UNIVERSITY COLLEGE LONDON

Antibody Therapeutic Prediction and Design: Immunogenicity and Stability.

by

Thomas Charles Northey

A thesis submitted to University College London
for the degree of Doctor of Philosophy

in the
Faculty of Life Sciences
Department of Structural and Molecular Biology

January 2017

Declaration of Authorship

I, Thomas Charles Northey, declare that this thesis titled, 'ANTIBODY THERAPEUTIC PREDICTION AND DESIGN: IMMUNOGENICITY AND STABILITY' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: ____

Abstract

One third of drugs currently in development are monoclonal antibody therapeutics. Key to the efficacy and safety of therapeutic antibodies is the avoidance of immunogenic reaction. Immunogenicity describes a patient's immune response to an antibody therapeutic leading to poor efficacy and allergic reaction that can vary in severity. Immunogenicity is influenced by the presence of T- and B-cell epitopes and by antibody stability. The aim of this thesis was to develop *in silico* methods to aid the selection and design of antibody therapeutics by focusing on the prediction B-cell epitopes and biophysical stability.

IntPred, a general protein-protein interface predictor, was applied as a B-cell epitope predictor. Tested on a set of antigen structures, IntPred was unable to predict B-cell epitopes. Thus, IntPred was amended to create the IntPred:Epi method. IntPred:Epi was able to outperform all methods tested, except SEPPA 2.0. However, further testing on a larger data set showed IntPred:Epi to outperform SEPPA 2.0.

Next, the influence of 'tolerated surfaces' on the selection of epitopes was considered. Libraries of surfaces were created from human and mouse PDB structures. From these, descriptors were generated to describe the tolerance state of antigen surfaces. These were then applied as a label in the B cell epitope prediction problem. Using tolerance labels as a filter on IntPred:Epi predictions, performance was improved on a test set of human antibody-bound antigens.

The efficacy and immunogenicity of a therapeutic antibody is also affected by its stability. Natural $\mathrm{V_H}$-$\mathrm{V_L}$ pair human Fabs were expressed and sequenced, and hydrophobicity and thermal stability data were generated. These data were then used to investigate the sequence determinants of the biophysical properties of antibodies. An exploratory analysis revealed correlations between sequence features and biophysical properties that can be applied in the future for the prediction of biophysical stability.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Andrew Martin for all his time, guidance, patience and support throughout the project — it is greatly appreciated.

This thesis was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) and UCB. I would like to thank both organisations for their support and in particular UCB for taking the time to organise regular meetings for all of their PhD students, therefore giving me an opportunity to share and discuss my work with people from universities across the country.

I would like to thank my thesis committee members Dr. Adrian Shepherd and Prof. Christine Orengo for their feedback and support throughout the project.

I want to express my gratitude to everyone who helped me at UCB, both throughout the project and during my time there. I would like to thank my industry supervisor Andy Popplewell for his help throughout the years. I especially need to thank Kerry Tyson for her patience and guidance, as well as Sarfaraj Topia for all his help. I'd also like to thank James Heads, who was kind enough to carry out a number of experiments for me.

I would like to thank all of the members of the Martin group, past and present. Thanks especially go to Francesco and Saba for keeping the Martin bay lively! I'd also like to thank everyone up in Room 636, in particular Sayoni, Su and Ivana.

I want to thank all my friends and family for all of their love and support, especially Bethel and all her esteemed associates (you know who you are). Finally, I want to thank Rosie for all her support during the last four years — I couldn't have done it without you.

Contents

List of Figures

List of Tables

Abbreviations

ASA	**A**ccessible **S**urface **A**rea
AUC	**A**rea **U**nder (receiver-operating characteristic) **C**urve
BCE	**B** **C**ell **E**eptiope
BCR	**B** **C**ell **R**eceptor
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
CDR	**C**omplementarity-**D**etermining **R**egion
CV	**C**ross-**V**alidation
Fab	**F**ragment **a**ntigen-**b**inding
FDR	**F**alse **D**iscovery **R**ate
FN	**F**alse **N**egative
FOSTA	**F**unctional **O**rthologues (from) **S**wissProt **T**ext **A**nalysis
FP	**F**alse **P**ositive
FPR	**F**alse **P**ositive **R**ate
GO	**G**ene **O**ntology
HIC	**H**ydrophobic **I**nteraction **C**hromatography
IEDB	**I**mmune **E**pitope **D**ata**B**ase
MCC	**M**atthews **C**orrelation **C**oefficient
MDS	**M**ulti-**D**imenional **S**caling
MSA	**M**ultiple **S**equence **A**lignment
PCA	**P**rincipal **C**omponent **A**nalysis
PDB	**P**rotein **D**ata **B**ank
PPI	**P**rotein-**P**rotein **I**nterface
PPV	**P**ositive **P**redictive **V**alue
rASA	**r**elative **A**ccessible **S**urface **A**rea
Sens.	Sensitivity

scFv	**s**ingle-**c**hain Fv
Spec.	**S**pecificity
SVM	**S**upport **V**ector **M**achine
TCE	**T** **C**ell **E**epitope
TCR	**T** **C**ell **R**eceptor
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
TSL	**T**olerated **S**urface **L**ibrary
UniProt	**Uni**versal **Prot**ein resource

Chapter 1

Introduction

In 1986, the first therapeutic monoclonal antibody was approved by the U.S Food and Drug Administration (FDA). Since then, therapeutic antibodies have become a corner stone of modern drug development. By end of 2014, 47 monoclonal antibodies had been approved in the U.S. or Europe, generating global revenue sales of $75 billion in 2013 (Ecker et al., 2015). As of September 2016, this has already climbed to 65 monoclonal antibodies[1].

Therapeutic antibodies are a part of a larger class of drugs known as biological therapeutics. A biological therapeutic is derived either completely or partly from living material. In comparison to small molecular weight drugs which can be thoroughly characterised, biological therapeutics tend to be larger in size and more complex in structure and are therefore harder to characterise. This complexity leads to difficulties in production, cost, pharmokinetics and safety that are unlike those faced with small-molecule drugs (Samaranayake et al., 2009). Therapeutic antibodies are the largest subclass of biological therapeutics and are the focus of this thesis.

Therapeutic antibodies offer the advantage of exquisite specificity to a broad range of targets. This has lead to their application in the treatment of a diverse range of conditions, including cancer, autoimmune diseases such as rheumatoid arthritis, viral infection and the prevention of transplant rejection. Despite this, companies face many challenges when developing and producing therapeutic antibodies. In particular, the failure of therapeutic antibodies in late-stage clinical trials due to the development of adverse immune responses is extremely costly (Ritter et al., 2001). The development of *in silico* methods to predict such adverse immune response at the early stages of drug development would help avoid such cases. This thesis focuses on the development of methods for the prediction of two important properties: immunogenicity and biophysical stability.

[1] Source available to members at `http://www.antibodysociety.org/`

This chapter begins with an introduction to the structure, function and application of antibodies. Due to their importance for the aims of the thesis, an introduction to the function, characterisation and prediction of epitopes then follows. The chapter concludes with the aims of the thesis.

1.1 Antibodies

One of the principle molecular components of the immune system, antibodies are crucial for detecting foreign material (antigen) in the body. Antibodies exhibit high specificity and affinity for their target antigen, properties that allow them to recognize the huge variety of pathogens that the human body encounters. Antibodies are produced by highly specialized B Lymphocyte Cells (B cells). In B cells, the immunoglobulin genes responsible for the production of antibodies undergo a remarkable reordering process that produces a huge reservoir of possible variants (see section 1.1.6). During the early stages of infection, B cells producing antibodies with a degree of specificity for the antigen are selected and undergo a process of somatic hypermutation that improves antibody affinity and maintains specificity (Rajewsky, 1996). These antibodies can be either attached to the B cell surface membrane in the form of a B cell Receptor (BCR), or secreted in soluble form. Upon binding to their target antigen, antibodies have the potential to neutralize the biological activity of its target antigen, or act as a flag for eliciting an immune response. Due to their remarkable specificity, antibodies have been the basis for a wide range of technologies, as well therapeutic applications.

1.1.1 Early discoveries

The study of antibodies began in the late 19th Century, when Emil von Behring and Shibasaburo Kitasato noted that animals injected with the soluble toxin of *Clostridium tetani* produced a *specific* neutralizing “antitoxin”. Behring also came to the same conclusion in regards to *Corynebacterium diphtheria.* This resulted in Kitasato proposing the theory of humoural immunity — that the serum contains an agent capable of reacting with foreign material (antigen). Work by Michael Heidelberger and Oswald Avery proved that this agent in the blood (by then termed antibody) was made of protein and was able to precipitate antigen (Van Epps and Heidelberger, 2006). In the 1940s, Linus Pauling characterised the antibody-antigen interaction following a ‘lock and key’ model, which highlighted the importance of complementarity between the shapes of antigen and antibody. In the same decade, Astrid Fagreaus identified B cells as the cellular origin of antibodies (Fagreaus, 1948). By the 1960s, scientists were starting to define details of antibody structure. Gerald Edelman discovered that antibodies consisted of light and

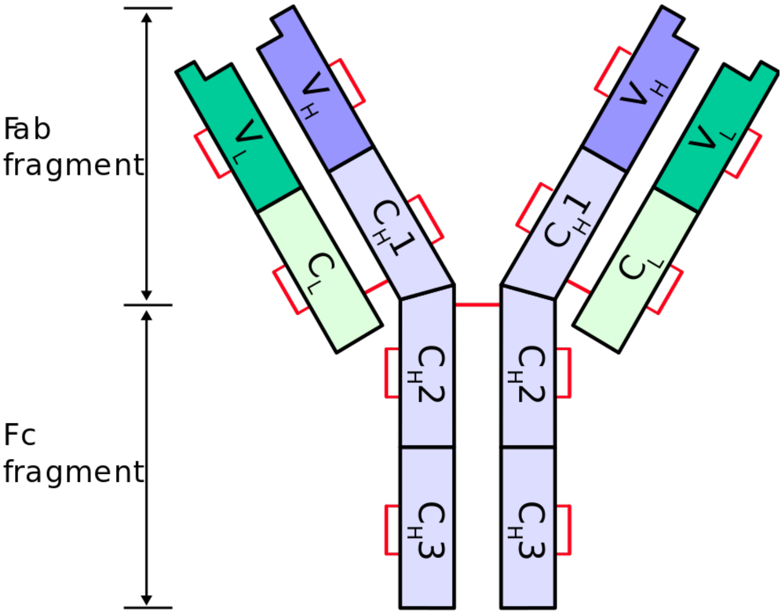

FIGURE 1.1: Antibody topology. Heavy-chain domains are shown in purple and light-chain domains are shown in green. Red bars indicates disulfide bonds. Fragment definitions are shown on the left. Image obtained under a Creative Commons Attribute-Noncommerical license, available at `https://commons.wikimedia.org/wiki/File:AntibodyChains.svg`

heavy chains linked together by disulfide bonds, while Rodney Porter helped to characterise the constituent Fab and Fc fragments. Together, Edelman and Porter went on to win the Nobel Prize in Physiology or Medicine in 1972 for their discoveries in elucidating the structure and complete amino acid sequence of IgG (Raju et al., 1999). In the same decade Hochman et al. (1973) characterised the Fv fragment, while (Hozumi and Tonegawa, 1976) showed that immunoglobulin genes undergo a somatic rearrangement that is partly responsible for the diverse range of antibodies observed.

1.1.2 Structure

An antibody consists of two identical heavy (H) chains and two identical light (L) chains (see figure 1.1). These chains form a Y-shaped structure held together with covalent disulfide bonds and non-covalent interactions. The light chains contain one variable (V_L) domain and one constant (CL) domain, while heavy chains contain one variable domain (V_H) and three or four constant domains (CH) depending on the antibody class. Treatment of antibody with papain cleaves both heavy chains in the flexible hinge region between C_H1 and CH2 domains (Porter, 1959), yielding two Fab fragments — each containing a light chain, V_H domain and C_H1 domain — and an Fc fragment containing the remaining heavy chain domains. Treatment with pepsin also cleaves in the flexible hinge

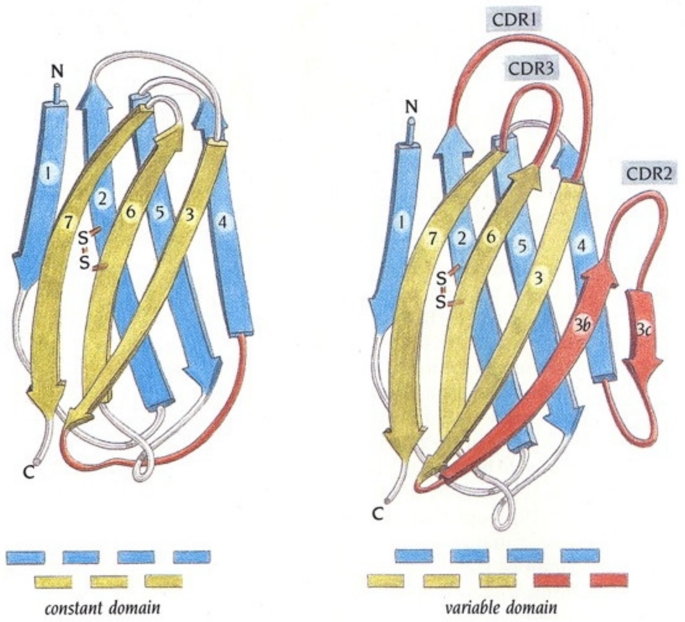

FIGURE 1.2: Variable and constant domain topologies. Both variable and constant domains have a β-sandwich fold formed from two anti-parallel β-sheets; both have a four-strand sheet (blue) paired with a three-strand sheet in the constant domain (yellow) or a five-strand sheet in the variable domain (yellow and red). Image taken from Branden and Tooze (1991).

region (Rossi et al., 1969) but below the two inter-chain disulfide bonds, resulting in one $\mathrm{F(ab')_2}$ fragment.

Variable and constant domains are homologous and share sequence and structural similarity; both consist of two anti-parallel β-sheets that form a β-sandwich (see figure 1.2). Constant domain β-sheets are three and four-stranded, while variable regions have two additional short strands forming four and five-stranded sheets. A disulfide bond between the two antiparallel sheets of the β-sandwich stabilizes the structure (Zouali, 2001). Wu and Kabat (1970) were the first to identify the hyper-variable sequence regions within $\mathrm{V_L}$ and $\mathrm{V_H}$ domains. These sequences form the complementary-determining regions (CDRs) of the variable domains that are the primary contacts for the antigen (see figure 1.3).

The first x-ray structure of a Fab fragment showed that the domains shared the same basic folding patterns and that the CDR sequences corresponded approximately to structural loops found in close spatial proximity to each other (Poljak et al., 1973). An x-ray structure of lysozyme in complex with a Fab fragment revealed the first antigen-antibody interface, confirming the role of CDRs in antigen interaction (Amit et al., 1985). Moreover, the structure showed the antibody-antigen interface to involve a larger area than the CDRs alone and that the correspondingly large surface on the antigen suggested a conformational, rather than sequential antigenic determinant (Amit et al., 1985). By

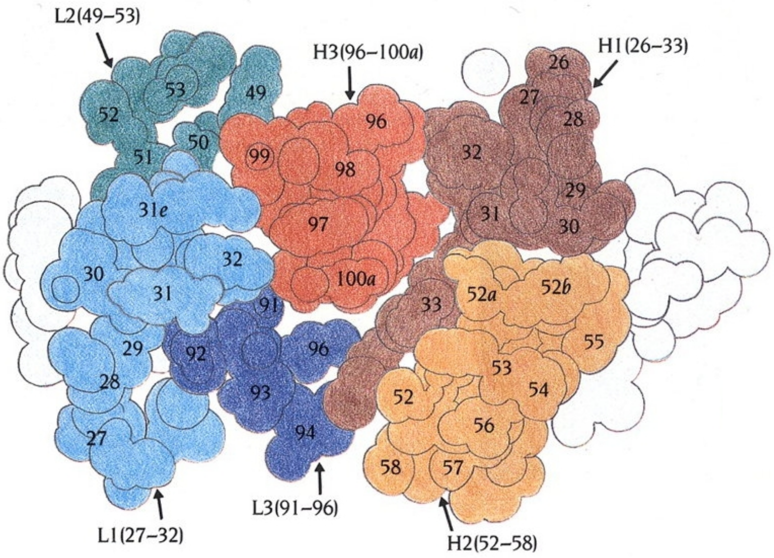

Figure 1.3: Antigen-binding site topology. Each colour represents the surface contributed by each CDR. H3 and L3 form the core of the site, with the remaining CDRs forming the periphery. H3 commonly has insertions that allow it to form more extensive contacts with the antigen. Image taken from Branden and Tooze (1991).

December 2015, over 2400 antibody-related structures had been deposited in the PDB, over 1180 of which are non-redundant.

1.1.3 Sequence position numbering schemes

A sequence position numbering scheme can be applied to a set of closely related proteins in order to simplify the description and comparison of members of the set. Antibodies are such a set. The first antibody numbering scheme to be introduced was the Kabat numbering scheme (Kabat et al., 1983). The Kabat numbering scheme is based on sequence alignment and is not based on any structural information. Importantly, it makes use of insertion codes (e.g. H100A) that can be used to accommodate the varying lengths of the CDR regions (though insertions are also found in the framework regions). Once antibody structures were available, Chothia and Lesk (1987) used structural comparison to define the Chothia numbering scheme, which is identical to the Kabat scheme, except that insertions in the CDRH1 and CDRL1 are numbered at structurally correct positions. The Chothia numbering scheme was extended by Abhinandan and Martin (2008) to create the enhanced-Chothia (Martin) scheme, which is identical to the chothia scheme but also places framework insertions correctly. Additionally, two numbering schemes are also available whose primary purpose is to unify numbering across antibody heavy and light chains and T cell receptor α and β chains (Honegger and Pluckthun, 2001, Lefranc et al., 2003). The first of these numbering schemes is not structurally correct, whilst the second is a modification of the first that is. However, these numbering schemes do not make use of insertion codes; instead, enough position numbers are provided to accommodate the expected range, based on current data. However, a lack of insertion codes means that these schemes would be problematic if sequences are found in the future with extreme-length insertions.

1.1.4 Function

Antibodies have a host of functions which mostly rely on their ability both to bind antigen at the antigen combining site and an effector receptor at the Fc region. The exception is their antibody neutralisation, whereby an antibody binds to an antigen in order to prevent it from performing its function. Otherwise, antibody function falls into two categories. The first is the activation of the complement system, one of the most important parts of the innate immune system. The classic complement pathway can be activated when C1q binds to the Fc region of IgG or IgM-class antibodies, the first step in a pathway of reactions that leads to formation the membrane attack complex — a transmembrane structure that causes osmotic lyis of target cells — as well as recruitment of macrophage cells for phagocytosis. The second major function of antibodies is the interaction of immune cells via FcR receptors, which bind to the Fc region. A diverse range of immune cells interact with antibodies via this mechanism and different processes are triggered according to the type of immune cell. Furthermore, different immune cells present different classes of FcR receptors on their surfaces, which, due to differences in the Fc regions of different antibody classes, determines which class of antibody can be bound. Once recruited by antibodies, immune cells can carry out a range of functions, including phagocytosis, degranulation and antibody-dependent cell-mediated cytotoxicity.

1.1.5 Development

In mammals, early B lymphocyte development occurs in the bone marrow. In these early stages, cell differentiation is accompanied by RAG1 and RAG2 gene expression. This expression is responsible for the re-arrangement of gene segments at the immunoglobulin heavy- and light-chain loci (see 1.1.6). After development in the bone marrow, B cells migrate to peripheral lymphoid tissues such as the lymph nodes, spleen and tonsils, where antigen may be encountered. Interaction with antigen activates B cells, leading to clonal proliferation and the formation of germinal centres in the peripheral lymphoid tissue. Following proliferation, a process of affinity maturation leads to the production of antibodies with increased affinity for antigen. Two coordinated processes lead to affinity mutation: somatic hypermutation and clonal selection.

The process of somatic hypermutation involves the variable regions of Ig genes undergoing extensive mutation after activation by antigen, resulting in sequences that differ from those of the original V(D)J recombinant (see 1.1.6). Rates of mutation have been estimated at 10^{-3} – 10^{-4} base pairs per cell generation, a rate approximately six orders of magnitude higher than the spontaneous mutation rate (Rajewsky et al., 1987). These mutations are almost always point mutations and in theory, mutation of any variable region residue can occur. In reality, the molecular spectrum of somatic hypermutation means that not all bases are equally likely to undergo mutation, with nucleotide type, strand type and sequence context all playing a role in determining mutation rate (Teng and Papavasiliou, 2007).

Clonal selection is the process whereby B cells are selected through affinity for their antigen. Follicular dendritic cells (FDCs) present antigen on their surface; through interaction with FDCs and T helper cells, those B cells with highest affinity for the antigen receive the strongest survival signals and go on to become plasma or memory B cells. Plasma cells travel from the germinal centres to the blood plasma and lymph systems and secrete large volumes of antibody. In contrast, memory B cells remain in the lymph node and persist over long time periods in order to generate a secondary immune response, which allows the immune system to react more quickly to re-infection.

Together, somatic hypermutation and clonal selection lead to the rapid development of antibodies that, through fine specificity for their target antigen, allow the identification and clearance of pathogens.

1.1.6 Genetics

The genes responsible for antibody heavy- and light-chains are found at the immunoglobulin heavy- and light-chain loci. The number of loci for each gene varies by species but all genes share a common form.

1.1.6.1 Heavy chain

At the heavy chain loci, V, D and J gene segments are recombined to form a full VDJ gene that encodes for the V_H domain (see figure 1.5). Multiple varying copies of V, D and J segments occur in sequential order at each locus. The removal of unwanted D and J segments occurs to allow DJ recombination before the removal of unwanted V and the remaining unwanted D segments to allow VDJ recombination. This process relies on the expression of RAG1 and RAG2 and the selection of V, D and J is random. Importantly, this recombination happens at the DNA level and results in the permanent loss of unused

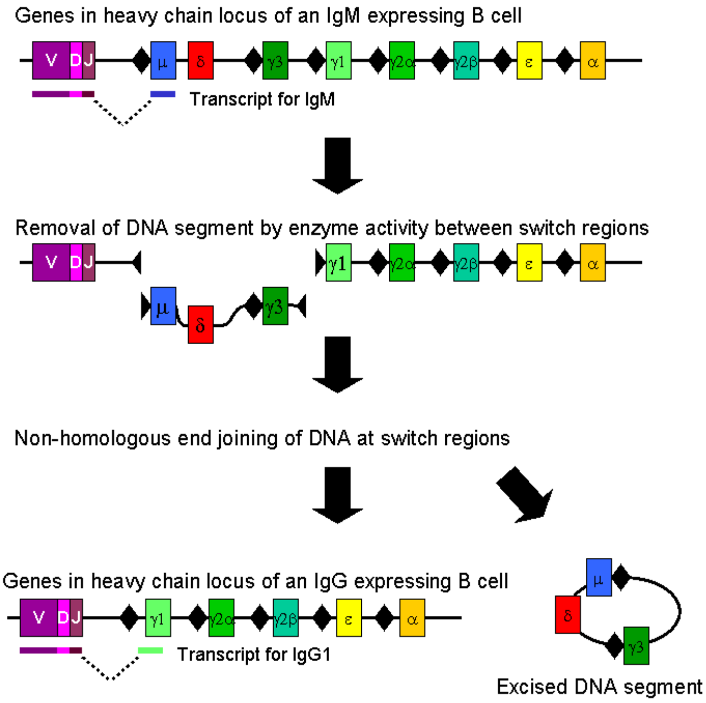
Genes in heavy chain locus of an IgM expressing B cell
V
DJ
μ
δ
γ3
γ1
γ2α
γ2β
ε
α
Transcript for IgM
Removal of DNA segment by enzyme activity between switch regions
Non-homologous end joining of DNA at switch regions
Genes in heavy chain locus of an IgG expressing B cell
Transcript for IgG1
Excised DNA segment

Figure 1.4: Class switch recombination to allow the expression of IgG1 antibodies. Note that there is no switch region between μ and δ, so transcription of loci occur followed by alternative splicing. Image obtained from `https://commons.wikimedia.org/wiki/File:Class_switch_recombination.png`

V, D and J gene segments; this ensures that each B cell can only produce a single VDJ recombinant.

Downstream of V, D and J segments are the constant region genes that code for the heavy chain constant regions. In humans, nine constant loci exist and they determine the class of antibody (class in parentheses); μ (IgM), δ (IgD), $\gamma1$ (IgG1), $\gamma2$ (IgG2), $\gamma3$ (IgG3), $\gamma4$ (IgG4), $\alpha1$ (IgA1), $\alpha2$ (IgA2) and ϵ (IgE). These loci are ordered as shown in figure 1.4. Initially, B cells transcribe an mRNA containing VDJ μ and δ sequences that can alternatively spliced to produce an IgM or IgD antibody. Later in B cell development, class switch recombination of the DNA can occur to bring a different heavy gene locus in proximity to the VDJ gene through recombination events at switch regions that proceed each constant genes except δ (see figure 1.4). Similarly to VDJ recombination, recombination events at switch regions occur at the DNA level and result in the permanent loss of intervening constant genes. Consequently, class switching is unidirectional and so, with the exception of IgM and IgG, B cells are not able to go from producing one class of antibody to producing an upstream class (e.g. IgE to IgD).

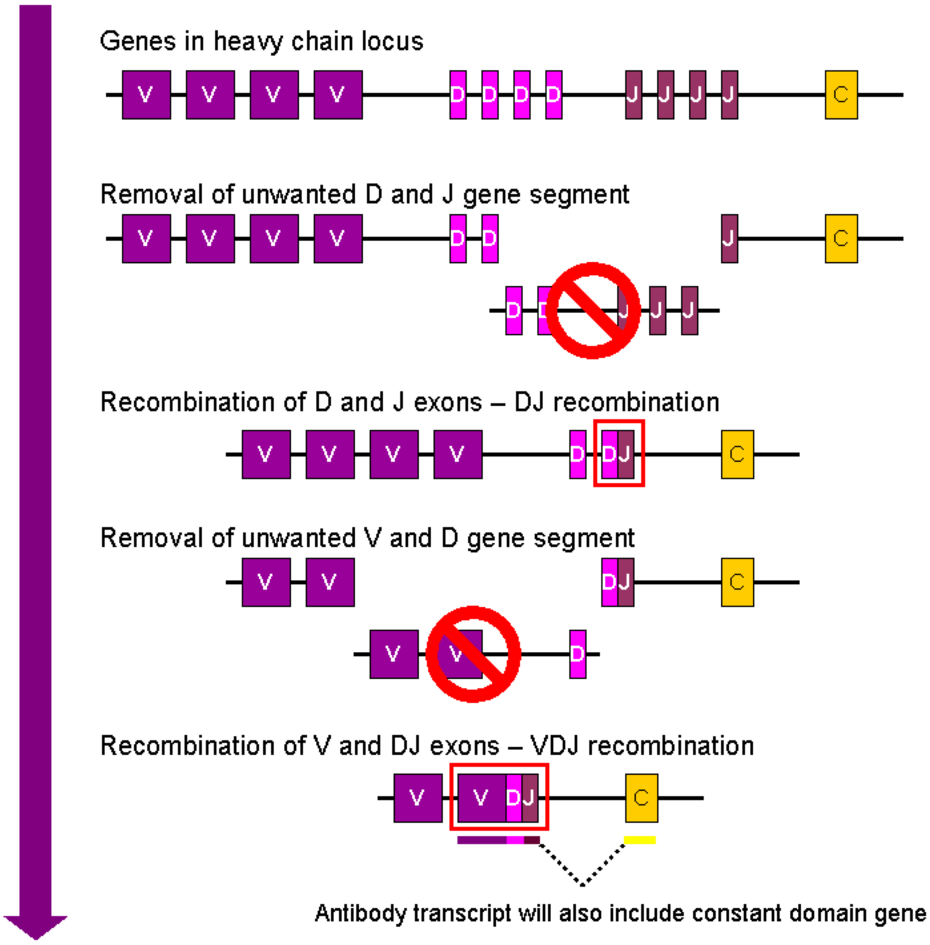
Genes in heavy chain locus
V
V
V
V
D
D
D
D
J
J
J
J
C
Removal of unwanted D and J gene segment
Recombination of D and J exons – DJ recombination
DJ
Removal of unwanted V and D gene segment
Recombination of V and DJ exons – VDJ recombination
Antibody transcript will also include constant domain gene

Figure 1.5: VDJ recombination at the heavy chain locus. DJ recombination occurs by removal of intervening D and J segments, before VDJ recombination through the removal intervening V and D segments. Recombination at the light chain locus follows the same process except that only V and J gene segments are present. Image obtained under a Creative Commons Attribute-Noncommerical license, available at `https://commons.wikimedia.org/wiki/File:VDJ_recombination.svg`

1.1.6.2 Light chain

In humans, two loci exist for the light chain: kappa (κ) and lambda (λ). At both loci, gene rearrangement similar to that at heavy chain loci occurs, the only difference being that κ and λ lack D segments. The lack of a D segment means that light chain sequences are less diverse than their heavy counterparts (see section 1.1.7). In contrast to the heavy chain, the constant domain of the light chain is joined to the VJ segment by splicing at the RNA level.

1.1.7 Diversity

Antibody diversity is a consequence of the huge number of possible V(D)J combinations. At the human heavy chain locus, 51 V, 6 J and 27 D segments exist (Alberts et al., 2002). This results in $51 \times 6 \times 27 = 8262$ different heavy-chain regions. Furthermore, 316 different $\mathrm{V_L}$ regions (200 κ and 116 λ) increase the combinatorial diversity of antibodies up to $316 \times 8262 \approx 2.6 \times 10^6$ (Alberts et al., 2002). A process called junctional diversification — whereby random loss and gain of nucleotides occurs the gene segment recombination site between the D and J segments — is estimated to increase the number of possibilities to approximately 10^{12}. There is also evidence that the D gene segment can be read in all six reading frames, adding more diversity (Darsley and Rees, 1985). Antibodies then see further diversification through somatic hypermutation (see section 1.1.5). Crucially, the sequence position for joining of V, D and J segments in the heavy chain corresponds to CDRH3, the most structurally diverse region of the antibody. This region is also the most involved in antigen binding and thus the area of most sequence diversity corresponds to the area which must be structurally diverse across antibodies in order to bind many potential antigen.

1.1.8 Production

The high specificity and affinity of antibodies against a target antigen are highly desirable properties both as a laboratory tool and in drug design. The production of antibodies for clinical uses originates from serum therapy developed in the 1860s, after it was recognized that the symptoms of diphtheria could be alleviated by treatment with the serum of rabbits inoculated with attenuated *Corynebacterium diphtheria* (Stockwin and Holmes, 2003). As well as therapeutic applications, antibodies were exploited as early on as 1942 to perform the first immunohistochemistry experiments, where a fluorescent antibody derivative was used to stain tissue infected with *Pneumococcus* (Coons et al., 1942).

Antibody populations derived directly from serum originate from many different B cell lineages and are therefore known as **polyclonal**. Polyclonal antibodies recognize many different epitopes; in contrast, monoclonal antibodies originate from a one unique cell line and as a consequence recognize a single epitope. In the 1970s, Köhler and Milstein (1975) revolutionized the study of monoclonal antibodies through the creation of hybridoma technology. This technique allows the creation of immortal cell lines that produce monoclonal antibodies raised against a specific antigen. The birth of hybridoma technology has helped antibody-based techniques to become common place in almost all molecular biology laboratories. Antibody-based technologies allow molecular biologists to carry fundamental tasks on a daily basis. Examples include immunoprecipitation to isolate and concentrate a target of interest; enzyme-linked immunosorbent assay (ELISA) for detection and quantification of a target and immunofluorescence for the visualisation of target molecules within biological samples.

1.1.9 Engineering

Since the advent of hybridoma technology, the manipulation of antibody sequence and structure in order to confer a desired property has been feasible. Antibody engineering has three main areas of focus: altering antigen-binding and specificity properties, bestowing novel function and improving the stability and efficacy of therapeutic antibodies and laboratory agents.

1.1.10 Antibody fragments

As well as the previously described Fab and $F(ab')_2$, more antibody fragments have been isolated and utilised. While Fab and $F(ab')_2$ can be generated by proteolysis of full antibody, these fragments are also generated using genetic engineering. The Fv fragment consists of one V_H and one V_L domain which, due to its low stability, is commonly expressed as a single-chain Fv (scFv) with a peptide linker between the two domains to prevent disassociation (Bird et al., 1988). The single-chain nature of the scFv fragment also simplifies expression and is therefore advantageous in techniques such as phage display (McCafferty et al., 1990). By modulating the length of the peptide linker, dimeric and trimeric form of scFvs can be induced, allowing bi- and tri-specificities (Holliger et al., 1993, Iliades et al., 1997).

The smallest variable fragment is the V_H domain. V_H domains without a partner V_L domain have been observed in nature in the form of V_HH domains found as heavy-chain antibodies in camelids and VNAR domains found in the IgNAR antibodies of cartilaginous fishes (Muyldermans et al., 1994, Greenberg et al., 1995). As well as the absence of hydrophobic residues that are found at the V_H/V_L interface in partnered V_H domains, V_HH and VNAR domains have longer CDRH3 loops that are thought to compensate for the loss of an antigen-binding V_L domain. Additionally, these long CDRH3 regions allow the penetration of cavities that are normally inaccessible for the typically planar surfaces of a full V_H/V_L paratope (Stanfield et al., 2004). These observations lead researchers to engineer stable and functional single variable domains that do not exhibit the problems with aggregation faced using natural human variable domains (Dudgeon et al., 2012).

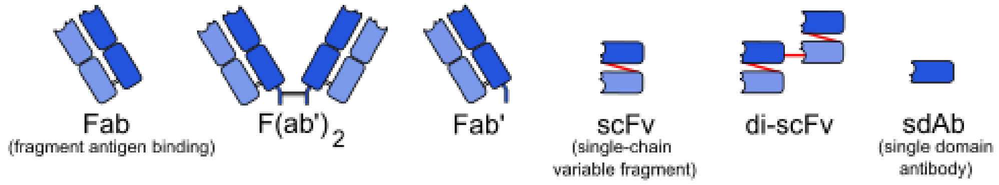

FIGURE 1.6: Antibody fragments. Light and heavy domains are coloured light and dark blue respectively. A red bar indicates a peptide linker. Image obtained and altered under a Creative Commons Attribute-Noncommerical license, available at `https://commons.wikimedia.org/wiki/File:Engineered_monoclonal_antibodies.svg`

1.1.11 Therapeutic antibodies

Original serum therapies had several inherent problems, including limited production, heterogeneity and side effects from the repeated injection of exogenous material eliciting an immune response (Stockwin and Holmes, 2003). While hybridoma technology significantly reduces production and heterogeneity issues, immunogenic problems still remain. Early hybridoma technology involves generating antibodies produced by B cells isolated from mice spleen cells. It was recognized that treatment with these mouse-derived antibodies sometimes elicited the production of human anti-mouse antibodies (HAMAs) in the patient, leading to the HAMA response (Klee, 2000). This response can reduce treatment efficacy and induce allergic reactions that range in their severity (Hwang and Foote, 2005). Significant progress has been made in improving the efficacy and safety of antibody-based therapeutics through the humanization of non-human antibodies, the development of transgenic mice with human immuunoglobulin genes (Tsurushita et al., 2005), the production of phage display human libraries (Pansri et al., 2009) and the development of rapid screening methods for the cloning of antibodies directly from a human blood sample (Smith et al., 2009).

1.1.11.1 Humanization techniques

Humanization refers to a process in which the sequence of a non-human antibody is altered in order to reduce immunogenicity, while maintaining binding specificity. The first humanization method was chimerization, in which the variable domain of a non-human donor antibody is joined to the human constant regions (Morrison et al., 1984). Though this method successfully reduces immunogenicity, a human anti-chimeric (HACA) response can still be raised (Hwang and Foote, 2005). Thus, many methods have followed since that aim to increase the human content of the antibody sequence further. CDR-grafting, a method in which the donor antibody CDRs are grafted on to a "scaffold" human framework sequence (Jones et al., 1986), has resulted in clinical successes (Hwang and Foote, 2005). However, CDR grafting alone may only partially confer the binding properties of the donor antibody, as the matching of certain framework residues between donor and scaffold has been shown to be necessary for antigen binding (Riechmann et al., 1988).

CDR-grafting was followed by resurfacing - a method where the antibody retains the CDRs and non-exposed residues, but surface exposed residues are changed to a human counterpart (Pedersen et al., 1994). More recent methods include Superhumanization, where the donor framework region is chosen from germline genes with similar canonical structures based upon homology of the CDR regions (Tan et al., 2002) and human string content optimization, which utilizes human germline sequences to determine the proportion of human-like sequence found in the target antibody sequence and to make changes accordingly (Lazar et al., 2007).

Nevertheless, immune responses can still be elicited and it cannot be assumed that humanized or even fully human antibodies are less immunogenic (Hwang and Foote, 2005).

1.2 Epitopes

An epitope (or antigenic determinant) is the part of an antigen that is recognised by the immune system through binding either to T cell receptors (TCRs), antibodies or antibodies in the form of B cell receptors (BCRs). The cognate partner defines the type of epitope; T cell epitopes (TCEs) bind TCRs, while B cell epitopes (BCEs) bind antibody or BCRs. For a given TCR/BCR/antibody and its antigen, epitope residues on the antigen form a geometrically and physico-chemically complementary surface to the cognate surface (paratope) that allows the formation of an energetically favourable interface. B and T cell epitopes both play a vital role in the recognition of foreign agents by the immune system. The binding of a B cell receptor to an antigen leads to its uptake by B cell via endocytosis. The antigen is then degraded within an endosome by proteolysis into peptides, some of which bind to MHC class II. The MHC class II-peptide complex is then presented on the surface of the B cell in order to bind to a TCR on the surface of a T cell. The peptide (or TCE) can only be recognised by the TCR in the context of the self MHC-peptide complex; this is known as MHC-restricted antigen recognition. The binding of MHC class II to a TCR allows the binding of CD40 on the surface of the B cell to CD40L on the T cell, activating the B cell and allowing it to undergo affinity maturation, class switching and differentiation into a memory cell. MHC class II binding also allows the binding of CD28 on the T cell to CD80/86 on the B cell, causing T cell activation and survival.

The characterisation of epitopes is important because it holds the promise of epitope prediction and design. The ability to predict epitopes would inform the selection of low immuogenic risk therapeutic candidates and conversely the rational design of vaccines.

1.2.1 T cell epitopes

As detailed above, TCEs are short linear peptides that can be the product of the MHC class II antigen-processing pathway. Additionally, TCEs can be the product of the MHC class I pathway which — in contrast to the MHC class II pathway — is responsible for the display of endogenous antigen. The MHC class I pathway is used for the processing of internal proteins for presentation to cytotoxic T cells. MHC class I typically presents peptides 8–11 amino acids in length, whereas MHC class II peptides are 13–17 amino acids in length. Both MHC I and II have an immunoglobulin-like structure and possess a large groove between two α-helices that allows peptide to bind (Bjorkman et al., 1987) (Stern et al., 1994). The primary difference between the MHC I and II grooves is that the MHC I groove is closed at both ends, whereas the MHC II groove is open at both ends; this difference is responsible for the difference in length of the binding peptide. The properties of the peptide-binding pocket are defined by the MHC gene isotypes, which are not only numerous but also known to be highly polymorphic (Mungall et al., 2003). This diversity means that different MHCs are able to bind different repertoires of peptides. Despite this diversity, pockets share common structural features that, in turn, define structural features that lead to TCE propensity (Spouge et al., 1987). Along with the fact that T cell epitopes are linear, this means that T cell epitope prediction is a tractable problem.

1.2.2 T cell epitope prediction

Many tools exist to predict TCEs. MHC I peptide binding predictors tend to perform well, due to the more restricted nature of the binding pocket; area under the the ROC curve (AUC, defined in section 2.5.3) scores of 0.87 have been reported (Yu et al., 2002). But this is also dependent on the isotype, as more difficult isotypes result in AUC scores of around 0.7 (Yu et al., 2002). In contrast, MHC II peptide binding prediction is more difficult. Again, prediction performance varies by MHC class II isotype, but in general AUC scores of around 0.7 are obtained (Wang et al., 2008). Despite this, the application of TCE prediction methods for the prediction of biologic therapeutic immunogenicity has been successful in some cases (Koren et al., 2007). However, *in silico* TCE prediction methods are limited by other factors contributing to the immunogenicity of a TCE that are currently outside of the model; these include the antigen-processing steps that reduce the number of peptides that are presented with MHC class II, affinity for TCR and accounting for T cell phenotype (Jawa et al., 2013).

1.2.3 B cell epitopes

In contrast to T cell epitopes, the B cell epitopes of proteins are formed from residues that are not necessarily proximal in sequence, brought together in space by the fold of the antigen to form a surface that binds to the antigen-binding site of an antibody. The non-linear nature of B cell epitopes makes their prediction much more challenging than T cell epitopes.

1.2.3.1 Characterisation

V(D)J recombination and somatic hypermutation leads to a vast potential antibody repertoire. Though many of these potential sequences cannot be expressed due to frame shift errors, this huge variety suggests that it may be possible to raise antibodies against any molecular surface. This begs the question: do the regions on a protein surface that form an interface with antibody have any features that distinguish them from the rest of the protein surface?

BCEs are classically divided into two groups - linear epitopes, that consist of contiguous stretches of residues, and conformational/discontinuous epitopes, formed from sequentially distant residues brought together by the protein's fold. Linear epitopes are often determined by the cross-reactivity of a peptide with a given antibody. However, it has been calculated that 90% of epitopes are discontinuous and, in reality, linear epitopes are likely part of larger conformational epitopes (Barlow et al., 1986). The reduced complexity of linear epitopes is desirable in applications such as vaccine design, where a peptide is easy to manufacture in comparison to a recombinant protein. To fully characterise epitope properties however, antigenic proteins must be considered structurally.

Over the past 30 years much work has been done on trying to find useful physical or biochemical properties to help distinguish epitope sequences and structures. The first epitope identification methods utilised properties that related to surface occurrence likelihoods such as hydrophobicity and secondary structure patterns (Hopp and Woods, 1981, Hopp, 1986). The first analysis of antigen surfaces followed, concluding that bound epiopes protruded significantly from the surface of the protein (Thornton et al., 1986). Since then, the expansion of structural data has allowed much more analysis of epitopic surfaces. Rubinstein et al. (2008) used computational methods to build upon previous work and thoroughly characterised BCEs in comparison to non-epitopic surfaces . They found that epitopes differed from normal surface in their physico-chemical, structural and geometrical nature and were thus able to describe a 'typical B cell epitope'. In summary, their typical BCE consists of 20 amino acids, with a small but significant over-representation of tryptophan, tyrosine, charged and polar residues, along with an under-representation of hydrophobic residues. Epitope residues are more solvent accessible and tend to be part of disorganised secondary structure and — in agreement with the early studies of Jones and Thornton (1997) — tend to be found on flat, convex surfaces. It is thought that these geometrical and structural properties allow epitopes to be highly accessible to the CDRs of antibodies, with unorganised secondary structure allowing the flexibility needed to find an energetically favourable conformation. Importantly, Rubinstein et al. (2008) also investigated the evolutionary conservation of BCEs and found that epitope residues are significantly less conserved than non-epitope surfaces.

Kringelum et al. (2013) recently undertook a further analysis of BCEs. Contrary to Rubinstein et al. (2008), they concluded that there is actually no significant difference in amino acid composition between epitopic and non-epitopic surfaces. By using epitopic and non-epitopic residue distributions with identical surface accessibility profiles, they only found a slight and insignificant over-representation of tyrosine and under-representation of valine and small hydrophobic residues. As well as this, they investigated geometric and spatial epitope properties. They found that epitopes tend to be elipsoid in shape. Investigating the spatial amino acid distribution of BCEs, they found that typically, an epitope was formed from a hydrophobic core, flanked by charged amino acids. Thus epitopes share similarity to regular protein binding sites, with an important difference being that they exhibit low evolutionary conservation.

Following on from Kringelum et al. (2013), Kunik and Ofran (2013) analysed BCE residues according to which CDR they were with in contact with. Using this classification, they found that, although overall epitope composition was not different from surface composition (in agreement with Kringelum et al. (2013)), each CDR had a distinct preference profile that diverged from the profile of all exposed residues by varying degree.

These studies support the hypothesis that there are differences between epitope and non-epitope surfaces, which suggests that B cell epitope prediction should be possible, although the magnitudes of the observed differences suggest that prediction may be difficult.

1.2.4 B cell epitope prediction

B cell epitope prediction methods can be split into two types: continuous and discontinuous prediction methods.

1.2.4.1 Continuous/linear BCE prediction methods

Continuous/linear BCE prediction methods simplify the prediction problem by focusing on contiguous regions of sequence that are used to approximate full structural BCEs. The data used to train and test these predictive models consist of antibody-peptide binding assays. The earliest linear BCE prediction methods relied on amino acid propensity scales, which assign a score to each amino acid based upon some physico- or biochemical parameter. These methods therefore rely on correlations between certain properties of amino acids and their presence in epitopic regions (El-Manzalawy and Honavar, 2010). A sequence profile can then be generated and used to determine stretches of sequence likely to be epitopic. Blythe and Flower (2005) exhaustively evaluated 484 amino acid scales and found that even the best set of scales performed only slightly better than random. The poor quality of purely scale-based methods has lead to the application of more sophisticated predictors that have incorporated Hidden Markov Models (HMMs), Support Vector Machines (SVMs) and other machine learning techniques to improve continuous BCE prediction methods, with varying degrees of success (El-Manzalawy and Honavar, 2010). The most successful of these recent methods have also reported good performance when applied to discontinuous BCE prediction (see section 1.2.4.5).

1.2.4.2 Structural/discontinuous BCE prediction methods

90% of BCEs are thought to be structural/discontinuous (Walter, 1986); thus the prediction of discontinuous BCEs is essential for the understanding of BCEs in general. Discontinuous prediction methods rely on structural information about the antigen in order to define residues or areas of surface that are likely to be part of an epitope. Invariably, x-ray crystal structures are used to train and test these methods. Structural information is often complemented by sequence information relating to the residues that constitute the epitope. Additionally, a number of methods have been tested on structural epitopes that only require sequence information (see section refsub:Predictors-using-Sequence-only).

1.2.4.3 Structural feature-based prediction

The majority of discontinuous epitope prediction methods make use of varied combinations of structure and sequence features to inform the residue classification decision.

The earliest structure-based prediction methods commonly used surface analysis as well as local structure information to inform prediction. The first method was CEP Kulkarni-Kale et al. (2005), which identifies residues with high solvent accessibility, before going through a number of explicit steps in order to include neighbouring residues into predicted BCEs. DiscoTope Haste Andersen et al. (2006), another early prediction method, takes propensity scores of surrounding residues into account when scoring residues. The propensity score is a combination of log-odds scores for each amino acid - calculated from a training data set of known epitope structures - and surface exposure values.

Later on, Rubinstein et al. (2009) utilized the features they had identified in their earlier paper (Rubinstein et al., 2008) to develop Epitopa. Notably, this was the first predictor to apply a machine learning method in an attempt to improve prediction. Using a feature selection method, they trained a naïve Bayes classifier on structural and sequence datasets. In contrast to previous methods that considered a prediction a success if it corresponded to the antibody-bound/unbound state seen in a x-ray crystal structure, the performance of Epitopia was evaluated by considering a prediction as successful on a whole antigen if the average score of the real epitope residues was higher than the overall average score for that antigen. This approach avoids the problems associated with using measures derived from a confusion table when the negative set is poorly defined (see sections 2.5.3 and 1.2.5).

Other methods have used whole-protein structural features to aid classification. The ElliPro (Ponomarenko et al., 2008) method uses whole-protein elipsoid shape approximation to determine residues that protrude highly on the protein surface which are then clustered to form potential epitopes. Scarabelli et al. (2010) also considered global structure features but took an non-traditional approach. Instead of relying on normal sequence and structural information, they attempted to predict epitopes through the use of molecular dynamics. Specifically, they proposed that epitopes consist of surface residues that tolerate mutations, are not involved in stabilizing the protein fold and are relatively flexible. On this basis, they utilised protein energetics determined from molecular dynamics simulations, along with topological information from a residue contact matrix, to determine residues that are minimally coupled to the the rest of the protein.

DiscoTope 2 (Kringelum et al., 2012) was released more recently, building upon the original method (Haste Andersen et al., 2006) by using half-sphere exposure measures to define surface accessibility as well as carefully defining benchmark datasets. The latter improvement involved properly defining multiple epitopes and including information about the biological unit of the antigen.

Other more recent methods include the BeTop method, which uses graph theory to describe the protein surface (Zhao et al., 2012). First, Delaunay triangulation is used to form a graph from the antigen surface residues. Markov clustering is then used to form subgraphs, each of which is input as a feature vector for an SVM classifier that learns an epitope/non epitope label.

1.2.4.4 Docking, antibody-specific and mimotope methods

Other BCE prediction methods that require either information about the antibody or some other experimental data are summarised here.

Docking methods attempt to define the surfaces involved in native protein-protein interactions, when structural information on both partners is known (Gabb et al., 1997). Docking methods have been shown to perform with slightly higher levels of accuracy than methods where only the epitope structure is available (Ponomarenko and Bourne, 2007). They have also been useful in elucidating certain properties of antigen-antibody interfaces, e.g. the asymmetric distribution of residue types on the paratrope and epitope (Chuang et al., 2008).

Additionally, some groups have worked on prediction of epitopes when the antibody sequence is known. Zhao and Li (2010) used epitope-paratope residue interaction patterns, along with epitope and paratope amino acid compositions, to produce the Bepar method. Bepar was shown to outperform DiscoTope on test set of fourteen samples not included in the training sets for either method. (Soga et al., 2010) used a similar methodology by extending a slightly modfied DiscoTope procedure to take into account paratope-epitope residue statistics in order to improve precision.

Finally, another class of methods focuses on the prediction of epitopes from specific biochemical assays of the antibody. Phage display libraries of random peptides can be used to determine mimotopes — peptides that bind with high affinity to the target antibody, but are not necessarily identical to any contiguous sequence in the antigen (Pizzi et al., 1995). Mimotope analysis attempts to map these peptides onto the antigen surface to determine the real epitope. While some methods use only sequence information to map mimotopes (Mumey et al., 2003), most methods rely on structural information (Huang et al., 2011).

1.2.4.5 Predictors using sequence-only information

Notably, two methods — CBTOPE (Ansari and Raghava, 2010) and BeePro — have reported superior performance in predicting discontinuous epitopes, yet only consider sequence features. CBTOPE was developed by training an SVM on windows of sequence using traditional sliding window techniques to capture simple amino acid frequencies and bio- and phyisco-chemical scale information. Despite the simplicity of the model, CBTOPE gave significantly higher performance than structure-based predictors when 5-fold cross-validation was performed on a benchmark set of antigen and their structural eptiopes.

For the BeePro method, Lin et al. (2013) used a combination of 16 properties, including physico-chemical and evolutionary descriptors, to build an SVM classifier. Specifically, they used position-specific scoring matrices output from psi-BLAST runs on each protein in the dataset to create PSSMs as input for the learning step. Similarly to CBTOPE, the predictor outperformed all previous methods on a conformational epitope benchmark dataset using 5-fold cross validation.

1.2.4.6 Meta-prediction

In a review of B cell epitope prediction methods by El-Manzalawy and Honavar (2010), the authors suggested that meta-predictors may improve performance by combining the results from several different methods. EPMeta is such a predictor, that combines the output of six predictors through a voting system. EPMeta was shown to outperform each of the six base predictors (Liang et al., 2010). More recently, Hu et al. (2014) demonstrated that meta-prediction could be approached as a machine learning problem, using output labels from eight base learners as input features for learning. Similarly to EPMeta, their meta-predictor was able to outperform each of the eight base learners.

1.2.5 Evaluation of discontinuous BCE prediction methods

The evaluation of discontinuous BCE prediction methods faces a number of challenges. These challenges are summarised, before attempts at evaluation are discussed.

Like any prediction problem, one of the problems of BCE predictor evaluation is the comparison of methods that have been trained and tested on different datasets. Testing a method on data used in the training step can lead to overstated performance. In an attempt to co-ordinate efforts, Ponomarenko and Bourne (2007) defined a benchmark dataset to use for the development of future methods. Though some methods that followed used this dataset (Ansari and Raghava, 2010, Lin et al., 2013), others have not done so (Sun et al., 2011, Kringelum et al., 2013). Furthermore, the definition of a benchmark dataset does not lead to a definition of training and testing; for example, in the two papers mentioned previously, the benchmark dataset was used to perform cross validation but not independent testing (Ansari and Raghava, 2010, Lin et al., 2013).

Additionally, the problem of training and test set definition is exacerbated by the variety of ways that epitope and non-epitope can be defined.The primary problem is the difficulty in defining a negative set. This is because all x-ray crystal structures consist of an antigen bound to a single monoclonal antibody. This is not representative of the reality *in vivo*, where the antigen is recognised by many antibodies as part of a polyclonal response. Thus there are likely to be surface epitopes that have not been defined in structural studies. Additionally, some surfaces may be precluded from antibody binding by location (e.g. intra-membrane regions) or by obligate binding partners. (Kringelum et al., 2012) were able to improve performance of the DiscoTope method by properly labelling multiple epitopes and using the biological unit data to augment predictions. As well as the problem of negative set definition, another consideration is what residues to include for prediction. Most predictors are evaluated on their predictive performance on all residues, including those that are within the core of the protein. Is it realistic to include fully buried residues when evaluating the performance of of a predictor, given that in a real-use-case scenario they would never be considered? It may be that the predictive power demonstrated by some methods may in fact simply be predicting surface/non-surface. This point will be returned to when the project aims are discussed (see section 1.3).

Despite these difficulties, method evaluations have been performed. In 2007, an evaluation of eight methods — including scale based, patch prediction and docking methods — on a benchmark dataset found that no methods that use antigen features alone were able to yield AUC values greater than 0.65, which is considered mediocre (Ponomarenko and Bourne, 2007). An evaluation of more current methods was performed by Hu et al. (2014) as part of an investigation into the efficacy of using meta-learning on top of existing methods . In order to evaluate the meta-learning methods developed, eight base learners were tested on the same data set; thus these performances are considered the most recent and most extensive evaluation of current methods. The testing by Hu et al. (2014) showed that the most recent methods available show improved scores, with the SEPPA 2.0 method giving the best AUC score of 0.765, as well as the next three best methods (DiscoTope 2.0, Bpredictor and ElliPro) giving AUC scores of close to 0.7.

1.3 Aims

As mentioned above, the overall aim of this project is the development of *in silico* methods to aid the selection and design of therapeutic antibodies. The work related to these aims falls into two sections: B cell epitope prediction and antibody biophysical stability prediction.

1.3.1 B cell epitope prediction

In order to raise an immunogenic reaction, therapeutic antibodies must be recognised primarily by host antibodies that bind to BCEs on their surfaces. B cell epitope prediction has the potential to inform the selection of candidate therapeutics by flagging those therapeutics with surfaces likely to bind antibody and therefore lead to an immunogenic response. Though TCE prediction has seen success in application to the prediction of immunogenicity, performance is not perfect. Though the exact relationship between TCEs, BCEs and immune response is complex and not completely understood, BCE prediction certainly holds the promise of being useful to drug design.

In order to predict BCEs on the surface of therapeutic antibodies, two aims were set: i) improve *general* BCE prediction and ii) specify BCE prediction for the prediction of human-host epitopes.

1.3.2 Improving general B cell epitope prediction

In a review written in 2010, El Manzalway et al. suggested BCE prediction may be improved by the utilization of advances in protein-protein interface prediction (El-Manzalawy and Honavar, 2010). In 2011, an in-house high performance general protein-protein interface predictor *IntPred* was developed in the Martin group (Baresic, 2011). The aim was to first test the performance of IntPred as a BCE predictor, before adapting the method in an attempt to improve BCE prediction in comparison to current methods.

1.3.3 Human-host B cell epitope prediction

Once a general BCE prediction had been developed, the aim was to improve BCE prediction in the context of predicting epitope bound to *human antibody* (human-host eptiopes). This is obviously the case when considering the response to therapeutic antibodies.

To create a B-cell epitope prediction method that is specific to a species (e.g. human), the specificity of that organism's immune response must be considered. One way in which organism immune responses differ is the nature of their antibody repertoire. Differences in the number and sequence of those genes responsible for the generation of antibodies contribute significantly to differences in antibody repertoire. In addition, functional differences between the proteins responsible for somatic hypermutation of these genes will alter the nature of stochastic sequence changes.

Another force by which the antibody repertoire is shaped is immune tolerance. Tolerance shapes the antibody repertoire via clonal deletion, anergy and receptor editing (see section 5.1.1 for a detailed introduction). The processes occur when B cell lymphocytes bind to self antigen. Because each species' 'self' consists of a different proteomic basis, their antibody repertoires are shaped in different manners accordingly. Presumably then, the surfaces of these self-proteins must be tolerated by the B cell population in order to avoid autoimmune reaction. If surfaces found on foreign protein are sufficiently similar to one of these tolerated surfaces then it can be assumed that no antibodies will be raised against this surface, as mechanisms of tolerance disallow an antibody to be raised that is significantly cross-reactive with self.

Human protein structures deposited in the PDB allow us to sample the human self-proteome. By using this sample of self-protein, the surfaces of these tolerated proteins can hopefully be described in such a way that allows surface-to-surface comparison. This method of comparison must be powerful enough to identify surfaces on non-human proteins that are either sufficiently similar to a human surface to be tolerated (and therefore non-immunogenic), or sufficiently different to be non-tolerated (and therefore immunogenic). These surfaces can be compared by breaking them into overlapping patches. By using a set of human surface patches as a reference, non-human proteins can be tested for the presence of potential tolerated and non-tolerated surfaces. This would help us filter BCE predictor results to avoid the prediction of BCEs that are in fact disallowed owing to tolerance.

1.3.4 Predicting the biophysical stability of therapeutic antibodies

The biophysical stability of a therapeutic antibody influences many properties important to its efficacy as a drug, including immunogenicity (an introduction to antibody biophysical stability is given in section 6.1). Though methods do exist to predict the relative effect of a point mutation on the stability of a single antibody, no methods exist to predict the absolute biophysical stability of an antibody. Using a dataset of human Fab sequences and biophysical stability measurements, we aim develop a method for the prediction of antibody biophysical stability.

Chapter 2

Tools and Resources

This chapter introduces the data resources, tools, algorithms and statistical methods used in the following chapters.

2.1 Data resources

This section introduces the data resources used in the following chapters. First the PDB, a major resource for structural data, is introduced, followed by the sequence data resource UniProtKB. The PDBSWS service is then introduced that allows mapping between UniProt and PDB data. The Gene Ontology resource, used for the labelling of UniProtKB data for form and function is then described. Finally the IEDB, a resource for epitope data, is described.

2.1.1 PDB

The Protein Data Bank[1] (**PDB**) is the central public repository for protein and other macromolecular structural data (Berman et al., 2000). The data comes in the form of plain-text PDB files, each of which contains information about a structure from a single experiment. The file is split into two components: the header, which contains annotations of the data (author, experimental conditions, etc.) and the body, which contains information on the structures resolved atoms and their co-ordinates.

As of September 2015, data on 122 583 structures have been deposited in the PDB. Of those structures, 93% are of proteins, whilst the remainder are mostly nucleic acids. 87%

[1] `http://www.rcsb.org/`

of structures are obtained from x-ray crystallography structures, 12% from NMR and the remaining 1% from electron microscopy.

2.1.2 UniProtKB/Swiss-Prot

The Universal Protein Resource[2] (**UniProt**, (The UniProt Consortium, 2009)) is the largest publicly available repository of protein sequence data. It is divided into four databases; UniParc, for protein sequence archives; UniRef, containing clustered sequences for rapid searching; UniMES, for metagenomic data and UniProt Knowledgebase (**UniProtKB**) the core database for protein sequences. UniProtKB is further divided into **UniProtKB/TrEMBL** — which contains unprocessed sequences — and **UniProtKB/SwissProt**, a significantly smaller set of manually curated, non-redundant sequences.

As of September 2016, UniProtKB/TrEMBL contains 66 905 753 sequences, comprised of approximately 22 billion amino acids, with an average sequence length of 336 amino acids per entry [3]. UniProtKB/SwissProt contains 551987 sequences comprised of approximately 200 million amino acids, with an average sequence length of 357 amino acids, taken from 246 580 references [4].

Every UniProt entry has a unique identifier known as the primary accession number, but previously known as accession code and therefore abbreviated to AC. Additionally, every entry has an entry name and optional secondary accession numbers. In the case of multiple entries being merged due to redundancy, one AC will be kept as the primary AC, with the remaining becoming secondary ACs. If an entry is to be split into multiple sequences, each new entry is given a new primary AC and assigned the old AC as their secondary AC.

2.1.3 PDBSWS

PDBSWS [5] is a relational database that provides mapping from a PDB residue to a UniProtKB residue, accessible via a RESTful web service (Martin, 2005). It primarily uses cross-references to UniProtKB entries found in PDB files, as well as cross-references to PDB files found in UniProtKB entries. Failing this, assignment is attempted by a brute-force sequence scan. PDBSWS also deals with mapping to and from secondary ACs. For each assignment, the UniProtKB sequence is aligned to the ATOM sequence of

[2] `http://www.uniprot.org/`

[3] `http://www.ebi.ac.uk/uniprot/TrEMBLstats`

[4] `http://web.expasy.org/docs/relnotes/relstat.html`

[5] `http://www.bioinf.org.uk/pdbsws/`

the PDB file; this alignment provides a residue-to-residue mapping that is then stored in the database. PDBSWS provides a unique mapping in the PDB-to-UniProtKB direction. It also provides a non-unique mapping in the opposite direction, allowing PDB chains assigned to the same UniProtKB entry to be collected easily. However, PDBSWS alone does not provide functionality to choose PDB entries according to any quality criteria.

PDBSWS is an in-house tool and is therefore regularly updated and easily accessible. Unlike other methods it is able to provide a residue-level mapping and also outperforms other methods in coverage and/or level of automation (Golovin et al., 2004). Therefore PDBSWS is used in this thesis for the mapping of UniProtKB to PDB sequences.

2.2 Gene Ontology

The Gene Ontology (GO) project[6] is a major initiative that aims to provide a set of structured, controlled vocabularies for the annotation of genes, gene products and sequences in order to ensure a consistent description of their attributes. The attributes of genes and gene products fall into three key biological domains: molecular function, biological process and cellular component. A further domain, the sequence ontology, describes sequence features.

The fundamental components of GO are *terms* used to describe functions, processes and components and *relationships* used to describe the relations between terms. The structure of GO can be described as a directed acyclic graph, where each term is a node and relationships between GO terms are represented by edges between nodes. GO is hierarchical — in the sense that child terms are more specialised than their parent terms — but not strictly so because child terms can have more than one parent term. Six relationships are commonly used: *is a* (*is a subtype of*); *part of*; *has part*; *regulates*; *negatively regulates* and *positively regulates*. An example of such a hierarchy is shown in figure 2.1.

GO has the major advantage of being applied for the annotation of UniProtKB entries due to the Gene Ontology Annotation (GOA) project (Camon et al., 2003). As well as manual annotations, the GOA have implemented a range of automatic annotation methods based on sequence similarity, orthology, domain information and pre-existing cross-references. As of September 2016[7], GOA provided GO annotations for almost 46 million gene products (of which 0.01% have manual annotation) across approximately 650 000 taxa. Thus GO annotations are used in this thesis for the identification of UniProtKB entries with a given cellular location.

[6] `http://www.geneontology.org/`

[7] `http://www.ebi.ac.uk/GOA/uniprot_release`

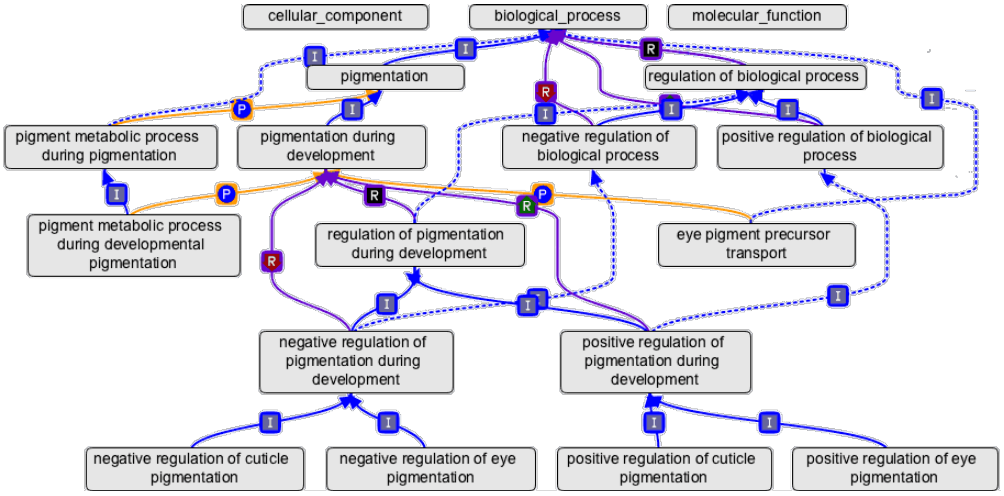
cellular_component
biological_process
molecular_function
pigmentation
regulation of biological process
pigment metabolic process during pigmentation
pigmentation during development
negative regulation of biological process
positive regulation of biological process
pigment metabolic process during developmental pigmentation
regulation of pigmentation during development
eye pigment precursor transport
negative regulation of pigmentation during development
positive regulation of pigmentation during development
negative regulation of cuticle pigmentation
negative regulation of eye pigmentation
positive regulation of cuticle pigmentation
positive regulation of eye pigmentation
I
R
P

FIGURE 2.1: Example GO ontology structure. In this example, the biological process of pigmentation is shown. Terms are shown as nodes in acyclic graph, with relationships between terms shown as edges between nodes. Arrow-heads point from child to parent. The letter at the center of each edge indicates the type of relationship, where I is for *is a subtype of*, R is for *regulates* and P is for *part of*. Image obtained from `http://geneontology.org/sites/default/files/u425/diag-ontology-graph.gif`

2.2.1 IEDB

The immune epitope database and analysis resource is a project hosted by La Jolla Institute for Allergy and Immunology that aims to provide data and resources for the analysis of T and B cell epitopes[8]. The main resource is the Immune Epitope Database (**IEDB**), a relational database of T and B cell epitope data collected from direct submissions and published literature. As of 2015, the IEDB contained manually curated data from more than 95% of the relevant published material, including 15 000 journal articles, covering 704 000 experiments (Vita et al., 2015). The IEDB can be queried by searching on a wide range of fields that relate to details about the epitope, antigen or experiment. Importantly, the IEDB also contains information from structural studies and cross-references the PDB (Ponomarenko et al., 2011). For this reason, the IEDB is used in this thesis to identify structures according to antigen and antibody annotation.

2.3 Tools and algorithms

In this section, tools and related algorithms are described that are used throughout the remainder of the thesis.

[8] `http://www.iedb.org/`

2.3.1 Solvent-accessible surface calculation

The solvent accessible area (or accessible surface area (**ASA**)) describes the area over which contact between an atom or residue of a protein and the solvent can occur. The concept of ASA and a method for its calculation was first described by Lee and Richards (1971). According to this method, the ASA is defined as the locus of the *centre* (rather than inward face) of a solvent molecule as it rolls over the var der Waals surface of a protein (see figure 2.2). Further to this, Lee and Richards also defined the relative-solvent accessibility of a residue $\boldsymbol{rASA}$

$$rASA = \frac{ASA(X)}{ASA_{av}(X)} \tag{2.1}$$

Where $ASA(X)$ is the observed ASA of residue X — termed the *absolute* solvent-accessibility — and $ASA_{av}(X)$ is the average ASA of residue X in the form of an Ala-X-Ala tripeptide. $rASA > 100$ indicates above average absolute ASA, common for the first or last residues of a chain, or residues with unusual (often erroneous) bond lengths or angles. $rASA$ is commonly represented as a percentage, as is the case in this thesis.

In this thesis the in-house program `pdbsolv`, part of the `BiopTools` package, was used to calculate ASA and $rASA$ values (Porter and Martin, 2015). Based on the method presented in Lee and Richards (1971), `pdbsolv` takes a PDB or PDML file as input and outputs atom ASA values, as well as residue and residue side-chain ASA and $rASA$ values. The probe radius can be set by the user: for this work, the default probe radius of 1.4 Å (corresponding to the radius of a water molecule) was used.

2.3.2 Surface patch creation

In order to calculate the properties of subsets of a protein surface, it has to be fragmented. Surface patches have been used in the prediction of general protein-protein interfaces (Baresic, 2011) as well as prediction of discontinuous B cell epitopes (Ponomarenko and Bourne, 2007). In this thesis, the program `pdbmakepatch` from the `BiopTools` tool set was used to form overlapping surface patches from the protein surface (Porter and Martin, 2015).

Before introducing the algorithm implemented by `pdbmakepatch`, the following terms must be introduced:

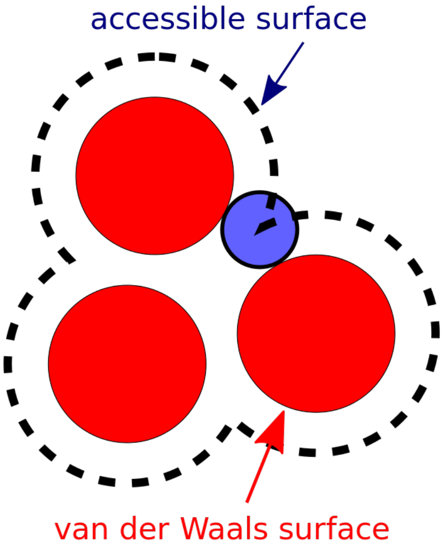

FIGURE 2.2: Accessible surface area (**ASA**). The ASA is defined as the locus of the centre of the solvent molecule (blue) as it rolls over the var der Waals surface of the protein (red). Image obtained under a Creative Commons Attribute-Noncommerical license, available at `https://commons.wikimedia.org/wiki/File:Accessible_surface.svg`

Patch centre atom is the central atom that is input to `pdbmakepatch` around which the patch is built. The residue to which the atom belongs is termed the patch centre residue.

Patch radius is the threshold distance from the patch centre atom used to select candidate residues for inclusion within the final patch.

Contact radius is defined for a pair of atoms as the sum of their var dar Waals radii, plus a tolerance (here set to 0.2 Å). Two atoms are *in contact* if the distance between them is less than the contact radius.

Residue geometry vector is a euclidean vector defined for a given residue with its initial point at the C_α and its terminal point at the centre of geometry of the 10 spatially closest neighbours. The centre of geometry is calculated as the average of the neighbours' C_α coordinates.

Residue solvent vector is also defined with its initial point at the C_α of a given residue, but points in the opposite direction to the residue geometry vector.

Solvent angle is defined between two residues and is the angle between the two residue solvent vectors.

For a given PDB file and a patch centre atom, `pdbmakepatch` iteratively builds a patch using the following procedure:

1. Define P as the set of atoms in the patch and add the patch centre atom to P.
2. Determine all residues with at least one atom centre within the patch radius from the patch centre atom. These are the set of residues C that are candidates for inclusion within the patch.
3. For each member of P, test if any of the members of C are in contact. If a member of C is in contact with a member of P and the solvent angle between them is less than 120° then move it to P.
4. Repeat step 3 until no more members of C are moved to P.
5. Label any residue with an atom in P as a patch residue.

The solvent angle test is used to avoid including residues from opposite sides of a pocket in the same patch. This prevents the creation of discontinuous patches (Jones and Thornton, 1997, Pettit et al., 2007).

2.3.2.1 Generating patches from a structure

For a given structure, a set of overlapping patches can be created to represent its surface. In order to create such a set, the set of highly solvent-accessible residues with an $rASA >$ 25% are identified. This is the set of patch centre residues. For each patch centre residue, the atom with the highest ASA is found. Each of these highly-solvent accessible atoms is a patch centre atom that is input into `pdbmakepatch`. In this thesis, a patch will be identified by the PDB code, chain ID and residue label of its central atom (e.g. *1djs:A:100*).

2.3.3 Protein sequence alignment and clustering

The comparison of protein sequences using alignment is a vital bioinformatic task. The clustering of sequences to prevent redundancy is similarly essential. As well as being part of the IntPred method (see section 2.3.4), alignment tools are used to select and label antigen structures (see section 3.4.2). In this thesis, cd-hit (Li and Godzik, 2006) is used for sequence clustering, whilst Clustal Omega (Thompson et al., 1994) is used for multiple sequence alignments.

2.3.3.1 cd-hit

cd-hit is a program widely used for the rapid clustering of large sequence data sets by sequence similarity (Li and Godzik, 2006). It works by counting the occurrence of 'words' (di or tri-peptides) and then estimating sequence similarity by comparing word counts between sequences, avoiding the time-consuming process of explicit sequence similarity calculation. cd-hit is particularly fast when clustering at high sequence similarity (Fu et al., 2012). cd-hit is used in this thesis for the clustering of antigens by sequence similarity for the construction of test and training data sets.

2.3.3.2 Clustal Omega

Clustal Omega is a dynamic programming multiple sequence alignment method that is the current implementation in the Clustal series (Sievers et al., 2011). The Clustal Omega algorithm has two main steps: i) use a modified version of the `mbed` algorithm to create a guide tree (Blackshields et al., 2010) ii) use the guide tree to carry out a multiple alignment. As the name suggests, the guide tree guides the creation of larger and larger sub-alignments by following the the branch order of the tree. This avoids the computation of an exact alignment, which is prohibitively time-consuming for all but the smallest numbers of sequences. The use of a modified `mbed` algorithm to build the guide tree also avoids the computation of a distance matrix (performed by the predecessor clustalw (Thompson et al., 1994)) by transforming each sequence into an n-dimensional vector (where n is proportional to the log of the total number of sequences), where each element is simply the distance between that sequence and one of n reference sequences. Clustal Omega exhibits better accuracy than other fast methods and also comparable accuracy to intensive slow methods (Sievers et al., 2011). Thus it is used in this thesis for the alignment of multiple sequences.

2.3.4 IntPred

IntPred is an in-house general protein-protein interface prediction method (Baresic, 2011). As IntPred is an important component in the work on BCE prediction presented in this thesis, the details of the method will be discussed in chapter 3. Briefly, IntPred is a method for the prediction of interface patches on the surface of any protein. The method was developed by generating a set of features for a large dataset of protein-protein complex structures. These features were then input for the training of a random forest algorithm (see section 2.5.5.1). Upon testing, IntPred was able perform well in comparison to other methods (see section 3.2.4). Importantly, because IntPred is in-house, it was amenable to the retraining, retesting and other amendments necessary in this thesis. IntPred is described in more detail in section 3.2.

2.4 Statistical methods

This section describes the statistical methods applied throughout this thesis. The two sample *t*-test and Mann-Whitney *U* test are applied for the comparison of two populations measured on a continuous scale; the former is used for normally distributed data whilst the former is used for non-normally distributed data. Fisher's exact test and the χ^2 test are used for the comparison of categorical variables between two and more than two populations respectively. Pearson's and Spearman's Rank correlation coefficients are used to test the correlation between normal and non-normal data respectively. Finally, PCA and the equivalent method of multi-dimensional scaling are used for the visualisation and investigation of high-dimensional data.

2.4.1 *t*-test

The t-test is a parametric statistical test measuring the significance in the difference in means of two normally-distributed populations. A *t*-test effectively tests the significance of the difference between two means. A number of related *t*-tests can be applied for different contexts.

In this thesis, a two-sample *t*-test is used for the comparison of two separate sets of independent samples, one from each of the populations being compared. The null hypothesis is that the means of the two populations are equal. In this thesis, Welch's *t*-test is used (Welch, 1947). For a Welch's *t*-test, the test statistic t is calculated as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2.2}$$

where $\bar{X}_i$, s_i^2 and n_i are the sample mean, variance and size for sample i. The degrees of freedom used to calculate a p-value for significance testing are approximated using the Welch-Satterwaite equation (equation 28 in Welch (1947)), which is based on the linear combination of degrees of freedom from each of the sample's variances.

Welch's *t*-test was originally developed for the testing of samples with unequal variance (Welch, 1947). However, studies since have demonstrated that Welch's *t*-test is preferable to student's t-test in almost all cases (Moser et al., 1989, Ruxton, 2006). Unlike Student's t-test, Welch's t-test maintains type I error rates close to nominal for unequal variances and for unequal sample sizes. Furthermore, Student's t-test is only slightly more powerful when variances are equal but the difference in sample size is very large; otherwise, the power of the two tests is equivalent. Using Welch's t-test also avoids the problems associated with running a two-step procedure of testing for equal variance before running a Student's t-test (Ruxton, 2006). Thus in this thesis, Welch's t-test is used. `t.test` implemented in the `R` language was used with the default settings of two-sided Welch's t-test.

2.4.2 Mann-Whitney U test

The Mann-Whitney U test is a non-parametric test of the null hypothesis that two samples are drawn from the same population. It is an alternative to a two-sample t-test when the distributions being compared are not normal. The test relies on the measurement being ordinal in order for measurements to be ranked. The test statistic U is defined as the smaller of U_1 and U_2

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{2.3}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{2.4}$$

where n_i and R_i are the size of, and the sum of, the ranks of sample i respectively. `wilcox.text` implemented in the `R` language was used with the default settings to run a two-sided Mann-Whitney U test when two samples are supplied.

2.4.3 Pearson's correlation coefficient

Pearson's correlation coefficient is a measure of the linear dependence between two variables that gives a value between -1 and 1, where 0 is no correlation, 1 is perfect positive correlation and -1 is perfect negative correlation. It is calculated as the covariance of the two variables divided by their standard deviations. For an n-size sample with two sets of measurements $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_n\}$, Pearson's correlation coefficient r is calculated as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.5}$$

where $\bar{x}$ and $\bar{y}$ are the means.

The interpretation of r is context-dependent and any set criteria is in some way arbitrary (Cohen, 1977). It is important to calculate a p-value for r to test the statistical significance of the coefficient. `cor.test` implemented in the `R` language was used with the 'pearson' method to calculate r and an associated two-sided p-value.

2.4.4 Spearman's rank correlation coefficient

Spearman's rank correlation coefficient, or r_s, is a non-parametric alternative to Pearson's coefficient that measures the statistical dependence between the ranking of two variables. Similarly to the Mann-Whitney U test, the measurements must be ordinal. For two sets of measurements X_i and Y_i, first their rank variables rg_X and rg_Y are calculated. Then r_s is defined as the Pearson correlation coefficient between the two rank variables

$$r_s = \rho_{rg_X, rg_Y} = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{2.6}$$

where σ_{rg_Y} are the standard deviations of the rank variables, ρ is the Pearson correlation coefficient applied to rank them and $\text{cov}(rg_X, rg_Y)$ is the covariance between them. Similarly to Pearson's coefficient, a p-value can also be calculated to test the significance of a correlation. `cor.test` implemented in the `R` language was used with the 'spearman' method to calculate r_s and an associated two-sided p-value.

2.4.5 χ^2 test

The Chi-squared (χ^2) (Mood et al., 1974) test is a non-parametric test used on nominal and categorical data. It can be interpreted either as a *goodness of fit* test, where it is used to compare a frequency distribution of a sample with a theoretical distribution, or as a *test of independence*, where two samples are compared with the null hypothesis that they are drawn from the same sample. To carry out the test, data for a which a categorical variable with k outcomes has been measured are divided into n groups. The category outcomes must be mutually exclusive and the frequency probabilities for each group, over all categories, has to sum to 1. χ^2 is defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2.7}$$

where O_i is the observed count and E_i is the expected count. The test has $(n-1)(k-1)$ degrees of freedom.

Table 2.1: Fisher's exact test. Equation 2.8 shows how the exact p-value is calculated from the table.

	A	B	
X	x_A	x_B	$x_A + x_B$
Y	y_A	y_B	$y_A + y_B$
	$x_A + y_B$	$x_B + y_B$	n

The χ^2 test assumes that underlying distribution of the data is the χ^2 distribution, a special case of the gamma distribution. When expected counts are low, care must be taken to avoid erroneously low p-values: 80% of cells should have an expected count of ≥ 5 and no cell should have an expected count of < 1 (Bewick et al., 2004). A common method to address this is the pooling of groups or categories (McDonalds, 2014), though this does reduce the degrees of freedom. Another option is the Yates Correction (Yates., 1934), where 0.5 is subtracted from the difference between O_i and E_i, but this is only applicable to 2×2 contingency tables (i.e. $n = 2, k = 2$). In this thesis, all 2×2 contingency tables are tested using Fisher's exact test (see section 2.4.6) and pooling will be used when expected cells counts are too low. `chisq.test` implemented in the `R` language was used in this thesis with the default parameters.

2.4.6 Fisher's exact test

Fisher's exact test is used as an alternative to the χ^2 test for 2×2 contingency tables (Fisher, 1922). The test is able to handle small cell expected values because — in contrast to the *estimated* p-value generated by a χ^2 — an *exact* p-value is calculated using the hypergeometric distribution

$$p = \frac{\binom{x_A+x_B}{x_A}\binom{y_A+y_B}{y_A}}{\binom{n}{x_A+y_A}} \tag{2.8}$$

where x_A, x_B, y_A and y_B are the cell values of the table (see 2.1) and $\binom{n}{k}$ is the binomial coefficient. Although the complexity of the calculation can make Fisher's exact test infeasible for large data sets, it was possible to use it in all of the 2×2 contingency tables used in this thesis. `fisher.test` implemented in the `R` language was used with the default parameters.

2.4.7 Brown-Forsythe test

A Brown-Forsythe test can be used to test for the equality of variance between measurements of two or more groups (Brown and Forsythe, 1974). Letting

$$z_{ij} = |y_{ij} - \tilde{y}_j| \tag{2.9}$$

where $\tilde{y}_j$ is the median of group j and y_{ij} is the i-th measurement from group j, the Brown-Forsythe test statistic F is calculated as

$$F = \frac{(N-p)}{p-1} \frac{\sum_{j=1}^{p} n_j (\tilde{z}_{.j} - \tilde{z}_{..})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n_j} (\tilde{z}_{ij} - \tilde{z}_{.j})^2} \tag{2.10}$$

where N is the total number of observations, p is the total number of groups, n_j is the number of observations in group j, $\tilde{z}_{.j}$ is the mean of group j and z_{ij} is the mean over all groups. In this thesis, the Brown-Forsythe test is only used to test for equality of variance between two groups.

2.4.8 Principal component analysis

Principal component analysis (**PCA**) is a statistical method used to transform data onto a new set of co-ordinates such that the greatest variance of some projection of the original data lies on the first co-ordinate (or *component*), the second greatest variance on the second component and so on (Jolliffe, 2014). Because the transformation is orthogonal, each component is orthogonal to its preceding components. An example in two dimensions is shown in figure 2.3.

Mathematically, PCA can be defined as the transformation of a data matrix $\boldsymbol{X}$ with n rows and p columns, defined by a set of p-dimensional vectors of weights (or *loadings*) $\boldsymbol{w}_k = (w_1, \cdots, w_p)_{(k)}$ that map each row vector $\mathbf{x}_i$ of $\boldsymbol{X}$ to a new vector $\boldsymbol{t}_{(i)} = (t_1, \cdots, t_k)_{(i)}$ given by $\boldsymbol{t}_{(k)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(i)}$, where each $\boldsymbol{t}_i$ successively inherits the maximum possible variance from $\mathbf{x}$ and each loading vector $\mathbf{w}$ is constrained to be a unit vector.

The principal components of a matrix can be found by calculating its covariance matrix. The principal components correspond to the eigenvectors of this covariance matrix. The primary component corresponds to the eigenvector of the largest eigenvalue, the second to the second largest eigenvalue and so on.

Because PCA maximises the variance within the first components, it acts to retain the maximum of the variance in the original data set within a reduced number of dimensions. In other words, PCA projects the original data into a smaller-dimensional subspace while minimising the reconstruction error. Because of this, PCA is commonly used to transform high-dimensional data into reduced dimensions for visualisation and exploratory analysis,

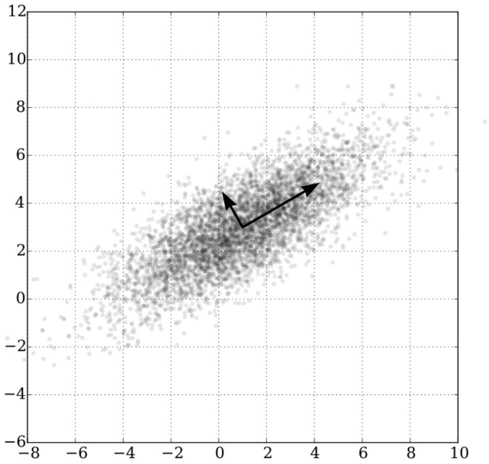

Figure 2.3: Example of PCA in two dimensions. The longer and shorter arrows show the first and second components respectively. These components are defined to have the greatest orthogonal variance when the data is projected onto them. Image obtained under a Creative Commons Attribute-Noncommerical license, available at `https://commons.wikimedia.org/wiki/File:GaussianScatterPCA.svg`

which is how it will be applied in this thesis. `prcomp` implemented in the `R` language was used in this thesis with the default parameters.

2.4.8.1 Multi-dimensional scaling

Multi-dimensional scaling (**MDS**, also known as Principal Coordinate Analysis) (Borg and Groenen, 2005) is used to visualise the level of similarity between objects in a set. Given a matrix of pairwise dissimilarity measures (or distances) for a set of objects, an MDS algorithm aims to place each object in an N-dimensional space where the pairwise distances are preserved. In this thesis, classical (Torgerson) MDS is used. Classical MDS is equivalent to PCA if the distances used are euclidean (Cox and Cox, 2001), which is the case for its application in this thesis. `cmdscale` implemented in the `R` language was used in this thesis with the default parameters.

2.5 Machine learning

The combination of exponentially increasing computational power and storage capabilities has lead to an explosion in the amount of data in all fields, including the biological sciences. This is evidenced by the growth seen in key central repositories such as UniProt and the PDB. Data in such unprecedented volume poses a challenge because *data* is not *information.* The elucidation of information or *structure* from large amounts of data is the central aim of machine learning.

Machine learning is defined by Mitchell (1997) in the following terms

> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.*

The most important and distinctive element of machine learning is the basis of *experience* to improve performance. Performance as a function of experience is why machine learning is well suited to problems for which large amounts of data (experiences) are available. Using Mitchell's terminology, experience comes in the form of training examples or **instances**. Each instance has a set of measured **attributes** (or **features**) that describes it. The number and choice of features is the decision of the experimenter and must be chosen with care for relevance to the task in mind. Ideally, each instance will have a value for each feature, though this is not always necessary (see section 2.5.2). Features can be one of two types: **numerical** features are expressed on a numeric scale, whilst **categorical** features are defined as finite set of mutually exclusive categories. These instances and their features are then the input for a learning process, the types of which will now be discussed.

The learning process can be split into two main types: supervised and unsupervised. In **supervised learning**, the aim is to predict how the features of an input instance affect some *outcome* feature. If the output feature is categorical, then the learning process is known as **classification** and the outcome feature is called the **class attribute**, or **class label**. Alternatively, if the outcome feature is numeric and continuous, then the process is a **regression**. In this thesis, only classification tasks are presented.

For classification, the learning process (or *learner*) requires a set of training instances known as the **training set**. For the training set, all of the features (chosen by the experimenter) for each instance are available to the learner, including the class attribute. The aim for the learner is to learn a set of rules or parameters such that the learner can assign class attributes to a **test set** — a group of instances for which the class attributes are not made available to the learner — as accurately as possible.

The problem of classification can be defined mathematically as

$$y = g(x|\Theta) \tag{2.11}$$

where the aim is to learn the mapping function $g(\cdot)$ that maps from the input space X to the output space Y by optimising the set of parameters Θ such that the error on the training set is minimised (Alpaydin, 2009). In this thesis, only binary classification problems are presented, where $y = 1$ for a positive outcome and $y = 0$ for a negative outcome.

Unsupervised learning can be applied when the instances available have no class attributes (unlabelled data). A lack of class attributes means that no approximation of a mapping function $g(\cdot)$ can be done and no test set is necessary. Instead, the aim of the task is to find hidden structure in the data. Examples include clustering, in which the task is to group instances in such a way that grouped instances are more similar to each other than they are to instances in other groups, and anomaly detection methods that seek to identify instances that do not conform to structures in the data set. PCA, introduced in section 2.4.8, is another example of unsupervised learning.

2.5.1 Data sampling

The efficacy of a machine learning method is dependent on the quality of the data available at training. The training set should be as representative as possible of the population concerned. This is a function of the size of the sample as well as the strategy used to collect it.

The collected data also need to be divided into training and test sets. How data are divided is a balancing act; on one hand, a larger training set may lead to a better model, but a smaller test set may not be sufficiently representative enough to give an accurate evaluation of performance; conversely, a smaller training set may lead to a poor learner that is evaluated thoroughly! One solution to the problem is **cross-validation** (**CV**), whereby all instances are used iteratively for training and testing. The data are partitioned into N non-overlapping subsets (or **folds**) of equal size, before training is done on all except one fold, which is used as a test set. This process is repeated, building N models each tested on a different fold. Evaluation of a cross-validated method is reported as the average of the N models.

Cross-validation is a form of sampling *without replacement*: once a sample is selected for training or testing, it cannot be selected again. Alternatively, *with replacement* sampling (or **bootstrapping**) allows the sampling of an instance more than once by replacing instances after selection. Sampling with replacement follows the binomial distribution, which means that if N samples are drawn with replacement from a set of N instances, then, on average, 63.2% will be chosen once or more, leaving 36.8% of the set for **out-of-bag** (OOB) testing.

Another important factor in the selection of data for learning is class balance. If the proportions of each class attribute value differ significantly in the data set then the data are said to be *imbalanced*. **Class imbalance** is problematic because most learners will be biased towards minimizing the error on the majority set (Japkowicz and Stephen, 2002). As an example, consider a set of 1000 instances where 10 instances have minority class label X and the remaining 990 instances have the majority class label Y. If the learner is aiming to maximise accuracy (see section 2.5.3) then it can simply assign the Y to the class attributes of all instances to score an accuracy of 0.99! Bias towards error minimization on the majority class is problematic because it is normally the minority class that is of interest (e.g. fraud detection, rare disease diagnosis).

A number of methods exist to tackle the class imbalance problem (He and Garcia, 2009), the simplest of which directly modulate the distribution of class values in the training set by **over-sampling** instances of the minority set or **under-sampling** those of the majority set. Over and under-sampling techniques are both known to be effective, the most effective being dependent on the data set (Estabrooks et al., 2004).

2.5.2 Missing data

Often in classification problems there will be instances that are without a value for some feature. Missing data can occur for a multitude of reasons, including malfunctioning measurement equipment, the collation of data from differently designed experiments or the refusal of a participant to answer a question. The cause of the absence of data should always be considered, as often there is a reason that is indicative of some structure in the data. As an example, it may be that those participants who refuse to answer a question about age tend to be older. If the aim of the analysis requires a representative sample, this non-random nature within the missing data may be problematic and should thus be a factor that is considered.

There are a number of methods to deal with missing values. The simplest is to remove all instances with missing attribute values. This might be feasible if a large number of instances do not have any missing values, but is problematic when data are limited. The alternative to excluding instances is data imputation, whereby missing data values are replaced by estimates that are based on information available from the training set. The most common imputation methods are the mean and mode imputation methods, which simply replace missing values with the nominal statistic. More sophisticated methods include the k-nearest neighbour method, as well as probabilistic methods, such as the fractional instances method implemented within the C4.5 learning method (Batista and

Table 2.2: Confusion table for a binary classification problem.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Table 2.3: Binary classification performance measures.

Measure	Formula	Range
Sensitivity / Recall / True Positive Rate	$\frac{TP}{TP+FN}$	$[0,1]$
Specificity / True Negative Rate	$\frac{TN}{FP+TN}$	$[0,1]$
Precision / PPV	$\frac{TP}{TP+FP}$	$[0,1]$
False Discovery Rate	$1-PPV$	$[0,1]$
False Positive Rate / Fall-out	$\frac{FP}{FP+TN}=1-Sp$	$[0,1]$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	$[0,1]$
F1	$\frac{2TP}{2TP+FP+FN}$	$[0,1]$
Matthews Correlation Coefficient	$\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	$[-1,1]$

Monard, 2003). The imputation methods used in this thesis are discussed in section 2.5.5.1.

2.5.3 Classifier evaluation

In order to evaluate a classification method, the agreement between the predicted class attributes and their real values must be assessed. For binary classifiers where one class label (normally the label of interest) is termed positive and the other negative, a prediction can have one of four outcomes: correctly predicted positive (true positive, **TP**), incorrectly predicted positive (false positive, **FP**), correctly predicted negative (true negative, **TN**) or incorrectly predicted negative (false negative, **FN**). These outcomes can be presented in a table commonly known as a **confusion table** (see table 2.2). The confusion table is the basis for the evaluation of a binary classifier's performance. A number of measures are widely used that each take into account different parts of the table to convey different properties of the classifier. The use of a variety of measures reflects the fact that different types of performance are desired for different types of problems. In the case of a diagnostic test for a serious health condition, false positives are less of a concern than false negatives. Conversely, in a problem like email spam filtering, false negatives are preferable to false positives (labelling legitimate email as spam). Table 2.3 defines some measures commonly used for binary classifier performance evaluation.

Sensitivity (or recall) tells us the proportion of positive cases correctly labelled as positive. Sensitivity is the most important measure when the avoidance of false negatives is the primary concern. A classifier with low sensitivity is said to be *under-predicting*. **Specificity** is the equivalent measure for negatives cases; it is the proportion of negative cases correctly labelled as negative. A classifier with low specificity is said to be *over-predicting*. **Precision** (or positive predictive value (**PPV**)) is the proportion of instances labelled positive that are truly positive, the inverse of which is the false discovery rate (**FDR**). The false positive rate (**FPR**) gives the proportion of negative cases incorrectly labelled as positive. There is normally a trade-off between sensitivity and precision that is modulated by the setting of a prediction threshold. When this is the case, a receiver-operating characteristic (**ROC**) curve can be plotted by plotting sensitivity against the false positive rate ($1 - Sp$). The area under the curve (**AUC**) can then be calculated, with 0.5 representing a random predictor and 1 a perfect predictor.

Accuracy is the proportion of all instances that have been correctly labelled. Accuracy takes into account all prediction outcomes but is misleading in the case of highly imbalanced data sets, where high accuracy can be obtained simply by labelling all instances with the majority class label. The **F1** is the harmonic mean of precision and sensitivity and as such is intended to give a single measure of how effective a classifier is. However, the F1 score does not take true negatives into account and therefore focuses on the positive class only. The better alternative is Matthews Correlation Coefficient (**MCC**), which is calculated using all four outcomes. MCC is essentially the correlation between the predicted and actual labels and takes a value between -1 (perfect negative correlation) and 1 (perfect positive correlation) and thus 0 means the classifier assignments are no better than random.

2.5.4 Benchmarking

Cross-validation is a useful method for giving an indication of performance. In particular, cross-validation is commonly used to assess the effect of parameters that can affect performance but are outside of the learning process (e.g. additional features, training set balancing, etc.). However, cross-validation is not a replacement for testing on a test set that is independent of the training set (or **benchmarking**). Benchmarking aims to provide an accurate estimate of future performance on unseen data. Benchmarking is difficult if data are scarce because researchers are likely to have used it during method development. This is exacerbated if researchers do not coordinate the use of training and test sets.

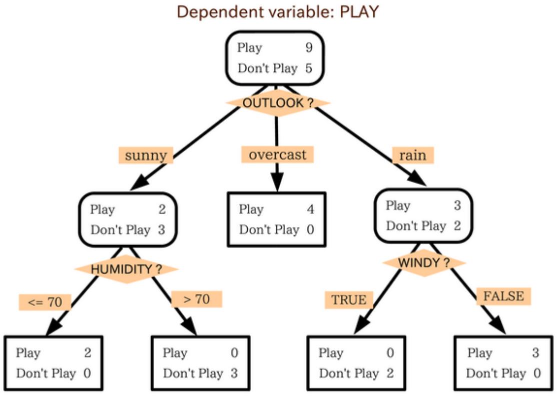

Figure 2.4: Decision tree example. In this example, the outcome is whether a sports game should be played. Nodes and attribute values are shown as orange parallelograms and rectangles respectively. At each node, the number of instances with each class attribute are shown. Image obtained and modified from `https://commons.wikimedia.org/wiki/File:Decision_tree_model.png`

2.5.5 Classifier algorithms

There is a wide range of algorithms available for the task of classification (see Kotsiantis (2007) for a review). However, in this thesis, only the random forest algorithm is presented. This is because the work here builds on a general protein-protein interface method IntPred (Baresic, 2011) that was developed in-house previously (see section 2.3.4). IntPred is a random forest classifier and thus this type of classifier will be introduced.

2.5.5.1 Random forest

Random forest is an extended decision tree method. Decision trees split instances by their attribute values (Breiman et al., 1984). Starting at the root node of the tree, instances are divided according to features used at the node (see figure 2.4). Instances then move down the tree to child nodes, where splitting happens again, and so forth until a terminal (or leaf) node is reached that determines which class attribute value is assigned. To build the tree during training, the feature used at each node is that which can best divide the data by class value. Nodes are successively added until each terminal node only contains instances of one class. Trees built in this manner are prone to over-fitting on the training data. To avoid over-fitting, the experimenter has the option to set a limit on the size of the tree, or prune the tree (see Breslow and Aha (1997) for a survey of such methods).

Random forests are simply a collection of decision trees, whereby the final class label is decided by a voting function on all predictions from all trees (Breiman, 2001) (see figure 2.5). To build a random forest, the experimenter must define the two parameters:

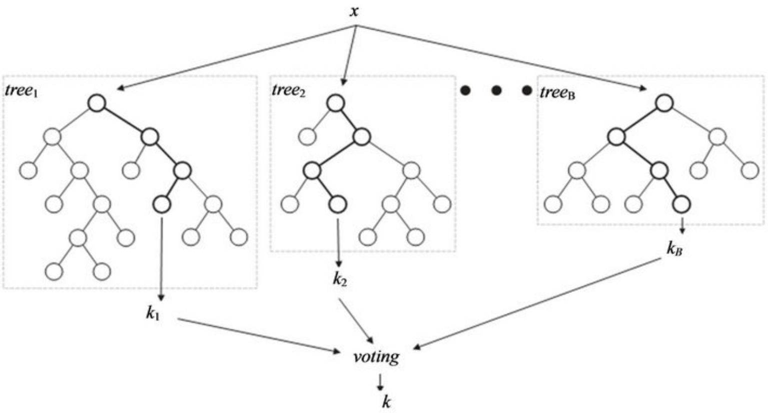

Figure 2.5: Random forest. An instance x is input into a forest with B trees. Each tree assigns a prediction label k_i to x according to its attribute values and the attributes used to divide instances at each node of the tree. B prediction labels are collected and a voting function decides the final prediction label k. Image reproduced from Nguyen et al. (2015) under a Creative Commons Attribution License.

the number of trees B and the feature bag size m_{try} (explained shortly). Each tree is trained on a bootstrapped sample of the data set, leaving 1/3 (on average) of instances as an out-of-bag test set for evaluation. Without feature bagging, this is essentially the closely related bagging algorithm (Breiman, 1996). However, the random forest method differs in that the feature used at each node is chosen from a randomly selected subset (or feature bag) of m_{try} features, rather than all p features. If trees are free to select from p features at each node, they will tend to select the same features and thus be correlated. Conversely a small m_{try} will lead to less informative trees. m_{try} is thus a tunable parameter, although it is believed that, as long its value is not p or 1, the effect on performance in minimal (Svetnik et al., 2003): typically $\lfloor\sqrt{p}\rfloor$ is chosen (Hastie et al., 2009). The other tunable parameter is B, which tends to show a benefit that levels off as it increases: hundreds or thousands of trees is typical.

Random forests have a number of advantages over other methods. Random forests have been shown to be more accurate than decision trees and comparable with, or better than, other popular machine learning algorithms, including artificial neural networks and support vector machines (Svetnik et al., 2003, Baresic, 2011). In comparison to decision trees, tree building and evaluation time are significantly reduced by feature subsetting and out-of-bag evaluation respectively. The avoidance of a tree-pruning step also simplifies model building. In fact, the random forest is one of the simplest models to build, as only two robust parameters (B and m_{try}) can be tuned, making it an easy method to apply.

In this thesis, two implementations of random forests are used, both of which are based on the algorithm presented by Breiman (2001). `RandomForest` implemented in `WEKA` (Hall et al., 2009) is used, as well as the `randomForest` package implemented in `R`. If it is not mentioned explicitly, it should be assumed that the `WEKA` implementation is being used.

The `WEKA` and `R` implementations differ in their handling of missing data. In the `WEKA` implementation, the fractional instances method is used by default. Using the fractional instances method, when a feature is used to split instances, any instances with missing features are sent to all child nodes, but weighted at each node according to the proportion of the number of instances at that node without a missing value and the total number of instances with no missing value across all child nodes. In `R`, no default method is supplied, so missing values were simply replaced by the median value of the feature.

Finally, it is noted that although random forest predictions are normally considered binary, a confidence score can be derived by utilising the level of consensus between trees. A consensus of half (0.5) or more needs to be reached in order for a random forest to assign a label. Thus in the `WEKA` implementation each prediction is output with a score that is guaranteed to be ≥ 0.5. This can be mapped to a score s with range of $[-1, 1]$ simply by transforming a consensus score c by

$$s = \begin{cases} (c - 0.5) \times 2 & \text{if (label is positive prediction)}, \\ -(c - 0.5) \times 2, & \text{if (label is negative prediction)}. \end{cases}$$

Chapter 3

IntPred as a BCE predictor

In this chapter, the general protein-protein interface (PPI) method IntPred is introduced. Methods for mapping the patch-level predictions of IntPred to residue-level predictions are then evaluated. A method for preparing antigen structural data sets in then presented before IntPred is evaluated for its performance as a BCE predictor.

3.1 Introduction

A review by El-Manzalawy and Honavar (2010) suggested that B cell epitope prediction may be improved by the utilization of advances in protein-protein interface prediction. This suggestion is directly addressed in this thesis by utilising IntPred, a previously developed in-house general PPI predictor (Baresic, 2011).

The next section introduces the IntPred method. IntPred is a general PPI method because because both obligate and transient interfaces are considered. It has been trained and tested on structural data obtained from x-ray crystal structures deposited in the PDB and utilises the features described in section 3.2.1.

3.2 IntPred method

IntPred is a random forest classifier[1] trained on a data set of 18 425 PDB chains (Baresic, 2011). IntPred makes predictions on overlapping *patches* of protein surface that are generated for a PDB structure using the method described in section 2.3.2.1. For a given

[1] Note that IntPred corresponds to the RF*(ALL) model presented in Baresic (2011), which showed the best performance when compared to other models developed in the same study.

Table 3.1: Summary of IntPred features. See main text for description of how these features are calculated.

Feature	Description	Type
	sequence	
`prop`	propensity score	Continuous numeric
`hpho`	hydrophobicity	Continuous numeric
`blast`	BLAST conservation score	Continuous numeric
`fosta`	FOSTA conservation score	Continuous numeric
	structural	
`SS`	disulphide bonds	Continuous numeric
`Hb`	hydrogen bonds	Continuous numeric
`helix`	α-helix secondary Structure	Binary categorical
`sheet`	β-sheet secondary Structure	Binary categorical
`mix`	mixed secondary Structure	Binary categorical
`coil`	coil secondary Structure	Binary categorical
`pln`	planarity	Continuous numeric
`intf`	class attribute	Binary categorical

PDB chain and set of patches, a set of 11 features are determined for each patch. These features and how they are calculated will now be described.

3.2.1 Features

IntPred uses 11 features for learning and prediction — these are summarised in table 3.1. These features can be divided into sequence features, which are based on sequence information, and structural features that require some structural information. The distributions of the *residue* features on which these *patch* features are based were all found to differ significantly between interface and non-interface (Baresic, 2011). Note that any ASA or $rASA$ values mentioned in the following sections were calculated as described in 2.3.1.

3.2.2 Sequence features

The following features only take sequence-based properties into account[2]. As these features are based on residue scores, the score of a patch is simply the average of the scores of its residues.

[2] In order to calculate the score of a patch however, structural information is obviously needed to determine which residues constitute a patch

Table 3.2: Consensus hydrophobicity values, as calculated in Kyte and Doolittle (1982).

Residue	Hydrophobicity Value
Ile	4.5
Val	4.2
Leu	3.8
Phe	2.8
Cys	2.5
Met	1.9
Ala	1.8
Gly	-0.4
Thr	-0.7
Ser	-0.8
Trp	-0.9
Tyr	-1.3
Pro	-1.6
His	-3.2
Glu	-3.5
Asn	-3.5
Gln	-3.5
Asp	-3.5
Lys	-3.9
Arg	-4.5

3.2.2.1 Hydrophobicity

The hydrophobicity of a residue is simply its hydrophobicity consensus score. The Kyte and Doolittle consensus scale is used (Kyte and Doolittle, 1982), the values for which are shown in table 3.2.

3.2.2.2 Propensity

In order to calculate propensities, a set of non-interface surface (*surf*) residues have to be collected as well as a set of interface (*intf*) residues. For IntPred, these were collected from 87 738 chains[3] (Baresic, 2011). For residue type X, the interface fraction F_{intf} is calculated as

$$F_{intf}(X) = \frac{ASA_{total}(X)}{ASA_{total}(intf)} \tag{3.1}$$

where $ASA_{total}(X)$ is the total ASA for all residues of type X in the interface set and $ASA_{total}(intf)$ is the total ASA of all residue in the interface set.

[3] This is a larger set than the IntPred training set — IntPred was trained on the subset of this larger set that did not have missing FOSTA or BLAST values (see section 3.2.2.3)

Similarly, the surface fraction $F_{surf}(X)$ is calculated as

$$F_{surf}(X) = \frac{ASA_{total}(X)}{ASA_{total}(surf)} \tag{3.2}$$

where $ASA_{total}(X)$ is the total ASA for all residues of type X in the surface set and $ASA_{total}(surf)$ is the total ASA of all residues in the surface set.

With these fractions, the propensity of a residue i of type X is calculated as

$$Pr(i) = \left(ln \frac{F_{intf}(X)}{F_{surf}(X)} \right) \times \frac{ASA(i)}{\overline{ASA_{surf}}(X)} \tag{3.3}$$

where $ASA(i)$ is the solvent-accessible area of i and $\overline{ASA_{surf}}(X)$ is the average ASA for all residues of type X in the surface dataset.

A positive $Pr(X)$ indicates over-representation of residue type X in the interface set, while a negative $Pr(X)$ indicates an under-representation.

3.2.2.3 Conservation scores

For each patch, two conservation scores are calculated: a FOSTA score and a BLAST score. Each score is calculated on the basis of an alignment produced using the matches generated by each tool.

In order to calculate FOSTA scores, PDBWS (Martin, 2005) is used to determine an associated UniProtKB/SwissProt entry for a given PDB chain. The FOSTA resource (McMillan and Martin, 2008) is then used to find the family of functionally-equivalent homologues of which the entry is a member. If this family contains at least nine other members, then the family is aligned using `Muscle3.7` (Edgar, 2004) with default parameters.

A BLAST search (Altschul et al., 1990) with the sequence of the PDB chain is also undertaken, using default parameters. Matches containing any of the terms *putative*, *predicted* or *hypothetical* are discarded, as are matches with an E-value > 0.01. If a minimum of 10 sequence matches are retained, then up to 200 of the top hits (ranked by lowest E-value) are kept. `Muscle3.7` is then used with default parameters to align these matches.

Each alignment is used to calculate residue conservation scores using the `Valdar01` method implemented in `scorecons`, part of the bioptools package (Porter and Martin, 2015). For both conservation scores, the score of a patch is the average of the score of its residues.

3.2.3 Structural features

The following features require structural information in order to be calculated.

3.2.3.1 Averaged features

Similarly to the features presented in section 3.2.2, these features are calculated at the residue level and calculated for a patch by averaging the scores of its residues.

Intra-chain **disulphide bonds** are identified by using the `pdblistss` tool from the `Bioptools` package (Porter and Martin, 2015). `pdblistss` identifies disulphide bonds by searching for S_γ-pair distances of less than 2.25 Å. This distance measure if based upon the average disulphide S_γ distance determined by Hazes and Dijkstra (1988), with an additional 10% tolerance for structure inaccuracy. A residue is given a score of 1 if it forms a disulphide bond or 0 otherwise.

Hydrogen bonds are identified using the `pdbhbond` tool from the `Bioptools` package (Porter and Martin, 2015). `pdbhbond` identifies hydrogen bonds using the rules of Baker and Hubbard (1984) by finding hydrogen-donor/acceptor pairs with a distance ≤ 2.5 Å and an angle between the hydrogen-donor/acceptor between 90° and 180°. Additionally, if a hydrogen atom is not defined, a hydrogen bond is assigned when the donor-acceptor distance is ≤ 3.35 Å. A residue is given a score of 1 if it forms a hydrogen bond and 0 otherwise.

3.2.3.2 Secondary structure

Secondary structure is assigned to a residue using the `pdbsecstr` tool from the `Bioptools` package (Porter and Martin, 2015). `pdbsecstr` assigns secondary structure according to the method by Kabsch and Sander (1983).

The secondary structure assignment of a patch SS_p follows

$$SS_p = \begin{cases} H & \text{if } \alpha > 20\% \text{ and } \beta < 20\%, \\ E & \text{if } \alpha < 20\% \text{ and } \beta > 20\%, \\ EH & \text{if } \alpha > 20\% \text{ and } \beta > 20\%, \\ C & \text{if } \alpha \leq 20\% \text{ and } \beta \leq 20\% \end{cases}$$

where α and β are the percentages of residues assigned to α-helix and β-sheet respectively. Becase SS_p is a nominal value that has four possible values, it is converted into four binary attributes SS'_p following

$$SS'_p = \begin{cases} (1,0,0,0) & \text{if } SS = H, \\ (0,1,0,0) & \text{if } SS = E, \\ (0,0,1,0) & \text{if } SS = EH, \\ (0,0,0,1) & \text{if } SS = C \end{cases}$$

3.2.3.3 Planarity

Patch **planarity** is calculated by finding the root mean squared distance of all atoms of the patch from a plane of best fit. The plane of best fit corresponds to the first primary component calculated by PCA on all patch atoms.

3.2.3.4 Class attribute value

The class attribute value of a patch is calculated by assessing the fraction of its total $rASA$ that is contributed by residues that have been defined as interface residues. A residue i is defined as interface if the following holds

$$rASA_i^n - rASA_i^c \geq 10\% \tag{3.4}$$

where $rASA_i^n$ and $rASA_i^c$ are the non-complexed and complexed $rASA$ values of i respectively. The interface fraction $rASA^f$ for a patch with a set of residues r_p and subset of interface residues r_{intf} is calculated as

$$rASA^f = \frac{\sum\limits_{j \in r_{intf}} rASA_j^n}{\sum\limits_{i \in r_p} rASA_i^n} \tag{3.5}$$

Using a class label threshold t_l, a class attribute value C is assigned as

$$C = \begin{cases} I, & \text{if } rASA^f \geq t_l, \\ S & \text{if } rASA^f = 0, \\ U, & \text{otherwise.} \end{cases}$$

where $t_l = 0.5$ (the value of this threshold will be discussed later in the context of B cell epitope prediction).

The value U corresponds to *unlabelled* that is assigned to patches that are on the rim of the interface. Patches with with class attribute value U are excluded from training and testing at patch level to ensure that classification remains a binary problem, but are included during testing when patch predictions are mapped to residue predictions (see section 3.3).

3.2.4 IntPred performance

IntPred was previously evaluated by comparing its performance to five other predictors (Baresic, 2011). The benchmark dataset used contains 4204 chains from 1306 protein complexes, the structures of which had all been solved after the publication of the predictors. These structures were also excluded from the IntPred training set, ensuring independence between training and test sets for all predictors tested. All five of the previously published methods make residue-level predictions and were therefore tested on all surface residues, which were defined as having $rASA^n > 5\%$. In contrast, IntPred predicts at a patch level. These predictions could be considered residue-level if the prediction label of the patch was assigned to its central residue. However, this mapping would be incomplete for two reasons: i) only residues with $rASA^n > 25\%$ were chosen to be patch centres and ii) U-labelled patches were removed in training and testing, leaving only I- and S-labelled patches. The former point could be addressed by instead using all surface residues as patch centres, although this would require reprocessing of the training set and retraining. However, the latter point is more problematic, because in a real testing scenario, class value labels are unknown and it thus not possible to discard U-labelled patches prior to testing. In this case, U-patches will be predicted which will affect real performance. For example, if the user is interested in knowing the likely proportion of positively-predicted patches that are true positives, they may consult a benchmarked PPV. However, this proportion is likely to be over-optimistic because in the real-case scenario, some U-labelled patches are likely to be positively-predicted. Thus, the comparison between IntPred and five other methods shown in table 3.3 is useful as an indicator of performance, but should be treated with caution as it is

Table 3.3: Benchmarking of IntPred and other previously published general PPI methods, reproduced from Baresic (2011). ACC=accuracy, PREC=precision, SPEC=specificity, SENS=sensitivity, MCC=Matthew's correlation coefficient, F=F-measure. The highest and the lowest score in every column are shown in blue and red, respectively.

Method	ACC	PREC	SPEC	SENS	MCC	F
ProMate	0.780	0.401	0.987	0.031	0.058	0.057
PIER	0.754	0.511	0.932	0.214	0.207	0.302
SPPIDER	0.759	0.472	0.783	0.676	0.410	0.556
PINUP	0.772	0.459	0.927	0.220	0.199	0.298
meta-PPISP	0.755	0.499	0.902	0.300	0.245	0.375
IntPred*	0.771	0.803	0.922	0.522	0.500	0.633

* tested on I and S-labelled *patches* only.

not a like-for-like comparison. In section 3.3 the problem of patch-to-residue mapping is addressed.

3.3 Mapping predictions from patch to residue

In order to compare the performance of IntPred fairly to other residue-based methods, prediction labels for every surface residue must be produced. This requires a mapping from patch-level predictions to residue-level predictions. Because surface patches can overlap for a given structure, a residue can occur in multiple patches. It is therefore not possible to devolve a patch label to residue labels directly, because a decision must be made as to which label a residue will take. Therefore, a decision-based function must be defined to map predictions from patch to residue.

3.3.1 Mapping functions

Four decision-based mapping functions were defined to map the patch labels to residues. Given that a patch can be predicted as either I (interface) or S (non-interface surface), the following functions were applied to determine the label of residue r:

Centre If r is a patch centre, give it the predicted label of the corresponding patch. If r is not found as a patch centre (i.e. it is a surface residue with an $rASA \leq 25$) then label it S.

Minimal If r is found in any patch labelled I, give r the label I, otherwise label it S.

TABLE 3.4: Patch-to-residue mapping. Four decision methods were used to map patch labels to residue labels for surface residues of the benchmark general PPI dataset (see main text for details).

Method	Sens.	Spec.	PPV	FDR	FPR	MCC
Centre	0.41	0.92	0.56	0.44	0.08	0.37
Minimal	0.84	0.53	0.32	0.68	0.47	0.31
Vote	0.41	0.88	0.48	0.52	0.12	0.31
Score	0.30	0.92	0.51	0.49	0.08	0.28

Vote Identify all the patches that contain r. Give r the majority label from this set of patches.

Score Identify all the patches that contain r. If the mean prediction score of these patches is greater than 0, then label the residue I, otherwise label it S.

Note that it is possible for a surface residue not to occur in any surface patches generated from the parent structure. This occurs when a surface residue is not close enough to a patch centre residue. If this is the case, the residue is labelled S.

These four methods were tested on the benchmark PPI test set described in section 3.2.4. Note that, in contrast to patch-level prediction, U-labelled patches are kept in the test set. Table 3.4 shows their performances. The *Centre* method gives the best performance, scoring the highest MCC of 0.37. The *Minimal* and *Vote* methods perform second best, with MCCs of 0.30, while the *Score* method performs slightly worse (MCC = 0.28. Although *Minimal* and *Vote* have the same MCC, they have different sensitivities and specificities. *Minimal* tends to over-predict, leading to a high sensitivity and medium specificity (0.84 and 0.53), while *Vote* tends to under-predict, resulting in a low sensitivity and high specificity (0.41 and 0.88). *Centre* performs with equal specificity to *Score* but a better sensitivity of 0.41, in comparison to 0.30.

3.3.2 Patch to residue-mapping with smaller patches

The *Minimal* method was one of the second best-performing mapping methods, but had the highest FPR of all three methods. It was hypothesised that the FPR could be reduced by altering the size of the patch used for mapping, while preserving the size of patches used for prediction. Although it was determined that patches with a 14 Å radius were optimal for prediction at a patch-level (Baresic, 2011), it may be that this relatively large radius leads to the over-labelling of residues as I when the *Minimal* method is used. One way to reduce over-prediction is to reduce the size of the patch used to map from patch to residue; a reduced-size **mapping patch** will have fewer residues than

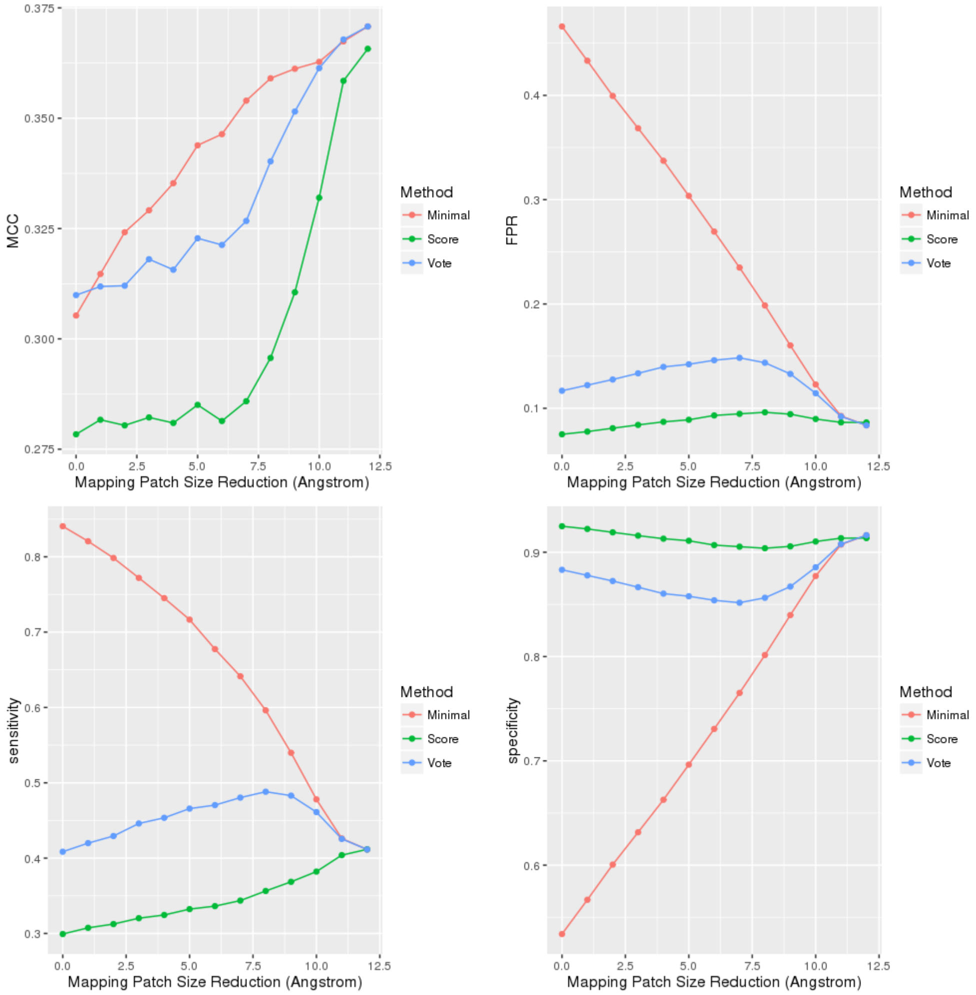

FIGURE 3.1: Reducing mapping patch size improves performance. These graphs show how performance changes when the size of the patch used to map predictions from patch to residue is reduced.

the original **prediction patch**, and so consequently I-labelled patches will devolve into fewer I-labelled residues, hopefully reducing the FPR.

The decision was made to use a series of smaller mapping patches in order to map from patch to residue for the *Minimal*, *Vote* and *Score* mapping methods. Because the *Centre* method only considers the patch centre residue it is not affected by the size of the patch used for mapping. However, it is shown shortly that the remaining three methods converge on the *Centre* method as mapping patch size is reduced.

Figure 3.1 shows the effect of reducing the patch mapping size on performance. All methods improve as mapping size reduces and that all methods perform best with a reduction of 12 Å. For the *Minimal* method, the FPR rates falls markedly from 0.47 to 0.08. Because the prediction patch radius is 14 Å, a reduction of 12 Å gives a mapping patch radius of 2 Å. At this radius, the mapping patch nearly always consists of the central residue only and thus all three methods are equivalent — or almost equivalent

— to the *Central* method (see table 3.4). Thus, MCC for the three remaining methods never exceeds the *Central* method. Note that there is a small variation in performance seen between methods when the radius reduction is 12 Å — this variation is due to five patches in the set that have two residues (rather than one) when their mapping radius is reduced by 12 Å.

3.4 Creating a structural epitope dataset

In order to test the performance of IntPred as a B cell epitope predictor, a test set of antigen structures with labelled epitope patches was required. This section presents a method for creating such a data set and uses it to define a test set that is used in section 3.5. Note that this method is the same that is used in chapter 4 to define additional antigen data sets.

Traditionally, homology within antigen structure datasets is controlled by setting a maximum sequence-identity cut-off between members. These sequence-identity cut-off controls lead to structures representing alternative epitope sites on the same antigen being omitted from the final set, resulting in epitope residues being incorrectly labelled non-epitope. To avoid this problem, a process was developed for identifying epitope residues amongst clusters of high-sequence identity chains which were then used these to label residues of a representative structure for each cluster.

3.4.1 Antibody-antigen complex identification

First, antibody-antigen complexes in the PDB had to be identified. Then, for each complex, the paratope residues of the antibody had to be identified in order to determine the epitope residues of the antigen. The following method was used:

1. A list of PDB files containing antibody structures was obtained from the SACS database (Allcorn and Martin, 2002).
2. Each PDB file was processed to identify light and heavy-chain pairs. Heavy and light chains were identified using the `idabchain` program (Allcorn and Martin, 2002). Heavy and light chains were then numbered using the `AbNum` program (Abhinandan and Martin, 2008), using the Chothia numbering scheme. CDR residues were then identified according to the Chothia numbering scheme. Heavy-light chain pairs were then identified by finding the heavy-light pairs with the largest number of inter-chain atom contacts, where contact is defined a distance of less than 4 Å. Note that a chain can only belong to one pair.

3. For each heavy-light chain pair, antigen chains were identified from the remaining chains in the PDB file. Any chains that had been identified as an antibody chain were ignored, as well those chains with a sequence length < 30. Any remaining chains in the PDB file that have at least one non-hydrogen atom less than 4 Å from any non-hydrogen CDR atom are defined as antigen chains for the antibody.

4. For each antibody-antigen complex, epitope residues are identified. Epitope residues are defined as any residue having a non-hydrogen atom < 4 Å from any non-hydrogen antibody atom and < 16 Å from any non-hydrogen CDR atom. This is in order to capture antigen residues in direct contact with CDR residues, as well as any residues that are also contacting the antibody and are part of the epitope-paratope interface due to their proximity to the CDR residues.

Note that most single chain antibodies such as VH and scFv antibodies were identified and processed, with the exception of some single-chain Fv fragments. This exception occurred when single-chain Fv fragments had been deposited in the PDB with a single chain identifier for the entire fragment. Currently `AbNum` is not able to number chains containing two variable regions and it is therefore not possible to identify their CDR residues. In the cases where a single-chain Fv fragment did have two chain identifiers, it was possible to identify their CDR residues and therefore process them.

This method allowed all antibody complexes in the PDB to be identified. This set can then be filtered to obtain a data set suitable for testing or training of a BCE prediction method.

3.4.2 Antigen clustering and surface patch creation

Antigen sequences must be clustered in order to identify groups of high-sequence identity chains that may share the same, overlapping or completely distinct epitopes. For each cluster, a representative member can be picked and its residues labelled using the epitope labels from all of the members, therefore minimizing the number of real epitope residues mislabelled as non-epitope.

The following method was applied to all antibody-antigen complexes with a resolution better than 3 Å identified using the method described in section 3.4.1

1. Antigen was clustered using cd-hit (Li and Godzik, 2006) using a sequence-identity cut-off of 90%.

2. For each cluster, members were aligned by sequence using Clustal Omega (Sievers et al., 2011). Each residue of each member was assigned an alignment sequence number that corresponded to its multiple sequence alignment position. The member which covered the largest sequence range across the alignment was chosen as the representative member. If more than one member covered the largest sequence range, the member with the best resolution was chosen. Any surface residue that was not labelled as epitope was labelled non-epitope.

3. For each representative member, epitope residue labels were assigned in addition to the residues that were already labelled as epitope in the representative. A representative member residue was labelled epitope if a residue with the same alignment sequence number in the group was labelled as epitope and was also the same amino acid type.

4. Surface patches were created for all representative members. Analogous to interface surface fraction (see equation 3.5) the epitope fraction $rASA_e^f$ for a patch with a set of residues r_p and subset of epitope residues r_{epi} was calculated as

$$rASA_e^f = \frac{\sum\limits_{j \in r_{epi}} rASA_j^n}{\sum\limits_{i \in r_p} rASA_i^n} \tag{3.6}$$

   Using a class value threshold t_l, class attribute value C was assigned as

$$C = \begin{cases} E, & \text{if } rASA_e^f \geq t_l, \\ N, & \text{if } rASA_e^f = 0.0, \\ U, & \text{otherwise.} \end{cases}$$

   where $t = 0.5$.

In the case of patch-based prediction, U-labelled patches were excluded from training and testing. In the following sections, different class value thresholds (t_l) will be investigated. It is important to reiterate that t_l controls the make-up of any final dataset. If t_l is increased then there will be fewer epitope patches and a greater number of U-labelled patches that are excluded from training and testing.

3.4.3 Feature calculation

Feature calculations for surface patches were undertaken as described in section 3.2.1, with the exception of class attribute values (described in section 3.4.2) and propensity scores. Propensity scores are calculated in the same manner but using epitope and

TABLE 3.5: IntPred performance on general PPIs and epitopes. IntPred was tested on the benchmark general PPI dataset and a test antigen set, the creation of which is described in section 3.4.

Test Set	Sensitivity	Specificity	PPV	MCC
Interface Test Set	0.51	0.92	0.81	0.50
Antigen Test Set	0.36	0.62	0.12	0.00

non-epitope residue sets which are analogous to the interface and surface sets defined in section 3.2.2.2. The exception to this is the testing presented in section 3.5, where propensity scores were calculated using the IntPred interface and surface sets, but then applied to a test set of antigen.

3.5 IntPred as a B cell epitope predictor

The performance of IntPred as B cell epitope predictor was tested by using IntPred with no amendments to the method described in section 3.2. In order to compare performance, all antibody-antigen complexes were removed from the general PPI benchmark test used to test IntPred previously (see section 3.2.4). This modified set was then run on IntPred, along with the antigen set described in section 3.4. In order to compare performance, prediction score posterior probability distributions were calculated for each class (surface or interface in the case of the general PPI benchmark test set; epitope or non-epitope in the case of our antigen test set). For a class C, the prediction score posterior probability for a score s is calculated as:

$$P(s|C) = \frac{P(C|s)}{P(C)} \tag{3.7}$$

Figure 3.2 shows the prediction score posterior probabilities for IntPred when tested on the antigen set, in comparison to the modified general PPI benchmark test set. The graph shows that while interface and surface score distributions differ markedly, the distributions for epitope and non-epitope are virtually identical. This indicates that, when trained on a set of general PPIs, IntPred is unable to distinguish between epitope and non-epitope — this is confirmed by the performance statistics shown in table 3.5.

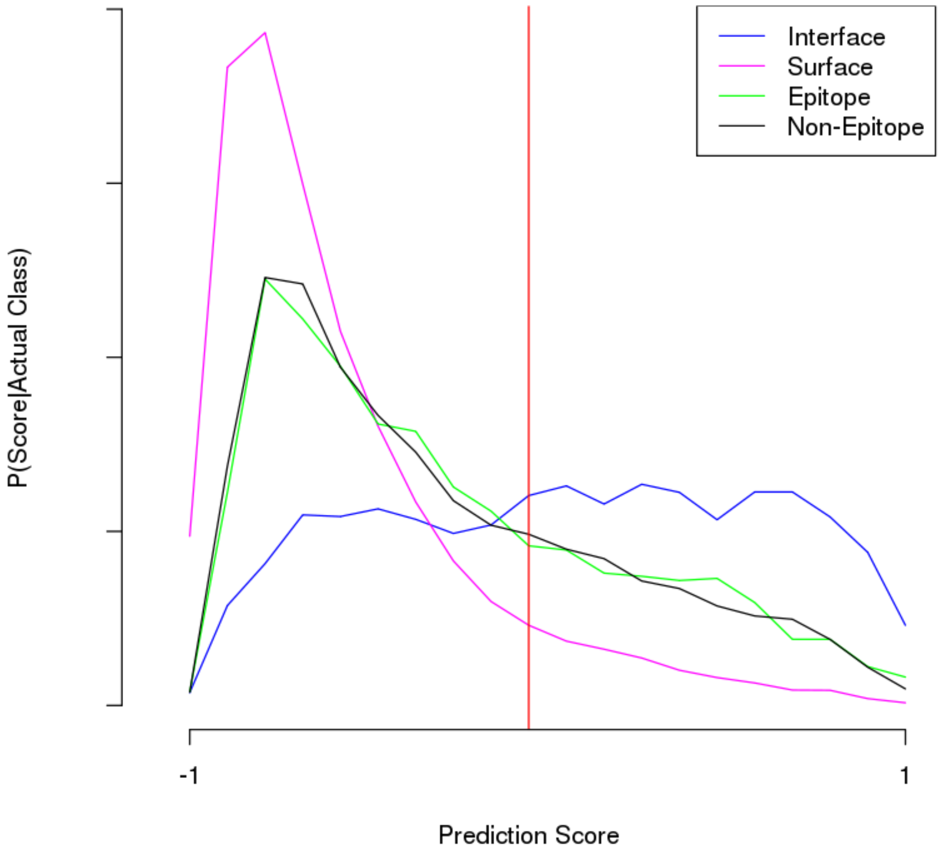

Figure 3.2: Prediction score posterior probability distributions of IntPred on general PPI and antigen test sets. The red line marks the surface/interface and non-epitope/epitope label threshold.

3.6 Discussion

As shown in previous studies (Jones and Thornton, 1997, Neuvirth et al., 2004), there are general similarities between other protein-protein interfaces and antibody-antigen interactions, such as an enrichment in tryptophan and tyrosine and a preference for unorganised secondary structure. However, overall residue distributions do differ significantly, with hydrophobic residues in particular being depleted, in comparison to with their enrichment in PPIs (Rubinstein et al., 2008, Krawczyk et al., 2013). The other notable difference is the extent of evolutionary conservation: PPIs tend to be highly conserved in comparison to the rest of the surface (Neuvirth et al., 2004) , whilst epitopes show a lack of conservation (Rubinstein et al., 2008). Baresic (2011) showed that IntPred relies most heavily on amino acids *propensities*, planarity and BLAST-based *conservation* — thus, it is perhaps unsurprising that IntPred is unable to predict BCEs, shown to differ markedly in two out of three of these features. Additionally, ProMate (Neuvirth et al., 2004), another structure based general PPI predictor, was tested on a data set of structural BCEs and, similarly to IntPred, was found to predict no better than random (Ponomarenko and Bourne, 2007).

Nevertheless, despite some properties differing markedly between PPIs and epitopes, it is clear that the *types* of features that are important for describing one are important in describing the other (e.g. amino acid propensity, evolutionary conservation). For that reason, it should be possible to retrain IntPred on a training set of antigen structures and their epitopes, using the same features as described in this chapter, in order to improve performance — this is carried out in the next chapter.

Ideally, the best-performing PPI predictor would be applied to BCE prediction. As shown in table 3.3, it was indicated that IntPred showed superior performance when a comparison was made between the patch-level predictions of IntPred and the residue-level predictions of existing methods. However, because U-labelled patches were discarded during testing, not all patches were tested on and thus the performance shown is not representative of performance in a real-case scenario, where patch labels are not known. In order to compare the methods fairly, patch-level predictions were mapped to residue-level predictions. In comparison to the performance stated in table 3.3, performance is not as good at the residue-level (MCC 0.5, compared with 0.37 for the best-performing mapping method). This means that in comparison to other methods, IntPred outperforms all methods except SSPIDER (MCC 0.41). SSPIDER (Porollo and Meller, 2007) is a neural network that, similarly to IntPred, is trained on a combination of structural and sequence-based features including, in contrast to IntPred, a solvent-accessibility feature. Though IntPred is outperformed by SSPIDER, it is IntPred that will be amended in the next chapter for use as a BCE predictor. This is because IntPred is in-house and therefore amenable to development and extension.

It may be that the method of mapping from patch to residue can be improved in the future. For example, using the consensus score produced by the random forest as a confidence score may prove useful. Additionally, a clustering method is presented in section 4.4.1 that is applied in the context of BCE predictions; this could also be applied to PPI predictions in the future.

3.7 Conclusion

In this chapter, the general protein-protein interface prediction method IntPred was presented and also modified to allow the production of residue-level predictions. IntPred was then tested on a test set of antigen structures and their epitopes; it was found that

IntPred is not able to distinguish between epitope and the rest of the protein surface. Thus in the following chapter, the IntPred method will be amended in an attempt to improve performance.

Chapter 4

Amendment of IntPred for the prediction of BCEs

This chapter presents the work that was undertaken to improve performance of IntPred as a B cell epitope predictor, by retraining and amending the method. The amended method is called IntPred:Epi. An analysis of the retrained random forest is also performed, as well as benchmark comparison between IntPred:Epi and existing methods.

4.1 Introduction

In the previous chapter, it was shown that IntPred is unable to predict B cell epitopes. As discussed in section 3.6, this is likely due to the considerable differences between other protein-protein interfaces and antibody-antigen interfaces. However, the same sorts of features used to train IntPred have been shown to be important in distinguishing epitope from non-epitope surface (Rubinstein et al., 2008). It was therefore hypothesised that retraining IntPred on a set of antigen structures with labelled epitopes should improve performance.

4.2 Amendment of IntPred method

In order to retrain and amend IntPred, training and test datasets of antigen structures had to be created. This was followed by retraining, as well as the inclusion of additional features and retuning of some learning parameters.

Table 4.1: Antigen data sets used for training or testing (see text for details). The sets have 90% maximum sequence identity between representative members.

Data set	Clustered Chains	Representative Chains	Surface Residues	Epitope Residues
cAtr	1603	310	53631	5433
mLtr	68	47	7771	667
mLts	-	15	3225	222

4.2.1 Data Sets

In order to assess the performance of IntPred for the prediction of epitopes, three antigen data sets were created using the method described in section 3.4. Note that two of the sets are taken from Hu et al. (2014). In this paper, the intersection of the training sets of eight predictors is collected to create a training set. The same eight BCE predictors were also tested on an independent test set of 14 antigen. These two sets can be used for easy comparison between our method and the eight methods tested in the paper. The three sets are as follows:

cAtr: Complete antigen training set. Created from all PDB files known to contain antibody chains (obtained via SACS (Allcorn and Martin, 2002)), except those specified in *mLts* (see below).

mLtr: Meta-learner training set. Created from the PDB chains of the training set described in Hu et al. (2014).

mLts: Meta-learner test set. Created from the PDB chains of the test set described in Hu et al. (2014).

Using *mLtr* as a training set allows us compare our cross-validated performance to the other predictors tested in Hu et al. (2014). *mLtr* contains 47 representative antigen chains. Using *cAtr*, which is a much larger data set than *mLtr* (310 antigen chains), allows us to investigate the effect of extra training data on performance. Finally, *mLts* can be used as an independent test set to compare method performance, as it is the same set that is used to assess the performance of the predictors described in Hu et al. (2014). The size of each data set is shown in table 4.1.

4.2.2 Retraining IntPred on epitope datasets

As shown in section 3.5, the standard version of IntPred is unable to distinguish epitope from non-epitope. Better performance from the method was sought by altering the learning process. The method was first retrained on a series of antigen data sets.

Table 4.2: Initial 10-fold CV. This table shows the 10-fold CV performance of four learners, each trained on a different data set. The training set for each learner is referenced in its name (see section 4.2.1 for definitions). Learners trained on a balanced training subset are indicated with a *b* suffix. For the calculation of performance measures, epitope and non-epitope patches are treated as positive and negative cases respectively.

Learner	Sens.	Spec.	PPV	FDR	FPR	MCC
RF:cAtr	0.04	1.00	0.89	0.10	0.00	0.17
RF:cAtr-b	0.66	0.61	0.34	0.38	0.39	0.28
RF:mLtr	0.11	1.00	0.89	0.11	0.00	0.30
RF:mLtr-b	0.68	0.73	0.69	0.31	0.27	0.40

4.2.2.1 Training and initial cross validation

IntPred was retrained on four different training sets to create four methods. Note that method names are prefixed with RF to indicate the random forest learning algorithm that is used in the machine learning stage. The first learner *RF:cAtr* was trained on *cAtr* and the second learner *RF:mLtr* was trained on *mLtr*. Both training sets are highly imbalanced, so two additional learners *RF:cAtr-b* and *RF:mLtr-b* were trained on balanced versions of these data sets. For both balanced sets, balancing was done by keeping all epitope patches and sampling at random from the subset of non-epitope patches. For this initial step, only one balanced subset was created for each training set, in contrast to the work done later (see section 4.2.2.3). All four methods were then validated using 10-fold CV (see section 2.5.1). Note also that for the learners trained on balanced sets, the testing partition in each fold of the CV is also a balanced set. The results are shown in table 4.2. It can be seen that all four methods yield some performance in comparison to IntPred alone. The best prediction appears to be *RF:mLtr-b*, with an MCC = 0.40. The unbalanced predictors differ markedly in their prediction profile in comparison to their balanced counterparts. Both unbalanced learners have very low sensitivity but good PPVs, meaning that they mislabel the majority of epitope patches as non-epitope, but do tend to correctly label an epitope in the rare event that they label a patch as epitope. In contrast, both balanced learners have a reasonable level of sensitivity (~ 0.67) but lower PPVs. However, this trade-off results in better performance when compared by MCC (0.30, 0.17 and 0.40, 0.28 for the *cAtr* and *mLtr* sets respectively). However, this apparent improvement in performance could be due to the balanced nature of the testing partitions, rather than a real difference in performance. In order to obtain a real difference, the performance of these four methods must be compared on an independent test set.

Table 4.3: Initial testing on *mLts*. This table shows the performance of the four learners on the independent test set *mLts*. Each learner's performance should be compared to its 10-fold CV performance shown in table 4.2. For the calculation of performance measures, epitope and non-epitope patches are treated as positive and negative cases respectively.

Learner	Sens.	Spec.	PPV	FDR	FPR	MCC
RF:cAtr	0.00	1.00	0.00	1.00	0.00	-0.01
RF:cAtr-b	0.47	0.71	0.09	0.91	0.29	0.09
RF:mLtr	0.09	1.00	0.11	0.89	0.00	0.01
RF:mLtr-b	0.73	0.56	0.10	0.90	0.44	0.14

4.2.2.2 Initial testing

The four learners described in section 4.2.2.1 were then tested on *mLts*. The results are shown in table 4.3.

For all four methods, the performance on *mLts* is poor in comparison to CV performances. For the two methods trained on balanced data sets, this may be because the test partitions used within cross-validation are also balanced, and therefore not representative of performance on an imbalanced set such as *mLts*. However, this is not the case for the two methods trained on imbalanced sets. This problem is addressed in section 4.2.2.3.

It is also observed that *RF:mLtr-b* outperforms *RF:cAtr-b*. This is surprising, considering the difference in the amount of input data (48 and 681 representative chains). The is counter to the the general model of machine learning — that is, as the learner is supplied with more data, it is better able to generalise the problem and as a consequence should be able to improve performance.

4.2.2.3 By-chain cross-validation

The difference in cross-validated and independent test performance lead to the concern that the discrepancy between 10-fold CV and testing metrics for the four learners was arising because of overlap between training and test folds during cross-validation. Two surface patches from the same antigen can share some residues, if their central residues are close to each other. This means that within a set, patches may exist which have similar features because they overlap.

To address the overlap between training and test sets, the by-chain CV method was created. In **by-chain CV**, the test partition of each fold consists of patches from a single chain and the training partition contains no patches from that same chain. This avoids overlapping patches occurring in train and test partitions. By-chain CV also allows

Table 4.4: By-chain CV performance. This table shows the performance of the four learners using by-chain CV. Values shown are averages from 20 balanced training subsets. Using MCC as a comparison, the by-chain CV performance of each learner is closer to the performance seen on the *mlTs* test set (shown in table 4.3) than the performance seen using 10-fold CV (shown in table 4.2). For the calculation of performance measures, epitope and non-epitope patches are treated as positive and negative cases respectively.

Learner	Sens.	Spec.	PPV	FDR	FPR	MCC
RF:cAtr	0.00	1.00	0.27	0.73	0.00	0.02
RF:cAtr-b	0.58	0.53	0.14	0.86	0.47	0.07
RF:mLtr	0.00	1.00	0.00	1.00	0.00	-0.02
RF:mLtr-b	0.53	0.57	0.11	0.89	0.43	0.06

imbalanced test partitions when performing CV for a learner trained on a balanced training set, giving us a more realistic performance estimation. Additionally, using a balanced training set, by-chain CV can also create a user-specified number of random balanced subsets of the full training set. These will be used to train a set of learners, each trained with a different balanced subset. This allows us to investigate whether having different balanced subsets of the training set affects performance.

By-chain CV was run on the two training sets and their balanced counter parts. The results are shown in table 4.4. Firstly, it is observed that the by-chain CV metrics are closer to the metrics given by testing on *mLts*, confirming the idea that by-chain CV is more representative of real performance than patch-based 10-fold CV.

In comparison to the 10-fold CV results shown in table 4.2, the non-balanced learners have zero sensitivity. This suggests that the sensitivities exhibited for 10-fold CV were the result of overlapping patches being present in training and test partitions.

For the balanced learners, the results are different in nature to those from the initial 10-fold CV, because by-chain CV allows the use of full unbalanced test partitions. For both balanced learners, PPV falls dramatically. This suggests that the inclusion of many more non-epitope patches during testing leads to many mislabelling events of non-epitope as epitope. However, for both learners sensitivity drops. Sensitivity is a function of the true positives and false negatives and is therefore not affected by the inclusion of many more non-epitope patches. Similarly to the unbalanced learners then, the drop in sensitivity must be the result of not having overlapping patches in training and test partitions.

The effect of non-epitope sampling on the results of by-chain CV was also investigated. Figure 4.1 shows the distribution of MCC scores from by-chain CV on *RF:mLtr-b* with 100 training subsets. The figure shows that the balanced subset can affect performance significantly; the average MCC score is 0.078, but scores range from 0.030 to to 0.110, depending on the training subset used to train the learner.

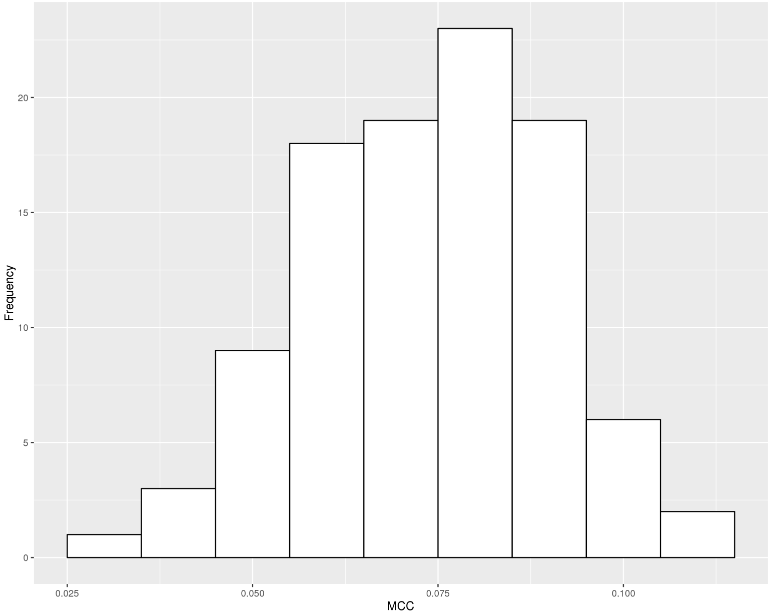

Figure 4.1: Performance variation owing to training data subset selection for balancing. This graph shows the by-chain CV MCC score distribution for the *RF:mLtr-b* learner trained on 100 balanced training subsets. Bin width = 0.01

4.2.2.4 Individual chain performance

Notably, there is a discrepancy between by-chain CV and test performance using *RF:mLtr-b* (MCC 0.06 and 0.11, see tables 4.4 and 4.3). The performance of the predictor based on each chain in the training and test sets was investigated to see if this would provide further insight. Looking at performance on each chain allows us to investigate the variance in prediction: is performance on each chain close to the average performance shown in tables 4.4 and 4.3, or does performance vary significantly for each chain? Predictions for the *RF:mLtr-b* learner were split by chain and an MCC score was calculated for each. The results are shown in figure 4.2. There is a large variation in performance when evaluated on a per-chain basis: 16 chains have an average MCC < 0 and 57 chains have an MCC > 0. The highest MCC score is 0.83 and the lowest is -0.62. This figure also further illustrates how the training subset selection can influence performance. For example, the performance on chain 1fnsA varies from -0.19 to 0.25, according to the training subset.

4.2.3 Method alterations

As well as retraining, a number of alterations were made to the method to try and improve performance. This included adding new ASA-based features, as well as investigating the effect of patch radius and training set class label distribution.

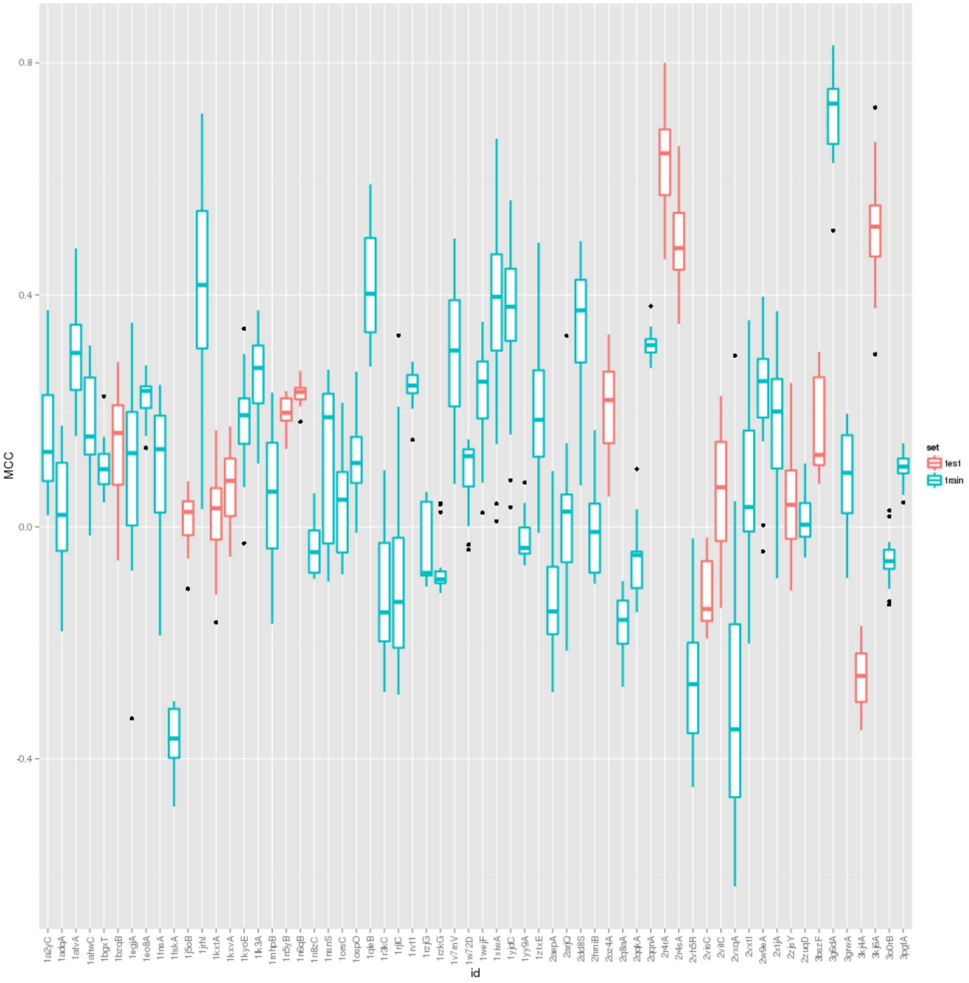

Figure 4.2: *RF:mLtr-b* performance on training and test set chains. *RF:mLtr-b* was trained on 20 balanced training subsets to create 20 learners. Each learner was then used to make predictions on each chain. For each chain, box and whisker plots show MCC score variability due to the balanced training subset used for training the learner.

4.2.3.1 Patch radius

Previously for general protein-protein interfaces, it was determined that 14 Å radius patches gave the best performance. I wanted to retest patch size to see if the same was true for epitope prediction. By-chain CV was performed on *mLtr-b* with a series of patch radii ($8, 9, \cdots, 14$ Å). Figure 4.3 shows that no other patch radius improves over the original of 14 Å.

4.2.3.2 ASA features

Solvent accessibility has previously been cited as a distinguishing feature of epitope and non-epitope surface (Rubinstein et al., 2008). Two ASA features were tested using by-chain CV to see if they improved performance. The first is the absolute solvent accessible

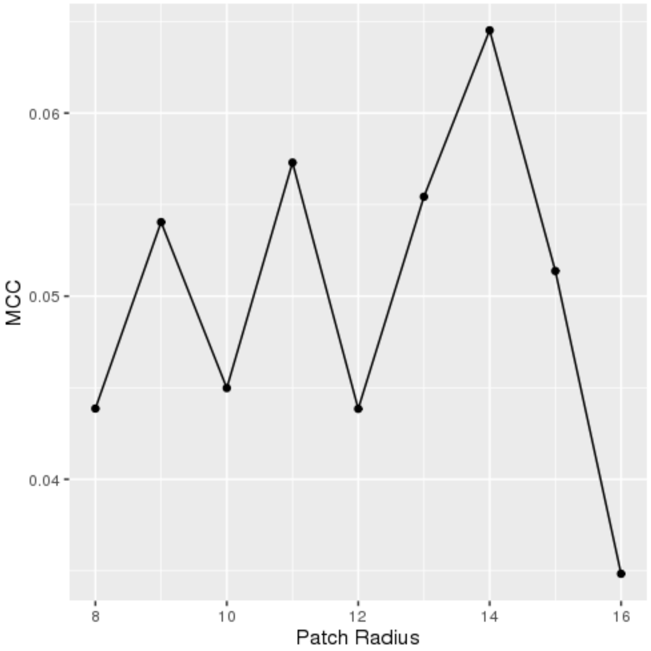

FIGURE 4.3: The effect of patch radius on performance. by-chain CV was run on *RF:mLtr-b* with a series of patch radii. As observed with general PPIs (Baresic, 2011), patch radius 14 Å gives the best performance.

TABLE 4.5: The effect of ASA features on performance. The effect of using solvent-accessibility features was investigated by including two features: absolute solvent-accessibility (aASA) and relative solvent-accessibility (rASA). These features were added into the by-chain CV process for *RF:mLtr-b* learner (see section 4.2), whose performance is included here as a baseline.

Learner	Sens.	Spec.	PPV	FDR	FPR	MCC
RF:mLtr-b	0.53	0.57	0.11	0.89	0.43	0.061
RF:mLtr-b + aASA	0.53	0.60	0.12	0.88	0.40	0.076
RF:mLtr-b + rASA	0.54	0.60	0.12	0.88	0.40	0.084

area of the patch (aASA). The second is the relative solvent-accessible area (rASA). Table 4.5 shows the results of including each feature separately. The inclusion of either solvent-accessibility feature improves specificity. rASA improves performance more than aASA by also very slightly improving sensitivity. This leads to an increase in MCC of 0.02 over the baseline learner *RF:mLtr-b*.

4.2.3.3 Balancing

The initial by-chain CV had shown that balancing of the training set at a 1:1 epitope:non-epitope ratio improved performance (see section 4.2.2.3) in comparison to unbalanced training. I decided to investigate the effect of training set class label distribution on performance by running by-chain CV on series of training sets with varying distributions. The graphs in figure 4.4 show that as epitope fraction increases, sensitivity increases

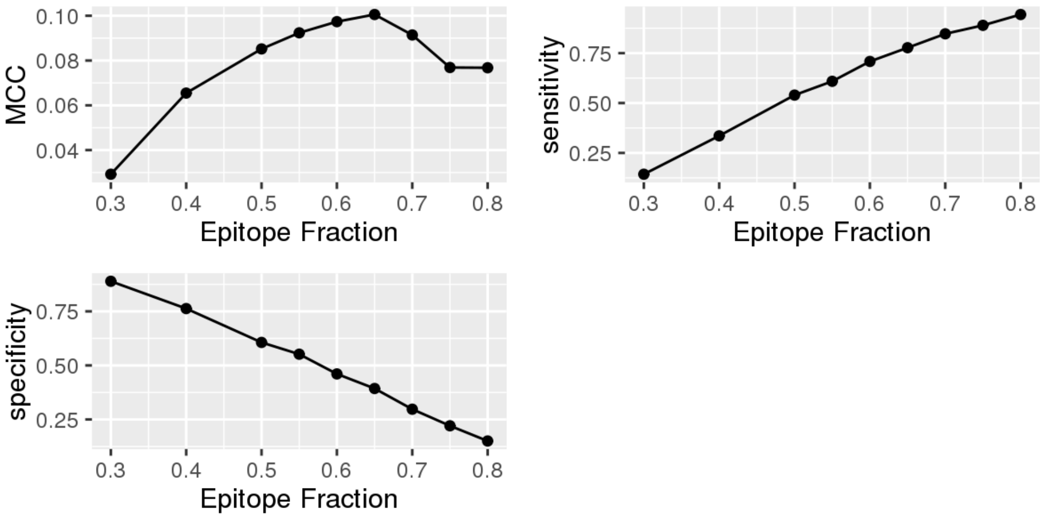

Figure 4.4: The effect of training set class label distribution on by-chain CV performance. As the fraction of training set labelled as epitope increases, sensitivity increases linearly while specificity decreases linearly, leading to an improvement in MCC that peaks at epitope fraction = 0.65.

linearly, while specificity falls linearly. An epitope fraction of 0.65 gives the best performance, with an MCC of 0.10. In comparison to the default balancing of 0.5, this is an improvement in MCC of 0.02.

Table 4.6: Correlations between features of *mLtr* set instances. For definition and description of these features, see table 3.1 and section 3.2.1. Note that this analysis was only done on the original IntPred features and does not include $rASA$ or ASA features (see section 4.2.3.2).

	`prop`	`hpho`	`pln`	`helix`	`mix`	`sheet`	`coil`	`SS`	`Hb`	`fosta`
`prop`	-	-	-	-	-	-	-	-	-	-
`hpho`	-0.72	-	-	-	-	-	-	-	-	-
`pln`	0.11	-0.03	-	-	-	-	-	-	-	-
`helix`	-0.10	0.15	-0.03	-	-	-	-	-	-	-
`mix`	0.07	-0.02	0.09	-0.20	-	-	-	-	-	-
`sheet`	0.08	-0.07	0.04	-0.60	-0.26	-	-	-	-	-
`coil`	-0.03	-0.07	-0.09	-0.30	-0.13	-0.39	-	-	-	-
`SS`	-0.21	0.12	0.10	-0.20	-0.08	0.11	0.15	-	-	-
`Hb`	0.11	0.00	-0.07	0.32	0.09	-0.10	-0.33	-0.19	-	-
`fosta`	0.04	-0.07	0.03	-0.23	0.06	0.21	-0.02	0.19	0.08	-
`blast`	-0.11	0.08	0.02	-0.03	-0.01	0.09	-0.09	0.07	0.14	0.35

4.3 Exploratory analysis of predictions

In order to gain more of an understanding of predictor performance, an exploratory analysis was carried out on the the features used, as well as an investigation into the random forest. For tis section, a random forest was trained using the `randomForest` implementation in `R` that was equivalent to the *RF:mLtr-b* learner described above.

4.3.1 Feature correlations

The decision was made to investigate the relationships between the features used for learning. Table 4.6 shows the correlations between features (see table 3.1 and section 3.2.1 for feature definitions). There is a strong correlation between `prop` and `hpho` (-0.72). This strong negative correlation indicates that the propensity statistic reflects a preference for hydrophilic amino acids within epitopes. Note than any correlations amongst `helix`, `mix`, `sheet` and `coil` must be disregarded, as these features are the result of creating four binary features from the original secondary structure feature, which was a nominal feature that could have one of four values (see 3.2.3.2); thus these features would be expected to correlate. Notably, `fosta` and `blast` features are only weakly correlated (0.35), despite both being evolutionary features.

Random forest methods are not significantly affected by feature correlations, but they are important when considering feature importance (see below).

4.3.2 Variable importance

A trained random forest is able to produce variable importance measures for each feature, which gives an indication of how important each feature is in helping distinguish between classes. One such measure is the mean decrease in Gini (MDG). This is calculated by summing and normalising the decrease in Gini index for each feature across all trees in the forest. The Gini index is a measure of class impurity in the nodes of a tree after using a chosen feature for splitting instances. It is calculated by summing the frequency of each class label multiplied by the frequency of a mistake in categorizing instances with the given class label. A higher mean decrease Gini indicates that of the feature is more effective for splitting instances by class. Figure 4.5 shows the mean decrease Gini for each feature. The two most important features are `prop` and `hpho`, which were shown to be anti-correlated in section 4.3.1. This correlation is likely to lead to an underestimation of variable importance for both features — this is because if two features are correlated, then the power of one feature to reduce Gini impurity subsequently will be reduced, as the training instances have already been split by the preceding correlated feature. The four secondary structure features make very little contribution to the prediction, as well as `SS` which only has a MDG of 10. Interestingly, `blast` seems to the more important of the two evolutionary features — `blast` has an MDG of 40 in comparison to `fosta` with just under 30. Similarly to `prop` and `hpho`, there is a correlation between `blast` and `fosta` (see table 4.6), though it is not as strong. This complicates the interpretation of the difference in MDG, but it is reasonable that BLAST conservation scores would be more useful as they are based on a larger alignment than FOSTA alignments, which only use functionally related matches.

4.3.3 Investigating proximity space

For a given pair of instances a and b that have been run through a random forest F consisting of T trees, the proximity pr between them can be calculated as

$$pr(a,b) = \sum_{t \in F} \frac{term(a,b,t)}{2T} \tag{4.1}$$

where

$$term(a,b,t) = \begin{cases} 1, & \text{if } tnode(a,t) = tnode(b,t) \\ 0, & \text{otherwise} \end{cases}$$

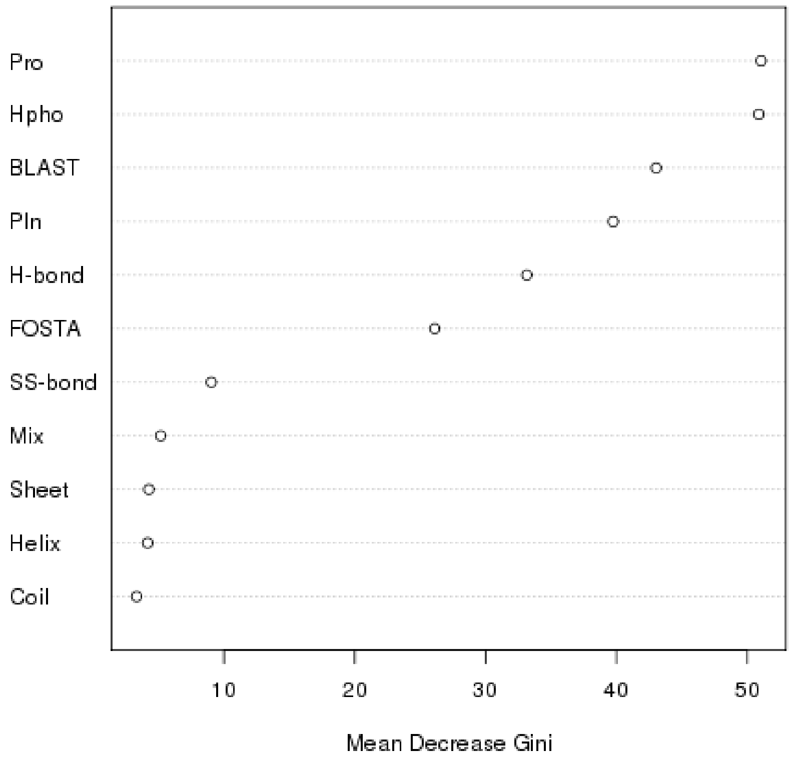

FIGURE 4.5: Variable importance. This graph shows the mean decrease Gini for each feature for the *RF:mLtr-b* learner. The mean decrease Gini is a measure of node impurity after splitting with each feature, across all trees in the random forest; the higher the mean decrease Gini, the more effective the feature is for separating instances by class.

$tnode(a, t) = tnode(b, t)$ if a and b are in the same terminal node of t.

These proximities therefore give us an idea as to what extent a random forest separates any pair of instances by giving us a distance between them. If a pair of instances have a small proximity value then the distance between them is small, so they are similar in the context of prediction. Conversely, if a pair of instances have a large proximity value, then the distance between them is large and they are less similar. Looking at the proximities between pairs can give us more of a understanding of how a random forest is separating instances. In particular, a matrix of proximities for a set of instances can be used as input for MDS. In its full form, a matrix of proximities will be n-dimensional, where n is the square of the number of instances. MDS seeks to reduce the number of dimensions while maintaining the distances between instances. MDS can be used to represent a set of proximities in two or three dimensions, thus allowing us to visualise the proximities. MDS was performed on the proximities of *mLtr*, run through the random forest trained on *mLtr-b*. Figure 4.6 shows the scaled proximities plotted in the first two component

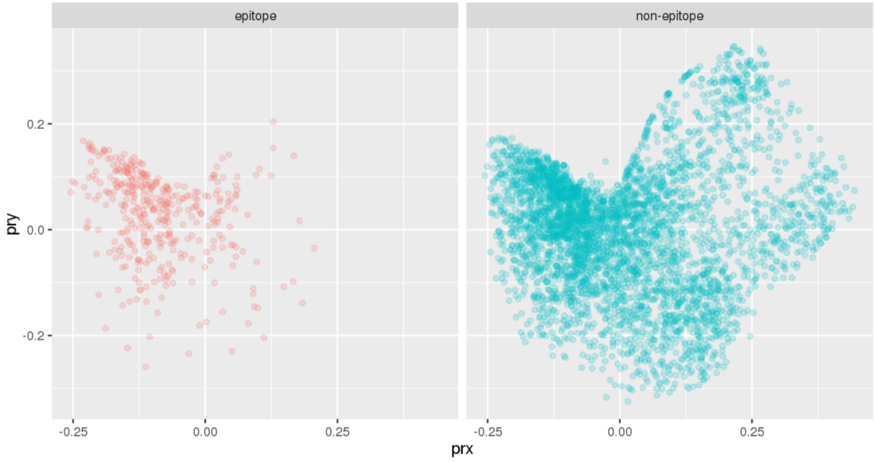

Figure 4.6: Epitope and non-epitope acrros the first and second MDS components. Proximities for the patches of the *mLtr* training set after being run through the *RF:mLtr-b* learner are plotted across the first two MDS components (prx and pry). Epitope and non-epitope instances are shown in red and blue respectively.

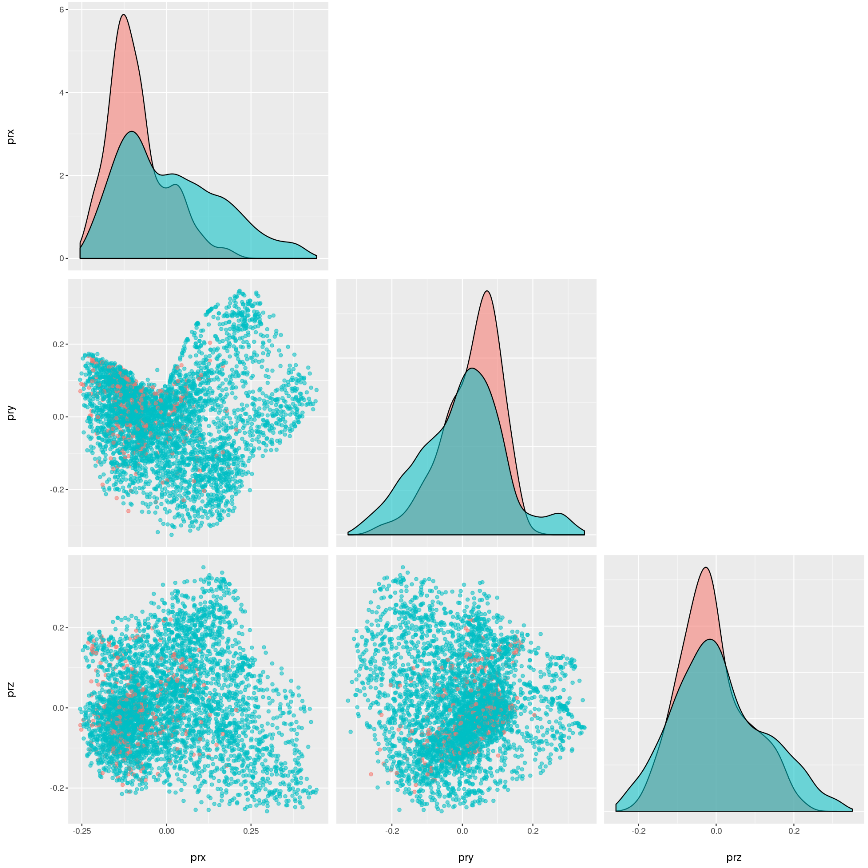

Figure 4.7: Proximities across the first three MDS components for the *mLtr* set. Epitope and non-epitope instances are shown in red and blue respectively.

dimensions prx and pry. The plot shows that epitope patches cluster together in the space. Non-epitope patches tend to fall in the same place, but their spread across the space is much wider along both axes. This is illustrated in the density distributions for each class across prx, pry and the third component prz, shown in figure 4.7. The modal prx value appears to be very similar for both classes, but non-epitope patches spread to create a distribution heavily skewed to higher prx. Similarly, the spread of non-epitope patches across pry is larger than for epitope patches, skewing towards lower pry and to some extent higher pry. The same skewing outwards of non-epitope in both directions can be seen in prz.

As shown in table 4.4, *RF:mLtr-b* has a high FDR combined with a fairly high FPR — this means that it tends to over-predict. The visualisation of the random forest predictions shown in figure 4.7 gives us insight into why this is: epitope patches appear to cluster together along the first three MDS components and therefore tend to end up in the same terminal nodes of the random forest. However, this cluster is also densely populated with non-epitope patches — this high overlap corresponds to many non-epitope patches being labelled epitope. But the plot also shows that in all three components, there are non-epitope patches that are distant from the epitope cluster.

The relationship between instances and their placement along the first three MDS components was investigated. Scaled components can be plotted against learning features in order to try and understand what is responsible for variance along an axis. The first component prx was found to correlate with `prop` and `hpho` respectively ($R^2 = -0.726$ and $R^2 = 0.657$, see figure 4.8). The second component, pry, was found to correlate with `pln` ($R^2 = -0.626$, see figure 4.9). The third component, prz, was found to correlate with `blast` score, although the correlation is weaker than the correlations with prx and pry ($R^2 = -0.479$, see figure 4.10). These correlations can be used to describe the distributions of epitope and non-epitope patch across the scaled space in terms of their features. In the prx-pry plot, the area of non-epitope patches at high prx, mid-to-high pry, corresponds to low propensity, non-planar patches. An area of high pry can be identified where epitope patches also do not occur — this would correspond to very planar patches. Looking at the prx-prz plot, it appears that two areas of non-epitope only occur; one at high prz and one at low prz. These would correspond to very low and very high blast conservation score respectively. Finally, the pry-plz plot shows an area of non-epitope at low pry, high prz. This corresponds to non-planar, low blast conservation scores.

The proximities for each chain can also be investigated. Figure 4.11 shows the the prx-pry space for each chain. The distribution of patches across the space can be seen to vary for each chain. Some chains, such as 1fnsA and 1nsnS, occupy small areas of the

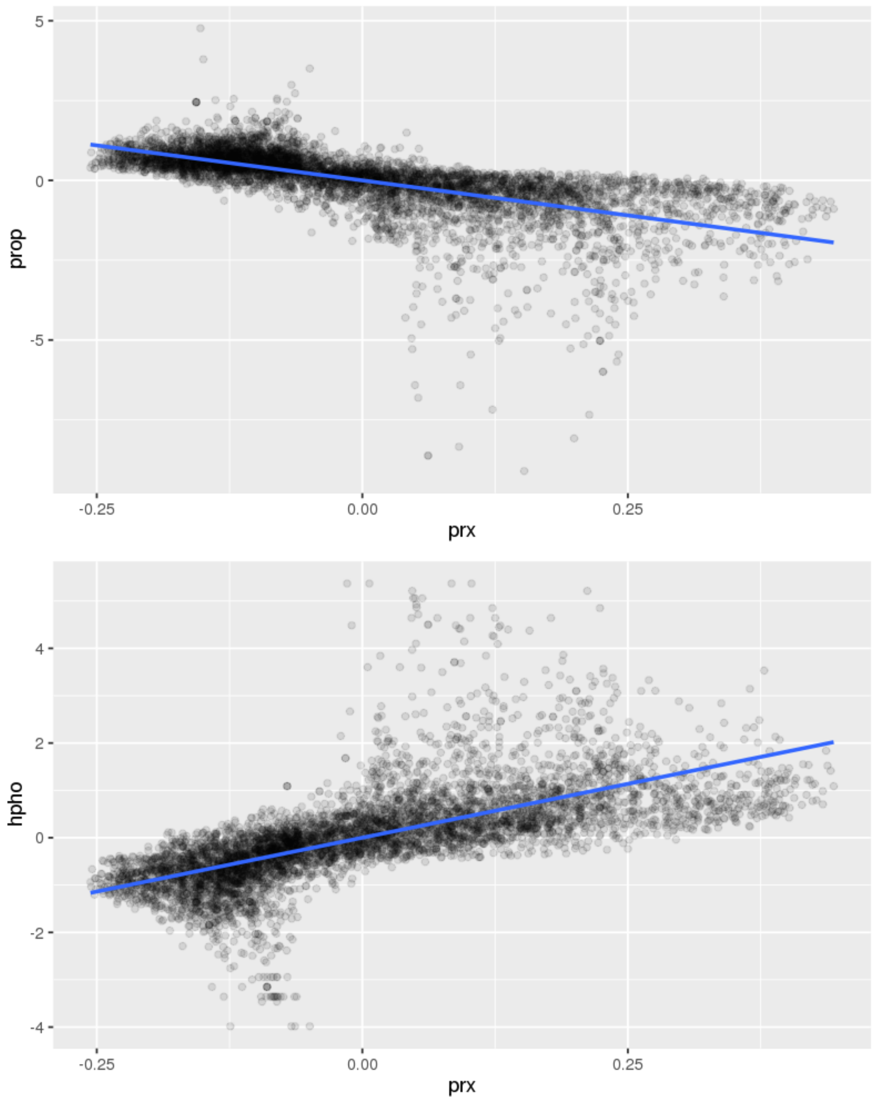

Figure 4.8: Propensity and hydrophobicity plotted against the first scaled component prx.

space; this is in contrast to 2dd8S, which occupies the entire space. All chains occupy the normal epitope space to a varying degree and in most cases their epitope patches occur in the epitope region ($prx < 0$, $prx > -0.1$). There are some exceptions, such as 1rzjG, 1rzkG, whose epitope patches are skewed away from epitope region, despite the presence of many patches in the epitope region. Interestingly, 1rzj:G and 1rzk:G are both structures of HIV envelope protein. The unusual distribution of patches may be explained by the fact that the majority of the envelope protein surface is covered by glycosylated carbohydrates; this may lead to the patches found in the epitope region being unavailable to the antibody.

Some chains are skewed towards the non-epitope space (top right corner: low propensity, high hydrophobicity, non-planar). One example is 2zuqD. This can be explained partly

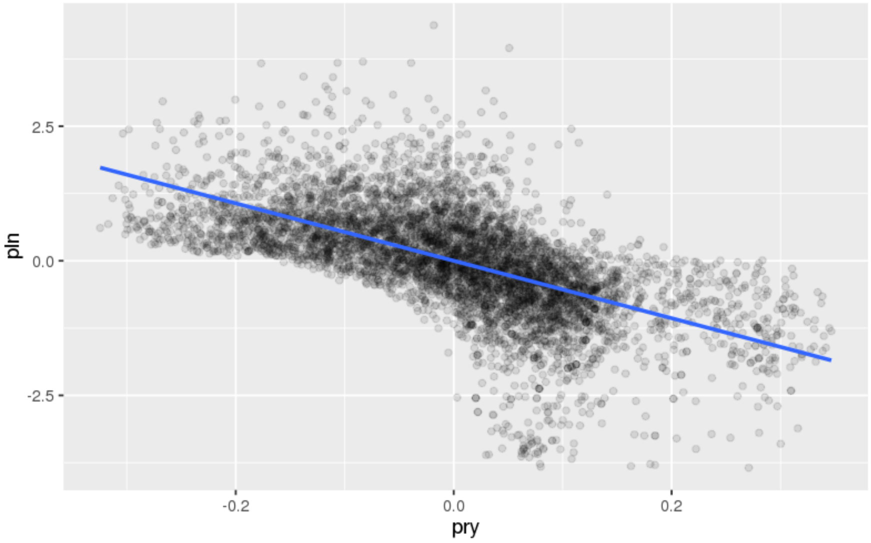

FIGURE 4.9: Planarity plotted against the second scaled component pry.

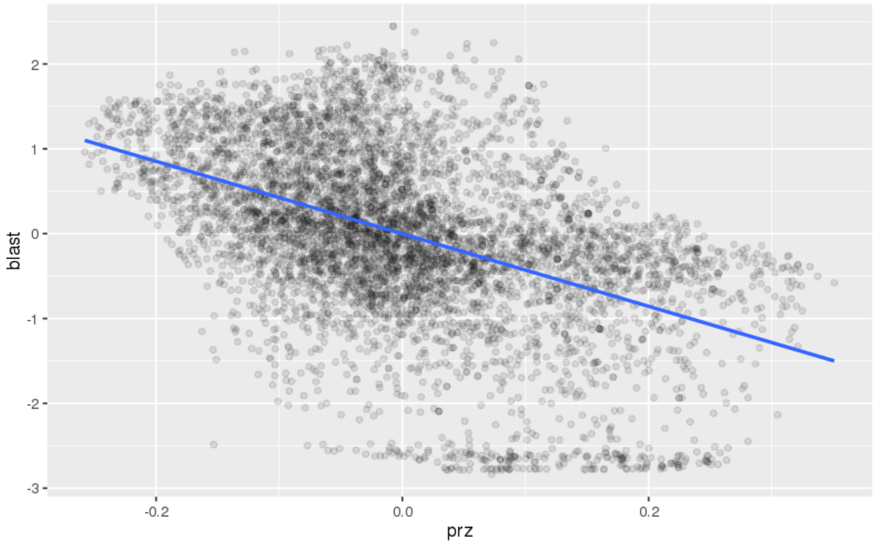

FIGURE 4.10: BLAST conservation scores plotted against the third scaled component prz.

by the fact that 2zuqD is disulfide bond formation protein B, a transmembrane protein that has to present a significant amount of hydrophobic surface in order to be integrated into the membrane.

4.3.4 Investigating outliers

Proximity matrices can be used to identify outliers for each class. For a set of instances of a class C, a non-normalized outlier score O can be calculated for each instance v_i as

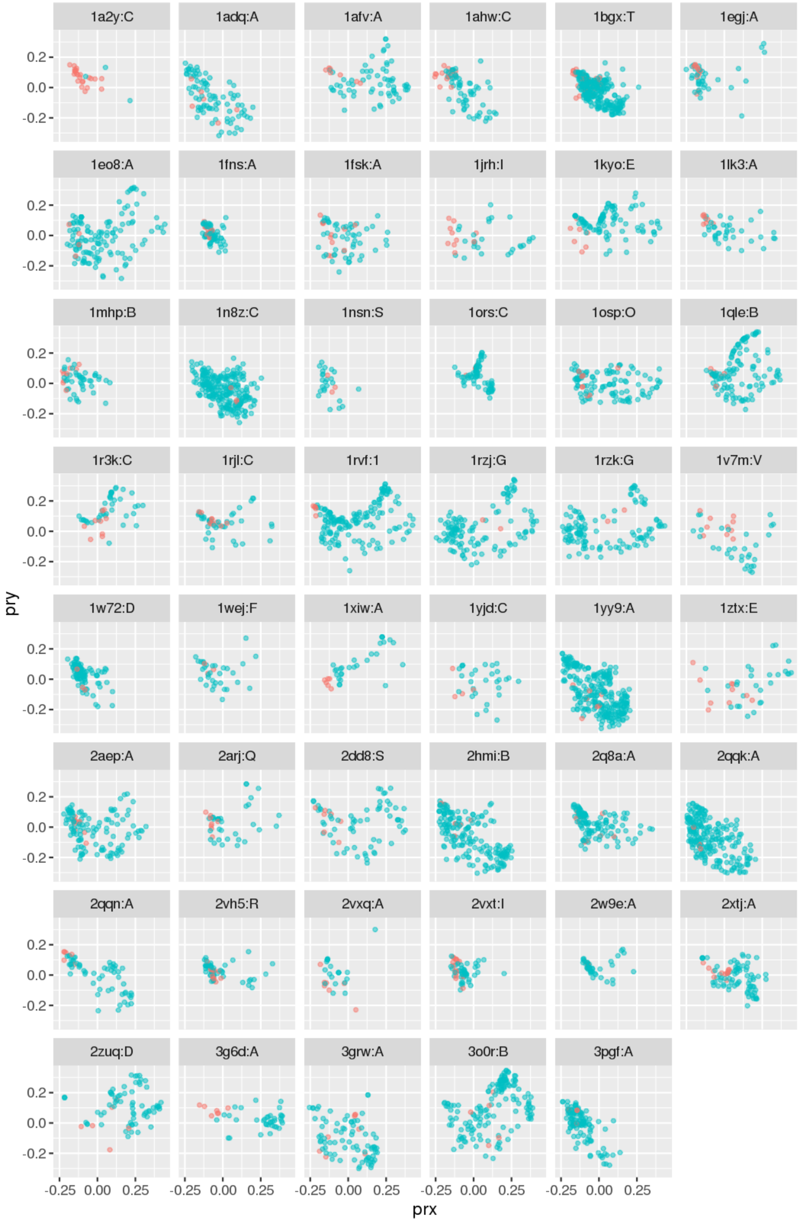

FIGURE 4.11: Patches from the *mLtr* set split by chain and plotted by the first two scaled components, prx and pry. Epitope and non-epitope patches shown in red and blue respectively.

$$O(v_i) = \frac{N}{\sum_{j=1}^{m} pr(v_i, v_j)^2} \tag{4.2}$$

where N is the total number of instances (from all classes) and m is the total number of instances of class C. The final normalised outlier score $\hat{O}$ for each instance v_i is calculated as

$$\hat{O}(v_i) = \frac{O(v_i) - \tilde{v}}{\tilde{d}} \tag{4.3}$$

where $\tilde{v}$ and $\tilde{d}$ are the median and the median absolute deviation of non-normalised outlier scores for the class C. $\tilde{d}$ is calculated as

$$\tilde{d} = \text{median}(|O(v_i) - \text{median}(O(v))|) \tag{4.4}$$

Intuitively, an outlier score gives a sense of how unusual an instance is, considering its class. A high outlier score indicates that an instance is unusual for its class in comparison to the majority of instances with the same class.

Figure 4.12 shows the absolute outlier scores for the epitope patches of *mLtr*, grouped by chain. The figure shows that for most chains, the majority of epitope patches have absolute outlier scores no greater than 2. This corresponds to what is observed when the epitope patches are plotted by the first two scaled components prx and pry, where they form a central cluster with a sparsely populated peripheral. Some chains have a small number of patches that have absolute outlier scores > 2, such as 1a2yC. The epitope patches of 2zuqD span a large range of outlier scores, with the mean at ~ 2.5. 1ztxE spans a similar range, but the mean lies at 0.5. 1rzkG and 1rzjG have some of the highest mean outlier scores, but also have small spreads.

A patch is labelled as epitope because it has a proportion of surface that is found bound to an antibody in a crystal structure, but it will likely have a remaining proportion of surface that is not found bound to an antibody. It was hypothesised that epitope patches with high outlier scores may tend to be those patches with a relatively low epitope surface area fraction $rASA_e^f$ (see equation 3.6 for how this is defined). It may be that some epitope patches have high outlier scores because they have a significant proportion of surface that is not in contact with the antibody that is distinctively 'non-epitope'. On the other hand, it may be that, considering the problem of defining a negative set (see section 1.2.5), the non-epitope surface is not much different from the epitope surface and therefore the patch does not have a high outlier score. Nevertheless,

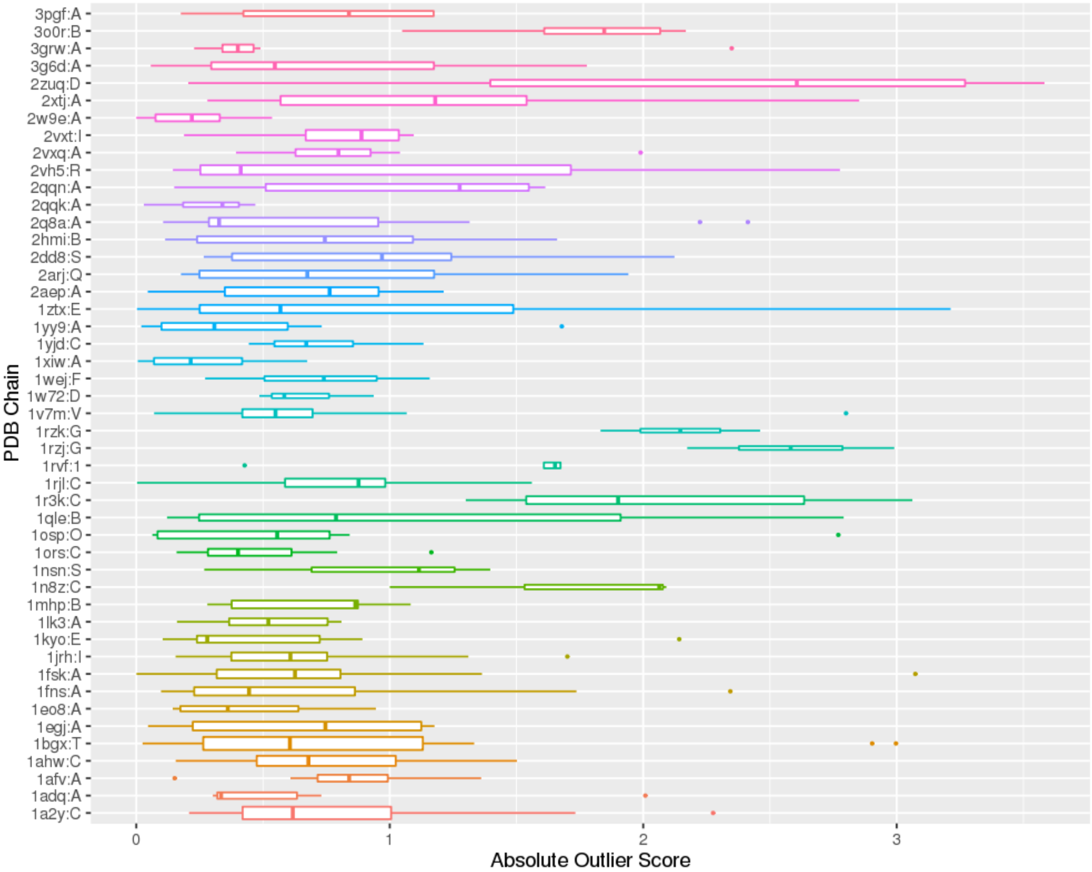

FIGURE 4.12: Absolute outlier scores of epitope patches from the *mLtr* set, grouped by chain.

it was hypothesised that there should be a weak, but significant, correlation between $rASA_e^f$ and the outlier score. To test this, a Spearman's rank correlation coefficient was calculated. As expected, this gave a weak, but significant, correlation ($\rho = -0.23, p = 2.46 \times 10^{-6}$).

With this correlation in mind, the decision was made to investigate those epitope patches with an absolute outlier score > 2. Figure 4.13 shows the outlier scores of these patches plotted against $rASA_e^f$. The figure shows that, for the majority of patches, the fraction bound < 0.6. As discussed above, these low-$rASA_e^f$ patches probably also have 'non-epitope'-like surface. However, for patches with higher $rASA_e^f$ (> 0.7), what could be leading to a higher outlier score? The decision was made to investigate these patches. These patches are from four chains: 1a2yC, 1fskA, 1r3kC and 2zuqD. The results of these investigations are discussed below.

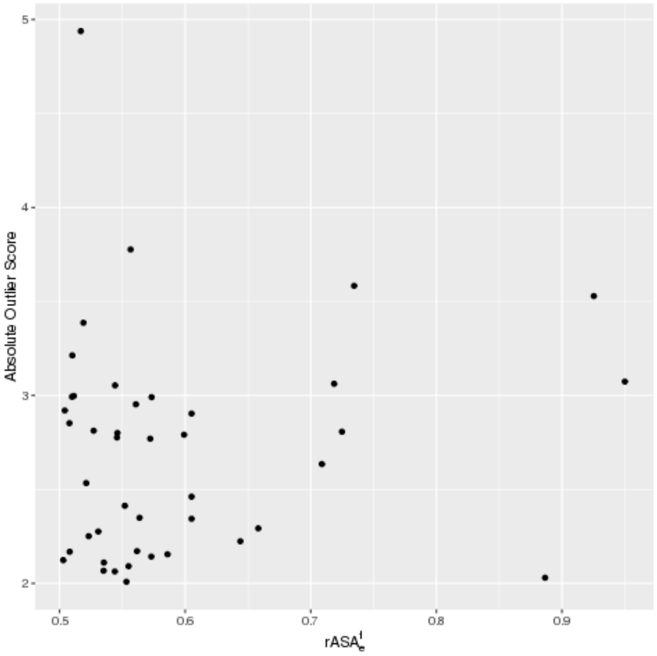

Figure 4.13: The absolute outlier scores of epitope patches with high absolute outlier scores (> 2) plotted against $rASA_e^f$. The majority of these patches have an $rASA_e^f$ < 0.6.

Table 4.7: Comparison between outlying epitope patches and their sub-patches (denoted by a *) that only contain epitope residues, where $|\hat{O}|$ is the absolute outlier score. The all-epitope sub-patch is less of an outlier in all three cases.

patchID	pro	hpho	pln	blast	$rASA_e^f$	$\|\hat{O}\|$
1a2y:C:63	-1.59	0.36	-1.31	0.30	0.72	2.75
1a2y:C:63*	-1.71	-1.09	-1.86	-0.28	1.00	0.93
1r3k:C:55	-0.67	0.41	0.05	-0.23	0.70	2.56
1r3k:C:55*	1.33	-0.80	-2.32	-0.62	1.00	1.41
2zuq:D:101	-0.41	0.84	-2.21	0.42	0.73	3.42
2zuq:D:101*	0.89	0.17	-2.52	0.46	1.00	1.18

4.3.4.1 All-epitope subpatches

As shown in table 4.7, patch 1a2y:C:63 has an outlier score of 2.75. It has low propensity (-1.59) and reasonable hydrophobicity (0.36). Figure 4.14 shows 1a2y:C:63 in relation to the cognate antibody. The patch contains residues that form the outer rim of the epitope, making contact primarily with CDRs H1 and H2. 1a2y:C:63 appears to represent a fraction of the epitope quite well. Further investigation of the patch revealed that it consists of nine residues; of these, five are labelled as epitope (TRP, ARG, ASN, LEU and ASN) and the remaining four are non-epitope (TRP, GLY, CYS, and ILE) (see figure

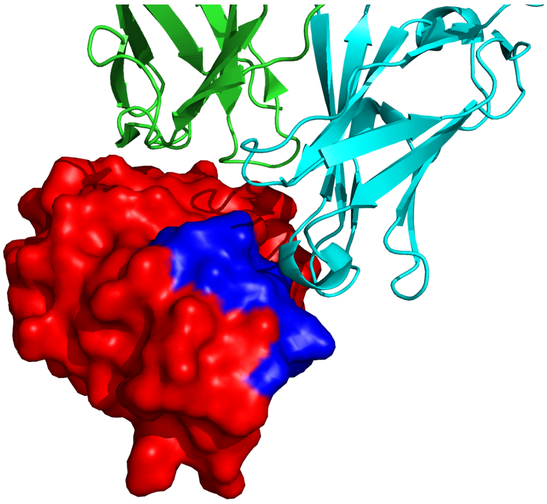

Figure 4.14: Patch 1a2y:C:63 (shown as blue surface) and the antibody heavy and light chains shown in green and cyan respectively.

4.15). It was hypothesised that the presence of the four hydrophobic non-epitope residues might be skewing the features of the patch. An 'all-epitope' sub-patch, consisting only of the epitope residues of the patch, was defined and its outlier score calculated. As shown in table 4.7, it was found that, when non-epitope residues were removed, the outlier score fell to 0.93.

This process was repeated for two more outlier patches, 1r3k:C:55 and 2zuq:D:101. Similarly to 1a2y:C:63, the all-epitope sub-patches of both patches had a reduced outlier score (see table 4.7).

These results support the hypothesis that outlier epitope patches result from non-epitope residues with a patch defined as epitope.

4.3.4.2 1fsk:A:48 and 1fsk:A:47

Patch 1fsk:A:48 (or '48') has an outlier score of 3.01. 48 is almost a complete subset of the larger patch, 1fsk:A:47 (or '47') as it contains only one residue that is not found in 47. 47 and 48 are both shown in figure 4.17. In contrast to 48, 47 has a low outlier score of 0.37 (see table 4.8). 47 includes all the interactions that 48 exhibits with the antibody as well as an interaction with a glutamate that is not captured by 48. Additionally, the residue in 48 that is not found in 47 is a valine that is distant from the antibody.

This example illustrates that it is possible for highly overlapping patches to have quite different outlier scores. In this case, the differences in the four most important features

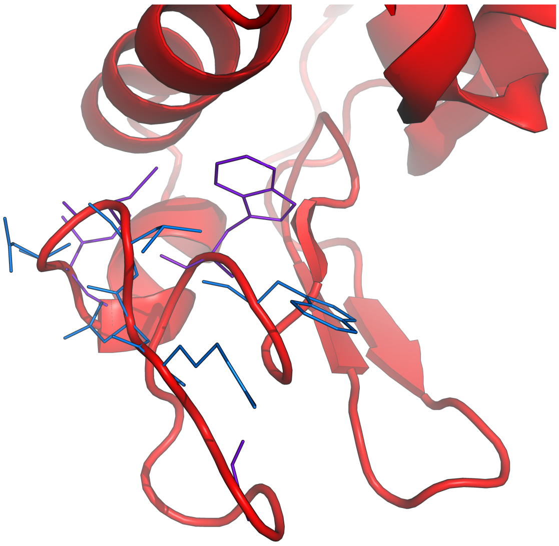

FIGURE 4.15: Patch 1a2y:C:63 residues. Epitope and non-epitope residues are shown in blue and purple respectively.

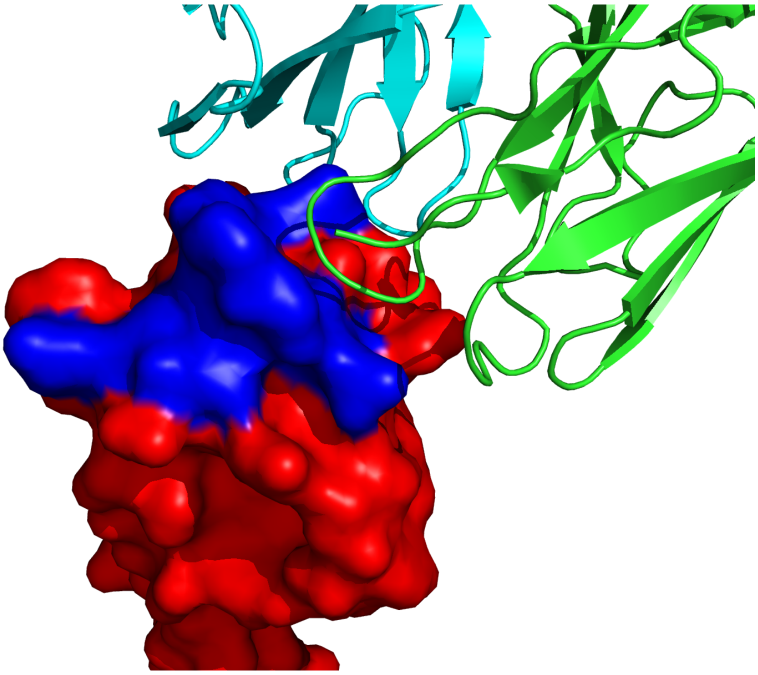

FIGURE 4.16: Patch 1r3k:C:55 (shown as blue surface) and antibody heavy and light chains (shown in green and cyan respectively).

TABLE 4.8: Comparison between outlier 1fsk:A:48 and the highly overlapping 1fsk:A47, where $|\hat{O}|$ is the absolute outlier score.

patchID	`pro`	`hpho`	`pln`	`blast`	$rASA^f_e$	$\|\hat{O}\|$
1fsk:A:47	0.36	0.09	-0.37	0.69	0.96	0.37
1fsk:A:48	0.04	0.16	-1.31	0.75	0.95	3.01

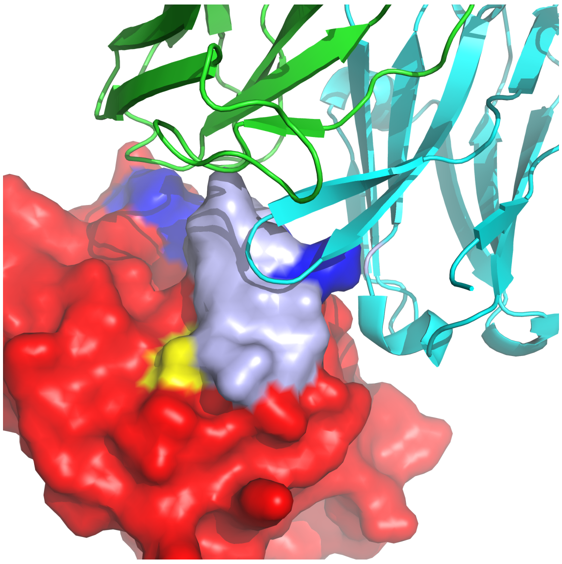

FIGURE 4.17: Outlier 1fsk:A:48 is an almost complete subset of 1fsk:A:47. The surface of 1fsk:A:47 is shown in dark blue; the subset contained within 1fsk:A:48 is shown in light blue; the remaining surface of 1fsk:A:48 that is not found in 1fsk:A:47 (a single valine) is shown in yellow. Antibody heavy and light chains are shown in green and cyan respectively.

seem to be quite small, with the exception of planarity (-0.37 in the outlier, -1.31 in the overlapping non-outlier)

4.3.5 Investigating epitope surface fractions

It was hypothesised that the inclusion of patches with relatively low $rASA^f_e$ (that is, epitope patches whose proportion of epitope surface to non-epitope surface is close to the class label threshold, t_l) may be affecting prediction performance. The benefit of including low $rASA^f_e$ patches is that all epitope residues within the patches labelled as epitope are likely to be included — if the $rASA^f_e$ threshold is higher, then there is risk of throwing away patches that are informative in the learning process. On the other hand, low $rASA^f_e$ patches can be thought of as noisy instances. This noise results from the inclusion of non-epitope residues that contribute to the features of the patch. Some patches may be noisier than others depending on the nature of the non-epitope residues it includes. If the non-epitope residues are in fact mislabelled epitope residues, then they

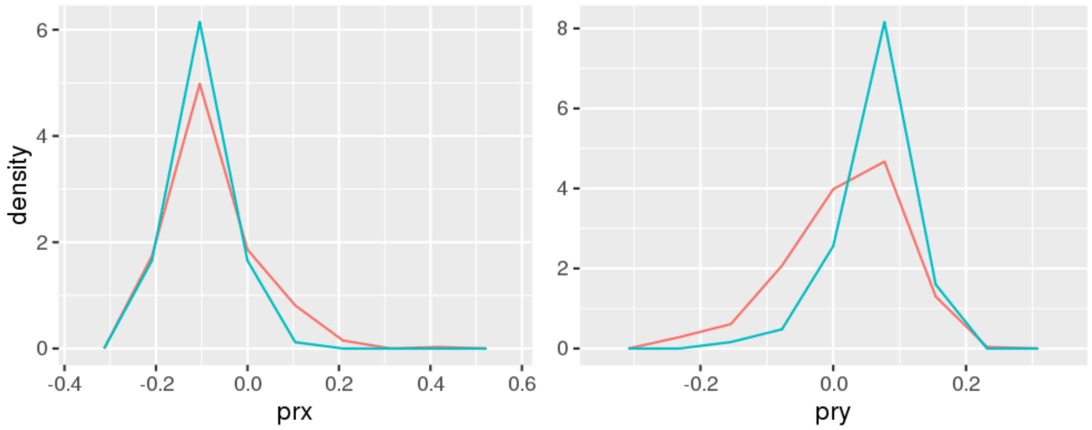

FIGURE 4.18: Distribution of low (0.5–0.8, red) and high (0.5–0.8, blue) $rASA_e^f$ epitope patches across the first two MDS proximity components, prx and pry. Higher $rASA_e^f$ patches have a lower spread across prx and pry.

may not be considered noisy; conversely, if the non-epitope residues are true non-epitopes then these may average out the epitope signal when features are calculated for the patch. This is exemplified by the all-epitope sub-patches (see section 4.3.4.1). The inclusion of these noisy patches during training may account for the low specificity that is observed, as the trees of the random forest must alter the feature values used to split instances at each node in order to accommodate this noise. With this in mind, the effect of $rASA_e^f$ on testing and training was investigated.

4.3.5.1 Low and high-$rASA_e^f$ epitope patches on the first MDS components

It was hypothesised that the distributions of low- and high-$rASA_e^f$ epitope patches on the first two MDS components would differ. In particular, if low-$rASA_e^f$ epitope patches are considered to be noisy patches, they would likely have a larger range of feature values and therefore be spread further across the proximity space. Epitope patches were split into low-$rASA_e^f$ (0.5–0.8) and high-$rASA_e^f$ (0.8–1) groups, which were then compared. Figure 4.18 shows their distribution across the first and second scaled proximity components, prx and pry. Along both components, spread appears to be reduced for patches with $rASA_e^f > 0.8$. Brown-Forsythe tests were performed between the two groups, which gave p-values of 0.001 and 7.825×10^{-5} for prx and pry respectively, confirming that low-$rASA_e^f$ values tend to be more spread across prx and pry than high-$rASA_e^f$ patches. As described in section 4.3.3 the MDS components approximate the proximity of instances within the terminal nodes of the trees of the random forest. Because high-$rASA_e^f$ epitoe patches are less spread across the first two components, it can be inferred that they tend to end up in the same terminals nodes more often than low-$rASA_e^f$ patches do. This suggests that the IntPred:Epi is able to classify less 'noisy', high-$rASA_e^f$ patches more

TABLE 4.9: *mLtr* chains sub-setted by their minimum $rASA_e^f$ value. The second column shows the range that a chain's minimum $rASA_e^f$ must be within to be included in the subset. Square brackets and parentheses indicate inclusive and exclusive ranges respectively.

subset	$\min(rASA_e^f)$	num. chains
s0.5	$[0.5, 0.6)$	2
s0.6	$[0.6, 0.7)$	10
s0.7	$[0.7, 0.8)$	12
s0.8	$[0.8, 0.9)$	7
s0.9	$[0.9, 1]$	16

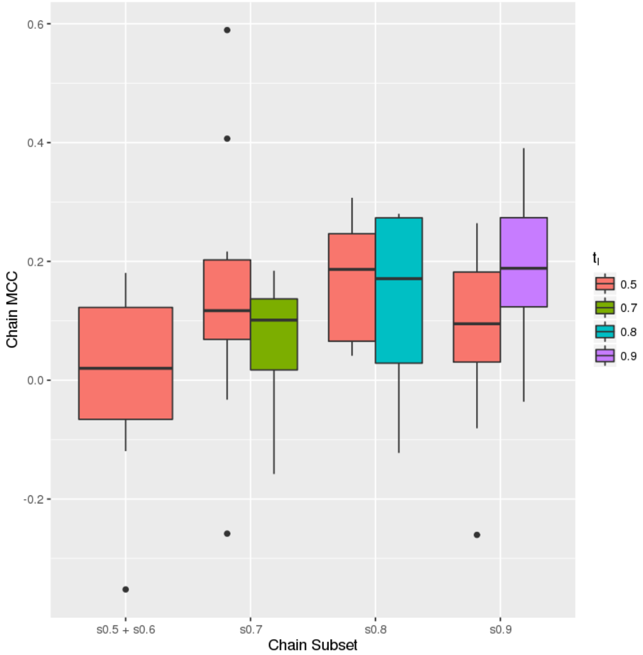

FIGURE 4.19: The effect of t_l in training and testing. Red bars show the performance of the baseline learner ($t_l = 0.5$) on each chain subset (defined in table 4.9). The remaining bars show the performance on each subset after retraining using each specified t_l. t_l is the $rASA_e^f$ threshold used to label patches as epitope. Chain MCC is the MCC for each chain in the subset. Chain subsets s0.5 and s0.6 are grouped together because s0.5 only has two members (see table 4.9).

easily than 'noisy', low-$rASA_e^f$ patches. This supports the idea that low-$rASA_e^f$ patches make the learning process more difficult and suggests that prediction performance may be improved by increasing the class label threshold t_l. This is investigated below.

4.3.5.2 The effect of increasing t_l

First, the relationship between $rASA_e^f$ and test performance was investigated. It was hypothesised that chains with higher $rASA_e^f$ values would be easier to predict. To test this, the chains of *mLtr* were split according to their minimum $rASA_e^f$ value — that is, the $rASA_e^f$ value of the patch having the minimum $rASA_e^f$ for the chain. These subsets

Table 4.10: Performance comparison of the default ($t_l = 0.5$) and $t_l = 0.9$ learner. Performance was compared on the s0.9 subset (see table 4.9). ‘Total’ refers to the total number of patches and ‘Epitope / Total’ refers to the fraction of patches labelled as epitope.

t_l	Total	Epitope / Total	Sens.	Spec.	FPR	FDR	PPV	MCC
0.5	1500	0.11	0.76	0.40	0.60	0.81	0.19	0.09
0.9	1364	0.02	0.84	0.74	0.26	0.86	0.12	0.19

are shown in table 4.9. The MCC for these subsets was then calculated from the by-chain CV results on the *mLtr* set. The results are shown in figure 4.19 (red boxes alone). As expected, median MCC on s0.7 is higher than ‘s0.5 + s0.6’ (grouped together owing to small sample size); a further improvement is also seen for s0.8 in comparison with s0.7. However, MCC is poorer for s0.9 in comparison with t0.8. Thus, there seems to be an overall trend of better MCC for chains with higher maximum MCC, up until maximum $rASA_e^f > 0.9$. In other words, the random forest tends to perform better when a chain has epitope patches with surfaces that are more epitopic — but this is not the case when a chain's epitope patches have very high proportions of epitopic surface.

Next, the effect of increasing t_l *before* training was investigated. A series of learners were trained on *mLtr* with $t_l = (0.7, 0.8, 0.9)$. For each learner, performance should be compared to the learner trained on *mLtr* with the default $t_l = 0.5$. Because some chains do not have any patches where $rASA_e^f$ reaches the given t_l, the by-chain CV step was amended so that in a training partition, any patch from any chain was allowed (except the chain in the test partition), but each test partition only consisted of patches from a chain with a minimum $rASA_e^f > t_l$. This meant that each learner was only cross-validated on the chains from the corresponding subset (see table 4.9). Thus, in order to compare performance, the default learner was also only cross-validated on the same subset. Figure 4.19 also shows these comparisons. As stated earlier, the red boxes show the performance of the baseline learner on the different chain subsets. The remaining boxes show the performance when t_l is altered. There appears to be no significant difference when the t_l is raised to 0.7 or 0.8, but the performance on the s0.9 subset appears to improve when t_l is raised to 0.9. Paired Wilcox tests were performed to assess the change in MCCs between the default learner and each other learner on the appropriate chain subset. The change in MCCs for t_l 0.7 and 0.8 were both found to be insignificant (p-value = 0.078 and 1 respectively), but there was a significant difference between the default and $t_l = 0.9$ learner (p-value = 0.003).

The performance of the default and the $t_l = 0.9$ learner is shown in table 4.10. The first thing to note is that when $t_l = 0.9$, the number of patches labelled epitope drops from 166 to 30 (~ 2 patches per chain), an 82% decrease. Consequently, the fraction of epitope patches in the set falls from 0.11 to 0.02. Despite this, the predictor manages to improve specificity markedly while also increasing sensitivity, which leads to MCC more than doubling (0.09 to 0.19). It must be noted that because the number of epitope patches changes with t_l, it is not straightforward to interpret the change in sensitivity. However, changing t_l does not affect the number of non-epitope patches in the set, as a patch is labelled non-epitope if its $rASA_e^f = 0$. Specificity is the number of true negatives over the the total number of negatives and can therefore be directly compared between the two t_l values. The jump in specificity seen when $t_l = 0.9$ can be interpreted as the classifier assigning far fewer false positive labels, when the surfaces of patches labelled as epitope consist almost completely of epitope residues. This suggests two possible avenues of further research. The first would be to investigate the possibility of using smaller patches with a higher t_l. In theory, this could lead to the improvement in performance seen in table 4.10, while also allowing the inclusion of chains that do not have any epitope patches with the current combination of patch size and t_l. The second would be to assess the effect of a higher t_l on patch to residue mapping (see section 4.4). It may be that the higher sensitivity and specificity seen for $t_l = 0.9$ would also lead to improved performance at the residue level.

4.4 Residue-level prediction

Similarly to general PPIs, most BCE prediction methods predict on a residue basis, rather than a patch basis. Therefore, in order to compare performance, patch predictions must be mapped to residue predictions. The same mapping methods applied to IntPred as described in section 3.3.1 were applied to the *RF:mLtr-b* leaner. The results are shown in table 4.11. Similarly to IntPred, the Minimal, Score and Vote methods all result in a drop in performance. However, in contrast to IntPred, the Centre method improves performance in comparison to patch-level performance: MCC increases from 0.06 to 0.11 (see table 4.4). The same mapping-patch analysis that was run for IntPred (see 3.3.2) was also repeated. The results are shown in figure 4.20; similarly to general PPIs, the Minimal, Score and Vote methods improve as mapping patch size is reduced, until all three methods converge at the smallest mapping patch size tested, which is equivalent to the Centre method.

4.4.1 Residue prediction clustering

Figure 4.20 illustrates that even after reducing the mapping patch size, there remains a high number of false positives (FPR = 0.46). In order to reduce the number of false positives, a method was implemented for clustering positively predicted residues. The idea was that if a positively predicted residue is surrounded by other positively predicted

Table 4.11: Residue-level prediction performance. This table shows the by-chain CV performance of the *RF:mLtr-b* learner on residues, using the four patch-to-residue mapping methods defined in section 3.3.1. Values shown are averages from 20 balanced training sub-sets. These performances can be compared to the by-chain CV performance of *RF:mLtr-b* on patches, shown in table 4.4.

Learner	Sensitivity	Specificity	Accuracy	MCC
Centre	0.65	0.54	0.59	0.11
Minimal	0.90	0.14	0.20	0.03
Score	0.23	0.81	0.76	0.03
Vote	0.50	0.59	0.58	0.04

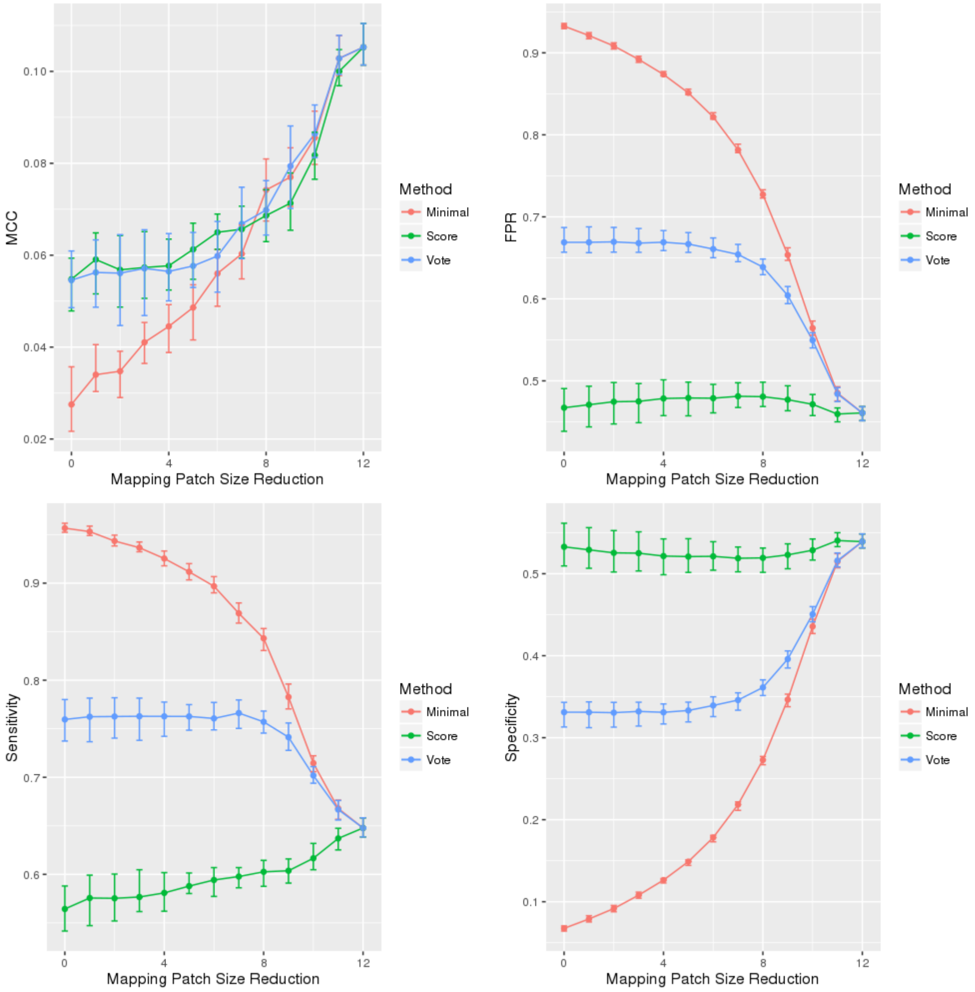

Figure 4.20: The effect of reducing mapping patch size. These graphs show how performance changes with size of the patch used to map predictions from patch to residue for surface residues of the *mLtr* set. The changes are similar to those seen for general PPIs (see figure 3.1).

residues, it is more likely to be a real epitope residue. The clustering method works by calculating a cluster score S_c for each positively predicted residue:

$$S_c = \frac{N_p}{N_T} \tag{4.5}$$

where N_T is the total number of neighbouring residues and N_p is the number of neighbouring residues predicted as positive. Neighbouring residues are defined as those found in the surface patch created by using the most solvent exposed atom from the positively predicted residue as a central atom. The surface patch has a radius r. A positive label will be changed to negative unless $S_c > t$, where t is a defined threshold. Thus the clustering method has two parameters: the threshold t and the neighbourhood patch radius r.

Clustering was applied to the by-chain cross-validated *mLtr* results with a series of t and r (patch predictions were mapped to residue using the Centre map method (see section 4.4)). r was incremented from 4 to 20 Å in steps of 1 Å. For each r, t was incremented from 0 to 1 in steps of 0.05; thus $17 \times 20 = 340$ different parameter pairs were tried.

To assess the effect of clustering, ΔMCC is used

$$\Delta MCC = MCC_{cls} - MCC_{orig} \tag{4.6}$$

where MCC_{orig} and MCC_{cls} is the MCC before and after clustering respectively.

Figure 4.21 shows the ΔMCC for every r, t combination. ΔMCC is negative for most (r, t), indicating that performance deteriorates with clustering. The (r, t) with the best performance improvement is $r = 11, t = 0.45$, where $\Delta MCC = 0.01$.

As clustering takes spatially proximal predictions into account, it will be influenced by chain-specific features of the prediction. To investigate this, different r and t combinations were performed on each chain. For each chain, ΔMCC was calculated across r, t. Figure 4.22 shows that for nearly all chains, performance is not altered at very low t, regardless of r. The effect of increasing t varies greatly for different chains. For some chains, such as 1rjzG and 3o0rB, performance is not greatly altered for any (r, t). In some cases, such as chain 1nsnS and 1kyoE, performance either remains the same or deteriorates — in some cases, the deterioration is marked ($\Delta MCC = -0.5$). For others, such as 1kl3A and 1v7mV, performance either remains the same or improves. This improvement can be marked in some cases ($\Delta MCC = 0.3$). For most chains, performance can improve or deteriorate across (r, t) and there is no pattern between them. This

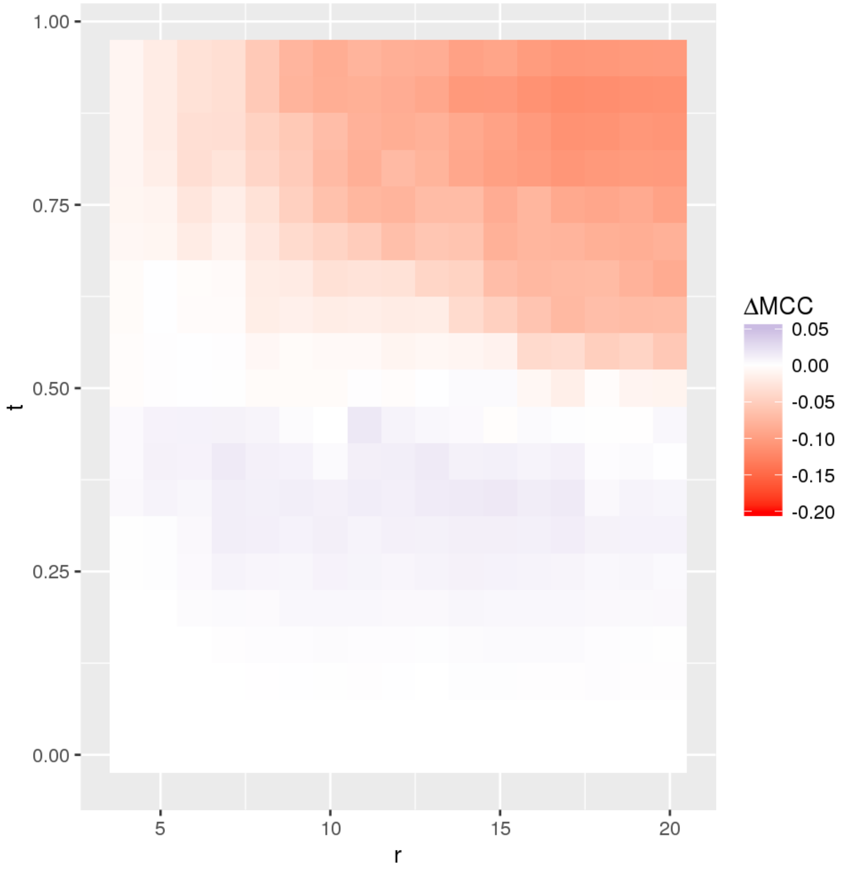

FIGURE 4.21: Clustering on the *mLtr* training set across parameters r and t. r = cluster patch radius, t = cluster threshold. A positive ΔMCC indicates an improvement in performance; negative indicates a deterioration.

explains why very small ΔMCC are observed for most (r, t) when an average is taken across the whole data set (see figure 4.21).

4.4.1.1 Identifying cluster performance correlates

In order to investigate further the relationship between cluster performance and chain-level features, the standard deviations of ΔMCC values for all chains at each (r, t) were calculated. Figure 4.23 shows that as both r and t increase, the standard deviation of chain ΔMCC increases. It was hypothesised that this variation might correlate with chain-level features (e.g. chain size). This might allow us to predict the performance of clustering at certain (r, t), given some chain-level feature(s). Furthermore, if different (r, t) correlated in different directions with the given features, then this could be used to selectively apply clustering at different (r, t) for different chains.

For a given chain-level feature, the correlation with ΔMCC was calculated for each (r, t). The following chain-level features were tested:

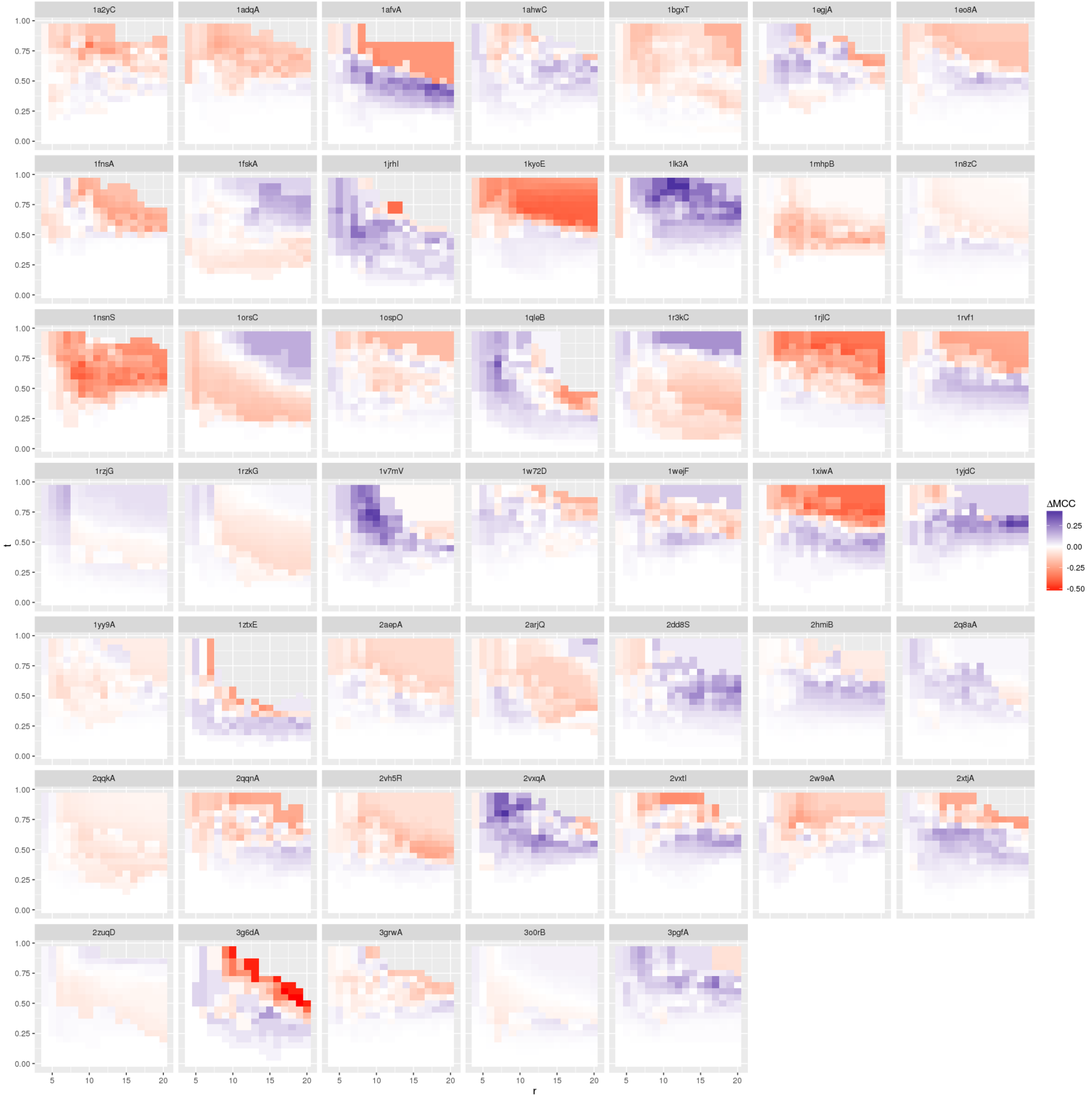

FIGURE 4.22: Clustering across r and t, by chain. r = cluster patch radius, t = cluster threshold. A positive ΔMCC indicates an improvement in performance and negative indicates a deterioration. Combinations of r, t where no residues are predicted as epitope are omitted.

- chain size (number of patches).
- total number of predicted positive patches.
- prediction rate (predicted positive / chain size).

It was found that none of these features were found to correlate significantly with performance. It was therefore decided to see if there were any non-predictive variables resulting from the pre-cluster prediction stage that could explain the chain-level variance in the clustering effect. For each (r, t), the correlation ρ_{MCC} between chain ΔMCC and original chain MCC (MCC_{orig}) was calculated. In order to account for bias when testing the

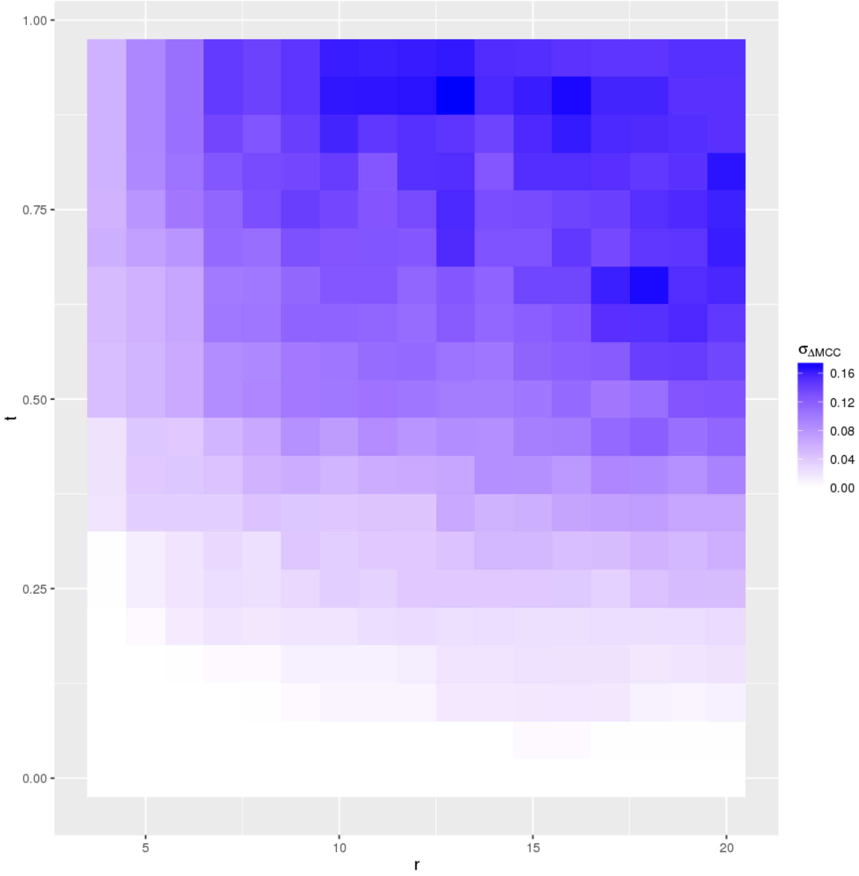

FIGURE 4.23: Chain $\sigma_{\Delta MCC}$ across (r, t).

correlation between a change and an initial value, Oldham's correction (Oldham, 1962) was applied to MCC_{orig} before calculating a correlation

$$MCC^c_{orig} = \frac{MCC_{orig} + MCC_{cls}}{2} \tag{4.7}$$

Correlations were then multiplied by $\sigma_{\Delta MCC}$ to obtain a score s_ρ

$$s_\rho = \sigma_{\Delta MCC} \times p_{MCC} \tag{4.8}$$

This helps to distinguish between correlations across small and large ranges of ΔMCC.

Figure 4.24 shows s_ρ for each (r, t). The highest s_ρ is seen $r = 18$, $t = 0.45$, where $\rho_{MCC} = 0.67$. Figure 4.25 shows MCC^c_{orig} plotted against ΔMCC for all chains using this parameter combination. Thus, the efficacy of clustering at certain (r, t) is to some extent dependent on the initial pre-clustering prediction performance. It can be reasoned that this dependence is due to the initial proportion of positively predicted residues that are epitope residues (i.e. true positives) — this proportion is the precision of a

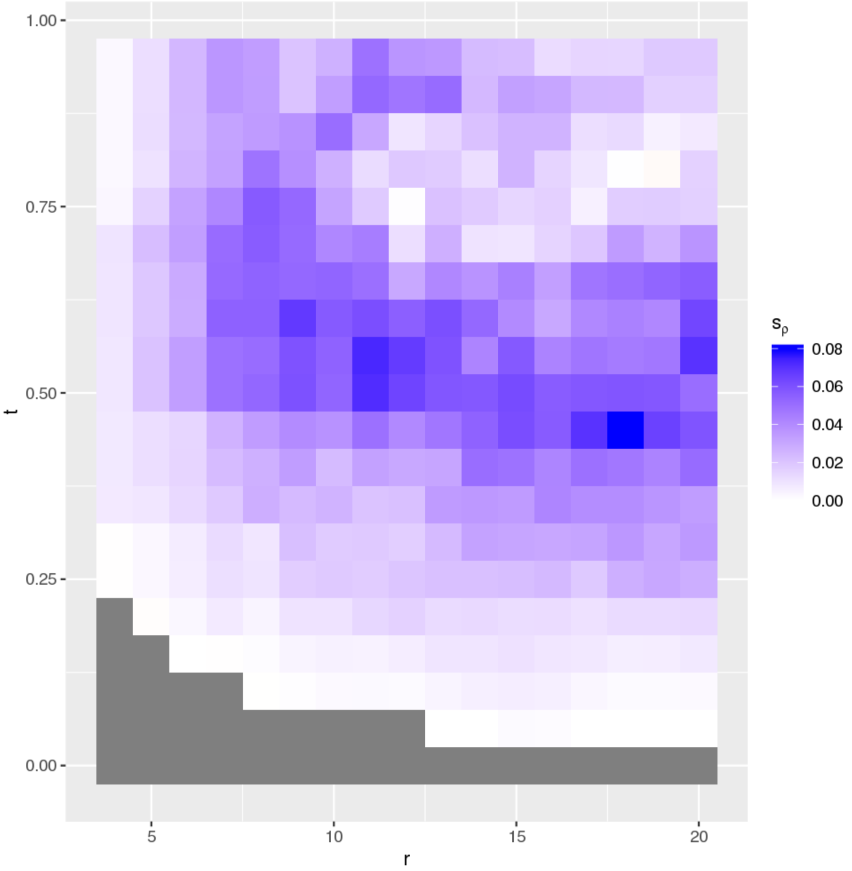

FIGURE 4.24: Correlations between MCC^c_{orig} and ΔMCC across (r, t) Correlations are multiplied by $\sigma_{\Delta MCC}$ to help identify correlations across large ranges of $DeltaMCC$.

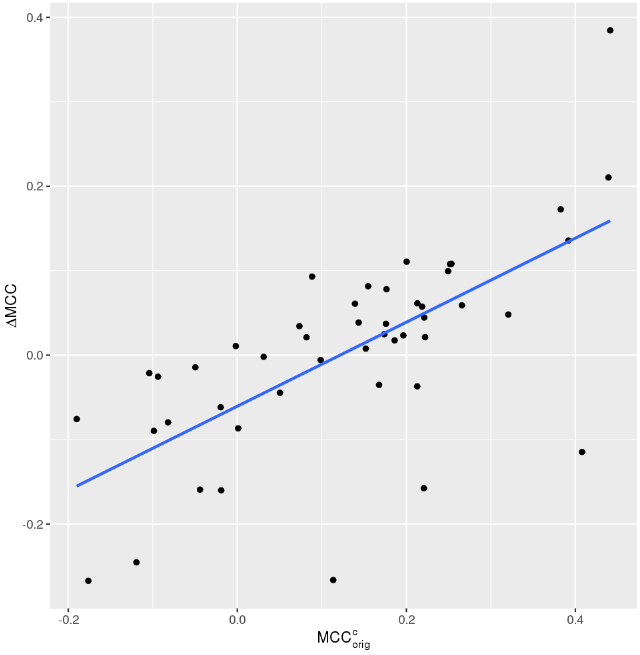

FIGURE 4.25: ρ_{MCC} between chain MCC^c_{orig} and ΔMCC across (r, t). MCC^c_{orig} is MCC_{orig} corrected using Oldham's correction (Oldham, 1962) (see main text for more details).

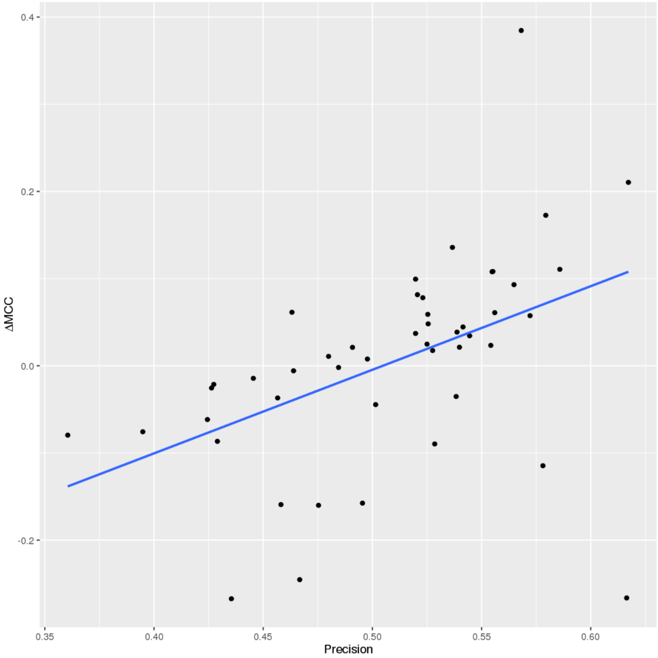

Figure 4.26: Pre-clustering precision against ΔMCC for $r = 18, t = 0.45$. For chains with pre-clustering precision > 0.5, most chains have a positive ΔMCC (performance improves). Conversely, those chains with a pre-clustering precision < 0.5 have a negative ΔMCC (performance deteriorates.)

predictor. If the true-positive:false-positive ratio is higher, then any cluster of positively predicted residue is more likely to contain true-positives. Therefore when the positive labels are retained for these clusters, it more likely that true-positives are retained. Figure 4.26 shows the relationship between precision and the ΔMCC for each chain at $r = 18$, $t = 0.45$. A positive relationship between the two can be seen: as precision increases, ΔMCC increases. This suggests that if pre-cluster prediction precision could be increased, then clustering could lead to improved performance. To demonstrate this, ΔMCC was recalculated for those chains with $MCC_{orig} > 0.1$. For this subset, mean $MCC_{orig} = 0.17$. Figure 4.27 shows that the best improvement for this subset is seen at $r = 15$, $t = 0.5$. Clustering results in an $\Delta MCC\ = 0.04$, giving a final $MCC = 0.21$.

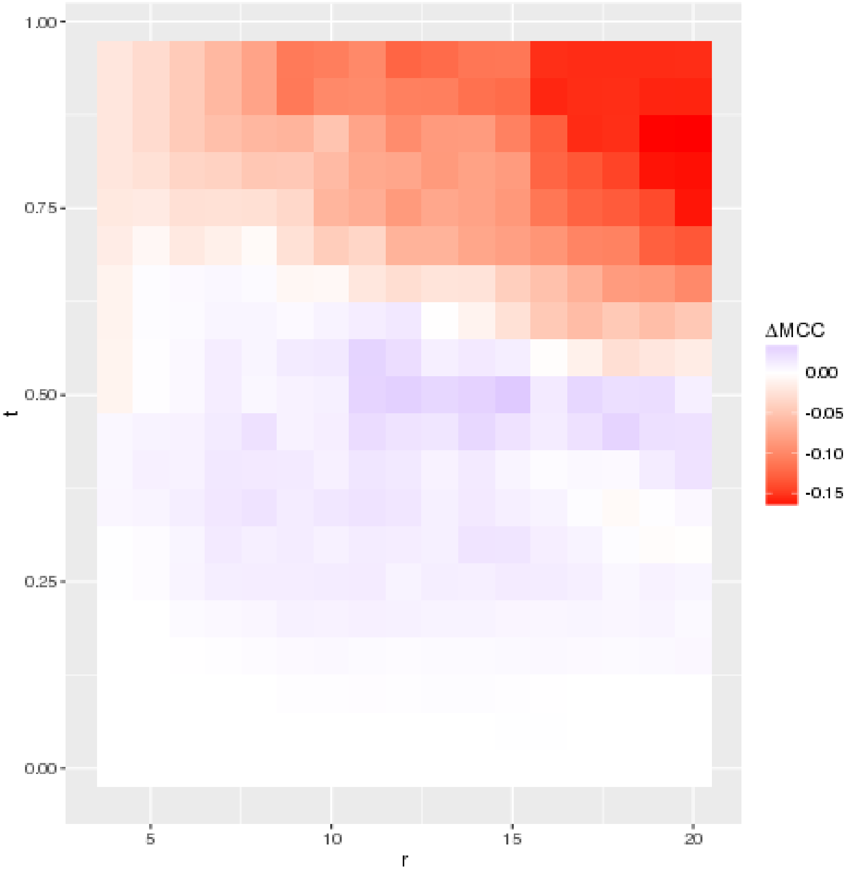

Figure 4.27: Clustering on the *mLtr* chains where $MCC_{orig} > 0.1$, across r and t. r = cluster patch radius, t = cluster threshold. A positive ΔMCC indicates an improvement in performance; negative indicates a deterioration.

4.5 IntPred:Epi definition and testing

The *RF:mLtr-b* learner, trained using a training set class distribution of 0.65 (see section 4.2.3.3) and with an additional $rASA$ feature (see section 4.2.3.2), was taken forward for testing and comparison to existing methods. This method is termed **IntPred:Epi**.

4.5.1 Correct evaluation of BCE predictor performance

Before testing IntPred:Epi, the decision was made to re-evaluate the performance of current methods. Nearly all BCE prediction methods work by assigning a prediction label to each residue of a structure, regardless of residue location. This means that prediction is performed on residues that form the core of the protein (and therefore could never be found in an epitope). It stands to reason that if a predictor simply predicts whether a residue is surface or not (which is trivial if one of your features is solvent accessibility), then it may perform reasonably at predicting epitopes when performance is judged on all residues. To illustrate this point, consider a predictor that simply labels any residue

Table 4.12: 'Surface Predictor' Results. The predictor simply labels any surface residue as epitope and any core residue as non-epitope. The resulting $MCC = 0.14$

	Positive	Negative
True	197	1485
False	3047	0

Table 4.13: Comparison between evaluation of current methods on all residues and surface residues only. (s) = surface residues only. Note that sensitivity does not change with the removal of non-surface core residues.

Method	Sensitvity	Specificity	Specificity (s)	MCC	MCC (s)
ABCPred	0.48	0.56	0.57	0.01	0.02
BCPREDS	0.83	0.30	0.31	0.06	0.07
Bepipred	0.76	0.50	0.46	0.10	0.10
Bpredictor	0.01	0.99	0.99	0.01	0.00
DiscoTope 2	0.93	0.27	0.20	0.08	0.08
ElliPro	0.82	0.47	0.33	0.12	0.08
SEPPA 2.0	0.27	0.95	0.94	0.19	0.19

as epitope if it has a rASA > 10%, and labels the remaining as non-epitope. Assessing the performance on the *mLts* antigen set (see section 4.2.1), this predictor obtains MCC score of 0.14 (the confusion table is shown in table 4.12).

With this in mind, the performance of competing methods was re-evaluated by predicting on surface residues ($rASA > 10\%$) of the *mLts* set only . Note that prediction scores produced by all the methods tested were obtained from the supplementary data supplied by Hu et al. (2014) and the prediction label thresholds stated in that study are used here. The performances of the methods are compared in table 4.13. It can be seen that, for most methods, testing on surface residues does not lead to a significant change in performance. The exception is ElliPro: its MCC score falls from 0.12 to 0.08. BCPREDS is the only predictor whose performance improves; its MCC score increases from 0.08 to 0.1. These surface-only predictions will be used for comparison with IntPred:Epi

4.5.2 Method comparison

The performance of IntPred:Epi on the surface residues of the *mLts* test set is compared with existing methods *mLts* in table 4.14. Using MCC score as a comparison, IntPred:Epi does reasonably, out-performing all methods except SEPPA 2.0, which beats our method by a margin of 0.08 (MCC). In comparison to SEPPA, IntPred:Epi has improved sensitivity (0.76 and 0.58) but poor specificity (0.48 and 0.82).

Table 4.14: BCE predictor method comparison. This table shows the performances of seven BCE methods in comparison with IntPred:Epi. All methods are tested on the surface residues of the *mLts* test set.

Method	Sensitivity	Specificity	FPR	FDR	PPV	MCC
ABCPred	0.48	0.57	0.43	0.93	0.07	0.02
BCPREDS	0.83	0.31	0.69	0.93	0.07	0.07
Bepipred	0.76	0.46	0.54	0.92	0.08	0.10
Bpredictor	0.01	0.99	0.01	0.94	0.06	0.00
DiscoTope 2	0.93	0.20	0.80	0.93	0.07	0.08
ElliPro	0.82	0.33	0.62	0.93	0.07	0.08
IntPred:Epi	0.74	0.48	0.52	0.92	0.08	0.11
SEPPA 2.0	0.27	0.94	0.06	0.78	0.22	0.19

4.5.3 Further comparison between IntPred:Epi and SEPPA 2.0

The performance of SEPPA 2.0 and IntPred:Epi were futher investigated. In the next chapter, an attempt is made to tailor general B cell epitope prediction to the prediction of epitopes for a given host species, e.g. the prediction of epitopes found bound to (host) human antibody. In order to provide a baseline for such methods, SEPPA 2.0 and IntPred:Epi was run on two test sets: human antibody-bound antigens and mouse antibody-bound antigens. Additionally, for this round of testing, IntPred:Epi was retrained on the SEPPA 2.0 training set. Retraining on this set guarantees fair comparison between SEPPA 2.0 and IntPred:Epi — as well as the the methods presented in the next chapter — while also minimizing the overlap between the training set and the data sets described in below, so that as many chains could be tested upon as possible.

4.5.3.1 Obtaining human- and mouse-host test antigen structures

For an antigen to be included in either the human- or mouse-host set, a number of criteria needed to be met. As stated above, the antigens were required to be bound by human or mouse antibody. Human/mouse proteins bound to human/mouse antibody were also to be avoided — this is because the concepts presented in chapter 5 only hold under the assumption that host tolerance is correctly regulated, i.e. human antibodies are not generated against self protein.

The Immune Epitope Database (IEDB) is an online resource for B and T cell epitopes (Vita et al., 2010) (see section 2.2.1). The IEDB was used to identify antigen structures because of the ease of obtaining antibody-antigen complex annotation related to source and antigen type. The IEDB Database was downloaded from the IEDB server as a MySQL Database export on 01/06/2016. The database was queried for antibodies with structures solved by x-ray diffraction, returning 1541 complexes. A structure was kept if

Table 4.15: Human- and mouse-host antigen sets. Surface residues are defined as those residues with an $rASA > 10\%$. The human-complement and mouse-complement sets are subsets of the larger sets that have had any chains found in the SEPPA 2.0 training set removed. See section 4.5.3.1 for more details on construction of the data sets.

Data Set	Chains	Surface Residues
human	18	2675
human-complement	10	1760
mouse	53	9172
mouse-complement	27	4159

the antibody origin was defined as host (human or mouse) and if antigen was defined as a single chain protein from a non-host source. This search returned 183 and 212 antigen chains for human and mouse-host respectively.

Next, any chains < 30 residues in length or with a resolution $> 3\,\text{Å}$ were removed. Epitope residues were defined as described in section 3.4. Antigens were then split into human and mouse-host sets. For each set, a subset was also created that did not include any of the structures found in the SEPPA 2.0 training set — these sets are termed *complement* sets. Similarly to the creation of previous data sets, structures were clustered at 90% sequence identity and epitope labels mapped to a representative structure (see section 3.4.2). However, in addition, representatives were clustered at 60% sequence identity and structures representing the largest number of aligned residues were kept. The four sets are summarized in table 4.15.

4.5.3.2 SEPPA 2.0 and IntPred:Epi testing

The *complement* sets were then used to test SEPPA 2.0 and IntPred:Epi (retrained on the SEPPA 2.0 training set).

The human-complement and mouse-complement sets were run on SEPPA 2.0 via web-server[1]. Each set was run using batch mode. Host species was set to 'Homo' or 'Mus' and prediction thresholds to 0.1 or 0.12 (as recommended) for human-complement and mouse-complement sets respectively. For both sets, sub-cellular localisation was set to 'unspecified'.

Note that for this evaluation, summary statistics were calculated by calculating them for each chain and then averaging (in contrast to aggregating predictions across all chains and then calculating). This is in agreement with the method used to test SEPPA 2.0 previously (Qi et al., 2014).

[1] Available at `http://badd.tongji.edu.cn/seppa/`

Table 4.16: IntPred:Epi and SEPPA 2.0 performance on human and mouse test sets. The sets referred to here are the human and mouse-complement sets (described in section 4.5.3.1). IntPred:Epi was retrained on the SEPPA 2.0 training set before testing (see section 4.5.3).

Test Set	Classifier	Sens.	Spec.	PPV	FDR	FPR	MCC
Human	IntPred:Epi	0.4725	0.6225	0.1502	0.8498	0.3775	0.0567
	SEPPA 2.0	0.5203	0.5392	0.1352	0.8648	0.4608	0.0264
Mouse	IntPred:Epi	0.5176	0.6443	0.1933	0.8067	0.3557	0.1104
	SEPPA 2.0	0.5235	0.5659	0.1557	0.8443	0.4341	0.0599

The results of the testing are shown in table 4.16. On the human test set, both methods do not perform well in comparison with previous testing — MCC 0.057 and 0.026 for IntPred:Epi and SEPPA 2.0 respectively, in comparison to MCC 0.11 and 0.19 on *mLts* set — but IntPred:Epi outperforms SEPPA 2.0. A similar drop in SEPPA 2.0 performance is seen on the mouse test set (MCC 0.06) but this is not the case for IntPred:Epi, which maintains the same performance as seen on *mLts* (MCC 0.11).

4.6 Discussion

As shown in section 4.2.2.1, a large difference in performance was observed between the initial cross-validation and testing stages. The suspicion that this was being caused by the presence of overlapping patches was confirmed by applying the by-chain CV method, which simply ensures that each test partition consists of all of the patches from one chain, thus ensuring that no overlapping patches are found across training and test partitions. Using the by-chain CV method, it was found that cross-validated and test performances were much more similar. It is important to consider the implications of this finding. Many predictors are evaluated on their cross-validation performance alone (Ansari and Raghava, 2010, Lin et al., 2013, Ren et al., 2014). In particular Lin et al. (2013) and Ren et al. (2014) report superior performance, despite being based on sequence information alone. Though none of these methods are patch-based, they all use some sort of sequence-window or neighbour-based function to calculate residue features; this would lead to the same over-optimistic performance measures seen for IntPred:Epi during normal cross-validation. It is therefore important to reiterate the importance of testing on an independent test set — without this, the real power of a predictor cannot be known. It should be noted that one of the very most recent methods applies an equivalent method to by-chain CV for cross-validation (Ren et al., 2015). However, no independent testing is done, which is vital to understand how a predictor performs on data that has not been used to tune parameters of the model.

As well as retraining, additional ASA-based features `ASA` and `rASA`, which were not implemented for IntPred, were found to improve prediction. In the future, these features should also be applied to IntPred. Different class label distributions were also tested to find the optimal balance. As well the distribution of class labels, the selection of the negative subset was found to influence by-chain cross-validated performance — this is discussed further below.

The analysis of the random forest performed in section 4.3 gives us some insights into the BCE prediction problem. Similarly to PPIs (Baresic, 2011), propensity, BLAST conservation score and planarity all seem to play an important role in predicion. In contrast, hydrophobicity seems to play more of a role in BCE prediction than it does PPI prediction — this corresponds to the depletion in BCEs of hydrophobic residues observed in other studies (Rubinstein et al., 2008, Krawczyk et al., 2013). FOSTA conservation seems to play less of a role in BCE prediction, which is reasonable considering the fundamental importance of functional conservation in PPIs.

The visualisation of random forest predictions gives us an insight into its behaviour. It is clear that IntPred:Epi finds it difficult to distinguish between epitope patches and a large proportion of the remaining surface. However, considering the labelling problem inherent to BCE prediction using x-ray crystal structure data — that is, positive and negative labelling based upon the interaction between an antigen and a single *monoclonal* antibody, rather than data from a *polyclonal* response — it is perhaps unsurprising that this is the case. In reality, the antibody response must be able to recognise and distinguish between a huge number of proteins and so it is expected for many surfaces to be recognised and therefore be potentially epitope. Thus, it seems likely that many patches currently labelled as non-epitope are, in reality, epitope. This has been recognised by Ren et al. (2015), who rather than approaching BCE prediction as a positive-*negative* prediction problem, treated it as a positive-*unlabelled* problem. By using an SVM to define ‘reliable negatives’ and then training another SVM on positive (epitope) and ‘reliable negatives’, they were able to improve performance when assessed on traditional positive and negative labels. This may be why a distribution of by-chain CV performance using different randomly selected subsets of the negative set was observed — the higher performance subsets may have more ‘reliable negatives’ and therefore be having the same effect on performance as observed by Ren et al. (2015).

The observation that, along the first three MDS clusters, non-epitope patches are present at a distance from the main cluster of epitope and non-epitope patches and are therefore rarely grouped together with epitopes by the random forest. This suggests that some non-epitope patches might be real non-epitopes, rather than mislabelled epitopes. Using the correlations between the first three MDS components and the `pro`, `hpho`, `pln` and `blast` `features`, it was possible to identify combinations of feature values that lead to patches being co-located with epitope patches in the terminal nodes of the forest less frequently. These were: low propensity and non-planarity; extreme planarity; very high evolutionary conservation and very low evolutionary conservation. This is interesting because, in most contexts, a researcher would be interested in predicting epitope (e.g. for rational vaccine design) — but for the application to therapeutic antibody candidate screening and design, a drug designer is more interested in finding a biological therapeutic that has a 'non-epitope' surface, in order to avoid immunogenic reaction upon administration. In the future, further investigation of these patches may help to identify what makes a surface 'immunologically silent'.

An analysis of the influence of $rASA_e^f$ and of the class label threshold t_l is also presented. It was observed that epitope patches found to be class outliers in the random forest could be amended such that, if only those residues defined as epitope were kept in the patch, a sub-patch could be defined that was less of an outlier. This, along with the reduced spread of high $rASA_e^f$ patches along the first two MDS components, suggests that lower $rASA_e^f$ patches may be reducing the 'epitope signal' during learning. Further to this, it was observed that chains with higher $rASA_e^f$ tend to be predicted on more easily, with the exception of chains with very high $rASA_e^f$ patches. This suggests that, in general, chains with 'less noisy' epitope patches are easier to predict. Interestingly, the drop in performance on chains with very high $rASA_e^f$ can be reversed when IntPred:Epi is retrained with a higher t_l — in fact, for the $\sim 1/3$ of *mLtr* chains with very high $rASA_e^f$ patches, performance roughly doubled from MCC 0.09 to 0.19. Of course, this comes with the caveat that as t_l is higher, more patches are labelled U and thus there may be more of a difference between stated and real-case performance. This needs to be assessed my mapping the patch predictions from the $t_l = 0.9$ learner to residue predictions. Furthermore, it is obviously not ideal to be unable to learn from $\sim 2/3$ of the dataset that has no patch labelled as epitope! This could be addressed by changing the patch radius, as there is a better chance that some patches will have sufficient $rASA_e^f$ when each patch is smaller. It was shown that reducing patch size leads to a deterioration in performance, but this analysis was performed with the default $t_l = 0.5$. An exploration of a range of t_l and patch radius combinations should be undertaken to see if a more optimal solution can be found.

Prediction clustering has been used previously to improve residue-level prediction (Ren et al., 2014) and so a similar method was also applied here. In contrast to the method presented by Ren et al. (2014), which takes into account the number of positively predicted and negatively predicted residues within a sequence window and then inputs these as features into a machine learning method, the method applied here is simpler — the prediction label of a positively predicted residue is changed to negative, unless the proportion of positively predicted residues within a given radius of it is over a given threshold. This approach was taken primarily because it was observed that IntPred:Epi was over-predicting, rather than under-predicting. It was found that clustering was not able to improve performance for the entire training set, but did improve performance for chains on which pre-clustering performance was relatively. This suggests that if better performance is obtained in the future, clustering may lead to further gains in performance. Importantly, the same principle of clustering holds for PPI residues — that is, a PPI residue is very likely to be found close to another PPI residue. If good pre-clustering performance is a prerequisite for effective clustering, then IntPred should be improved by the application of a similar clustering method.

Before their comparison to IntPred:Epi, seven existing methods were re-evaluated for their performance on surface residues *only*, after it was shown that a predictor could give competitive performance across all residues, simply by labelling all surface residues as epitope. The performance of these methods was mostly unchanged by this re-evaluation, with the exception Ellipro (Ponomarenko et al., 2008), where MCC fell from 0.12 to 0.08. When compared to the seven existing methods on a test set of 14 antigens, it was found that IntPred:Epi outperformed all methods except SEPPA 2.0. Further testing between IntPred:Epi and SEPPA 2.0 on two test sets of 10 human antibody-bound and 27 mouse antibody-bound antigens — primarily for the purposes of creating a baseline for the host-specific methods presented in the next chapter — showed that IntPred:Epi was able to maintain its performance on the mouse set, whilst SEPPA 2.0 performance fell markedly. Furthermore, although IntPred:Epi performance fell on the human set, SEPPA 2.0 performance fell dramatically. It is unclear why this drop in performance is seen. It may be that the published version of SEPPA 2.0 is actually trained on a data set that includes the initial test set: it is not unreasonable for researchers to develop a method using a training set, state its performance on a test set and then combine both sets to create a published version that utilises all the available data. However, if this is the case, then it makes retesting the method by other researchers more difficult. Nevertheless, the human and mouse test sets together combine to more than double the size of the initial set, suggesting that the performance measures on these sets is more representative than the initial testing.

Finally, it is unclear why IntPred:Epi and SEPPA 2.0 methods perform worse on the human-host set than on the mouse-host set. The only previous comparison of BCE predictor performance on human and mouse-host test sets found that, of the four methods tested, two performed slightly better on the human-host set, one performed better and the final method showed similar performance on both sets Qi et al. (2014). Is there some intrinsic property of human-host epitopes that makes them more difficult to for IntPred:Epi to predict? Or is there a bias for researchers to co-crystallise ‘unusual’ epitopes with human antibody? Further work is needed to confirm the relationship between host and prediction performance.

4.7 Conclusion

In this chapter, the general PPI method IntPred was retrained and extended to produce the IntPred:Epi method for B cell epitope prediction. The performance of IntPred:Epi on BCEs is not as good the performance of IntPred on PPIs, but the comparison of IntPred:Epi performance to existing methods shows that BCE prediction is a much more difficult task. An exploratory analysis of the IntPred:Epi random forest indicates that there may be a set of features that are distinctly ‘non-epitope’, which is promising for the application of a BCE predictor for the screening, or design, of biological therapeutics. The exploratory analysis also helped to identify a relationship between the epitope surface fraction of epitope patches and their prediction that suggests better performance could be obtained by altering the class label threshold. Furthermore, clustering promises to become effective once an improvement in performance is seen.

Despite SEPPA 2.0 outperforming IntPred:Epi in the initial round of testing, further testing of the two methods in order to provide a baseline for the host-specific methods presented in the next chapter showed that IntPred:Epi outperformed SEPPA 2.0 on a larger test set. Thus, IntPred:Epi will be used in the next chapter as a foundation for the tailoring of the BCE prediction problem to a certain host.

Chapter 5

Development of Tolerance Labels for B-Cell Epitope Prediction

In this chapter, an attempt is made to build upon the *general* BCE prediction method IntPred:Epi (presented in chapter 4) by applying the concept of immune tolerance to augment its predictions. First immune tolerance is introduced, before the creation of libraries of tolerated human and mouse surface patches is described. These libraries are then utilised to produce features for BCE prediction, the efficacy of which is evaluated.

5.1 Introduction

When attempting to predict the presence of epitopes on the surface of a biological therapeutic, it is important to think about the context of the therapeutic. Rather than the prediction of any epitope, a drug (or possibly vaccine) designer is interested in predicting epitopes that are recognised by the *human* immune system. Despite this, nearly all BCE prediction methods do not take the host immune system into account and predict, for example, on test antigen bound to human antibody the same way that they treat antigen bound to mouse antibody. Is it certainly reasonable that human and mouse could produce quite different antibody responses and therefore recognise different epitopes — the difference in antibody gene loci is just one example of a factor that differs considerably between the two species (see section 1.1.6). Considering the difficulty of BCE prediction, any useful relationships between epitope selection and host-species could prove very valuable.

Almost all B cell epitope prediction methods do not take the host species (that is, the species that produced the antibody found bound to the antigen) into account. The only method to have done so is SEPPA 2.0 (Qi et al., 2014). As well as a general BCE prediction method, SEPPA 2.0 provides models that have been built on the human- and mouse-host subsets of its main training set[1]. However, comparison between the general SEPPA 2.0 method and the human- and mouse-host models showed that neither model is able to improve over the general method (Qi et al., 2014). This likely because each model is trained on less data which in general leads to poorer performance on independent test data. It is therefore unlikely that any gains in performance for a given host can be obtained simply from retraining on a subset of the data. Instead, a different approach is taken here, that considers something that is truly unique for each species and is therefore likely to influence the selection of epitopes: immune tolerance. The work presented in this chapter aims to incorporate the concept of immune tolerance into the prediction of B cell epitopes. Thus immune tolerance will be introduced before moving on to how it can be applied to B cell epitope prediction.

5.1.1 Immune tolerance

The immune system must be able to distinguish between harmless and pathogenic molecules. Dysregulation of this distinction can lead to autoimmunity on one hand — where an immune response is raised against non-pathogenic 'self' molecules — or the spreading of an infection if a pathogen is not regarded as a danger. The ability of the immune response to avoid reactions against self molecules is known as **tolerance**. and is the result of a number of mechanisms that help to regulate the repertoire of the immune system.

5.1.1.1 Central tolerance

The observation that 55% to 75% of antibodies expressed by early immature B cells exhibit self-reactivity is testament to the challenge the immune system faces to maintain tolerance (Wardemann et al., 2003). Central tolerance refers to three processes that occur in the primary lymphoid tissue that help to avoid the generation of mature self-reactive lymphocytes: clonal deletion, clonal anergy and receptor editing.

Both B and T cells undergo selection processes during their development that help to avoid self-reactivity — this is known as clonal **deletion**. In the thymus, T cells go through a positive selection — whereby only those that recognise self-MHC are sent survival signals — followed by a negative selection, where T cells binding MHC-bound self-antigen die (Kappler et al., 1987). Antigen presenting cells are capable of inducing

[1] Additional models are also provided that are trained on data subsets grouped by cellular location ('membrane' and 'secretory')

this deletion. Similarly, B cells undergo selection during development in the bone marrow — for B cells, ligation of the BCR to self antigen results in deletion (Hartley et al., 1993).

Clonal deletion is complemented by **clonal anergy**. Anergy refers to a cellular state in which induced impairment of signal transduction pathways results in a functionally inactive lymphocyte, thus maintaining tolerance. Within B cells, this modulation of signal transduction results in a reduction in proliferation rate, BCR expression and lifespan (Goodnow et al., 1988). A similar process of anergy occurs in T cells, activated through TCR ligation in the absence of co-stimulation and reversed through exposure to IL-2 (an autocrine factor expressed by activated, but not anergic T cells) (Schwartz et al., 1989).

Central tolerance is further applied through **receptor editing** (Gay et al., 1993). This is a process whereby self-reactive B cells undergo rearrangement of heavy- and light- chain genes in order to redefine their specificity. B cells can undergo receptor editing multiple times and, if auto-reactivity persists, apotosis can be induced. A similar process also occurs in developing T cells (McGargill et al., 2000).

Tolerance is aided through the expression of tissue-specific antigens within the thymus. The observation that humans expressing a defective form of the autoimmune regulator (AIRE) protein develop multi-organ autoimmune disease lead to the discovery that AIRE is involved in the ectopic expression of peripheral-tissue antigens within the medullary epithelial cells of the thymus (Anderson et al., 2002). AIRE therefore plays an important role in inducing T cell tolerance to tissue-specific antigens. The effect of AIRE on gene expression is complex but a sample of genes it up-regulates shows that both intracellular and extracellular tissue-specific antigens are targeted (Anderson et al., 2002).

5.1.1.2 Peripheral tolerance

Though central tolerance leads to the removal, or functional depression, of self-reactive lymphocytes in the primary lymphoid tissues, there are populations of lymphocytes that exist outside these tissues that are reactive to self-antigen. Peripheral tolerance is important in ensuring that tolerance is maintained in tissues beyond the primary lymphoid tissues. The main mechanism of B-cell peripheral tolerance is a lack of T cell help, which leads to reduction in B cell lifespan (Fulcher et al., 1996). This reduction is also dependent on the level of Ig receptor binding to the self-antigen. Similarly, peripheral T cells can undergo anergy through a lack of co-stimulation (Lechler et al., 2001). Tolerogenic dendritic cells also exist that deliver co-inhibitory signals that prevent T cell activation and proliferation (Gallucci et al., 1999).

Immune privilege allows the expression of an antigen within a tissue without eliciting an immune response. Immune privilege was thought to be a natural consequence of blood-tissue barriers and a lack of lymphatic drainage within certain tissues (e.g. the cornea of the eye). Current opinion describes immune privilege as a more dynamic and active system of interaction between lymphocytes and specialised tissue. In the eye, for example, expression of ligands such as CD95L induces apoptosis of CD95$^+$ T cells that is critical for maintaining tolerance (Griffith et al., 1996).

5.1.1.3 Autoantibodies

It must be noted that even with the mechanisms of central and peripheral tolerance, antibodies that bind to self-antigen (autoantibodies) still occur. Although autoantibodies are a common hallmark of autoimmune diseases where tolerance mechanisms have broken down, they are not only a property of pathogenic states. Antibodies that react with self-antigen in healthy individuals are known as natural autoantibodies — these are mainly IgM and, in contrast to pathogenic autoantibodies, exhibit no, or few, mutations in variable regions and have low affinity for antigen (Elkon and Casali, 2008). Most natural autoantibodies are also polyreactive, binding to several unrelated antigens, generally with low to moderate affinity (Elkon and Casali, 2008).

5.1.2 Does the immune system tolerate intracellular antigen?

Like proteins in general, the ability of antibody to bind to intracellular protein will vary. However, it may be that the seclusion of intracellular protein away from circulating antibody and B lymphocytes disallows tolerance to be developed. Certainly, autoantibodies to intracellular antigens can be generated. In particular, these autoantibodies have been shown to be present in a number of disease states, such as ischemic heart attack (Nussinovitch and Shoenfeld, 2010) and cardiomyopathy (Nussinovitch and Shoenfeld, 2011). Anti-intracellular antigen autoantibodies are also seen in cancer states — it has been shown that in some human medullary breast carcinomas, B lymphocytes expressing anti-actin BCRs are found at the tumoral cell surface (Hansen et al., 2002).

How these disease-state autoantibodies are generated is unclear and four non-mutually exclusive hypotheses have been put forward (Racanelli et al., 2011). General dysregulation of the immune system, a character of most autoimmune diseases, is one hypothesis — though this is not supported by observations that in nearly all cases, anti-intracellular antigen autoantibodies show significant somatic mutation of variable region genes, indicating continuous contact between antibody and antigen (Racanelli et al., 2011).

Dysregulation of apoptotic proccesses has been implicated in the generation of these autoantibodies. Studies in several autoimmune diseases including systemic lupus erythematosus (SLE) have shown that impaired clearence of apoptotic blebs — small microbodies that express intracellular antigen on their membrane — generate strong immune reponses against the intracellular antigen that play a role in pathogenesis (Munoz et al., 2010). A similar occurance is seen in some cancer states — the anti-actin autoantibodies seen in some human medullary breast carcinomas were found bound to the tumoral cell surface of early phase apoptotic blobs (Hansen et al., 2002).

Epigenetics also seems to play a role in anti-intracellular autoantibody generation. Epigenetic changes including DNA methylation, histone modifications and miRNA expression are associated with altered gene expression that favours autoimmune disease. For example, human and murine CD4$^+$ T cells treated with hydralazine and procainamide — DNA methyltransferase inhibitors that invoke SLE-like symptoms in patients — increases expression of B cell co-stimulatory factors. Co-culturing of these co-stimulatory factors with B cells lead to increased antibody secretion (Oelke et al., 2004). In another case, patients with vasclitudes caused by anti-neutrophil cytoplasmic antibodies exhibited a loss of epigenetic silencing that contributed to an increase in expression of the antigen of intracellular self-antigens myeloperoxidase and proteinase-3 (Ciavatta et al., 2010).

The final proposed mechanism of anti-intracellular antigen autoantibody generation is molecular mimicry. In molecular mimicry, the generation of antibodies against foreign proteins that cross-react with intracellular self-antigen results in autoantibodies. Although a handful of studies have tried to test this hypothesis, discrepencies in results have highlighted the difficulty in choosing an appropriate methodology (Racanelli et al., 2011).

It is important to note that proteolytic processing of intracellular antigens may be important in the generation of autoantibodies in some disease states. The granzyme family of proteases are a component of cytotoxic lymphocyte granule-mediated cell death that have been implicated in catalysing structural changes to self-antigen that may be important in autoimmunity. Granzyme B in particular has been shown to have many targets that are autoantigenic — these targets are also diverse in respect to their localisation, though many are intracellular (Darrah and Rosen, 2010). Notably, granzyme B (GrB) was found to process autoantibody target actin on the surface of apoptotic human medullary breast carcinoma cells (Hansen et al., 2002). Human GrB has a broad tetrapeptide specificity and consequently many consensus target sequences can be found — in reality, most of these sites are not cleaved (Gahring et al., 2001). This is likely to be due to structural considerations — one study shows that GrB cleavage of extracellular matrix proteins vitronetin and fibronectin only occurs when they are in matrix-associated conformations

(Buzza et al., 2005). It is likely that granzyme B cleavage influences autoimmunity by revealing cryptic epitopes, or destroying normally dominant epiotopes, as is seen for similar proteases such as asparaginyl endopeptidase (AEP) which cleaves tetanus toxoid C fragment (TTCF) (Manoury et al., 1998). Moreover, a review of known GrB cleavage sites and known autoimmune T and B cell epitopes shows that the two are commonly colocated (Darrah and Rosen, 2010). However it must also be noted that there is no direct evidence that shows GrB cleavage altering the autoimmune responses or disease propagation (Darrah and Rosen, 2010).

There are very few studies that specifically investigate the effect of antigen location on immunogenicity. By using transgenic mouse models, Ferry et al. (2003) tested the difference in response to a tolerogenic cell surface protein, hen egg lysozyme (HEL), upon sequestration to the endoplasmic reticulum through addition of a two amino acid retention signal. Rather than induce tolerance as the cell surface antigen does, intracellular sequestration lead to failure in tolerance and autoimmunogenesis upon administration of HEL. The intracellular antigen induced B cells to differentiate into B1 cells and produce large numbers of IgM autoantibodies in a T cell independent manner. Further work on intracellular antigen by Ferry et al. (2007) suggests a functional role for the immunogenicity of intracellular antigen during dead cell clearance. Using the mouse model of sequestered HEL, they found that autoreactive B1 cells seemed to play an important role in IgM and C1q-dependent cell clearance through autoantibody binding to HEL exposed on the surfaces of dying cells. In this way, moderate autoimmunity appears to limit exposure of conventional B cells to self antigen by aiding the clearance of potentially immunogenic intracellular antigen.

5.1.3 Tailoring the B cell epitope prediction problem

Recent work has addressed how general B cell epitope prediction performance is affected when the species of pathogen is specified (Resende et al., 2012), but as mentioned above, there has only been one attempt at restricting BCE prediction to a certain host which, upon testing, appeared to not improve performance compared with a general BCE prediction model (Qi et al., 2014). Host-specific B cell epitope predictions could be applied in the development of vaccines, as well as in the prediction of protein therapeutic immunogenicity. The decision was made to take a novel approach to the prediction of host-specific BCEs, by considering the influence of tolerance on the selection of epitopic surfaces.

Our initial concept was that the surfaces of tolerated self-protein influences the selection of antigenic sites on foreign antigen. In the early stages of an antibody response, antibodies may be raised that cross-react with self protein via the similarity between surfaces of self-protein and surfaces of foreign protein. However, due to mechanisms of tolerance (as described above), the B cells producing these self-reactive antibodies are either deleted (clonal deletion), rendered functionally silent (clonal anergy), or altered in their specificity (receptor editing). This allows the antibody response to produce high-affinity, high-specificity antibodies, while maintaining tolerance to self. This process of tolerance maintenance implies that there are some surfaces that, due to their similarity to self-surfaces, are unlikely to be the target of an antibody response. Therefore, it was hypothesised that the tolerance state of a protein surface should anti-correlate with the epitope state, i.e. a tolerated surface is unlikely to be an epitopic surface. Moreover, if the tolerance state of a protein surface can be predicted, then this could aid B cell epitope prediction.

It was hypothesised that a collection of self protein surfaces could be used to inform the process of epitope prediction on an antigen by effectively ruling out similar surfaces that would otherwise be predicted as epitope due to phyisco-chemical, structural or evolutionary features. This collection could be used to create a classifier to label surfaces as 'tolerated' or 'non-tolerated'. Such a classifier could then be incorporated into an epitope prediction method, in the hope that the tolerance state of a surface has a significant enough effect on epitope selection to result in good predictive performance.

With this in mind, the following work falls into four tasks:

1. Creation of a library of tolerated self surface patches.
2. Development of a method to compare test patches against library of self patches.
3. Creation of a classifier to predict the tolerance state of a patch by comparison against the self patch library.
4. Incorporation of the tolerance state classifier into a general B cell epitope prediction method.

Note that, in theory, this method could be used to tailor a general B cell epitope prediction method to any species. However, two types of data must be available: protein structures from the host organism in order to create a library and antibody-antigen complex structures in order to test the resulting method. The majority of complexed antibody structures are human and mouse so these were focused on in the following work.

The above method was implemented by constructing human and mouse-tolerated surface libraries (TSLs) and apply these libraries by creating a number of different methods in an attempt to improve B-cell epitope prediction when applied to human and mouse antibody-bound antigen.

5.2 Creation of Tolerated Surface Libraries (TSLs)

In order to create libraries of tolerated surfaces, the following was undertaken for each host (human and mouse):

1. Collect host SwissProt entries.
2. Select resolved PDB and model structures to represent these entries.
3. Create surface patches from chosen structures.
4. Create surface patch descriptions.

In section 5.1.2, previous studies were described that explored the relationship between immune tolerance and cellular location of self-antigen. Considering the evidence that some intra-cellular antigens were not tolerated, it was hypothesised that the inclusion of intracellular proteins within a TSL would lead to test surfaces falsely labelled as tolerated. The decision was made to test the effect of including intracellular structures by creating extracellular-only TSLs, as well as libraries including all structures.

Additionally, the effect of including modelled structures within libraries was to be tested. Models increase the structural coverage of a TSL and therefore increase the likelihood of identifying tolerated surfaces on foreign antigen. However, model quality may affect labelling through the inclusion of poorly modelled surfaces. This may lead to test surfaces being falsely labelled as tolerated, due to matching with such a surface.

With both of these considerations in mind, four libraries were created for each host that agreed with the following criteria (the emphasised terms will be used in the following sections):

ec: Extracellular proteins only, no models.

ec-m: Extracellular proteins only, including models.

all: All proteins, no models.

all-m: All proteins, including models.

The selection of solved and modelled structures for inclusion in these libraries is explained in the following sections.

Table 5.1: GO terms used to search the GO term tree. Any child of any of these terms having the relationship *is_ a*, *intersection_ of* or *relationship* to its parent was used to select extra cellular SwissProt entries.

GO Term	Name
GO:0031012	extracellular matrix
GO:0044420	extracellular matrix part
GO:0005576	extracellular region
GO:0044421	extracellular region part
GO:0016020	membrane
GO:0044425	membrane part

5.2.1 Selection of extracellular SwissProt entries

In order to create the extracellular-only libraries, candidate structures must be filtered by cellular location. This can be accomplished by referring to the GO annotation of a SwissProt entry.

5.2.1.1 Selection of GO Terms

As described in section 2.2, GO terms have been comprehensively applied for the annotation of UniProtKB-SwissProt entries (Harris et al., 2004). GO terms can be used to describe the cellular location of proteins and therefore allow us to determine if a protein is to be included in our extracellular-only libraries. The GO term scheme was downloaded on 12/04/2013. As described in section 2.2, GO terms are arranged in a semi-hierarchical structure whereby child terms relate to their parents via a specified relationship. To identify proteins annotated as either extracellular or membrane associated, six cellular_component GO terms were chosen from which to find all child terms that had either of the following relationships: *is_ a*, *intersection_ of* or *relationship* (see table 5.1). These six GO terms were picked manually by starting at the root term and moving downwards, therefore guaranteeing that no relevant parent terms were missed.

Note that although GO:0016020 (membrane) and GO:004442 (membrane) annotated proteins may be associated with intracellular membranes, the selection of extra-cellular sequence regions avoids the inclusion of these in the final libraries (see section 5.2.1.2).

5.2.1.2 Selection of SwissProt entries

For the extracellular-only libraries, SwissProt entries were selected if they were annotated with at least one of our set of GO terms. If the matching GO term was a term relating to extra-cellularity, the entire entry sequence was selected. If the matching GO term was related to membrane-association, the sequence regions annotated as extra-cellular were selected.

5.2.2 Selection of PDB structures

In order to select representative PDB chains for a given SwissProt entry, RepPDB was developed (see below). RepPDB was used to select a subset from all structures assigned to a human or mouse SwissProt entry. In the case of the extracellular-only libraries, if a SwissProt entry was selected due to its annotation with a membrane-associated GO term (see table 5.1), then only those PDB chains with sequences 'sufficiently aligned' to an extra-cellular sequence region were kept. A chain is 'sufficiently aligned' if its sequence meets any of the following criteria:

- It is contained within an extracellular region.
- It is an extracellular sequence region.
- The start and end points are each less than 20 residues away from the start and end points respectively of an extracellular region.

A sequence length cut-off of 30 was used to remove peptides from the set. A resolution cut-off of 3 Å was also used.

5.2.2.1 RepPDB

For a given SwissProt entry, RepPDB seeks to create a subset of PDB chains that maximises the both the structure quality and coverage of the sequence, while minimizing redundancy between structures.

The first step is to assign structures to SwissProt entries and then align structure sequences with entry sequences. This is done by PDBSWS Martin (2005). The second step is to then gather information about each PDB chain (resolution, R factor, and experimental method) and its alignment to a SwissProt sequence (matches, mismatches, insertions, deletions and alignment start/end positions). Given a PDB chain with a sequence s that is aligned from position i to position j of a SwissProt entry, the coverage c of the chain is calculated as

$$c = j - i - (m + n + d) \tag{5.1}$$

where m, n and d are the total mismatches, insertions and deletions respectively.

Using this information, it can be decided which PDBs are optimal for a given SwissProt entry by determining which PDBs overlap in their structural representation of the entry's sequence.

For those overlapping chains, a decision has to be made about which to keep. If two chains have the same alignment start and end positions, then the chain with the best coverage is chosen; if coverage is the same, then the chain with the best resolution is chosen and if resolutions are also the same then the chain with the lowest R-factor is chosen.

Any chains that are complete subsets of another chain are removed. If two chains overlap but each chain represents a sufficiently large sequence area (> 20 residues) not represented by the other, then the higher-coverage PDB is chosen while the other is labelled as a 'partially redundant'. This allows the user to include these if they value additional coverage over redundancy.

If the alignment start and ends of two chains are both close to each other (within 20 residues at each end and having a difference in coverage ≤ 20) and one chain has better resolution but worse coverage then a decision boundary that balances coverage and resolution is used (see figure 5.1).

For the creation of the libraries, RepPDB was used to select structures and partially redundant chains were kept.

5.2.3 Including SWISS-MODEL models

Modelled structures improve the structural coverage of the human/mouse proteome and therefore may improve performance of the method. SWISS-MODEL models were obtained for all human and mouse SwissProt entries that had no solved structure in the PDB (Biasini et al., 2014). For each entry, the model with the template having the highest sequence identity to the entry was kept. No models were kept if the highest template sequence identity was $< 60\%$. 924 and 956 models were kept for human and mouse respectively.

5.2.4 Surface patch creation

Surface patches have been previously been effectively applied to predict general protein-protein interfaces (Baresic, 2011). The same process of patch creation as in previous

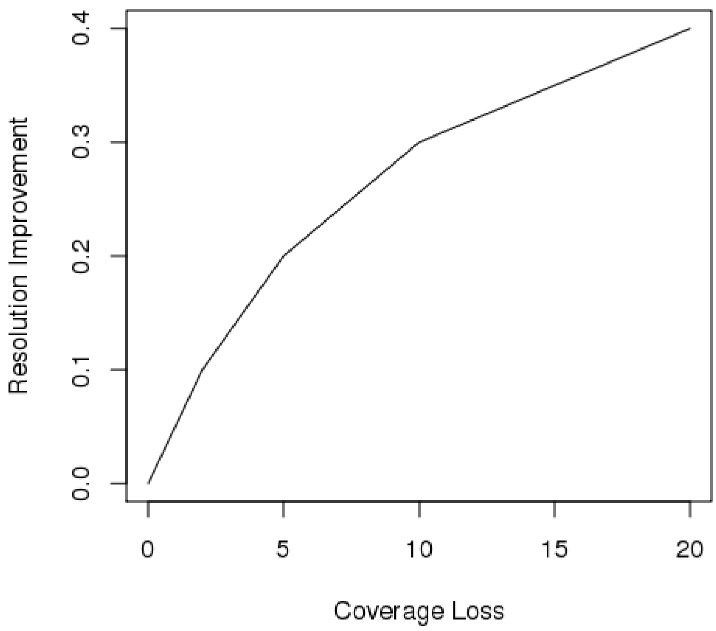

Figure 5.1: Decision boundary used to decide between a pair a and b of overlapping structures, where b has poorer coverage but better resolution. b is selected if it leads to a loss of coverage but improvement in resolution that places it above the line — otherwise a is selected.

work was used (described in section 2.3.2), with a slight modification. Surface patches are created by identifying residues within a given radius that are in contact with a chosen central residue. This differs from the previous method, in which residues are included that are in contact with at least one member residue and within the given radius, but not necessarily in contact with the central residue. The method was modified in this manner because a single layer of residues around the central residue was required for the purposes of patch description (see below). In this work, a patch radius of 8 Å was chosen in order to ensure that a full layer of residues was included around the central residue.

5.2.5 Tolerated Surface Libraries (TSLs)

The final TSLs are described in table 5.2. As defined in section 5.2, the *ec* libraries include extracellular proteins only, whereas the *all* libraries include all proteins regardless of cellular location. The *m* libraries also include SWISS-MODEL models.

Table 5.2: Tolerated Surface Libraries (TSLs).

Host	Data Set	Cellular Location	Structures			Patches
			Resolved	Modelled	Total	
Human	human-ec	Extracellular	1354	0	1354	108,448
	human-ec-m	Extracellular	1354	159	1513	123,327
	human-all	All	4993	0	4993	363,418
	human-all-m	All	4993	924	5917	474,794
Mouse	mouse-ec	Extracellular	385	0	385	27,441
	mouse-ec-m	Extracellular	385	302	687	58,822
	mouse-all	All	1288	0	1288	87,469
	mouse-all-m	All	1288	956	2244	200,420

5.2.6 Surface patch description

Antibodies recognise antigens due to the geometric and physico-chemical complementarity of paratope and epitope surfaces. The antibody response relies on differences between surfaces in order to distinguish antigen. In the case that a target surface is similar to another surface, cross-reactivity may occur. Thus, the aim of patch description is to describe patches in a way that allows the identification of cross-reactive patches. Due to the specific nature of epitope recognition, cross-reactivity between surfaces will not be the result of general features that may be shared by geometrically and physico-chemically distinct patches, such as secondary structure composition or surface accessibility. Rather, cross-reactivity will be decided by features that are specific to those patches.

A method of patch description was required in order to test for matches against a given TSL. The decision was made to describe patches using a simple ‘clock-face’ type description, where the central residue is surrounded by an ordered layer of peripheral residues. This description can be represented by a string of characters, where the first character represents the central residue and the remaining characters represent the peripheral residues. These description strings allow rapid comparison of a test patch against a TSL.

For a given patch, the following procedure occurs:

1. Patch atom co-ordinates are translated so that the centroid of the C_α atoms of the patch lies at the origin (0, 0, 0).

2. PCA is performed on patch atom co-ordinates.

3. An observation point is defined at (1, 0, 0). An observation vector is defined with its initial point at the observation point and its terminal point at the origin. All atoms are rotated so that the plane formed by the first and second principle component vectors lays orthogonal to the observation vector. This plane is the patch plane.

4. All non-patch atoms that are contacting patch atoms are identified. If an equal or greater number of non-patch atoms lay in front of the patch plane than behind it, then a new observation vector is created that is the opposite of the original observation vector. Unless an equal number of non-patch atoms lay on either side, the original observation point is discarded.

5. The patch is formed from the patch centre residue and peripheral residues surrounding it. For the C_α of each peripheral residue, a projection onto the patch plane is calculated. For each observation vector, the order of these projections is calculated using their angle around the vector.

6. For each observation vector, the order of residues is recorded as a string. Each patch residue is represented by its one-letter amino acid code. The first character of the string represents the central residue. The remainder of the string represents the order of peripheral residues around the central residue. As this peripheral order is circular (i.e. only the position of a residue relative to the position of the other residues is important), it must be made linear in a way that allows strings to be compared coherently. A peripheral residue order containing n residues can be represented by n unique but overlapping strings, by starting each string at a different peripheral residue. These strings are sorted alphabetically and the first is chosen to represent the peripheral order. This ensures that patches with the same central residue and peripheral residue order match when compared.

5.3 Tolerance classifiers

A foreign patch can be classified as 'tolerated' or 'untolerated' by comparing it to the patches in a self-tolerated label. Specifically, if the description string of a foreign patch is found in the self-tolerated library, it is labelled 'tolerated'; otherwise it is labelled 'untolerated'. Different tolerance classifiers can be created not only by changing our patch library but also by introducing an amino acid grouping scheme. The twenty amino acids can be grouped according to various physico-chemical or evolutionary relationships. These groupings can simplify comparative analysis between proteins, leading to insights into the principles behind protein structure, function and evolution. In this case, grouping of amino acids may allow us to identity foreign patches that are sufficiently similar to a self patch to be tolerated. If amino acid grouping can be used to identify these patches then it will improve the performance of our method.

5.3.1 Amino acid grouping and patch description string translation

To compare patch description strings, it is important to consider the number of unique strings there are. This number is dependent on the number of residues in a patch and the number of amino acids there are; given these two numbers, it is possible to calculate how many unique patch description strings there are. This is useful for understanding the likelihood of a match when comparing the description strings of two patches.

The total number of different patch descriptions N for a patch with n residues, given that there are g amino acids, can be calculated as

$$N = g \times \sum_{k \in F_x} \frac{g^k - k - \sum\limits_{j \in \bar{F}_x} (g^j - j)}{k} \tag{5.2}$$

where

- $k = n - 1$
- F_x is the set of all positive divisors of x
- $\bar{F}_x$ is the set of all proper divisors of x (i.e. all positive divisors of x except itself)

For patch size of 8 residues, given that there are 20 amino acids, $\sim 3 \times 10^9$ different patch description strings are possible. It follows that it is very unlikely that the patch description strings of two patches will match, given the large number of possible configurations. To address this, it was sought to reduce the possible number of patch description strings. Amino acids can be grouped to reduce the number of distinct combinations. For example, if the twenty amino acids are grouped into four groups (i.e. $g = 4$ in equation 5.2), the number of distinct patch description strings for patch size = 8 residues drops from $\sim 3 \times 10^9$ to 9376, making it far more likely that the patch descriptions of two patches will match. A **grouping scheme** can be used to determine which amino acids are grouped together in order to reduce the number of patch description strings. This process of applying an amino acid grouping to a patch description string is termed **translation**.

5.3.2 Taylor groupings

In order to group the amino acids, it is sensible to use a grouping scheme that groups them according to some sort of similarity. Taylor defined a classification of the twenty amino acids, based on physico-chemical and mutation data (Taylor, 1986). As shown in

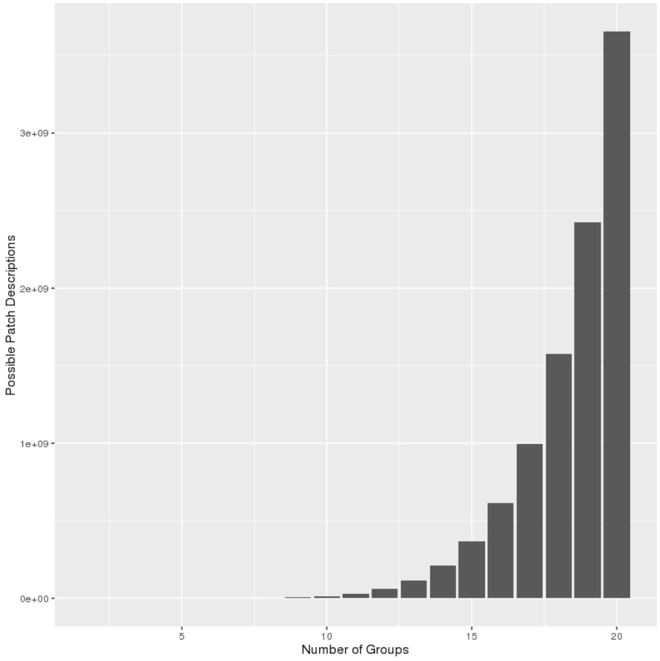

Figure 5.2: Unique patch description strings as a function of the number amino acid groups, when string length = 8. Calculated using equation 5.2

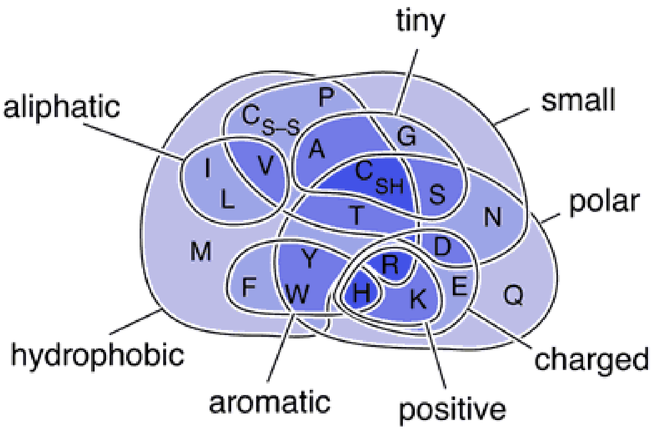

Figure 5.3: Venn diagram of the 20 amino acids placed according to the properties defined by Taylor (1986). Image obtained from `https://commons.wikimedia.org/wiki/File:Amino_Acids_Venn_Diagram.png`

figure, the resulting classification can be viewed as a Venn-diagram where eight circles, each representing a physico-chemical property, intersect to form seventeen groups in which the amino acids are placed. Thus, amino acids are described by one or more properties which they may share with other amino acids.

The number of groups can be altered by using a subset of the eight properties. A method was devised to create a set of groupings by iteratively adding properties. Starting with all the properties, the property associated with the smallest number of amino acids is removed. This forms one unique grouping scheme. The process is then repeated until only

```
0: [A][V][F][IL][M][G][P][C][T][H][K][WY][D][N][S][E][R][Q]
1: [A  V][F][IL  M][G  P][C  T][H][K][WY][D][N][S][E][R][Q]
2: [A  V][F][IL  M][G  P][C  T][H][K][WY][D][N][S][E  R][Q]
3: [A  V][F  IL  M][G  P][C  T][H  K][WY][D][N][S][E  R][Q]
4: [A  V][F  IL  M][G  P][C  T][H  K  WY][D  N][S][E  R  Q]
5: [A  V  F  IL  M][G  P][C  T  H  K  WY][D  N][S  E  R  Q]
6: [A  V  F  IL  M][G  P][C  T  H  K  WY  D  N][S  E  R  Q]
7: [A  V  F  IL  M  G  P][C  T  H  K  WY  D  N  S  E  R  Q]
```

Figure 5.4: Taylor grouping schemes, based upon the Taylor groupings (Taylor, 1986). In group scheme 0, twenty amino acids are initially grouped by the properties described by Taylor. Properties are successively removed until two groups remain. Each grouping scheme provides a unique way to represent amino acids.

two groups remain. This results in eight unique grouping schemes. This set of groupings schemes can be thought of as groupings of decreasing resolution between amino acids.

5.3.3 Method summary

Figure 5.4 summarises the process of tolerance labelling by a TSL classifier. Given a test protein, surface patches are created. A surface patch is described as a patch description string. This string is then translated using a grouping scheme. The patch description string is then used to search against the TSL which has been translated using the same grouping scheme. If this string is found in the library, it is labelled ‘tolerated’; otherwise it is labelled ‘untolerated’.

Thus, a **TSL classifier** consists of two components: the TSL library against which to search patch description strings and the grouping scheme used to translate the library and patch to be classified.

5.4 Testing tolerance classifiers

It was hypothesised that anti-correlation between the tolerance and epitope state of antigen surface would allow a TSL classifier to improve epitope prediction when in combination with IntPred:Epi. The set of classifiers must be tested on a set of antigen bound to human or mouse-antibody. In the following sections, the TSL classifiers are tested both on their own and in combination with IntPred:Epi. All of the TSLs described in table 5.2 are tested in combination with grouping schemes 1–7 shown in figure 5.4. Grouping scheme 0 was omitted due to its similarity to using no grouping scheme.

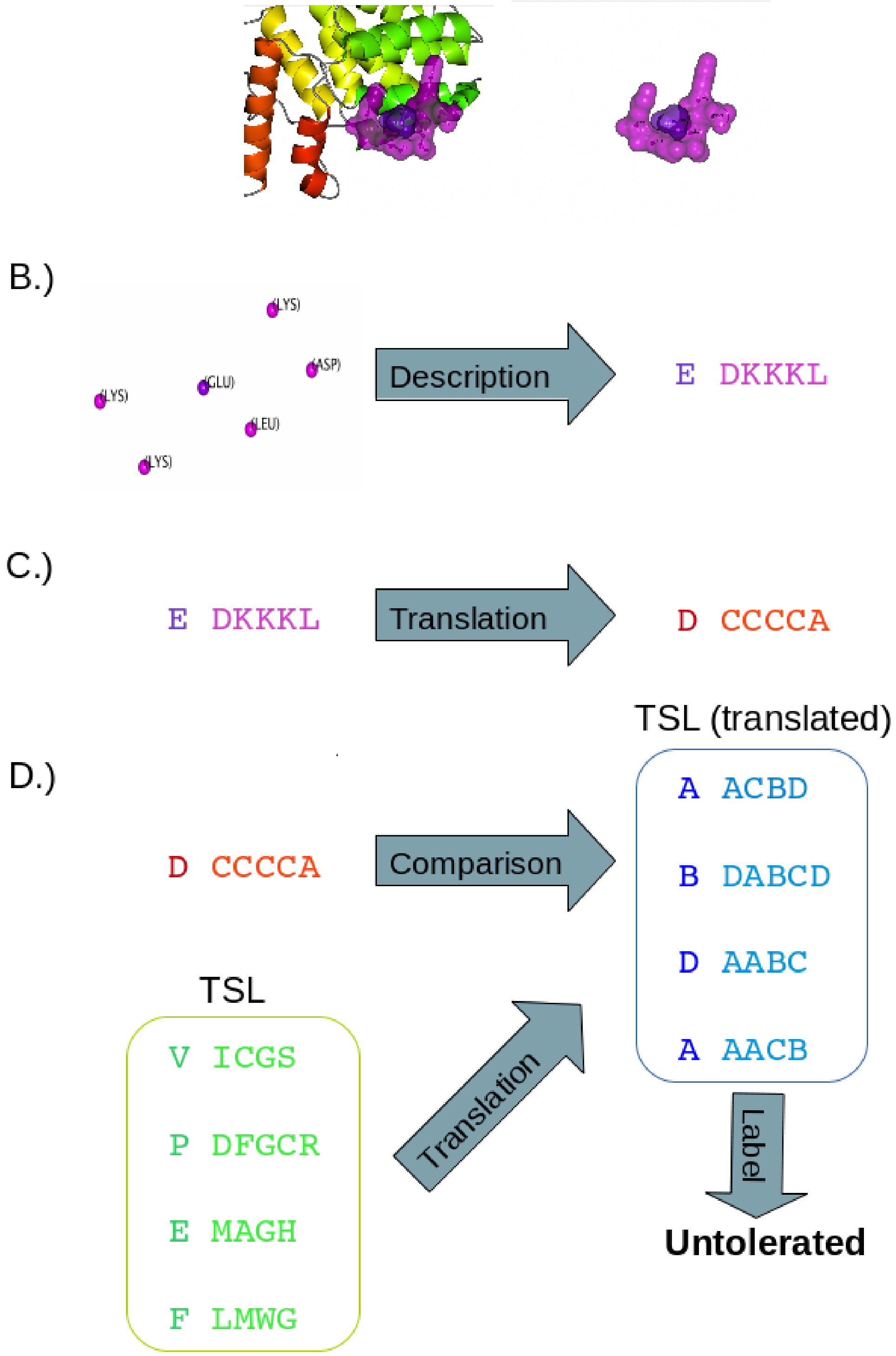

Figure 5.5: TSL classifier labelling. A.) Surface patches are created from the the test antigen. B.) A patch is described as a patch description string. C.) The patch description string is translated using a Taylor grouping scheme. D.) The translated patch description is compared against the patch description strings of a tolerated surface library that been translated using the same Taylor grouping scheme. If the patch description string is found in the library, the patch is labelled tolerated; otherwise it is labelled untolerated.

5.4.1 Antigen test sets

A dataset of antigens was required to test the performance of our TSL classifiers. The antigens were required to be bound by human antibody as TSL classifiers are host-species specific. The creation of such sets is described in section 4.5.3.1. As mentioned in the previous section, structures of human/mouse proteins bound to human antibody were not included — this is because our concept only holds under the assumption that tolerance is correctly regulated, i.e. host antibodies are not generated against self protein.

As shown in table 4.15, two data sets were created for each host: one larger set that overlaps with the training set used to train the baseline learners (IntPred:Epi and SEPPA 2.0) and a subset of this set — known as the *complement* set — that does not overlap. Testing on the *complement* sets gives us a measure of performance that can be compared fairly with SEPPA 2.0 and IntPred:Epi, whilst the larger sets can be used for assessing the performance of the TSL classifiers alone (see section 5.4.3), as none of these data have been used to train them in any way. The larger sets can also be used to assess the effect of combining IntPred:Epi with TSL classifiers (see section 5.4.4). Though the IntPred:Epi predictions will be over-fitted to those chains found in the training set, its performance when combined with the TSL classifiers on the larger serves as a useful indicator of their utility.

5.4.1.1 Avoiding unnatural antibody-antigen interactions

One of our aims was to remove any antigen bound to ‘unnatural’ antibodies from the antigen sets. This is because the method is relevant to the natural human or mouse immune response and thus the testing of methods on non-natural interfaces involving point mutated or phage display antibodies should be avoided. This selection is more difficult than it first appears. For a given antibody-antigen complex, it was hoped that parsing the PDB header for the ‘engineered’ line would tell us if the antibody had been engineered in some way. However, the definition of ‘engineered’ includes modifications to constant-region sequence such as the use of a Fab fragment rather than a full antibody (techniques which are quite common for the crystallisation of antibodies). Unlike those techniques that lead to a non-natural antibody specificity, these engineering techniques are not of concern. Thus, identifying antibodies with non-natural specificities is non-trivial. Going back to the literature would be the most comprehensive method, though this is laborious and impractical for the purposes of automation. It may be that the IEDB annotation schema contains the information required to discern human and mouse antibodies with non-natural specificity. This could be investigated in future work.

5.4.2 TSL classifier patch/residue labelling

A TSL Classifier is used to label surface patches as 'tolerated' or 'untolerated', as described in section 5.3.3. As well as patch labels, residue labels can generated. For a given chain, surface patches can be overlapping which means that a residue may occur in more than one patch. A residue can be individually labelled by considering the set of patches that contains that residue. A residue is labelled as 'tolerated' if it is found in any patch labelled 'tolerated' — otherwise it is labelled as 'untolerated'.

5.4.3 TSL classifiers as BCE prediction methods

Before combining the TSL classifiers with IntPred:Epi, their performance alone was investigated. Here, a residue is predicted non-epitope if it is labelled tolerated and epitope if it is labelled untolerated. For each TSL, labels produced by applying Taylor groupings 1–7 were applied to the human and mouse test sets. The results are shown in figure 5.6.

For the human set, TSL *all-m* classifier gives the best performance. In fact, this performance is comparable with the performance of some general BCE prediction methods (MCC = 0.08, compared with the performances shown in table 4.5.2). TSL *all-m* outperforms the remaining three libraries when used with all grouping schemes except scheme 6. Focusing on the first five schemes, a general pattern of increasing performance with increasing library size is observed (see table 5.2). The pattern does not hold for schemes 6 and 7. However, none the TSLs perform well with these schemes. Each successive grouping scheme is more likely to label a patch as non-epitope — thus, moving from grouping scheme 1 to 7, positive (i.e. epitope) prediction rate falls and higher specificity is traded for lower sensitivity. For grouping scheme 1-5, this trade-off results in similar overall performance. For grouping schemes 6 and 7, sensitivity is too low for further gains in specificity to maintain any performance.

Performance is markedly lower for the mouse test set. No correlation can be seen between TSL library size and performance. It should be noted that all mouse TSLs except mouse-all-m are smaller than the smallest human TSL, human-ec (108,448 patches, see table 5.2). TSL mouse-all-m is approximately twice the size as human-ec but is still less than half the size of human-all-m (474,794 patches).

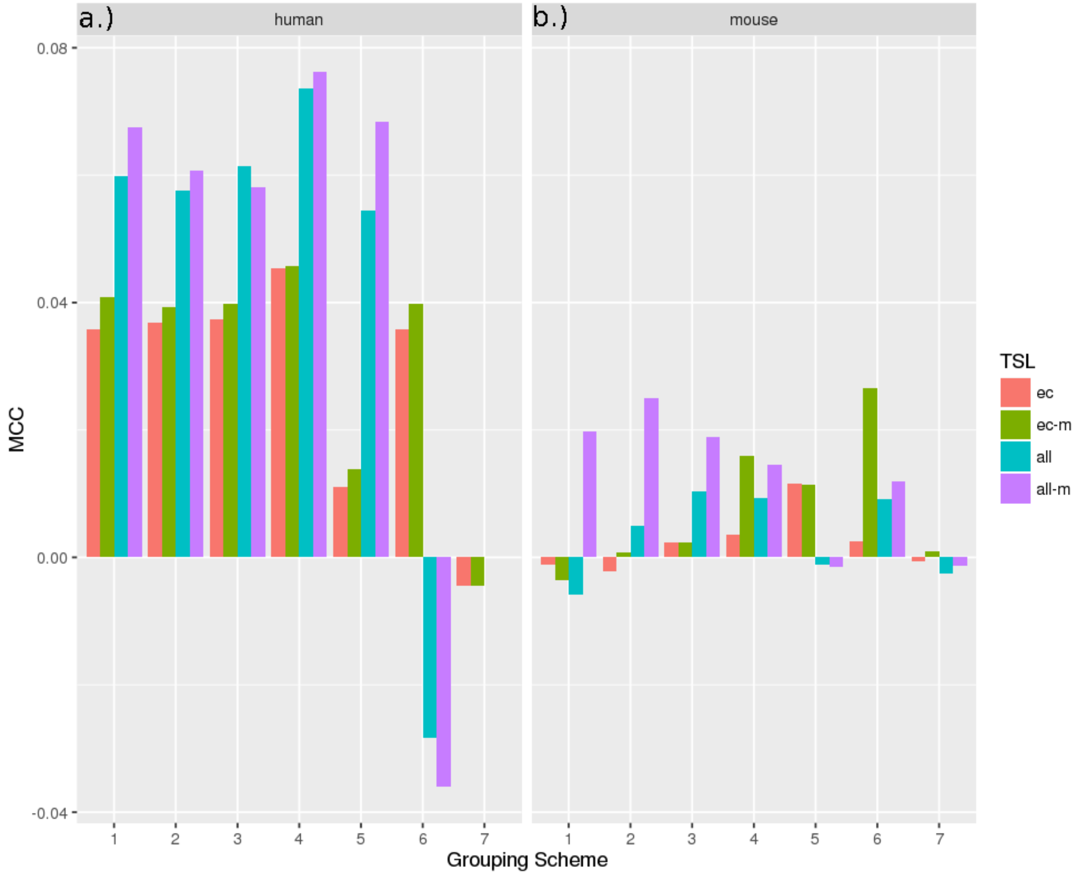

Figure 5.6: Performance of TSL classifiers on the a.) human and b.) mouse sets using the (*human-* or *mouse-*) *ec* (red), *ec-m* (green), *all* (blue) and *all-m* (purple) TSLs (see table 5.2 for more details).

5.4.4 TSL classifiers as BCE prediction filters

It was hypothesised that a TSL classifier would improve the performance of IntPred:Epi if it was used to filter output prediction labels. Specifically, all residues labelled tolerated would be predicted non-epitope and the remaining residues would be predicted on by IntPred:Epi as normal. This process should decrease the false positive rate and therefore improve overall performance.

The *human-all-m* and *mouse-all-m* TSLs were taken forward for testing, as they had in general performed the best on their own. The seven grouping schemes tested in section 5.4.3 were tested with each TSL. For each host-species, two test sets were used for prediction: one larger set that included chains found in the SEPPA 2.0 training set and a smaller set with these chains removed (see table 4.15).

The results of the filtering method are shown in figure 5.7. For the human and human-complement sets, performance is improved with all grouping schemes except scheme 7 when combined with IntPred:Epi. On the human set, schemes 1–5 in combination with

Table 5.3: Filtered IntPred:Epi performance on human and mouse test sets. The sets referred to here are the human and mouse-complement sets as described in section 4.5.3.1. For the human complement set, Filter refers to the human-all-m TSL combined with grouping scheme 5. For the mouse set, Filter refers to the mouse-all-m TSL combined with grouping scheme 1.

Test Set	Classifier	Sens.	Spec.	PPV	FDR	FPR	MCC
Human	IntPred:Epi	0.4725	0.6225	0.1502	0.8498	0.3775	0.0567
	IntPred:Epi + Filter	0.3340	0.7948	0.1518	0.8482	0.2052	0.0845
Mouse	IntPred:Epi	0.5176	0.6443	0.1933	0.8067	0.3557	0.1104
	IntPred:Epi + Filter	0.4972	0.6639	0.1967	0.8033	0.3361	0.1118

IntPred:Epi manage to exceed the performance of IntPred:Epi alone, with group 5 showing the biggest improvement (MCC = 0.0966, compared with 0.0728 for IntPred:Epi alone). The same is seen for the human-complement set, where classifier 5 in combination with IntPred:Epi has an MCC = 0.0845, in contrast to 0.0567 for IntPred:Epi alone. Although the same improvement in performance in comparison to the classifiers alone is seen for the mouse sets, no marked improvement is seen in comparison to IntPred:Epi alone. This is perhaps unsurprising, considering the poor performance of the TSL classifiers alone.

Although TSL *human-all-m* performed best when combined with grouping scheme 4 as a classifier, better performance was given by grouping scheme 5 when used as a filter. Considering that a TSL filter reduces the number of false positive labels (as well the number of true positive labels), this means that the scheme 5 filter removes a more optimal number of false positives to complement IntPred:Epi performance.

5.4.5 TSL labels as features in a machine learning method

As well as using TSL Classifier labels as a simple filter, they may also be used as an additional feature in the machine learning process. It was hypothesised that inclusion of a tolerance feature in the random learning process should yield more performance gains than a TSL filter. This is because a TSL filter acts like a new node added to the top of every tree in the random forest — any instance labelled as tolerated is labelled non-epitope by the tree at the 'filter node' and the remaining instances flow through the pre-existing tree. Properly incorporating a tolerance feature into the random forest learning process allows it to be treated like any other feature, allowing the established advantages of the random forest method to be gained.

For each of the training sets, IntPred:Epi was retrained twice: once with a tolerance feature and once without to provide a baseline. As IntPred:Epi predicts on patches, patch tolerance values had to be calculated. The TSL tolerance labels of human-all-m TSL

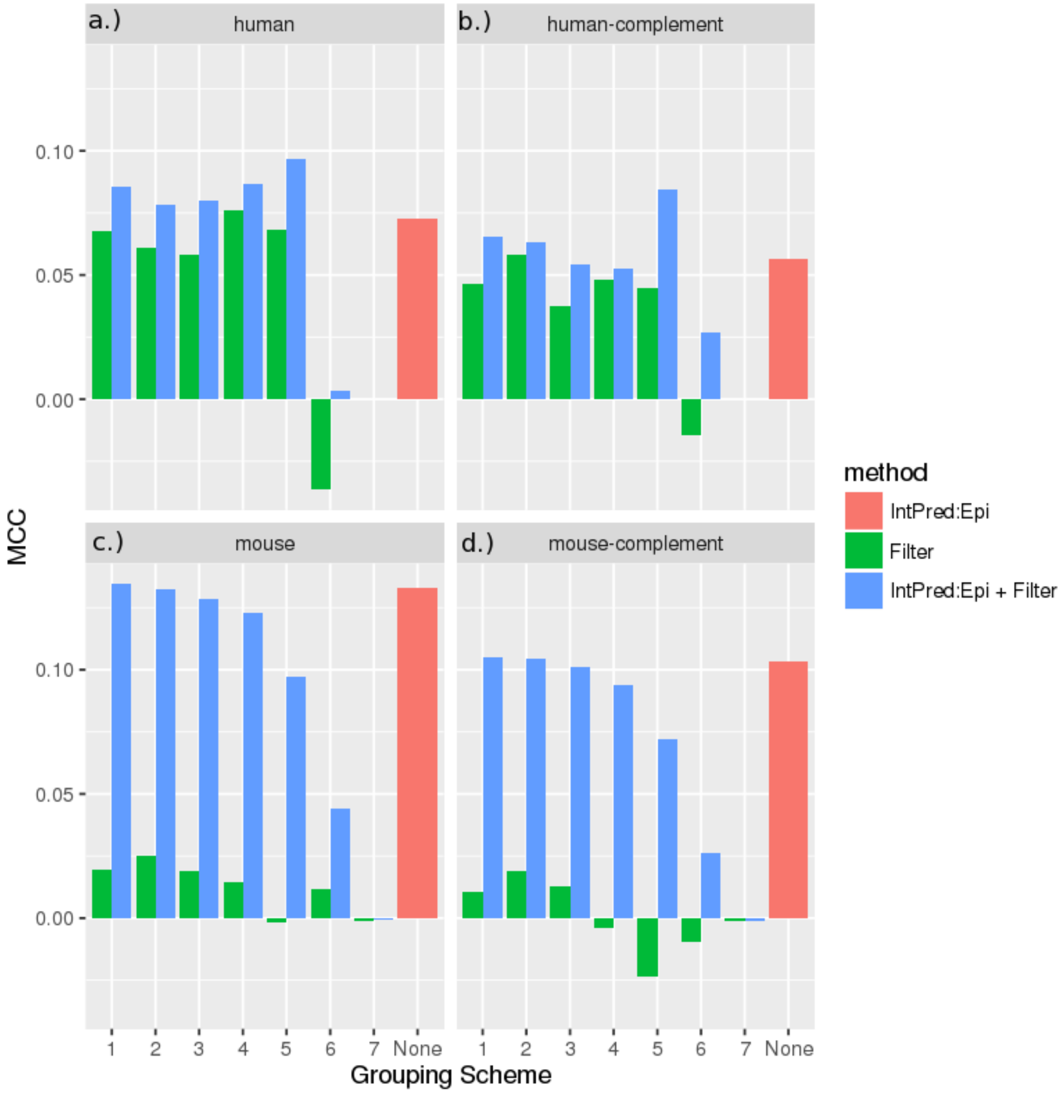

Figure 5.7: TSL classifier, TSL filter and IntPred:Epi performance. Each of the TSL classifiers is shown in green, with respective TSL filters in blue. The unfiltered performance of IntPred:Epi is shown in red. TSL *human-all-m* was used on a.) human b.) human-complement test sets; *mouse-all-m* was used on c.) mouse and d.) mouse-complement test sets.

classifier 5 and mouse-all-m TSL classifier 1 were used to calculate patch tolerance values for the human and mouse sets respectively. Patch tolerance values were calculated by dividing the number of patch residues labelled as tolerated by the total number of patch residues. by-chain CV was used to assess performance of the learners. The performances are shown in table 5.4. Surprisingly, the human baseline learner shows virtually no performance, with an MCC close to 0. This is in contrast to the mouse baseline, which performs well in comparison with the performance shown by IntPred:Epi during by-chain CV (see section 4.2.2.3). For both sets, including patch tolerance values as a feature fails to improve performance.

Table 5.4: TSL tolerance labels applied as a feature in random forests. Random forests were trained using a patch tolerance value derived from TSL human-all-m grouping scheme 5 and mouse-all-m grouping scheme 2 residue tolerance labels.

Training Set	Tolerance Labels	Sens.	Spec.	PPV	FDR	FPR	MCC
human	No	0.5448	0.4801	0.1081	0.8919	0.5199	0.0094
	Yes	0.5354	0.4855	0.1082	0.8918	0.5145	0.0071
mouse	No	0.6736	0.5091	0.1592	0.8408	0.4909	0.1086
	Yes	0.6773	0.5048	0.1594	0.8406	0.4952	0.1089

5.5 Discussion

A review of the mechanisms of tolerance allows proper evaluation of what a TSL represents. If crystal structures for all host proteins within the PDB are included, then all the self patches available to us are included. This will include soluble intracellular and extracellular proteins, along with membrane associated proteins (both cell surface and other). The first issue concerns the tolerance state of intracellular proteins. It may be that some intracellular proteins are presented to developing B cells via their release from apoptotic cells — in which case, clonal deletion or anergy may lead to tolerance of these proteins. On the other hand, studies mentioned previously indicate that certain intracellular proteins can indeed be immunogenic, though a number of these are in the context of an autoimmune disease or cancer state. Assuming intracellular proteins escape tolerance, an immune response could be raised against a foreign extracellular protein with a similar surface patch. If the foreign protein surface is displayed with high valence (e.g. in multimeric form) then it is possible that a T cell independent response could be raised. A T cell dependent response would require stimulation from helper T cells. Would the T cell population be tolerant to intracellular proteins? Although intracellular proteins are processed and presented as peptides to T cells via the MHC I pathway and therefore tolerated, B cells would present via the MHC II pathway. MHC II has different peptide affinities and therefore would present different peptides of a protein, compared with MHC I processing. Therefore it is possible that a T cell dependent response could be raised. Considering these factors, it was not initially clear how the inclusion of intracellular antigen would affect classifier performance. However, the superior performance of TSLs *all all-m* over *ec* and *ec-m* on the human test set suggests that including intracellular antigen is indeed beneficial for the identification of tolerated patches.

Considering the presence of anergised B cells and natural autoantibodies, it is clear that a TSL includes patches that can bind to host antibodies in a healthy context. It also contains patches that are tolerated through clonal deletion and are therefore not bound by host antibodies. Thus a TSL contains proteins that are immunogenic and/or antigenic. It is apparent from the study of autoimmune diseases that the distinction between these categories is dynamic and context dependent. Further limitations must be considered. Post-translational modifications are important factors in immunogenicity, as illustrated by granzyme B processing. Crystal structures do not allow us to include this structural multi-dimensionality. Immune privilege also adds an additional layer of complexity by defining a subset of self proteins that are not tolerated. Considering all these factors, it seems likely that mislabelling events will occur that set a limit on potential performance. Despite this, predictive power was exhibited by the TSL *all* and TSL *all-m* classifiers on the human set that is similar to the performance of IntPred:Epi alone. On the human sets, IntPred:Epi performance was improved when combined with TSL *all-m* grouping scheme 5 as a filter. The same wasn't seen on the mouse sets. This may be because the predictive power of the TSL classifier alone wasn't enough to helpfully complement the predictions made by IntPred:Epi. The poor performance of the TSL classifiers on the mouse data sets may be due to the relatively small sizes of the TSL libraries. However, even TSL *mouse-all-m* yields poor performances in comparison to TSL *human-ec*, despite being approximately double its size.

It was hypothesised that IntPred:Epi retrained with an additional tolerance label would outperform IntPred:Epi in combination with a TSL filter. However, IntPred:Epi failed to show any predictive power when retrained on the human set and no improvement when retrained on the mouse set. Although the lack of improvement when retraining on the mouse-set with an additional tolerance label is perhaps unsurprising considering the lack of improvement using mouse-library TSLs filters, is it unclear why IntPred:Epi does no better than random when trained on the human-host set. It might be that IntPred:Epi fails to learn from the human set because the data set is simply too small (18 chains, in comparison to 53 chains in the mouse set). On the other hand, the failure to learn from the human set, as well as the poorer baseline performance of IntPred:Epi without retraining on the human-complement set (see section 4.5.3.2), may be due to some feature of human-host epitopes (see section 4.6 for a brief discussion).

5.5.1 Future directions

Future work should include investigating successful and unsuccessful predictions, in order to identify any patterns amongst false positive and true negative predictions. Specifically, it is possible to identify which host protein surfaces match with the antigen tested. Is there a relationship between the host proteins that false positives match with? Do true negatives match to certain types of proteins? If there is a protein with a similar structure to the test antigen within the TSL, are true negatives found that match to sequence identical areas of the structures? If there are true positives (i.e labelled untolerated), do these correspond to areas of sequence difference between the test antigen and the similar structure? Future work should address these questions. Additionally, the impact of including more modelled structures should be assessed, particularly for mouse TSLs where very low predictive power was seen.

The method of patch comparison through description string matching is rapid but unrefined. The similarity of patches with matching patch description strings, both before and after translation, should be investigated. A simple example measure of similarity would be RMSD. Additionally, a more sophisticated method of surface comparison could be implemented. One study has used graph theory to compare protein surfaces in order to identify determinates of cross-reactivity (Iakhiaev and Iakhiaev, 2010). The method consists of creating an association graph between the surface accessible residues of two proteins and finding the maximum cliques (i.e. the most similar surfaces between the two). As a proof of concept, it was shown that the method could identify known cross-reactive sites on coagulation factor VII and anticoagulant protein C, proteins with high sequence similarity. Furthermore, the method was used to compare a group of coagulation proteins to beta2-glycoprotein-I (beta2-GPI), the antigen known to induce antiphospholipid antibodies in antiphospholipid syndrome (APS). The autoantibody epitopic sites of beta2-GPI have been well-characterised and although coagulation proteins are known to be cross-reactive with beta2-GPI, low sequence similarity makes identifying cross-reactive epitope sites difficult. The method was able to identify similar sites across all of the coagulation proteins in question and beta2-GPI. Moreover, the similar site on beta2-GPI overlapped with known autoantibody epitope sites, revealing potential epitope sites on the group of coagulation proteins. Thus this work demonstrates that cross-reactivity can be identified via surface comparison. Scaling up of the method presented by Iakhiaev and Iakhiaev (2010) to compare many human proteins against non-human proteins could be a viable alternative to the patch description string comparison.

5.6 Conclusion

In this chapter, a method was presented that tailors the BCE prediction problem to a host — in this case, either human or mouse. Applying the concepts of immune tolerance, libraries of tolerated human and mouse protein surface patches were created, as well as methods that allow the description and comparison of surface patches in order to label test antigen patches as either tolerated or non-tolerated. These labels were then applied to BCE prediction in a number of ways: on their own, as a filter and as features for learning. Although BCE prediction could not be improved for mouse-host epitopes, an improvement was seen for human-host epitopes. Development of the patch description and comparison methods, as well as extension of TSLs with more solved and/or modelled structures may improve future performance. Additionally, careful investigation of patches, their prediction labels and the library structures that have influenced the labelling process may help to elucidate other useful patterns.

Chapter 6

Investigating Sequence Determinants of Antibody Stability

This chapter presents an analysis of a set of natural V_H-V_L pair human Fabs. Using sequence and biophyiscal data, sequence features are identified that have the potential to be applied in a future machine learning method that predicts the biophysical stability of an antibody from its sequence.

6.1 Introduction

The biophysical stability of any protein therapeutic is important because of its influence across many aspects of its behaviour. It can influence the efficacy and immunogenicity of a therapeutic, as well as affecting shelf-life and storage requirements, potentially limiting its use in areas where a cold chain process is not possible. Thus the prediction of biophysical properties would allow the identification of candidate therapeutic antibodies that have a higher risk of failing early and late-phase drug development.

6.1.1 Biophysical stability

A number of interrelated properties exist that describe biophysical stability. The most primary properties of biophysical stability relate to the free energy of the folded state of the antibody: a greater free energy of folding results in a more stable antibody. The free energy of folding influences the behaviour of an antibody under changes in temperature (thermal stability) and solvent. The **melting temperature** (**Tm**) is commonly used as a measurement of thermal stability of a protein and is the temperature at which 50% of the protein is unfolded. Another commonly used measure is the denaturation midpoint, which is the concentration of denaturant required to unfold 50% of the protein.

Another important measure of biophysical stability is the surface hydrophobicity of an antibody. Hydrophobic interactions can lead to the formation of aggregates as well as problems with high viscosity. However, hydrophobic interactions are important in the binding of antigen (Soltis and Hasz, 1982). Thus an antibody commonly presents some amount of hydrophobic surface but this will vary between antibodies. The surface hydrophobicity of an antibody can be measured by performing a **hydrophobic interaction chromatography** (**HIC**) assay, which measures the amount of time that an antibody spends on a hydrophobic column under decreasing salt concentration. The presentation of more hydrophobic surfaces will naturally lead to greater interaction with the hydrophobic column and therefore the antibody will take longer to elute as salt concentration is decreased.

Other properties can be described that relate to antibodies in the context of their use as therapeutics. Therapeutic antibodies must be administered at high concentrations, which can lead to problems with high viscosity, making solutions difficult to manufacture and administer (Shire et al., 2004). Resistance to chemical degradations such as asparagine deamidation and asparate isomerization is also important as it affects the shelf-life of a therapeutic (Beck et al., 2013). Recent work has provided evidence for these more specific properties being significantly correlated with basic properties such as hydrophobicity, net charge and residue solvent exposures (Sharma et al., 2014).

6.1.2 Biophysical properties of antibodies

Despite sharing a common structural framework, the biophysical properties of antibodies can differ markedly. This is unsurprising considering the high sequence variability between antibodies, a consequence of domain variability, V(D)J recombination and somatic hypermutation. Despite this high variability, studies have found correlations between sequence and structural features and biophysical properties. In particular, much work has been done assessing the biophysical properties of variable domains, both isolated and in combination in the form of an scFv fragment. Ewert et al. (2003) studied the biophysical properties of single antibody variable domains having consensus sequences of the seven major human germline subclasses. They found that the most stable single V_H and V_L domains were V_H3 and $V_\kappa3$ respectively, whilst V_H6, $V_\kappa2$ and V_λ were the least stable. An assessment of a range of structural factors found that low stability seemed to be correlated with a host of small defects, including poor hydrophobic core packing, disruption of ionic interactions and exposed hydrophobic residues. They then went on to assess the effect of pairing each V_H and V_L domain with the most stable single V_L and V_H domains. They found that pairing with the most stable partner reduced the inherent difference between domains, with low stability V_λ domains in particular being rescued by V_H3. This illustrates the influence of the V_H/V_L pairing on determining the stability of an antibody.

The influence of V_H/V_L pairing on biophysical stability was more comprehensively investigated by Tiller et al. (2013). They constructed a fully synthetic Fab library with favourable biophysical properties. In order to narrow down library selection from the potential 3200 pairs available, they chose 20 V_H and 20 V_L subclasses that were found with high frequency in natural and engineered repertoires. From 400 V_H/V_L combinations, 36 were chosen for their superior biophysical and expression profiles. 12 V_H and 15 V_L subclasses make up these 36 pairs. Interestingly, they found that two of the four V_H1 members paired with $V_\kappa3$, which corresponds to one of the pair preferences shown by Jayaram et al. (2012). They also found that pairs containing V_κ tended to have higher Tms, in comparison to V_λ-containing pairs.

Despite the superior stability of certain variable domain subclasses, this does not guarantee a stable antibody. A study by Honegger et al. (2009) showed that the grafting of CDRs from a murine V_H9 domain onto a V_H3 framework resulted in a poor stability scFv. They were able to develop a high-stability scFv by producing a hybrid containing murine V_H9, human V_H1 and and V_H5 elements and concluded that the optimal framework was CDR-dependent. This illustrates that the diverse range of structural frameworks available to the human immune system is partly to increase the range of stable framework-CDR combinations.

The nature of the antibody fragment type is also important in influencing biophysical stability. The simplest fragments, such as single V_H domains, tend to face problems with aggregation due to the exposure of hydrophobic patches normally involved in the V_H/V_L interface (Barthelemy et al., 2008). Moreover, a comparative study of the same variable regions in Fab and scFv formats found that the V_H/V_L and C_H1/C_L interfaces were mutually stabilised in the Fab fragment (Rothlisberger et al., 2005). Consequently the Fab was more kinetically stable, in particular having a much slower unfolding rate. Furthermore, the disulfide linking C_H1 and C_L domains in the Fab was essential for this effect. This illustrates that the factors determined as being important for the stability of a given antibody fragment type may not necessarily be as important in other types.

6.1.3 Approaches to engineering antibody stability

A number of different approaches exist for the engineering of stable antibodies and related domains/fragments. One of the earliest approaches considered the sequence statistics of the V_κ domain in order to predict stabilizing mutations (Steipe et al., 1994). Here the principle is that selection can be inferred from the amino acid frequencies at each sequence position in the V_κ domain: if an amino acid is seen at a position less than expected by chance, then it may be selected against to maintain stability. Point mutations were made in the hydrophobic core, V_H/V_κ interface, V_κ-C_H interface, CDRL1 and CDRL3 to change low-frequency residues to their high-frequency consensus counterparts. From 9 positively predicted and 1 negatively predicted mutation, they obtained a 60% success rate when the direction of the prediction was considered (rather than the magnitude). Since then, the same principle has been applied to a number of different proteins, including intrabodies (Steipe, 2004). Though the consensus approach has proven to be useful, it is limited. In the case of one antibody, three stabilising mutations were found that mutated their respective positions away from the consensus (Wang et al., 2013). This suggests that some mutations are stabilising in the context of other mutations that have been made to the antibody.

Directed evolution methods have also been applied to increase antibody stability. Directed evolution methods such as phage or yeast display are commonly used for the identification of antibody fragments with some desired specificity. The general principle is to apply some selective pressure (e.g. the ability to bind an immobilised antigen) on to a population (e.g. M13 bacteriophage) whose members each express an antibody fragment and then display it on its surface. During each round of replication, each member has a random chance of mutating so that it binds with higher affinity to the selected antigen. Directed evolution methods are most commonly used to identify antibody fragments that bind with high affinity to some target, but the same principle can also be applied to select for biophysical stability. Stability selection can be carried out by adding a stress step (e.g. heating, guanidium chloride) to inactivate low stability members before the selection step. This has been applied to a number of different targets to select for different biophysical properties (Jermutus et al., 2001, Jespers et al., 2004). Furthermore, in the case of single V_H domains, the sequences of selected (biophysically stable) clones have been compared to unselected clones to identify sequence determinants of biophysically stability (Barthelemy et al., 2008, Dudgeon et al., 2009).

6.1.4 Somatic hypermutation and antibody stability

The combination of somatic hypermutation and clonal selection leads to the production of high affinity antibodies. However, somatic hypermutation has also been implicated in the maintenance of antibody stability. Studying a small hapten-binding antibody, Wang et al. (2013) grouped somatic mutations into those directly involved in antigen binding and those peripheral mutations that are not. By making antigen binding mutations to the germline precusor, they found that melting temperature decreased by 9 °C in comparison to germline. They then introduced the peripheral mutations and found that thermal stability returned close to germline. A structural analysis found that while the mutations involved in antigen binding were clustered close to the V_H/V_L interface, the peripheral stabilising mutations were found distal to the interface. Importantly, the peripheral mutations seemed mainly to affect the structures of the loop regions connecting the β-strands, which themselves play an important role in the V_H/V_L interface. This study provides strong evidence to support the idea that clonal selection selects for both affinity to antigen and antibody stability. This is reasonable, considering that protein expression levels often correlate with thermal stability and that the level of antibody on the surface of a B cell affects the avidity of which it is able to bind to antigen (which in turn leads to greater activation).

6.1.5 Aim

Previous studies have been able to elucidate mechanisms by which sequence can influence the stability of an antibody. In particular, existing methods seem to be effective when the aim is to improve the stability of an antibody via the introduction of one or more point mutations (Monsellier and Bedouelle, 2006). However, these methods are inherently limited because they make predictions about the effect of mutations to a *single* antibody. Their predictions cannot tell us anything useful about, for example, two very different antibodies.

Directed evolution methods have the potential to generate large datasets of sequence data correlated with biophysical stability (e.g. Dudgeon et al. (2009), though only 80 data points are obtained). However, to illustrate why directed evolution methods selecting for biophysical stability are limited, let us consider two selection methods where i) stability alone is selected for and ii) stability is selected for whilst maintaining affinity to a target antigen. In the former case, improvements in stability may be to the detriment of antigen binding ‘potential’. In other words, it is hypothetically possible that very stabilising mutations identified in this fashion may not be compatible with mutations required for antigen binding, either in the sense that a position mutated for stability is required to have a different mutation for antigen binding, or in the sense that some of the effect of a stabilising mutant is negated by an antigen-binding mutation at some other position. In the latter case, mutations that lead to the improvement of stability may only be applicable in the context of antibodies that bind to the specific antigen being used. It may be that mutations found to be stabilising in a set of antibodies specific for one antigen do not have the same effect in a set of antibodies specific for a different antigens. Over the course of many directed evolution experiments with a range of different antigen, it may be possible to infer general mechanisms of increased stability, but this is obviously time consuming. Furthermore, phage display methods in particular are limited by how effectively different subtypes can be expressed. For example, V_H2, V_H4 and $V_\kappa4$ are known to express poorly in phage and therefore are often absent from libraries (Tiller et al., 2013). This further limits the insights that can be taken from phage display studies.

Further to this, many studies focus on the improving the biophysical stability of fragments or domains of antibodies (e.g. V_H, scFv). In some cases, the insights from these studies will be limited to the domain or fragment studied. For example, in the cases of single domains, it is clear that the replacement of hydrophobic residues that are normally involved in the V_H/V_L interface are very unlikely to improve the stability of an scFv or Fab! Therefore the patterns seen for smaller domains and fragments cannot be confidently extrapolated to full antibodies or a larger fragment (e.g. Fab). Moreover, studies such as Ewert et al. (2003) and Monsellier and Bedouelle (2006) limit their explorations so that variable region sequences are fixed. It is clear however that the variable regions are also important in modulating stability (Wang et al., 2013).

Fab fragments are advantageous in that they provide multi-valency and also stability. Despite the work done on single domains and scFvs, studies on Fab stability are limited. With this in mind, the aim of this study was to analyse a data set of human antibody Fab fragments with known sequence on which a set of biophysical assays had been performed (see section 6.2.1). These Fab fragments are naturally paired, so the data set captures the biophysical space explored *in vivo*. Furthermore, the dataset used is derived from the serum of healthy human donor memory IgG B cells and as such is without bias towards an antigen. In contrast to previous methods, antibodies are being sampled randomly and thus variable regions will vary and this will allow us to study the relationship between the variable region and stability. The aim of the analysis was to determine sequence features that correlate with biophysical stability that could be used to train a machine learning algorithm, thus giving us a model to predict biophysical parameters of any human Fab. Importantly, certain biophysical parameters such as thermal stability are not expected to be predicted with high accuracy - this is because the thermal stability can be influenced greatly by a single point mutation. However, if general trends can be identified across a large, diverse set of Fabs, then these can be used to select candidates from large panels; this is becoming more important with the advent of next-generation sequencing that vastly increases the space that can be searched. Moreover, the combination of narrowing down candidates and then applying the methods discussed above to apply point mutations predicted to be stabilising promises to be particularly powerful.

6.2 Data generation and processing

This section provides a summary of the Fab data generation method and details the data processing steps required for analysis.

6.2.1 Data generation

A set of human naturally paired variable regions were sequenced and expressed, and hydrophobicity and thermal stability data were generated according the method summarised here (and also in figure 6.1) — more detail is given in appendix A.)

Individual IgG B cells are isolated from the serum of a non-immunized human donor using fluorescence-activated cell sorting. Cells are sorted individually into wells, where each well should contain one B cell. For each isolated B cell, a reverse transcriptase reaction is used to generate cDNA from cell mRNA. Using cDNA as a template, two separate primary PCR reactions are undertaken to amplify the antibody variable regions — one for heavy and one for light. Secondary PCR reactions are used to amplify the variable regions further using a distinct set of primers, leading to cleaner product. Additionally, the reverse primers of the secondary reaction have overhanging ends that are complementary between heavy and light-chain variable region fragments. This allows a final tertiary PCR step in which the heavy and light-chain variable region fragments anneal to create a template for amplification. As a result, a fragment containing the heavy and light-chain variable regions is obtained (see figure 6.2).

V-region fragments are pooled and then ligated into a vector containing leader sequences and promoters for expression. Ligation requires purification of V-region fragments by gel electrophoresis followed by gel extraction. This process is impractical for many samples, so pooling of fragments is required to increase throughput. In order to re-isolate successfully ligated vectors, competent *E.coli* is then transformed with the vector pool and plated on medium that selects for successful transformation. Colonies are picked and cultured overnight, before plasmid DNA is extracted following a mini-prep protocol.

In order to obtain a full Fab fragment, the C_L of the light chain and C_H1 of the heavy chain must be inserted. This is done by inserting a fragment containing the two domains into

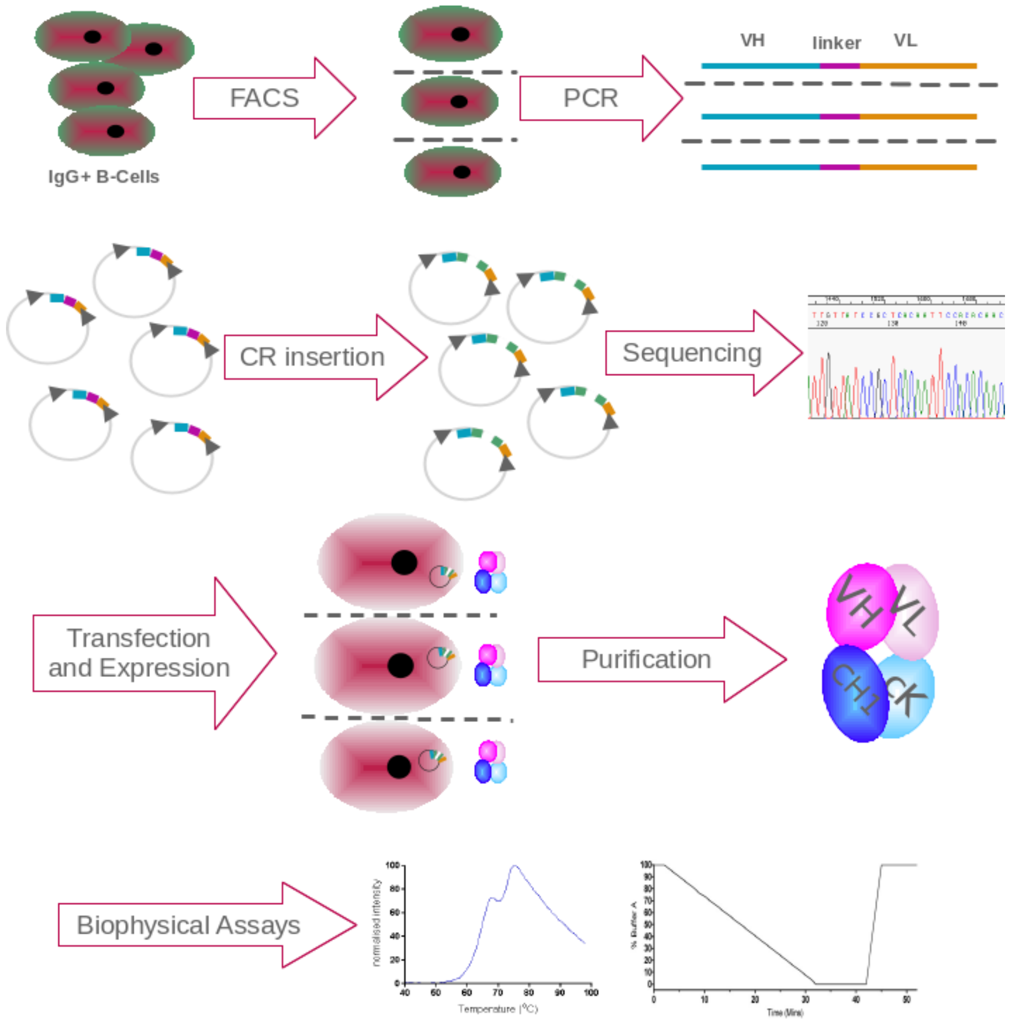

FIGURE 6.1: A schematic of the Fab data generation method. See main text for a summary and appendix A for details.

the V-region vector by ligation. As with the first ligation, V-region vectors are pooled to increase throughput. Complete V+C vectors are isolated by *E.coli* transformation followed by culturing and plasmid DNA preparation. Note also that the C_H1 domain has a polyhistidine-tag required for purification. At this point the V-regions are sequenced.

A mammalian expression system is used to express Fab fragments. For each V+C vector, ExpiHek cells are transfected with the vector and cultured. The supernatant is then isolated and purified for Fab fragment using the PhyNexus purification protocol for high throughput. Fab concentration and purity is then measured. As this point, a Fab is ready to be assayed for biophysical properties.

For each Fab a thermoflour assay is performed to obtain a Tm value. A Fab sample is mixed with a florescent dye that interacts with hydrophobic residues. By increasing the temperature and measuring the change in fluorescence, the temperature at which half of the Fab is unfolded (melting temperature) can be determined.

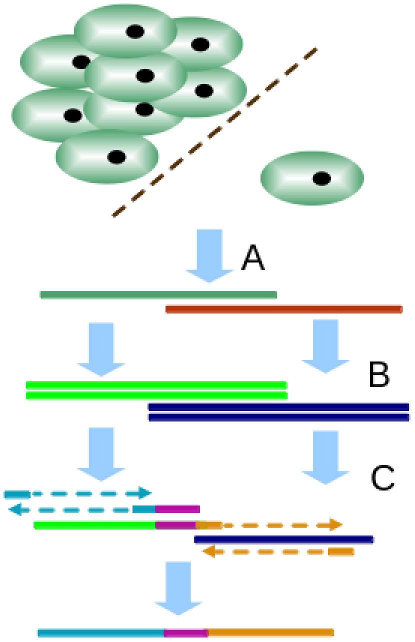

FIGURE 6.2: A schematic of V-region fragment PCR. A. Reverse transcriptase reaction on individual B-cell isolate, producing cDNA from cell mRNA. B. Primary and secondary PCR reactions amplify heavy and light-chain V-regions. C. Annealing of heavy and light-chain V-region fragments via overhanging ends of the reverse primers from secondary PCR reaction, followed by amplification.

A hydrophobic interaction chromatography (**HIC**) assay is also performed to obtain a HIC retention time. For this assay, a Fab sample is run onto a hydrophobic column under high salt concentration. As the salt concentration is gradually reduced, the Fab will elute from the column: the more hydrophobic the surface of the Fab is, the longer it is retained on the column before elution.

6.2.2 Data processing

Human Fabs data were processed for analysis. V-segment germlines were assigned to sequences using IgBLAST (Ye et al., 2013). V-segment germline sequences were obtained from IMGT (Lefranc et al., 2009). Standard numbering was then applied to the sequences using the *AbNum* program and the Chothia numbering scheme (Abhinandan and Martin, 2008). Sequence, biophysical and other experimental data were stored in a PostgreSQL database (see figure 6.3).

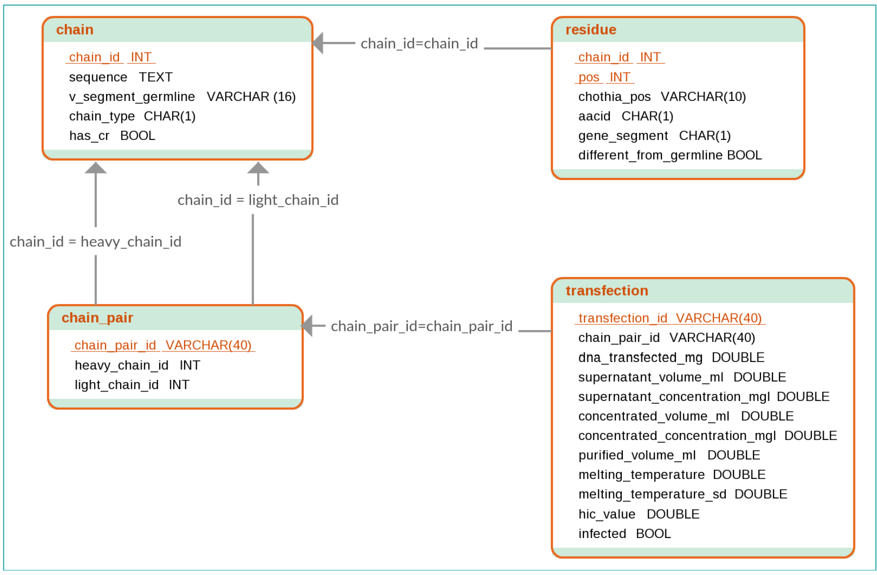
chain
chain_id INT
sequence TEXT
v_segment_germline VARCHAR (16)
chain_type CHAR(1)
has_cr BOOL
chain_id=chain_id
residue
chain_id INT
pos INT
chothia_pos VARCHAR(10)
aacid CHAR(1)
gene_segment CHAR(1)
different_from_germline BOOL
chain_id = light_chain_id
chain_id = heavy_chain_id
transfection
transfection_id VARCHAR(40)
chain_pair_id VARCHAR(40)
dna_transfected_mg DOUBLE
supernatant_volume_ml DOUBLE
supernatant_concentration_mgl DOUBLE
concentrated_volume_ml DOUBLE
concentrated_concentration_mgl DOUBLE
purified_volume_ml DOUBLE
melting_temperature DOUBLE
melting_temperature_sd DOUBLE
hic_value DOUBLE
infected BOOL
chain_pair
chain_pair_id VARCHAR(40)
heavy_chain_id INT
light_chain_id INT
chain_pair_id=chain_pair_id

FIGURE 6.3: Fab database physical data model. Database is implemented in PostgreSQL.

6.3 Sequence analysis

In total, 61 unique antibodies were sequenced. The V-gene assignments for the heavy and light chains are shown in figure 6.4. Note that only kappa-light primers are used during the PCR of sequences from isolated B cells, so no lambda-light chains are observed. The majority of heavy chains are assigned to IGHV3, while light chains are mostly either IGKV1 or IGKV3. This leads to most pairings being IGHV3-IGKV1 or IGHV3-IGKV3. A chi-squared test was performed on the pairings to test for any preferences. Because of the limited data, light V genes IGKV2, 2D and V3 were grouped, as well as heavy V genes IGHV1, 2, 4 and 5. From this grouped table, a Chi-squared test results in $\chi^2 = 12, d.f. = 9, p = 0.21$. This does not provide any evidence that the null hypothesis that V-gene germlines show pairing preferences is incorrect. However, future data will increase the sample size and therefore increase the statistical power of the test.

Figure 6.5 shows the distribution of V-gene identities, where V-gene identity of an antibody is calculated as the average identity of its heavy and light V-gene regions to germline. The median V-gene identity is approximately 90%, which is around 20 mutations across both V-gene segments.

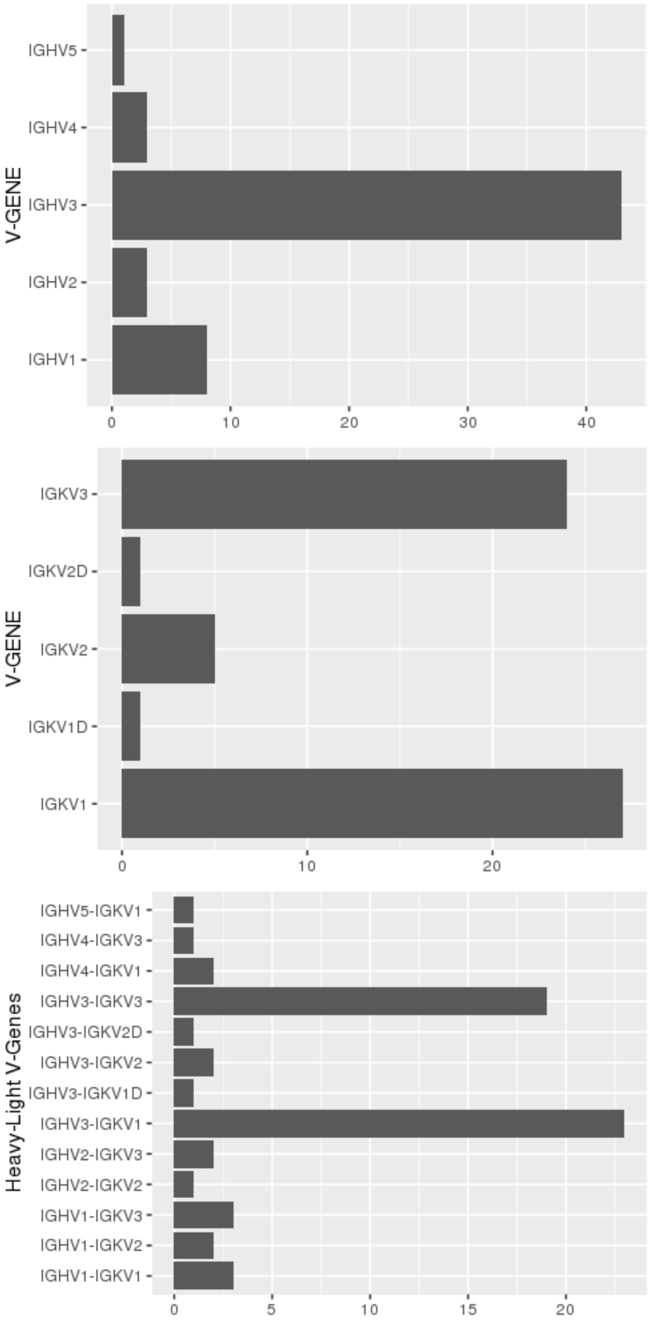

Figure 6.4: V-gene assignments. V-genes were assigned to each V_H and V_L sequence using IgBLAST (Ye et al., 2013), with V-Gene sequences obtained from IMGT (Lefranc et al., 2009).

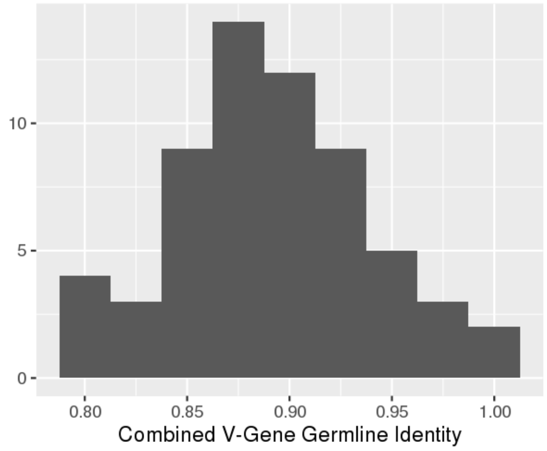

FIGURE 6.5: Combined V-gene germline identity. For each Fab, V_H and V_L V-Gene germline identity is averaged. Bin width = 0.0.25.

6.4 Biophysical data analysis

6.4.1 Measurements

For each antibody, two assays were undertaken to obtain measurements of biophysical stability. A thermoflour assay gives us the melting temperature (Tm), which corresponds to the temperature at which 50% of a protein is unfolded and is therefore a measure of thermal stability. A hydrophobic interaction chromatography (HIC) assay gives us the time for which a protein is retained on a hydrophobic column. The HIC retention time is therefore a measurement of how hydrophobic the surface of a protein is (see appendix A for more details on these assays).

37 antibodies were successfully assayed to obtain Tm values. Of these, 36 were also successfully assayed to obtain HIC retention times. Additionally, one antibody was assayed successfully for HIC retention time, but not a Tm.

Figure 6.6 shows the distribution of Tm values. Tm ranges from approximately 55 to 80 °C, with around 66% of antibodies having a Tm between 63 and 75 °C. A further 27% have a Tm greater than 75 °C, while only two antibodies ($\sim$ 5%) have a Tm less than 60 °C. As every Tm value is the mean of three repeated repeated measurements, each has an associated standard deviation. Figure 6.7 shows the Tm standard deviations. With the exception of one antibody, all standard deviations are less than 0.75 °C. In the following sections, antibodies will be grouped into low, medium and high-Tm groups; these are shown in figure 6.6. These groupings were simply decided according to the semi-discrete intervals seen in the data.

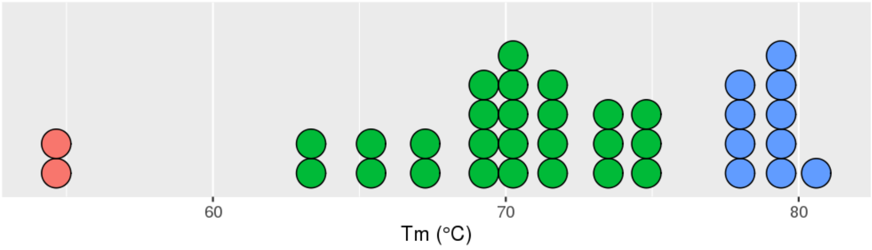

Figure 6.6: Tm data. Fabs are divided into low (red), medium (green) and high (blue) Tm-groups. These groups will be used in the proceeding analysis. Bin width = 1.

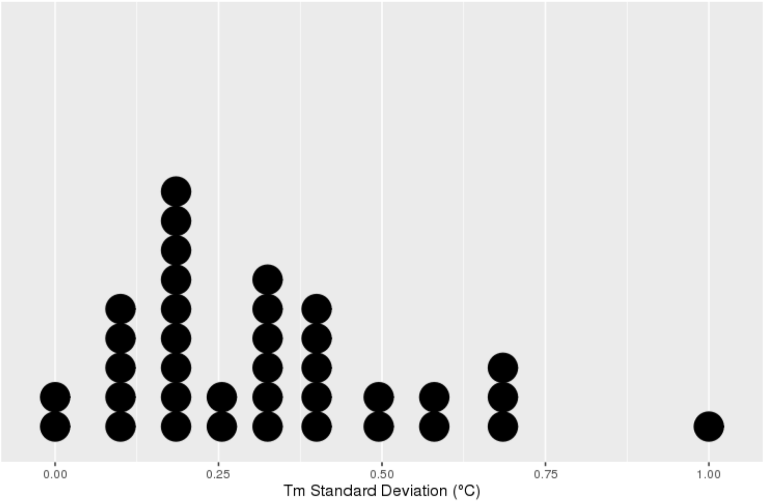

Figure 6.7: Tm Data Standard Deviations. Tm values are the average of three measurements and therefore have standard deviations. Bin width = 0.06.

Figure 6.8 shows the distribution of HIC values. The majority (approximately 80%) of antibodies have a retention time of less than 5 min. A further four antibodies have a retention time between 5 and 7.5 min, while the remaining four antibodies have a retention time greater than 10 min. Retention time is only measured once per antibody, so there is no associated standard deviation. Similarly to Tm, antibodies were grouped into short and long-retention groups according to the large interval seen in the data.

Note that no correlation was found between Tm and HIC retention time for those 36 antibodies where both assays were successful.

6.4.2 V-Gene germlines

The relationship between V-gene germline and hydrophobicity and HIC was investigated. Figure 6.9 shows the distribution of HIC values amongst each heavy and light-chain V-gene germlines. Three of the four long-retention antibodies have IGHV1 V-genes, despite this V-gene being assigned to only 8 of 37 (approx. 20%) of our antibodies. If the

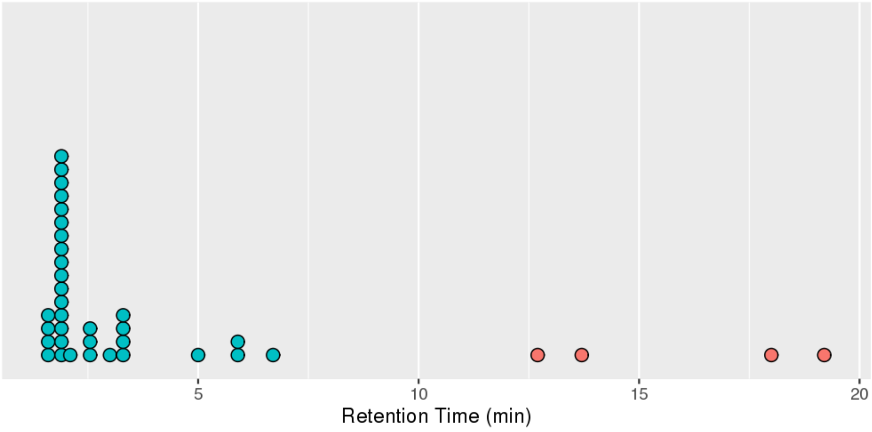

Figure 6.8: HIC retention time data. A longer HIC time corresponds to a longer elution time from a hydrophobic column. Fabs are split into short (blue) and long (red) HIC retention time Fabs. These groups are used in the proceeding analysis. Bin width = 0.3.

antibodies are split into two groups, IGHV1 assigned and not IGHV1 assigned, and then use retention time group as a categorical variable, it possible to perform a fisher's exact test which gives $p = 0.03$, suggesting that there is a correlation between hydrophobicity and IGHV1 germline.

Light V-gene and combined V-genes are also shown in figure 6.9. In contrast to heavy V-genes, there seems to be no relationship between light V-gene and retention time. Although three of the four long-retention time antibodies are assigned to IGKV1, approximately 50% of the antibodies have been assigned to this V-gene. Similarly, there does not appear to be a relationship between pairing and retention time. Considering the relatively large number of potential pairings, the limited sample size and the distribution of retention times (i.e. very few high-retention time antibodies), it is not possible to perform an effective statistical test on these data; in the future more data should provide the statistical power required to test for an effect.

The relationship between V-gene and Tm is shown in figure 6.10. In contrast to HIC retention time, no heavy V gene appears to be associated with Tm, with the two most represented heavy V-genes IGHV1 and IGHV3 both displaying the full range of observed Tm. However, for light V-genes it is observed that both low-Tm antibodies (Tm <60 °C) have V-gene IGKV2, despite only four antibodies (10%) having this V-gene. The same procedure that was applied to retention time and heavy V-gene was applied to IGKV2 and not-IGKV2 groups, using Tm as a categorical variable (low Tm <60 °C and not-low Tm $\geq$ 60 °C). A Fisher exact test gives a p-value of 0.01, suggesting there is a correlation between thermal stability and IGKV2 germline.

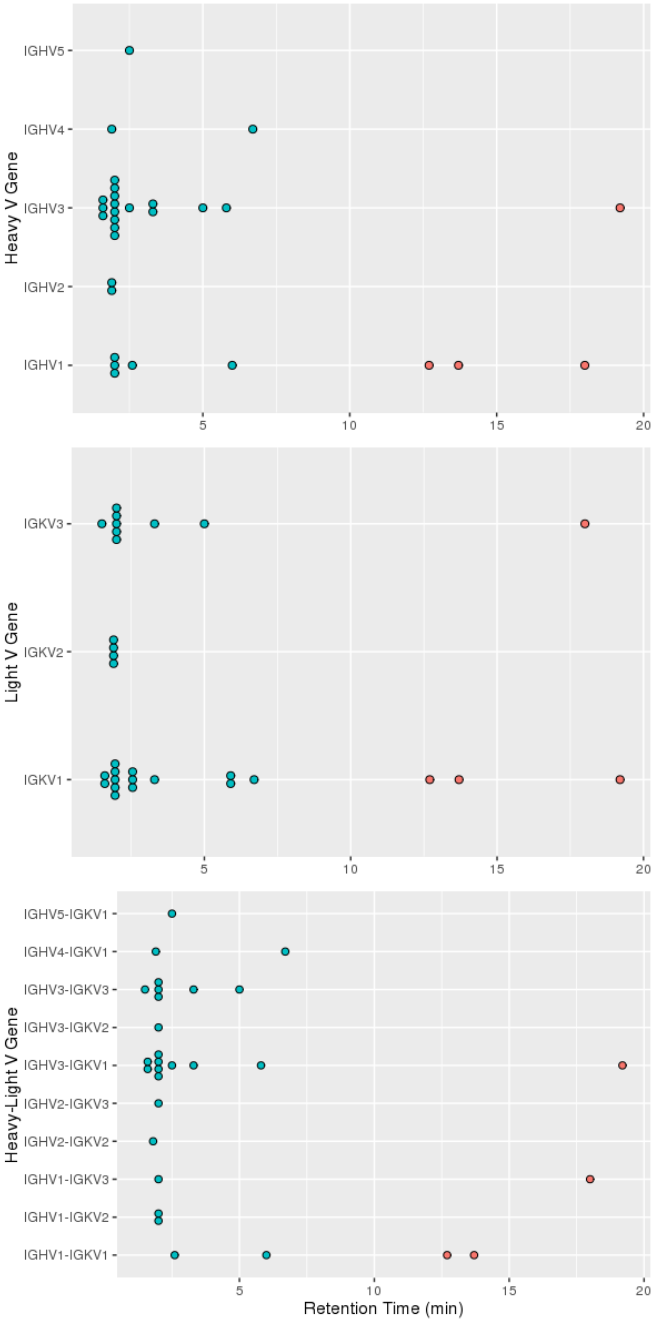

Figure 6.9: HIC retention times across V_H and V_L V-genes. Points are coloured blue and red for short and long-HIC retention time Fabs respectively. Bin width = 0.3.

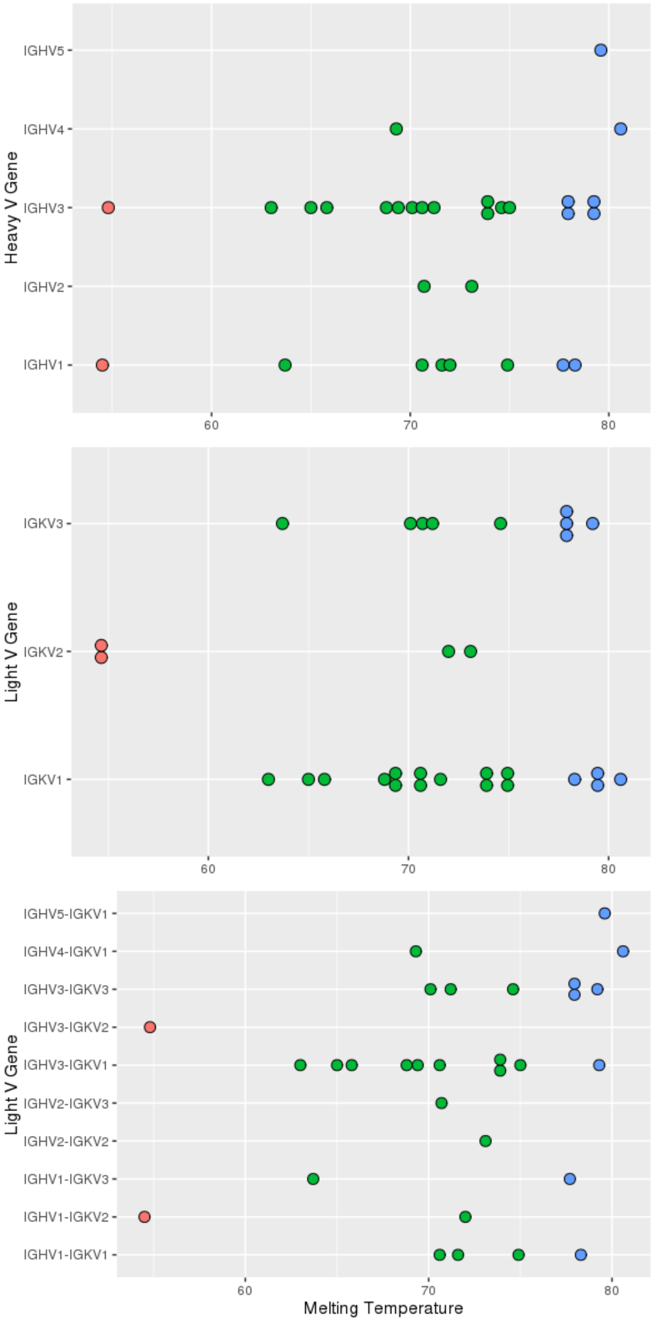

FIGURE 6.10: Tm across V_H and V_L V-Genes. Points are coloured red, green and blue for low, medium and high-Tm Fabs respectively. Bin width = 0.4.

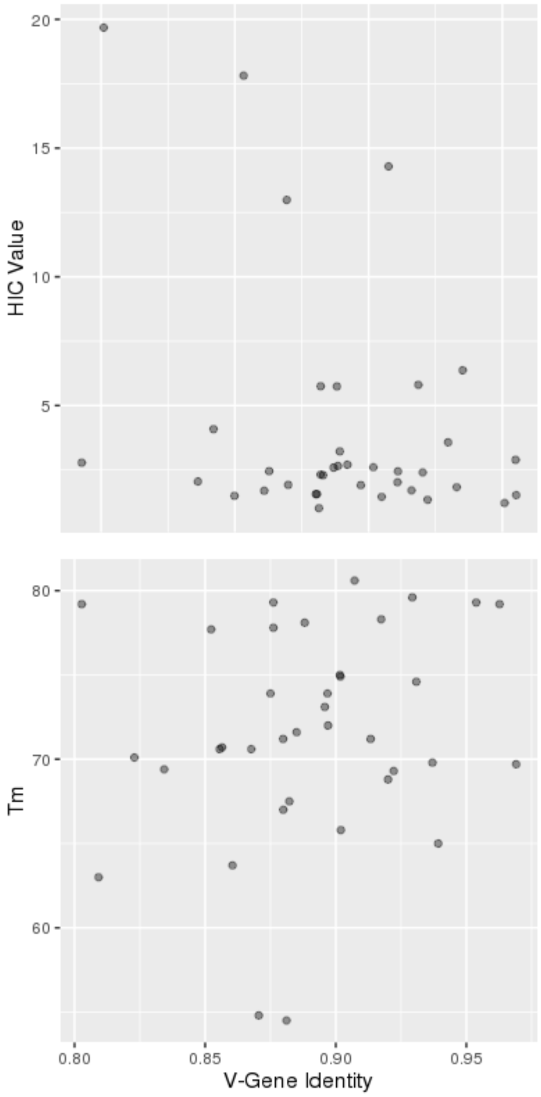

Figure 6.11: Combined V-Gene germline identity correlates with neither Tm nor HIC retention time.

Figure 6.10 also shows the distribution of Tm values across heavy-light V-gene pairs. However, considering the number of potential pairs, more data is required to undertake an effective analysis on the relationship between pairing and Tm.

The relationship between V-gene germline identity and hydrophobicity and was also investigated. Figure 6.11 shows V-gene germline identity plotted against Tm and HIC retention time. The graphs show that neither Tm nor retention time correlates with V-gene germline identity. If it is assumed that V-gene germline identity correlates with overall somatic hypermutation, this suggests that biophysical stability is not dependent on the extent of somatic hypermutation.

TABLE 6.1: Consensus hydrophobicity values, as calculated by Eisenberg et al. (1982).

Residue	Hydrophobicity Value
Ile	0.73
Phe	0.61
Val	0.54
Leu	0.53
Trp	0.37
Met	0.26
Ala	0.25
Gly	0.16
Cys	0.04
Tyr	0.02
Pro	-0.07
Thr	-0.18
Ser	-0.26
His	-0.40
Glu	-0.62
Asn	-0.64
Gln	-0.69
Asp	-0.72
Lys	-1.1
Arg	-1.8

6.4.3 Surface hydrophobicity

It was hypothesised that the HIC retention time of an antibody is related to the hydrophobicity of the residues on its surface. To test this hypothesis, the set of antibody surface residues first had to be defined. A set of 529 Chothia-numbered, non-redundant free antibody crystal structures was obtained from AbDb (Ferdous and Martin, in press). For each antibody structure, residue $rASA$ was calculated using the *pdbsolv* program from the bioptools package (Porter and Martin, 2015). For each Chothia position, the mean rASA was then calculated. If mean $rASA > 10$, then the Chothia position is defined as surface. From 278 Chothia positions, 191 positions were defined as surface. For a given antibody and Chothia position, the hydrophobicity of the residue at that position was then taken according to the hydrophobicity scale shown in table 6.1.

First the relationship between the total surface hydrophobicity of an antibody and its HIC retention time was investigated. Total surface hydrophobicity is simply the sum of the hydrophobicity values of the surface residues of the antibody. Figure 6.12 shows total surface hydrophobicity plotted against HIC retention time. Although the antibody with the longest retention time value also has the most hydrophobic surface, there does not appear to be an overall correlation across the sample.

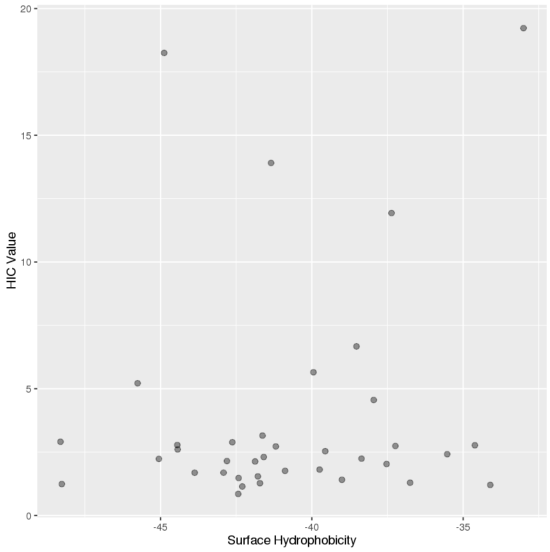

Figure 6.12: Total surface hydrophobicity does not correlate with HIC retention time. The total surface hydrophobicity is simply the sum of the hydrophobicity values of all residues with $rASA > 10\%$.

The relationship between counts of hydrophobic or hydrophilic surface residues and retention time was then examined. In the case of hydrophobic surface residues, a surface residue was only counted if it's hydrophobic value was greater than a threshold t, where $t = 0.0$, 0.1, 0.2, 0.3, 0.5, 0.6 and 0.7. The same was then repeated for hydrophilic residues, where a surface residue was counted if its hydrophobic value was less than $t = -1.7$, -1, -0.7, -0.6, -0.3, -0.2, -0.1 and 0.0. The results for each threshold are shown in figure 6.13. For each threshold, a test for correlation was undertaken. Because of non-normal distribution of retention times, a non-parametric measure must be used. Therefore, the Spearman's rank correlation coefficient and its p-value was calculated. Using a significance value of 0.05, no correlations were found to be significant.

Next, the relationship between hydrophobic/hydrophilic surface residue clusters and retention time was investigated. It was hypothesised that either the presence of hydrophobic clusters or the absence of hydrophilic clusters would lead to higher retention time.

The in-house program `clusterResidues` (Martin, unpublished) was used to cluster surface residues. `clusterResidues` uses a distance matrix and a distance cut-off to identify neighbouring residues. It then identifies residues that meet specified property criteria (e.g. $rASA > 10\%$ and hydrophobicity value > 0.2). Starting with a cluster containing an initial member that meets the criteria, it adds neighbouring residues to the cluster that also meet the criteria and then repeats the process recursively with the new members. Once no more residues can be added, the cluster is output if the number of member residues is equal to or larger than the specified cluster size threshold.

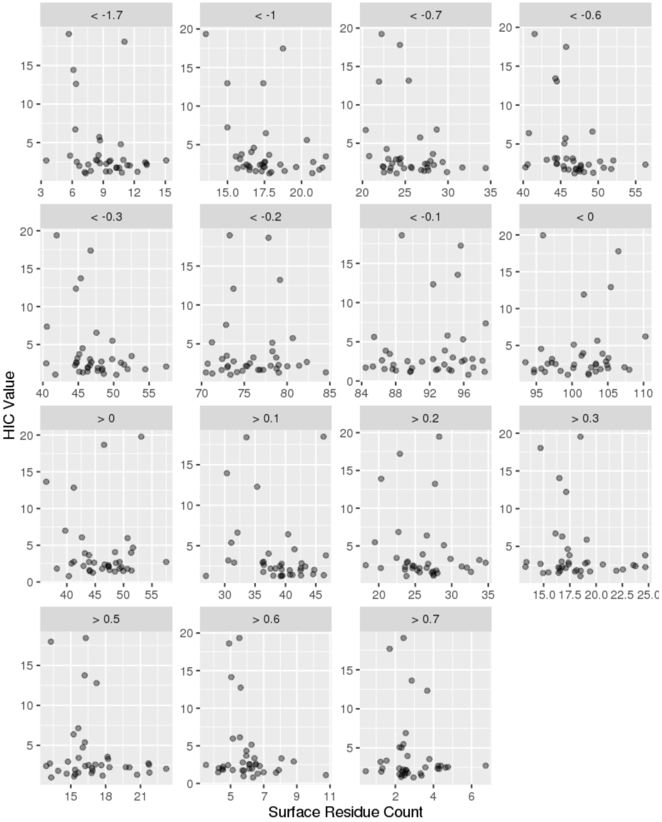

FIGURE 6.13: Surface residue counts against HIC retention times. Each graph is titled with the threshold used. No significant correlations were found using any of the thresholds.

To identity hydrophobic/hydrophilic clusters for a given antibody, the following input was supplied to `clusterResidues`:

- A C$_\alpha$-C$_{alpha}$ distance matrix for all Chothia positions, calculated by averaging the distances from all antibody crystal structures deposited in the PDB as of December 2011.
- A neighbour distance cut-off of 4 Å.
- Residue property criteria of $rASA > 20\%$ and hydrophobicity t.
- Minimum cluster size of 3.

where t is the same as defined above.

Several metrics can be calculated from the output of `clusterResidues`. The metrics investigated were: mean cluster size (number of residues), total number of clusters, total clustered hydrophobicity and total number of clustered residues. Figure 6.14 shows the mean cluster size across t. The testing for correlation as applied previously to the residue

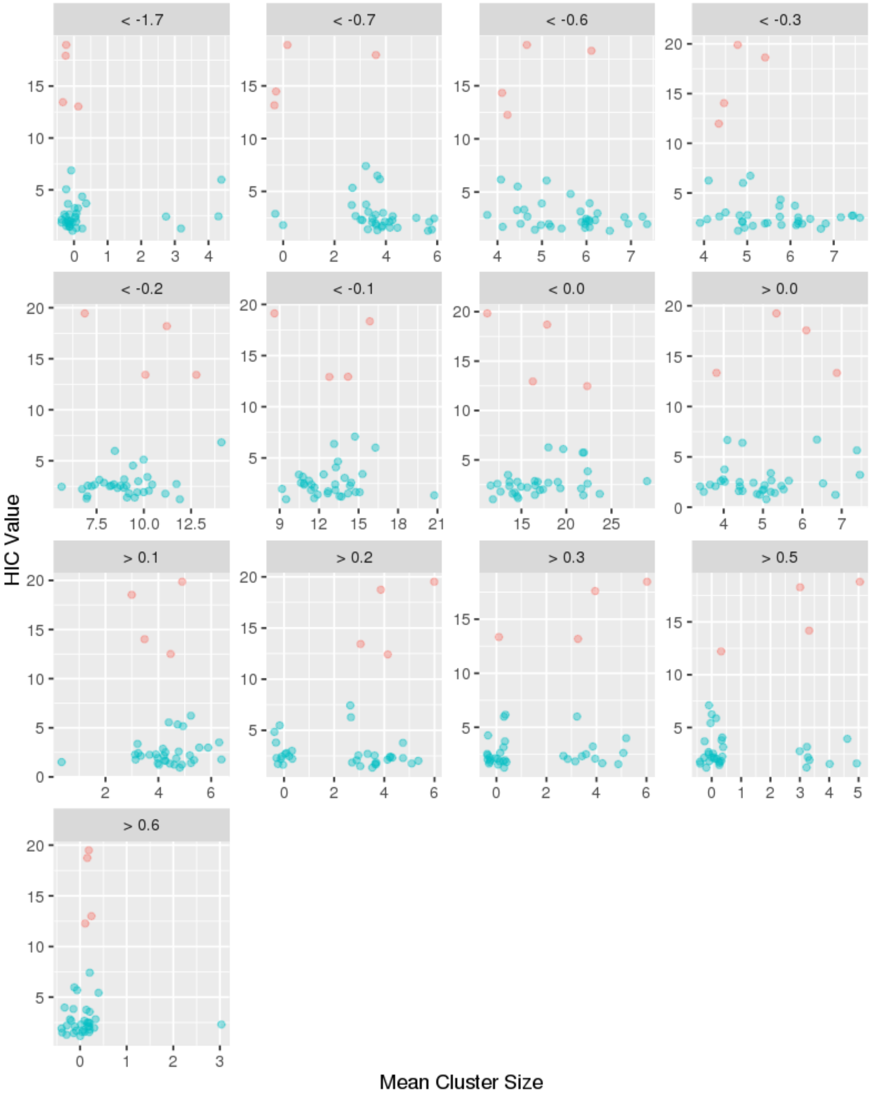

FIGURE 6.14: Surface residue clusters against HIC Retention Time. Each graph is titled with the threshold used to run *clusterResidues*. Short and long-HIC retention time points are coloured in blue and red respectively. See table 6.2 for correlation coefficients and p-values for each threshold.

count results was applied here on the four metrics. The results are shown in table 6.2. It was found that no correlations were significant for hydrophobic t. However, a subset of the hydrophilic t did produce significant correlations. For hydrophilic $t = -0.3$, -0.6 and -0.7, mean cluster size significantly negatively correlates with retention time, while total clustered hydrophobicity significantly positively correlates. Clustered residue count significantly correlates when $t = -0.7$, and total clustered hydrophobicity significantly correlates when $t = -0.0$. However, the number of clusters does not correlate significantly at any t.

TABLE 6.2: Spearman's rank correlation coefficients between cluster measurements r_s and HIC retention time. p-values for each coefficient are shown in brackets. Coefficients with $p < 0.05$ are highlighted. t refers to the hydrophobicity threshold used during clustering; a negative value is used as a hydrophilic threshold (e.g. < -1.7) and a positive value is used as a hydrophobic threshold (e.g. > 0.5). Thus -0.0 refers to threshold < 0.0.

	Number of Clusters	Clustered Residue Count	Mean Cluster Size	Total Clustered Hydrophobicity
t	r_s (p-value)			
0.6	-0.20 (0.25)	-0.20 (0.24)	-0.20 (0.24)	-0.20 (0.24)
0.5	0.04 (0.80)	0.04 (0.80)	0.09 (0.59)	0.028 (0.87)
0.3	-0.03 (0.86)	-0.02 (0.90)	0.04 (0.81)	-0.03 (0.88)
0.2	-0.02 (0.92)	-0.03 (0.87)	-0.03 (0.88)	-0.02 (0.89)
0.1	-0.02 (0.89)	-0.19 (0.27)	-0.02 (0.89)	-0.19 (0.27)
0.0	-0.25 (0.13)	-0.10 (0.57)	0.09 (0.60)	-0.15 (0.39)
-0.0	-0.18 (0.28)	0.16 (0.33)	0.22 (0.20)	**0.33 (0.048)**
-0.1	-0.19 (0.26)	0.17 (0.30)	0.26 (0.13)	0.32 (0.054)
-0.3	0.09 (0.59)	-0.26 (0.13)	**-0.38 (0.020)**	**0.35 (0.032)**
-0.6	0.09 (0.60)	-0.24 (0.15)	**-0.38 (0.020)**	**0.34 (0.040)**
-0.7	-0.24 (0.15)	**-0.38 (0.022)**	**-0.56 (0.00031)**	**0.36 (0.029)**
-1.7	-0.09 (0.59)	-0.08 (0.64)	-0.08 (0.64)	0.08 (0.64)

6.4.4 Tm and surface/core hydrophobicity

The relationship between Tm, surface hydrophobicity (Φ_s) and core hydrophobicity (Φ_c) was investigated. It is generally believed that Φ_s and Φ_c affect the thermal stability of an antibody: hydrophobic residues in the core interact favourably to stabilise the fold of a protein, while hydrophilic residues on the surface favourably interact with the solvent for further stability. As antibodies have a common framework, the set of antibodies are expected to have similar Φ_s and Φ_c. However, it was hypothesised that small changes in either value will lead to changes in Tm.

First, Φ_s and Φ_c were investigated to assess if these values alone showed any correlation with Tm. Figures 6.15 and 6.16 illustrate that neither Φ_s nor Φ_c alone correlate with Tm. Next, the relationship between a combination of these values and Tm was investigated. Figure 6.17 shows Φ_s plotted against Φ_c, with antibodies grouped by low, medium and high-Tm. An area of medium-Tm is found in the center of graph, centred at approximately -40 Φ_s, 17 Φ_c. Within ±5 Φ_s, ±2 Φ_c of this center, no low or high-Tm antibodies are found. High-Tm antibodies appear on either side of this center, although 5 of 10 appear directly to the left (i.e. similar Φ_c to the medium-Tm core, but lower Φ_s).

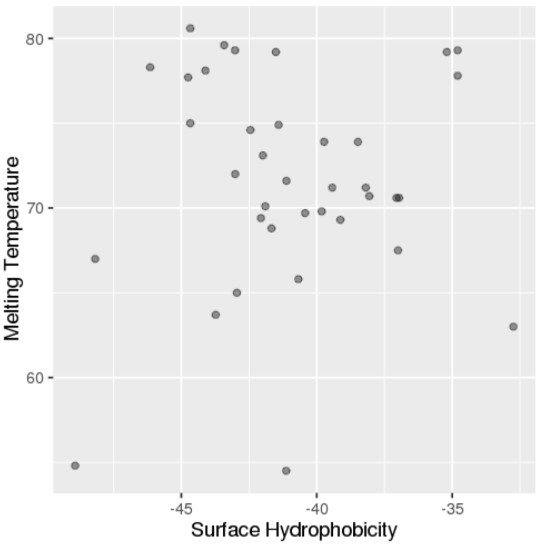

FIGURE 6.15: Surface hydrophobicity vs. Tm. A significant (p-value < 0.05) correlation between the two variables was not found.

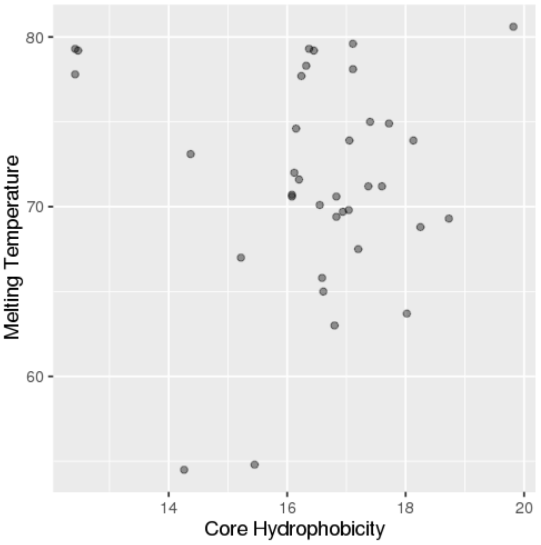

FIGURE 6.16: Core hydrophobicity vs. Tm. Similarly to surface hydrophobicity (see figure 6.15), no significant correlation was found.

This suggests that the balance of Φ_c and Φ_s may indeed influence Tm. To investigate this further, the normalised surface hydrophobicity $\tilde{s}$ of each antibody was calculated as

$$\tilde{s} = \frac{\Phi_s}{\Phi_c} \tag{6.1}$$

In figure 6.18, $\tilde{s}$ is plotted against low, medium and high-Tm antibody groups. There appears to be a difference in $\tilde{s}$ distributions between medium and Tm antibodies and the other two groups. High-Tm antibodies appear on average to have a lower $\tilde{s}$ than medium-Tm antibodies. A two-sample t-test between the two groups (medium-Tm $\mu =$

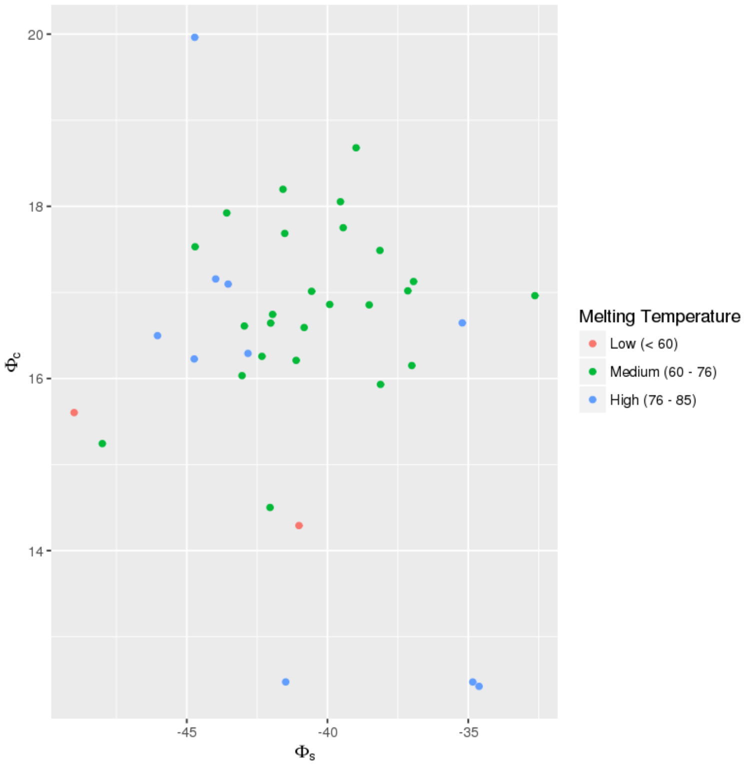

Figure 6.17: Surface (Φ_s) and core (Φ_c) hydrophobicity of Tm groups. An area of medium-Tm is found in the center of graph, centred at approximately -40 Φ_s, 17 Φ_c.

$-2.41, \sigma = 0.26$ and high-Tm $\mu = -2.67, \sigma = 0.33$ gives a p-value $==$ 0.02. Interestingly, the two low-Tm antibodies also have lower $\tilde{s}$ in comparison to the medium-Tm group. This suggests that a low $\tilde{s}$ might not necessarily lead to higher Tm but more generally a deviation from medium-Tm.

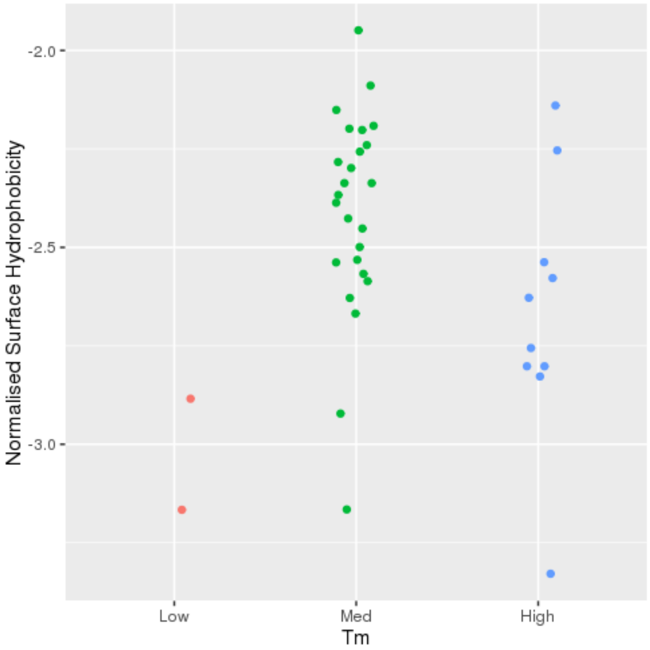

Figure 6.18: Normalised surface hydrophobicity ($\tilde{s}$) of low, medium and high-Tm antibodies (coloured red, green and blue respectively). A two-sample t-test between medium and high-Tm samples gives a p-value = 0.02, suggesting that high-Tm antibodies have a more negative (i.e. more hydrophilic) $\tilde{s}$. The two low-Tm samples also have more negative $\tilde{s}$ in comparison to the medium-Tm samples.

6.4.5 Investigating residue-level features

In the previous sections, whole-sequence features were investigated in order to identify correlates with thermal stability and surface hydrophobicity. In this section, residue-level features are investigated. There are 278 antibody variable region sequence positions and therefore 278 dimensions of interest. However, this far outnumbers our sample size. In order to overcome this, PCA can be used to reduce the number of dimensions.

In order to perform PCA, an $m \times n$ matrix H is formed, where $m =$ number of Chothia positions and $n =$ number of antibody sequences. Each cell of the matrix H_{ij} is filled with the hydrophobicity value of the residue at the j-th Chothia position of the i-th sequence. If a position is missing from a sequence (e.g. CDRH3 insert positions), then the hydrophobicity value for that position is set to 0. This matrix is then the input for PCA. Note that the 278 dimensions are not normalized before PCA to account for the fact that the units of each dimension are the same.

6.4.5.1 HIC Retention Time

For the analysis of HIC retention time, all 37 sequences with a retention time were input to the PCA. The first three components are plotted in figure 6.19. Though there seems

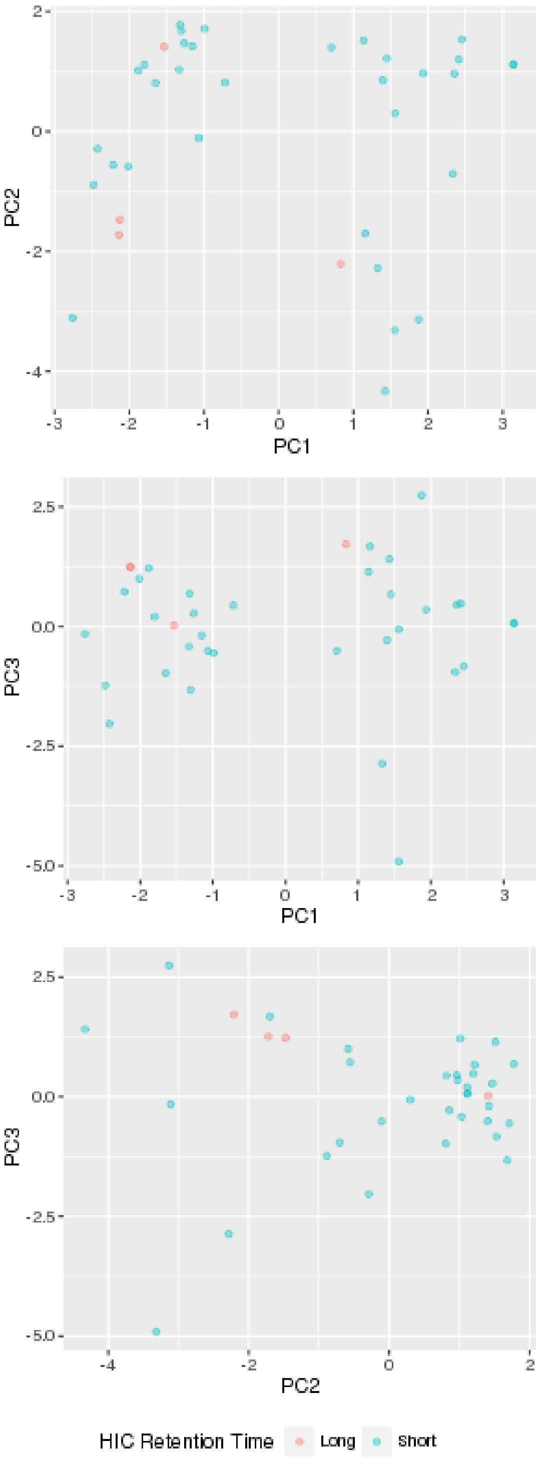

Figure 6.19: Principal Components (PC) 1–3 vs. HIC retention time. Points are coloured blue and red for short and long-HIC retention time groups respectively. The clustering of long-HIC antibodies on PC2 and 3 was tested and a p-value of 0.006 was obtained (see main text for more details).

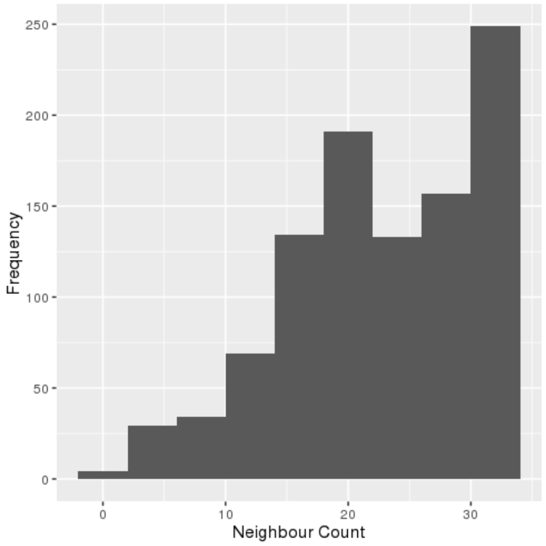

Figure 6.20: HIC PCA: PC 2 and 3 neighbour count distribution. Neighbour counts were sampled 1000 times as described in the main text to generate this distribution. The neighbour count for the

to be no relationship between HIC retention time and the first component, there does appear to be a clustering of 3 of 4 of the long retention time antibodies when plotted on the second and third component. To test the statistical significance of this, the probability of finding such a group of points was calculated. In order to do this, the following was simulated

1. Choose four random points. Select the closest three-point subset S (i.e. those points which have the minimum total distance between them.)
2. Find the maximum distance d between points in S. Then find n by counting the number of points that are within d from any point in S.

The above process was repeated 1000 times to create a distribution of n. To find a p-value, the proportion of the distribution where $n \leq h$ for $h = 4$ (the value of n for the long-HIC retention time group) is found. This gives us a p-value of 0.006, suggesting that the clustering of hydrophobic points on PC2 and PC3 is significant.

In order to ascertain the positions responsible for the distribution of points across PC2 and PC3, the loadings of each position onto the two components were identified. The positions with the highest loadings for PC2 and PC3 are H71 and H12 respectively. The hydrophobicity values of H71 and H12 are shown in figure 6.21. The plot shows that similarly to PC2 and PC3, the hydrophobicity values of H71 and H12 are able to separate three of the four long-retention antibodies from the majority of the remaining antibodies; these three Fabs have a relatively hydrophobic H71 and a hydrophilic H12.

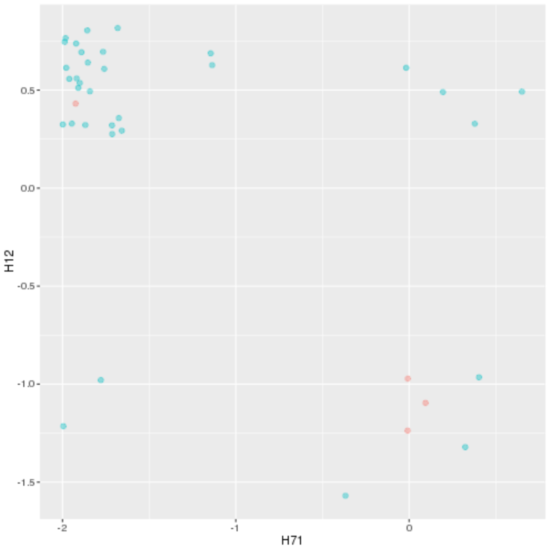

Figure 6.21: H71 and H12 hydrophobicity values vs. HIC-retention time. Short and long-HIC retention time antibodies are coloured blue and red respectively. H71 and H12 were identified the positions with the largest loading on PC2 and 3 respectively.

6.4.5.2 Tm

PCA was also used to investigate Tm. For this analysis, a subset of the Fabs with Tm values was used. It was hypothesised that above a certain core hydrophobicity value (Φ_c), an antibody can be stabilised by decreasing surface hydrophobicity (Φ_s). This is based on figure 6.17, where a subset of high-Tm antibodies are found that have similar Φ_c to the majority of medium-Tm antibodies, but lower Φ_s. With this in mind, only the 30 antibodies with $\Phi_c > 16$ were used to undertake PCA. Furthermore, because it seems that surface hydrophobicity is important, only surface positions ($rASA > 10\%$) were chosen. This results in 191 positions. Additionally, any position with 0 variance (i.e. the same residue in all sequences) was removed, leaving 157 variable surface positions.

The results of the PCA are shown in figure 6.22. Unlike the HIC retention time, no clustering seems to occur for high-Tm Fabs. However, for most components, a high-Tm Fab seems to occur at one or both ends of the distribution (e.g. the positive end of PC2 and the negative end of PC4). To see this more clearly, the median absolute deviation (MAD) of each Fab was plotted for the first twelve principal components. It can be seen that high-Tm antibodies tend to occur at one or both ends of PC1, 2, 4, 7, 9 and 11. To test this, MADs for each component were first normalised. Then the maximum normalised MAD was identified for each Fab from any of the first twelve components. Maxmimum MADs for medium and high-Tm are shown in figure 6.24. The maximum MADs for medium and high-Tm Fabs were then tested using the Mann-Whitney U test, which resulted in $W = 122.5, p = 0.02$. This supports the idea that high-Tm antibodies tend to be found at the ends of components. This suggests that

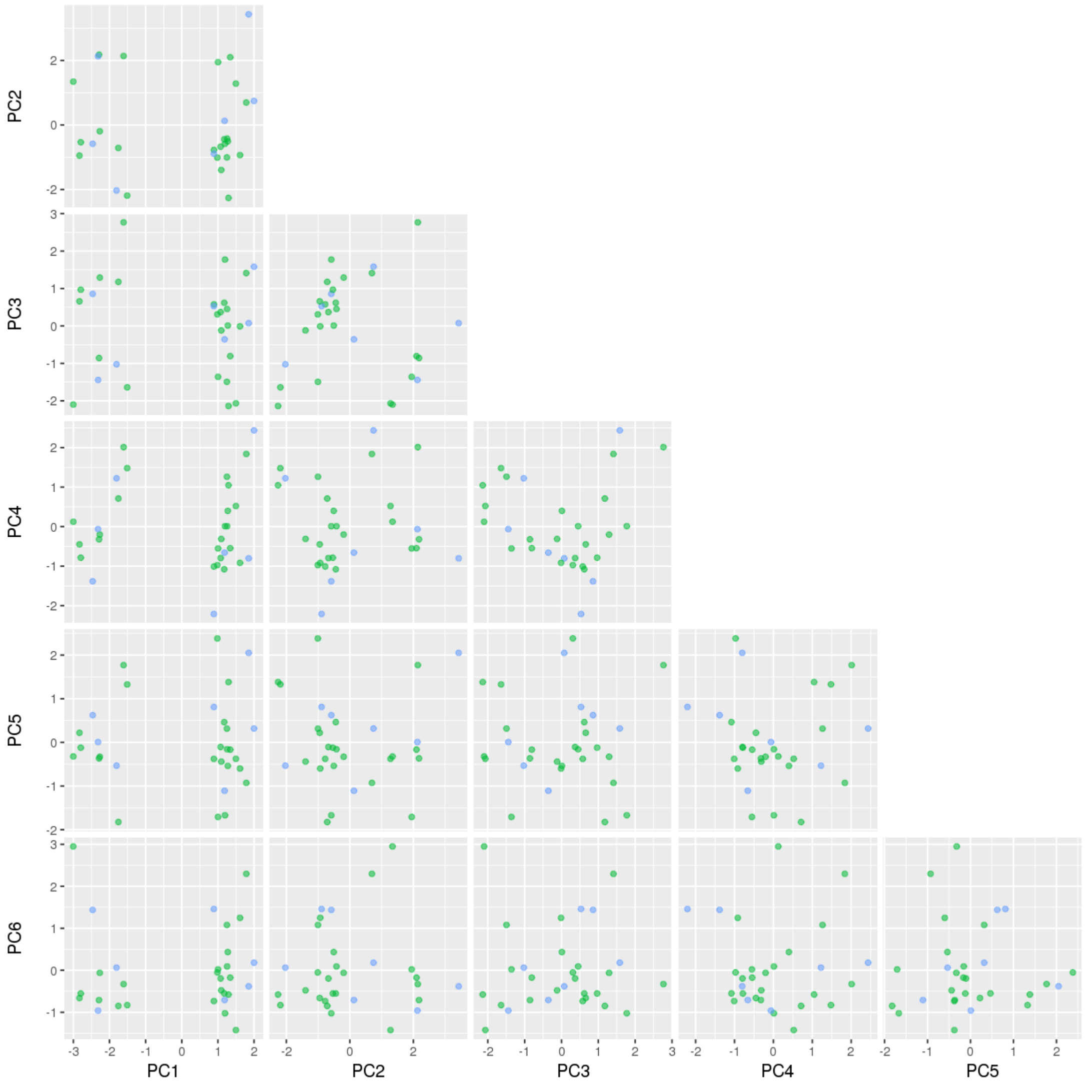

Figure 6.22: Principal Components (PC) 1–6 and Tm. PCA was performed on a subset of 30 out of 37 Fabs, using only surface residue positions (see main text for more detail). Medium and high-Tm antibodies are coloured green and blue respectively. On PC1, 2 and 4 a high-Tm instance appears at one or both ends of the component.

some aspect of outlyingness is correlated with higher Tm. This was investigated further by ascertaining which positions lead to this outlyingness.

The outlyingness observed for the high-Tm Fabs occurs across many components, rather than a single component. Each component is influenced by a different set of positions. This suggests that rather than a single set of positions, thermal stability may be correlated with many sets of positions, with the common feature being unusual (i.e. outlying) residues at these positions. For each outlier, the positions that are the main contributors to its outlyingess can be identifed. To illustrate this, figure 6.25 shows five components with a high-Tm antibody at one end. The components are reconstructed by iteratively adding positions and their loadings. The loaded positions are ordered by decreasing

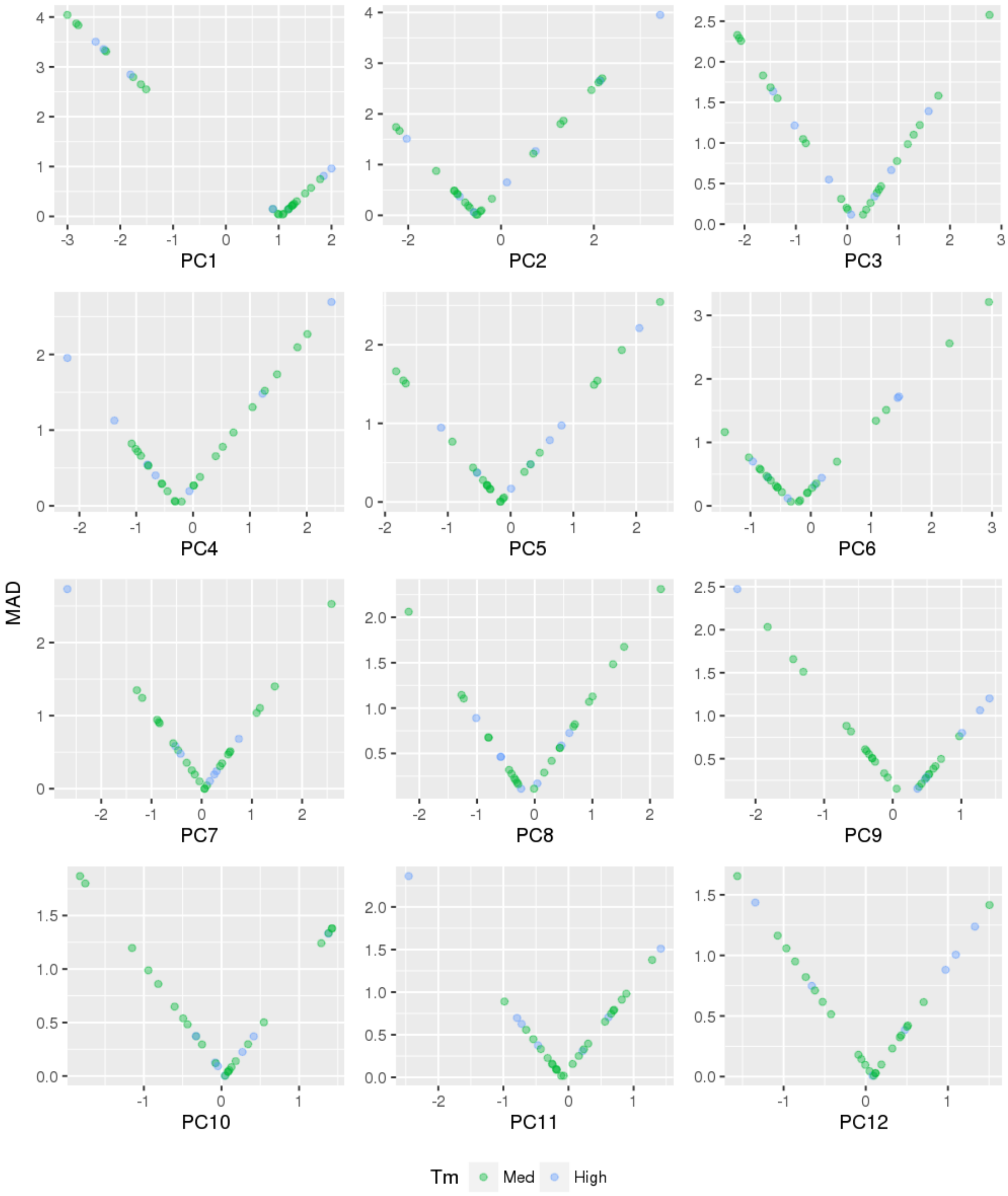

Figure 6.23: Median Absolute Deviations (MAD) for principal components 1–12. High-Tm Fabs (blue) are found at one or both ends of PC1, 2, 4, 7, 9, and 11.

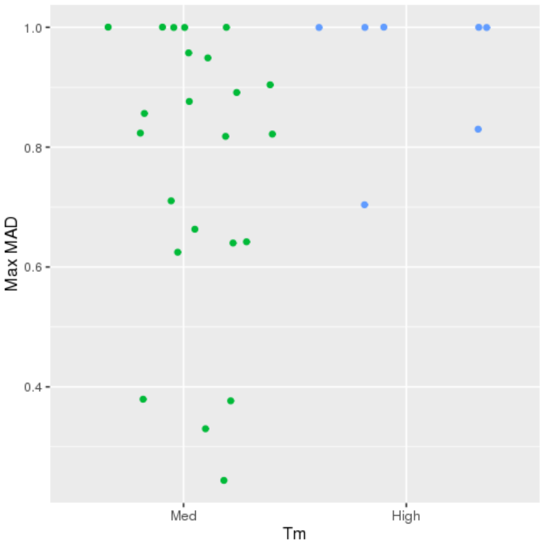

FIGURE 6.24: Maximum MADs. For each antibody, the maximum MAD after normalisation was found. Max MADs were compared between medium and high-Tm antibody groups using a Mann-Whitney U test, which resulted in a p-value $= 0.02$.

difference between the median of the loaded position and the value of the loaded position for the high-Tm outlier. These graphs shows that only a small number of positions are required before each high-Tm outlier approximately reaches its position on the final component. Furthermore, as the loaded positions are ordered by the magnitude of the difference between the median of loaded position and the value of loaded position for the high-Tm outlier, the positions can be considered to be ordered by importance for making it such an outlier. With this in mind, the ten most outlying positions were identified for those high-Tm outliers with a maximum MAD of 1 (i.e. those outliers at the very end of a component). In the cases where an outlier had a maximum MAD of 1 on more than one component, the component on which the outlier had the highest non-normalised MAD was used. The same was done for medium-Tm outliers in order to compare any differences between the two groups. The results are shown in table 6.3. The first observation is that in three of the five high-Tm outliers, the primary outlier position is an H100 insert, in comparison to only two of the nine medium-Tm outliers. Furthermore, it was observed that H100 positions seemed to be slightly over-represented in the first 10 outlying positions of the high-Tm group ($11/50 = 0.22$), in comparison to the medium-Tm group ($9/90 = 0.1$).

To test this further, the distribution of H100 positions across all 157 positions ordered by outlyingness was investigated. If H100 positions have more of an influence over high-Tm outlyingness, more of them should be observed in the most primary positions of the position order, in comparison to the medium-Tm group. To test this, the rank of each H100 position in the position order for each outlier was determined. For high and medium-Tm groups, all H100 position ranks were then ordered and plotted in sequence,

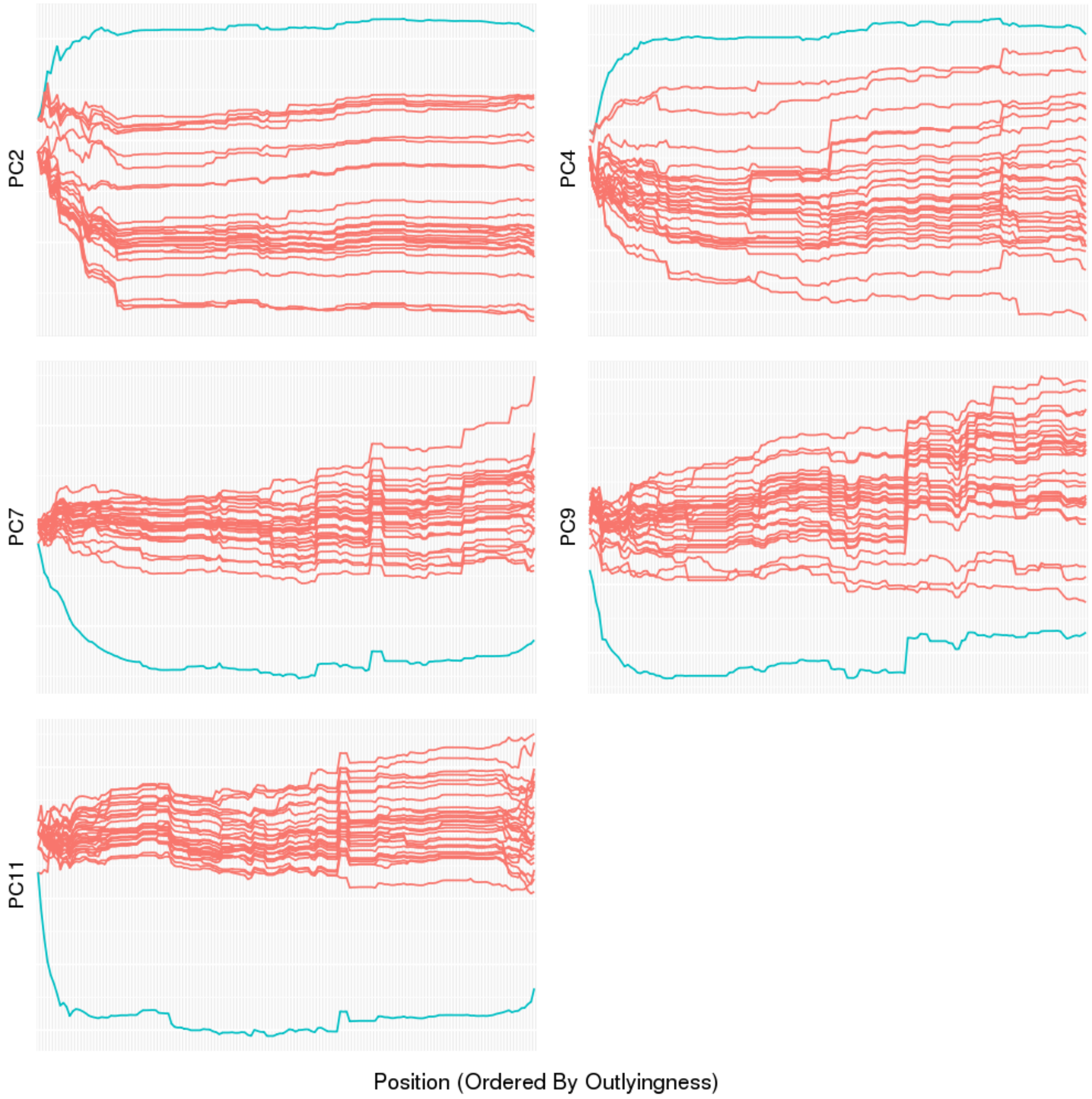

Figure 6.25: Reconstructed principal components (PC) for five high-Tm outliers. The five high-Tm outliers (blue) are those found at one end of a PC; if a high-Tm antibody is found at the end of more than PC, then the PC on which it has the largest MAD is shown. The final position of each antibody sequence on the principal component is reconstructed across the x axis by cumulatively adding each sequence position hydrophobicity weighted by the loading of that position on the component. Positions are ordered by the largest absolute difference between the position median hydrophobicity and the position hydrophobicity for the high-Tm outlier in question; this gives us the most outlying positions in order. This figure illustrates that only a small subset of positions is needed to approximate the final outlyingess of the high-Tm antibody in question.

Table 6.3: Top 10 most outlying sequence positions. For each component, the most outlying antibody was taken at each end. If an antibody was found more than once, the component on which the antibody had the highest MAD was kept. For each antibody and component, the top ten most outlying positions on the component are shown. H100 and insert positions are highlighted.

1	2	3	4	5	6	7	8	9	10
					high-Tm				
H12	H75	H23	H98	L18	L60	H58	H19	H10	H54
H83	H19	H5	H96	H77	**H100C**	H98	H85	H16	H60
H100C	**H100F**	L65	L3	H99	H61	H76	H84	L95	L93
H100A	L30	H82B	**H100**	H64	H56	H62	**H100D**	H33	H60
H100E	L95	**H100C**	H31	**H100F**	L54	**H100**	H56	H53	**H100H**
					medium-Tm				
H5	H16	**H100B**	H85	L77	H83	**H100E**	H19	H101	H58
H58	H97	L77	H57	L56	L15	H76	L50	L24	H52
H108	H58	H52	H101	L18	**H100C**	H31	H97	H98	L30
L65	L3	L42	H30	H98	H46	L93	L32	H10	L31
H30	H75	H3	H97	**H100C**	H58	L10	**H100B**	L92	L31
H100B	H96	H30	L69	H52	L92	L100	L94	H77	H84
H16	L91	L54	H57	L94	H52	H98	L55	L15	L85
H58	H31	H44	**H100E**	L65	H23	H52	H54	H53	H98
H100F	**H100E**	H3	H83	H19	L50	H53	H98	L93	H5

as shown in figure 6.26. The figure shows that when the most primary H100 ranks in high-Tm to medium-Tm are compared, those ranks are numerically lower in high-Tm for approximately the first third of ranks. To test if this difference is significant, the first third of high-Tm and medium-Tm ranks (sample sizes 31 and 57 respectively) were tested using a Mann Whitney U test. This resulted in a p-value = 0.001.

It is clear then that H100 and its insert positions play a role in defining the outlyingness of our high-Tm outliers. Specifically, outlying hydrophobicity values at these positions seem to be associated with high-Tm outliers. Consider that for a position in a sequence where that position is absent, hydrophobicity = 0 (see section 6.4.5). Thus, it is possible that a sequence can have an outlying value at an insertion position that is commonly absent from sequences, simply by the presence of that insertion in the sequence. With this in mind, it was hypothesised that the length of the CDRH3 region may correlate with Tm. Figure 6.27 shows the distribution of CDRH3 lengths for high and medium-Tm groups (here the 30 antibodies used to carry out the PCA are included). As expected, CDRH3 length appears to be larger in high-Tm antibodies. A Mann Whitney U on the two groups gives a p-value of 0.04.

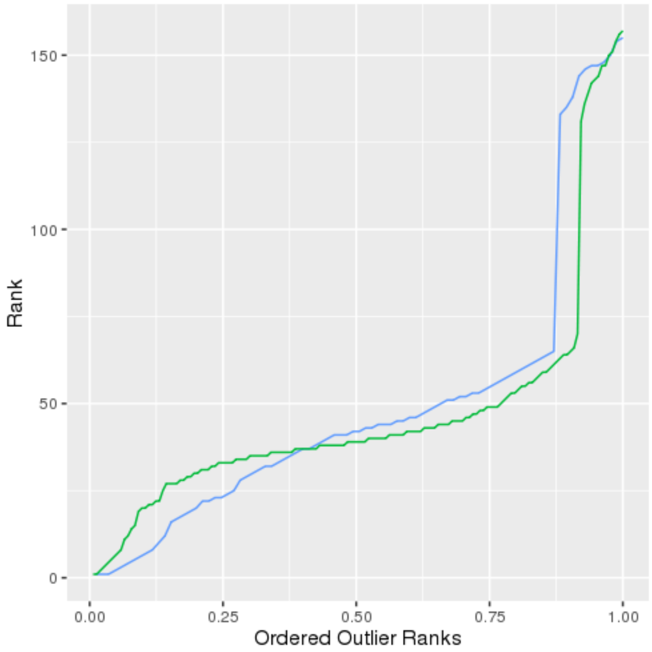

FIGURE 6.26: H100 and H100 insertion outlier ranks for medium (green) and high (blue)-Tm outliers. For each medium and high-Tm outlier, the outlier rank for H100 and each H100 insertion position was found. For each Tm-group, these ranks were grouped and sorted *numerically* low to high. In order to control for the difference in the sizes of the medium and high-Tm groups, the ranks are then plotted with the distances between ranks along the x-axis such that both sets of ranks cover the length along the axis. If H100 and insertion positions were ranked the same in medium and high-Tm groups, these lines would be expected to be the same; instead, medium-Tm ranks reach numerically higher values more quickly the high-Tm. This illustrates that H100 and H100 insertion positions are more outlying in high-Tm than medium-Tm outliers.

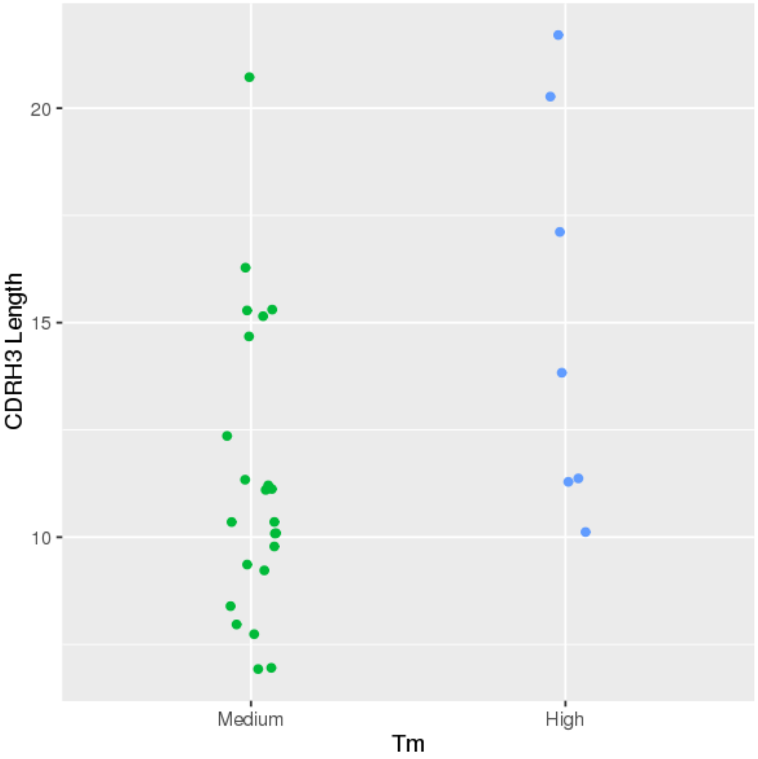

Figure 6.27: CDRH3 lengths of medium and high-Tm antibodies. Note that this set of Fabs corresponds to the subset defined at the start of section 6.4.5.2. A Mann Whitney U test on the two groups results in $W = 121.5, p = 0.04$

6.5 Discussion

In this chapter, sequence and biophysical assay data from 37 human Fabs were analysed in order to identify correlates between sequence features and biophysical stability.

The distribution of Tms obtained corresponds well to previous studies (Garber and Demarest, 2007, Tiller et al., 2013). It was observed that the light chains of both low-Tm Fabs have V-gene IGKV2. This corresponds with the observation that IGKV2 is not only the least stable isolated V_κ, domain but also that, in contrast to other low-stability domains, is not rescued by partnering to a high-stability V_H within an scFv (Ewert et al., 2003).

Because of the limited sample size and the fact that any primer bias due to use of different primers for each V-gene is unknown, a complete analysis of V-gene usage cannot be done. However, the observation that the majority of V_H are assigned to IGV_H3 and V_κ to IGKV1 or IGKV3 corresponds to the prevalence of those V-genes in existing data (Zhao and Lu, 2011). No V-gene pairings were found to be preferred or disfavoured in our set of 61 variable-region sequence pairs, in contrast to previous studies (Jayaram et al., 2012). However, the limited sample size as well as small effect size expected (see Jayaram et al. (2012)) means that the performed test had low statistical power. More data in the future will allow pair preferences to be tested more comprehensively.

The problem of multiple testing can arise when a large number of statistical tests are undertaken simultaneously (McDonald, 2009). For example, if 100 statistical tests are undertaken using a significance threshold of $p < 0.05$, then five tests would be expected to be significant due to chance alone and thus five false positives would be obtained. Multiple testing was undertaken occasionally in this chapter (e.g. the testing for correlation between HIC retention time and the number of hydrophobic surface clusters, repeated at different residue hydrophobicity thresholds). However, no method to account for multiple testing (e.g. the Bonferroni correction (Dunn, 1961) or the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995)) was applied. Correction for multiple testing was not considered appropriate given the exploratory nature of the work and the possibly over-conservative nature of multiple testing correction (Perneger, 1998). Rather than being used to make any firm conclusions, the p-values stated throughout serve to highlight interesting hypotheses that can be followed up in further experiments.

Owing to the limited size of the dataset, the training of a machine learning algorithm was not performed. Thus, the features identified are to be treated as preliminary until the availability of more data allows us to evaluate their utility as machine learning features through cross-validation and independent testing.

Previous studies have shown that the removal of hydrophobic surfaces from antibodies and antibody fragments can lead to reduced aggregation propensity (Chennamsetty et al., 2009, Dudgeon et al., 2009, Wu et al., 2010). Though Wu et al. (2010) were able to improve solubility by decreasing overall surface hydrophobicity, a relationship between HIC retention time and overall surface hydrophobicity was not found in this study. However, it was found that the absence of hydrophilic surface residue clusters correlated significantly with long HIC retention time. Interestingly, Chennamsetty et al. (2009) were able to use molecular dynamic simulations to identify sites on the surface of an antibody that tended to be exposed and hydrophobic over the course of a simulation. The engineering of these sites on two therapeutic antibodies to reduce hydrophobicity lead to decreased aggregation propensity. Currently, the program used to identify surface residue clusters (`clusterResidues`) defines structural neighbours by using average residue-residue distances from a large set of antibody PDB structures; this obviously limits the accuracy of the identified clusters. In the future, either structural modelling, with our without molecular dynamics simulations, could increase structural accuracy.

PCA on a matrix of residue hydrophobicity values revealed it possible to separate three of four long-HIC retention time Fabs from all but three short-time Fabs by observing the hydrophobicity at positions H71 and H12. Three of the four long-HIC retention time antibodies have a hydrophobic H71. H71 has been identified as an important position in influencing the conformation of CDRH1 and CDRH2: it sits in a position where it can influence the packing of CDRH1 and CDRH2 by interacting with CDRH2 residues H51 and H52A and FR1 residue H29 (Xiang et al., 1995). So as well as presenting a more hydrophobic side chain itself, H71 may be influencing how CDRH1 and CDRH2 are structured, thus influencing their surfaces. H12 was found to be more *hydrophilic* in three of the four high-HIC retention time Fabs. Thus the possible influence of H12 is more difficult to guess. H12 is close to, but not part of, the V_H/C_H1 interface (Wang et al., 2009). It is also close to H6, H7 and H10 which together define the structural subtype of the heavy-chain framework and play an important part in stabilising the domain (Jung et al., 2001). Finally, it is also close to H13, which can influence the packing of the lower core of the domain (Ewert et al., 2003). If further data confirm the correlation between H71 and H12 and HIC retention time, then structural modelling should give more of an insight into a mechanism.

Wu et al. (2010) found that decreasing overall surface hydrophobicity lead to a modest improvement in thermal stability. In the current study, no correlation between thermal stability and overall surface hydrophobicity, nor core hydrophobicity was found. However, it was observed that medium-Tm Fabs tend to have a combination of average core and surface hydrophobicity, whilst deviation away from an average combination leads correlates with both high and low-Tm. In particular, it seems as if retaining an average core hydrophobicity while decreasing surface hydrophobicity may lead to high-Tm. Furthermore, studies on the engineering of antibody and non-antibody proteins suggest a link between increasing negative surface charge and resistance to heat-induced aggregation and inactivation (Lawrence et al., 2007, Shaw et al., 2008, Dudgeon et al., 2009). As well as this, the comparative study of hyperthermophilic and mesophilic proteins suggests networks of surface ion-pairs are important for increasing thermal stability (Goldman, 1995). If surface charge and/or networks of surface ions are important to antibody thermal stability, it would be expected that the more general property of surface hydrophobicity might loosely correlate with Tm in the manner observed in this study. This suggests that future work should focus on overall, as well as networks of, surface charge.

CDRH3 residues are commonly found at the V_H/V_L interface, which is known to be important for fold stability (Ewert et al., 2003, Abhinandan and Martin, 2010) . Additionally, Ewert et al. (2003) found that H100 formed the center of a network of salt-bridge interactions that tended to be disrupted in low-stability germline subtypes. Thus relationships between features of CDRH3 and thermal stability are expected. The observation that (within a subset of Fabs with average core hydrophobicity) average CDRH3 length differs between medium and high-Tm Fabs shows that a very general description of CDRH3 is sufficient to capture some of its influence on stability. More data in the future will allow more nuanced descriptions that help distinguish high-Tm Fabs more accurately, particularly by focusing on those positions known to be important in the V_H/V_L interface.

6.5.1 Conclusion

This chapter presents the results of an analysis of a set of natural V_H-V_L pair Fab sequences and biophysical properties. Though the sample size was limited, a number of sequence features were identified that can be validated in the future when more data are available. In particular, more data points will allow a machine learning method to be built using proper cross-validation and independent testing. Owing to simplicity of the features described, their application as features for a machine learning method will be straightforward.

Future work should also include the investigation of more sophisticated features. An obvious avenue of research is the building and analysis of structural models, as well molecular dynamics. Although the modelling of CDRH3 is still proving to be challenging, the remainder of the Fab structure can be modelled with high accuracy. Void analysis should be applied to structural models as it is believed the packing within variable domains effects stability (Ewert et al., 2003). Beyond this, the literature contains a wealth of observations that can be used as starting points for the identification of structural features (Ewert et al., 2003, Monsellier and Bedouelle, 2006, Honegger et al., 2009). The V_H-V_L interface is another area of interest that should certainly be included in a structural investigation.

Chapter 7

Conclusions and Future Directions

The aim of this thesis was to develop *in silico* methods to aid the design of therapeutic biologic drugs and, in particular, antibodies. Two important and related properties were chosen for investigation: immunogenicity and biophysical stability. This chapter concludes the investigation of these properties that was carried out in this thesis.

7.1 B cell epitope Prediction

A biological therapeutic, like any foreign antigen, is recognised via B and T cell epitopes. Because of their linear nature, T cell epitopes are simpler to predict Yu et al. (2002), Wang et al. (2008). In contrast, B cell epitopes are structural in nature — being formed from distant sequence elements brought together by the protein fold — and are therefore harder to predict. This thesis focused on the prediction of B cell epitopes.

A review by El-Manzalawy and Honavar (2010) suggested that B cell epitope prediction may be improved by the utilization of advances in protein-protein interface prediction. Our first aim was to address this suggestion directly by applying the in-house IntPred method previously developed by (Baresic, 2011) for PPI prediction.

7.1.1 Development of IntPred:Epi

In chapter 3, IntPred was introduced. IntPred provides predictions on surface patches, so in order to compare it fairly to existing PPI predictors that predict on residues, a selection of methods for mapping from patch- to residue-level predictions were tested. Even the best mapping method lead to a drop in performance in comparison to patch-level predictions, but it is important to remember that the manner in which patches are labelled means that a set of patches are excluded from training and testing and thus any related performance measures are not representative of a real-case scenario. Nevertheless, IntPred performs well at residue-level — the only method that showed better performance was SSPIDER (Porollo and Meller, 2007) (MCC 0.37 and 0.41 respectively). Furthermore, IntPred performance could be improved in the future by applying the changes made to IntPred:Epi (presented in chapter 4), such as dataset balancing, inclusion of ASA-based features and clustering of residue predictions.

Despite the fact that IntPred was outperformed by SSPIDER, IntPred was chosen to carry forward for testing as a BCE predictor, because its previous development in the Martin group meant that source code and documentation was readily available, making method development feasible. Preliminary testing of IntPred on a set of antigen structures revealed that, without any amendment, IntPred showed no predictive performance. This supports the hypothesis that B cell epitopes are significantly different from other types of protein-protein interface. Thus the next step, presented in chapter 4, was to create IntPred:Epi, by amending the IntPred method to improve BCE prediction performance. The principal amendment was to retrain IntPred on a dataset of antigen structures. It was observed that performance measures from 10-fold cross-validation of the learner were very over-optimistic in comparison to performance on an independent test set. Rather than cross-validating on random subsets of patches, by-chain CV was developed that simply ensures that all of the patches of one chain are found in the test partition for each fold of validation. By-chain cross-validated performance is much closer to test performance and thus shows that the presence of overlapping patches in training and test partitions during cross-validation can lead to overstated performance measures. This illustrates that independent testing is crucial for a true measure of predictor performance and that any method which only reports cross-validated performance measures should be treated with caution.

An initial round of testing to compare IntPred:Epi to seven existing methods indicated the IntPred:Epi was only outperformed by SEPPA 2.0 (Qi et al., 2014) (MCC 0.11 and 0.19 respectively). However, a second round of testing on a human and mouse-host test sets — performed for the purposes of providing a baseline for the host-tailed methods presented in chapter 5 — gave contrary results. IntPred:Epi was able to maintain its performance on the mouse-host set (MCC 0.11), whilst performance dropped for the human-host set (MCC 0.06). In contrast, SEPPA 2.0 performance dropped markedly on mouse and human-host sets (MCC 0.03 and 0.06 respectively). The difference in SEPPA 2.0 performance on the different training sets highlights the difficulty of BCE predictor evaluation. It is not certain that the first test set — created specifically by Hu et al. (2014) to consist only of those antigen structures not found in any of the training sets of the seven predictors tested — was not included in the final SEPPA 2.0 model available to run via a web-server. Ideally, all methods available to run either locally or via web-server would also be available to retrain. Retraining all methods on the same data set would allow performance comparisons to be completely fair. The source code required to train and test IntPred and IntPred:Epi is available via GitHub[1].

An exploratory analysis of random forest predictions gave us insight into the difficulty of B cell epitope prediction. The random forest is unable to separate B cell epitopes from a large proportion of the remaining non-epitope surface. The reasons for this may be two-fold: firstly, the small differences observed between epitope and non-epitope surfaces (Rubinstein et al., 2008) indicate that epitopes are only subtly different from the rest of the surface and therefore hard to distinguish; secondly, the negative set definition problem: the nature of structural epitope data — that is, x-ray crystal structures of single monoclonal antibody-antigen complexes — does not represent the polyclonal response well, which leads to epitope surface being mislabelled as non-epitope. Recent approaches have sought to address these problems. Ren et al. (2015) treated the BCE prediction as a positive-unlabelled learning problem, leading to an improvement in performance when applied to positive-negative labelled data. As well as this, a method has been developed that uses a combination of bacterial surface display and structural modelling to map the epitopes of a polyclonal antibody response (Rockberg et al., 2008), thus significantly reducing the extent of the negative set problem. Because these methods are fairly new, data are scarce in comparison to the antibody-antigen complex data available in the PDB. However, combining both types of data could lead to improved BCE prediction in the future.

Patches are labelled according the fraction of their surface that is contributed by epitope residues. If a patch has an epitope surface fraction greater than the chosen class label threshold, then it is labelled as epitope. The exploratory analysis of the random forest also helped us identify a relationship between performance and the class label threshold used to label patches as epitope. Using a low threshold appears to produce ‘noisy’ epitope patches that, in terms of their features, are dissimilar to other epitope patches, owing to the presence of non-epitope patch residues. It was then shown that retraining IntPred:Epi on patches labelled using a high class label threshold was able to improve performance, with the caveat that it remains to be seen if this improvement is carried on to residue-level predictions.

As well as retraining, a handful of other amendments were made to IntPred:Epi. ASA-based features helped improve performance, as well as finding the optimal balance of class labels in the training set. Clustering residue-level predictions was found not to

[1] `https://github.com/ACRMGroup`

improve performance, though the results indicated that, if other performance gains can be made, clustering might start to become effective.

7.1.2 Development of TSLs and application to BCE prediction

In chapter 5, the development of libraries of human- and mouse-tolerated surface libraries (TSLs) was presented and their application to B cell epitope prediction analysed. By using a novel method of surface description and comparison, antigen surface patches can be searched for within a library, in order to give them tolerance labels. The application of these labels to predict BCEs on human antibody-bound antigen proved successful, showing performance on their own that was comparable with most general BCE predictors (MCC 0.08). Furthermore, the combination of IntPred:Epi with these tolerance labels by using them as a simple filter managed to improve its performance, increasing MCC from 0.0567 to 0.0845 on the human-host test set. However, the same was not seen for the mouse-host set.

The TSL methods are novel and relatively underdeveloped. The method used to describe and compare surface patches is likely to be critical for prediction performance. An analysis of the patches defined as similar by the method should help to guide its development. Additionally, a more sophisticated combination of tolerance labels within a machine learning method should be investigated, as well the inclusion of additional structural models by lowering the required template sequence identity threshold.

7.1.3 Future directions

In the near future, work should focus on finding an optimal combination of class label threshold, patch radius, class balance and residue-mapping parameters in order to improve BCE prediction performance. This improved predictor should then be combined with a more developed TSL method in a way that takes advantage of machine learning, in order to improve performance of human-host epitopes.

Further ahead, the positive-negative set problem should be considered. In particular, the utilisation of alternative data sources that represent the polyclonal response more completely should be considered.

7.2 Biophysical stability prediction

Biophysical stability is important for a biological therapeutic because it influences its shelf life, efficacy and immogenicity. In chapter 6, the sequence and biophysical properties of 37 natural $\mathrm{V_H}$-$\mathrm{V_L}$ pair Fabs were analysed in an attempt to find sequence determinants of Fab stability. The size of the data set meant that a conclusive analysis was not possible, but an exploratory analysis of the data revealed some interesting sequence features that correlated with either melting temperature or HIC retention time, such as the absence of hydrophilic surface patches, the ratio of core to surface hydrophobicity, CDRH3 length and V-gene germline. In the future, these features can be utilised with a machine learning method, in order to make biophysical property predictions from sequence. In this thesis, only two biophysical properties were analysed, but it is hoped that future data will also include other properties, such as aggregation propensity. Future work should also investigate the wealth of features observed to have an influence on biophysical stability (Ewert et al., 2003, Monsellier and Bedouelle, 2006, Honegger et al., 2009).

7.3 Application to therapeutic antibody design

From the testing performed in this thesis, it is clear that B cell epitope prediction has a long way to go until it can be utilised for therapeutic selection or design: currently, it is likely the performance is too low to be useful. Nevertheless, if potential BCEs are identified on the surface of therapeutic antibody, these could potentially be targeted for mutagenesis, particularly if they represent clusters of unusual residues, not regularly seen in human antibody.

The analysis of biophysical data performed for this thesis was exploratory in nature and no predictive method was developed owing to the paucity of the data. Nevertheless, it can certainly be said that the exploratory analysis showed promising signs for the development of a method in the future, once more data are available.

Appendix A

Appendix: Human Fab Data Generation

In this appendix the method used to generate a set of human natural V_H-V_L pair Fabs is described.

A.1 Method

The following sections describe the method in detail — for a summary, see section 6.2.1. This method is mostly consistent with that presented by Pashiardis (2015) with the exception of constant-region insertion (see section A.1.3.2 for details).

A.1.1 IgG$^+$ Memory B Cell Isolation

A 100 ml sample of blood was taken from a single human volunteer. Blood was diluted to a 1 : 1 ratio with phosphate-buffered saline (PBS) before cells were separated by centrifugation in LeucoSep tubes (Greiner Bio One) at 1000 rpm for 10 minutes. Peripheral blood mononuclear cells (PBMCs) were then harvested and transferred for centrifugation at 1200 rpm for 5 min before being re-suspended in fluorescence activated cell sorting (FACS) buffer (1% FCS, 1 mM EDTA, 25 mM HEPES and Hanks Balanced Salt Solution). For every 10^8 cells (determined by cell count), 1 µg of each of the following stains were applied for FACS isolation of IgG$^+$ memory B cells: mouse anti-human IgG Fc fragment-specific-APC, mouse anti-human CD19-PerCP, mouse anti-human light chain lambda-FITC, mouse anti-human light chain kappa-PE and mouse anti-human IgM-BV 421. Singles cells were sorted into cold 96-well plates using a BD FACS ARIA III with a 100 µm nozzle.

A.1.2 Reverse Transcription

In preparation for reverse transcription (RT), the B cell plates were kept cold under a hood and each well of each plate was made up to contain the following: 4 µl 5X buffer, 4 µl 2.5 µM oligo deoxy-thymidine (dT), 1 µl dithiothreitol, 1 µl 10% NP-40 detergent, 1 µl dNTPs (2.5 µM each), 0.5 µl RNasin (Promega), 1 µl Superscript III reverse transcriptaste (Invitrogen) and 7.5 µl diethylpyrocarbonate (DEPC) H_2O.

RT was performed at 50 °C for 60 minutes before heating at 70 °C to inactive the enzyme. Plates were then stored at −80 °C.

A.1.3 PCR

In order to produce a fragment containing naturally paired V_H and V_L regions, a three-step PCR method was undertaken. For the primary and secondary steps V_H and V_L amplification was undertaken separately, before being combined in the final step to allow the formation of a fragment containing both regions.

A.1.3.1 Primary

For each well, one primary PCR reaction was undertaken for each variable region. Each well contained a final volume of 25 µl that included the following: 1 µl of 10 µM heavy (or kappa) chain forward primer set solution (see tables A.1 and A.2), 1 µl of 10 µM of the reverse primer (heavy 5′ `CACTGTACTTTGGCCTCTCTGG` 3′, kappa 5′ `CGACACCGTCACCGGTTCGGG` 3′), 0.25 µl Herculase polymerase (Agilent), 5 µl 5X Herculase buffer, 0.25 µl dNTPs (25 mM each), 1 µl cDNA, 0.5 µl dimethylsulfoxide (DMSO) and 16 µl DEPC H_2O. The forward primer sets are designed to bind to the leader peptide sequences found upstream of the V-region genes. The reverse primers anneal within the C_H1 and CK sequences respectively.

For the PCR reaction, samples were initially held at 94 °C for 2 min followed by 40 cycles of: denaturation at 94 °C for 30 s, primer annealing at 50 °C for heavy or 58 °C for kappa for 30 s and elongation at 72 °C for 1 min. A final elongation step at 72 °C for 2 min then completes the reaction. Samples are then held at 4 °C.

A.1.3.2 Secondary

For the secondary PCR step, V_H and V_κ fragments are further amplified with a set of nested primers in order to improve the fidelity of the reaction. The forward heavy and

TABLE A.1: Primary PCR forward heavy chain primers. All primers are written in the 5′ to 3′ direction.

Sequence	Subgroup
GACACGAAGCTTGCCACCATGGACTGGACCTGGAGGATC	V_H1
GACACGAAGCTTGCCACCATGGACTGGACCTGGAGCATC	V_H1
GACACGAAGCTTGCCACCATGGACTGCACCTGGAGGATC	V_H1
GACACGAAGCTTGCCACCATGGACTGCACCTGGACGATC	V_H1
GACACGAAGCTTGCCACCATGGACTGGACCTGGAGAATC	V_H1
GACACGAAGCTTGCCACCATGGACTGGACCTGGAGGGTC	V_H1
GACACGAAGCTTGCCACCATGGACTGGACCTGGAGGTTC	V_H1
GACACGAAGCTTGCCACCATGGACTGGACCTGGACGTTC	V_H1
GACACGAAGCTTGCCACCATGGACTGGACCTGGAGGACC	V_H1
GACACGAAGCTTGCCACCATGGACTGGATTTGGAGGATC	V_H1
GACACGAAGCTTGCCACCATGGACACACTTTGCTCCACG	V_H2
GACACGAAGCTTGCCACCATGGACACACTTTGCTACACA	V_H2
GACACGAAGCTTGCCACCATGGAATTGGGGCTGAGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTGGGACTGAGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCTGAGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCTGAGTTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCTGAGATGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTGGGGCTGAGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAACTGGGGCTCCGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGACTGAGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTGGGGCTGTGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCTTAGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTTGGCTGAGCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCTGAACTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCTGAACTTG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCTGACCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTGGGGCTGTGCCGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTGGGGCTGTTCTGG	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCTGAACTGC	V_H3
GACACGAAGCTTGCCACCATGGAGTTTGGGCCGAGCTGG	V_H3
GACACGAAGCTTGCCACCATGACGGAGTTTGGGCTGAGCTGG	V_H3
GACACGAAGCTTGCCACCATGAAACACCTGTGGTTCTTC	V_H4
GACACGAAGCTTGCCACCATGAAGCACCTGTGGTTCTTC	V_H4
GACACGAAGCTTGCCACCATGAAACATCTGTGGTTCTTC	V_H4
GACACGAAGCTTGCCACCATGAAGCACCTGTGGTTTTTC	V_H4
GACACGAAGCTTGCCACCATGAAACCCCTGTGGTTCTCC	V_H4
GACACGAAGCTTGCCACCATGAAGCACCTGTGGTTTTCC	V_H4
GACACGAAGCTTGCCACCATGGGGTCAACCGCCATCCTC	V_H5
GACACGAAGCTTGCCACCATGGGGTCAACCGCCATCCTT	V_H5
GACACGAAGCTTGCCACCATGGGGTCAACCGCCATCTTC	V_H5
GACACGAAGCTTGCCACCATGGGGTCAACCGCCATCGTC	V_H5
GACACGAAGCTTGCCACCATGTCTGTCTCCTTCCTCATC	V_H6

Table A.2: Primary PCR forward kappa chain primers. All primers are written in the $5'$ to $3'$ direction.

Sequence	Subgroup
CCTAAAAGCCACGAATTCGCCACCATGGACATGAGGGTCCCCGCT	$V_\kappa 1$
CCTAAAAGCCACGAATTCGCCACCATGGACATGAGGGTCCCTGCT	$V_\kappa 1$
CCTAAAAGCCACGAATTCGCCACCATGGACATGAGGGTCCCCGTT	$V_\kappa 1$
CCTAAAAGCCACGAATTCGCCACCATGGACATGAGAGTCCTCGCT	$V_\kappa 1$
CCTAAAAGCCACGAATTCGCCACCATGGACATGAGGGTCCTCGCT	$V_\kappa 1$
CCTAAAAGCCACGAATTCGCCACCATGGACATGAGGGTGCCCGCT	$V_\kappa 1$
CCTAAAAGCCACGAATTCGCCACCATGAGGGTCCCCGCTCAGCTC	$V_\kappa 1$
CCTAAAAGCCACGAATTCGCCACCATGAGCGTGCCTACCCAGGTC	$V_\kappa 1$
CCTAAAAGCCACGAATTCGCCACCATGAGGCTCCCTGCTCAGCTC	$V_\kappa 2$
CCTAAAAGCCACGAATTCGCCACCATGAGGCTCCTTGCTCAGCTT	$V_\kappa 2$
CCTAAAAGCCACGAATTCGCCACCATGAGGCTCCCTGCTCAACTC	$V_\kappa 2$
CCTAAAAGCCACGAATTCGCCACCATGGAAACCCCAGCGCAGCTT	$V_\kappa 3$
CCTAAAAGCCACGAATTCGCCACCATGGAAGCCCCAGCGCAGCTT	$V_\kappa 3$
CCTAAAAGCCACGAATTCGCCACCATGGAAGCCCCAGCTCAGCTT	$V_\kappa 3$
CCTAAAAGCCACGAATTCGCCACCATGGAACCATGGAAGCCCCAG	$V_\kappa 3$
CCTAAAAGCCACGAATTCGCCACCATGGAAGCCCCAGCGCAGCTC	$V_\kappa 3$
CCTAAAAGCCACGAATTCGCCACCATGGAAGCCCCAGTTCAGCTC	$V_\kappa 3$
CCTAAAAGCCACGAATTCGCCACCATGGTGTTGCAGACCCAGGTC	$V_\kappa 4$
CCTAAAAGCCACGAATTCGCCACCATGGGGTCCCAGGTTCACCTC	$V_\kappa 5$
CCTAAAAGCCACGAATTCGCCACCATGTTGCCATCACAACTCATTGGG	$V_\kappa 6$
CCTAAAAGCCACGAATTCGCCACCATGGTGTCCCCGTTGCAATTC	$V_\kappa 6$

kappa primers include *NheI* and *MfeI* restriction sites respectively, which are used for the later ligation step. The reverse primers also contain complementary 22-base overhanging ends that will be used to anneal V_H and V_κ fragments in the tertiary step. This step differs from the method presented in Pashiardis (2015), where a different set of overhanging ends serve to anneal to a C_H1/CL-containing fragment in the tertiary PCR step — instead the C_H1 and CL regions are inserted by ligation after PCR (see section A.1.4).

Each 25 µl reaction mixture contained 1 µl of the primary PCR reaction product, 1 µl of 10 µM heavy (or kappa) chain forward primer set mixture (see tables A.3 and A.5), 1 µl of 10 µM of the reverse primer mixture (see tables A.4 and A.6), 0.25 µl Herculase polymerase (Agilent), 5 µl 5X Herculase buffer, 0.25 µl dNTPs (25 mM each), 0.5 µl DMSO and 16 µl DEPC H_2O.

For the secondary PCR reaction, samples are initially heated at 94 °C for 2 min, followed by 40 cycles of: denaturation at 94 °C for 30 s, primer annealing at 50 °C for 30 s and elongation at 72 °C for 30 s. The reaction is completed with a final elongation step at 72 °C for 2 min and is then held at 4 °C.

Table A.3: Secondary PCR forward heavy chain primers. All primers are written in the $5'$ to $3'$ direction.

Sequence	Subgroup
TTCCTGCTAGCTGCAGCCACAGGTGCCCACTCC	$V_H1/7$
TTCCTGCTAGCTGCAGCTACAGGCACCCACGCC	V_H1
TTCCTGCTAGCTGCAGCCACAGGTGCCTACTCC	V_H1
TTCCTGCTAGCTGCAGCTACAGGTGTCCAGTCC	V_H1
TTCCTGCTAGCTATCCCTTCATGGGTCTTGTCC	V_H2
TTCCTGCTAGCTGTCCCGTCCTGGGTCTTATCC	V_H2
TTCCTGCTAGCTATTTTAAAAGGTGTCCAGTGT	V_H3
TTCCTGCTAGCTATTTTAAAAGGTGTCCAATGT	V_H3
TTCCTGCTAGCTGCTCCCAGATGGGTCCTGTCT	V_H4
TTCCTGCTAGCTGCTCCCAGATGGGTCCTGTCC	V_H4
TTCCTGCTAGCTGTTCTCCAAGGAGTCTGTTCC	V_H5
TTCCTGCTAGCTCTCCCATGGGGTGTCCTGTCA	V_H6

Table A.4: Secondary PCR reverse heavy chain primers. All primers are written in the $5'$ to $3'$ direction.

Sequence	Subgroup
CGAAGCTAGTCACGATCGCATGCTCGAGACGGTGACCAGGGTGCCCTGGCC	V_H1
CGAAGCTAGTCACGATCGCATGCTCGAGACAGTGACCAGGGTGCCACGGCC	V_H2
CGAAGCTAGTCACGATCGCATGCTCGAGACGGTGACCATTGTCCCTTGGCC	V_H3
CGAAGCTAGTCACGATCGCATGCTCGAGACGGTGACCAGGGTTCCCTGGCC	$V_H4/5$
CGAAGCTAGTCACGATCGCATGCTCGAGACGGTGACCAGGGTTCCTTGGCC	$V_H4/5$
CGAAGCTAGTCACGATCGCATGCTCGAGACGGTGACCAGGGTCCCTTGGCC	$V_H4/5$
CGAAGCTAGTCACGATCGCATGCTCGAGACGGTGACCGTGGTCCCTTGGCC	V_H6
CGAAGCTAGTCACGATCGCATGCTCGAGACGGTGACCGTGGTCCCTTGCCC	$V_\kappa6$

Table A.5: Secondary PCR forward kappa chain primers. All primers are written in the $5'$ to $3'$ direction.

Sequence	Subgroup
CCATCAATTGCTGGGGCTCCTGCTGCTCTGG	$V_\kappa1$
CCATCAATTGCTGGGGCTCCTGCTGCTCTGT	$V_\kappa1$
CCATCAATTGCTGGGGCTCCTGCAGCTCTGG	$V_\kappa1$
CCATCAATTGCTGGGGCTCCTGCTACTCTGG	$V_\kappa1$
CCATCAATTGCTGGGGCTGCTAATGCTCTGG	$V_\kappa2$
CCATCAATTGCTCTTCCTCCTGCTACTCTGG	$V_\kappa3$
CCATCAATTGTTCATTTCTCTGTTGCTCTGG	$V_\kappa4$
CCATCAATTGCTCAGCTTCCTCCTCCTTTGG	$V_\kappa5$
CCATCAATTGATTGGGTTTCTGCTGCTCTGG	$V_\kappa6$

Table A.6: Secondary PCR reverse kappa chain primers. All primers are written in the $5'$ to $3'$ direction.

Sequence	Subgroup
CATGCGATCGTGACTAGCTTCGTGGGGCCGCTACCGTACGTTTGATTTCC	$V_\kappa1$
CATGCGATCGTGACTAGCTTCGTGGGGCCGCTACCGTACGTTTGATCTCC	$V_\kappa2/4$
CATGCGATCGTGACTAGCTTCGTGGGGCCGCTACCGTACGTTTGATATCC	$V_\kappa3$
CATGCGATCGTGACTAGCTTCGTGGGGCCGCTACCGTACGTTTAATCTCC	$V_\kappa5$

A.1.3.3 Tertiary

For the final PCR step, secondary V_H and V_κ products were combined and diluted 1 : 10 in H_2O in order to reduce reagent carry-over. The complementary sequences in the reverse primers of the secondary reaction allow the fragments to anneal, producing a fragment with V_H and V_κ on opposite strands. The secondary forward heavy and forward kappa are used as forward and reverse primers in the reaction.

Each 25 µl reaction mixture contained 1 µl of the combined secondary product mixture, 1 µl of 10 µM secondary PCR heavy chain forward primer set mixture (see table A.3), 1 µl of 10 µM secondary PCR kappa chain forward primer set solution (see table A.5), 0.25 µl Herculase polymerase (Agilent), 5 µl 5X Herculase buffer, 0.25 µl dNTPs (25 mM each), 0.5 µl DMSO and 16 µl DEPC H_2O.

For the secondary PCR reaction, samples are initially heated at 94 °C for 2 min, followed by 30 cycles of: denaturation at 94 °C for 30 s, primer annealing at 50 °C for 30 s and elongation at 72 °C for 40 s. The reaction is completed with a final elongation step at 72 °C for 2 min and is then held at 4 °C.

A.1.4 Expression Plasmid Construction

Variable region fragments were then inserted into a Fab-expression plasmid. This expression plasmid contains a chEF1a promoter followed by V_H leader sequence upstream from a *MfeI* restriction site on one strand and a hCMV promoter followed by V_κ leader sequence upstream from a *NheI* restriction site. The variable-region fragment is inserted between these sites by ligation. The expression plasmid also contains a *kan* gene for kanamycin resistance and a pUC origin site for replication.

Successful tertiary products were then pooled for the ligation step. 43 µl of the pool was digested with restriction enzymes MfeI and NheI (both 1 µl) in 5 µl 10X NEB4 buffer at 37 °C for 90 min, before being purified by gel electrophoresis. Fragments were then ligated by T4 ligation into a Fab-expression vector previously treated with MfeI, NhweI and calf intestinal alkaline phosphatase (CiP). Ligation was done with 1 µl of the digested vector, 2 µl H_2O, 2 µl digested fragment, 0.5 µl T4 ligase and 1 µl T4 buffer at room temperature for 60 min.

The ligation mixture was then used to transform competent XL1-Blue supercompetent cells by heat shock following the manufacturer's instructions (Agilent). Single colonies of transformants were selected from Luria broth agar plates containing 30 µg ml^{-1} kanamycin and incubated for 37 °C for 24 hours. Plasmid DNA was then extracted by mini-prep according to the manufacturer's instructions (Qiagen).

Extracted plasmid DNA was then pooled for C-region fragment insertion. The C-region fragment contains the CK sequence and the C_H1 sequence on opposite strands, with a bidirectional poly(A) signal between them. The C_H1 sequence is also followed by a poly-his tag. Owing to the reverse primers used in the PCR steps, every V-region fragment contains partial C-region sequence that contain *Bsi* and *Xho* sites. Thus the V-region plasmid and C-region fragment can be digested and ligated together. For the digestion, 20 µl of the plasmid DNA pool was added to 5 µl buffer H, 1 µl BsiW1, 1 µl Xho1 and 23 µl H_2O. The mixture was kept at 37 °C for 1 h then 55 °C for 1 h before linearised plasmids were purified by gel electrophoresis. C-region fragments previously treated with Bsi, Xho and CiP were then ligated by T4 ligation into the plasmids using the same ligation protocol as above. Transformation with the ligation mixture, plating and incubation was then undertaken following the same procedure as above.

A.1.5 Transfection and Expression

Fab fragment plasmids were transiently transfected into HEK-293 cells. For each Fab fragment plasmid, 25–30 µg of plasmid DNA was diluted in 1.5 ml OpitiMEM media. 80 µl ExpiFactamine 293 was then diluted in 1.5 ml OptiMEM, incubated for 5 min and then mixed with the DNA solution and incubated for 30 min. This mixture was then transferred to a flask containing 25.5 ml of HEK-293 cells at a density of 30×10^6 cells per ml. After incubation at 37 °C for 24 hours, 150 µl of transfection enhancer 1 and 1.5 ml transfection enhancer 2 were added to the mixture. The mixture was then incubated for 6 days before being spun down for 10 min at 1400 rpm. The supernatant was collected and sterilised by filtering it with a 0.22 µm filter. Sterilised supernatant was stored at 4 °C.

A.1.6 His-tag purification

Purification of HEK-293 supernatants was carried out in a 2 ml 96-well plate using the automated Phynexus system. First, 1 ml PhyTips containing 10 µl Nickel-Sep resin to bind his-tag were equilibrated with PBS pH 7.4. Each supernatant was then aspirated and dispensed eight times before two wash steps were carried out: the first wash buffer was a solution of 50 mM NaP and 0.5 M NaCl pH 6.0 and the second was a solution of 500 mM NaCl. Elutions were done in four 100 µl aliquots. All samples were eluted in 50 mM acetate and 1 M NaCl pH 4.0, neutralised with 60 µl 0.75 M Na_2HPO_4, before being concentrated to above 0.1 $\mu g\,ml^{-1}$ using 10 000 Da molecular weight cut off filters. Final concentrations were determined by UV-spectroscopy at 280 nm using a NanoDrop (Thermo Scientific). Purified samples were stored at 4 °C.

A.1.7 HPLC purity

The purity of purified samples was checked by analytical size exclusion using a BEH200 150 mm column. Elution was detected by fluorescence with excitation at 280 nm and emission at 340 nm.

A.1.8 Biophysical assays

Two biophysical assays were performed on each successfully expressed and purified Fab. Thermofluor assays were performed to determine melting temperature (Tm) values for each Fab. Hydrophobic interaction chromatography assays were also performed to determine the retention time of Fabs on a hydrophobic column, thus giving us a measure of surface hydrophobicity.

A.1.8.1 Thermofluor assay

Purified Fabs were diluted to 0.11 $\mathrm{mg\,ml^{-1}}$ in PBS pH 7.4 on a 96-well plate. 45 µl of each sample was mixed with 5 µl of 30X SPYRO orange (Invitrogen), a fluorescent dye which binds hydrophobic residues but is quenched when interacting with water. 10 µl of the mixture was dispensed in quadruplicate into a 384 PCR optical well-plate, which was run on a 7900HT Fast Real-Time PCR System (Agilent). In a peltier-based thermal cycling system, the proteins were initially exposed to a temperature of 20 °C which then increased up to 99 °C at a rate of 1.1 $\mathrm{°C\,min^{-1}}$ As the temperature increases and the protein unfolds, core hydrophobic residues become exposed and bind the dye, increasing fluorescence. Fluorescence within the wells was monitored by a charge-coupled device. For each Fab, fluorescence intensity was plotted and the inflection point of the slope was taken as the melting temperature.

A.1.8.2 Hydrophobic interaction chromatography assay

A 100 mm × 4.6 mm Dionex ProPac HIC-10 column was used for all runs. An Agilent 1260 HPLC equipped with a fluorescence detector was equilibrated for 2 min with 50 mM sodium phosphate pH 7.4 at 20 °C using a flow rate of 0.8 $\mathrm{ml\,min^{-1}}$. The mobile phase consists of 0.8 M $(NH_4)_2SO_4$ and 50 mM Na_2HPO_4. 50 µl of each sample diluted to 0.11 $\mathrm{mg\,ml^{-1}}$ was injected into the column. Starting with 2 min at 0% 50 mM Na_2HPO_4, the Fab was eluted using a linear gradient from 0% to 100%. The elution is detected by fluorescence with excitation at 280 nm and emission at 340 nm. Between samples, the column was washed with 100% 50 mM Na_2HPO_4 for 2 min and re-equilibrated with 0% 50 mM for 10 min.

Bibliography

Abhinandan, K. R. and Martin, A. C. (2008), 'Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains', *Mol. Immunol.* **45**(14), 3832–3839.

Abhinandan, K. R. and Martin, A. C. (2010), 'Analysis and prediction of VH/VL packing in antibodies', *Protein Eng. Des. Sel.* **23**(9), 689–697.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002), The generation of antibody diversity, *in* 'Molecular biology of the cell', 4 edn, Garland Science.

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

Alpaydin, E. (2009), *Introduction to Machine Learning*, 2 edn, The MIT Press.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990), 'Basic local alignment search tool', *J. Mol. Biol.* **215**(3), 403–410.

Amit, A. G., Mariuzza, R. A., Phillips, S. E. and Poljak, R. J. (1985), 'Three-dimensional structure of an antigen-antibody complex at 6 A resolution', *Nature* **313**(5998), 156–158.

Anderson, M. S., Venanzi, E. S., Klein, L., Chen, Z., Berzins, S. P., Turley, S. J., von Boehmer, H., Bronson, R., Dierich, A., Benoist, C. and Mathis, D. (2002), 'Projection of an immunological self shadow within the thymus by the aire protein', *Science* **298**(5597), 1395–1401.

Ansari, H. R. and Raghava, G. P. (2010), 'Identification of conformational B-cell Epitopes in an antigen from its primary sequence', *Immunome Res* **6**, 6.

Baker, E. N. and Hubbard, R. E. (1984), 'Hydrogen bonding in globular proteins', *Prog. Biophys. Mol. Biol.* **44**(2), 97–179.

Baresic, A. (2011), Structural analysis of single amino acid polymorphisms, PhD thesis, University College London.

Barlow, D. J., Edwards, M. S. and Thornton, J. M. (1986), 'Continuous and discontinuous protein antigenic determinants', *Nature* **322**(6081), 747–748.

Barthelemy, P. A., Raab, H., Appleton, B. A., Bond, C. J., Wu, P., Wiesmann, C. and Sidhu, S. S. (2008), 'Comprehensive analysis of the factors contributing to the stability and solubility of autonomous human VH domains', *J. Biol. Chem.* **283**(6), 3639–3654.

Batista, G. E. A. P. A. and Monard, M. C. (2003), 'An analysis of four missing data treatment methods for supervised learning', *Applied Artificial Intelligence* **17**(5-6), 519–533.

Beck, A., Wagner-Rousset, E., Ayoub, D., Van Dorsselaer, A. and Sanglier-Cianferani, S. (2013), 'Characterization of therapeutic antibodies and related products', *Anal. Chem.* **85**(2), 715–736.

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the royal statistical society. Series B (Methodological)* pp. 289–300.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000), 'The Protein Data Bank', *Nucleic Acids Res.* **28**(1), 235–242.

Bewick, V., Cheek, L. and Ball, J. (2004), 'Statistics review 8: Qualitative data - tests of association', *Crit Care* **8**(1), 46–53.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L. and Schwede, T. (2014), 'SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information', *Nucleic Acids Res.* **42**(Web Server issue), W252–258.

Bird, R. E., Hardman, K. D., Jacobson, J. W., Johnson, S., Kaufman, B. M., Lee, S. M., Lee, T., Pope, S. H., Riordan, G. S. and Whitlow, M. (1988), 'Single-chain antigen-binding proteins', *Science* **242**(4877), 423–426.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C. (1987), 'Structure of the human class I histocompatibility antigen, HLA-A2', *Nature* **329**(6139), 506–512.

Blackshields, G., Sievers, F., Shi, W., Wilm, A. and Higgins, D. G. (2010), 'Sequence embedding for fast construction of guide trees for multiple sequence alignment', *Algorithms Mol Biol* **5**, 21.

Blythe, M. J. and Flower, D. R. (2005), 'Benchmarking B cell epitope prediction: Underperformance of existing methods', *Protein Sci* **14**, 246–248.

Borg, I. and Groenen, P. J. F. (2005), *Modern Multidimensional Scaling*, 2 edn, Springer-Verlag New York.

Branden, C. and Tooze, J. (1991), *Introduction to protein structure*, Introduction to Protein Structure, Garland Pub.

Breiman, L. (1996), 'Bagging predictors', *Mach. Learn.* **24**(2), 123–140.

Breiman, L. (2001), 'Random forests', *Mach. Learn.* **45**(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. (1984), *Classification and Regression Trees*, 1 edn, Wadsworth International Group.

Breslow, L. A. and Aha, D. W. (1997), 'Simplifying decision trees: A survey', *The Knowledge Engineering Review* **12**(01), 1–40.

Brown, M. B. and Forsythe, A. B. (1974), 'Robust tests for the equality of variances', *Journal of the American Statistical Association* **69**(346), 364–367.

Buzza, M. S., Zamurs, L., Sun, J., Bird, C. H., Smith, A. I., Trapani, J. A., Froelich, C. J., Nice, E. C. and Bird, P. I. (2005), 'Extracellular matrix remodeling by human granzyme B via cleavage of vitronectin, fibronectin, and laminin', *J. Biol. Chem.* **280**(25), 23549–23558.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A. and Apweiler, R. (2003), 'The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro', *Genome Res.* **13**(4), 662–672.

Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. and Trout, B. L. (2009), 'Design of therapeutic proteins with enhanced stability', *Proc. Natl. Acad. Sci. U.S.A.* **106**(29), 11937–11942.

Chothia, C. and Lesk, A. M. (1987), 'Canonical structures for the hypervariable regions of immunoglobulins', *J. Mol. Biol.* **196**(4), 901–917.

Chuang, G. Y., Kozakov, D., Brenke, R., Comeau, S. R. and Vajda, S. (2008), 'DARS (Decoys As the Reference State) potentials for protein-protein docking', *Biophys. J.* **95**(9), 4217–4227.

Ciavatta, D. J., Yang, J., Preston, G. A., Badhwar, A. K., Xiao, H., Hewins, P., Nester, C. M., Pendergraft, W. F., Magnuson, T. R., Jennette, J. C. and Falk, R. J. (2010), 'Epigenetic basis for aberrant upregulation of autoantigen genes in humans with ANCA vasculitis', *J. Clin. Invest.* **120**(9), 3209–3219.

Cohen, J. (1977), The significance of a product moment rs, *in* J. Cohen, ed., 'Statistical Power Analysis for the Behavioral Sciences (Revised Edition)', revised edition edn, Academic Press, pp. 75 – 107.

Coons, A., Creech, H., Jones, R. and Berliner, E. (1942), 'Demonstration of pneumoccocal antigen in tissues by use of fluorescent antibody.', *J. Immunol.* **45**, 159–170.

Cox, T. F. and Cox, M. A. A. (2001), *Multidimensional Scaling*, 2 edn, CRC/Capman and Hall.

Darrah, E. and Rosen, A. (2010), 'Granzyme B cleavage of autoantigens in autoimmunity', *Cell Death Differ.* **17**(4), 624–632.

Darsley, M. J. and Rees, A. R. (1985), 'Nucleotide sequences of five anti-lysozyme monoclonal antibodies', *EMBO J.* **4**(2), 393–398.

Dudgeon, K., Famm, K. and Christ, D. (2009), 'Sequence determinants of protein aggregation in human VH domains', *Protein Eng. Des. Sel.* **22**(3), 217–220.

Dudgeon, K., Rouet, R., Kokmeijer, I., Schofield, P., Stolp, J., Langley, D., Stock, D. and Christ, D. (2012), 'General strategy for the generation of human antibody variable domains with increased aggregation resistance', *Proc. Natl. Acad. Sci. U.S.A.* **109**(27), 10879–10884.

Dunn, O. J. (1961), 'Multiple comparisons among means', *Journal of the American Statistical Association* **56**(293), 52–64.

Ecker, D. M., Jones, S. D. and Levine, H. L. (2015), 'The therapeutic monoclonal antibody market', *MAbs* **7**(1), 9–14.

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

Eisenberg, D., Weiss, R. M., Terwilliger, T. C. and Wilcox, W. (1982), 'Hydrophobic moments and protein structure', *Faraday Symp. Chem. Soc.* **17**, 109–120.

El-Manzalawy, Y. and Honavar, V. (2010), 'Recent advances in B-cell epitope prediction methods', *Immunome Res* **6 Suppl 2**, S2–S2.

Elkon, K. and Casali, P. (2008), 'Nature and functions of autoantibodies', *Nat Clin Pract Rheumatol* **4**(9), 491–498.

Estabrooks, A., Jo, T. and Japkowicz, N. (2004), 'A multiple resampling method for learning from imbalanced data sets', *Computational Intelligence* **20**(1), 18–36.

Ewert, S., Huber, T., Honegger, A. and Pluckthun, A. (2003), 'Biophysical properties of human antibody variable domains', *J. Mol. Biol.* **325**(3), 531–553.

Fagreaus, A. (1948), 'Antibody production in relation to the development of plasma cells', *Acta. Med. Scand.* **Suppl. 204**.

Ferdous, S. and Martin, A. (in press), 'AbDb: Antibody structure database - a database of PDB derived antibody structures'.

Ferry, H., Jones, M., Vaux, D. J., Roberts, I. S. and Cornall, R. J. (2003), 'The cellular location of self-antigen determines the positive and negative selection of autoreactive B cells', *J. Exp. Med.* **198**(9), 1415–1425.

Ferry, H., Potter, P. K., Crockford, T. L., Nijnik, A., Ehrenstein, M. R., Walport, M. J., Botto, M. and Cornall, R. J. (2007), 'Increased positive selection of B1 cells and reduced B cell tolerance to intracellular antigens in c1q-deficient mice', *J. Immunol.* **178**(5), 2916–2922.

Fisher, R. A. (1922), 'On the interpretation of chi-square from contingency tables, and the calculation of p', *Journal of the Royal Statistical Society* **85**(1), 87–94.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012), 'CD-HIT: accelerated for clustering the next-generation sequencing data', *Bioinformatics* **28**(23), 3150–3152.

Fulcher, D. A., Lyons, A. B., Korn, S. L., Cook, M. C., Koleda, C., Parish, C., Fazekas de St Groth, B. and Basten, A. (1996), 'The fate of self-reactive B cells depends primarily on the degree of antigen receptor engagement and availability of T cell help', *J. Exp. Med.* **183**(5), 2313–2328.

Gabb, H. A., Jackson, R. M. and Sternberg, M. J. (1997), 'Modelling protein docking using shape complementarity, electrostatics and biochemical information', *J. Mol. Biol.* **272**(1), 106–120.

Gahring, L., Carlson, N. G., Meyer, E. L. and Rogers, S. W. (2001), 'Granzyme B proteolysis of a neuronal glutamate receptor generates an autoantigen and is modulated by glycosylation', *J. Immunol.* **166**(3), 1433–1438.

Gallucci, S., Lolkema, M. and Matzinger, P. (1999), 'Natural adjuvants: endogenous activators of dendritic cells', *Nat. Med.* **5**(11), 1249–1255.

Garber, E. and Demarest, S. J. (2007), 'A broad range of Fab stabilities within a host of therapeutic IgGs', *Biochem. Biophys. Res. Commun.* **355**(3), 751–757.

Gay, D., Saunders, T., Camper, S. and Weigert, M. (1993), 'Receptor editing: an approach by autoreactive B cells to escape tolerance', *J. Exp. Med.* **177**(4), 999–1008.

Goldman, A. (1995), 'How to make my blood boil', *Structure* **3**(12), 1277–1279.

Golovin, A., Oldfield, T. J., Tate, J. G., Velankar, S., Barton, G. J., Boutselakis, H., Dimitropoulos, D., Fillon, J., Hussain, A., Ionides, J. M., John, M., Keller, P. A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Pajon, A., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, G. J., Tagari, M., Tromm, S., Vranken, W. and Henrick, K. (2004), 'E-MSD: an integrated data resource for bioinformatics', *Nucleic Acids Res.* **32**(Database issue), D211–216.

Goodnow, C. C., Crosbie, J., Adelstein, S., Lavoie, T. B., Smith-Gill, S. J., Brink, R. A., Pritchard-Briscoe, H., Wotherspoon, J. S., Loblay, R. H. and Raphael, K. (1988), 'Altered immunoglobulin expression and functional silencing of self-reactive B lymphocytes in transgenic mice', *Nature* **334**(6184), 676–682.

Greenberg, A. S., Avila, D., Hughes, M., Hughes, A., McKinney, E. C. and Flajnik, M. F. (1995), 'A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks', *Nature* **374**(6518), 168–173.

Griffith, T. S., Yu, X., Herndon, J. M., Green, D. R. and Ferguson, T. A. (1996), 'CD95-induced apoptosis of lymphocytes in an immune privileged site induces immunological tolerance', *Immunity* **5**(1), 7–16.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009), 'The weka data mining software: An update', *SIGKDD Explor. Newsl.* **11**(1), 10–18.

Hansen, M. H., Nielsen, H. V. and Ditzel, H. J. (2002), 'Translocation of an intracellular antigen to the surface of medullary breast cancer cells early in apoptosis allows for an antigen-driven antibody response elicited by tumor-infiltrating B cells', *J. Immunol.* **169**(5), 2701–2711.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. and White, R. (2004), 'The Gene Ontology (GO) database and informatics resource', *Nucleic Acids Res.* **32**(Database issue), D258–261.

Hartley, S. B., Cooke, M. P., Fulcher, D. A., Harris, A. W., Cory, S., Basten, A. and Goodnow, C. C. (1993), 'Elimination of self-reactive B lymphocytes proceeds in two stages: arrested development and cell death', *Cell* **72**(3), 325–335.

Haste Andersen, P., Nielsen, M. and Lund, O. (2006), 'Prediction of residues in discontinuous B-cell epitopes using protein 3D structures', *Protein Sci.* **15**(11), 2558–2567.

Hastie, T., Tibshirani, R., Friedman and J. (2009), *The Elements of Statistical Learning*, 2 edn, Springer-Verlang New York.

Hazes, B. and Dijkstra, B. W. (1988), 'Model building of disulfide bonds in proteins with known three-dimensional structure', *Protein Engineering* **2**(2), 119–125.

He, H. and Garcia, E. A. (2009), 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284.

Hochman, J., Inbar, D. and Givol, D. (1973), 'An active antibody fragment (Fv) composed of the variable portions of heavy and light chains', *Biochemistry* **12**(6), 1130–1135.

Holliger, P., Prospero, T. and Winter, G. (1993), '"Diabodies": small bivalent and bispecific antibody fragments', *Proc. Natl. Acad. Sci. U.S.A.* **90**(14), 6444–6448.

Honegger, A., Malebranche, A. D., Rothlisberger, D. and Pluckthun, A. (2009), 'The influence of the framework core residues on the biophysical properties of immunoglobulin heavy chain variable domains', *Protein Eng. Des. Sel.* **22**(3), 121–134.

Honegger, A. and Pluckthun, A. (2001), 'Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool', *J. Mol. Biol.* **309**(3), 657–670.

Hopp, T. P. (1986), 'Protein surface analysis. Methods for identifying antigenic determinants and other interaction sites', *J. Immunol. Methods* **88**(1), 1–18.

Hopp, T. P. and Woods, K. R. (1981), 'Prediction of protein antigenic determinants from amino acid sequences', *Proc. Natl. Acad. Sci. U.S.A.* **78**(6), 3824–3828.

Hozumi, N. and Tonegawa, S. (1976), 'Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions', *Proc. Natl. Acad. Sci. U.S.A.* **73**(10), 3628–3632.

Hu, Y. J., Lin, S. C., Lin, Y. L., Lin, K. H. and You, S. N. (2014), 'A meta-learning approach for B-cell conformational epitope prediction', *BMC Bioinformatics* **15**, 378.

Huang, J., Ru, B. and Dai, P. (2011), 'Bioinformatics resources and tools for phage display', *Molecules* **16**(1), 694–709.

Hwang, W. Y. and Foote, J. (2005), 'Immunogenicity of engineered antibodies', *Methods* **36**(1), 3–10.

Iakhiaev, M. A. and Iakhiaev, A. V. (2010), 'Graph-theoretical comparison of protein surfaces reveals potential determinants of cross-reactivity and the molecular mimicry', *Mol. Immunol.* **47**(4), 719–725.

Iliades, P., Kortt, A. A. and Hudson, P. J. (1997), 'Triabodies: single chain Fv fragments without a linker form trivalent trimers', *FEBS Lett.* **409**(3), 437–441.

Japkowicz, N. and Stephen, S. (2002), 'The class imbalance problem: A systematic study', *Intell. Data Anal.* **6**(5), 429–449.

Jawa, V., Cousens, L. P., Awwad, M., Wakshull, E., Kropshofer, H. and De Groot, A. S. (2013), 'T-cell dependent immunogenicity of protein therapeutics: Preclinical assessment and mitigation', *Clin. Immunol.* **149**(3), 534–555.

Jayaram, N., Bhowmick, P. and Martin, A. C. (2012), 'Germline VH/VL pairing in antibodies', *Protein Eng. Des. Sel.* **25**(10), 523–530.

Jermutus, L., Honegger, A., Schwesinger, F., Hanes, J. and Pluckthun, A. (2001), 'Tailoring in vitro evolution for protein affinity or stability', *Proc. Natl. Acad. Sci. U.S.A.* **98**(1), 75–80.

Jespers, L., Schon, O., Famm, K. and Winter, G. (2004), 'Aggregation-resistant domain antibodies selected on phage by heat denaturation', *Nat. Biotechnol.* **22**(9), 1161–1165.

Jolliffe, I. (2014), 'Principal Component Analysis', Wiley StatsRef: Statistics Reference Online. Accessed: 2016-09-21.

Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. and Winter, G. (1986), 'Replacing the complementarity-determining regions in a human antibody with those from a mouse', *Nature* **321**(6069), 522–525.

Jones, S. and Thornton, J. M. (1997), 'Analysis of protein-protein interaction sites using surface patches', *J. Mol. Biol.* **272**(1), 121–132.

Jung, S., Spinelli, S., Schimmele, B., Honegger, A., Pugliese, L., Cambillau, C. and Pluckthun, A. (2001), 'The importance of framework residues H6, H7 and H10 in antibody heavy chains: experimental evidence for a new structural subclassification of antibody V(H) domains', *J. Mol. Biol.* **309**(3), 701–716.

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

Kabsch, W. and Sander, C. (1983), 'Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features', *Biopolymers* **22**(12), 2577–2637.

Kappler, J. W., Roehm, N. and Marrack, P. (1987), 'T cell tolerance by clonal elimination in the thymus', *Cell* **49**(2), 273–280.

Klee, G. G. (2000), 'Human anti-mouse antibodies', *Arch Pathol Lab Med* **124**, 921–923.

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

Koren, E., De Groot, A. S., Jawa, V., Beck, K. D., Boone, T., Rivera, D., Li, L., Mytych, D., Koscec, M., Weeraratne, D., Swanson, S. and Martin, W. (2007), 'Clinical validation of the "in silico" prediction of immunogenicity of a human recombinant therapeutic protein', *Clin. Immunol.* **124**(1), 26–32.

Kotsiantis, S. B. (2007), Supervised machine learning: A review of classification techniques, *in* 'Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies', IOS Press, Amsterdam, The Netherlands, The Netherlands, pp. 3–24.

Krawczyk, K., Baker, T., Shi, J. and Deane, C. M. (2013), 'Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking', *Protein Eng. Des. Sel.* **26**(10), 621–629.

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

Kringelum, J. V., Nielsen, M., Padkjaer, S. B. and Lund, O. (2013), 'Structural analysis of B-cell epitopes in antibody:protein complexes', *Mol. Immunol.* **53**(1-2), 24–34.

Kulkarni-Kale, U., Bhosle, S. and Kolaskar, A. S. (2005), 'CEP: a conformational epitope prediction server', *Nucleic Acids Res.* **33**(Web Server issue), W168–171.

Kunik, V. and Ofran, Y. (2013), 'The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops', *Protein Eng. Des. Sel.* .

Kyte, J. and Doolittle, R. F. (1982), 'A simple method for displaying the hydropathic character of a protein', *Journal of Molecular Biology* **157**(1), 105 – 132.

Lawrence, M. S., Phillips, K. J. and Liu, D. R. (2007), 'Supercharging proteins can impart unusual resilience', *J. Am. Chem. Soc.* **129**(33), 10110–10112.

Lazar, G. A., Desjarlais, J. R., Jacinto, J., Karki, S. and Hammond, P. W. (2007), 'A molecular immunology approach to antibody humanization and functional optimization', *Mol. Immunol.* **44**(8), 1986–1998.

Lechler, R., Chai, J. G., Marelli-Berg, F. and Lombardi, G. (2001), 'The contributions of T-cell anergy to peripheral T-cell tolerance', *Immunology* **103**(3), 262–269.

Lee, B. and Richards, F. M. (1971), 'The interpretation of protein structures: estimation of static accessibility', *J. Mol. Biol.* **55**(3), 379–400.

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

Lefranc, M. P., Pommie, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V. and Lefranc, G. (2003), 'IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains', *Dev. Comp. Immunol.* **27**(1), 55–77.

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

Liang, S., Zheng, D., Standley, D. M., Yao, B., Zacharias, M. and Zhang, C. (2010), 'EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results', *BMC Bioinformatics* **11**, 381.

Lin, S. Y., Cheng, C. W. and Su, E. C. (2013), 'Prediction of B-cell epitopes using evolutionary information and propensity scales', *BMC Bioinformatics* **14 Suppl 2**, S10.

Manoury, B., Hewitt, E. W., Morrice, N., Dando, P. M., Barrett, A. J. and Watts, C. (1998), 'An asparaginyl endopeptidase processes a microbial antigen for class II MHC presentation', *Nature* **396**(6712), 695–699.

Martin, A. C. (2005), 'Mapping PDB chains to UniProtKB entries', *Bioinformatics* **21**(23), 4297–4301.

McCafferty, J., Griffiths, A. D., Winter, G. and Chiswell, D. J. (1990), 'Phage antibodies: filamentous phage displaying antibody variable domains', *Nature* **348**(6301), 552–554.

McDonald, J. H. (2009), *Handbook of biological statistics*, Vol. 2.

McDonalds, J. (2014), *Handbook of Biological Statistics*, 3 edn, Sparky House Publishing.

McGargill, M. A., Derbinski, J. M. and Hogquist, K. A. (2000), 'Receptor editing in developing T cells', *Nat. Immunol.* **1**(4), 336–341.

McMillan, L. E. and Martin, A. C. (2008), 'Automatically extracting functionally equivalent proteins from SwissProt', *BMC Bioinformatics* **9**, 418.

Mitchell, T. M. (1997), *Machine Learning*, 1 edn, McGraw-Hill, Inc., New York, NY, USA.

Monsellier, E. and Bedouelle, H. (2006), 'Improving the stability of an antibody variable fragment by a combination of knowledge-based approaches: validation and mechanisms', *J. Mol. Biol.* **362**(3), 580–593.

Mood, A., Graybill, F. A. and Boes, D. C. (1974), *Introduction to the Theory of Statistics*, 3 edn, McGraw-Hill.

Morrison, S. L., Johnson, M. J., Herzenberg, L. A. and Oi, V. T. (1984), 'Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains', *Proc. Natl. Acad. Sci. U.S.A.* **81**(21), 6851–6855.

Moser, B., Stevens, G. and Watts, L. (1989), 'The two-sample t test versus Satterthwaite's approximate F test', *Communications in Statistics - Theory and Methods* **18**(11), 3963–3975.

Mumey, B. M., Bailey, B. W., Kirkpatrick, B., Jesaitis, A. J., Angel, T. and Dratz, E. A. (2003), 'A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins', *J. Comput. Biol.* **10**(3-4), 555–567.

Mungall, A. J., Palmer, S. A., Sims, S. K., Edwards, C. A., Ashurst, J. L., Wilming, L., Jones, M. C., Horton, R., Hunt, S. E., Scott, C. E., Gilbert, J. G., Clamp, M. E., Bethel, G., Milne, S., Ainscough, R., Almeida, J. P., Ambrose, K. D., Andrews, T. D., Ashwell, R. I., Babbage, A. K., Bagguley, C. L., Bailey, J., Banerjee, R., Barker, D. J., Barlow, K. F., Bates, K., Beare, D. M., Beasley, H., Beasley, O., Bird, C. P., Blakey, S., Bray-Allen, S., Brook, J., Brown, A. J., Brown, J. Y., Burford, D. C., Burrill, W., Burton, J., Carder, C., Carter, N. P., Chapman, J. C., Clark, S. Y., Clark, G., Clee, C. M., Clegg, S., Cobley, V., Collier, R. E., Collins, J. E., Colman, L. K., Corby, N. R., Coville, G. J., Culley, K. M., Dhami, P., Davies, J., Dunn, M., Earthrowl, M. E., Ellington, A. E., Evans, K. A., Faulkner, L., Francis, M. D., Frankish, A., Frankland, J., French, L., Garner, P., Garnett, J., Ghori, M. J., Gilby, L. M., Gillson, C. J., Glithero, R. J., Grafham, D. V., Grant, M., Gribble, S., Griffiths, C., Griffiths, M., Hall, R., Halls, K. S., Hammond, S., Harley, J. L., Hart, E. A., Heath, P. D., Heathcott, R., Holmes, S. J., Howden, P. J., Howe, K. L., Howell, G. R., Huckle, E., Humphray, S. J., Humphries, M. D., Hunt, A. R., Johnson, C. M., Joy, A. A., Kay, M., Keenan, S. J., Kimberley, A. M., King, A., Laird, G. K., Langford, C., Lawlor, S., Leongamornlert, D. A., Leversha, M., Lloyd, C. R., Lloyd, D. M., Loveland, J. E., Lovell, J., Martin, S., Mashreghi-Mohammadi, M., Maslen, G. L., Matthews, L., McCann, O. T., McLaren, S. J., McLay, K., McMurray, A., Moore, M. J., Mullikin, J. C., Niblett, D., Nickerson, T., Novik, K. L., Oliver, K., Overton-Larty, E. K., Parker, A., Patel, R., Pearce, A. V., Peck, A. I., Phillimore, B., Phillips, S., Plumb, R. W., Porter, K. M., Ramsey, Y., Ranby, S. A., Rice, C. M., Ross, M. T., Searle, S. M., Sehra, H. K., Sheridan, E., Skuce, C. D., Smith, S., Smith, M., Spraggon, L., Squares, S. L., Steward, C. A., Sycamore, N., Tamlyn-Hall, G., Tester, J., Theaker, A. J., Thomas, D. W., Thorpe, A., Tracey, A., Tromans, A., Tubby, B., Wall, M., Wallis, J. M., West, A. P., White, S. S., Whitehead, S. L., Whittaker, H., Wild, A., Willey, D. J., Wilmer, T. E., Wood, J. M., Wray, P. W., Wyatt, J. C., Young, L., Younger, R. M., Bentley, D. R., Coulson, A., Durbin, R., Hubbard, T., Sulston, J. E., Dunham, I., Rogers, J. and Beck, S. (2003), 'The DNA sequence and analysis of human chromosome 6', *Nature* **425**(6960), 805–811.

Munoz, L. E., Lauber, K., Schiller, M., Manfredi, A. A. and Herrmann, M. (2010), 'The role of defective clearance of apoptotic cells in systemic autoimmunity', *Nat Rev Rheumatol* **6**(5), 280–289.

Muyldermans, S., Atarhouch, T., Saldanha, J., Barbosa, J. A. and Hamers, R. (1994), 'Sequence and structure of VH domain from naturally occurring camel heavy chain immunoglobulins lacking light chains', *Protein Eng.* **7**(9), 1129–1135.

Neuvirth, H., Raz, R. and Schreiber, G. (2004), 'ProMate: a structure based prediction program to identify the location of protein-protein binding sites', *J. Mol. Biol.* **338**(1), 181–199.

Nguyen, C., Wang, Y. and Nguyen, H. (2015), 'Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic', *Journal of Biomedical Science and Engineering* **6**, 551–560.

Nussinovitch, U. and Shoenfeld, Y. (2010), 'Anti-troponin autoantibodies and the cardiovascular system', *Heart* **96**(19), 1518–1524.

Nussinovitch, U. and Shoenfeld, Y. (2011), 'The Clinical and Diagnostic Significance of Anti-myosin Autoantibodies in Cardiac Disease', *Clin Rev Allergy Immunol* .

Oelke, K., Lu, Q., Richardson, D., Wu, A., Deng, C., Hanash, S. and Richardson, B. (2004), 'Overexpression of CD70 and overstimulation of IgG synthesis by lupus T cells and T cells treated with DNA methylation inhibitors', *Arthritis Rheum.* **50**(6), 1850–1860.

Oldham, P. D. (1962), 'A note on the analysis of repeated measurements of the same subjects', *J Chronic Dis* **15**, 969–977.

Pansri, P., Jaruseranee, N., Rangnoi, K., Kristensen, P. and Yamabhai, M. (2009), 'A compact phage display human scFv library for selection of antibodies to a wide variety of antigens', *BMC Biotechnol.* **9**, 6.

Pashiardis, A. (2015), Polymerase chain reaction (PCR) optimisation for the successful amplification of human immunoglobulin genes and the influence of the variable region on the expression yield and biophysical properties of igg1 antibody fab fragments, Master's thesis, King's College London.

Pedersen, J. T., Henry, A. H., Searle, S. J., Guild, B. C., Roguska, M. and Rees, A. R. (1994), 'Comparison of surface accessible residues in human and murine immunoglobulin Fv domains. Implication for humanization of murine antibodies', *J. Mol. Biol.* **235**(3), 959–973.

Perneger, T. V. (1998), 'What's wrong with Bonferroni adjustments', *BMJ* **316**(7139), 1236–1238.

Pettit, F. K., Bare, E., Tsai, A. and Bowie, J. U. (2007), 'HotPatch: a statistical approach to finding biologically relevant features on protein surfaces', *J. Mol. Biol.* **369**(3), 863–879.

Pizzi, E., Cortese, R. and Tramontano, A. (1995), 'Mapping epitopes on protein surfaces', *Biopolymers* **36**(5), 675–680.

Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P. and Saul, F. (1973), 'Three-dimensional structure of the Fab' fragment of a human immunoglobulin at 2,8-A resolution', *Proc. Natl. Acad. Sci. U.S.A.* **70**(12), 3305–3310.

Ponomarenko, J., Bui, H. H., Li, W., Fusseder, N., Bourne, P. E., Sette, A. and Peters, B. (2008), 'ElliPro: a new structure-based tool for the prediction of antibody epitopes', *BMC Bioinformatics* **9**, 514.

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

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

Porollo, A. and Meller, J. (2007), 'Prediction-based fingerprints of protein-protein interactions', *Proteins* **66**(3), 630–645.

Porter, C. T. and Martin, A. C. (2015), 'BiopLib and BiopTools–a C programming library and toolset for manipulating protein structure', *Bioinformatics* **31**(24), 4017–4019.

Porter, R. R. (1959), 'The hydrolysis of rabbit y-globulin and antibodies with crystalline papain', *Biochem. J.* **73**, 119–126.

Qi, T., Qiu, T., Zhang, Q., Tang, K., Fan, Y., Qiu, J., Wu, D., Zhang, W., Chen, Y., Gao, J., Zhu, R. and Cao, Z. (2014), 'SEPPA 2.0–more refined server to predict spatial epitope considering species of immune host and subcellular localization of protein antigen', *Nucleic Acids Res.* **42**(Web Server issue), 59–63.

Racanelli, V., Prete, M., Musaraj, G., Dammacco, F. and Perosa, F. (2011), 'Autoantibodies to intracellular antigens: generation and pathogenetic role', *Autoimmun Rev* **10**(8), 503–508.

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

Rajewsky, K., Forster, I. and Cumano, A. (1987), 'Evolutionary and somatic selection of the antibody repertoire in the mouse', *Science* **238**(4830), 1088–1094.

Raju, T. N., Edelman, G. M. and Porter, R. R. (1999), 'The Nobel chronicles. 1972: Gerald M Edelman (b 1929) and Rodney R Porter (1917-85)', *Lancet* **354**(9183), 1040.

Ren, J., Liu, Q., Ellis, J. and Li, J. (2014), 'Tertiary structure-based prediction of conformational B-cell epitopes through B factors', *Bioinformatics* **30**(12), i264–273.

Ren, J., Liu, Q., Ellis, J. and Li, J. (2015), 'Positive-unlabeled learning for the prediction of conformational B-cell epitopes', *BMC Bioinformatics* **16 Suppl 18**, S12.

Resende, D. M., Rezende, A. M., Oliveira, N. J., Batista, I. C., Correa-Oliveira, R., Reis, A. B. and Ruiz, J. C. (2012), 'An assessment on epitope prediction methods for protozoa genomes', *BMC Bioinformatics* **13**, 309.

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

Ritter, G., Cohen, L. S., Williams, C., Richards, E. C., Old, L. J. and Welt, S. (2001), 'Serological analysis of human anti-human antibody responses in colon cancer patients treated with repeated doses of humanized monoclonal antibody A33', *Cancer Res.* **61**(18), 6851–6859.

Rockberg, J., Lofblom, J., Hjelm, B., Uhlen, M. and Stahl, S. (2008), 'Epitope mapping of antibodies using bacterial surface display', *Nat. Methods* **5**(12), 1039–1045.

Rossi, G., Choi, T. K. and Nisonoff, A. (1969), ‘Crystals of fragment Fab’: preparation from pepsin digests of human IgG myeloma proteins’, *Nature* **223**(5208), 837–838.

Rothlisberger, D., Honegger, A. and Pluckthun, A. (2005), ‘Domain interactions in the Fab fragment: a comparative evaluation of the single-chain Fv and Fab format engineered with variable domains of different stability’, *J. Mol. Biol.* **347**(4), 773–789.

Rubinstein, N. D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J. M. and Pupko, T. (2008), ‘Computational characterization of B-cell epitopes’, *Mol. Immunol.* **45**(12), 3477–3489.

Rubinstein, N. D., Mayrose, I. and Pupko, T. (2009), ‘A machine-learning approach for predicting B-cell epitopes’, *Mol. Immunol.* **46**(5), 840–847.

Ruxton, G. D. (2006), ‘The unequal variance t-test is an underused alternative to Student’s t-test and the Mann-Whitney u test’, *Behavioral Ecology* **17**(4), 688–690.

Samaranayake, H., Wirth, T., Schenkwein, D., Raty, J. K. and Yla-Herttuala, S. (2009), ‘Challenges in monoclonal antibody-based therapies’, *Ann. Med.* **41**(5), 322–331.

Scarabelli, G., Morra, G. and Colombo, G. (2010), ‘Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping’, *Biophys. J.* **98**(9), 1966–1975.

Schwartz, R. H., Mueller, D. L., Jenkins, M. K. and Quill, H. (1989), ‘T-cell clonal anergy’, *Cold Spring Harb. Symp. Quant. Biol.* **54 Pt 2**, 605–610.

Sharma, V. K., Patapoff, T. W., Kabakoff, B., Pai, S., Hilario, E., Zhang, B., Li, C., Borisov, O., Kelley, R. F., Chorny, I., Zhou, J. Z., Dill, K. A. and Swartz, T. E. (2014), ‘In silico selection of therapeutic antibodies for development: viscosity, clearance, and chemical stability’, *Proc. Natl. Acad. Sci. U.S.A.* **111**(52), 18601–18606.

Shaw, B. F., Schneider, G. F., Bilgicer, B., Kaufman, G. K., Neveu, J. M., Lane, W. S., Whitelegge, J. P. and Whitesides, G. M. (2008), ‘Lysine acetylation can generate highly charged enzymes with increased resistance toward irreversible inactivation’, *Protein Sci.* **17**(8), 1446–1455.

Shire, S. J., Shahrokh, Z. and Liu, J. (2004), ‘Challenges in the development of high protein concentration formulations’, *J Pharm Sci* **93**(6), 1390–1402.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D. and Higgins, D. G. (2011), ‘Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega’, *Mol. Syst. Biol.* **7**, 539.

Smith, K., Garman, L., Wrammert, J., Zheng, N. Y., Capra, J. D., Ahmed, R. and Wilson, P. C. (2009), 'Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen', *Nat Protoc* **4**(3), 372–384.

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

Soltis, R. D. and Hasz, D. (1982), 'Studies on the nature of intermolecular bonding in antigen-antibody complexes', *Immunology* **46**(1), 175–181.

Spouge, J. L., Guy, H. R., Cornette, J. L., Margalit, H., Cease, K., Berzofsky, J. A. and DeLisi, C. (1987), 'Strong conformational propensities enhance T cell antigenicity', *J. Immunol.* **138**(1), 204–212.

Stanfield, R. L., Dooley, H., Flajnik, M. F. and Wilson, I. A. (2004), 'Crystal structure of a shark single-domain antibody V region in complex with lysozyme', *Science* **305**(5691), 1770–1773.

Steipe, B. (2004), 'Consensus-based engineering of protein stability: from intrabodies to thermostable enzymes', *Meth. Enzymol.* **388**, 176–186.

Steipe, B., Schiller, B., Pluckthun, A. and Steinbacher, S. (1994), 'Sequence statistics reliably predict stabilizing mutations in a protein domain', *J. Mol. Biol.* **240**(3), 188–192.

Stern, L. J., Brown, J. H., Jardetzky, T. S., Gorga, J. C., Urban, R. G., Strominger, J. L. and Wiley, D. C. (1994), 'Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide', *Nature* **368**(6468), 215–221.

Stockwin, L. H. and Holmes, S. (2003), 'Antibodies as therapeutic agents: vive la renaissance!', *Expert Opin Biol Ther* **3**, 1133–1152.

Sun, J., Xu, T., Wang, S., Li, G., Wu, D. and Cao, Z. (2011), 'Does difference exist between epitope and non-epitope residues? analysis of the physicochemical and structural properties on conformational epitopes from b-cell protein antigens.', *Immunome Res.* **7**, 1–11.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. and Feuston, B. P. (2003), 'Random forest: a classification and regression tool for compound classification and QSAR modeling', *J Chem Inf Comput Sci* **43**(6), 1947–1958.

Tan, P., Mitchell, D. A., Buss, T. N., Holmes, M. A., Anasetti, C. and Foote, J. (2002), '"Superhumanized" antibodies: reduction of immunogenic potential by complementarity-determining region grafting with human germline sequences: application to an anti-CD28', *J. Immunol.* **169**(2), 1119–1125.

Taylor, W. R. (1986), 'The classification of amino acid conservation', *J. Theor. Biol.* **119**(2), 205–218.

Teng, G. and Papavasiliou, F. N. (2007), 'Immunoglobulin somatic hypermutation', *Annu. Rev. Genet.* **41**, 107–120.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994), 'CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice', *Nucleic Acids Res.* **22**(22), 4673–4680.

Thornton, J. M., Edwards, M. S., Taylor, W. R. and Barlow, D. J. (1986), 'Location of 'continuous' antigenic determinants in the protruding regions of proteins', *EMBO J.* **5**(2), 409–413.

Tiller, T., Schuster, I., Deppe, D., Siegers, K., Strohner, R., Herrmann, T., Berenguer, M., Poujol, D., Stehle, J., Stark, Y., Hessling, M., Daubert, D., Felderer, K., Kaden, S., Kolln, J., Enzelberger, M. and Urlinger, S. (2013), 'A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties', *MAbs* **5**(3), 445–470.

Tsurushita, N., Hinton, P. R. and Kumar, S. (2005), 'Design of humanized antibodies: from anti-Tac to Zenapax', *Methods* **36**(1), 69–83.

Van Epps, H. L. and Heidelberger, M. (2006), 'Michael Heidelberger and the demystification of antibodies', *J. Exp. Med.* **203**(1), 5.

Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A. and Peters, B. (2015), 'The immune epitope database (IEDB) 3.0', *Nucleic Acids Res.* **43**(Database issue), D405–412.

Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A. and Peters, B. (2010), 'The immune epitope database 2.0', *Nucleic Acids Res.* **38**(Database issue), D854–862.

Walter, G. (1986), 'Production and use of antibodies against synthetic peptides', *J Immunol Methods* **88**, 149–161.

Wang, F., Sen, S., Zhang, Y., Ahmad, I., Zhu, X., Wilson, I. A., Smider, V. V., Magliery, T. J. and Schultz, P. G. (2013), 'Somatic hypermutation maintains antibody thermodynamic stability during affinity maturation', *Proc. Natl. Acad. Sci. U.S.A.* **110**(11), 4261–4266.

Wang, N., Smith, W. F., Miller, B. R., Aivazian, D., Lugovskoy, A. A., Reff, M. E., Glaser, S. M., Croner, L. J. and Demarest, S. J. (2009), 'Conserved amino acid networks involved in antibody variable domain interactions', *Proteins* **76**(1), 99–114.

Wang, P., Sidney, J., Dow, C., Mothe, B., Sette, A. and Peters, B. (2008), 'A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach', *PLoS Comput. Biol.* **4**(4), e1000048.

Wardemann, H., Yurasov, S., Schaefer, A., Young, J. W., Meffre, E. and Nussenzweig, M. C. (2003), 'Predominant autoantibody production by early human B cell precursors', *Science* **301**(5638), 1374–1377.

Welch, B. L. (1947), 'The generalisation of student's problems when several different population variances are involved', *Biometrika* **34**(1-2), 28–35.

Wu, S. J., Luo, J., O'Neil, K. T., Kang, J., Lacy, E. R., Canziani, G., Baker, A., Huang, M., Tang, Q. M., Raju, T. S., Jacobs, S. A., Teplyakov, A., Gilliland, G. L. and Feng, Y. (2010), 'Structure-based engineering of a monoclonal antibody for improved solubility', *Protein Eng. Des. Sel.* **23**(8), 643–651.

Wu, T. T. and Kabat, E. A. (1970), 'An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity', *J. Exp. Med.* **132**(2), 211–250.

Xiang, J., Sha, Y., Jia, Z., Prasad, L. and Delbaere, L. T. (1995), 'Framework residues 71 and 93 of the chimeric B72.3 antibody are major determinants of the conformation of heavy-chain hypervariable loops', *J. Mol. Biol.* **253**(3), 385–390.

Yates., F. (1934), 'Contingency tables involving small numbers and the chi-square test', *Supplement to the Journal of the Royal Statistical Society* **1**(2), 217–235.

Ye, J., Ma, N., Madden, T. L. and Ostell, J. M. (2013), 'IgBLAST: an immunoglobulin variable domain sequence analysis tool', *Nucleic Acids Res.* **41**(Web Server issue), 34–40.

Yu, K., Petrovsky, N., Schonbach, C., Koh, J. Y. and Brusic, V. (2002), 'Methods for prediction of peptide binding to MHC molecules: a comparative study', *Mol. Med.* **8**(3), 137–148.

Zhao, L. and Li, J. (2010), 'Mining for the antibody-antigen interacting associations that predict the B cell epitopes', *BMC Struct. Biol.* **10 Suppl 1**, S6.

Zhao, L., Wong, L., Lu, L., Hoi, S. C. and Li, J. (2012), 'B-cell epitope prediction through a graph model', *BMC Bioinformatics* **13 Suppl 17**, S20.

Zhao, S. and Lu, J. (2011), 'A bioinformatics pipeline to build a knowledge database for in silico antibody engineering', *Mol. Immunol.* **48**(8), 1019–1026.

Zouali, M. (2001), *Antibodies*, John Wiley and Sons.