known to bind the target. However, the sequences and the structures grafted have not been seen in known antibodies. Therefore, the validation pipeline is heavily focused on identifying if the grafted loop will fold in the correct conformation on the antibody and will not cause the rest of the antibody to destabilise. After we validate that the loop can be

grafted we also validate whether a computational docking tool is able to identify the expected binding pose. In this chapter we provide a set of validations tools to achieve these goals.

For all of the methods described above the antibody designs obtained after the initial *de novo* design step are further refined. AbDesign (Lapidoth et al., 2015) and OptCDR (Pantazes and Maranas, 2010) both perform sequence optimisation by swapping residue types in the CDR to achieve backbone structure and sequence refinement for better interactions with their targeted antigens. In both cases the processes are guided by their own energy functions to score favourable versus unfavourable mutations. Liu et al. (2017) take a different approach, instead of making point mutations they attempt to change the entire structure of CDR H3 to a more favourable conformation, with two of the changes resulting in an affinity improvement.

In SAbDesigner we adopt an approach that encompasses both of these paradigms. Firstly, we make point mutations to the grafted loops, but not as a means of increasing binding affinity, but as a means to change residues which are detrimental to the design. Secondly, for affinity maturation, we swap one of the other CDRs, which is not used for grafting a binding loop, to a more favourable conformation.

4.1.1 Validation

The Cβ thresholds introduced in the previous chapter are useful to gauge if a pair of residues are found closer to each other than normally expected. However, this does not account for a lack of compatibility between a residue and its environment, which can make areas of a protein susceptible to instability and aggregation (Clark et al., 2008). If the designed antibody is prone to aggregation it has a low chance of exerting its engineered functionality and in a therapeutic scenario aggregates can also elicit an immunogenic response (Roberts, 2014).

Computational methods for therapeutic protein aggregation prediction have been developed (Chaudhri et al., 2013), but few of them have been validated by experimental studies and they have not yet shown a systematic applicability (Buck et al., 2012). In our design scenario we know that the antibody framework is soluble, and the receptor from which the binding loop has been extracted is also soluble. We then rely on the hypothesis that only a change in the environment of either the loop or the antibody can cause instability. For this we quantified the accessible surface area (ASA) change for individual residues in their new environment, either on the grafted loop or the rest of the antibody (due to the presence of the grafted loop). SAbDesigner allows manual inspection of these results via an inbuilt Pymol plugin.

The methods described so far all consider the structures as rigid bodies. We therefore also analysed the impact of allowing the structure to move, using Rosetta FastRelax relaxation (Conway et al., 2014). For each design SAbDesigner quantifies the difference in movement of the grafted loops between their native environment and the design environment, using Rosetta energy and RMSD. The movement in the individual cases is affected differently by the respective environment, and our interest is to select designs where this is at similar levels between the two environments. To establish what are permissible differences for the energy we used a biologically inspired antibody example, by swapping the CDR canonical forms on an antibody and measuring the change in Rosetta energy units.

The grafted binding loops dictate what the expected binding pose of the designed antibody will be via matched molecular pairs (see Section 3.2.3.3). We analyzed the possibility of validating this pose by using *docking*, a method through which the binding pose is predicted from separate structures of the two binding partners. According to the latest CAPRI assesment of computational docking there is no method that currently shows a broad applicability at

predicting accurately the binding interface of a complex (Lensink et al., 2017). We, therefore, do not use docking as an absolute validation method, but instead use it as a relative measure. We compare the difference in the docking protocol's ability to identify the designed pose versus the natural pose of the receptor. For this we selected ZDock (Mintseris et al., 2007) as it was a highly ranked tool in the latest assessment of docking methods (Lensink et al., 2017), it is fast at generating binding poses, and is a standalone tool which can be run automatically by SAbDesigner. ZDock has also been successfully used to predict antibody binding sites when the possible binding area for the pose generator was restricted to either the individual binding residues (Krawczyk et al., 2013) or the CDRs (Chen et al., 2003). We implemented a similar approach by running ZDock with the epitope defined *ex ante*.

4.1.2 Refinement

In terms of refining the structures our goals were two fold. One was to mutate residues which have the potential to destabilise the antibody, and the other to increase the affinity of a design but without altering the shape of the grafted binding loop.

Lippow et al. (2007) developed an affinity maturation pipeline where they mutate all the residues from the CDR regions of an antibody to all other possible amino acid types, except cysteine and Proline. They then evaluate favourable mutations using a physics based energy function. OptMaven (Li et al., 2014) and AbDesign (Lapidoth et al., 2015) adopt a similar approach, but restrict the possible amino acid types to ones which antibodies would normally use through affinity maturation. A more detailed analysis of maturation pipelines for antibodies can be found in Sirin et al. (2016). Our aim for a single residue mutation is not to increase the affinity, but to identify a better option for a residue that is deemed detrimental to the design, while preserving the existing

backbone structure. As the loop grafted is of non-antibody origin the amino acid options that antibodies chose for affinity maturation are not applicable. SAbDesigner uses an approach were it restricts the possible mutations to amino acid types that can easily adopt the conformation of the original residue based on neighbour dependant Ramachandran probability distributions (Ting et al., 2010). Barthelemy et al. (2008) managed to increase the solubility (i.e. quantity of monomeric antibodies at high concentrations and thermostability at high temperatures) of VH domains by replacing hydrophobic residues with ones which are more hydrophilic. We incorporate this knowledge in the mutation protocol by testing only non-hydrophobic mutations for a hydrophobic residue.

The engineered binding loop is placed in the CDR area, and therefore it is likely that the other CDRs are found close to the target. We therefore also examined our ability optimise the other CDRs in order to increase the affinity of the complex based on the same principle of CDR canonical class replacement described in Section 3.2.5. Liu et al. (2017) used this approach on an antibody designed by loop grafting. Their initial design had a binding fragment grafted on CDR L2 and after experimental validation it was identified to have low mM affinity. They then grafted computationally other existing CDR H3 conformations, and ranked the best performing ones using the Rosetta energy score. Two of these refined designs resulted in low nM affinities against their target after experimental validation. Pantazes and Maranas (2010) identify the best combination of CDR structures for a specific target from a library of known CDR conformations. We developed a similar method for SAbDesigner. Our aim was to test all possible CDRs for other canonical conformations (not just one as in Liu et al. (2017)), but we do not want to optimise more than one CDR as in Pantazes and Maranas (2010) because we want at this stage to minimise the changes that are made to the original scaffold. In terms of optimisation criteria

the values that we use are higher buried surface area of the designed antibody in the docked position and a lower energy in terms of Rosetta energy units.

4.1.3 Summary of results

The validation and refinement pipelines were used on the initial anti-IL5 designs from the previous chapter (see Table 3.2). The validation showed that all but one of the initial designs have residues which are affected by the change in environment. These were hydrophobic residues which are more exposed to solvent in their new environment as opposed to their original native environment. For two of those designs the refinement pipeline identified possible non-hydrophobic point mutations, which are also predicted to maintain the native structure of the loop. For the rest of the designs no possible refinements were identified.

Another issue, which was identified by manual inspection, was that two of the designs contain a cysteine in a unbounded state. That cysteine residue formed a disulphide bridge with another cysteine on the original receptor. For these the point mutation tool was also used, but it did not predict any favourable mutations. In these cases we proposed refined designs that are based on the advice of experimentalists to change the cysteine to serine.

The refinement pipeline generated two other designs by optimising the CDRs of one of the initial designs, one of which was validated also by docking to show an increase which passes the threshold for validation (as opposed to the initial design which did not). These designs are predicted to have a significant increase in the buried surface area in comparison to the initial design. Finally, a further design was manually added that incorporates two independent refinements, one of the point mutations and one of the CDR optimisations. The designs that are described in this chapter are the ones which we have

selected to send for experimental validation along with the original designs (see Figure 4.6 for a full list of designs).

After the submission of those designs we also added to SAbDesigner the docking validation pipeline, which recapitulates the binding pose for four of the designs, indicating that they have the highest probability of binding the target in the matched molecular pairs position.

4.2 Methods

4.2.1 Validation

4.2.1.1 Changes in Accessible Surface Area

SAbDesinger quantifies changes in ASA for residues on the grafted loop and the antibody scaffold. The ASA is calculated for the native structure of the receptor, for the wild-type antibody framework, and for the final designed antibody (in all cases without the target present). We used NACCESS (Hubbard et al., 1992) for calculating the surface area with the default options. For the grafted loop the change that is considered is between the designed antibody and the native receptor. For the rest of the antibody the changes considered are between the wild-type framework and the native antibody. Large changes are then flagged up to the user using SAbDesigner's PyMol plugin, where residues are coloured to either orange or red.

A hydrophobic residue is coloured orange if the ASA change from the original environment to the new environment is in between +30Å^2 and +60Å^2 (see Figure 4.2.1.1), or it is coloured red if the change is greater than +60Å^2. The hydrophobic residues considered are alanine (ALA), valine (VAL), leucine (LEU), isolecuine (ILE), phenylalanine (PHE), methionine (MET) and tryptophan (TRP). For a polar residue a residue is coloured orange if the change is in between -30Å^2 and -60Å^2, or it is coloured red change if the change is lower than -60Å^2.

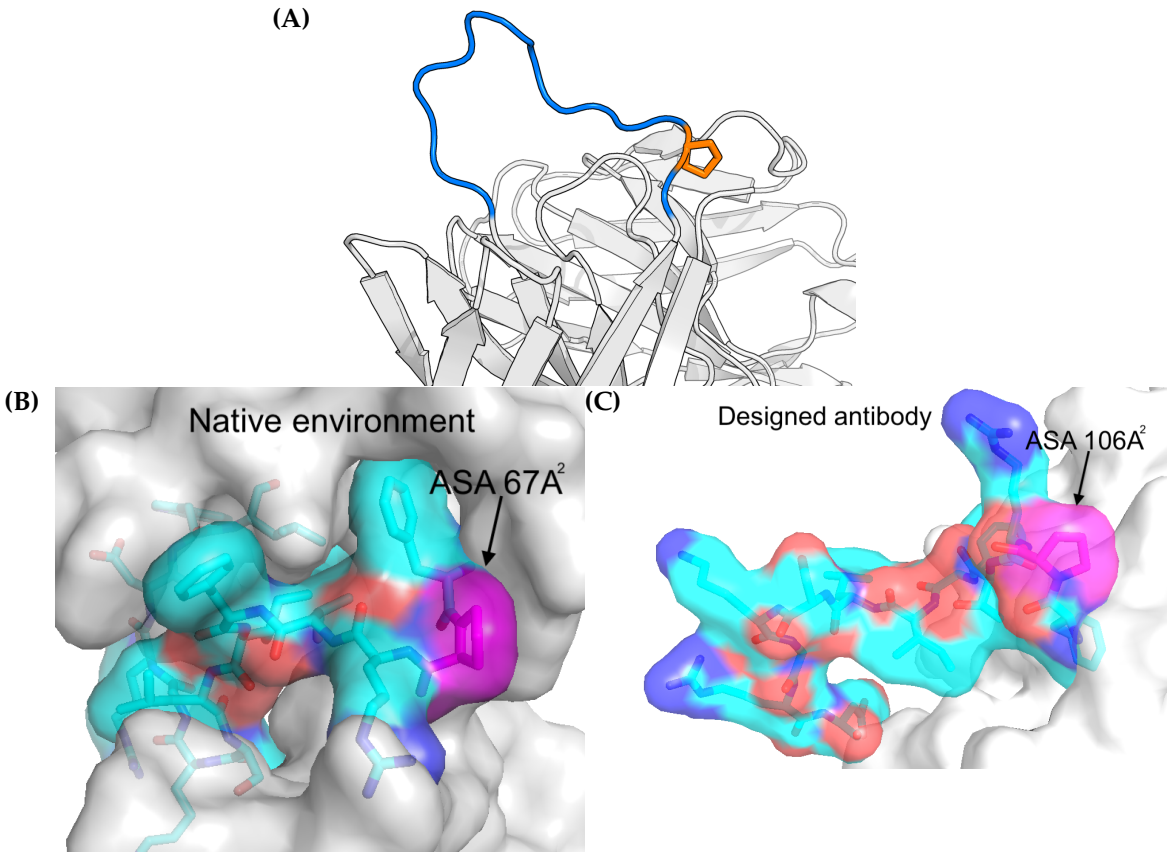

Figure 4.1: (A) Designed antibody with the grafted loop highlighted with blue, and the antibody scaffold shown in grey. Residue 26A Proline is shown in orange because of a positive change in ASA of $39A^2$.
(B) The grafted loop in its in original environment. The surface of the loop is shown in cyan, with the Proline residue highlighted in magenta, while the rest of the structure is shown in grey. The Proline residue can be seen packing against the surrounding structure.
(C) The grafted loop in the designed environment. The surface of the loop is shown in cyan, with the Proline residue highlighted in magenta, while the rest of the antibody is shown in grey. The Proline residue is not packing against the surrounding structure and has increased ASA.

The polar residues considered are glutamine (GLN), asparagine (ASN), histidine (HIS), serine (SER), threonine (THR), tyrosine (TYR) and cysteine (CYS).

4.2.1.2 Relaxation test

SAbDesigner also tests the behaviour of the grafted loop when it is allowed to move. This is performed using the Rosetta FastRelax relaxation protocol (Conway et al., 2014). This particular protocol was chosen because it only

performs local conformation perturbation and packing in iterative runs, in each run ramping up the weight of the repulsive forces. Each conformation is scored using the Rosetta energy function, and the one with the lowest energy is retained. The Rosetta energy function in its latest form uses statistical potentials and models that describe the energy landscape generated by local area interactions between residues, atom-atom interactions, solvation etc. (Alford et al., 2017). For our purposes the Rosetta energy function is particularly well suited due to the terms focusing on local backbone hydrogen bonding, side-chain hydrogen bonding and Ramachadran preferences which are well suited for local perturbations and high-resolution refinement.

We expect that if during the relaxation process opposing residues are found next to each other the ramped repulsive forces will identify a conformation with significant displacement for the grafted loop.

In the same fashion as in section 4.2.1.1 Rosetta FastRelax is run on the native structure of the receptor without the target present, for the wild-type antibody framework, and for the final designed antibody. The relaxation protocol step is stochastic, so we run it ten times and the median value is used for comparisons. The individual comparisons that are made are the following:

- The median Rosetta energy score of the native antibody is compared to that of the designed antibody.
- The median RMSD of the grafted loop on the designed antibody is compared to the RMSD on the native receptor.
- The median RMSD of the wild-type antibody is compared to that of the designed antibody.

We identified allowable thresholds for change in Rosetta energy. Clark et al. (2008) and Söderlind et al. (2000) have shown that it is possible to swap a CDR

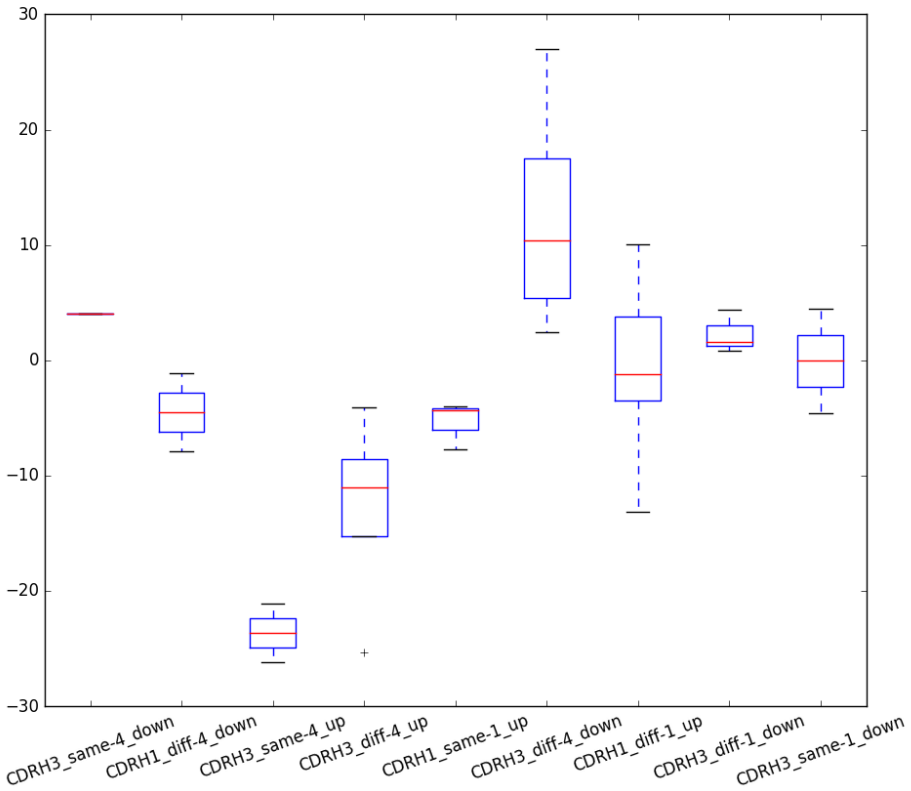

Figure 4.2: Rosetta energy difference between the structures before and after swapping the canonical forms. Each boxplot reflects one swap, with the label detailing the CDR and the length difference of the swap. The label is formatted as [CDR ID]_[framework sequence identity]_[change in length]. Same indicates that the framework sequence identity is above 90% between the donor antibody and the baseline antibody, while "diff" indicates it is below. "change in length" is the length difference between the swapped canonical forms.

on an antibody for a different canonical form from another antibody, and the antibody and the CDRs will still fold correctly. We used this premise to define allowable thresholds for change in Rosetta energy.

We used the antibody with PDB id 3mxv (chains H + L) and swapped one canonical form at a time for CDRs using the same grafting protocol as in the previous chapter. We performed a total of nine swaps. We then ran Rosetta relaxation and calculated the change in Rosetta energy, observing that the highest difference in medians being 23 Rosetta energy units (see Figure 4.2).

We used this threshold when grafting loops, retaining only the designs that have a change within this boundary. For the RMSD the threshold is set to 1Å from the values seen in the native environment.

4.2.1.3 Comparison to known antibodies

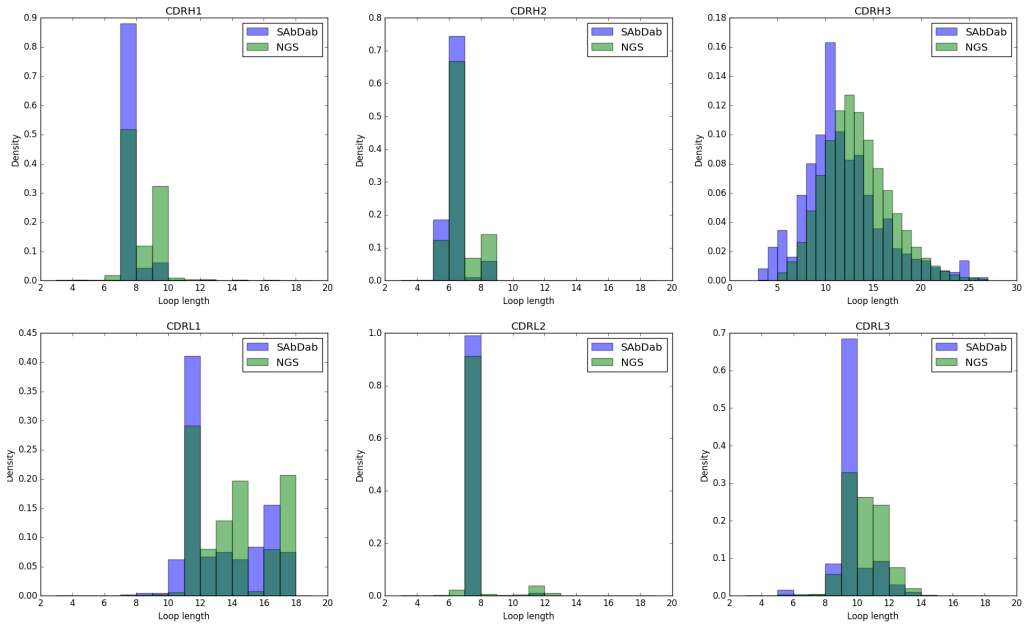

Figure 4.3: CDR loop length distribution comparison between SAbDab and the NGS dataset. Each subplot shows the length distribution of unique sequence for a particular CDR in SAbDab(blue) and the NGS dataset (green)

In order to test the viability of a design we also compared the length of the grafted loop versus the distribution of possible loop lengths for that particular CDR. Experimentalists have suggested that a grafted loop of an unusual length for a CDR could have the potential of destabilising the antibody, making it prone to aggregation. We considered two sources for computing the distribution of loop lengths, the CDR structures in SAbDab and the CDR sequences in the NGS data set (see Section 2.2.1.1). We computed both and found that they are different, the CDRs in the NGS dataset tending to have longer loop lengths than those

in the SAbDab data set (see Figure 4.3). We made the conservative decision of using the distribution from SAbDab because it reflects known soluble antibodies.

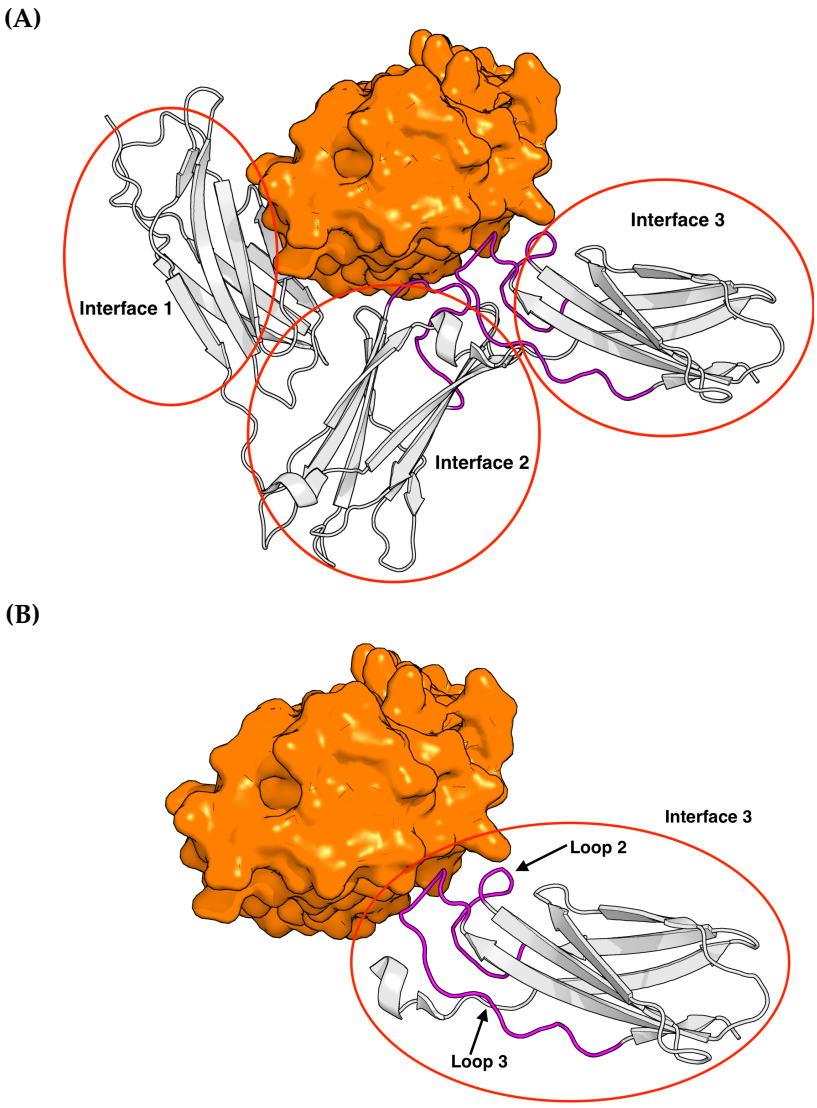

Figure 4.4: IL-5 (orange) in complex with its receptor (gray) from PDB file 3VA2
(A) The complete receptor of the IL-5 receptor, with the interfaces labelled.
(B) The reduced interface of the IL-5 receptor for loops 2 and 3.

4.2.1.4 Molecular docking

ZDock is a rigid body docking tool that generates possible docking poses for two individual proteins (Pierce et al., 2011). It uses the Fast Fourier Transfer

(FFT) to quickly generate a set of poses, which it then clusters and further refines using an energy function. Here we tested whether ZDock could identify the docked pose of the grafted antibody.

We use the 3.0.2 linux x64 stand-alone package, generating 3600 poses. The resulting docking file is transformed into individual PDB files containing the docked poses, which are then superpositioned to the matched molecular pairs docked pose of the design, using the receptor as reference. The RMSD is calculated for the residues that are part of the epitope (i.e. residues on the target which have at least one atom within 5.0Å of an antibody/receptor atom in the matched molecular pairs pose). The pose with the lowest RMSD is retained and considered the final result.

ZDock is run for both the receptor-target complex and the antibody-target complex, and the RMSD of the interface for both runs are compared. The residues of the binding loop(s) are provided as being part of the binding site pre-docking, by setting the ACE type to 19 in the pre-processed file, as indicated in the ZDock documentation.

The IL-5 receptor has characteristics which make it facile for ZDock to identify the correct binding pose. This is because the receptor is formed of three interfaces which together have a high geometric complementarity for IL-5 (see Figure 4.4A), and the FFT algorithm of ZDock identifies this very easily. A designed antibody will not be able to achieve this particular complementarity, primarily because of its size but also because it can only mimic one interface. This also aligns with our set out expectation that by mimicking a subset of the receptor the designed antibody will have a reduced affinity in comparison with the original receptor (see Section 3.4). Therefore, our standard of comparison should not be the entire receptor. We consider that a more suitable standard for the designs would be one in which the receptor is reduced to only the interface of the grafted loop(s). For example for IL-5 if the grafted loops are Loop 2 and

Loop(s) grafted	Interface	ZDock interface RMSD
Loop 0	Interface 2	1.09Å
Loop 3	Interface 3	3.01Å
Loops 2 and 3	Interface 3	3.01Å

Table 4.1: The binding interface RMSD obtained after computationally docking the receptor to the target with the loop(s) pre-defined as the binding site. These are used as validation standards for the designs that have the respective loops grafted. The loop ids used here are from Figure 3.4 and Figure 4.4

3, the receptor used should be interface 3 (see Figure 4.4B). In this context the standards identified using ZDock for the combinations of grafted loops are listed in Table 4.1. We compare these values with the values obtained for our designs.

Another method of evaluating the results of docking is by using the CAPRI criteria for an acceptable prediction of the binding interface on the ligand (i.e. the antigen in our case), which is 4.0Å (Lensink et al., 2017). This value, in the context of CAPRI, is used as an acceptable range in which a docking pose is correctly predicted. In a similar way we will use it as a cut-off for where a design binding pose is considered to be recapitulated correctly by docking. This threshold is also used as validation by two other *de novo* antibody designs methods, AbDesign (Lapidoth et al., 2015) and OptMaven (Li et al., 2014).

We also performed the same docking test as for the standards with ClusPro (Kozakov et al., 2017) and Rosetta Dock (Weitzner et al., 2017), two of the other best performers in the CAPRI docking assessment (Lensink et al., 2017). Unfortunately, neither of the tools were able to predict a pose for an interface of the native complex that falls within the 4.0Å CAPRI threshold. This suggests that ZDock is the best tool for this particular application.

4.2.2 Refinement

4.2.2.1 Point mutations

The first method used for refining a design identifies favourable mutations for residues which have been predicted through validation to have a destabilising effect on the designed antibody. These can be the ones automatically identified by SAbDesigner through the methods in section 4.2.1, or ones manually annotated by the user. Favourable mutations are identified by mutating the residue to a list of potential residue types and relaxing the structure using Rosetta FastRelax relaxation (Conway et al., 2014). The energy of the full design and RMSD displacement of the loop is calculated after each relaxation and compared to the original design, with the ones decreasing the energy and maintaining similar displacement being retained.

Point mutations are generally used by computationally mutagenesis tools to increase the affinity of an antibody. In the case of SAbDesigner our primary aim for using mutations is to increase the stability of the backbone given its new environment, and not to necessarily to increase affinity (although it is possible for one to affect the other). SAbDedsigner therefore restricts the list of potential residue types it tests to those that have been observed to allow the conformation of the original residue based on the constraints introduced by neighbouring residues. For this we used neighbour dependent Ramachandran probability distributions (Ting et al., 2010), which have been generated from existing structures. There is one distribution for each neighbour, left (upstream) and right (downstream) with the probability of the combination being given by Formula 4.1:

$$p(phi, psi|C, L, R) = \frac{p(phi, psi|C, L)p(phi, psi|C, R)}{p(phi, psi|C, R = ALL)} \tag{4.1}$$

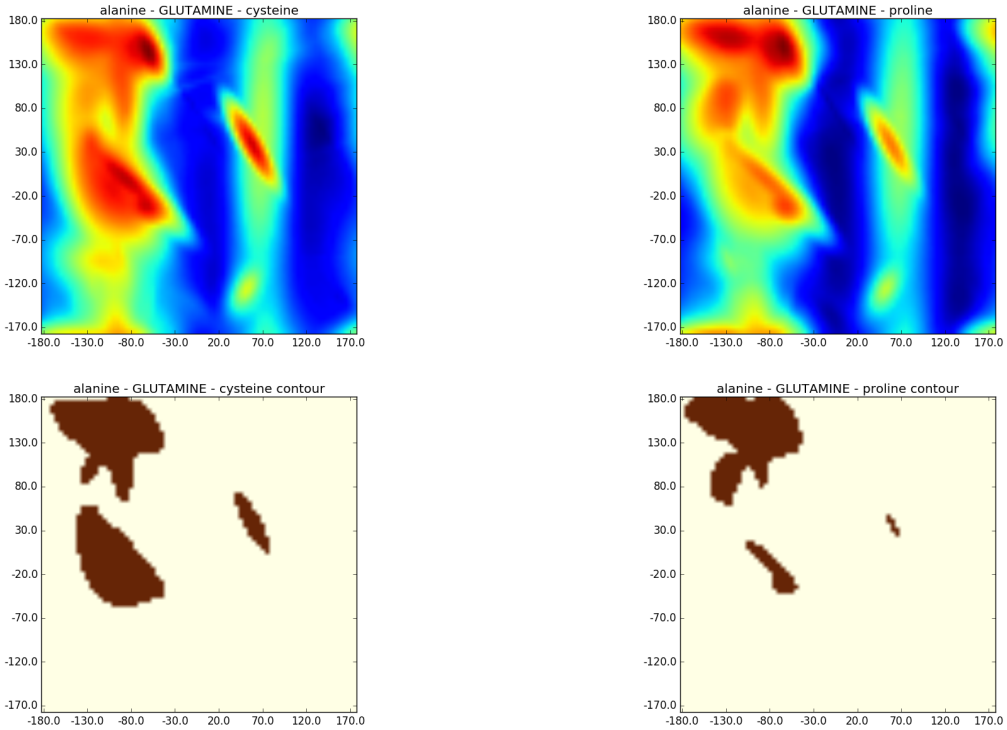

Figure 4.5: Neighbour dependent Ramachandran distributions and their 90% contours. The first line contains the probability density heat map for the neighbour dependent Ramachadran plots of GLUTAMINE, using two different neighbour pairs. The residues in the title indicate the residue and its neighbour, with the first being the left neighbour residue, 2^{nd} the central residue, and 3^{rd} the right residue. The 2^{nd} line contains the 90% contours of the above distributions with brown indicating areas part of the contour.

In formula 4.1 C is the amino acid type of the central residue, L is the left amino acid type and R is the right amino acid type. Ting et al. (2010) produced multiple versions of the probability distribution, each one being specific to a particular type of secondary structure pattern. In our case we used the TCB dataset (Turns, Coils and Bridges) which was created from all possible secondary structures of proteins except helices. This dataset contains not just loops but also beta strands, however, these are necessary as some loops exhibit the bridge like conformations that form a beta strand as part of their irregular secondary structure pattern (Ting et al., 2010).

For our purposes a residue type is only included in the possible mutation

list if the required dihedral angle combination (from the native residue) falls within the 90% contour of the neighbour dependent Ramachandran probability distribution (see Figure 4.5).

4.2.2.2 CDR optimisation

The second refinement method optimises a designed antibody for the target by identifying other more favourable canonical forms for the other CDRs (i.e. the CDRs that have not been mutated to a binding loop), or conformations in the case of CDR H3. For this, SAbDesigner uses the database of canonical forms described in section 3.2.5. For each CDR, every possible canonical form of its type is in turn grafted to the initial design in order to create a refined design. The canonical forms grafted in each refinement are checked for the same Cβ thresholds as the original grafted loop, both inter and intra protein, with the ones that do not pass being discarded. The refinements remaining have their BSA calculated, using the docking pose of the original design, with the ones resulting in an increase from the original design being retained.

4.3 Results & Discussion

4.3.1 Design 1

Design 1 is the design where SAbDesigner was able to graft two loops (i.e. Loop 2 and Loop 3) from the receptor on the antibody (see Figure 3.6B). Figure 4.7 shows the validation report for this design. It shows that Loop 3 is unusual as an antibody L3 CDR in terms of length (panel A), while Loop 2 is normal in length for CDR L1 (panel B) . It also highlights in panel D that there are four hydrophobic residues (three on Loop 2 and one on Loop 3) that are more exposed to solvent, compared to their original environment. Figure 4.7C shows that the full antibody energy and RMSD after relaxation increases with the grafted binding loop and the displacement of the grafted loop is higher than

Design	Scaffold	Chains	Grafted Loops	Other changes
1	3cfk	I,G	L3: Loop 3 (-4, -4)	None
			L1: Loop 2 (-1, 4)	None
		Refinement ID	**Refinement details**	
		1A	Phenylalanine(Phe)-27 mutated to Arginine	
		1B	Phenylalanine(Phe)-27 mutated to Glutamine	
		1C	Leucine(Leu)-95 mutated to Threonine	
		1D	Valine(Val)-25A mutated to Serine	
2	4rd7	O,P	L3: Loop 3 (-4, -4)	None
		Refinement ID	**Refinement details**	
		2A	Cysteine(Cys)-92A mutated to Serine	
3	1zlu	M,K	H1: Loop 0 (-1,2)	H3: 3u0t chain B
				H2: 2ojz chain H
4	5cil	H,L	H1: Loop 0 (-1,4)	H3: 1mjj chain B
5	4lss	H,L	H3: Loop 3(-2,-4)	None
		Refinement ID	**Refinement details**	
		5A	Cysteine-97(A) mutated to Serine	
		5B	Cysteine-97(A) mutated to Alanine	
		5C	CDR L1 swapped with L1 from 3bd3 chain B	
		5D	CDR L3 swapped with L3 from 4ht1 chain L	
		5E	CDR L3 swapped with L3 from 4ht1 chain L	
			Cysteine-97(A) mutated to Serine	

Figure 4.6: The designs from Table 3.2 (yellow background colour) along with their refined versions resulted from applying the methods in this chapter

in its native environment, but there are all within the allowable thresholds set out (see Section 4.2.1.2).

Our primary concern for this design was the several exposed hydrophobic residues. We then attempted to improve the design by mutating these residues to other residue types. Although the phenylalanine PHE-27 had the lowest increase in ASA out of the four flagged residues (i.e. coloured in orange), we concentrated on it first because it is a large residue and has the potential for non-specific interactions. Figure 4.8C shows the results of mutating this residue to the other allowable residue types (see Section 4.2.2.1). The mutation to arginine

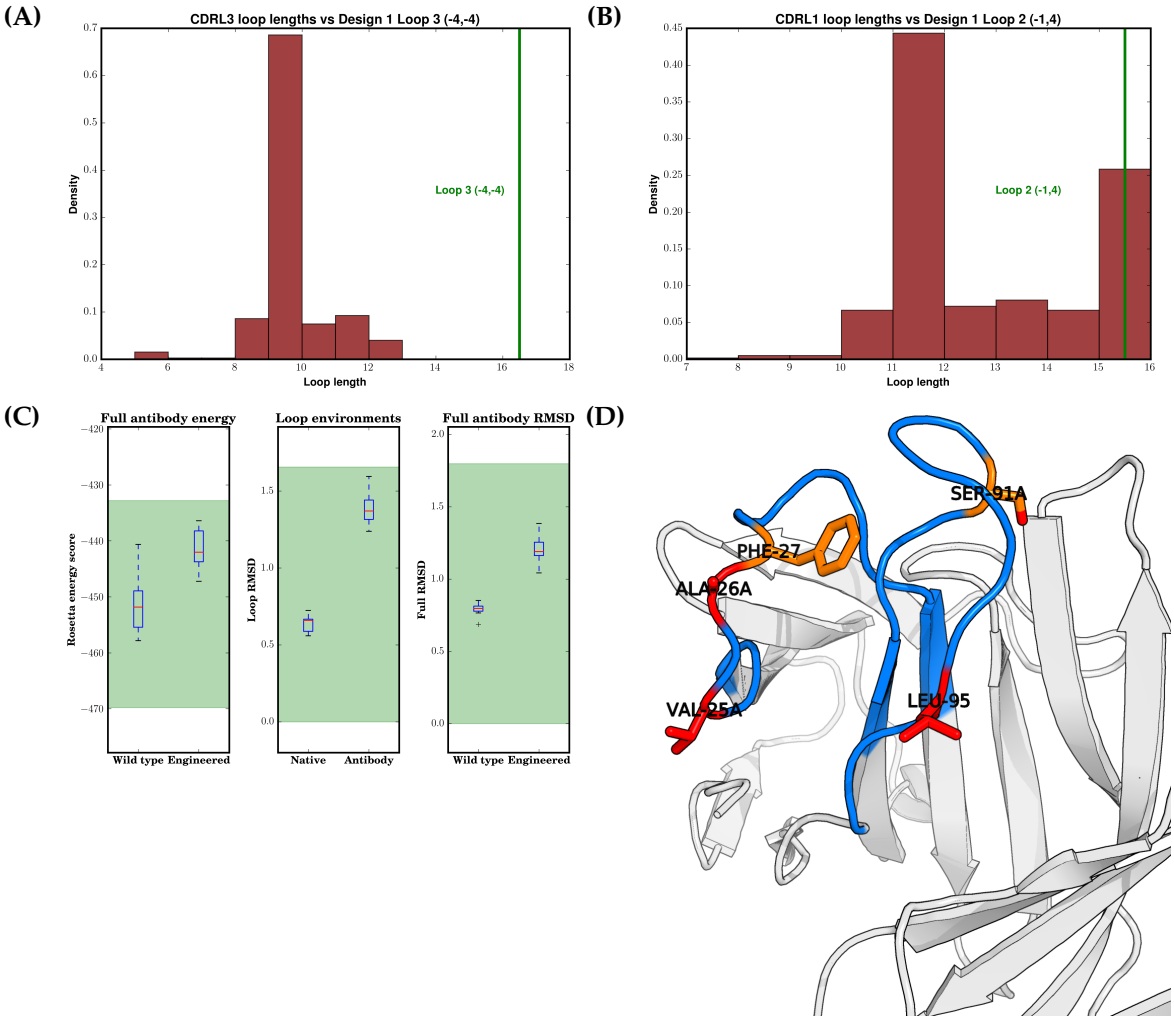

Figure 4.7: Validation report for Design 1.
(A) and (B) show the length of the grafted loops with a green bar, Loop 3 and Loop 2 respectively, compared to the distribution of loop lengths for the particular CDR (i.e. CDR L3 and CDR L1 respectively).
(C) Shows the results of performing the Rosetta FastRelax validation. The first subplot shows the difference in Rosetta energy units for the grafted loop between the native environment and the designed antibody. The second subplot shows the RMSD displacement of the loop between the two environments. The 3^{rd} subplot compares the RMSD after relaxation of the entire antibody before and after grafting. The ranges which fall within the thresholds are shown with a green background.
(D) Output from SAbDesigner's Pymol plugin that shows the differences in ASA for residues on the designed antibody. The grafted loops are shown in blue, and significant changes in ASA are shown with red and orange (see Section 4.2.1.1).

and glutamine are the best options, given that the structure has lower energy with them and the loop shows lower RMSD displacement from the required

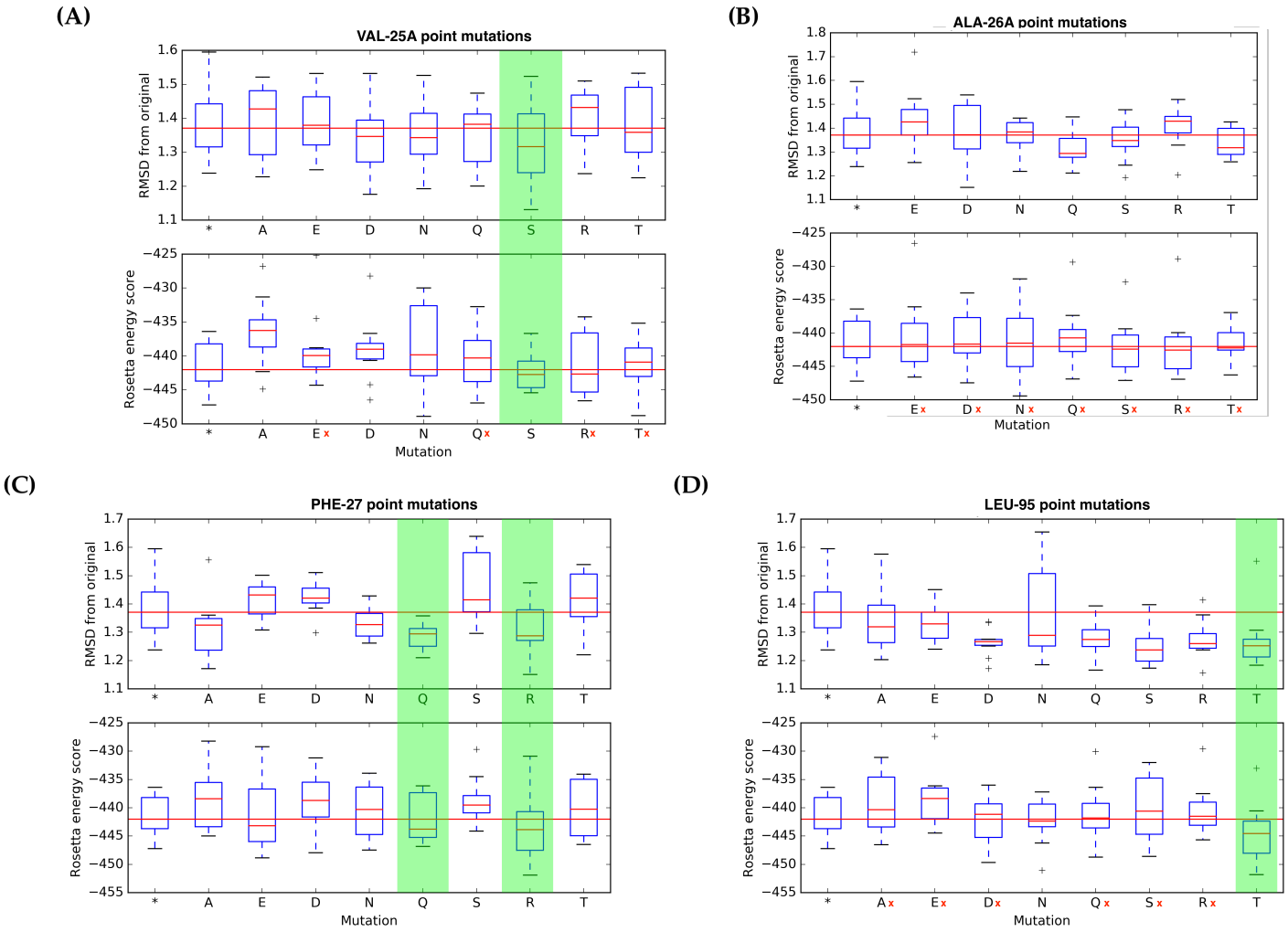

Figure 4.8: Refinement report for Design 1. In each subplot the results for mutating a particular residue on the designed antibody are shown. The upper part of each subplot shows the loop RMSD after relaxation for the individual residue type mutations, while the bottom subplot shows the change in total energy for the whole design (* is the current amino acid, and the red line on each plot is drawn on the median values of the current). The mutagenesis study was performed for residues 25A (A), 26A (B), 27 (C) and 95 (D). The selected mutations are highlighted in green, and a red cross indicates that the neighbour dependant dihedral combination is not allowed for the mutation

conformation. We therefore created the refined Design 1A and Design 1B that contain the mutation to arginine and glutamine respectively (see Figure 4.6).

Based on the same principle we decided to create Design 1C and 1D which contain a mutation of LEU-95 to threonine and VAL-25A to serine respectively, from the results of Figure 4.8D and A respectively. There was no favourable mutation identified for ALA-26A (see Figure 4.8B).

4.3.2 Design 2

Design 2 is a version of Design 1 without Loop 2. Without this loop there are fewer amino acids that have significant differences in terms of their ASA in their new environment (see Figure 4.9). However, there is now a cysteine in an unbonded state. We considered this to be an a more significant issue due to the cysteine's potential to form covalent bonds with other cysteines found in the antibody, and potentially denature the antibody. Therefore, instead of mutating the small hydrophobic residues we concentrated on mutating the cysteine.

The mutagenesis study did not find any viable options (see Figure 4.10). The only options that show lower or same level of RMSD deviation (glutamic acid and alanine) unfortunately also show increased energy values for the antibody. We made the decision, based on experimentalist advice, to create a manual refined design (Design 2A) in which the cysteine is replaced with a serine. The justification for this is that serine is polar, and it has a similar size through the similar side-chain (see Figure 1.3).

4.3.3 Design 3 and Design 4

In Design 3, Loop 0 is grafted on CDR H1. Loop 0 has 14 residues for its anchor variant, which is an unusual length for CDR H1. It can also be seen from the structural representation of the protein (see Figure 4.11C) that the loop is highly exposed, without any other structure to pack against. These issues have

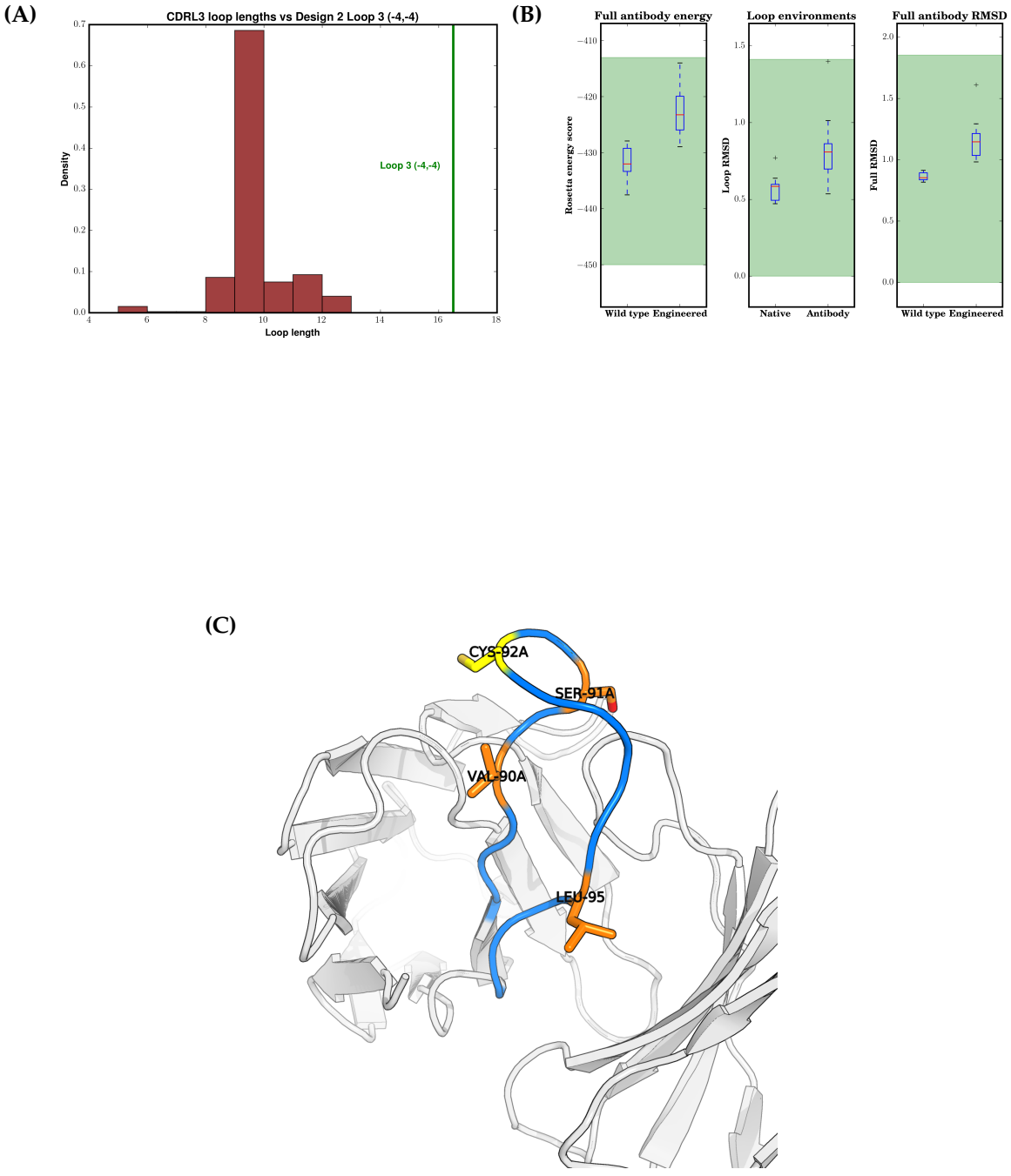

Figure 4.9: Validation report for Design 2. The individual subfigures follow the same pattern as Figure 4.7.
(A) How the length of the grafted loop compares with CDR L3 loop lengths.
(B) The results from the relaxation test.
(C) ASA changes report.
A new colour, yellow, is introduced in (C), where a cysteine is in an unbonded state.

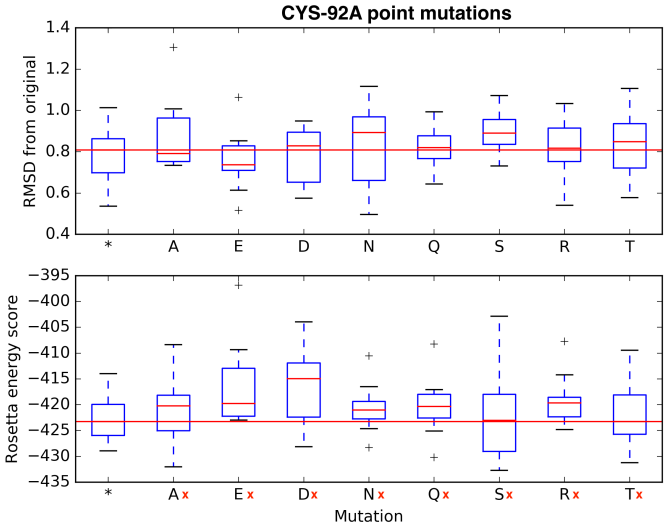

Figure 4.10: Refinement report for Design 2. The plot follows the same pattern as in Figure 4.8

also been raised by the ASA analysis (see Figure 4.11C), with many residues flagged. Two out of the relaxation runs showed the loop adopting conformations which were outside the allowable thresholds, although the median is within the allowable thresholds (see Figure 4.11B).

The result of running the refinement algorithm did not yield any mutations that can improve the design, and therefore this design was not further refined.

Design 4 also has loop 0 grafted on CDR H1 (see Figure 4.6), the only difference being that the anchor variant is different and the loop is two residues longer. After validation we have found that the issues that have been flagged up for Design 3 are also applicable to Design 4.

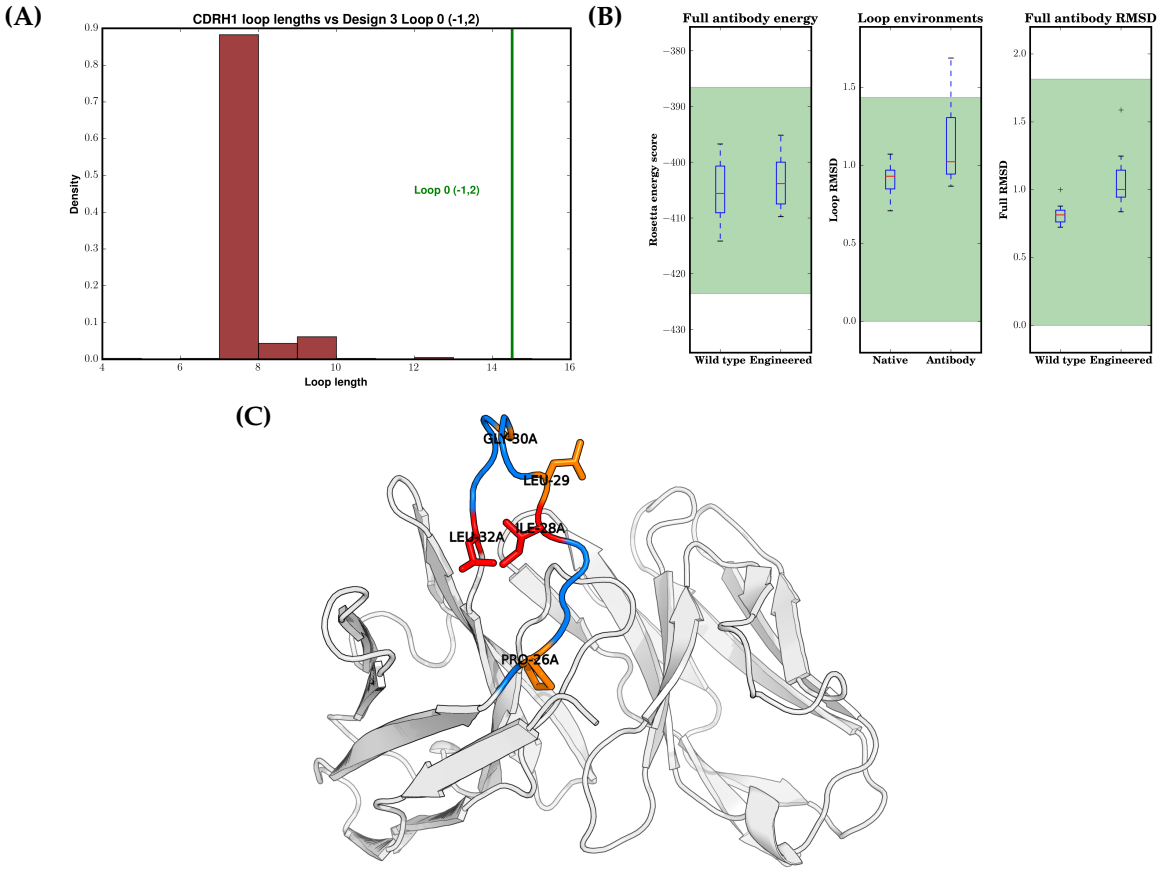

Figure 4.11: Validation report for Design 3. The individual subfigures follows the same pattern as Figure 4.7

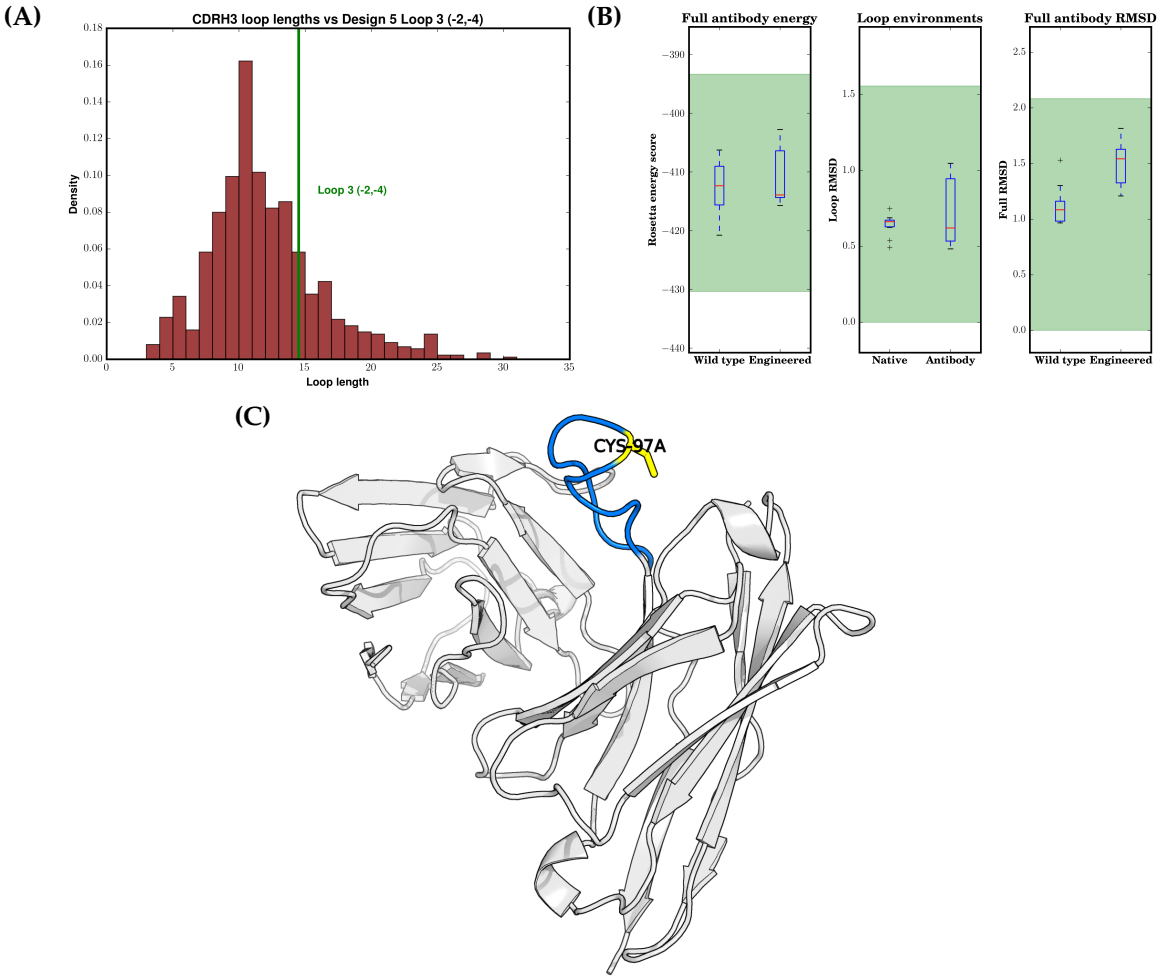

Figure 4.12: Validation report for Design 5. The individual subfigures follows the same pattern as Figure 4.7, with yellow added for the unbonded cysteine. (A) How the length of the grafted loop compares with CDR H3 loop lengths.
(B) The results from the relaxation test.
(C) ASA changes report.

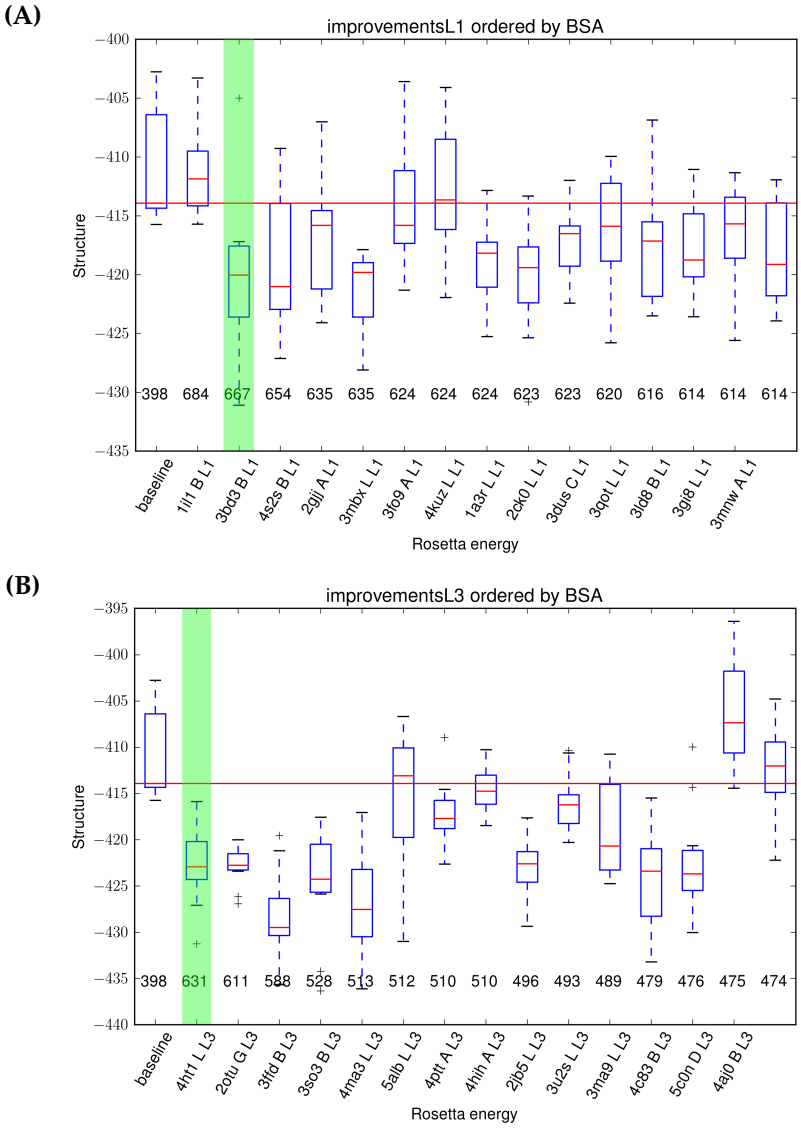

Figure 4.13: Refinement report for Design 5. Each boxplot represents a canonical form swap for L1 CDR (A) and L3 CDR (B). The y axis shows the Rosetta energy after relaxing the refined antibody, and at the bottom of each boxplot the total surface area of the antibody is reported. The first boxplot in each subplot are the values of the initial designed antibody (baseline). The red line is the median of the initial designed antibody. The selected mutations are highlighted in green

4.3.4 Design 5

Design 5 contains Loop 3 grafted on CDR H3. The validation report has shown that there is no residue that has a significant change in environment, and its length is common for H3 loops (see Figure 4.12C and A respectively). The relaxation test does not show any potential issues (see Figure 4.12B). This design has also the added benefit that the novel loop is grafted on CDR H3, the area of the antibody that naturally shows the greatest structural diversity. From this perspective we expect this design to have the highest probability of being expressed correctly.

The only issue flagged up is the cysteine, which without the presence of the cysteine from Loop 2 is an unbonded state. The mutagenesis study also did not yield any suitable mutations. We therefore chose as a possible refinement the default option from Design 2A of mutating the residue to serine (which is now Design 5A), but also to the alanine conservative mutation which is known for preserving backbone structure (Design 5B).

As this design has the lowest amount of potential expression issues we used it to run the affinity maturation protocol developed for SAbDesigner (see Section 4.2.2.2). We replaced the canonical forms on all the non-H3 CDRs to all the other possible ones, with positive results identified for CDR L1 and L3 (see Figure 4.13A and B respectively). To create new designs we have selected for both CDR L1 and L3 the top performing canonical form (best scoring BSA improvement, with Rosetta energy under baseline) and generated Design 5C and Design 5D respectively (see Figure 4.6).

At this point we had one more slot in the experimental validation quota and we chose to further refine design 5D by replacing the cysteine in the unbonded state with serine (Design 5E).

4.3.5 Docking

In terms of docking the Design 1 pose was recapitulated by ZDock with an average interface RMSD of of 4.65Å , with the standard being 3.01Å for the specific loops (Loop 2 + Loop 3) grafted on the design (see Table 4.1). In the case of Design 2 this went lower to 6.75Å , but this was expected as this Design 2 contains a grafted loop less than Design 1.

In the case of Design 3 and 4, for both runs, ZDock recapitulated the proposed binding site with 2.08Å RMSD, with the standard being 1.09Å for this grafted loop (see Table 4.1). This design is closest in terms of docking results to the standard, and also passes the 4.0Å threshold, and we therefore expect that this design is the one with highest probability to bind to the target if it is expressed correctly.

For Design 5, the docking pose was recapitulated by ZDock with 6.11Å RMSD. The canonical form swap in Design 5D and 5C, however, shows a substantial increase, with the binding pose being recapitulated with 3.24Å and 3.46Å RMSD. These results further suggest that the changes in canonical forms increase the potential binding affinity of design 5.

The docking validation based on the 4Å cut-off would suggest that Designs 3 and 4 have the greatest chance of replicating the binding site. Based on the similarity of values to the docking standard we expect Designs 5D and 5C to have the highest chance of replicating binding as their values are closest.

4.4 Conclusion

A semi-automated pipeline for validating and refining the initial designs produced by SAbDesigner was developed. The pipeline flags-up residues which pose a risk to the stability of the designed antibody, and then also suggests

refinements that remove those risk factors. Also, a protocol for affinity maturation based on optimising the other native CDRs has also been implemented. The combination of these methods have lead to the proposal of a further 10 designs for experimental validation (see Figure 4.6).

The original designs in the previous chapter have been further validated and refined using the above methods. Out of all the designs, Design 5 shows the greatest potential for expression due to a reduced set of risks being identified, and the loop being grafted on the region which is known to accept the greatest structural variability, CDR H3. This design was further refined to remove the only risk factor identified, the cysteine in an unbonded state. We have also suggested refinements where the canonical forms of the CDRs were swapped in order to increase BSA.

The designs that have been identified to provide the greatest risk for expression are Designs 3 and 4, and unfortunately they did not have any predicted positive mutation in the mutagenesis study. On the other hand, docking has predicted that these designs have the greatest chance of replicating the binding site.

All the designs listed in Figure 4.6 have been submitted for experimental validation to GSK.

There is no plan for no deal because we are going to get a great deal

— Boris Johnson MP

5 Conclusion and future work

Contents

Antibodies are a crucial part of the immune system and their modularity, low toxicity, and high affinity & specificity towards their targets make them a robust platform for the engineering of bio-therapeutics. The antibody therapeutics development landscape is heavily based on experimental methods that simulate the diversity generation and affinity maturation processes in other organisms, while rational computational methods are less established. As a result, the development process for the majority of antibody therapeutics developed so far do not contain a rational computational step.

In this thesis we analysed the existing computational antibody design

methods, focusing on two classes in particular: antibody modelling and *de novo* antibody design. In the case of antibody modelling we estimated how much of the antibody sequence space can be modelled using existing structures using an NGS dataset of 15 million sequences. We concluded that all but the H3 have sequences for which we already know the structure. We then further analysed why H3 is hard to model, and concluded that this is not method dependent but in fact that the H3 subset of loops has unique characteristics from the rest of the non-antibody loops.

In the case of *de novo* antibody design we developed SAbDesign a novel automated pipeline for computational design of antibodies. This is based on mimicking on an antibody the binding area of a receptor of the target protein. We used as an example target the IL-5 protein, an important therapeutic protein target and generated a set of 15 possible antibodies. We also conducted an applicability study across other important therapeutic protein targets.

SAbDesigner has also a validation pipeline which tests if the grafted loop will fold in the correct conformation, and a refinement pipeline to mutate residues which are detrimental to the design or increase affinity through canonical class changes of the other CDRs.

This chapter summarises the major results and conclusions of the thesis, and also proposes further research venues that can be explored in the future.

5.1 Chapter 2. Antibody CDR loops structural diversity

In Chapter 2 we analysed the proportion of the antibody sequence space which can be modelled in an NGS dataset of 15 million naïve antibody sequences. To model this dataset we modified the loop modelling algorithm FREAD to allow the prediction of the structure of a loop without the presence of the rest of

structure, which was a requirement in the original algorithm. After adapting the algorithm we retested the threshold values for the ESS score. We identified that it is best for this measure to be length dependent, and proposed length dependent values. The results showed that for CDRs L1, L2, L3 and H1 the existing structures cover the majority of the AA-constrained sequence space (in terms of identical matches), and in the case of the unique sequence space the majority of sequences can be modelled using an existing solved structure from the PDB.

In the case of CDR H2 over 50% of the unique sequences can not be modelled. We believe this to be an artefact of the method, which is generated by the depth at which the the sequence space at length six is explored by the antibodies in the data set. This essentially is a result of a recall problem for the algorithm, and we believe is not reflective of an actual increased structural diversity. We recently conducted a Metropolis Hastings parameter sweep for the revised version of FREAD, where we allowed multiple ranges of loop lengths to have independent thresholds. At each step either the ESS threshold was changed for a loop length interval, or the boundaries of one of the intervals was changed randomly. We discovered that it is possible to increase recall for shorter loop lengths, while maintaining the FPR. There is a risk of over fitting the dataset, and this will have to be further analysed in the future. This could potentially solve the issue with the sequences of length six, and can also result in higher coverage overall for the algorithm.

The other CDR for which a large percentage of the unique sequence space can not be modelled is CDR H3. We further analysed what makes H3 hard to model, and our results suggest that the issue is not the method, but rather it stems from the fact that the CDR H3 has unique structural characteristics in comparison to the loops in the non-antibody world. This comes in the form of unique full loop structures, unique four residue fragments and unique dihedral combinations for Tyrosine and Glycine. We believe that these results can be

used in future structure prediction methods which are specifically tailored for CDR H3. This can come in the form of an *ab initio* method that would extend the Ramachandran areas used for sampling in the case of Tyrosine and Glycine. In a similar fashion the unique fragments can be incorporated as part of the input list for the fragment database for Rosetta (Gront et al., 2011), where the fragment library is compiled using fragments that contain the unusual Tyrosine and Glycine conformations.

FREAD can also be modified to compress the unmodellable sequences (i.e. cluster the unmodellable loops in sequences that potentially share the same structure). The original version of FREAD also contains a substitution table that is dihedral angle independent. This means it could be used as a similarity measure between two sequences, without any sequence information. The thresholds for similarity can be recalculated, but set to minimise the amount of false positives for the top prediction. This will result in stricter cut-offs. An initial analysis that we have conducted showed that it is possible to reduce the amount of unmodellable sequences to around a potential 50%, with the possibility of extending this even further. This approach can be important for prioritising the H3 sequences that should be solved experimentally in the future, with a higher priority towards the sequences that would results in a greater coverage in the AA-constrained sequence space. The strict cut-offs can, however, result in a large proportion of false negatives and this quantity should be closely monitored.

5.2 Chapter 3. Designing antibodies using non-antibody protein loops or fragments

In Chapter 3 we presented SAbDesigner, an automatic pipeline for designing antibodies using non-protein fragments. SAbDesigner mimics the interface of a non-antibody receptor of the target by grafting loops from the binding interface

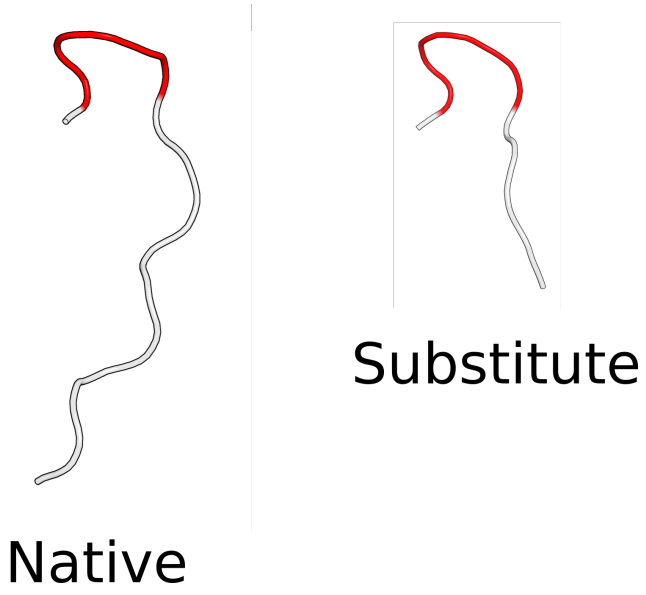

Figure 5.1: An example of a substitute loop that shares the same binding motif as Loop 3 (Native). The substitute has different anchor characteristics than the native.

on an antibody framework. The important loops for binding are automatically identified, and also an antibody framework that can accommodate those loops is automatically identified. The loop is checked to fit within the constraints of the antibody using pairwise residue specific Cβ thresholds. Through this designing pipeline five different designs were proposed for targeting IL-5, by mimicking the binding loops of its alpha receptor.

Part of the criteria for establishing if a loop can be grafted on a specific CDR of a framework is that the four residue anchors (two upstream and two downstream) of the native loop, and the anchors of the antibody framework are similar to within 1.0Å RMSD. There is evidence from (Hu et al., 2007) that this value can relaxed to a two residue anchor (one upstream and one downstream), and a 3.0Å RMSD threshold can be used. Using a relaxed anchor criteria can potentially increase the number of binding loops for which a scaffold can be identified, and therefore increasing the number of potential designs. As a result of the relaxed cut-offs greater care will have to be taken to make sure that the loops pack against the framework as in Hu et al. (2007).

If a loop can not be modelled on an antibody because of anchor constraints or the DSSP definition, a different approach can be used. The specific motif within the loop that modulates binding (the linear peptide that contains all the contacts with the antigen) can be isolated. The PDB can then be queried for a loop structure that contains the same motif, but has different anchors which can potentially allow for the loop to be grafted (see Figure 5.1). This approach is similar to the work of (Liu et al., 2017), but in their case they limit their search to known antibody CDRs.

The areas of an antibody which are used by SAbDesigner for grafting novel loops are the CDRs. OptMaven (Li et al., 2014) proposed different modular parts for antibodies, using instead of CDRs the gene segments that make up the Fv area as areas to design (see Figure 1.14A). Their claim is that this separation would allow for better conservation of interactions that maintain the stability of the antibody, although they do not detail their experimental justification that proves that this separation is optimal. We do, however, see as beneficial to preserve native interactions in the fringe areas of the grafted loop (Clark et al., 2008). We believe that in future work it would be useful to include the native interactions between the anchor residues and the loop, in the framework selection algorithm. The algorithm would favour antibody frameworks that have similar residue types in the anchor region, or residue types that would allow for the formation of similar interactions. If a framework structure does not exist with such residues, the NGS data set of sequences could also used to identify frameworks with high sequence identity that have these residues.

One aspect that was not considered when we chose the specific IL-5 receptor target complex was the affinity of the initial complex, which we have been told by experimentalists afterwards is in the mMol range. As we have discussed in the results, because SAbDesigner is mimicking only a subset of the receptor binding area, it would be important to start from high affinity complexes in

the proof-of-concept examples. As such, we have proposed a list of designs for three other important therapeutic targets for which the receptor has a higher binding affinity against the target.

There are other applications that we envisage for SAbDesigner. Aside from proposing an antibody binder it can also be used to enhance experimental methods. One of the scenarios in which phage display is used is on a universal library of antibodies with rearranged segments (see Section 1.3.2.4). This approach does not guarantee a success, as there might not be any antibody in that library that has required affinity against the target. This in turns leads to the library further being mutated. Instead of mutating what is essentially a blind library against the target, the starting library could be the designed antibody supplemented by a broad spectrum of mutations made to the CDR areas of that antibody. Our expectation is that on average a library with a designed antibody as a starting point would take fewer iterations before reaching a high affinity binder than a blind library, due to the fact that the antibody already contains a known binding motif to the target.

5.3 Chapter 4. SAbDesigner validation and refinement

5.3.1 Validation

In Chapter 4 we described the computational validation pipeline for SabDesigner. This includes testing if the loop displacement for the native environment and the antibody environment is within similar values, changes in ASA between the environments, changes in Rosetta energy on the designed antibody versus the original and molecular docking. The goal for the first three tests was to ensure that the loop will maintain its structure in the new environment on the antibody, and therefore be capable of modulating binding to the target.

The movement made by the Rosetta FastRelax algorithm are an approximation of movement over a short period of time. A different approach to this type of validation is allowing the structure to explore conformational using a molecular dynamics simulation (MD) (Bös and Pleiss, 2009). There are two possible validations scenarios that could be followed when using MD. The first scenario involves simulating the loop both its native environment and its designed environment, and then performing a differential analyse on the conformational space they explore. The second scenario involves simulating the designed structure to identify if it is likely to maintain its expected conformation, or if it is likely to form interactions that might alter it shape. The problem with MD simulations is that they are computationally intensive, and require multiple runs to establish statistical significance. These would only be run on a reduced short-list of possible designs, as it would not be feasible to run them systematically on all potential designs.

5.3.2 Refinement

SAbDesigner also contains refinement methodologies for the designs, in the form of single residue mutations and optimisation of the other CDRs for the target. Single point mutations are performed for those amino acids which were flagged up by during validation as being detrimental to stability. These residues are mutated in turn to all the other residue types that allow for the same conformation to be maintained based on a neighbour dependent Ramachandran distribution. The effect of each mutation was measured using a Rosetta FastRelax protocol, and the favourable ones were retained. The single residue mutation approach is generally used for affinity maturation, the reason why we did not choose this path is that we wanted to validate through experimental methods designs that are close as possible to the original. In the future, we do however envisage adding that to the pipeline to increase the affinity of the antibodies

by including residues that would form more favourable interactions with the target in the matched molecular pairs docked pose.

In the same way the replacement of all the other CDRs on the antibody can be extended on SaBDesigner like in OptCDR (Pantazes and Maranas, 2010) where all the other CDRs are optimised against the target. This would, however, require the exploration of more frameworks with different angles between the VH and VL and therefore require further optimisations to deal with the increased computational time.

5.4 Experimental validation

GSK are currently in the process of running validation experiments for all the designs in Figure 4.6. By the 4^{th} of October, the time of submission of this thesis, they have attempted to express designs 1A-D (not including the original design), 2A, 5, 5A-E. They have not detected expression levels in designs 1A-1D and 5D, low expression in designs 5C and 5E, and suitable expression in designs 5, 5A and 5B.

The results for the refined designs of initial design 1 suggest that changing only one residue was not enough to mitigate the issues around the many hydrophobic residues exposed to solvent. In hindsight, we believe that in the future one refined design should contain all the hydrophobic residues mutated to allowable polar ones.

The expression data from designs 5C-5E suggest that the swapping of canonical forms on other CDR loops needs to be further validated, although the number of cases were not broad enough to provide a clear picture of the suitability of this method, especially that there are cases where this has succeeded (Liu et al., 2017).

The results from 5, 5A and 5B are very encouraging, as we have concluded before the data was received that design 5 and its point mutation designs have the highest chance of being expressed.

The results as a whole suggest that a sensible cut-off for validation would be at having no orange or red flagged residues on the designed loop. We are still awaiting for expression data from Design 3 and 4 (which we have predicted through docking to have the highest potential of replicating binding), and affinity data for designs 5, 5A and 5B. A separate, more thorough, analysis would have to be conducted at that point.

Bibliography

Bissan Al-Lazikani, Arthur M Lesk, and Cyrus Chothia. Standard conformations for the canonical structures of immunoglobulins. *Journal of molecular biology*, 273(4):927–948, 1997. (Cited on pages 21 and 89.)

Bruce Alberts. *Molecular biology of the cell*. Garland science, 2017. (Cited on page 23.)

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J OâĂŹMeara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 2017. (Cited on page 118.)

Juan C Almagro, Alexey Teplyakov, Jinquan Luo, Raymond W Sweet, Sreekumar Kodangattil, Francisco Hernandez-Guzman, and Gary L Gilliland. Second antibody modeling assessment (ama-ii). *Proteins: Structure, Function, and Bioinformatics*, 82(8):1553–1562, 2014. (Cited on pages 36, 37, 46, and 78.)

Mads Hald Andersen, David Schrama, Per thor Straten, and Jürgen C Becker. Cytotoxic t cells. *Journal of Investigative Dermatology*, 126(1):32–41, 2006. (Cited on page 24.)

Amy C Anderson. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003. (Cited on page 13.)

Francesco A Aprile, Pietro Sormanni, Michele Perni, Paolo Arosio, Sara Linse, Tuomas PJ Knowles, Christopher M Dobson, and Michele Vendruscolo.

Selective targeting of primary and secondary nucleation pathways in aβ42 aggregation using a rational antibody scanning method. *Science advances*, 3(6): e1700488, 2017. (Cited on page 84.)

Ryutaro Asano, Yukiko Sone, Koki Makabe, Kouhei Tsumoto, Hiroki Hayashi, Yu Katayose, Michiaki Unno, Toshio Kudo, and Izumi Kumagai. Humanization of the bispecific epidermal growth factor receptor$\times$ cd3 diabody and its efficacy as a potential clinical reagent. *Clinical cancer research*, 12(13):4036–4042, 2006. (Cited on page 32.)

Mariana Babor and Tanja Kortemme. Multi-constraint computational design suggests that native sequences of germline antibody h3 loops are nearly optimal for conformational flexibility. *Proteins: Structure, Function, and Bioinformatics*, 75(4):846–858, 2009. (Cited on pages 35 and 47.)

Lies Baeten, Joke Reumers, Vicente Tur, François Stricher, Tom Lenaerts, Luis Serrano, Frederic Rousseau, and Joost Schymkowitz. Reconstruction of protein backbones from the brix collection of canonical protein fragments. *PLoS Comput Biol*, 4(5):e1000083, 2008. (Cited on page 67.)

Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. How cryo-em is revolutionizing structural biology. *Trends in biochemical sciences*, 40(1):49–57, 2015. (Cited on page 15.)

Pierre A Barthelemy, Helga Raab, Brent A Appleton, Christopher J Bond, Ping Wu, Christian Wiesmann, and Sachdev S Sidhu. Comprehensive analysis of the factors contributing to the stability and solubility of autonomous human vh domains. *Journal of Biological Chemistry*, 283(6):3639–3654, 2008. (Cited on page 114.)

HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The protein data bank nucleic acids research, 28, 235-242. *View Article PubMed/NCBI Google Scholar*, 2000. (Cited on pages 15 and 16.)

Bibiana Bielekova, Nancy Richert, Thomas Howard, Gregg Blevins, Silva Markovic-Plese, Jennifer McCartin, Jens Würfel, Joan Ohayon, Thomas A Waldmann, Henry F McFarland, et al. Humanized anti-cd25 (daclizumab) inhibits disease activity in multiple sclerosis patients failing to respond to interferon β. *Proceedings of the National Academy of Sciences of the United States of America*, 101(23):8705–8708, 2004. (Cited on page 32.)

Sara Birtalan, Yingnan Zhang, Frederic A Fellouse, Lihua Shao, Gabriele Schaefer, and Sachdev S Sidhu. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *Journal of molecular biology*, 377(5):1518–1528, 2008. (Cited on page 65.)

Fabian Bös and Jürgen Pleiss. Multiple molecular dynamics simulations of tem β-lactamase: dynamics and water binding of the ω-loop. *Biophysical journal*, 97(9):2550–2558, 2009. (Cited on page 146.)

Carl Ivar Branden et al. *Introduction to protein structure*. Garland Science, 1999. (Cited on page 11.)

Patrick M Buck, Sandeep Kumar, Xiaoling Wang, Neeraj J Agrawal, Bernhardt L Trout, and Satish K Singh. Computational methods to predict therapeutic protein aggregation. *Therapeutic proteins: Methods and protocols*, pages 425–451, 2012. (Cited on page 112.)

Alexander Bujotzek, James Dunbar, Florian Lipsmeier, Wolfgang Schäfer, Iris Antes, Charlotte M Deane, and Guy Georges. Prediction of vh–vl domain orientation for antibody variable domain modeling. *Proteins: Structure, Function, and Bioinformatics*, 83(4):681–695, 2015. (Cited on page 39.)

J Callaway, M Cummings, B Deroski, P Esposito, A Forman, P Langdon, M Libeson, J McCarthy, J Sikora, D Xue, et al. Protein data bank contents guide:

Atomic coordinate entry format description. *Brookhaven National Laboratory*, 1996. (Cited on page 17.)

Paul J Carter. Potent antibody therapeutics by design. *Nature Reviews Immunology*, 6(5):343–357, 2006. (Cited on pages 30 and 82.)

Patrick Chames, Marc Van Regenmortel, Etienne Weiss, and Daniel Baty. Therapeutic antibodies: successes, limitations and hopes for the future. *British journal of pharmacology*, 157(2):220–233, 2009. (Cited on page 31.)

Anuj Chaudhri, Isidro E Zarraga, Sandeep Yadav, Thomas W Patapoff, Steven J Shire, and Gregory A Voth. The role of amino acid sequence in the self-association of therapeutic monoclonal antibodies: insights from coarse-grained modeling. *The Journal of Physical Chemistry B*, 117(5):1269–1279, 2013. (Cited on page 112.)

Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics*, 26(5):689–691, 2010. (Cited on page 86.)

Jieming Chen, Nicholas Sawyer, and Lynne Regan. Protein–protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Science*, 22(4):510–515, 2013. (Cited on page 85.)

Rong Chen, Li Li, and Zhiping Weng. Zdock: an initial-stage protein-docking algorithm. *Proteins: Structure, Function, and Bioinformatics*, 52(1):80–87, 2003. (Cited on page 113.)

Yoonjoo Choi and Charlotte M Deane. Fread revisited: accurate loop structure prediction using a database search algorithm. *Proteins: Structure, Function, and Bioinformatics*, 78(6):1431–1440, 2010. (Cited on pages 38, 46, 50, 51, and 88.)

Yoonjoo Choi and Charlotte M Deane. Predicting antibody complementarity determining region structures without classification. *Molecular Biosystems*, 7 (12):3327–3334, 2011. (Cited on pages 50 and 53.)

Cyrus Chothia and Arthur M Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823, 1986. (Cited on pages 34 and 36.)

Cyrus Chothia and Arthur M Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of molecular biology*, 196(4):901–917, 1987. (Cited on pages 21, 23, 38, 46, and 67.)

Jo-Lan Chung, Wei Wang, and Philip E Bourne. Exploiting sequence and structure homologs to identify protein–protein binding sites. *Proteins: Structure, Function, and Bioinformatics*, 62(3):630–640, 2006. (Cited on page 85.)

Louis A Clark, Skanth Ganesan, Sarah Papp, and Herman WT van Vlijmen. Trends in antibody sequence changes during the somatic hypermutation process. *The Journal of Immunology*, 177(1):333–340, 2006. (Cited on pages 46 and 63.)

Louis A Clark, P Ann Boriack-Sjodin, Eric Day, John Eldredge, Christopher Fitch, Matt Jarpe, Stephan Miller, You Li, Ken Simon, and Herman WT Van Vlijmen. An antibody loop replacement design feasibility study and a loop-swapped dimer structure. *Protein Engineering, Design & Selection*, 22(2):93–101, 2008. (Cited on pages 111, 118, and 144.)

UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014. (Cited on page 92.)

Patrick Conway, Michael D Tyka, Frank DiMaio, David E Konerding, and David Baker. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science*, 23(1):47–55, 2014. (Cited on pages 112, 117, and 124.)

Saulo HP de Oliveira, Jiye Shi, and Charlotte M Deane. Building a better fragment library for de novo protein structure prediction. *PloS one*, 10(4): e0123998, 2015. (Cited on page 67.)

Charlotte M Deane and Tom L Blundell. Coda: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Science*, 10(3):599–612, 2001. (Cited on page 35.)

Brandon J DeKosky, Oana I Lungu, Daechan Park, Erik L Johnson, Wissam Charab, Constantine Chrysostomou, Daisuke Kuroda, Andrew D Ellington, Gregory C Ippolito, Jeffrey J Gray, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proceedings of the National Academy of Sciences*, 113(19):E2636–E2645, 2016. (Cited on page 49.)

Zygmunt S Derewenda. Rational protein crystallization by mutational surface engineering. *Structure*, 12(4):529–535, 2004. (Cited on page 14.)

William S DeWitt, Paul Lindau, Thomas M Snyder, Anna M Sherwood, Marissa Vignali, Christopher S Carlson, Philip D Greenberg, Natalie Duerkopp, Ryan O Emerson, and Harlan S Robins. A public database of memory and naive b-cell receptor sequences. *PloS one*, 11(8):e0160853, 2016. (Cited on page 49.)

Márcio Dorn, Mariel Barbachan e Silva, Luciana S Buriol, and Luis C Lamb. Three-dimensional protein structure prediction: Methods and computational strategies. *Computational biology and chemistry*, 53:251–276, 2014. (Cited on pages 35 and 38.)

J Dunbar, A Fuchs, J Shi, and CM Deane. Abangle: characterising the vh–vl orientation in antibodies. *Protein Engineering, Design & Selection*, 26(10): 611–620, 2013a. (Cited on page 39.)

James Dunbar and Charlotte M Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2015. (Cited on pages 21, 50, 55, and 89.)

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2013b. (Cited on pages 88, 89, and 91.)

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014. (Cited on pages 53, 55, and 57.)

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Claire Marks, Jaroslaw Nowak, Cristian Regep, Guy Georges, Sebastian Kelm, Bojana Popovic, and Charlotte M Deane. Sabpred: a structure-based antibody prediction server. *Nucleic acids research*, 44(W1):W474–W478, 2016. (Cited on page 82.)

Liam J Fanning, Alison M Connor, and Gillian E Wu. Development of the immunoglobulin repertoire. *Clinical immunology and immunopathology*, 79(1): 1–14, 1996. (Cited on page 46.)

Alexey N Fedorov and Thomas O Baldwin. Cotranslational protein folding. *Journal of Biological Chemistry*, 272(52):32715–32718, 1997. (Cited on page 6.)

Narcis Fernandez-Fuentes and András Fiser. Saturating representation of loop conformational fragments in structure databanks. *BMC structural biology*, 6(1): 15, 2006. (Cited on page 175.)

Krzysztof Fidelis, Peter S Stern, David Bacon, and John Moult. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein engineering*, 7(8):953–960, 1994. (Cited on page 67.)

András Fiser, Richard Kinh Gian Do, et al. Modeling of loops in protein structures. *Protein science*, 9(9):1753–1773, 2000. (Cited on page 36.)

Martin F Flajnik and Masanori Kasahara. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics*, 11(1):47–59, 2010. (Cited on page 23.)

André Frenzel, Thomas Schirrmann, and Michael Hust. Phage display-derived human antibodies in clinical development and therapy. In *MAbs*, volume 8, pages 1177–1194. Taylor & Francis, 2016. (Cited on page 33.)

Karen S Frese, Hugo A Katus, and Benjamin Meder. Next-generation sequencing: from understanding biology to personalized medicine. *Biology*, 2(1):378–398, 2013. (Cited on page 48.)

Harvey Friedman. Programming pearls. *Linux Journal*, 2000(73es):18, 2000. (Cited on page 60.)

Jacob Glanville, Wenwu Zhai, Jan Berka, Dilduz Telman, Gabriella Huerta, Gautam R Mehta, Irene Ni, Li Mei, Purnima D Sundar, Giles MR Day, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences*, 106(48):20216–20221, 2009. (Cited on pages 28 and 46.)

Julian Gough and Cyrus Chothia. Superfamily: Hmms representing all proteins of known structure. scop sequence searches, alignments and genome assignments. *Nucleic acids research*, 30(1):268–272, 2002. (Cited on page 11.)

Julian Gough, Kevin Karplus, Richard Hughey, and Cyrus Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of molecular biology*, 313 (4):903–919, 2001. (Cited on page 55.)

Dominik Gront, Daniel W Kulp, Robert M Vernon, Charlie EM Strauss, and David Baker. Generalized fragment picking in rosetta: design, protocols and applications. *PloS one*, 6(8):e23294, 2011. (Cited on page 142.)

Ortho Multicenter Transplant Study Group et al. A randomized clinical trial of okt3 monoclonal antibody for acute rejection of cadaveric renal transplants. *New Eng J Med*, 313:337–342, 1985. (Cited on page 31.)

Rebecca Hamer, Qiang Luo, Judith P Armitage, Gesine Reinert, and Charlotte M Deane. i-patch: Interprotein contact prediction using local network information. *Proteins: Structure, Function, and Bioinformatics*, 78(13):2781–2797, 2010. (Cited on page 85.)

Steven A Hofmeyr. An interpretative introduction to the immune system. *Design principles for the immune system and other distributed autonomous systems*, 3:28–36, 2001. (Cited on page 26.)

Daniel Holtby, Shuai Cheng Li, and Ming Li. Loopweaver: loop modeling by the weighted scaling of verified proteins. *Journal of Computational Biology*, 20 (3):212–223, 2013. (Cited on page 35.)

Annemarie Honegger and Andreas Plückthun. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *Journal of molecular biology*, 309(3):657–670, 2001. (Cited on page 21.)

Xiaozhen Hu, Huanchen Wang, Hengming Ke, and Brian Kuhlman. High-resolution design of a protein loop. *Proceedings of the National Academy of Sciences*, 104(45):17668–17673, 2007. (Cited on page 143.)

Simon J Hubbard, Janet M Thornton, and Simon F Campbell. Substrate recognition by proteinases. *Faraday discussions*, 93:13–23, 1992. (Cited on pages 85 and 116.)

Howook Hwang, Thom Vreven, Joël Janin, and Zhiping Weng. Protein–protein docking benchmark version 4.0. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3111–3114, 2010. (Cited on page 91.)

William Ying Khee Hwang, Juan Carlos Almagro, Timothy N Buss, Philip Tan, and Jefferson Foote. Use of human germline genes in a cdr homology-based approach to antibody humanization. *Methods*, 36(1):35–42, 2005. (Cited on page 31.)

James A Irving, James C Whisstock, and Arthur M Lesk. Protein structural alignments and functional genomics. *Proteins: Structure, Function, and Bioinformatics*, 42(3):378–382, 2001. (Cited on page 67.)

Peter T Jones, Paul H Dear, Jefferson Foote, Michael S Neuberger, and Greg Winter. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*, 321(6069):522–525, 1986. (Cited on pages 31 and 83.)

W Kabsch and C Sander. Secondary structure definition by the program dssp. *Biopolymers*, 22:2577–2637, 1983a. (Cited on pages 55 and 57.)

Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 34(5):827–828, 1978. (Cited on pages 17, 86, and 90.)

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983b. (Cited on page 85.)

Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39): 15674–15679, 2013. (Cited on page 34.)

Simon King. The best selling drugs of all time; humira joins the elite. *Pharma & Healthcare. Forbes. January*, 28, 2013. (Cited on page 33.)

Georges Köhler and Cesar Milstein. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256(5517):495–497, 1975. (Cited on page 30.)

Rahul M Kohli, Shaun R Abrams, Kiran S Gajula, Robert W Maul, Patricia J Gearhart, and James T Stivers. A portable hot spot recognition loop transfers sequence preferences from apobec family members to activation-induced cytidine deaminase. *Journal of Biological Chemistry*, 284(34):22898–22904, 2009. (Cited on page 83.)

Dima Kozakov, David R Hall, Bing Xia, Kathryn A Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda. The cluspro web server for protein-protein docking. *nature protocols*, 12(2):255–278, 2017. (Cited on page 123.)

Konrad Krawczyk, Terry Baker, Jiye Shi, and Charlotte M Deane. Antibody i-patch prediction of the antibody binding site improves rigid local antibody–antigen docking. *Protein Engineering, Design & Selection*, 26(10):621–629, 2013. (Cited on page 113.)

Daisuke Kuroda, Hiroki Shirai, Matthew P Jacobson, and Haruki Nakamura. Computer-aided antibody design. *Protein Engineering Design and Selection*, 25 (10):507–522, 2012. (Cited on page 47.)

Seisuke Kusano, Mutsuko Kukimoto-Niino, Nobumasa Hino, Noboru Ohsawa, Masashi Ikutani, Satoshi Takaki, Kensaku Sakamoto, Miki Hara-Yokoyama, Mikako Shirouzu, Kiyoshi Takatsu, et al. Structural basis of interleukin-5 dimer recognition by its α receptor. *Protein Science*, 21(6):850–864, 2012. (Cited on page 95.)

Gideon D Lapidoth, Dror Baran, Gabriele M Pszolla, Christoffer Norn, Assaf Alon, Michael D Tyka, and Sarel J Fleishman. Abdesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins: Structure, Function, and Bioinformatics*, 83(8):1385–1406, 2015. (Cited on pages 40, 83, 110, 111, 113, and 123.)

Jinwoo Leem, James Dunbar, Guy Georges, Jiye Shi, and Charlotte M Deane. Abodybuilder: Automated antibody structure prediction with data–driven accuracy estimation. In *MAbs*, volume 8, pages 1259–1268. Taylor & Francis, 2016. (Cited on pages 36, 37, and 38.)

Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, 2003. (Cited on pages 21 and 49.)

Marc F Lensink, Sameer Velankar, and Shoshana J Wodak. Modeling protein–protein and protein–peptide complexes: Capri 6th edition. *Proteins: Structure, Function, and Bioinformatics*, 85(3):359–377, 2017. (Cited on pages 113 and 123.)

Cyrus Levinthal. How to fold graciously. *Mossbauer spectroscopy in biological systems*, 67:22–24, 1969. (Cited on page 35.)

Tong Li, Robert J Pantazes, and Costas D Maranas. Optmaven–a new framework for the de novo design of antibody variable region models targeting specific antigen epitopes. *PloS one*, 9(8):e105954, 2014. (Cited on pages 113, 123, and 144.)

Shaun M Lippow, K Dane Wittrup, and Bruce Tidor. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nature biotechnology*, 25(10):1171–1176, 2007. (Cited on pages 83 and 113.)

Tao Liu, Guangsen Fu, Xiaozhou Luo, Yan Liu, Ying Wang, Rongsheng E Wang, Peter G Schultz, and Feng Wang. Rational design of antibody protease inhibitors. *Journal of the American Chemical Society*, 137(12):4042–4045, 2015. (Cited on page 84.)

Xiaofeng Liu, Richard D Taylor, Laura Griffin, Shu-Fen Coker, Ralph Adams, Tom Ceska, Jiye Shi, Alastair DG Lawson, and Terry Baker. Computational design of an epitope-specific keap1 binding antibody using hotspot residues grafting and cdr loop swapping. *Scientific Reports*, 7:41306, 2017. (Cited on pages 41, 83, 110, 111, 114, 144, and 147.)

N Lonberg. Human monoclonal antibodies from transgenic mice. In *Therapeutic Antibodies*, pages 69–97. Springer, 2008. (Cited on page 32.)

Robert M MacCallum, Andrew CR Martin, and Janet M Thornton. Antibody-antigen interactions: contact analysis and binding site topography. *Journal of molecular biology*, 262(5):732–745, 1996. (Cited on page 47.)

Shikha Malhotra, Susan Kovats, Weiguo Zhang, and K Mark Coggeshall. B cell antigen receptor endocytosis and antigen presentation to t cells require vav and dynamin. *Journal of Biological Chemistry*, 284(36):24088–24097, 2009. (Cited on page 24.)

C Marks and CM Deane. Antibody h3 structure prediction. *Computational and structural biotechnology journal*, 15:222–231, 2017. (Cited on pages 38 and 83.)

Claire Marks, Jaroslaw Nowak, Stefan Klostermann, Guy Georges, James Dunbar, Jiye Shi, Sebastian Kelm, and Charlotte M Deane. Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics*, 33(9):1346–1353, 2017. (Cited on page 35.)

BW Matthews, H Nicholson, and WJ Becktel. Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proceedings of the National Academy of Sciences*, 84(19):6663–6667, 1987. (Cited on page 13.)

Tatiana Maximova, Ryan Moffatt, Buyong Ma, Ruth Nussinov, and Amarda Shehu. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS computational biology*, 12(4):e1004619, 2016. (Cited on page 35.)

John McCafferty, Andrew D Griffiths, Greg Winter, and David J Chiswell. Phage antibodies: filamentous phage displaying antibody variable domains. *nature*, 348(6301):552–554, 1990. (Cited on pages 32 and 33.)

Kenneth Merz and Scott M LeGrand. *The protein folding problem and tertiary structure prediction*. Springer Science & Business Media, 2012. (Cited on page 9.)

Mario Abdel Messih, Rosalba Lepore, and Anna Tramontano. Looping: a template-based tool for predicting the structure of protein loops. *Bioinformatics*, 31(23):3767–3772, 2015. (Cited on page 35.)

S Miersch and SS Sidhu. Synthetic antibodies: concepts, potential and practical considerations. *Methods*, 57(4):486–498, 2012. (Cited on page 82.)

Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. BionumbersâĂŤthe database of key numbers in molecular and cell biology. *Nucleic acids research*, 38(suppl_1):D750–D753, 2009. (Cited on page 3.)

Julian Mintseris, Brian Pierce, Kevin Wiehe, Robert Anderson, Rong Chen, and Zhiping Weng. Integrating statistical pair potentials into protein complex prediction. *Proteins: Structure, Function, and Bioinformatics*, 69(3):511–520, 2007. (Cited on page 113.)

Sherie L Morrison, M Jacqueline Johnson, Leonard A Herzenberg, and Vernon T Oi. Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *Proceedings of the National Academy of Sciences*, 81(21):6851–6855, 1984. (Cited on page 31.)

Mary Munson, Suganthi Balasubramanian, Karen G Fleming, Athena D Nagi, Ronan O'Brien, Julian M Sturtevant, and Lynne Regan. What makes a protein a protein? hydrophobic core designs that specify stability and structural properties. *Protein Science*, 5(8):1584–1593, 1996. (Cited on page 6.)

Kenneth M. Murphy. *Janeway's immunobiology*. Garland Science, 2008. (Cited on page 28.)

Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995. (Cited on pages 11 and 55.)

Kiyotaka Nakano, Takahiro Ishiguro, Hiroko Konishi, Megumi Tanaka, Masamichi Sugimoto, Izumi Sugo, Tomoyuki Igawa, Hiroyuki Tsunoda, Yasuko Kinoshita, Kiyoshi Habu, et al. Generation of a humanized anti-glypican 3 antibody by cdr grafting and stability optimization. *Anti-cancer drugs*, 21(10):907–916, 2010. (Cited on pages 31 and 83.)

Magali Nicaise, Marielle Valerio-Lepiniec, Philippe Minard, and Michel Desmadril. Affinity transfer by cdr grafting on a nonimmunoglobulin scaffold. *Protein science*, 13(7):1882–1891, 2004. (Cited on pages 31 and 83.)

N Nishimoto and T Kishimoto. Humanized antihuman il-6 receptor antibody, tocilizumab. *Therapeutic Antibodies*, pages 151–160, 2008. (Cited on page 31.)

A Nissim and Y Chernajovsky. Historical development of monoclonal antibody therapeutics. *Therapeutic Antibodies*, pages 3–18, 2008. (Cited on pages 30 and 31.)

Benjamin North, Andreas Lehmann, and Roland L Dunbrack. A new clustering of antibody cdr loop conformations. *Journal of molecular biology*, 406(2):228–256, 2011. (Cited on pages 23, 38, 46, and 67.)

Jaroslaw Nowak, Terry Baker, Guy Georges, Sebastian Kelm, Stefan Klostermann, Jiye Shi, Sudharsan Sridharan, and Charlotte M. Deane. Length-independent structural similarities enrich the antibody cdr canonical class model. *mAbs*, 8 (4):751–760, 2016. (Cited on pages 23, 38, 46, and 67.)

Christine A Orengo, AD Michie, S Jones, David T Jones, MB Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997. (Cited on page 11.)

World Health Organization et al. General policies for monoclonal antibodies: Inn working document 09.251. *Update*, 18:12, 2009. (Cited on page 31.)

RJ Pantazes and CD Maranas. Optcdr: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Engineering, Design & Selection*, 23(11):849–858, 2010. (Cited on pages 39, 83, 110, 111, 114, and 147.)

Robert J Pantazes and Costas D Maranas. Maps: a database of modular antibody parts for predicting tertiary structures and designing affinity matured antibodies. *BMC bioinformatics*, 14(1):168, 2013. (Cited on page 40.)

S Parthasarathy and MRN Murthy. Analysis of temperature factor distribution in high-resolution protein structures. *Protein science*, 6(12):2561–2567, 1997. (Cited on page 58.)

Linus Pauling, Robert B Corey, and Herman R Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide

chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951. (Cited on page 6.)

Yingjie Peng, Wenwen Zeng, Hui Ye, Kyung Ho Han, Venkatasubramanian Dharmarajan, Scott Novick, Ian A Wilson, Patrick R Griffin, Jeffrey M Friedman, and Richard A Lerner. A general method for insertion of functional proteins within proteins via combinatorial selection of permissive junctions. *Chemistry & biology*, 22(8):1134–1143, 2015. (Cited on page 84.)

Alan S Perelson and George F Oster. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of theoretical biology*, 81(4):645–670, 1979. (Cited on page 46.)

Brian G Pierce, Yuichiro Hourai, and Zhiping Weng. Accelerating protein docking in zdock using an advanced 3d convolution library. *PloS one*, 6(9): e24657, 2011. (Cited on page 121.)

Alessandro Pini, Francesca Viti, Annalisa Santucci, Barbara Carnemolla, Luciano Zardi, Paolo Neri, and Dario Neri. Design and use of a phage display library human antibodies with subnanomolar affinity against a marker of angiogenesis eluted from a two-dimensional gel. *Journal of Biological Chemistry*, 273(34):21769–21776, 1998. (Cited on page 33.)

Venkata Giridhar Poosarla, Tong Li, Boon Chong Goh, Klaus Schulten, Thomas K Wood, and Costas D Maranas. Computational de novo design of antibodies binding to a peptide with high affinity. *Biotechnology and bioengineering*, 114(6): 1331–1342, 2017. (Cited on page 40.)

Gopalasamudram Narayana Ramachandran, Chandrasekharan Ramakrishnan, and V Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7(1):95–99, 1963. (Cited on page 6.)

Steven L Reiner. Development in motion: helper t cells at work. *Cell*, 129(1): 33–36, 2007. (Cited on page 24.)

Christopher J Roberts. Therapeutic protein aggregation: mechanisms, design, and control. *Trends in biotechnology*, 32(7):372–380, 2014. (Cited on page 111.)

Stephen I Rudnick and Gregory P Adams. Affinity and avidity in antibody-based tumor targeting. *Cancer Biotherapy and Radiopharmaceuticals*, 24(2):155–161, 2009. (Cited on page 26.)

Sarah Sirin, James R Apgar, Eric M Bennett, and Amy E Keating. Ab-bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016. (Cited on page 113.)

George P Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228:1315–1318, 1985. (Cited on page 82.)

George P Smith and Valery A Petrenko. Phage display. *Chemical reviews*, 97(2): 391–410, 1997. (Cited on page 32.)

Jeffrey W Smith, Kathy Tachias, and Edwin L Madison. Protein loop grafting to construct a variant of tissue-type plasminogen activator that binds platelet integrin αiibβ3. *Journal of Biological Chemistry*, 270(51):30486–30490, 1995. (Cited on page 83.)

MS Smyth and JHJ Martin. x ray crystallography. *Molecular Pathology*, 53(1):8, 2000. (Cited on pages 13, 14, and 16.)

Eskil Söderlind, Leif Strandberg, Pernilla Jirholt, Norihiro Kobayashi, Vessela Alexeiva, Anna-Maria Åberg, Anna Nilsson, Bo Jansson, Mats Ohlin, Christer Wingren, et al. Recombining germline-derived cdr sequences for creating diverse single-framework antibody libraries. *Nature biotechnology*, 18(8):852, 2000. (Cited on page 118.)

Pietro Sormanni, Francesco A Aprile, and Michele Vendruscolo. Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, 112(32):9902–9907, 2015. (Cited on page 84.)

Amelie Stein and Tanja Kortemme. Improvements to robotics-inspired conformational sampling in rosetta. *PLoS One*, 8(5):e63090, 2013. (Cited on page 36.)

Alexey Teplyakov, Galina Obmolova, Thomas J. Malia, Jinquan Luo, Salman Muzammil, Raymond Sweet, Juan Carlos Almagro, and Gary L. Gilliland. Structural diversity in a human antibody germline library. *mAbs*, 8(6):1045–1063, 2016. (Cited on page 47.)

Daniel Ting, Guoli Wang, Maxim Shapovalov, Rajib Mitra, Michael I Jordan, and Roland L Dunbrack Jr. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLoS computational biology*, 6(4):e1000763, 2010. (Cited on pages 114, 124, and 125.)

Susumu Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909): 575–581, 1983. (Cited on pages 46 and 47.)

Maria Trott, Svenja Weiß, Sascha Antoni, Joachim Koch, Hagen von Briesen, Michael Hust, and Ursula Dietrich. Functional characterization of two scfv-fc antibodies from an hiv controller selected on soluble hiv-1 env complexes: a neutralizing v3-and a trimer-specific gp41 antibody. *PloS one*, 9(5):e97478, 2014. (Cited on page 33.)

Sameer Velankar, Glen van Ginkel, Younes Alhroub, Gary M Battle, John M Berrisford, Matthew J Conroy, Jose M Dana, Swanand P Gore, Aleksandras Gutmanas, Pauline Haslam, et al. Pdbe: improved accessibility of macromolecular structure data from pdb and emdb. *Nucleic acids research*, 44(D1): D385–D395, 2015. (Cited on page 92.)

Gestur Vidarsson, Gillian Dekkers, and Theo Rispens. Igg subclasses and allotypes: from structure to effector functions. *Frontiers in immunology*, 5, 2014. (Cited on page 28.)

Flavio Vincenti, Marianne Lantz, Jytte Birnbaum, Marvin Garovoy, Diane Mould, John Hakimi, Keith Nieforth, and Susan Light. A phase i trial of humanized anti-interleukin 2 receptor antibody in renal transplantation1. *Transplantation*, 63(1):33–38, 1997. (Cited on page 32.)

Guoli Wang and Roland L Dunbrack Jr. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003. (Cited on pages 57 and 90.)

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005. (Cited on page 93.)

Brian D Weitzner and Jeffrey J Gray. Accurate structure prediction of cdr h3 loops enabled by a novel structure-based c-terminal constraint. *The Journal of Immunology*, 198(1):505–515, 2017. (Cited on page 38.)

Brian D Weitzner, Roland L Dunbrack, and Jeffrey J Gray. The origin of cdr h3 structural diversity. *Structure*, 23(2):302–311, 2015. (Cited on page 47.)

Brian D Weitzner, Jeliazko R Jeliazkov, Sergey Lyskov, Nicholas Marze, Daisuke Kuroda, Rahel Frick, Jared Adolf-Bryfogle, Naireeta Biswas, Roland L Dunbrack Jr, and Jeffrey J Gray. Modeling and docking of antibody structures with rosetta. *Nature protocols*, 12(2):401–416, 2017. (Cited on pages 82 and 123.)

Bernard L Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, pages 28–35, 1947. (Cited on page 64.)

Greg Winter, Andrew D Griffiths, Robert E Hawkins, and Hennie R Hoogenboom. Making antibodies by phage display technology. *Annual review of immunology*, 12(1):433–455, 1994. (Cited on page 33.)

Alexander Wlodawer, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski. Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *The FEBS journal*, 280 (22):5705–5736, 2013. (Cited on page 13.)

Andreas Wollenberg and Thomas Bieber. Antigen presenting cells. *Atopic dermatitis. Marcel Dekker, New York*, pages 267–283, 2002. (Cited on page 24.)

Tai Te Wu and EA Kabat. An analysis of the sequences of the variable regions of bence jones proteins and myeloma light chains and their implications for antibody complementarity. *The Journal of experimental medicine*, 132(2):211–250, 1970. (Cited on pages 19 and 20.)

K Wüthrich. Nuclear magnetic resonance spectroscopy of proteins. *Encyclopedia of Life Sciences*, 2001. (Cited on page 14.)

Zhexin Xiang, Cinque S Soto, and Barry Honig. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Sciences*, 99(11):7432–7437, 2002. (Cited on page 36.)

Jian Ye, Ning Ma, Thomas L Madden, and James M Ostell. Igblast: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research*, 41(W1):W34–W40, 2013. (Cited on page 49.)

Michael Zemlin, Martin Klinger, Jason Link, Cosima Zemlin, Karl Bauer, Jeffrey A Engler, Harry W Schroeder, and Perry M Kirkham. Expressed murine and human cdr-h3 intervals of equal length exhibit distinct repertoires

that differ in their amino acid composition and predicted range of structures. *Journal of molecular biology*, 334(4):733–749, 2003. (Cited on pages 46 and 65.)

Feng Zhu, Zhe Shi, Chu Qin, Lin Tao, Xin Liu, Feng Xu, Li Zhang, Yang Song, Xianghui Liu, Jingxian Zhang, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic acids research*, 40(D1):D1128–D1136, 2011. (Cited on page 92.)

Appendices

I would be in the upper half of the 10, so we'll be looking at 7, 7.5 ... maybe 7

— Jeremy Corbyn MP, when asked to rate on a scale of 1-10 how passionate he is about staying in the EU

A Appendix

A.1 Unique loop fragments clustering algorithm

This section further details the algorithm used to cluster four residue fragments from non-antibody proteins presented in section 2.2.2.8.

The theoretical complexity of this method is O(nk), where n is the number of fragments and k is the number of clusters. In the worst case scenario If k=n, and the complexity of this method is O(n^2), which is the same as generating the distance matrix for a standard clustering algorithm. However, we know that the number of clusters is expected to be much lower as there is a high redundancy at the level of four residue fragments (Fernandez-Fuentes and Fiser, 2006).

This method can also be parallelised by splitting the list of fragments equally to individual cores. This is enabled by keeping the list of clusters in a shared memory so all the clusters can access it (see Figure A.1). The parallel implementation is an approximation of the sequential algorithm described above, with the difference being that it can occur that at the same time two individual fragments which are identical are made clusters. This happens when two fragments which are identical are processed by two separate cores at the same time, and the list of clusters does not contain an identical fragment. This

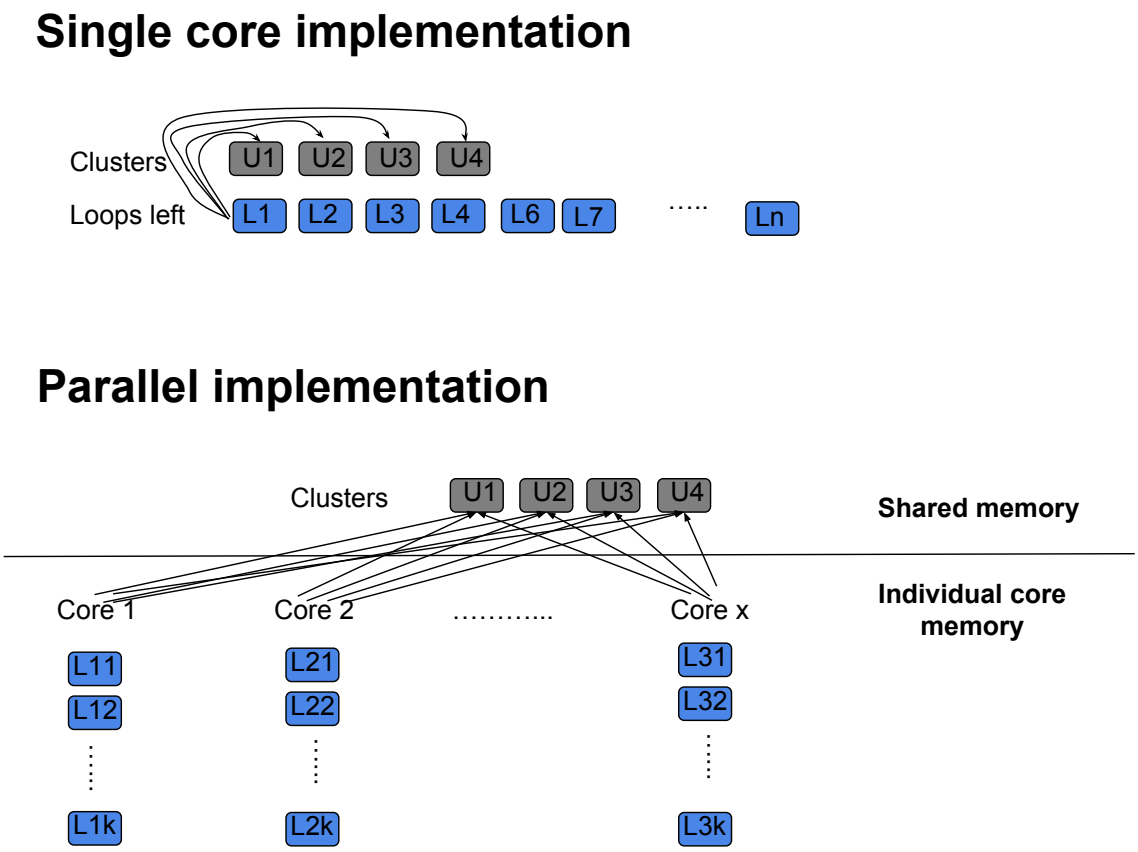

Figure A.1: Difference between the parallel implementation and the single-core implementation of the clustering algorithm. An individual fragment is represented by a blue rectangle, and a cluster is represented by a grey rectangle. In the parallel implementation the input is split equally in chunks of length k among x cores, with k=n/x

results in a potential over-estimation of the number of unique fragments, but it is not an issue for our purpose, because the goal is to reduce the list of $12*10^6$ fragments to a lower number.

Superposition is not transitive and by using a 1.0Å cut-off for clustering we might not capture some of the unique shapes. Therefore, when clustering the non-Ig fragments we chose a stricter uniqueness cut-off of 0.7Å. This is expected to result in a possible overestimation of the number of non-Ig shapes,

but also reduces the possibility of generating false positives when performing our intended classification of H3 fragments.

A.2 Cβ thresholds tables

Residue combination	Cut-off (Å)	Residue combination	Cut-off (Å)	Residue combination	Cut-off (Å)
ALA - ALA	3.5	ALA - ARG	3.75	ALA - ASN	3.75
ALA - ASP	3.5	ALA - CYS	3.5	ALA - GLN	3.75
ALA - GLU	3.75	ALA - GLY	3.5	ALA - HIS	3.75
ALA - ILE	3.75	ALA - LEU	3.75	ALA - LYS	3.75
ALA - MET	3.75	ALA - PHE	3.5	ALA - PRO	3.5
ALA - SER	3.5	ALA - THR	3.75	ALA - TRP	3.5
ALA - TYR	3.5	ALA - VAL	3.75	ARG - ARG	4.0
ARG - ASN	4.0	ARG - ASP	3.75	ARG - CYS	4.0
ARG - GLN	4.0	ARG - GLU	4.0	ARG - GLY	3.75
ARG - HIS	3.75	ARG - ILE	4.25	ARG - LEU	4.0
ARG - LYS	4.0	ARG - MET	4.0	ARG - PHE	3.75
ARG - PRO	3.75	ARG - SER	3.75	ARG - THR	4.0
ARG - TRP	3.75	ARG - TYR	3.75	ARG - VAL	4.0
ASN - ASN	3.75	ASN - ASP	3.75	ASN - CYS	3.75
ASN - GLN	3.75	ASN - GLU	3.75	ASN - GLY	3.75
ASN - HIS	3.75	ASN - ILE	4.0	ASN - LEU	3.75
ASN - LYS	4.0	ASN - MET	3.75	ASN - PHE	3.75
ASN - PRO	3.75	ASN - SER	3.75	ASN - THR	4.0
ASN - TRP	4.0	ASN - TYR	3.75	ASN - VAL	4.0
ASP - ASP	3.75	ASP - CYS	3.75	ASP - GLN	4.0
ASP - GLU	3.75	ASP - GLY	3.75	ASP - HIS	3.75
ASP - ILE	4.0	ASP - LEU	3.75	ASP - LYS	3.75
ASP - MET	4.0	ASP - PHE	3.75	ASP - PRO	3.75
ASP - SER	3.75	ASP - THR	3.75	ASP - TRP	4.0
ASP - TYR	3.75	ASP - VAL	3.75	CYS - CYS	3.25
CYS - GLN	4.0	CYS - GLU	3.75	CYS - GLY	3.75
CYS - HIS	3.75	CYS - ILE	4.0	CYS - LEU	3.75
CYS - LYS	4.0	CYS - MET	3.75	CYS - PHE	3.75
CYS - PRO	3.75	CYS - SER	3.5	CYS - THR	3.75
CYS - TRP	3.75	CYS - TYR	3.75	CYS - VAL	4.0
GLN - GLN	4.0	GLN - GLU	3.75	GLN - GLY	4.0
GLN - HIS	4.0	GLN - ILE	4.25	GLN - LEU	4.0
GLN - LYS	4.0	GLN - MET	4.0	GLN - PHE	3.75
GLN - PRO	3.75	GLN - SER	3.75	GLN - THR	3.75
GLN - TRP	3.75	GLN - TYR	3.75	GLN - VAL	4.0
GLU - GLU	3.75	GLU - GLY	3.75	GLU - HIS	3.75
GLU - ILE	4.0	GLU - LEU	4.0	GLU - LYS	4.0
GLU - MET	4.0	GLU - PHE	3.75	GLU - PRO	3.75
GLU - SER	3.75	GLU - THR	4.0	GLU - TRP	3.75
GLU - TYR	4.0	GLU - VAL	4.0	GLY - GLY	3.5
GLY - HIS	3.75	GLY - ILE	4.0	GLY - LEU	3.75
GLY - LYS	3.75	GLY - MET	3.75	GLY - PHE	3.75
GLY - PRO	3.75	GLY - SER	3.75	GLY - THR	3.75
GLY - TRP	3.75	GLY - TYR	3.75	GLY - VAL	3.75
HIS - HIS	3.75	HIS - ILE	4.0	HIS - LEU	3.75
HIS - LYS	3.75	HIS - MET	3.75	HIS - PHE	3.75
HIS - PRO	3.75	HIS - SER	3.75	HIS - THR	4.0
HIS - TRP	3.75	HIS - TYR	3.75	HIS - VAL	4.0
ILE - ILE	4.25	ILE - LEU	4.0	ILE - LYS	4.25
ILE - MET	4.0	ILE - PHE	4.0	ILE - PRO	4.0
ILE - SER	3.75	ILE - THR	4.25	ILE - TRP	3.75
ILE - TYR	4.0	ILE - VAL	4.25	LEU - LEU	4.0
LEU - LYS	4.0	LEU - MET	4.0	LEU - PHE	3.75
LEU - PRO	3.75	LEU - SER	3.75	LEU - THR	4.0
LEU - TRP	4.0	LEU - TYR	3.75	LEU - VAL	4.0
LYS - LYS	3.75	LYS - MET	4.0	LYS - PHE	4.0
LYS - PRO	3.75	LYS - SER	3.75	LYS - THR	4.0
LYS - TRP	3.75	LYS - TYR	3.75	LYS - VAL	4.0
MET - MET	4.0	MET - PHE	3.75	MET - PRO	3.75
MET - SER	3.5	MET - THR	3.75	MET - TRP	4.0
MET - TYR	3.75	MET - VAL	4.0	PHE - PHE	3.75
PHE - PRO	3.75	PHE - SER	3.75	PHE - THR	4.0
PHE - TRP	3.75	PHE - TYR	3.75	PHE - VAL	4.0
PRO - PRO	3.5	PRO - SER	3.75	PRO - THR	3.75
PRO - TRP	3.75	PRO - TYR	3.75	PRO - VAL	4.0
SER - SER	3.5	SER - THR	3.75	SER - TRP	3.75
SER - TYR	3.75	SER - VAL	3.75	THR - THR	4.0
THR - TRP	3.75	THR - TYR	4.0	THR - VAL	4.25
TRP - TRP	3.75	TRP - TYR	3.75	TRP - VAL	3.75
TYR - TYR	3.75	TYR - VAL	4.0	VAL - VAL	4.25

Table A.1: List of intra-protein Cβ thresholds for each amino acid combination

Residue combination	Cut-off (Å)	Residue combination	Cut-off (Å)	Residue combination	Cut-off (Å)
ALA - ALA	3.5	ALA - ARG	3.5	ALA - ASN	3.25
ALA - ASP	3.5	ALA - CYS	4.0	ALA - GLN	3.75
ALA - GLU	3.75	ALA - GLY	3.5	ALA - HIS	3.75
ALA - ILE	4.0	ALA - LEU	3.75	ALA - LYS	3.5
ALA - MET	3.75	ALA - PHE	3.5	ALA - PRO	3.75
ALA - SER	3.25	ALA - THR	3.75	ALA - TRP	4.25
ALA - TYR	3.25	ALA - VAL	4.0	ARG - ARG	4.5
ARG - ASN	3.5	ARG - ASP	3.75	ARG - CYS	4.75
ARG - GLN	4.0	ARG - GLU	4.0	ARG - GLY	3.5
ARG - HIS	3.75	ARG - ILE	4.25	ARG - LEU	3.75
ARG - LYS	4.0	ARG - MET	3.25	ARG - PHE	3.5
ARG - PRO	3.5	ARG - SER	4.0	ARG - THR	3.25
ARG - TRP	4.0	ARG - TYR	3.75	ARG - VAL	4.75
ASN - ASN	4.25	ASN - ASP	4.0	ASN - CYS	3.75
ASN - GLN	4.0	ASN - GLU	4.0	ASN - GLY	3.75
ASN - HIS	3.75	ASN - ILE	4.25	ASN - LEU	3.75
ASN - LYS	4.0	ASN - MET	3.5	ASN - PHE	4.25
ASN - PRO	3.75	ASN - SER	3.75	ASN - THR	4.0
ASN - TRP	3.75	ASN - TYR	3.5	ASN - VAL	4.0
ASP - ASP	4.75	ASP - CYS	4.0	ASP - GLN	4.25
ASP - GLU	4.0	ASP - GLY	4.0	ASP - HIS	4.75
ASP - ILE	4.25	ASP - LEU	3.75	ASP - LYS	4.0
ASP - MET	3.5	ASP - PHE	4.0	ASP - PRO	3.75
ASP - SER	3.5	ASP - THR	3.75	ASP - TRP	5.0
ASP - TYR	3.5	ASP - VAL	4.0	CYS - CYS	3.75
CYS - GLN	6.0	CYS - GLU	5.25	CYS - GLY	3.75
CYS - HIS	4.5	CYS - ILE	3.75	CYS - LEU	3.75
CYS - LYS	5.5	CYS - MET	5.0	CYS - PHE	3.5
CYS - PRO	4.5	CYS - SER	4.0	CYS - THR	5.0
CYS - TRP	4.25	CYS - TYR	4.75	CYS - VAL	4.75
GLN - GLN	4.25	GLN - GLU	3.25	GLN - GLY	4.25
GLN - HIS	4.0	GLN - ILE	4.5	GLN - LEU	4.25
GLN - LYS	3.5	GLN - MET	5.25	GLN - PHE	3.5
GLN - PRO	3.75	GLN - SER	3.75	GLN - THR	3.5
GLN - TRP	4.25	GLN - TYR	3.5	GLN - VAL	4.5
GLU - GLU	4.75	GLU - GLY	3.75	GLU - HIS	3.0
GLU - ILE	4.0	GLU - LEU	3.75	GLU - LYS	4.0
GLU - MET	4.5	GLU - PHE	4.5	GLU - PRO	3.75
GLU - SER	3.75	GLU - THR	4.0	GLU - TRP	4.75
GLU - TYR	3.75	GLU - VAL	4.5	GLY - GLY	3.5
GLY - HIS	3.75	GLY - ILE	3.75	GLY - LEU	3.75
GLY - LYS	3.75	GLY - MET	3.75	GLY - PHE	3.25
GLY - PRO	3.75	GLY - SER	3.5	GLY - THR	4.0
GLY - TRP	4.25	GLY - TYR	3.5	GLY - VAL	3.75
HIS - HIS	5.0	HIS - ILE	4.0	HIS - LEU	4.25
HIS - LYS	3.75	HIS - MET	4.0	HIS - PHE	3.75
HIS - PRO	3.5	HIS - SER	3.5	HIS - THR	4.0
HIS - TRP	5.25	HIS - TYR	3.75	HIS - VAL	4.75
ILE - ILE	4.5	ILE - LEU	4.25	ILE - LYS	4.0
ILE - MET	4.25	ILE - PHE	3.75	ILE - PRO	4.25
ILE - SER	3.75	ILE - THR	4.0	ILE - TRP	4.25
ILE - TYR	4.25	ILE - VAL	4.5	LEU - LEU	4.25
LEU - LYS	3.75	LEU - MET	3.75	LEU - PHE	4.0
LEU - PRO	4.0	LEU - SER	4.0	LEU - THR	4.25
LEU - TRP	3.75	LEU - TYR	4.25	LEU - VAL	4.25
LYS - LYS	5.5	LYS - MET	4.0	LYS - PHE	4.5
LYS - PRO	4.0	LYS - SER	3.75	LYS - THR	3.5
LYS - TRP	5.0	LYS - TYR	3.75	LYS - VAL	3.5
MET - MET	5.75	MET - PHE	3.75	MET - PRO	3.5
MET - SER	3.75	MET - THR	4.25	MET - TRP	5.0
MET - TYR	4.0	MET - VAL	4.25	PHE - PHE	4.75
PHE - PRO	4.25	PHE - SER	3.5	PHE - THR	4.5
PHE - TRP	3.5	PHE - TYR	4.0	PHE - VAL	4.75
PRO - PRO	3.5	PRO - SER	3.5	PRO - THR	3.75
PRO - TRP	3.75	PRO - TYR	3.75	PRO - VAL	4.5
SER - SER	4.25	SER - THR	3.75	SER - TRP	4.0
SER - TYR	3.75	SER - VAL	4.25	THR - THR	4.0
THR - TRP	3.75	THR - TYR	4.0	THR - VAL	4.25
TRP - TRP	4.5	TRP - TYR	3.75	TRP - VAL	4.25
TYR - TYR	4.25	TYR - VAL	4.0	VAL - VAL	3.75

Table A.2: List of inter-protein Cβ thresholds for each amino acid combination

A.3 Extended Target database algorithm

- 1. The UNIPROT ID was used to retrieve the GENE ID of the protein using the API query "http://www.uniprot.org/uniprot/?query=accession:{UNIPROT ID}&columns=genes&format=tab"

- 2. The GENE ID is then used to retrieve the list of UNIPROT IDs of the protein in other species using the API query http://www.uniprot.org/uniprot/?query=gene:{GENE ID}&format=tab"

A.4 Initial IL-5 designs

A.4.1 Design 1

Heavy Chain

EVKLLESGGGLAQPGGSLKLSCAASGFDFRRYWMTWVRQAPGKGLEWIGEINPD
SRTINYMPSLKDKFIISRDNAKNSLYLQLSRLRSEDSALYYCVRLDFDVYNHYY
VLDYWGQGTSVTVSS

Area	Sequence
FK1	EVKLLESGGGLAQPGGSLKLSCAAS
CDR H1	GFDFRRY
FK2	WMTWVRQAPGKGLEWIGEI
CDR H2	NPDSRT
CFK3	NYMPSLKDKFIISRDNAKNSLYLQLSRLRSEDSALYYCVR
CDR H3	LDFDVYNHYYVLDY
FK4	WGQGTSVTVSS

Table A.3: Design 1 - Heavy chain

Light Chain

VVTQESALTTSPGETVTLTCEKPVSAFPIHCFDYEWVQEKPDHLFTGLIGGTNK

RAPGVPARFSGSLIGDRAALTITGAQTEDEAIYFCRAAVSSMCREAGLWSEFGG
GTKLE

Area	Sequence
FK1	VVTQESALTTSPGETVTLTC
CDR L1	EKPVSAFPIHCFDYE
FK2	WVQEKPDHLFTGLIGG
CDR L2	TNKRAP
CFK3	GVPARFSGSLIGDRAALTITGAQTEDEAIYFC
CDR L3	RAAVSSMCREAGLWSE
FK4	FGGGTKLE

Table A.4: Design 1 - Light chain

A.4.2 Design 2

Heavy Chain

VQLQESGPGLVKPSDTLSLTCAVSGYSITGGYSWHWIRQPPGKGLEWMGYIHYS
GYTDFNPSLKTRITISRDTSKNQFSLKLSSVTAVDTAVYYCARKDPSDAFPYWG
QGTLVTVSS

Area	Sequence
FK1	VQLQESGPGLVKPSDTLSLTCAVS
CDR H1	GYSITGGY
FK2	SWHWIRQPPGKGLEWMGYI
CDR H2	HYSGY
CFK3	TDFNPSLKTRITISRDTSKNQFSLKLSSVTAVDTAVYYCAR
CDR H3	KDPSDAFPY
FK4	WGQGTLVTVSS

Table A.5: Design 2 - Heavy chain

Light Chain

IVLTQSPDFQSVTPKEKVTITCRASQSISDHLHWYQQKPDQSPKLLIKYASHAI
SGVPSRFSGSGSGTDFTLTINSLEAEDAATYYCRAAVSSMCREAGLWSEFGGGT
KVEIK

Area	Sequence
FK1	IVLTQSPDFQSVTPKEKVTITC
CDR L1	RASQSISDHLH
FK2	WYQQKPDQSPKLLIK
CDR L2	YASHAIS
FK3	GVPSRFSGSGSGTDFTLTINSLEAEDAATYYC
CDR L3	RAAVSSMCREAGLWSE
FK4	FGGGTKVEIK

Table A.6: Design 2 - Light chain

A.4.3 Design 3

Heavy Chain

EVQLVESGGGLVKAGGSLILSCGVSFPRTFILSKGRDWLTMNWVRRVPGGGLEW
VASIYPGSRDYADAVKGRFTVSRDDLEDFVYLQMHKMRVEDTAIYYCARLYSLP
VYWGPGTVVTVS

Area	Sequence
FK1	EVQLVESGGGLVKAGGSLILSCGVS
CDR H1	FPRTFILSKGRDWL
FK2	TMNWVRRVPGGGLEWVASI
CDR H2	YPGS
CFK3	RDYADAVKGRFTVSRDDLEDFVYLQMHKMRVEDTAIYYCAR
CDR H3	LYSLPVY
FK4	WGPGTVVTVS

Table A.7: Design 3 - Heavy chain

Light Chain

VVMTQSPSTLSASVGDTITITCRASQSIETWLAWYQQKPGKAPKLLIYKASTLK
TGVPSRFSGSGSGTEFTLTISGLQFDDFATYHCQHYAGYSATFGQGTRVEIK

Area	Sequence
FK1	VMTQSPSTLSASVGDTITITC
CDR L1	RASQSIETWLA
FK2	WYQQKPGKAPKLLIY
CDR L2	KASTLKT
FK3	GVPSRFSGSGSGTEFTLTISGLQFDDFATYHC
CDR L3	QHYAGYSAT
FK4	FGQGTRVEIK

Table A.8: Design 3 - Light chain

A.4.4 Design 4

Heavy Chain

VQLVQSGAEVKRPGSSVTVSCKASFPRTFILSKGRDWLAVALSWVRQAPGRGLE
WMGGVIPLLTITNYAPRFQGRITITADRSTSTAYLELNSLRPEDTAVYYCARKL
GWFPYWGQGTLVTVSS

Area	Sequence
FK1	VQLVQSGAEVKRPGSSVTVSCKAS
CDR H1	FPRTFILSKGRDWLAV
FK2	ALSWVRQAPGRGLEWMGGV
CDR H2	IPLLTI
CFK3	TNYAPRFQGRITITADRSTSTAYLELNSLRPEDTAVYYCAR
CDR H3	KLGWFPY
FK4	WGQGTLVTVSS

Table A.9: Design 4 - Heavy chain

Light Chain

EIVLTQSPGTQSLSPGERATLSCRASQSVGNNKLAWYQQRPGQAPRLLIYGASS

RPSGVADRFSGSGSGTDFTLTISRLEPEDFAVYYCQQYGQSLSTFGQGTKVEVK

Area	Sequence
FK1	EIVLTQSPGTQSLSPGERATLSC
CDR L1	RASQSVGNNKLA
FK2	WYQQRPGQAPRLLIY
CDR L2	GASSRPS
FK3	GVADRFSGSGSGTDFTLTISRLEPEDFAVYYC
CDR L3	QQYGQSLST
FK4	FGQGTKVEVK

Table A.10: Design 4 - Light chain

A.4.5 Design 5

Heavy Chain

QVQLVQSGGQMKKPGESMRISCRASGYEFIDCTLNWIRLAPGKRPEWMGWLKPR

GGAVNYARPLQGRVTMTRDVYSDTAFLELRSLTVDDTAVYFCTRAVSSMCREAG

LWSEWGRGTPVIVSS

Area	Sequence
FK1	QVQLVQSGGQMKKPGESMRISCRAS
CDR H1	GYEFIDC
FK2	TLNWIRLAPGKRPEWMGWL
CDR H2	KPRGGA
CFK3	VNYARPLQGRVTMTRDVYSDTAFLELRSLTVDDTAVYFCTR
CDR H3	AVSSMCREAGLWSE
FK4	WGRGTPVIVSS

Table A.11: Design 5 - Heavy chain

Light Chain

EIVLTQSPGTLSLSPGETAIISCRTSQYGSLAWYQQRPGQAPRLVIYSGSTRAA

GIPDRFSGSRWGPDYTLTISNLESGDFGVYYCQQYEFFGQGTKVQV

Area	Sequence
FK1	EIVLTQSPGTLSLSPGETAIISC
CDR L1	RTSQYGSLA
FK2	WYQQRPGQAPRLVIY
CDR L2	SGSTRAA
FK3	GIPDRFSGSRWGPDYTLTISNLESGDFGVYYC
CDR L3	QQYEF
FK4	FGQGTKVQV

Table A.12: Design 5 - Light chain