PHD

Molecular modelling of immunoglobulin folds

Searle, Stephen M. J.

Award date:
1997

Awarding institution:
University of Bath

Molecular Modelling of Immunoglobulin Folds

Submitted by Stephen M J Searle
for the degree of
Doctor of Philosophy
Department of Biology and Biochemistry
University of Bath

November 24, 1997

Abstract

Molecular Modelling of Immunoglobulin Folds

Stephen M. J. Searle

Ph.D Thesis
November 1997

The crystallisation of the T Cell $\alpha\beta$ Receptor (TCR) proved difficult to accomplish. Many TCR clones have been sequenced since the receptor was first identified in 1984. The main aim of this thesis is the prediction of TCR variable region structures using homology modelling techniques and gaining some insight into their interaction with MHC-peptide complexes.

The modelling was approached in a series of steps. TCR sequences were compared to antibody sequences. These two classes of structures are both members of the immunoglobulin superfamily. Evidence is presented that both TCR α and β chains show greater homology to antibody κ light chains than to antibody heavy chains, particularly in the Vα-Vβ interface residues. A similarity index for comparing one set of sequences to two others was devised to aid in the comparison of

TCRs and antibody sequences. An algorithm had previously been developed in the laboratory for modelling antibody variable domains. Certain modifications were made to this algorithm to allow the use of an antibody light chain dimer framework and also a V_LV_H hybrid framework for modelling the CDR $\beta2$ region which appeared more similar to the antibody heavy chain. Before models of the TCR were made the modifications to the algorithm were tested on the non-antibody immunoglobulin CD4, which has a known crystal structure, so that the accuracy of the model could be judged. The algorithm was then used to model TCRs. First a set of TCR sequences were modelled for which data relating sequence and function were available. This provided a means of testing the ability of the algorithm to generate models which were consistent with these known facts. Then a TCR was modelled and used to predict the identity of a peptide contacting residue in a TCR CDR $\alpha3$ loop which was then confirmed to be important for binding. The experimental work was done by L. Wedderburn at the Imperial Cancer Research Fund Laboratories (London, UK).

Also during the course of this work many antibody models were produced. It became apparent that the CAMAL modelling algorithm did not always produce accurate results especially for long CDR H3 loops. Adaptations to the algorithm CAMAL which increase its accuracy in modelling antibody CDRs are also presented.

A brief description of the work done on a new method for humanising murine antibodies (resurfacing) is also presented. This method generally requires fewer changes to be made to the mouse sequence than CDR-grafting and is therefore less likely to adversely affect the binding affinity of the antibody.

Some of the work described in this thesis has been published elsewhere:

Pedersen, J.T., Searle, S.J., Henry, A.H., and Rees, A.R. Antibody modelling: Beyond homology. *Immunomethods*, 1:126–136, 1992.

Rees, A.R., Staunton, D., Webster, D.M., Searle, S.J., Henry, A.H., and Pedersen, J.T. Antibody design: Beyond the natural limits. *Trends Biotech.*, 12:199–206, 1994.

Roguska, M.A., Pedersen, J.T., Keddy, C.A., Henry, A.H., Searle, S.J., Lambert, J.M., Goldmacher, V.S., Blattler, W.A., Rees, A.R., and Guild, B.C. Humanization of murine monoclonal antibodies through variable domain resurfacing. *Proc. Natl. Acad. Sci. USA*, 91:969–973, 1994.

Pedersen, J.T., Henry, A.H., Searle, S.J., Guild, B.C., Roguska, M., and Rees, A.R. Comparison of surface accessible residues in human and murine immunoglobulin F_v domains. Implication for humanization of murine antibodies. *J. Mol. Biol.*, 235:959–973, 1994.

Rees, A.R., Searle, S.J., Pedersen, J.T., and Webster, D.M. Antibody Structure: X-ray cyrstallography and molecular modelling. In van Oss, C.J. and van Regenmortel, M.H.V., editors, *Immunochemistry*, pages 615–652. Marcel Dekker, 1994.

Wedderburn, L.R., Searle, S.J.M., Rees, A.R., Lamb, J.R., and Owen, M.J. Mapping T cell recognition: The identification of a T cell receptor residue critical to the specific interaction with an influenza hemagglutinin peptide. *Eur. J. Immun.*, 25:1654–1662, 1995.

Rees, A.R., Pedersen, J.T., Searle, S.J., Henry, A.H., and Webster, D.M. Antibody Structure and Function. In Borrebaeck, K., editor, *Antibody Engineering: A Manual*, pages 1–59. Oxford University Press, 1995.

Roguska, M.A., Pedersen, J.T., Henry, A.H., Searle, S.M.J., Roja, C.M, Avery, B., Hoffee, M., Cook, S., Lambert, J.M., Blattler, W.A., Rees, A.R., and Guild, B.C. A comparison of two murine monoclonal antibodies humanized by CDR-grafting and variable domain resurfacing. *Protein Eng.*, 9:895–904, 1996.

Searle, S.J., Henry, A.H., Pedersen, J.T., and Rees, A.R. Antibody Combining Sites. In Sternberg, M.J.E., editor, *Protein Structure Prediction: A Practical Approach*, pages 141–172. Oxford University Press, 1996.

Acknowledgements

Firstly, I would like to thank Professor Anthony R. Rees, my supervisor, for three interesting years of the PhD and his understanding during the difficult times since. I am grateful to Jan Pedersen for guiding me into the subject in my first year, and to Andrew Henry. Together the three of us made a good team.

Thanks, also, to the other members of the Rees group, particularly David Webster ,and to David Osguthorpe's molecular graphics group.

David Osguthorpe, a true expert in his field and in Unix, deserves a special thank you for allowing the close interaction between Prof. Rees's modelling group and his own, and to him and his wife Pnina for many interesting lunchtime discussions.

None of this work would have been possible without financial support from SERC (the funding agency for the studentship), and from Immunogen (who provided six months money for writing up). Also I would not have been able to survive financially since, without the support of my parents.

Finally a very special thankyou goes to Alison for supporting me through this thesis, and for putting up with me through the many difficult times during the

write-up. Without her this thesis would not have been completed.

Abbreviations

Ab Antibody.

AbM An antibody modelling program by Oxford Molecular Ltd.

APC Antigen presenting cell.

CAMAL Combined algorithm for modelling antibody loops.

C region Constant region.

CD Cluster determining.

CDR Complementarity determining region.

D gene segment Diversity gene segment.

DAG Diacyl glycerol.

DNA Dioxyribonucleic acid.

ER Endoplasmic reticullum.

GM-CSF Granulocyte-macrophage colony stimulating factor.

HIV Human immunodeficiency virus.

HLA Human leucocyte antigen.

ICAM Intercellular adhesion molecule.

IFN Interferon.

Ig Immunoglobulin.

IL Interleukin.

IP3 Inositol (1,4,5) triphosphate.

IUIS International Union of Immunological Societies

J gene segment Joining gene segment.

LCMV Lymphocytic Choriomeningitis virus

LFA Lymphocyte function associated.

MAC Membrane attack complex.

MCC Moth cytochrome c.

MHC Major histocompatibility complex.

mRNA Messenger ribonucleic acid.

PDB Protein Data Bank.

PIP2 Phospatidyl inositol 4,5 bisphosphate.

RAG Recombination activating gene.

SCR Structurally conserved region.

SH src homology.

TAP Transporter in antigen processing.

TCR T cell antigen receptor.

TNF Tumour necrosis factor.

V gene segment Variable gene segment.

VR Variable region.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aim of the project was the molecular modelling of **T cell receptor** (TCR) α/β dimers, as well as their interactions with their ligands, the **major histocompatibility complex** (MHC) -peptide complexes. At the time (1990) no structural data existed for any TCR because of difficulties in solubilising the protein for crystallography. It is only recently that the structure of a complete TCR V region was solved [1,2], although the subunits had been crystallised separately in 1994 [3,4]. The modelling was an extension to previous work by the Bath group on antibody modelling [5–7].

In this introduction the immune system is briefly described and the role of the T cell and its receptor within it is explained. In Chapter 2 the sequences of TCRs and antibodies are compared. Chapter 3 describes the antibody molecular modelling method and adaptations that were made to it in an attempt to improve accuracy as well as to allow it to be used for the modelling of TCRs. In Chapter

4 an examination is made of the published work on the TCRs, and the modelling studies on them are described. Chapter 5 describes the application of antibody modelling to humanisation studies. In the final chapter a discussion is made of the value and accuracy of molecular modelling in this system. Also the future of modelling for TCRs is discussed.

1.1 The Immune System

There are many potential pathogenic microorganisms in the environment, including bacteria, viruses and protozoa. Multicellular organisms have developed increasingly more advanced mechanisms through evolution to protect them from infection by these entities.

As well as the physical barriers to invasion such as the skin and mucosa, the circulation contains many different cells and molecules involved in the elimination of foreign entities. Some are always present and do not vary in concentration in response to an infection. These form the **natural immunity** (or **innate immunity**) of the body. Others recognise very specific foreign molecules, which are termed **antigens**, and increase in concentration to enable the rapid elimination of the foreign entity. These cells and molecules form the **specific immunity** of the organism [8].

On entry of a foreign organism, the initial response is by the natural immune mechanisms, before the specific immune response can be activated. These mechanisms include phagocytosis by phagocytes and macrophages and the activation of the **complement system** by the **alternative pathway**. The complement system is a group of proteins which form a cascade proteolytic pathway to produce a protein complex called the **Membrane Attack Complex** (MAC) which creates holes in the cell membrane of invading microbes causing their death (see figure 1.1). Complement components can also mark cells for **opsonisation** by phagocytes and macrophages. Natural Killer (NK) cells are involved in immune surveillance, screening for tumour or virus infected cells [9]. They cause cell death of these abnormal cells by creating pores in their cell membranes.

The specific immune system consists of two main parts. The first is **humoral immunity**. This can be transferred between individuals in the humors (serum and plasma), that is without the transfer of cells. The second arm is the **cell mediated immune response** (cellular immunity) which can only be transferred between individuals by the transfer of blood leukocytes [11].

The specific immune response has several characteristics. The first is **specificity** which is the ability to identify uniquely a particular protein. The second is **memory** which is the capability of the organism to retain the ability to fight the same infection again without having to redevelop the response. This **secondary response** has massively increased effectiveness and often the micro-organism is eliminated before symptoms develop. A third is **self non-self discrimination** which is the ability to discriminate self molecules from non self foreign molecules. These can include foreign micro-organisms, and also now tissue grafts or transplants. Failure of self recognition leads to **autoimmunity** [12, 13]. In the case of tissue grafts self non-self discrimination presents a problem to successful

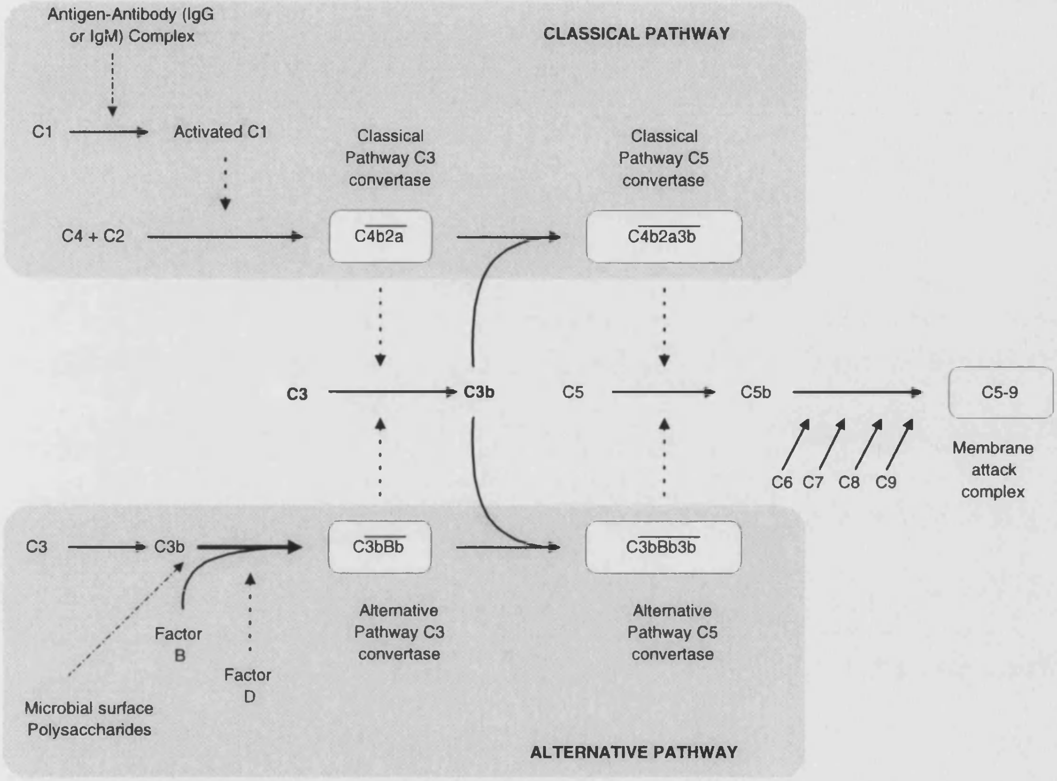

Antigen-Antibody (IgG or IgM) Complex
CLASSICAL PATHWAY
C1
Activated C1
Classical Pathway C3 convertase
Classical Pathway C5 convertase
C4 + C2
C4b2a
C4b2a3b
C3
C3b
C5
C5b
C5-9
Membrane attack complex
C6 C7 C8 C9
C3
C3b
C3bBb
C3bBb3b
Factor B
Factor D
Alternative Pathway C3 convertase
Alternative Pathway C5 convertase
Microbial surface Polysaccharides
ALTERNATIVE PATHWAY

Figure 1.1: The complement system. There are two different mechanisms by which the complement system can be activated; the Classical and the Alternate Pathways. However the two pathways both eventually lead to the same final result, the formation of the membrane attack complex (MAC). The classical pathway is activated by multivalent antibody antigen complexes, while the alternate pathway can be activated by several molecules including microbial cell surface polysaccharides. Both pathways lead to the production of C3 convertases although these are made up of different components in the two pathways. C3b is a component of both C5 convertases which proteolyse C5 into C5a and C5b. C5b associates loosely with the cell membrane. This association is strengthened by the formation of the C5bC6C7 complex which is highly lipophilic. C8 and multiple copies of C9 then bind to the C5bC6C7 complex resulting in the pore forming MAC. (Dashed line = proteolytic activity, Line above text = proteolytic complex, Dashed and dotted line = activating agent) (adapted from [10]).

grafting. A fourth is **self regulation**, which is the capacity to initiate the correct response at the right time and terminate the response once the infective agent has been eliminated. Failure of self regulation can lead to **hypersensitivity** or **allergy** [14]. The last characteristic is **diversity**, which ensures that an effective response can be mounted against almost any entity [15].

The specific immune system involves two types of cells from the circulatory system, **B lymphocytes** (B cells) and **T lymphocytes** (T cells). These have, on their surface, clonally distributed receptors which can recognise foreign molecules. B lymphocytes produce antibodies that act as specific receptors for intact foreign macromolecules which can be free in solution or cell bound. T lymphocytes use a different receptor, the T-cell receptor which is more restricted in specificity, only recognising antigen when presented on another cell. In fact only fragments of the original antigen can be recognised when presented along with a **major histocompatibility complex** (MHC) antigen on the surface of another cell [16, 17].

1.1.1 The Role of B Lymphocytes

B cells have antibodies expressed on their cell surface. These can bind antigen which can be internalised and processed for presentation by the MHC. This makes the B cell a very good **antigen presenting cell** (APC) as the antigen is specifically taken up. T cell help causes the proliferation of the B cell and differentiation into plasma cells which can secrete soluble antibody. The membrane bound antibodies on B cells can also act as the recognition component of a protein signalling complex known as the **B Cell Receptor** (BCR) [18]. This can stimulate activation of the B cell when antigen is bound if accompanied by T cell help.

Soluble antibody can bind to the antigen in solution and cause the activation of complement via the **classical pathway**. Alternatively it can bind to Fc receptors on phagocytes and hence target them to specific antigens in a process known as **opsonization.**

1.1.2 The Role of T Lymphocytes

T lymphocytes are primarily involved in the elimination of virally infected cells, although they also have an important regulatory capacity in the B-cell antibody response to micro-organisms. T lymphocytes are subtyped according to their function. Those which bind to cells and cause them to die are called **cytotoxic T cells,** while those which bind to specific types of cells (B cells and other lymphocytes) termed APCs and enhance B cell proliferation and immunoglobulin production by **lymphokine** secretion are called **T helper cells.** There are also T cells which, on binding to APCs, suppress the immune response to a particular antigen. These are called **T suppressor cells.**

T cells develop from **haemopoietic stem cells** in the **bone marrow**. These enter the thymus where they complete their development. The T cells are divided into two groups distinguished by whether they express α/β TCRs or γ/δ TCRs on their cell surface. The α/β TCR expressing T cells constitute about 95% of the T cells in the body.

T helper and suppressor cells, as well as recognising antigen, must also be able to recognise the cell type as an APC. A set of **polymorphic** molecules, the MHC antigens, are involved in presentation of the antigen. Distributed almost ubiquitously among the tissues of the body are the **MHC class I** molecules which mark cells as “self”. “Non self” cells entering the body such as those from transplants are recognised as foreign because they generally have different MHC class I molecules expressed on their cell surface. This is known as an **allotypic** response. **MHC class II** molecules have a much more restricted distribution, only being found on APCs. The structures of several MHC class I molecules have been determined by X-ray crystallography [19–21]. The structure comprises a glycosylated polypeptide chain of 45kD in non-covalent association with β_2 **microglobulin**, a 12kD polypeptide which is also found non-associated in serum. The 45kD chain is divided into five domains: three extracellular domains, a transmembrane region and a cytoplasmic domain. The two N-terminal domains of this chain form the region which is involved in antigen presentation to T cells. This consists of a groove with α-helix sides and a β sheet base which binds a peptide from the processed antigen. Figure 1.2 shows an MHC class I molecule with bound peptide. This is the region of the molecule where the polymorphic variations are centred.

The tertiary structure of MHC class II is similar to the class I molecules with the antigen presented in a similar way. However the primary structure is different, with two chains of approximately equal length each forming half of the peptide binding site. Several MHC class II proteins have been crystallised [22, 23] and found to form dimers. There is now some evidence that the dimeric state may be important for their signalling function [24].

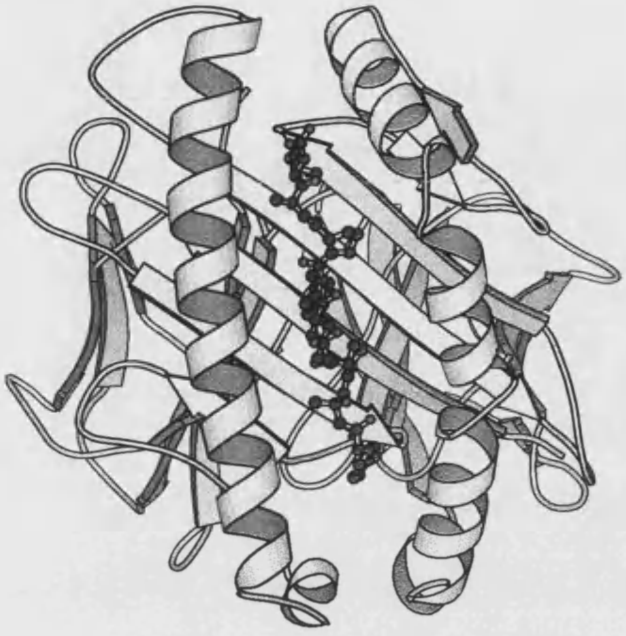


Figure 1.2: A schematic diagram of the structure of the MHC class I molecule H-2K^b in complex with a peptide from Sendai virus nucleoprotein (residues 324-332) [25]. Diagram created with molscript [26].

Pockets have been identified in the binding clefts of both types of MHC which are specific for certain amino acid types at fixed positions in the peptides which bind.

The breakdown of the antigens and association of peptide occurs within the APC [27]. Proteins produced within the cell, such as those of viral origin are presented in association with MHC class I molecules whereas those which have been endocytosed into the cell are associated with MHC class II on APCs. This allows the different types of cells to be distinguished by the subsets of T cells. The cytotoxic T cells mainly recognise cells with MHC class I. They have on their surface a receptor, CD8, which binds to a non polymorphic region of the MHC class I molecule. The T helper cells on the other hand are mainly class II restricted and have a receptor, CD4, which binds to the MHC class II molecule. Although CD4 and CD8 are involved in determining which cells are bound, these receptors bind to non-polymorphic regions on the MHC molecules [28]. To give specificity for a particular MHC and peptide a receptor is needed which binds to both the polymorphic regions on the MHC and to the bound peptide. This is the function of the TCR.

1.1.3 γ/δ TCR expressing T Cells

During the early sequencing experiments to find the α chain a third TCR chain type, the γ chain, was discovered [29]. The role of the γ chain remained a mystery until the δ chain was identified and it was discovered that there were two distinct types of TCR (α/β and γ/δ) [30]. TCR γ/δ expressing T cells form a distinct lineage from α/β TCR expressing T cells. The γ and δ TCR sequences are similar to the α and β and rearrange in a similar manner. However the function of γ/δ T cells is uncertain [31]. They do respond to stimuli in similar ways to α/β T cells, for instance by secreting lymphokines or producing cytotoxic responses. However they do not, in general, seem to bind to classical MHC molecules. There is some evidence they may bind to **non classical MHC** possibly without bound peptide [32]. Also γ/δ cells often do not express either CD4 or CD8 (**double negative** cells) , but the TCR chains do appear to be associated with the CD3 signal transduction complex .

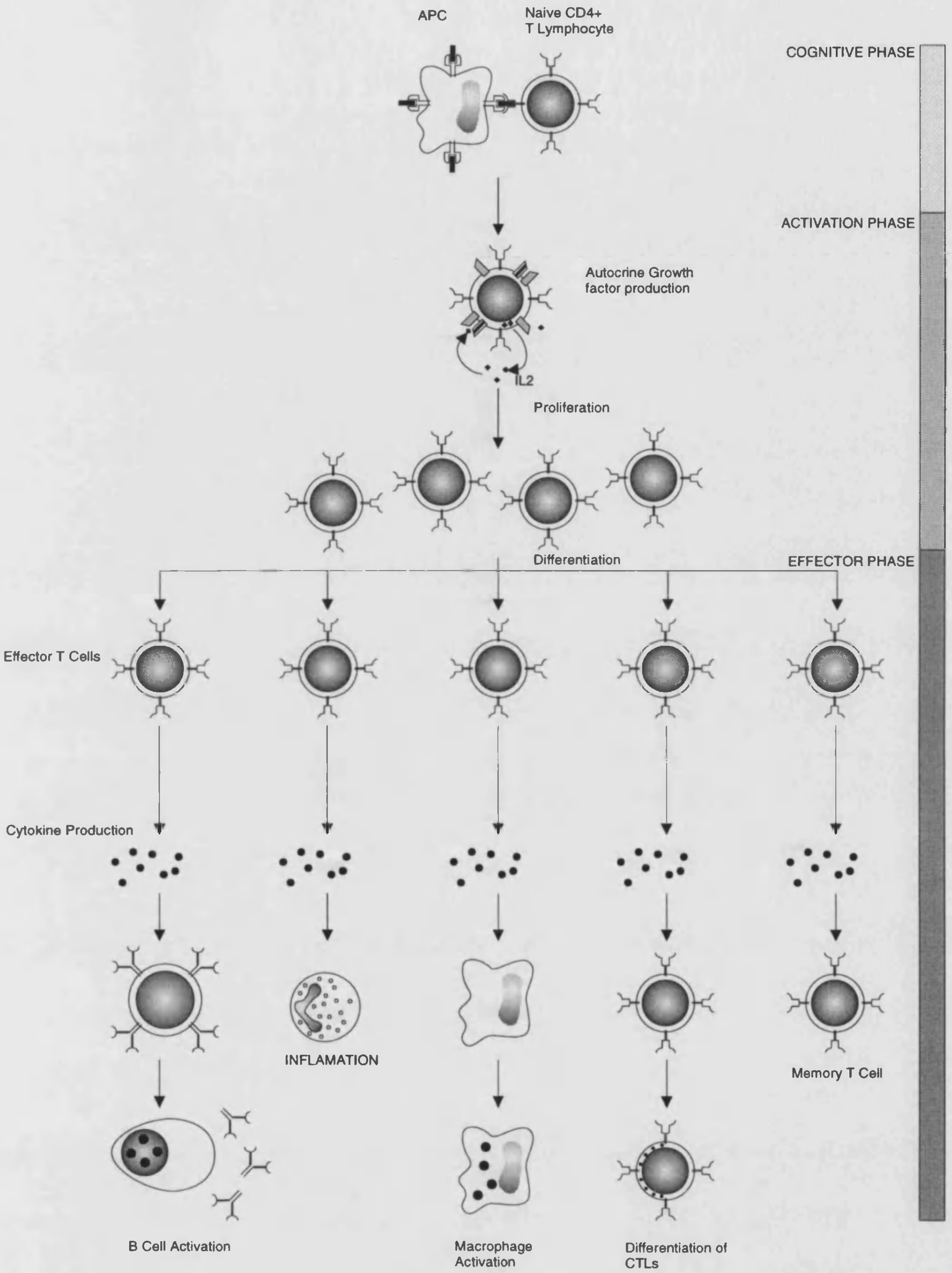


Figure 1.3: The response to antigen in CD4^{+} T cells is divided into three stages; the cognitive, activation and effector phases. The cognitive phase is the stage at which recognition of MHC bound peptide on an APC occurs. Signal transduction through the CD3 complex activates signalling pathways within the T cell. This leads to stimulation of interleukin-2 growth factor secretion by the T cell that binds to receptors on the cell causing proliferation and differentiation of the T cell clone. In the effector phase these cells bring about responses in other cells of the immune system by the release of cytokines. Adapted from [10].

The γ/δ expressing cells may be involved in responses to stress, possibly recognising **heat shock proteins** [33]. It has been suggested that this might produce an early non-specific response to micro-organism infection before the TCR α/β response develops [34]. Another possibility is that they may be involved in immune surveillance such as the elimination of tumour cells. In the epithelia, there are populations of γ/δ T cells expressing a single TCR. The role of these receptors is not certain [31].

1.1.4 Mechanism of Action of T Cells

The first stage in T cell activation is the cell cell contact between the T cell and the APC or target cell which initiates signalling [35]. The TCR is probably not the main protein responsible for the initial binding phase, this being performed by the monomorphic receptors on the cell surface such as CD2 which binds LFA3 on the APC and ICAM1 which binds LFA1 on the APC or *vice versa.*

The signal transduction pathway in T cells is shown in figure 1.5. The binding of MHC-Ag induces a signal which is transduced by the CD3 protein complex (see figure 1.4) [36]. CD4 has on its C terminus a binding site for a phosphorylation enzyme $p56^{lck}$ which phosphorylates several **src homology 2** (SH2) sites on the C terminal domains of CD3. An SH2 site on CD3 ζ chain binds phospholipase C (PI-PLCγ1) which catalyses the conversion of **phosphatidyl inositol 4,5 bisphosphate** (PIP2) to **inositol (1,4,5) triphospate** (IP3) and **diacylglycerol** (DAG) . IP3 stimulates the release of calcium from intracellular vesicles and, possibly, also the uptake of calcium from the exterior and the removal of potassium from

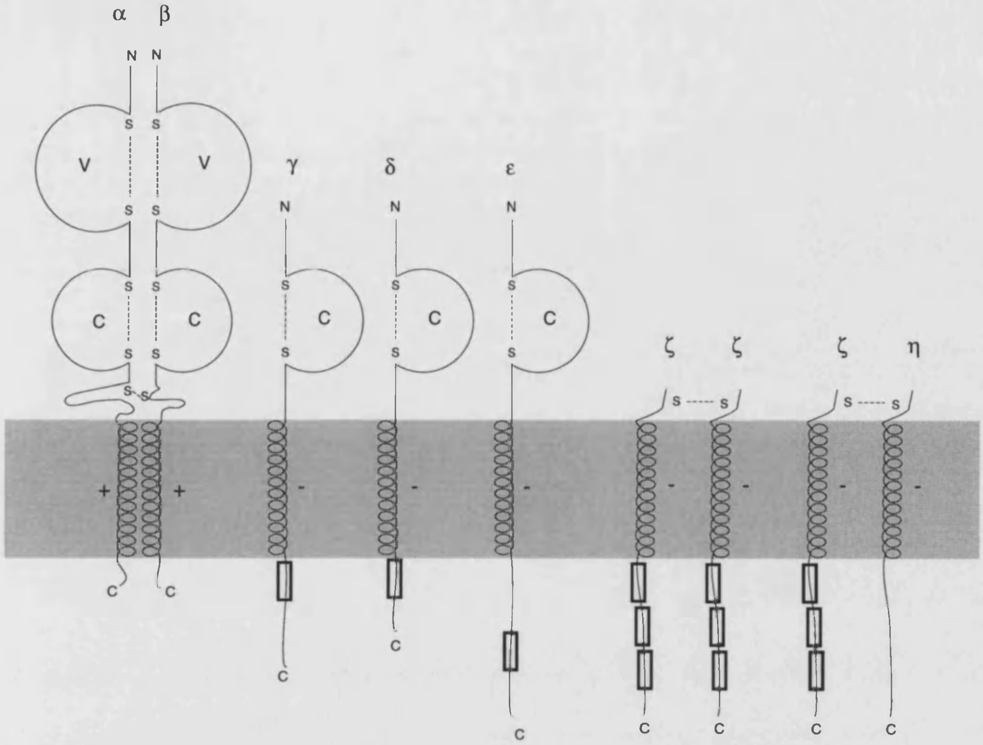


Figure 1.4: The TCR CD3 protein complex. The TCR α/β and γ/δ chain pairs both form a complex with the CD3 complex. CD3 acts as a signal transduction complex detecting binding to the recognition TCR component and producing intra-cellular responses. CD3 consists of a γ, δ and ϵ chains noncovalently associated with the TCR heterodimer. There are conserved complementary charged residues in the presumed membrane spanning regions of the chains which may be involved in maintaining the complex. There is either a ζ chain homodimer or a ζ chain η chain heterodimer associated with CD3 and the TCR.

the cell. DAG causes the activation of protein kinase C which catalyses various enzyme phosphorylations leading to gene activation and new protein production.

The genes which are activated depend on the type of T cell. In mouse the T helper cells can be divided into two groups based on the cytokines they produce. These are TH1 and TH2. Table 1.1 shows the profiles of cytokines produced by TH1, TH2 and cytotoxic T cells. Within the two T helper cell types IL-2, IFN$_\gamma$ and TNFβ are produced exclusively by TH1 while IL-4, IL-5, IL-6 and IL-10 are

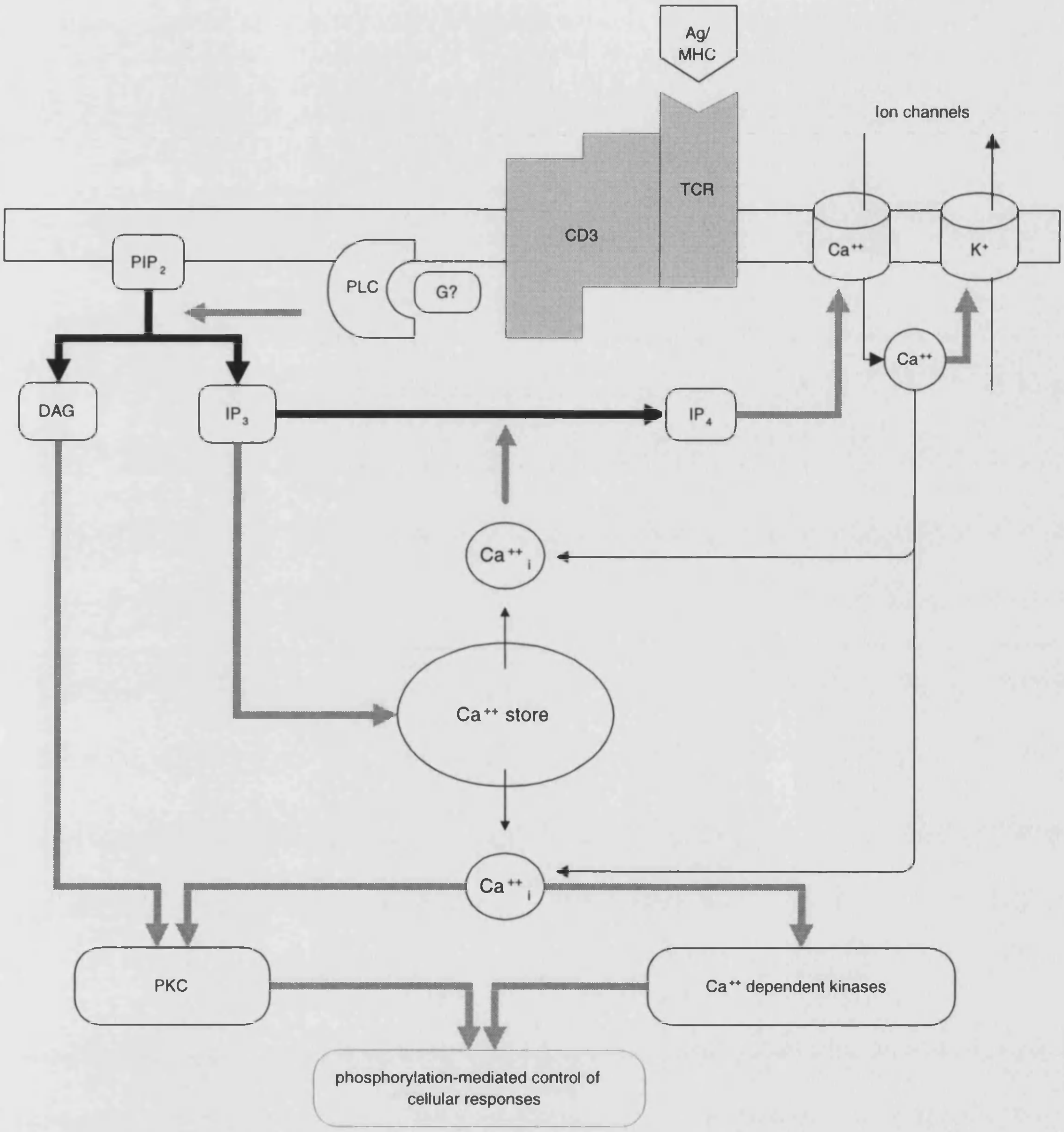


Figure 1.5: The signal transduction pathway in T cells. The binding to MHC with antigenic peptide (Ag) to the TCR induces signalling pathways within the T Cell. These are described in detail in the text (adapted from [37]).

produced only by TH2 cells. These differences lead to different functions for the two types of cells. They differ in the isotypes of antibodies with which they can give T cell help. TH2 provide help for IgM, IgG3, IgA and particularly IgE. The production of IgE requires IL-4. TH1 can provide help for IgM, IgG3, IgG1 and IgG2a. IgG2a requires IFN$_\gamma$ for its production. Parasitic worms elicit responses which are mainly TH2 based leading to the production of IgE, while responses to some bacterial and most virus infections are TH1 based leading to the production of IgG2a. IFN$_\gamma$ is also involved in macrophage activation, and so this is a role of TH1 cells. IL-3 and IL-4 are involved in mast cell growth and IL-5 in eosinophil maturation. These are therefore functions of TH2 cells.

In humans, cells with similar cytokine expression profiles to mouse TH1 and TH2 cells do occur. However many human T cells show another pattern of expression and are classed as TH0 cells.

IL-2 induces proliferation of cytotoxic T cells. A close association between T cell and target cell is required before killing can occur. There may be more than one mode of killing. One mechanism is pore formation, produced by the secretion of perforins into the intercellular space which form polyperforin multimers in the target cell membrane, puncturing it. TNFα and TNFβ can also be secreted and these cause **apoptosis** (programmed cell death) of the target cell. **Proteoglycan** (condroitin-5-sulphate) may help to stabilise perforins and protect the T cell from their pore forming action during an attack on a target cell. Calcium is also required for perforin polymerisation.

Cytokine	CTL	TH1	TH2
IFN$_{\gamma}$	++	++	-
IL-2	+/-	++	-
TNF$_{\beta}$	+	++	-
GM-CSF	++	++	+
TNF$_{\alpha}$	+	++	+
IL-3	+	++	++
IL-4	-	-	++
IL-5	-	-	++
IL-6	-	-	++
IL-10	-	-	++

Table 1.1: This table indicates the cytokines which are produced by the three mouse T cell types TH1, TH2 and CTL [37].

1.1.5 Antigen Processing and Presentation

Endogenously produced proteins are proteolysed in the cytoplasm by a **proteasome** encoded in the MHC, or in **endoplasmic reticulum** (ER) . Peptides produced in the cytoplasm are transported across the ER membrane by the **transporter in antigen processing** (TAP) before binding to MHC class I molecules. Only peptide bound MHC is expressed at the cell surface (see figure 1.6).

Exogenously produced proteins are endocytosed into **early endosomes** where they are proteolysed. These bind to class II storage vesicles where the MHC class II molecules have their binding grooves protected by **CLIP** protein. The CLIP protein is degraded in the vesicle allowing the proteolysed foreign peptides to bind in the MHC groove [38].

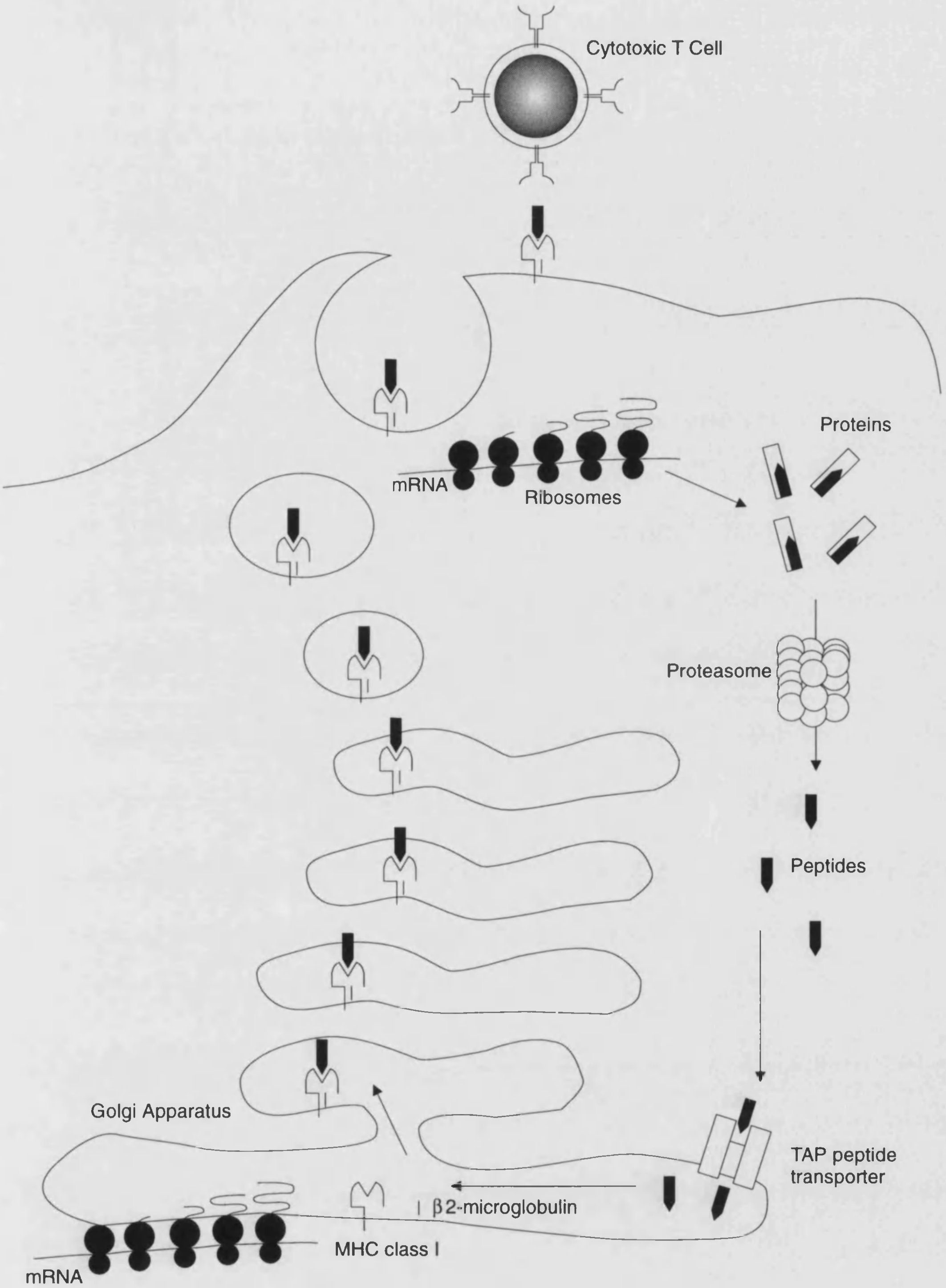


Figure 1.6: Endogenously produced protein antigens are broken down by the proteasome into peptides which tranfer through the endoplasmic reticulum membrane using the TAP peptide transporter. There they become associated with MHC class I molecules and are transferred via the Golgi apparatus to the cell surface where they are screened by T cells. Adapted from [10].

1.2 The Immunoglobulin Superfamily

A **superfamily** is a group of proteins which show sequence homology (at least 15%), usually encoded by a single exon and forming a compact structural unit. Each member of the superfamily probably derives from a common precursor by divergent evolution. Some examples of superfamilies are the globins, the growth factors, the neurotransmitter receptor ion channels and the immunoglobulins.

Membership of the immunoglobulin superfamily is indicated by the presence of one or more Ig domains (homology units) of 70-100aa's homologous to Ig Variable (V) or Constant (C) domains. The Ig superfamily comprises proteins which in general have receptor functions, many of which are involved in immune recognition but some in areas such as endocrine and neuronal recognition. Figure 1.7 shows some of the members of the Ig superfamily. There are three types of Ig domain, the V type (like an antibody V region), the C type (like an antibody C region) and the H type (hybrid sharing V and C character). Some proteins contain several Ig domains of one or more type [39].

The ancestry of the different types of domain is still a matter of some speculation. However it is thought that V and C domains split early in Ig evolution and that the gene rearrangement mechanism shown in Ab and TCR V regions developed early [40]. The low homology between Ig V domains and TCR V domains suggests an early split. The study of Ig evolution is hampered by the gene rearrangement mechanism. It is thought that Ig genes are evolving more rapidly than other proteins, probably because the antigen recognition role Ab and TCR

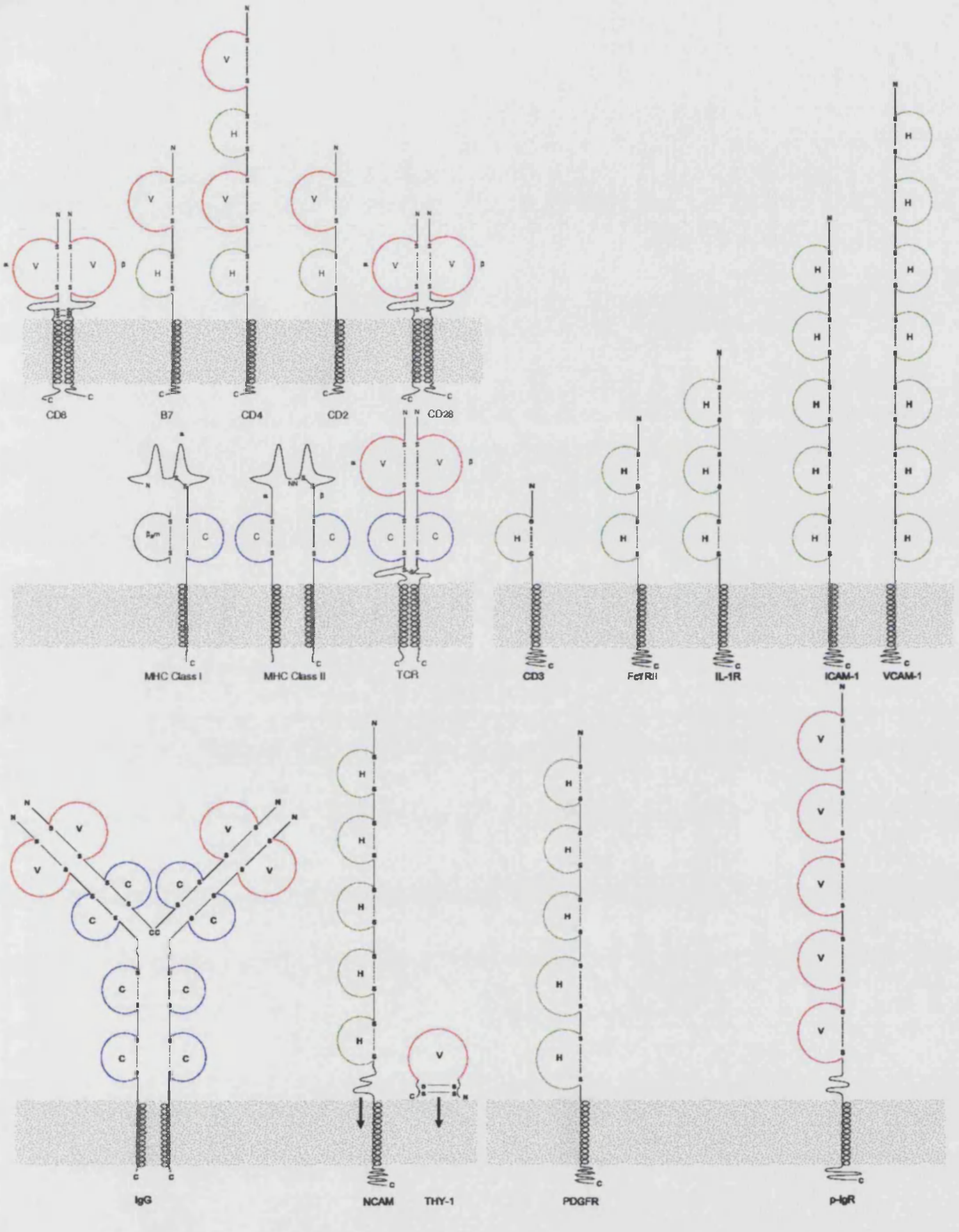


Figure 1.7: Some proteins containing immunoglobulin domains are shown. Each domain is represented by an arc. The three types of domain are C type (like Ab constant region) which are coloured blue, V type (like Ab variable region) which are coloured red and H type (hybrid containing both constant region and variable region character) which are coloured green.

perform makes rapid adaptation to new antigens vital [41].

1.2.1 Immunoglobulin Structure

The structure of an antibody F_{ab} was first solved by X-ray crystallography in 1973 by Poljak *et al* [42]. Since then the structures of many antibodies and several other members of the Ig superfamily have been determined (CD4 [43, 44], CD8 [45], NCAM, ICAM and MHC). Table 1.2 lists published antibody structures. The Ig fold consists of two antiparallel β-sheets of 4 and 3 (C domain) or 4 and 5 (V domain) strands as shown in figure 1.8. They form what is known as a "Greek Key" motif. These are brought together to form a compact structure with the hydrophobic residues in the β strands forming a stable core [46]. The two sheets are bonded together by a disulphide bond between two very highly conserved cysteines in strands B and F.

Brookhaven Entry	Name	Resolution (Å)	Chain Type	Reference
3hfl	HyHEL-5	2.65	κ/γII	[47]
3hfm	HyHEL-10	3.0	κ/γI	[48]
1bji/2bji	LOC	2.8	κ/κ	
2fbj	J539	1.95	κ/IGA	[49]
3fab/7fab	NEW	2.0	κ/γII	[50]
4fab	4-4-20	2.7	κ/γII	[51]
5fab/6fab	36-71	1.9	κ/γI	[52]
1mcp/2mcp	McPC603	3.0	κ/γIII	[53]
3mcg	MCG	2.0	$\lambda1/\lambda1$	[54]
1mcw	WEIR/MCG	3.5	$\lambda1/\lambda1$	[55]
2rhe	RHE	1.6	$\lambda1/\lambda1$	[56]
1rei	REI	2.0	κ/κ	[57]
2fb4/2ig2	KOL	1.9	$\lambda1/\gamma$III	[58]
1f19	R19.9	2.8	κ/γII	[59]
1fdl	D1.3	2.5	κ/γII	[60]
1mam	YS*T9.1	2.5	κ/γII	

Brookhaven Entry	Name	Resolution (Å)	Chain Type	Reference
8fab	HIL	1.8	$\lambda 1/\gamma$I	[61]
1baf	AN02	2.9	κ/γI	[62]
1hil/1hin/1him	17/9	2.0	κ/γII	[63]
1igf/2igf	B13I2	2.8	κ/γI	[64]
1dfb	3D6	2.7	κ/γI	[65]
1igm	POT	2.3	κ/?	[66]
1bbd	8F5	2.8	κ/?	[67]
1ncd	NC41	2.9	κ/?	[68]
1igi	26-10	2.7	κ/γIIA	[69]
1ggi	50.1	2.8	κ/γII	[70]
1acy	59.1	3.0	κ/γI	[71]
1bbj	B72.3	3.1	κ/γIV	[72]
1bre	BRE	2.0	κ/κ	[73]
1cbv	bv04-01	2.66	κ/?	[74]
1eap	17E8	2.5	κ/?	[75]
1fbi	F19.3.7	3.0	κ/γI	[76]
1fgv	HUH52-AA	1.9	κ/?	[77]
1fig	1F7	3.0	κ/γI	[78]
1for	FAB17-IA	2.75	κ/γIIA	[79]
1fpt	C3	3.0	κ/γIIA	[80]
1frg	FAB26/9	2.8	κ/γIIA	[81]
1fvc	4D5	2.2	κ/?	[82]
1gig	HC19	2.3	λ/γI	[83]
1igc	MOPC21	2.6	κ/γI	[84]
1ikf		2.5	κ/γI	[85]
1ind	CHA255	2.2	λ/γI	[86]
1ivl	M29B	2.17	κ only	[87]
1jel	JE142	2.8	κ/?	[88]
1jhl	D11.15	2.4	κ/γI	[89]
1lmk	L5MK16	2.6	κ/?	[90]
1mlb/1mlc	D44.1	2.1	κ/γI	[91]
1nbv	BV04-01	2.0	κ/?	[74]
1nma/1nmb	NC10	3.0/2.5	κ/?	[92]
1nsn	N10	2.9	κ/γI	[93]
1opg	OPG2	2.0	κ/?	[94]
1rmf	R6.5	2.8	κ/?	[95]
1tet	TE33	2.3	κ/γI	[96]
1vfa	D1.3	1.8	κ/γI	[97]
1wtl	WAT	1.9	κ/κ	[98]
2cgr		2.2	κ/γII	[99]

Brookhaven Entry	Name	Resolution (Å)	Chain Type	Reference
2dbl	DB3	2.9	κ/γIIA	[100]
2f19	R19.9	2.8	κ/γIIB	[101]
2gfb	CNJ206	3.0	κ/γIIA	[102]
3bjl	LOC	2.3	λI/λI	[103]

Table 1.2: List of antibody structures and their fragments whose X-ray coordinates are currently available at the time of writing from the Brookhaven database of protein structures.

Three of the four loops between strands at one end of the core in the V domain of each chain form the antigen combining site in antibodies [104] and probably the MHC-peptide binding site in TCRs. These are the loops between B and C strands, C' and C" strands and the F and G strands.

In the TCR α chain V region structure the C" strand switches from one sheet to the other [4].

A single protein chain often contains more than one immunoglobulin domain separated by short connecting regions. The chains themselves are often associated into multimeric proteins; for example IgG contains four chains, two heavy and two light. The heavy chain contains four domains V_H C_{H1} C_{H2} and C_{H3}. The light chain contains two domains V_L and C_L. There are interactions between the V_H and V_L, the C_L and C_{H1}, the C_{H2} and C_{H2}, and the C_{H3} and C_{H3} domains. The interactions between the V domains are between the five strand sheets. In the C regions the four stranded sheets interact. Other members of the immunoglobulin superfamily such as CD8 are multimers with similar pattern of association. Some, however, are monomers, for example CD4 which is made up of four Ig domains in a single chain.

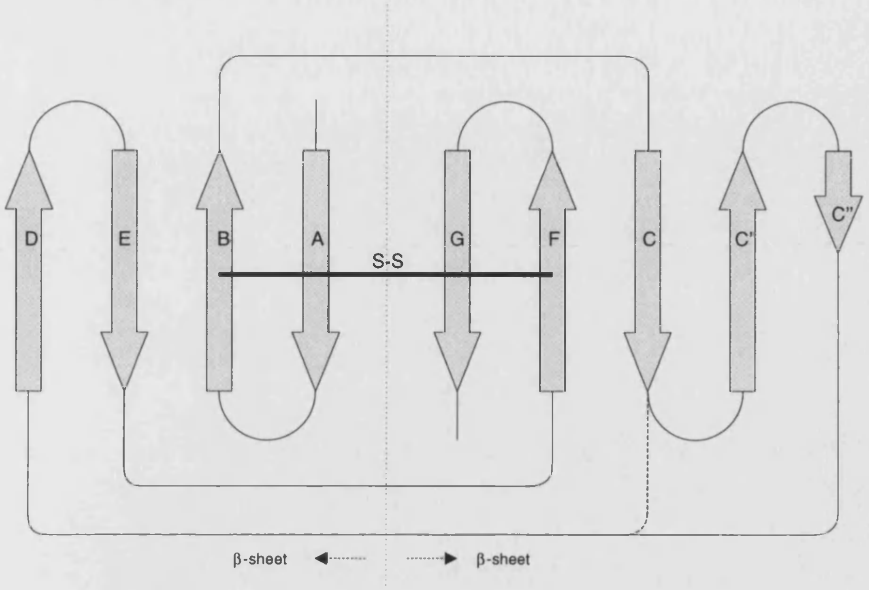


Figure 1.8: The arrangement of β strands in antibody V and C domains. V Domains have one sheet consisting of 5 and one of 4 strands and C domains one of 3 and one of 4 strands. Also shown is the disulphide bond between the two sheets.

1.2.2 TCR Gene Arrangement

The polymorphic TCR chains are encoded in four gene loci **TCRA** (α chain), **TCRB**(β chain), **TCRG** (γ chain) and **TCRD** (δ chain).

The TCRB chain locus consists of 25 V region segments followed by two tandemly arranged groups of 1 D, 6 J and 1 C region. 3' to the second C region is an inverted V region [105]. The TCRG locus in humans is similar to the β chain locus overall, having two groups of J and C segments. It however lacks D regions and has fewer V regions (8). The two tandemly arranged groups of J and C regions also contain different numbers of J regions (2 and 3). The TCRD locus is unusual in that it is completely enclosed in the TCRA locus. The TCRD locus V region gene segments are interspersed amongst the more than 75 α chain V region genes.

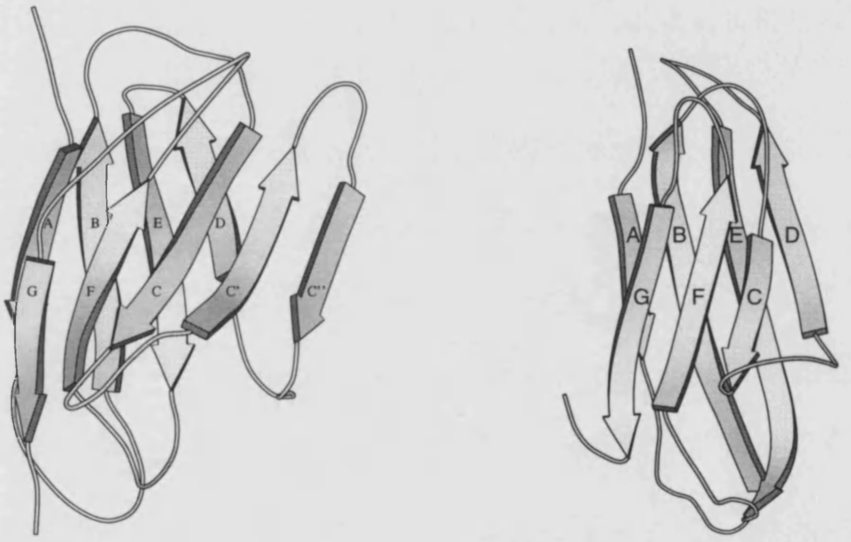


Figure 1.9: The arrangement of β strands in the antibody V_H and C_{H1} domains of 4-4-20 [51]. V Domains have one sheet consisting of 5 and one of 4 strands (left). C domains are missing the C' and C" strands having 4 and 3 stranded sheets (right). Diagram created with molscript [26].

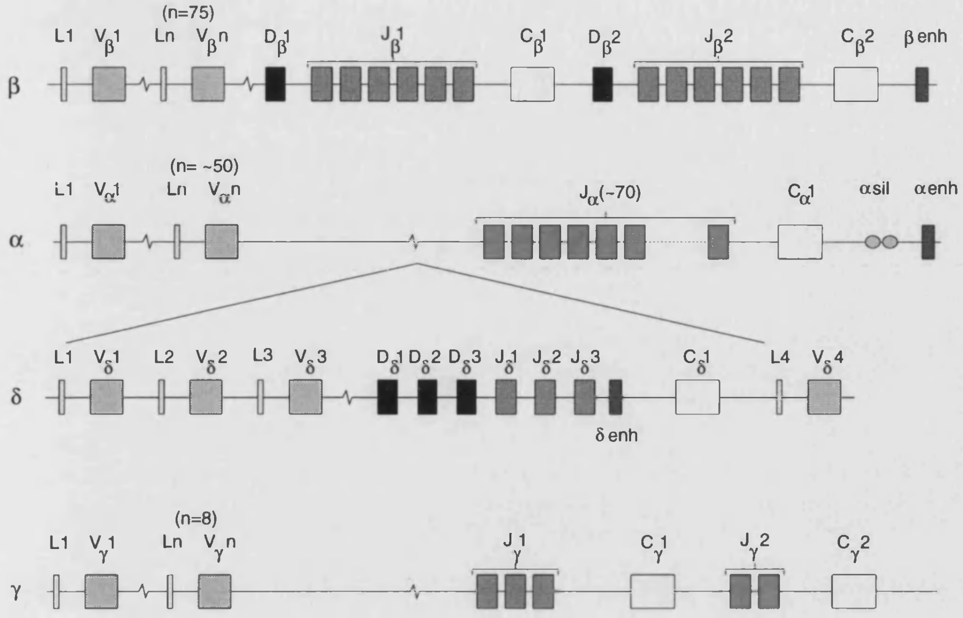

(n=75)
L1 $V_\beta 1$ Ln $V_\beta n$ $D_\beta 1$ $J_\beta 1$ $C_\beta 1$ $D_\beta 2$ $J_\beta 2$ $C_\beta 2$ βenh
β
(n= ~50)
L1 $V_\alpha 1$ Ln $V_\alpha n$ $J_\alpha(\sim 70)$ $C_\alpha 1$ αsil αenh
α
L1 $V_\delta 1$ L2 $V_\delta 2$ L3 $V_\delta 3$ $D_\delta 1$ $D_\delta 2$ $D_\delta 3$ $J_\delta 1$ $J_\delta 2$ $J_\delta 3$ $C_\delta 1$ L4 $V_\delta 4$
δ
δenh
(n=8)
L1 $V_\gamma 1$ Ln $V_\gamma n$ $J_\gamma 1$ $C_\gamma 1$ $J_\gamma 2$ $C_\gamma 2$
γ

Figure 1.10: Schematic of the human TCR gene complexes. The top line shows the β chain complex. The second and third the α/δ complex and the bottom line the γ chain complex. Adapted from [10].

There are 2 D_δ regions followed by two J_δ regions and a C_δ region. 3' to this is an inverted V_δ region [106, 107]. The 75 V region genes of the α locus are 5' to the TCRD locus whilst the 70 J_α regions are 3' to it. Some of the V region gene segments can be used in both α and δ chains. The single C_α region is positioned 3' to the last J_α region. The arrangement of the loci is illustrated in figure 1.10.

1.2.3 TCR Gene Rearrangement

Gene rearrangement is the general mechanism by which TCRs generate their diversity [108]. The first two stages occur at the DNA level by somatic recombination, initially between a D and a J gene segment (β and δ chains only) and subsequently between a V and the rearranged D-J (β and δ) or V and J (α and

γ) gene segments. The DNA between the particular V, D and J regions utilised is usually excised during the recombination processes. The processes involved in TCR gene rearrangement are shown in figure 1.11.

A primary mRNA transcript consisting of the leader exon, the combined VDJ exon, and the region 3' to this, including the C region gene exons, is transcribed. RNA processing (splicing) then generates the LVDJC mRNA transcript which is translated to protein.

The mechanism of V, D, J somatic recombination has been studied in detail. It involves the protein products of two genes known as **recombination activating genes** (RAG) . RAG1 and RAG2 recognise conserved heptamer and nonamer base sequences which are separated by either 12 or 23 base pairs. These segments are present 3' to each V region, 5' and 3' to each D region and 5' to the J region segments. Hybridisation between a pair of heptamer and a pair of nonamer sequences leads to recombinase enzyme recognition, and excision of the DNA between the two gene segments, as long as the loop contains one 12 and one 23. Excision will not occur if two 12's or two 23's are between the two heptamer nonamer pairs.

1.2.4 Immunoglobulin Sequence Diversity

A comparison of the TCR and antibody V domain diversity is shown in table 1.3. The smaller number of V segments in TCR compared to antibodies reduces the variability possible due to combinatorial association. By contrast, diversity within the junctional region is greater for TCRs than for antibodies. This difference in distribution of variability led to the suggestion that the junctional region in

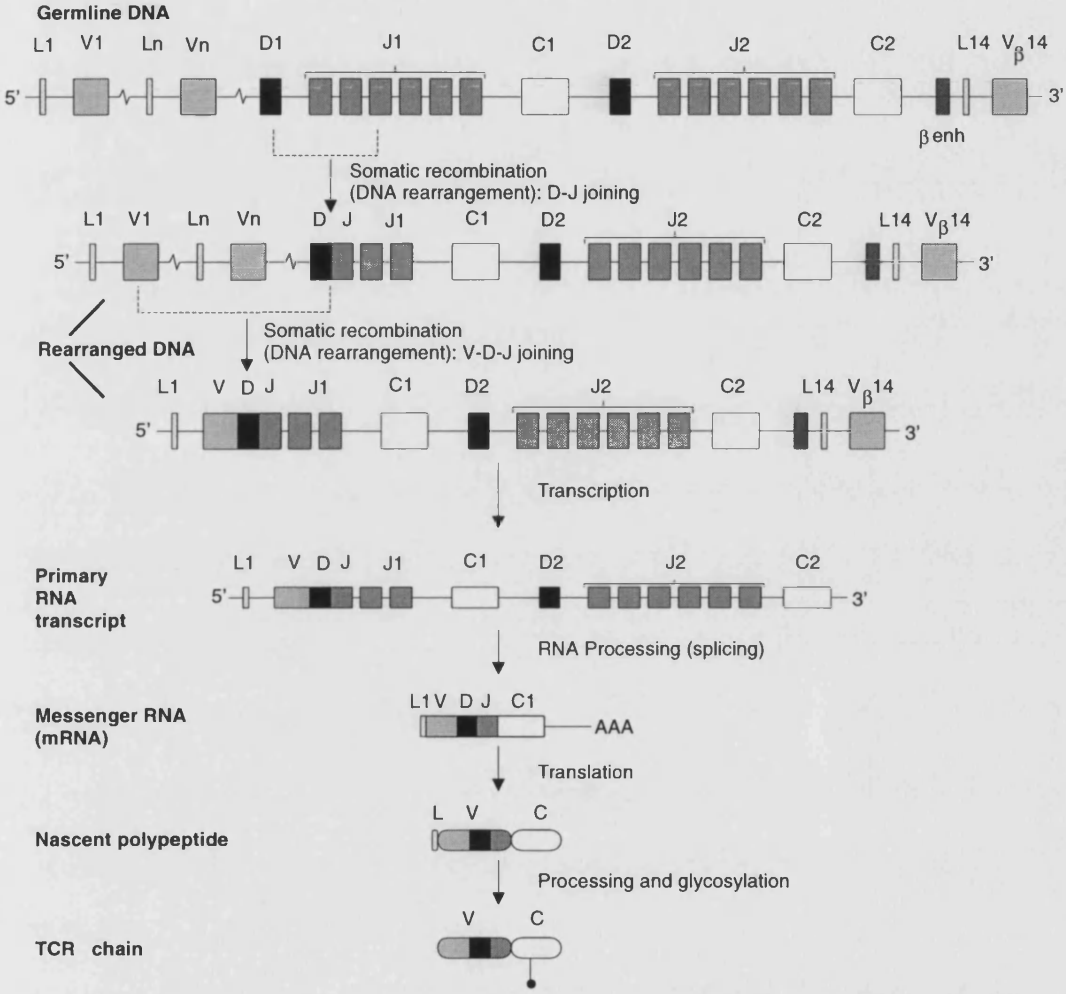


Figure 1.11: The scheme represents the stages involved in producing translated TCR protein from the chromosomal DNA. First the DNA in the cell recombines (described in the text) to splice together a V, for β D, J and a C region into a transcribable gene. This is transcribed and translated. Adapted from [10].

TCRs, that codes for the CDR 3 region in antibodies, is responsible for binding to the peptide antigen, while the V gene segments, which have less possibility for variability, are involved in binding the MHC molecule. Several investigators have examined this possibility by sequencing the receptors from several T cell clones reactive against the same MHC and antigen, to determine whether V segment usage was conserved [109–113]. The consensus from these studies is that TCRs recognising the same antigen often express a limited repertoire of V and J gene segments but that there is no complete dependence on a single segment for each MHC.

The number of possible TCR sequences is very large and includes many which could be reactive against self components and produce adverse autoimmune responses in the body. To prevent such problems there is a very restrictive selection process. This occurs in the thymus, and the major mechanism appears to be **clonal deletion**. It is obviously important that the TCR should be sensitive to the composition of the peptide as well as the MHC, so it is necessary to eliminate T cells which express TCRs which bind too strongly to MHC. However TCRs with low affinity for host MHC would also need to be eliminated as these might not be able to discriminate self from non-self MHC alleles [37].

1.3 Summary

T cells form an integral part of the immune system, having two major functions; the killing of virally infected cells (cytotoxic T cells) and the stimulation of spe-

		Ig		TCR	
		H	κ	α	β
	variable(V)	250-1000	250	100	25
	diversity(D)	10	0	0	2
	joining(J)	4	4	50	12
Variable region combinations		62,500–250,000		2500	
Junctional Diversity	usage of different D and J segments	yes	yes	yes	yes
	variability in 3' joining of V and J	rarely	rarely	yes	no
	D joining in all three reading frames	rarely	-	-	often
	N region diversity	V-D, D-J	none	V-J	V-D, D-J
Junctional combinations		$\sim 10^{11}$		$\sim 10^{15}$	
Total repertoire		$\sim 10^{11}$		$\sim 10^{15}$	

Table 1.3: Table comparing the mechanisms involved in the generation of antibody and TCR diversity. Estimates of the total diversity are shown.

cific B cells to produce antibodies (T helper cells). The TCR is vital to the T cell's functioning, acting as a specific receptor for both an antigenic peptide and self MHC. Many different TCRs can be generated by the mechanisms of junctional and combinatorial diversity, but each T cell expresses only one TCR sequence. This provides the specificity of the T cell response. How the TCR itself signals that it has bound to an MHC is under intense investigation. However, the cascade of signalling reactions which occur within the T cell have been widely studied. They cause the T cell to express interleukins as well as receptors for them. These compounds cause proliferation of the T cells, and initiate B cell proliferation and differentiation.

1.4 Aims and Objectives

The TCR is a member of the immunoglobulin superfamily. Methods for the accurate prediction of the structure of antibody variable domains exist. This thesis describes an approach taken to use homology modelling to produce models of TCRs based on methods used to model antibodies. The initial work concentrated on making comparisons of TCR and antibody sequences and improving antibody modelling. Information from these studies was then used to devise a method for modelling TCRs and this method was used to create models of TCRs. The recent publication of the TCR structures has enabled some comparisons to be made between models and structure. This work is described in the subsequent chapters.

Chapter 2

Sequence Analysis of Immunoglobulins.

2.1 Introduction

The sequence is the starting point for any homology modelling approach. For antibodies and TCRs, the diversity generation mechanisms give rise to many different sequences. The structure of all antibody Fv's so far determined have a very similar core structure, varying mainly in the hypervariable loops which form the antibody combining site (three from each chain). However, not all the regions of sequence between these loops is completely conserved in different V and J regions. Those residues which *are* highly conserved are therefore likely to be important for maintaining the structure, or to have some other important functional role. There are several thousand antibody V region sequences published and several hundred TCR sequences. Therefore sets of homologous sequences can provide invaluable information in producing models of the structure.

There are known relationships between sequence and structure and sequence and function for antibodies [46,104,114–117]. Studies on antibodies indicate that only a proportion of the residues in the sequence interact directly with antigen. Almost all of these residues fall into regions of the sequence which are hypervariable between antibody sequences. There are three such regions in antibodies; CDR1, CDR2 and CDR3. CDRs 1 and 2 are coded for in the V segment of germline DNA while CDR 3 encompasses the junctional region including the 3' end of the V segment, the D segment in the antibody heavy chain and TCR β and δ chains, and the 5' end of the J segment.

Some residues in antibodies, outside the CDRs in the more conserved framework regions, are almost completely conserved between sequences. Structural studies have shown that some of these residues are involved in structurally important interactions such as those between the strands and those in the V-C and V_L-V_H interfaces [117,118].

Some structurally important interactions between framework and CDR residues have also been found. These lead to the ability to use sequence patterns to divide some CDRs into **canonical** groups [119–122].

The frequencies of residues in CDRs of antibodies is distinctly different to the framework regions. Residues such as tyrosine, tryptophan and asparagine are more common in the CDRs than the framework. The aromatic residues in CDRs are more exposed than is usual while the asparagines are more buried. It is thought that the asparagines are involved in structural interactions and the aromatic residues provide an important contribution to the binding energy for the antibody-antigen interaction [123].

As there are few structures of TCRs, structure-function relationships are less certain. However the sequence homology of the two groups of sequences enables the information (such as the positions of CDRs and the roles of conserved residues) and techniques (such as variability plots) that have been applied to antibody sequences to be transferred to studying TCRs.

The TCR sequences are divided into non overlapping **subfamilies** based on a homology criterion of 75%. These classes have had various naming schemes applied to them over the years but recently the mouse and human TCR sequences have been assigned a standardised naming scheme by the nomenclature sub-committee of the International Union of Immunological Societies (IUIS) [124–126].

A comparison of TCR sequences to antibody sequences has been made by Chothia and Lesk [127]. This showed that the TCR V domain sequences were consistent with an antibody V type fold. Variability analysis, which identifies the regions of family of sequences that are most variable in terms of the residues which occur, was performed on TCR sequences by Jores *et al* [128]. This showed peaks of variability in TCRs in regions equivalent to antibody CDRs. An extra peak of variability was also seen between the CDR 2 and CDR 3 equivalent regions in the TCR β chain. This chapter, rather than repeating this work concentrates on specifically identifying which of the antibody chain types is most similar to each TCR chain type, in order to identify the best framework to use in molecular modelling of TCRs.

2.2 Methods

2.2.1 Sequence Selection

Sequence alignments of antibody light and heavy chains, TCR α and β chains and MHC chains are maintained by Kabat and Wu [129]. The "fixed length" file versions of these alignments were obtained via anonymous ftp from ncbi.nlm.nih.gov (mirrored at ftp.ebi.ac.uk) and converted into NBRF format using the program MOL [130]. The converter checked for data consistency and several errors were found and notified to the database maintainer who posted updates. The NBRF format files were analysed in the program SR [131].

Wherever possible sequences were obtained from the Kabat database. When a sequence was not obtainable from Kabat the Genbank database was used. Three different groups of sequences were selected, the CDR3 regions, the complete antibody V region and the complete TCR V regions. There were many more CDR3 regions in the Kabat database than complete V regions. When examining the properties of these regions it is better to include all those which have a CDR3 region rather than just those with the complete V region.

2.2.1.1 CDR3 Sequences

All the sequences of a particular chain type were selected, for mouse and human separately, from the Kabat database. The next stage in the selection process was to discard all sequences in the Kabat database containing any unknown or ambiguous residues, for example containing 'X', 'B', 'Z' or '?' residues. The rest of the sequences were selected.

A subset of sequences was also created in which identical sequences were eliminated. This subset was used to check for any possible bias in the original Kabat data. It also provided the possibility of doing a parallel analysis, the results of which could be compared with the original set.

2.2.1.2 Complete Antibody V Region Sequences

The sequences were again selected from the Kabat database after entries with ambiguous or unknown residues had been discarded. The selection was also restricted to all the sequences which started within 5 residues of the start and ended within 5 residues of the end of the standard Kabat alignment. Sequences containing an identical V segment to other selected sequences were also eliminated.

2.2.1.3 TCR V Region Sequences

The paper describing the IUIS classification scheme for TCR V segment sequences contains information on which clone names belong to each class in the scheme. These clone names were initially used to identify the sequences to be retrieved from the database. The Kabat entry number for each of these clones in the database was identified. The corresponding sequence was retrieved from Kabat and was then used as the basis of a further search for similar sequences. Sequences which matched in either residues 1-50, 50-93 or 30-70 (Kabat residue numbering) were identified.

V segment sequences which had been extracted directly from the published classification scheme were then aligned in the same way as those extracted from Kabat. These also were then used as probes for further searches of the Kabat database (again using the same three sequence ranges).

It was possible that there were further matching entries in Kabat which had not been retrieved because they contained sequence errors, inconsistent clone names or were only partial sequences. In an attempt to retrieve these sequences, a selection was made of all sequences containing 30 residues or more, but excluding the ones already retrieved. These were examined manually and an exact class or, failing this, the closest matching class, was determined for each.

Any sequences in the classification table which had still not been found by this procedure and also where the Kabat sequence identified was very short, were searched for in the Genbank database. The SR program was used to translate the Genbank Nucleotide sequence into a protein sequence using the feature entries present in the Genbank file. These translated regions for each sequence were stored on file in case the process needed to be repeated.

Retrieved sequences from all sources were then compared to the classification scheme to look for errors. Inconsistencies or obvious errors were noted and the sequence was discarded. For each class, excluding non functioning or pseudogene classes, the best sequence was identified. In most cases this was the longest, but sequences which had known TCR α and β pairings were favoured, whereas pseudogene sequences in a non pseudogene class were considered unsuitable.

The chosen sequences were then extracted from the various sources into a NBRF format file. All comments in the original format files were retained. Any Genbank sequences thus selected were aligned manually according to the Kabat alignment.

The complete list of sequences is shown in tables B.2 to B.7 in appendix B.

2.2.1.4 Obtaining Pairs

Paired sequences were then obtained from all the selected sequences. They were identified by the 'AAName' field from the Kabat database. In paired sequences this has the same value in both the α and the β sequence entries for the TCR, or in the light and heavy chain sequence entries, for antibodies

2.2.2 Sequence alignments

Various programs exist which attempt to automatically align sequences (AMPS [132], CLUSTALW [133], PILEUP [134]). It was initially attempted to use these programs to align the TCR and antibody sequences against one another. However, these programs failed to produce sensible alignments, probably due to the large number of sequences, the low homology between antibodies and TCRs and the high variability in length of the CDR 3 regions. It was therefore decided to use a more manual approach to align the sequences, based on structural comparisons between the antibody light and heavy chains.

For the T cell receptor groups the first step in the alignment process was to manually align the sequences extracted from the Genbank database to fit the Kabat alignment.

The Kabat database contains inserted residue fields for some of its sequences. These are sequences which do not fit the Kabat alignment because they have extra residues between the standardised Kabat positions. To incorporate these residues in the data, extra positions were added to the antibody and TCR alignments (automatically in the program SR).

For the CDR3 region of the heavy chains the length of the alignment expanded to approximately 70 residue positions although the longest sequence was only 32 residues. This was because similar regions of sequence were inserted at several different positions. The alignment was edited in this region to consolidate the gaps. The heavy and light chain structures of 50 antibodies were all superimposed onto one another using conserved residues in the B and C β strands. From this it was possible to determine the positions where the light and heavy chains differed structurally. The alignments were altered at these positions to create a structural alignment. This was achieved by inserting or moving gaps as required.

The next step was to match the TCR alignments to this structural alignment for antibodies. This was achieved by aligning the conserved residue motifs which define the limits of the CDRs for antibodies and TCRs.

2.2.3 Sequence variability

Sequence variability was calculated for each chain type for sequences which contained V, D and J segments using the method of Wu and Kabat [135]. The variability at each alignment position is calculated as the number of residue types occurring at the position divided by the frequency of the most commonly occurring residue at the position.

$$Variability = \frac{number\ of\ different\ amino\ acids\ that\ occur\ at\ site}{frequency\ of\ most\ common\ amino\ acid\ at\ site} \qquad (2.1)$$

2.2.4 Sequence Similarity

Comparisons were made between TCR α and β chains and human light and heavy chains using a similarity score comparing one set of sequences to two other sets of sequences. The score at each position is defined as the difference in frequencies of occurrence of each amino acid at each position compared to each set. The score ranges from -2 to 2 with 0 indicating equal similarity to each chain type. Positive values indicate greater similarity of the comparison set to one of the sets and negative values greater similarity to the other set.

$$SimScore = \sum_{res=1}^{20} (Freq1_{res} - Freq2_{res}) \qquad (2.2)$$

A dissimilarity score was also calculated for each set of frequencies which was the sum rather than the difference of the two frequency sets. The two scores together indicate which positions are useful in distinguishing which chain type the comparison set is more similar to.

Loop	Light	Heavy	α	β
CDR 1	24 - 34	23 - 35B	23 - 33	24 - 33
CDR 2	50 - 60	50 - 65	49 - 61	48 - 64
CDR 3	89 - 97	95 - 102	91 - 105	93 - 107

Table 2.1: Kabat residue numbering [129] ranges used in CDR length calculations. Using the residue numbering in figure 2.2 the ranges are 24–41 (CDR 1), 57–74 (CDR 2) and 105–137 (CDR 3).

2.2.5 CDR Ranges

The residue ranges used to delimit the CDRs in the length calculations are shown in table 2.1. TCR α-β sequence pairs used in the combined CDR 3 length calculation were generated by linking sequences with the same clone name. Highly conserved sequence motifs are present at each of these limits, ensuring that the length of the equivalent piece of structure is being compared. The structure of regions around the three CDRs in the antibodies with solved X-ray structures, as well as CD4 and CD8, were examined (see section 4.5).

2.2.6 Sequence Environments

The similarity of the structural environment of TCRs and antibodies was examined by calculating a normalised similarity score for each position which included all residues within 5 Å of each aligned position in an antibody and aligned TCRs. The residues within 5 Å of each residue were determined using the average distance between the centres of gravities of all residues.

$$EnvScore = (\sum_{res=1}^{NClose} SimScore_{res})/NClose \qquad (2.3)$$

2.3 Results

2.3.1 Framework comparisons

Table 2.2 shows the average sequence identity between TCR α and β chains and the sequences of antibodies of known structure. The overall level of identity is low at 15-22 %. The identity to light chains is higher than to heavy chains for both TCR α and β chains.

Such low identities between the sequences would often indicate that they had different structures. However, other immunoglobulin superfamily members such as CD4 and CD8 show high structural homology to antibodies but low sequence homology ($<$ 20%).

An analysis of the frequencies of occurrence of residues at positions thought to be vital for maintaining the structure of antibodies was carried out to further quantify the light and heavy chain character of the two types of TCR sequences (table 2.3). The positions chosen were those which are distinct to either antibody light or heavy chains. Positions 145 and 147 (residue numbering as in figure 2.2) are involved in V_H - C_H contacts. They are also very conserved positions in light chains. TCR β chains were most similar to λ light chains at these two positions, while the TCR α chains showed greater variation at these positions

Antibody Sequence	TCR α	TCR β
gloop2 light	21.6	20.3
D1.3 light	22.2	20.6
Hyhel 5 light	21.7	20.6
Hyhel10 light	22.0	20.7
J539 light	20.3	19.5
Kol light	20.3	21.6
MCPC603 light	22.7	20.5
New light	22.9	22.9
Rei light	22.8	20.9
Rhe light	22.0	22.7
R19.9 light	21.4	20.2
gloop2 heavy	16.1	14.7
D1.3 heavy	14.4	14.0
Hyhel 5 heavy	17.1	15.8
Hyhel 10 heavy	16.8	16.9
J539 heavy	17.1	17.8
Kol heavy	17.5	18.2
MCPC603 heavy	17.0	17.5
New heavy	18.8	16.3

Table 2.2: The average sequence identities between α and β chains and antibodies of known structure. The identity score is calculated across the whole V domain sequence including the CDR regions.

Res. Num.	Frequency					Interface
	TCR α	TCR β	Ab Heavy	Ab κ	Ab λ	
43	Y 0.77 F 0.22	Y 1.00	V 0.83 I 0.11	Y 0.79 F 0.14	Y 0.92	V_L-V_H
51	P 0.53 L 0.40	L 0.75 P 0.13	L 0.99	P 0.87	P 1.00	
53	L 0.45 F 0.16	L 0.41 F 0.37	W 0.93	L 0.75	L 0.78	
138	F 0.96	F 1.00	W 0.99	F 0.99	F 0.98	
145	T 0.25 S 0.24 Q 0.12	T 0.62 L 0.17 S 0.17	T 0.98	E 0.98	T 0.97	V_H-C_H
147	K 0.24 I 0.16 S 0.16	L 0.86	S 0.99	K 0.96	L 0.96	

Table 2.3: Table of residue frequencies at conserved interface residues. The residue numbering is as in figure 2.2. All residue frequencies greater than 0.1 are shown at each position. Only positions which are present in both light and heavy chains and differ in residue type between the two chain types are shown. The list of interface residues was taken from [127].

than the antibody chain types. Positions 43, 51, 53 and 138 are involved in V_L - V_H contacts. In both α and β these residues show greater homology to light chains than heavy chains. At three out of these four positions there is virtually no heavy chain character while at position 51 it is interesting to note that both α and β chains show some light and some heavy chain character.

The lengths of the conserved framework loops A-A', A'-B and E-F, as noted by Chothia *et al* [127] are similar to the antibodies. Of these regions, only the A-A' loop has a different length in κ light (3 residues) and heavy (2 residues) chains. In both α and β chains the majority of sequences (94% murine TCR α, 93% murine TCR β) show the same length as the κ light chains.

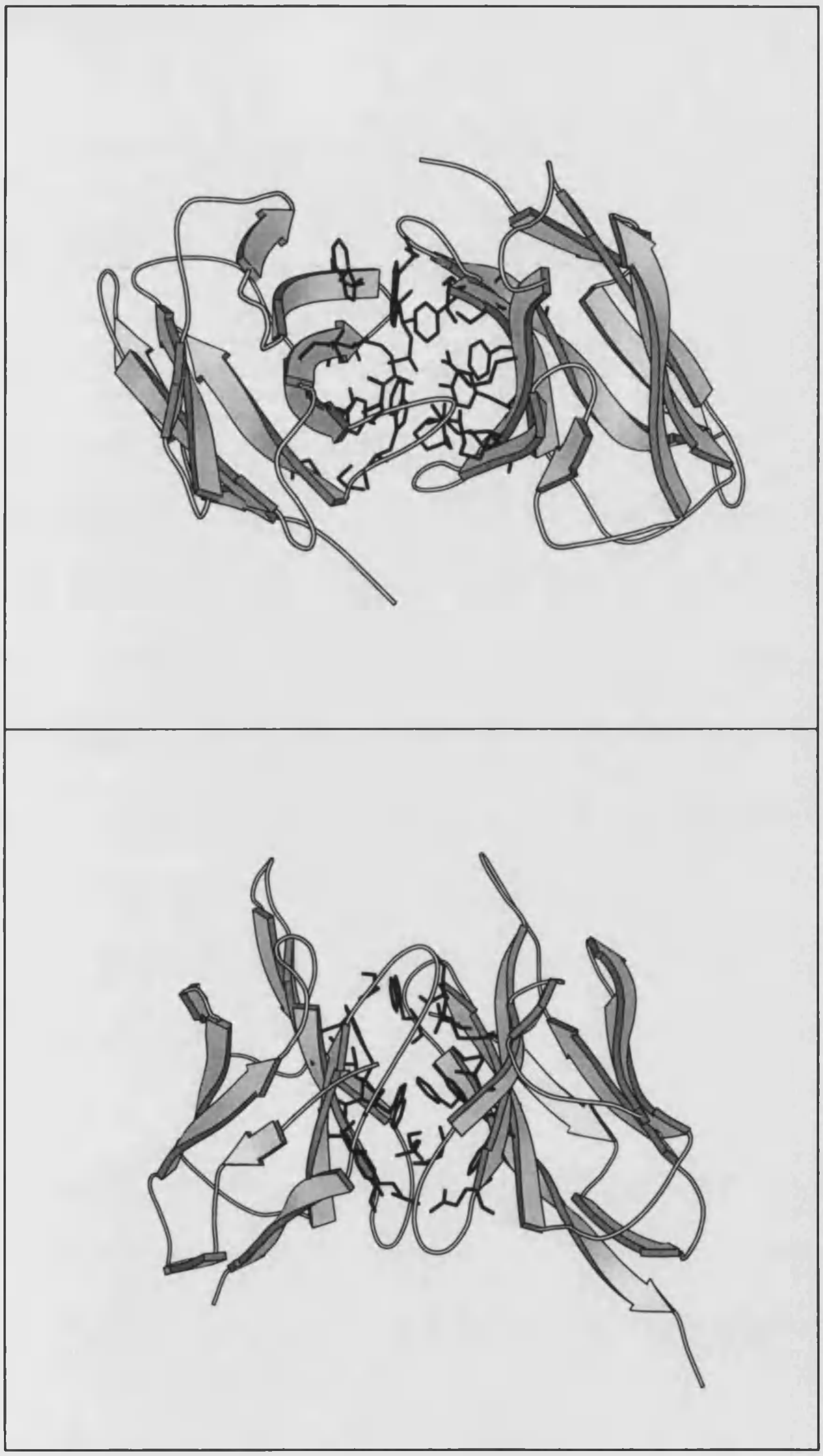


Figure 2.1: The V domain of the antibody 4-4-20 [51] showing the structurally important interface residues.

2.3.2 Alignments

A sample of twenty sequences from the alignments of TCR α and TCR β chains are shown in figures 2.2 and 2.3 compared to samples of twenty sequences from the antibody κ, λ and γ chain alignments.

Figure 2.4 shows the RMS at each residue position along the chain for antibody κ, λ and heavy chains. The plots highlight the high structural homology between the β-strand framework regions of the antibody structures. The plot including all light and heavy chain structures shows that the strands are structurally conserved between light and heavy chains.

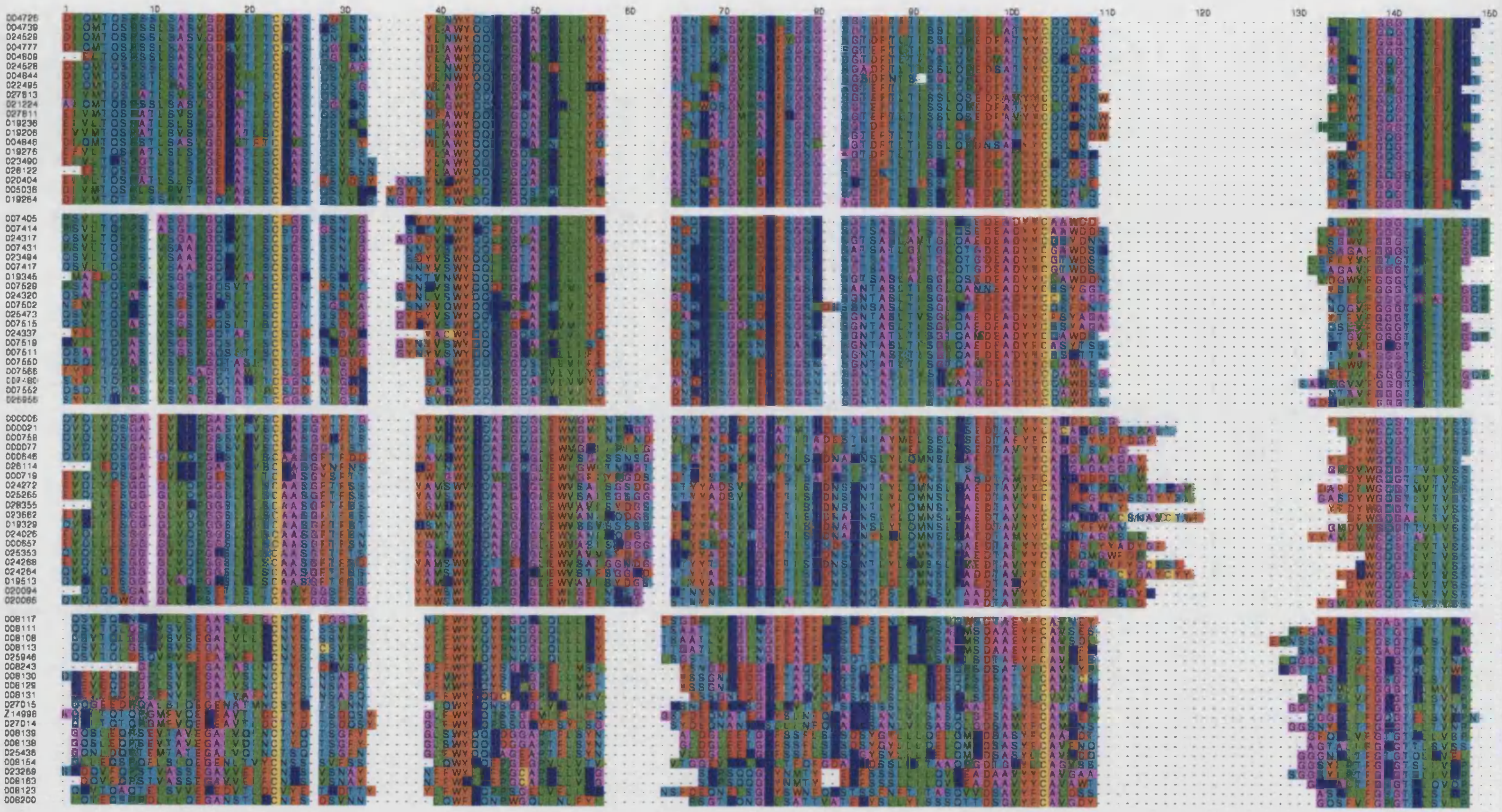


Figure 2.2: An extract of the Ig κ, Ig λ, Ig γ and TCR α chain alignment. Twenty sequences from each group are shown with κ chains at the top and α chains at the bottom. The residues are coloured according to type. The all numeric identification codes are Kabat database entry IDs, while the others are Genbank database entry IDs.

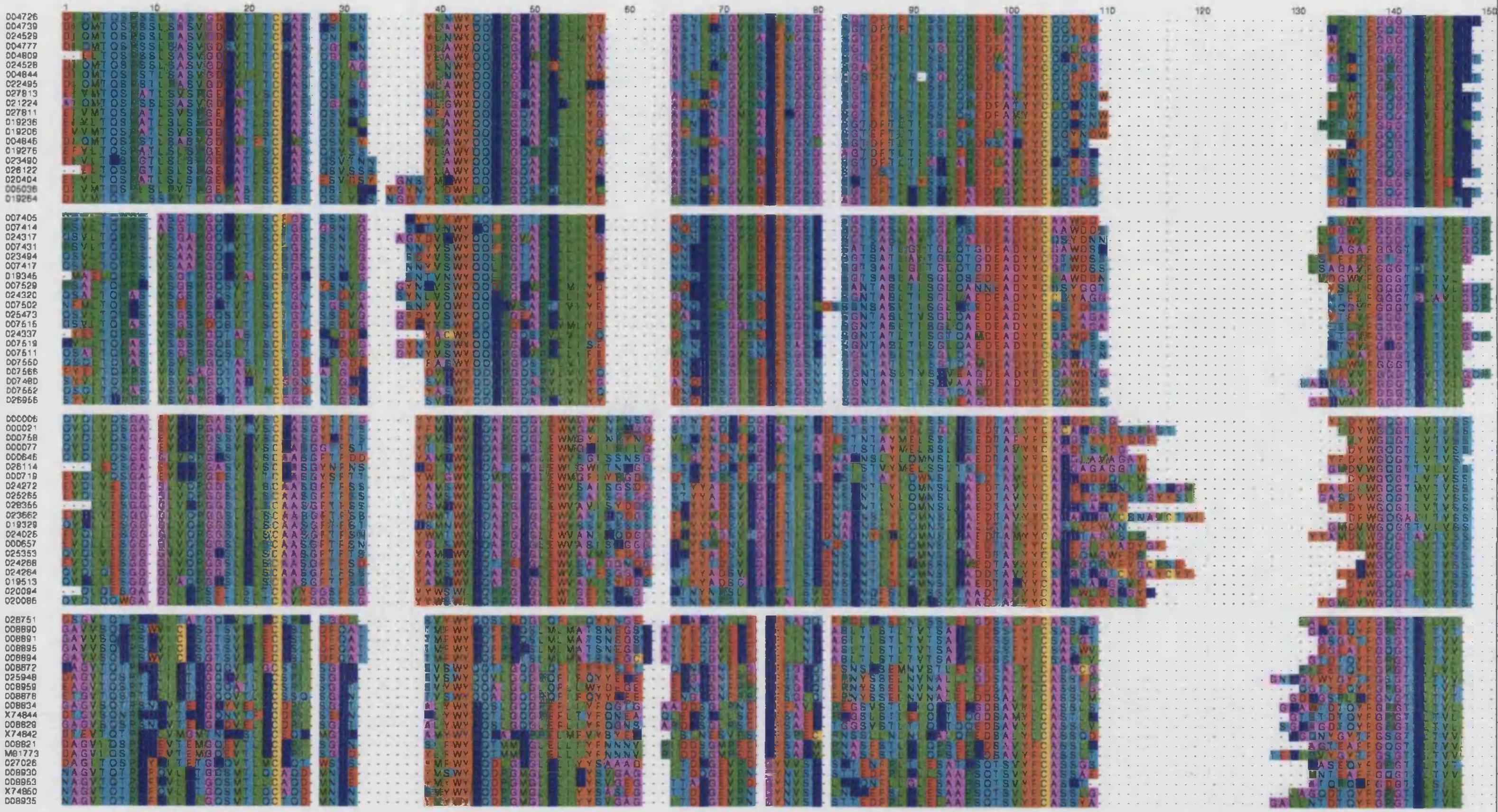


Figure 2.3: An extract of the Ig κ, Ig λ, Ig γ and TCR β chain alignment. Twenty sequences from each group are shown with κ chains at the top and β chains at the bottom. The residues are coloured according to type. The all numeric identification codes are Kabat database entry IDs, while the others are Genbank database entry IDs.

2.3.3 Sequence Variability

Variability analyses using the Wu and Kabat variability index, for the α and β chains as well as for the antibody light and heavy chains, were performed. The results, for mouse and human, are shown in figure 2.5 and figure 2.6 respectively.

These two sets of results were first compared and it was noted that the variability of the λ chain is much less in mouse than in human. This can be explained by the smaller repertoire of mouse λ chains. There is also some difference in the D-E loop - E strand region (sometimes referred to as the 4th hypervariable region) of the TCR β chain. In all other respects the mouse and human plots showed a similar variability pattern.

Overall, the plots show higher variability in the TCR sequences than in the antibody sequences. In the α chain there are discernible peaks in all three CDR equivalent regions. Although the β chain has peaks in the CDR2 and CDR3 equivalent regions, it has no discernible peak in CDR1. This may indicate that it is not involved in binding to the MHC molecule or peptide or that it interacts with a highly conserved MHC region.

The following observations on the variability in the β strands can be made. In the α chain there is a small peak in variability in the region between the A and A' loops. In the antibody chains only the κ chain shows such a peak although it has a much smaller variability. The pattern of variability in the C - C' region is similar in both the TCR α and β chains and both antibody light chain types, although again the TCR shows a higher variability. In both human and mouse the β chain

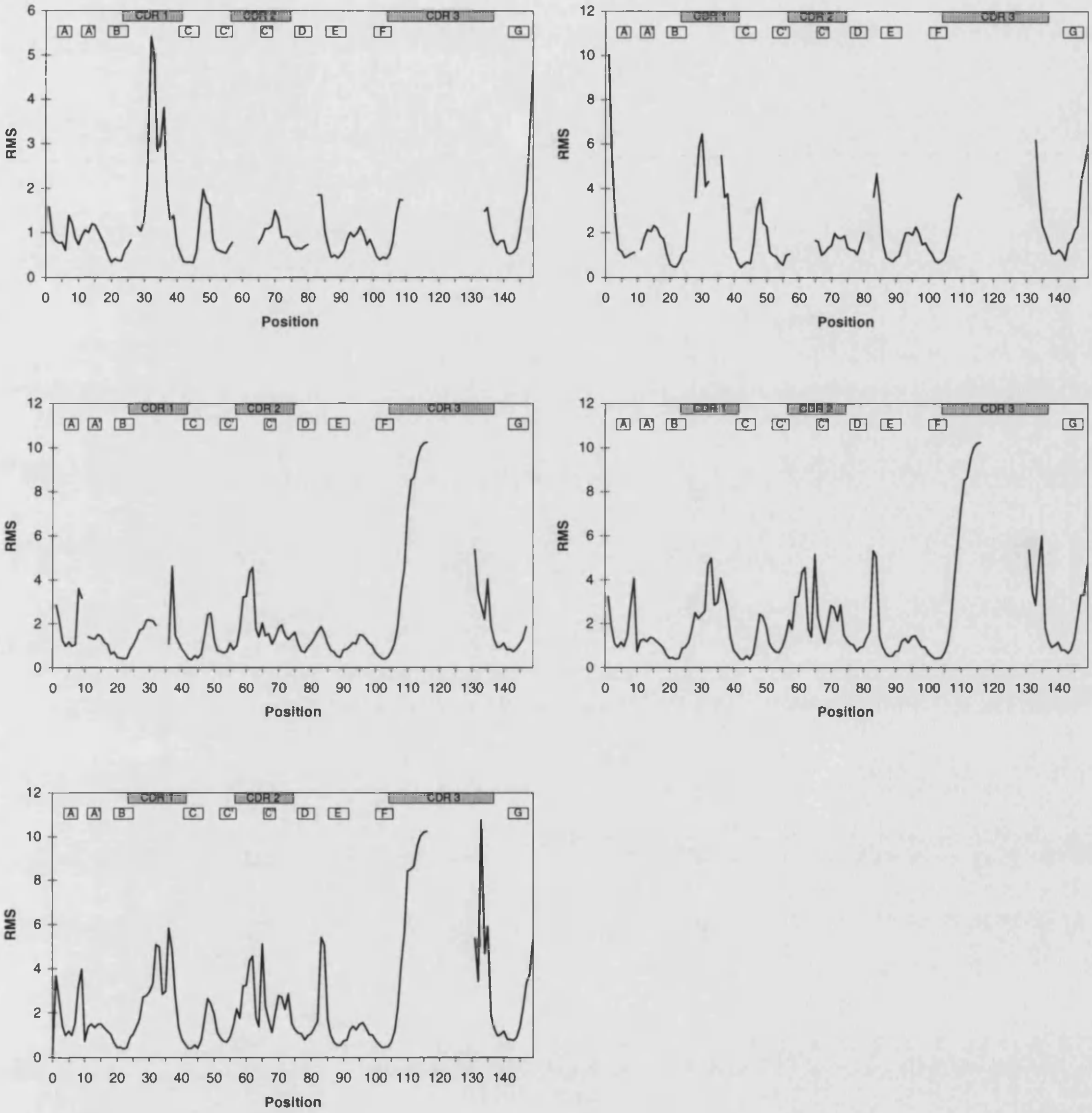


Figure 2.4: RMS plots for antibody κ chains(top left), λ chains (top right), heavy chains(middle left), κ and heavy chains (middle right) and all antibody chains (bottom left) are shown. The CDR equivalent regions of the sequence are marked in black above the plot. The conserved β strand positions are indicated by white boxes above the plot. The numbering is as in figure 2.2. The structures used were those in table 1.2. They were multiply fitted on conserved residues in strands B and C using MULFIT [136].

D strand shows fairly low variability, whereas the D loop and E strand show much more variability than any antibody chain type. Also, in the α chain, the D strand shows more variability than the E strand whereas in all other chain types it appears to be reversed. In the F strand region the variability is similar for all chain types. In both species there is a large amount of variability in the G strand of the α chain. This is not seen in any of the antibody chain types.

2.3.4 Lengths of CDRs

Plots showing the CDR lengths for antibody κ, λ and heavy chains and TCR α and β chains are shown in figures 2.7-2.14.

2.3.4.1 CDR1

In both human and mouse the distribution of TCR β chains is limited to lengths 10 and 11, with the majority (90%) being of length 10. Also in both species the most common TCR α chain length is 11 (90% of TCR α chains). There is a slight difference in the range, however. In humans it is 10 - 12, whereas in mouse there are also some chains with length 13. In both species the κ chains have two peaks, one at 11 and one at 16. The lower length group shows the greatest similarity of the antibody chain types to the TCR α chain apart from the 13 residue chains present in mouse. This length is very rarely seen in κ chains. The higher length group of κ chains represents a set of sequences which have a bulge in the side of the loop (see figure 4.3).

In both mouse and human the TCR β chain peak at 10 residues represents a

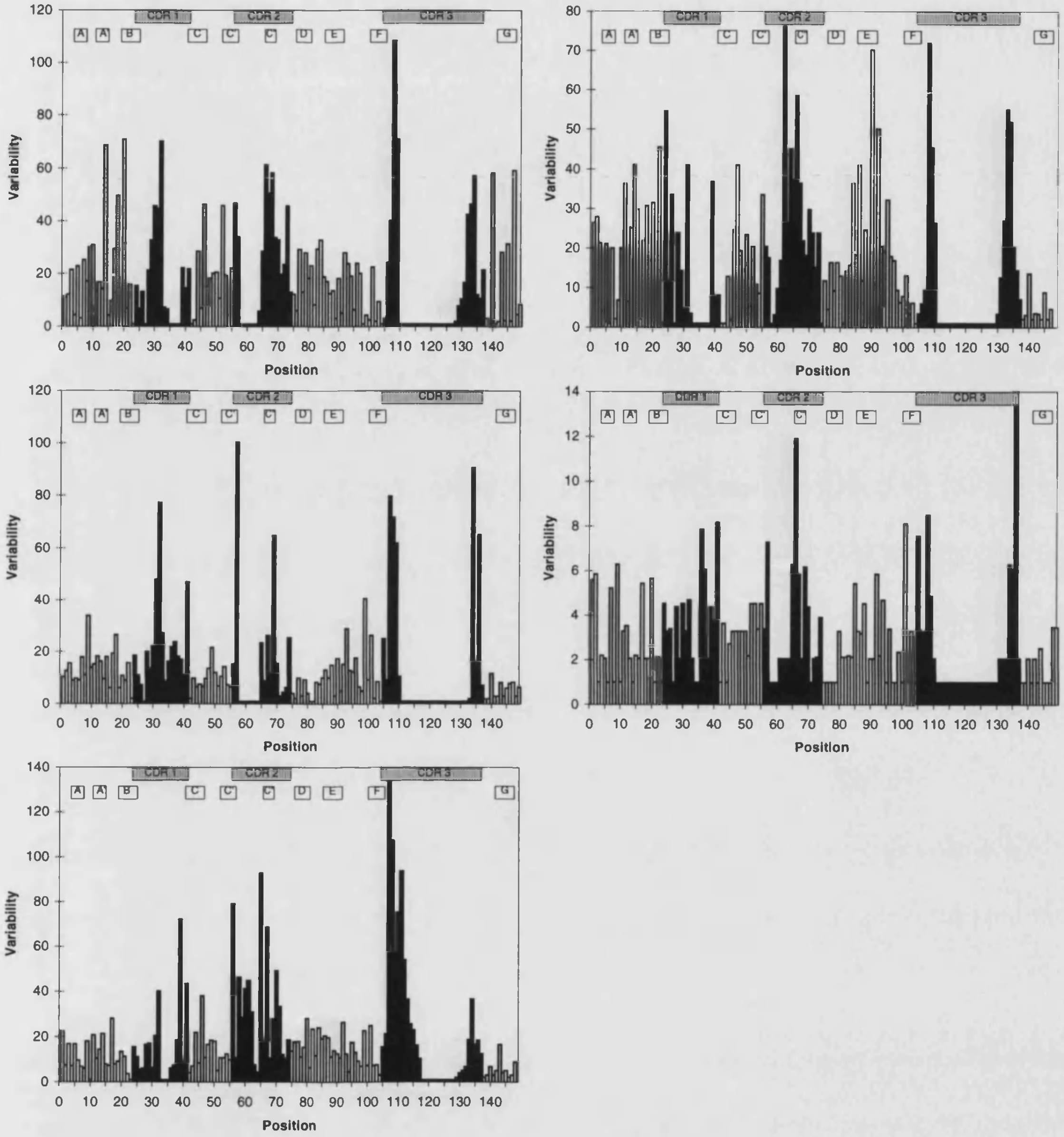


Figure 2.5: Variability plots for Mouse TCR α (top left) and β chains (top right), and antibody κ (middle left), λ (middle right) and γ (bottom left) chains are shown. The CDR equivalent regions of the sequence are marked in black above the plot. The conserved β strand positions are indicated by white boxes above the plot. The numbering is as in figure 2.2.

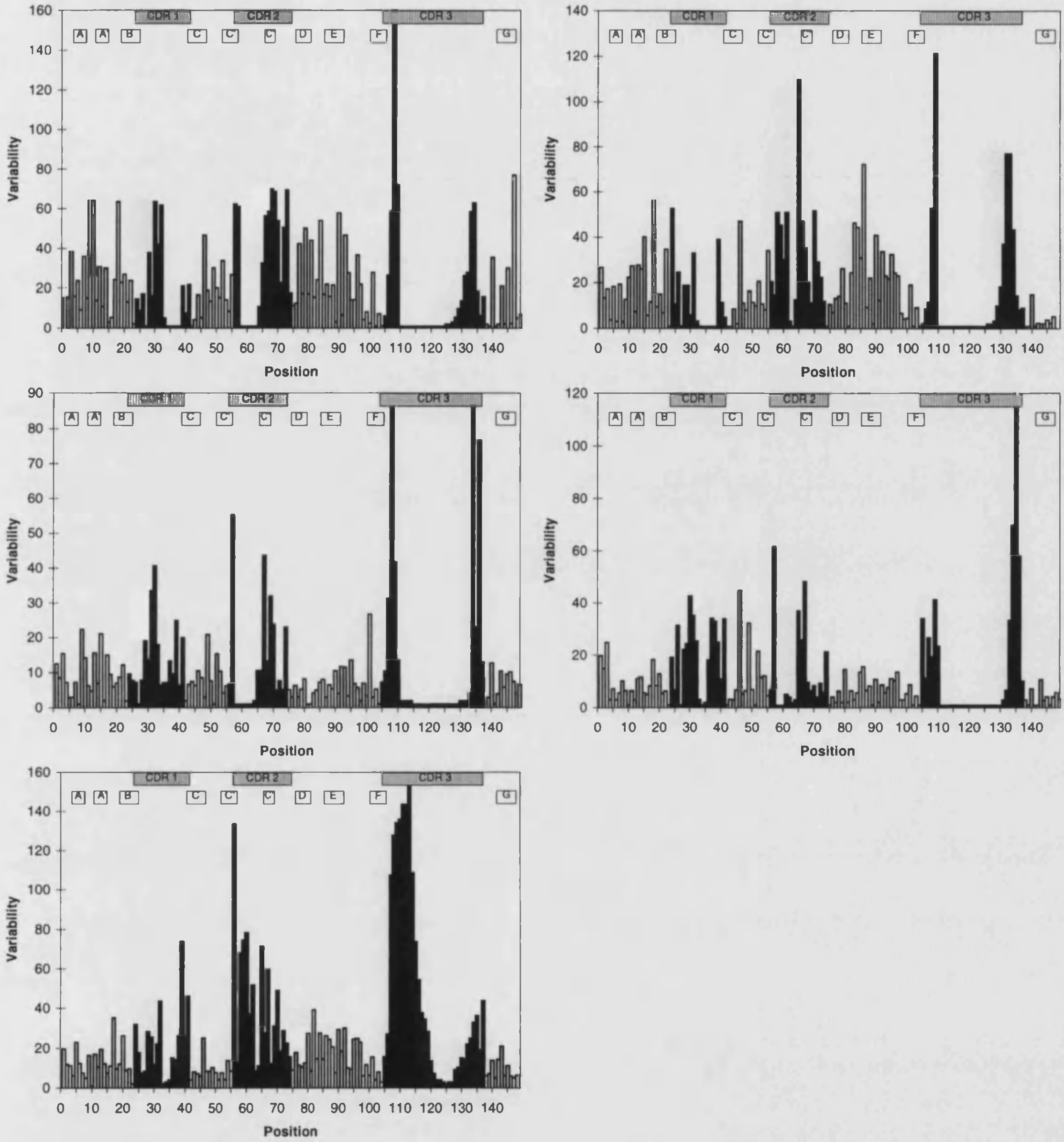


Figure 2.6: Variability plots for Human TCR α (top left) and β chains (top right), and antibody κ (middle left), λ (middle right) and γ (bottom left) chains are shown. The CDR equivalent regions of the sequence are marked in black above the plot. The conserved β strand positions are indicated by white boxes above the plot. The numbering is as in figure 2.2.

similar percentage to the antibody heavy chain peak at 13 residues.

2.3.4.2 CDR2

The length distribution for the TCR β chain is very similar in both mouse and human having a range of 13 - 16 and a modal value of 14 residues. The distribution of α chains is also similar in mouse and human, apart from some shorter (7 and 8 residue) chains in human. The distribution in both α and β is very different to the antibody light chains where virtually all the sequences are the same length (11 residues) and belong to the same canonical class (having the same backbone conformation).

In α chains the modal value (11 residues) is the same length as in κ. However the broader range of the α chain distribution indicates that it may have a different structure. The modal value for the β chains is different (14 residues) to either the κ and λ (11) or heavy chains (16) but the shape of the distribution shows some similarities to the heavy chain.

2.3.4.3 CDR3

The distribution of both antibody heavy and TCR chains is broader in human than in mouse. A possible explanation for this is that the majority of sequences for mouse come from inbred strains, whereas the human sequences come from the general, outbred population. In both species the minimum TCR β length is 8 residues, but in mouse the range extends to 15 residues whereas in humans it extends to 18. Again for the TCR α lengths, both species show the same minimum length of 4 residues but with the maximum for mouse being 15 and that for human 18 (although this is quite rare). The modal value for mouse is 12 and for human is 10.

The κ chains have a modal value of 9 residues in both mouse and humans, most chains being of this length although the range is 4 - 11 in human and 6 - 12 in mouse. The human λ chain distribution shows more similarity to the TCR chain than any of the other chain types, although the range is more restricted (from 9 - 13 with one outlier of 6 residues). The distribution of TCRs is more restricted than the antibody heavy chain distribution (5 - 22 mouse, 5-33 human). The CDR3 length plots for the eliminated set confirmed these results, with just slight changes to some of the frequencies.

The combined length of TCR α and β chain CDR3s in the known pairs is shown in table 2.4 and in figure 2.15. In all cases at least one of the chains always has a CDR 3 of nine or more residues. This may indicate that this is the minimum length required to reach the peptide in the MHC groove. There is a narrower distribution of total length than in κ/heavy antibody pairs and the same range is in λ/heavy pairs (18-34 α/β, 13-39 κ/heavy, 18-34 λ/heavy).

2.3.5 Sequence Similarity Plots

The similarity plots for mouse and human showed very similar results except for the λ chain which has a small repertoire in the mouse. The results are shown in figures 2.16–2.27 and are described below for each.

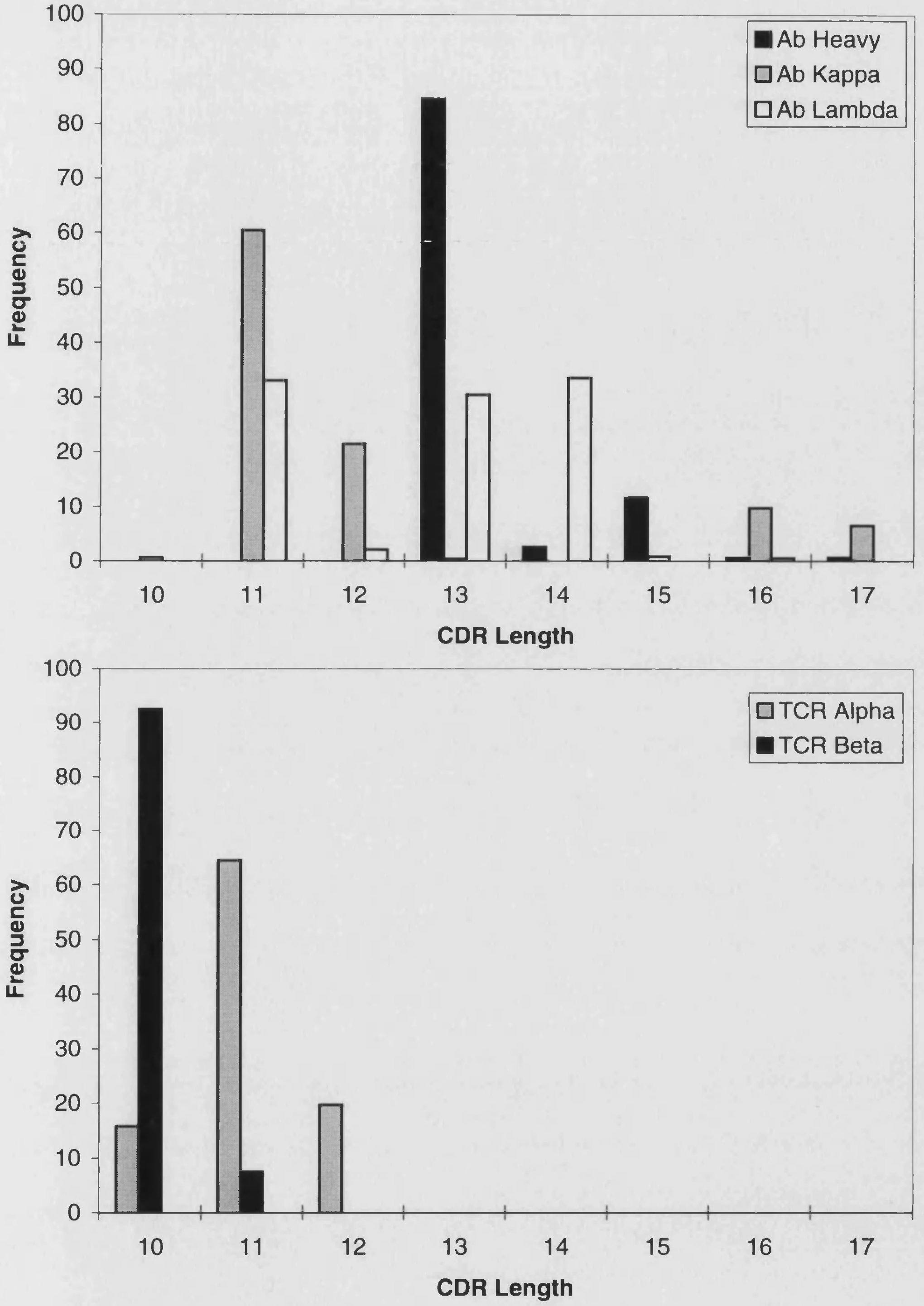


Figure 2.7: Length frequencies of the CDR 1 regions of human TCR α and β chains, and human antibody κ, λ and heavy chains are shown.

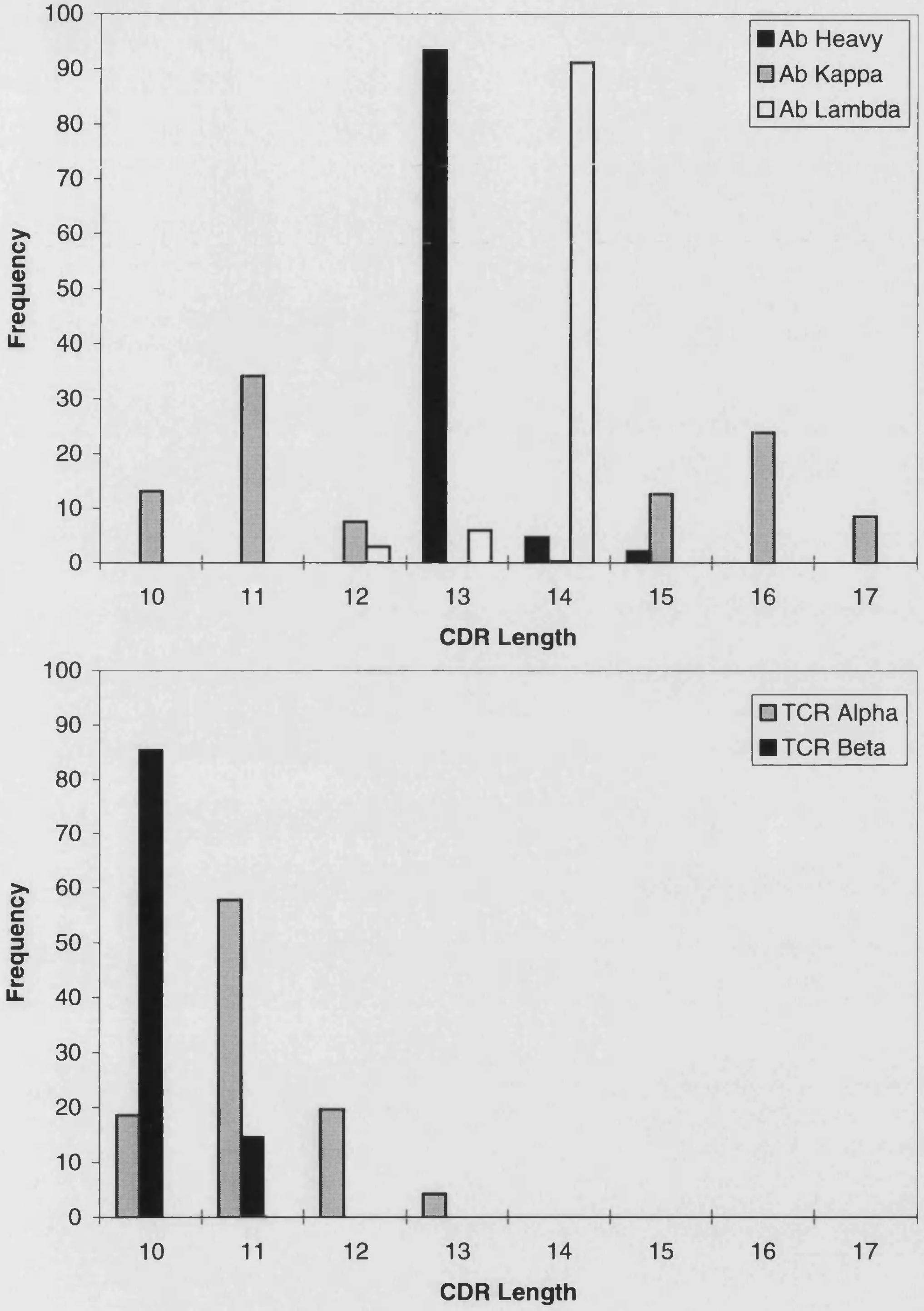


Figure 2.8: Length frequencies of the CDR 1 regions of mouse TCR α and β chains, and mouse antibody κ, λ and heavy chains are shown.

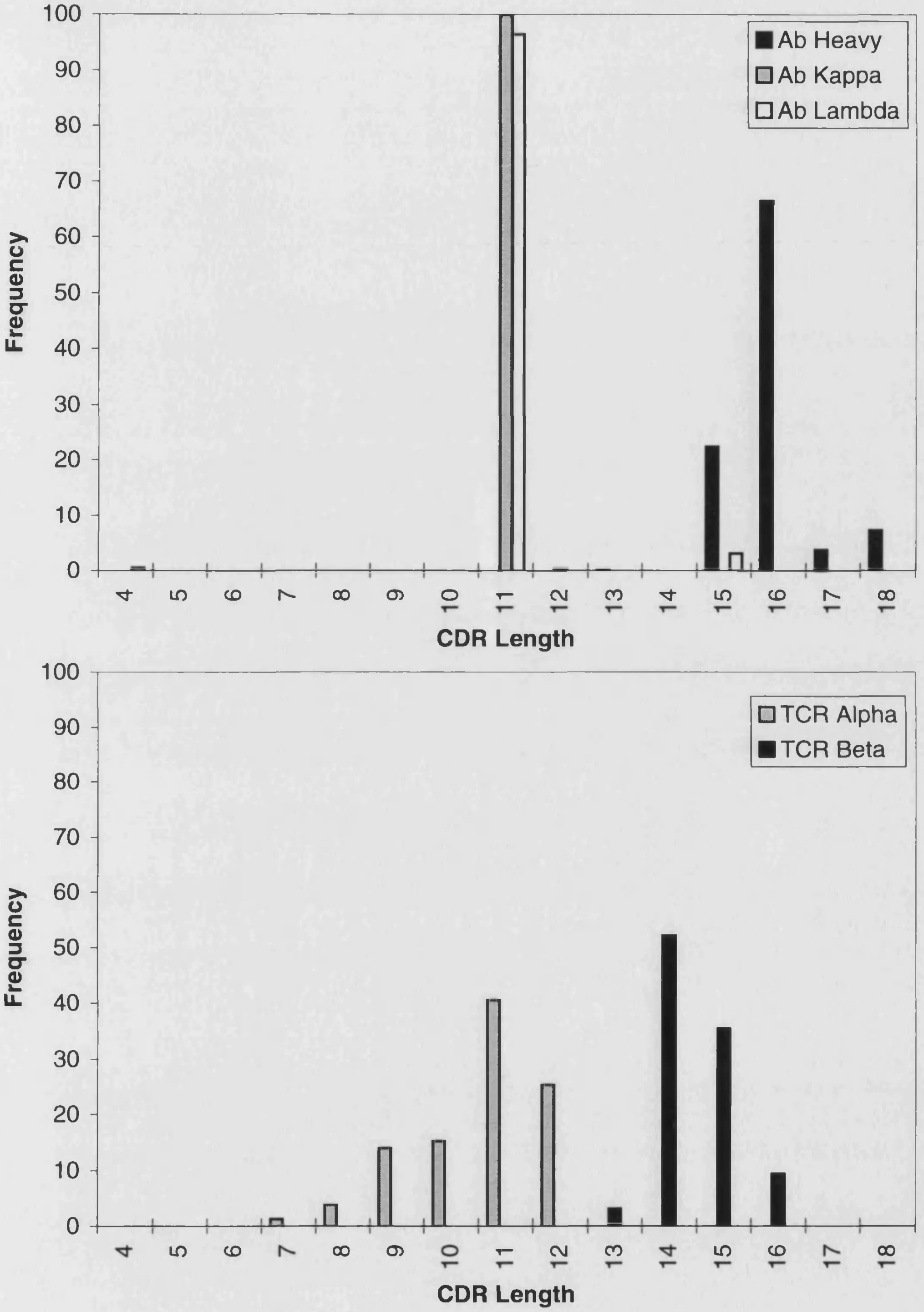


Figure 2.9: Length frequencies of the CDR 2 regions of human TCR α and β chains, and human antibody κ, λ and heavy chains are shown.

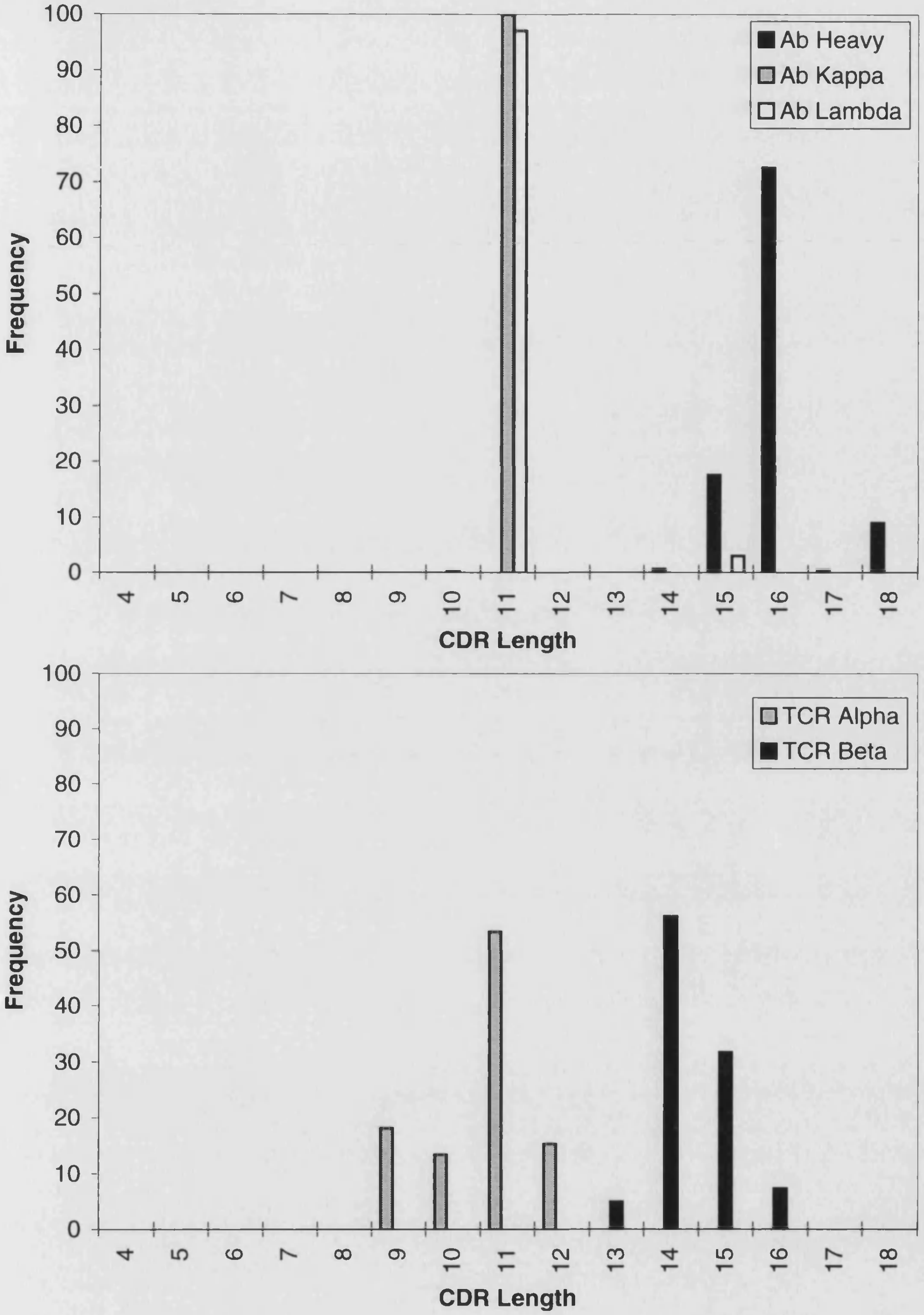


Figure 2.10: Length frequencies of the CDR 2 regions of mouse TCR α and β chains, and mouse antibody κ, λ and heavy chains are shown.

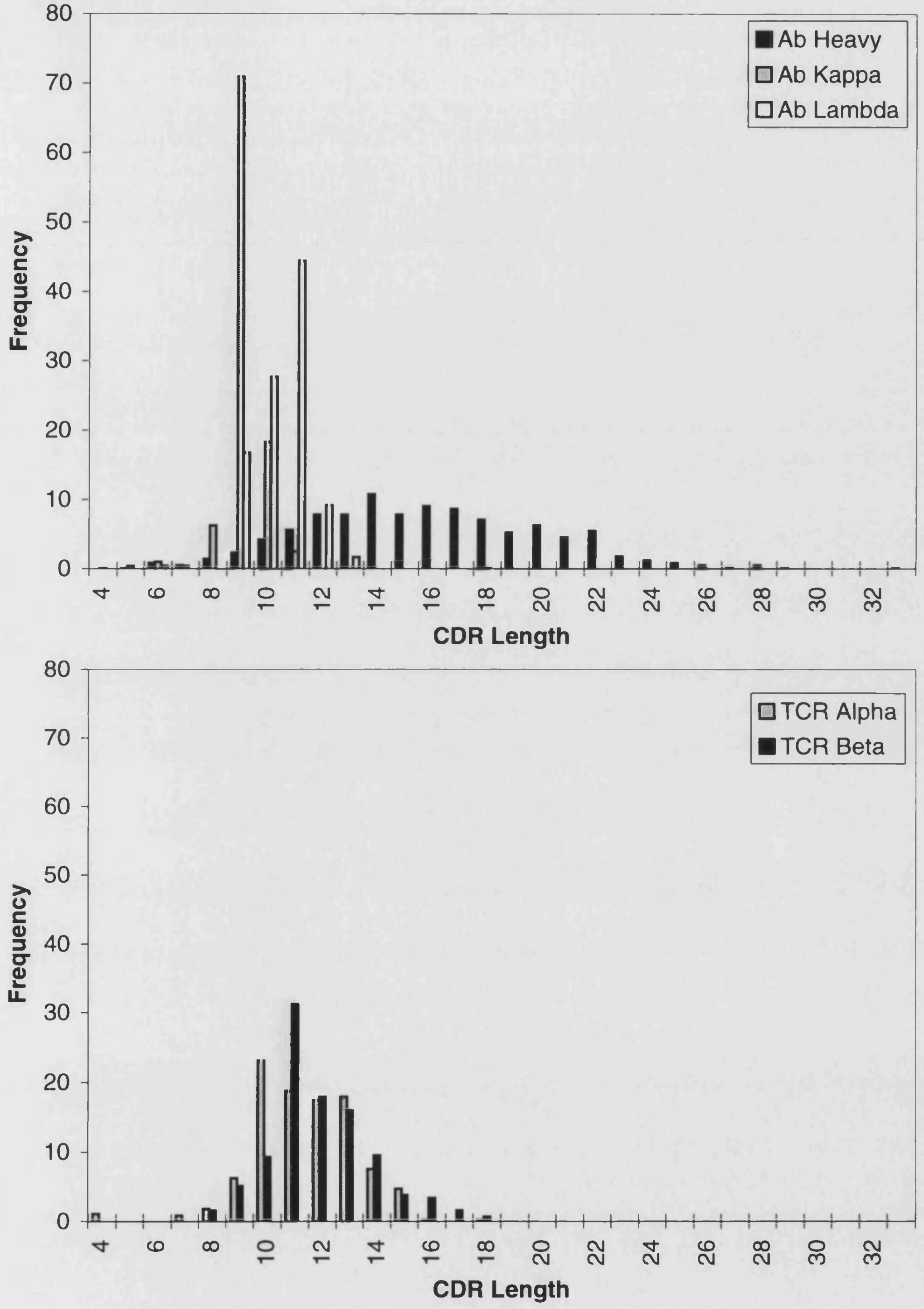


Figure 2.11: Length frequencies of the CDR 3 regions of human TCR α and β chains, and human antibody κ, λ and heavy chains are shown.

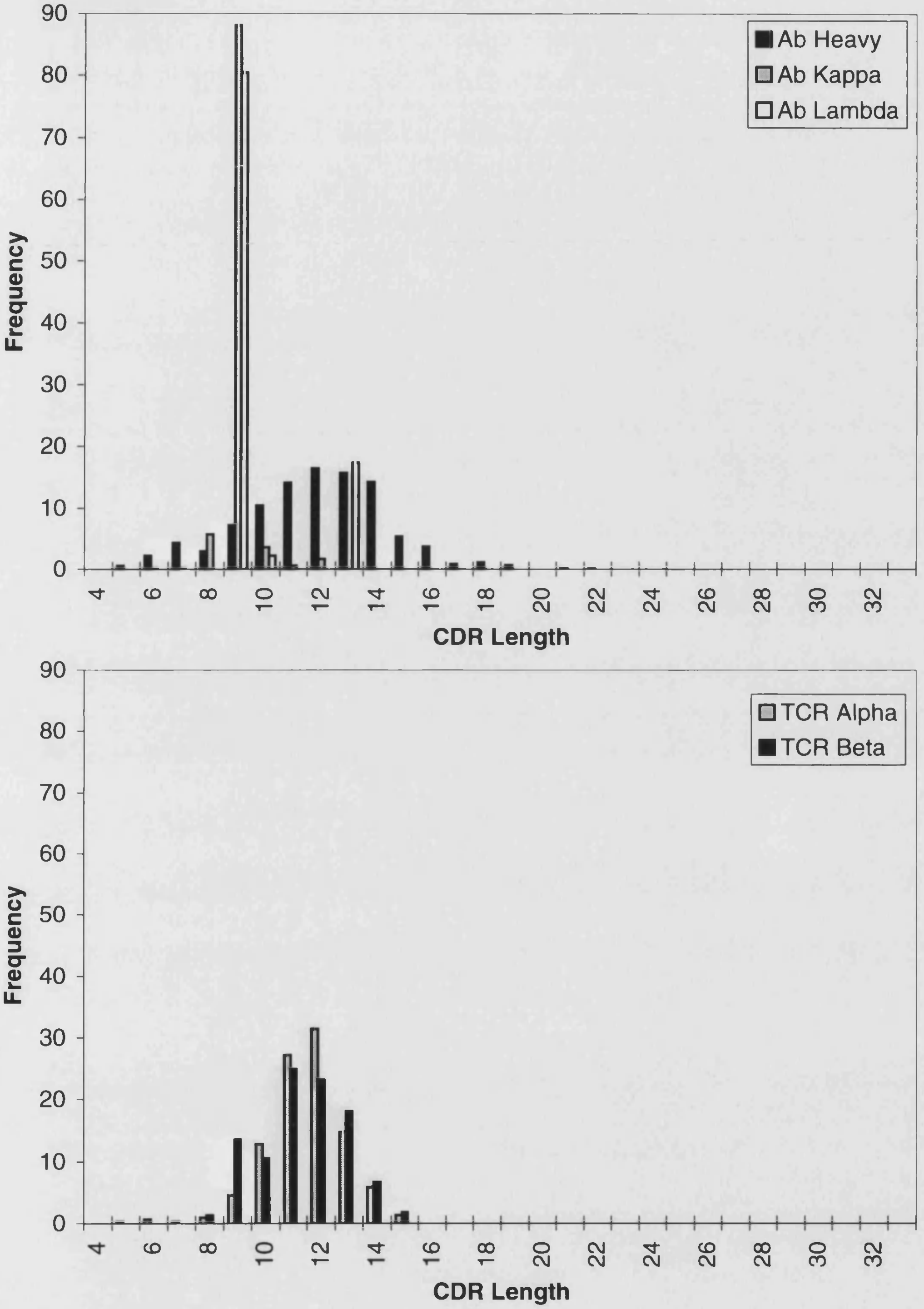


Figure 2.12: Length frequencies of the CDR 3 regions of mouse TCR α and β chains, and mouse antibody κ, λ and heavy chains are shown.

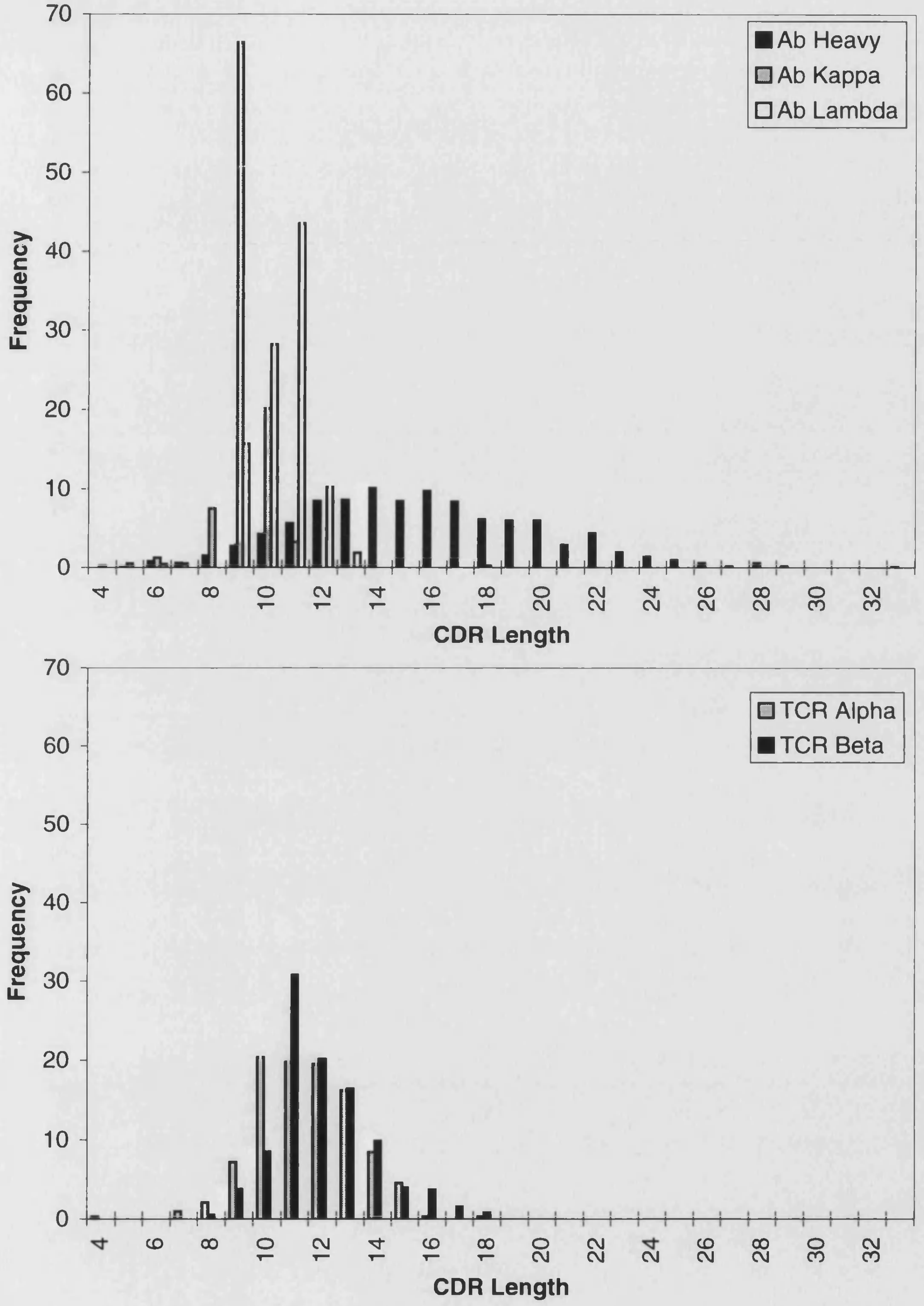


Figure 2.13: Length frequencies of the CDR 3 regions of human TCR α and β chains, and human antibody κ, λ and heavy chains, after elimination of identical sequences, are shown.

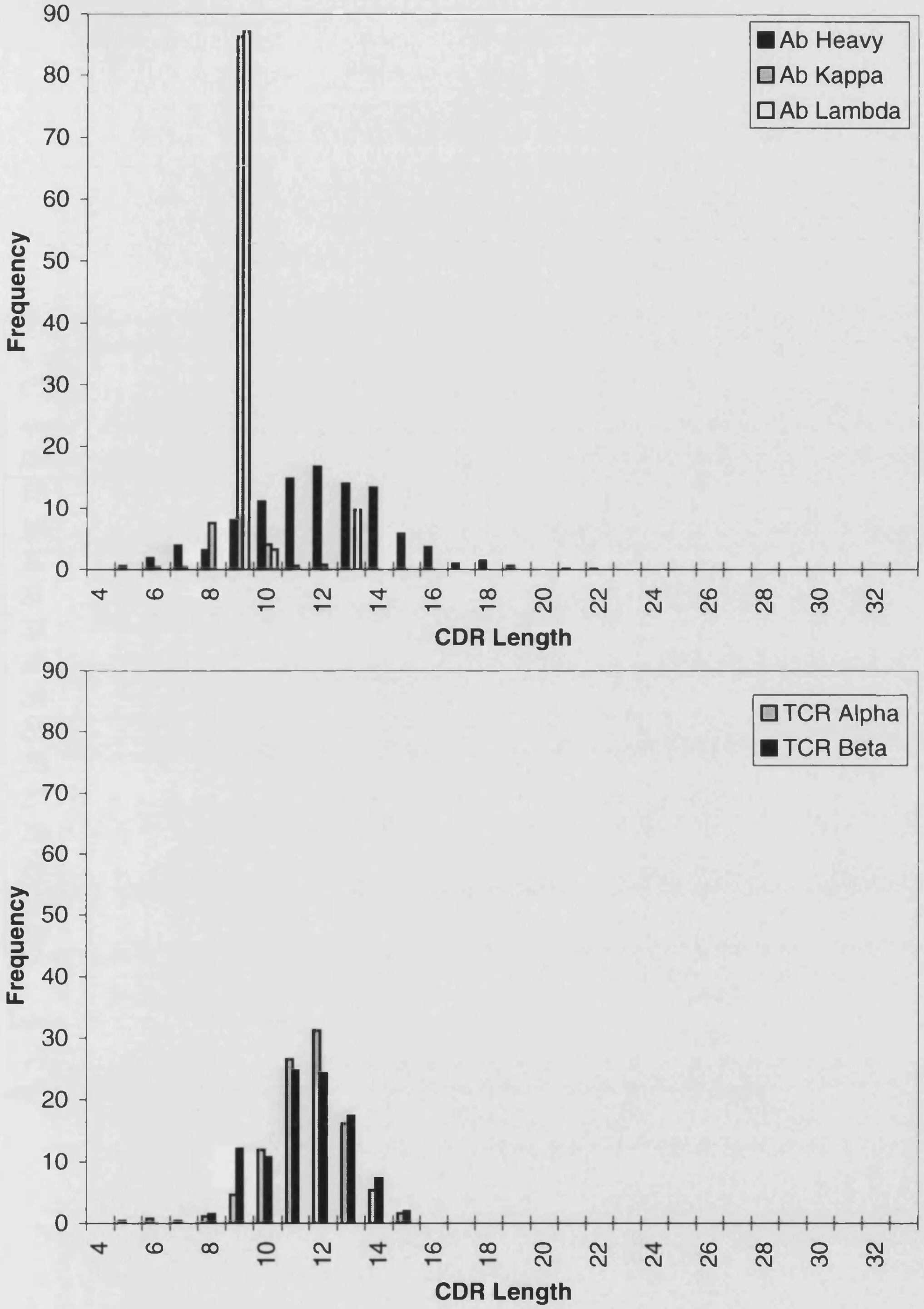


Figure 2.14: Length frequencies of the CDR 3 regions of mouse TCR α and β chains, and mouse antibody κ, λ and heavy chains, after elimination of identical sequences, are shown.

Combined Length of CDR $\alpha 3$ and CDR $\beta 3$	Lengths of loops (α 3 first)					
18	6 12 (1)	9 9 (2)	10 8 (8)			
19	5 14 (1)	7 12 (1)	8 11 (2)	9 10 (3)	10 9 (1)	11 8 (1)
20	9 11 (2)	10 10 (9)	11 9 (5)			
21	9 12 (1)	10 11 (11)	11 10 (4)	12 9 (17)		
22	9 13 (7)	10 12 (8)	11 11 (29)	12 10 (2)	13 9 (12)	
23	8 15 (1)	9 14 (1)	10 13 (8)	11 12 (23)	12 11 (20)	13 10 (5)
24	9 15 (3)	10 14 (6)	11 13 (15)	12 12 (19)	13 11 (13)	
25	10 15 (1)	11 14 (7)	12 13 (21)	13 12 (9)	14 11 (6)	
26	11 15 (1)	12 14 (5)	13 13 (11)	14 12 (5)	15 11 (3)	
27	10 17 (1)	11 16 (4)	13 14 (4)	14 13 (4)	15 12 (1)	16 11 (1)
28	12 16 (1)	13 15 (2)	14 14 (1)	15 13 (1)		
29	13 16 (1)	14 15 (1)	15 14 (2)			
30	13 17 (1)					
34	18 16 (1)					

Table 2.4: The combined length of CDR $\alpha 3$ and CDR $\beta 3$ for the sequences of mouse and human clones containing the CDR 3 regions of both α and β chains in the Kabat database (a total of 336 entries). The numbers in parentheses are the number of occurrences of the combination of lengths.

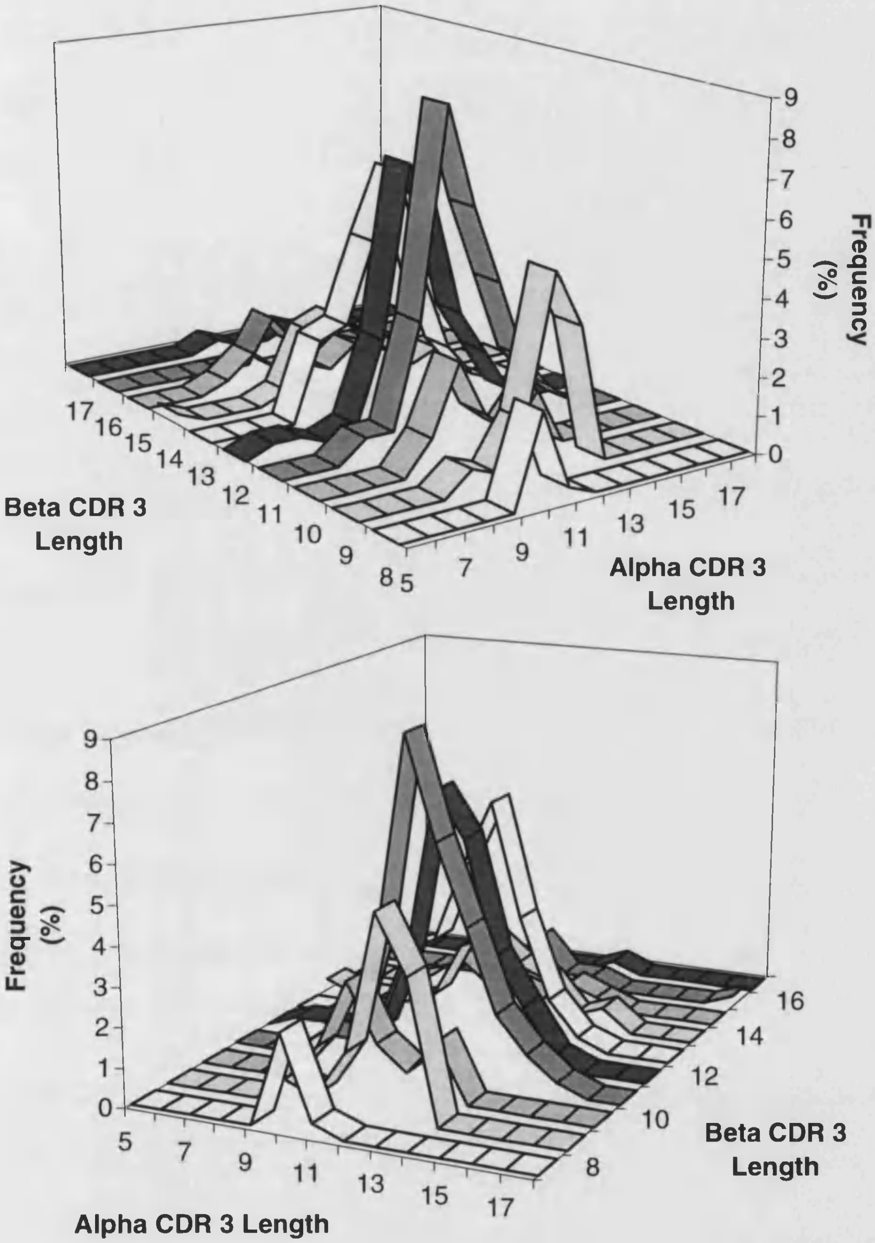


Figure 2.15: Frequencies of the CDR regions of human TCR α and β chains CDR, and human antibody κ, λ and heavy chains are shown.

2.3.5.1 TCR α chain

In the A and the A' strands the α chains shows a pattern of residue types more similar to that of the κ chains than either the λ or the heavy chains. They are least similar to the heavy chains.

In the B strand the differences are not large but the α chain again appears to be more like κ than λ or heavy.

In the C strand, where the α chains contain the conserved tyrosine (92% of sequences) at position 43, there is a large single residue peak favouring the light chains (this position is valine (75%) or isoleucine (22%) in heavy chains). Other than that there is very little discrimination in this strand.

In the C' strand there is a definite discernible peak in favour of both light chain types. It is not possible to determine which is the most similar as the κ/λ plot shows little discrimination in this strand.

In the D strand the α chain looks more like the heavy chain whereas in the E strand the residue pattern is again more similar to that of the light chains.

There is high conservation of every second residue in all chain types in the regions either side of the CDR3 (the F strand and the region before the G strand), consistent with the known strand interactions in antibodies. At residue 138 light chains almost always have a phenylalanine whereas the heavy chains have a tryptophan. In the α chain a phenylalaline also occurs at this position. This can be seen on the plots as a single residue, light chain favouring, peak at 138.

The G strand shows some heavy chain and also some λ chain character.

2.3.5.2 TCR β chains

Overall the β chain plots show fewer discernible peaks in all regions. This indicates that the β chains show more equal similarity or dissimilarity to both light and heavy chains.

In the A strand the β chains are again more like κ chains than heavy or λ. However in the A' strand the β chains show more similarity to the heavy chains, although it is quite different to any antibody chain type. This region is close to the G strand which also shows some heavy chain character.

In the loop between A' and B the β chain definitely appears to have most similarity to λ but also shows some similarity to κ. In the B strand itself it is not possible to see any obvious preference although the differences are small.

In the C strand, where the β chain sequences contain the same tyrosine as the α and light chain, the same single residue peak occurs at position 43. The overall residue pattern has greater heavy chain character. However the dissimilarity plots show that this region has one of the lowest difference values, showing greater conservation between the different chain types.

In the C' strand the β chain similarity plots are more similar to light chains although it is not possible to determine between λ and κ.

In the D and E strands it is difficult to determine any discrimination. The E strand is close to the B, the D, E, B and A together forming a β sheet, so they would be expected to have similar properties.

As in the α chain the regions either side of the CDR3 are conserved.

Comparison	Chain Type 1	Chain Type 2
Human α		
κ/heavy	-16.66	10.45
λ/heavy	-17.41	10.23
κ/λ	-7.64	8.62
Human β		
κ/heavy	-21.17	12.26
λ/heavy	-21.73	12.46
κ/λ	-9.20	9.55
Mouse α		
κ/heavy	-20.68	12.66
λ/heavy	-16.88	20.73
κ/λ	-20.90	9.03
Mouse β		
κ/heavy	-20.48	13.80
λ/heavy	-14.56	19.37
κ/λ	-20.29	8.80

Table 2.5: The summed similarity score across the alignment to each of the comparison antibody chain types.

Finally in the G strand the β chain is very similar to the λ chain and shows some similarity to the heavy but none, or very little to the κ chain.

Table 2.5 summarises the similarity scores across the entire framework regions of the alignment. It shows that both the TCR α and β chains appear to have significantly more similarity to the light chain types than to the heavy.

2.3.6 Environment scores

Environment score schematics for mouse TCR α and β chains are shown in figures 2.30 and 2.31 respectively. The pattern of homology is similar to that seen for the similarity scores, although some regions are less well defined. The conserved

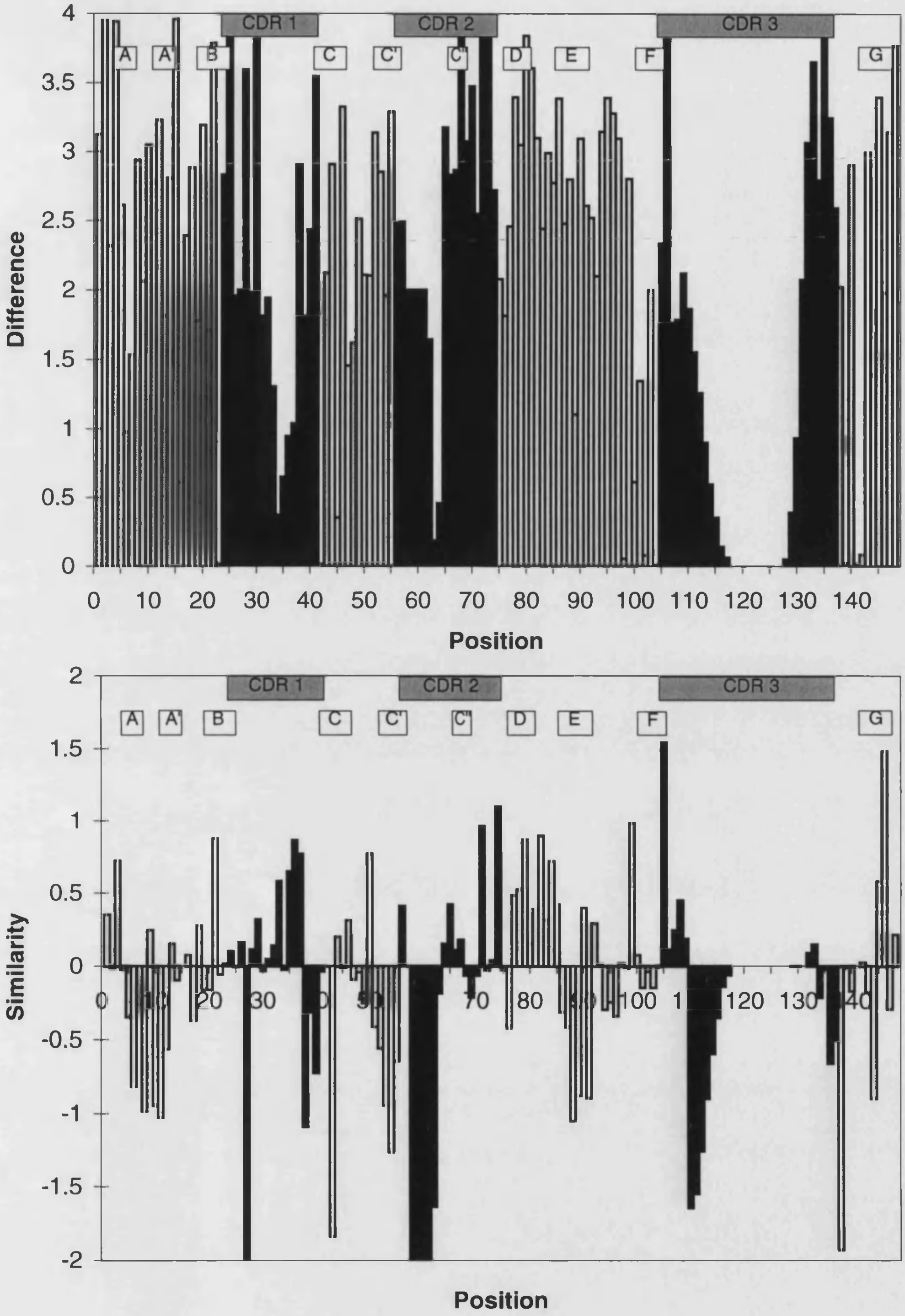


Figure 2.16: Similarity plots for mouse α chains against antibody heavy and κ chains. Negative values indicate greater similarity to antibody κ chains than heavy chains and positive values greater similarity to antibody heavy chains.

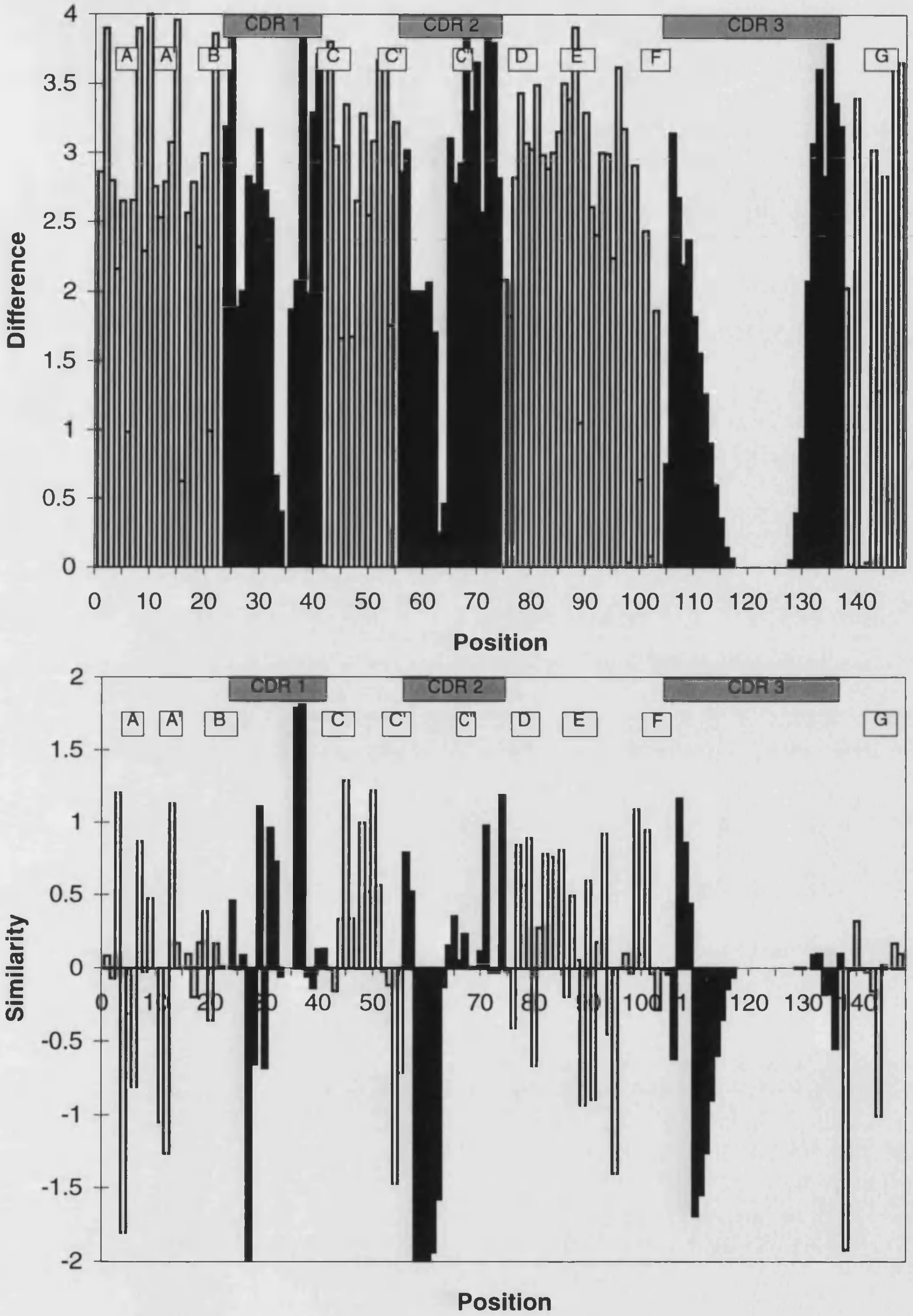


Figure 2.17: Similarity plots for mouse α chains against antibody heavy and λ chains. Negative values indicate greater similarity to antibody λ chains than heavy chains and positive values greater similarity to antibody heavy chains.

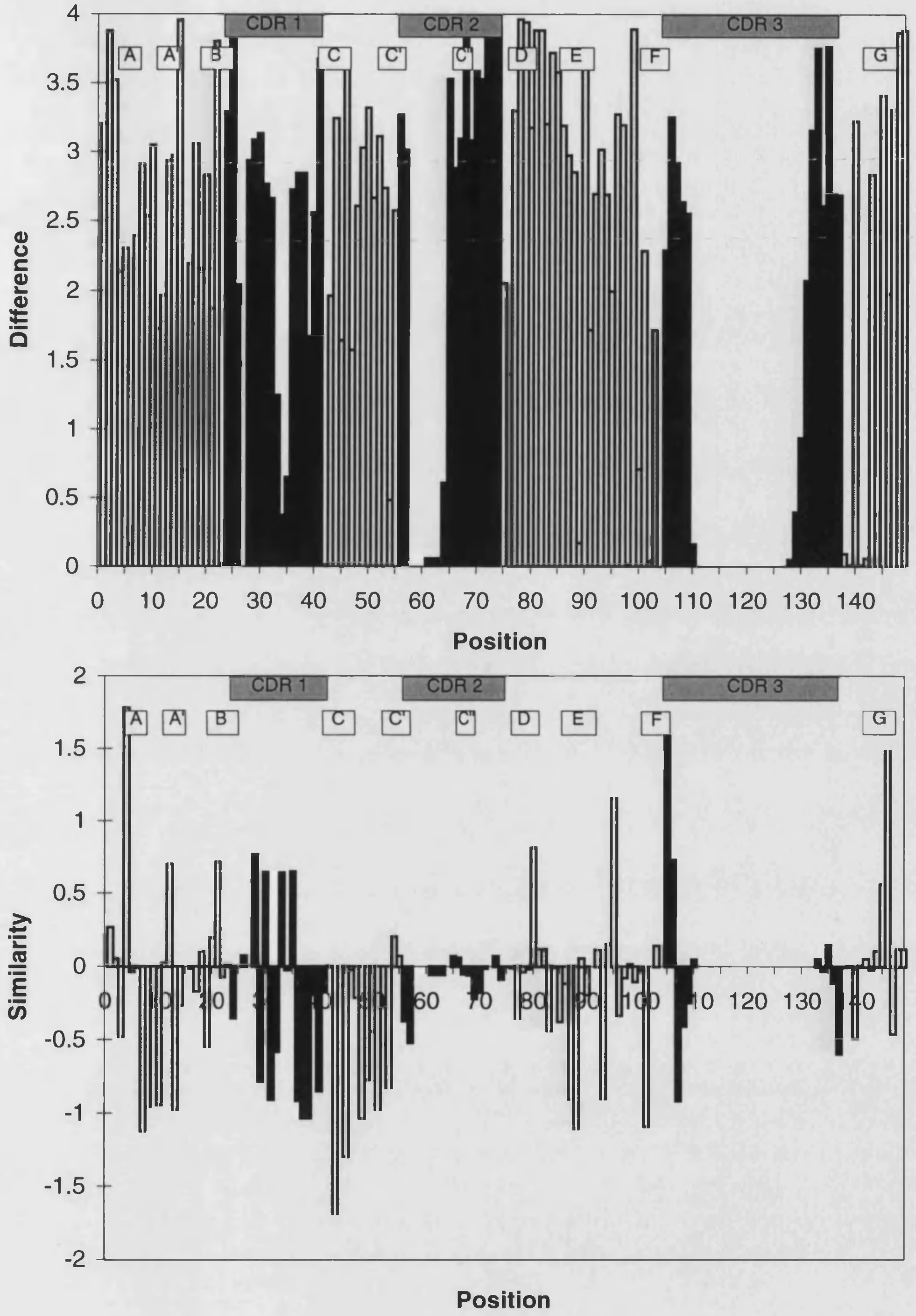


Figure 2.18: Similarity plots for mouse α chains against antibody κ and λ chains. Negative values indicate greater similarity to antibody κ chains than λ chains and positive values greater similarity to antibody λ chains.

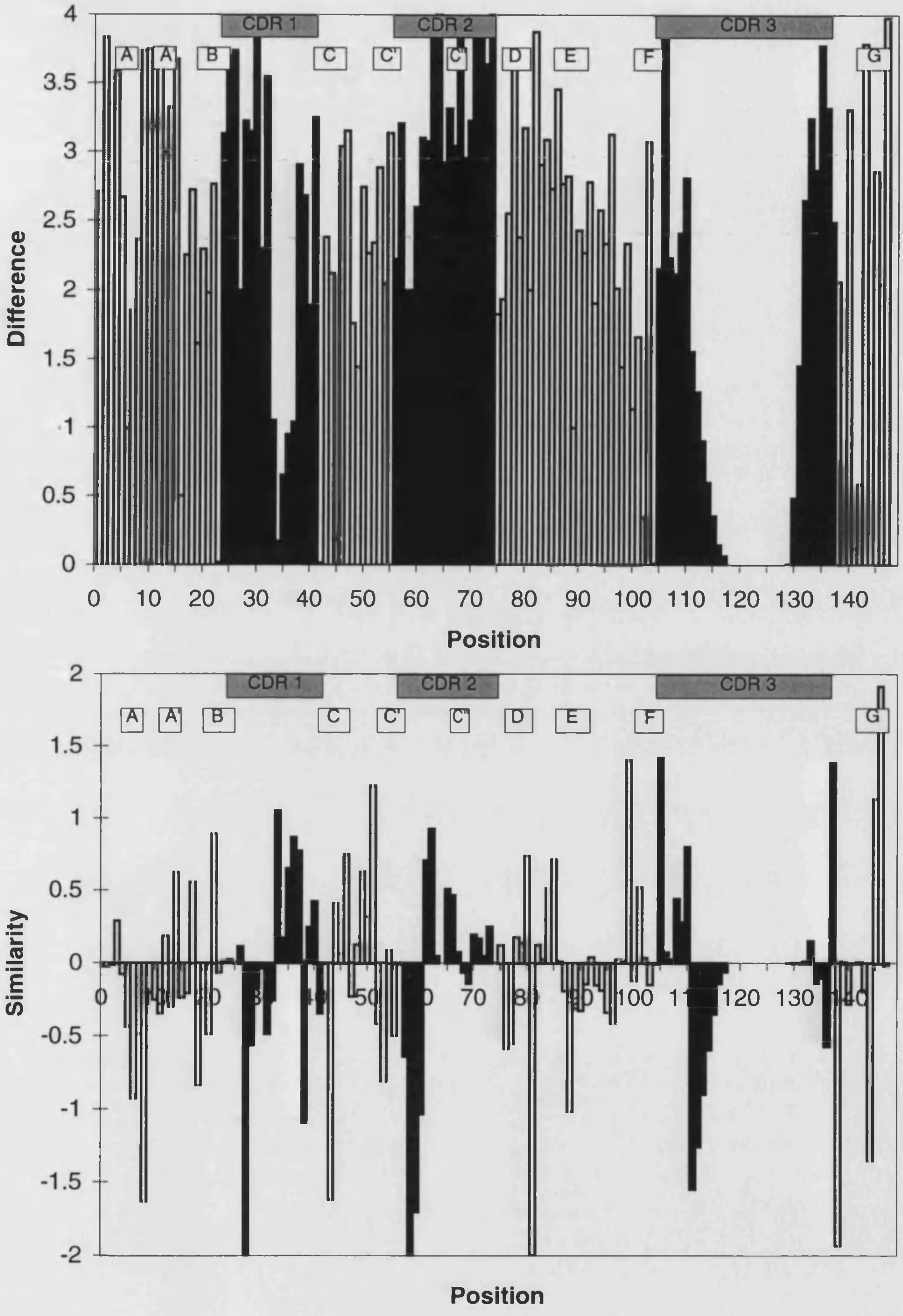


Figure 2.19: Similarity plots for mouse β chains against antibody heavy and κ chains. Negative values indicate greater similarity to antibody κ chains than heavy chains and positive values greater similarity to antibody heavy chains.

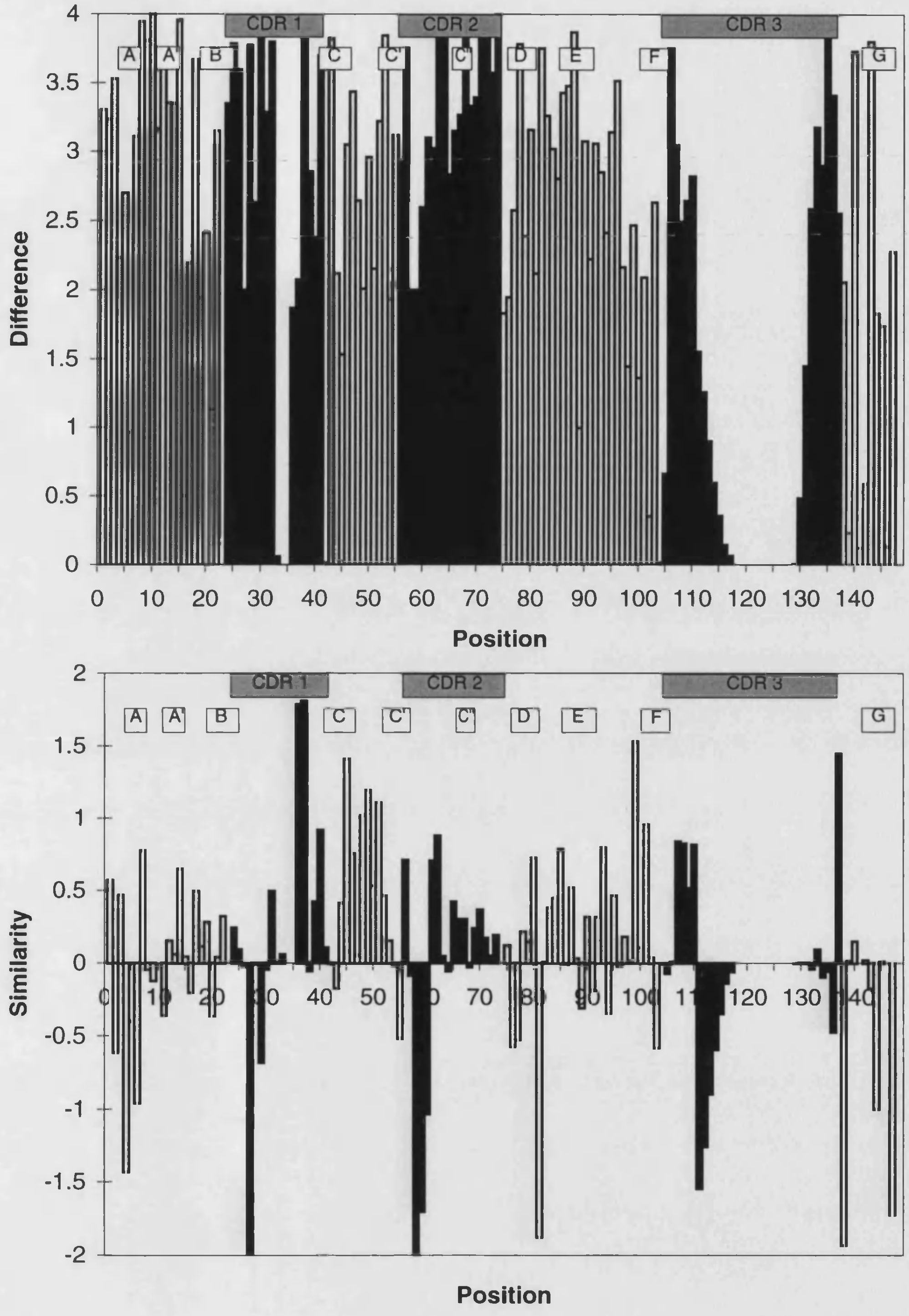


Figure 2.20: Similarity plots for mouse β chains against antibody heavy and λ chains. Negative values indicate greater similarity to antibody λ chains than heavy chains and positive values greater similarity to antibody heavy chains.

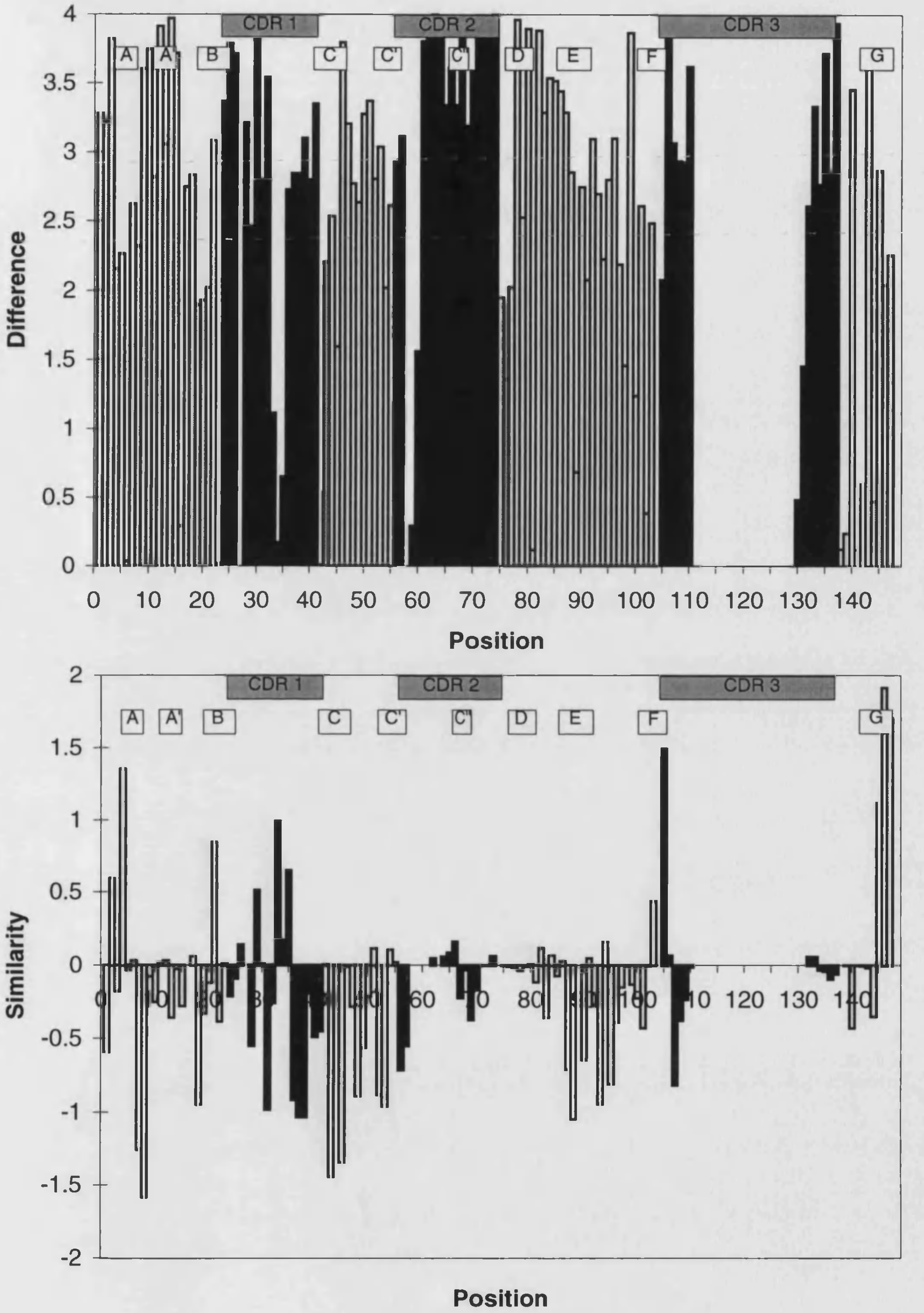


Figure 2.21: Similarity plots for mouse β chains against antibody κ and λ chains. Negative values indicate greater similarity to antibody κ chains than λ chains and positive values greater similarity to antibody λ chains.

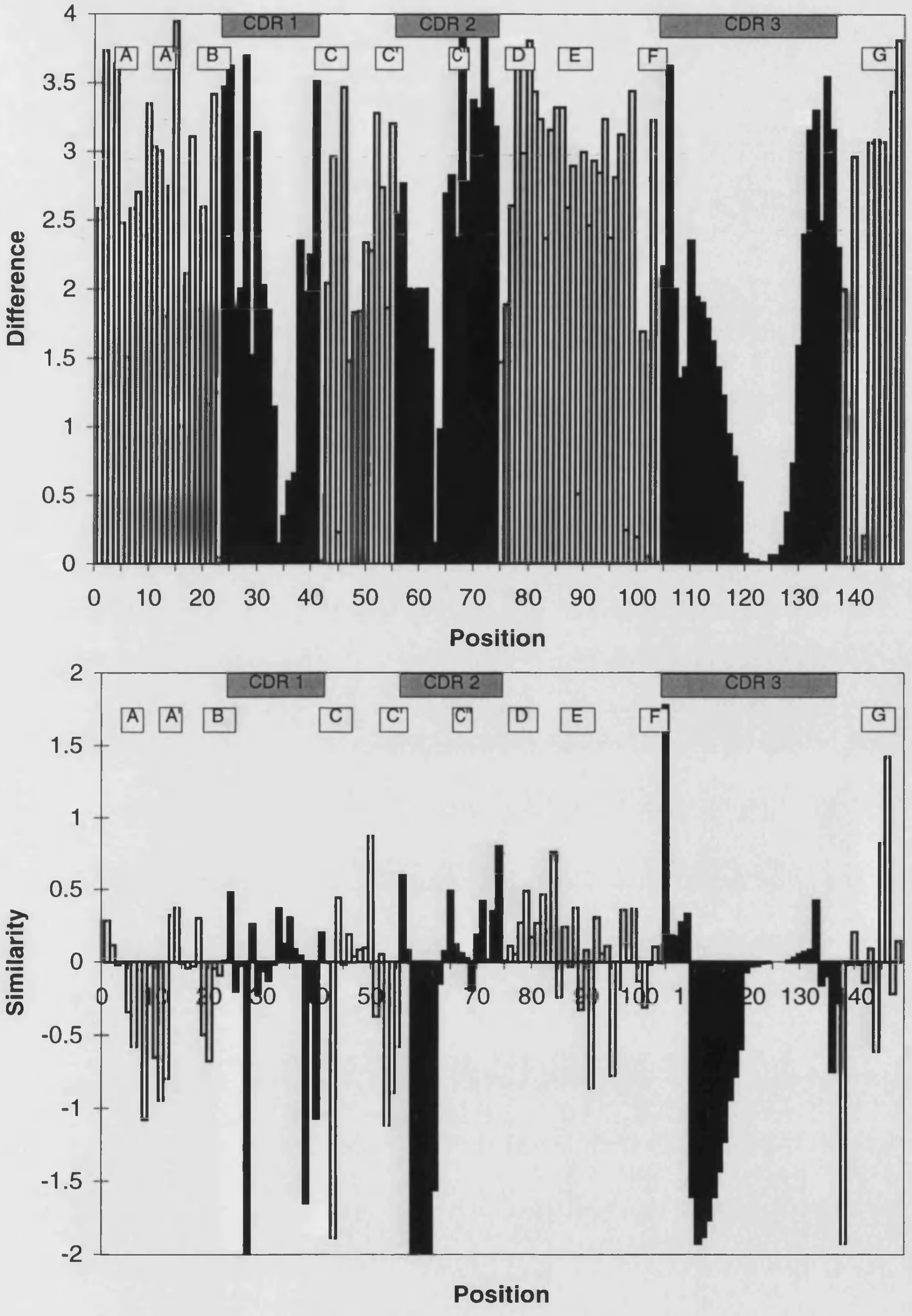


Figure 2.22: Similarity plots for human α chains against antibody heavy and κ chains. Negative values indicate greater similarity to antibody κ chains than heavy chains and positive values greater similarity to antibody heavy chains.

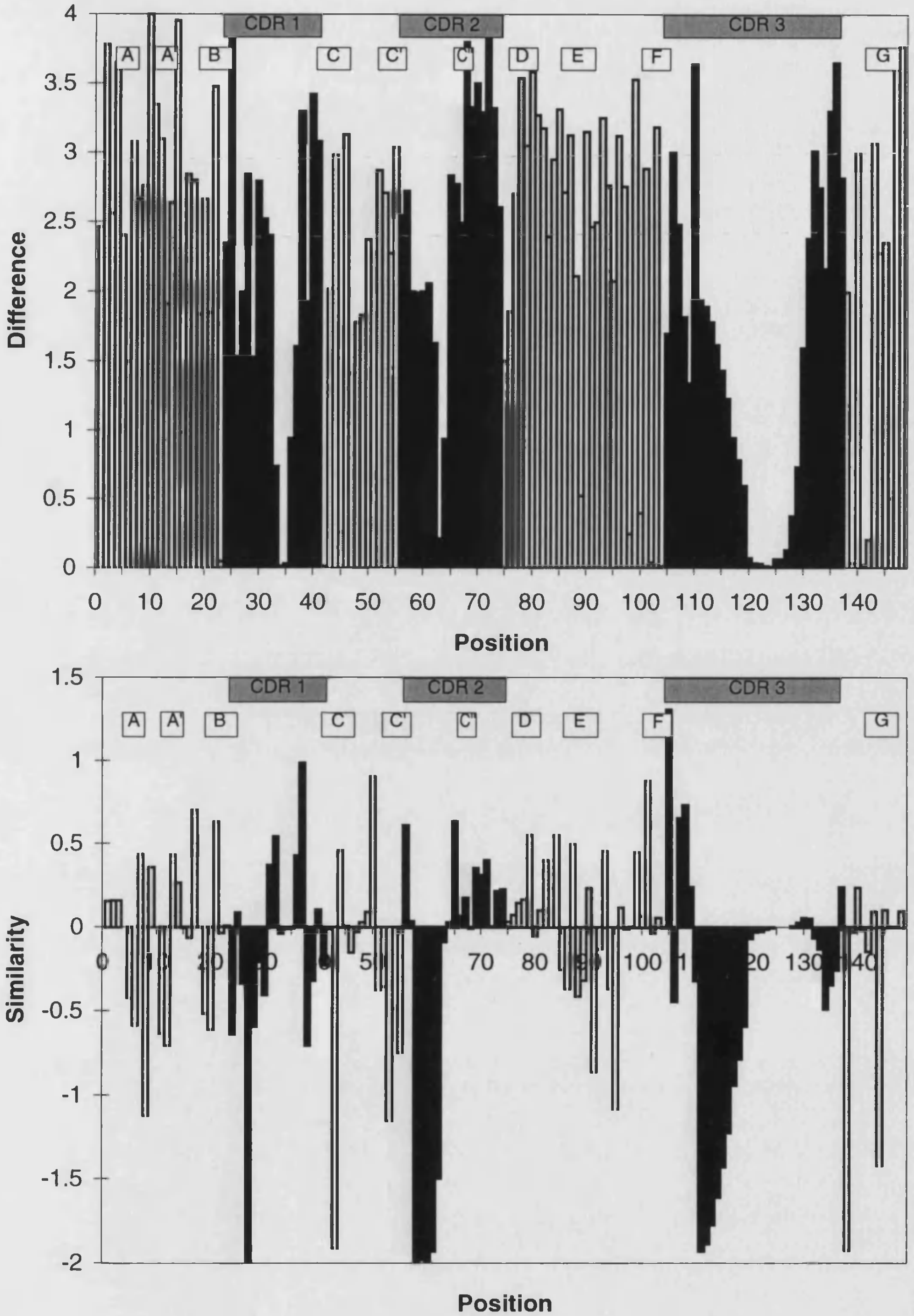


Figure 2.23: Similarity plots for human α chains against antibody heavy and λ chains. Negative values indicate greater similarity to antibody λ chains than heavy chains and positive values greater similarity to antibody heavy chains.

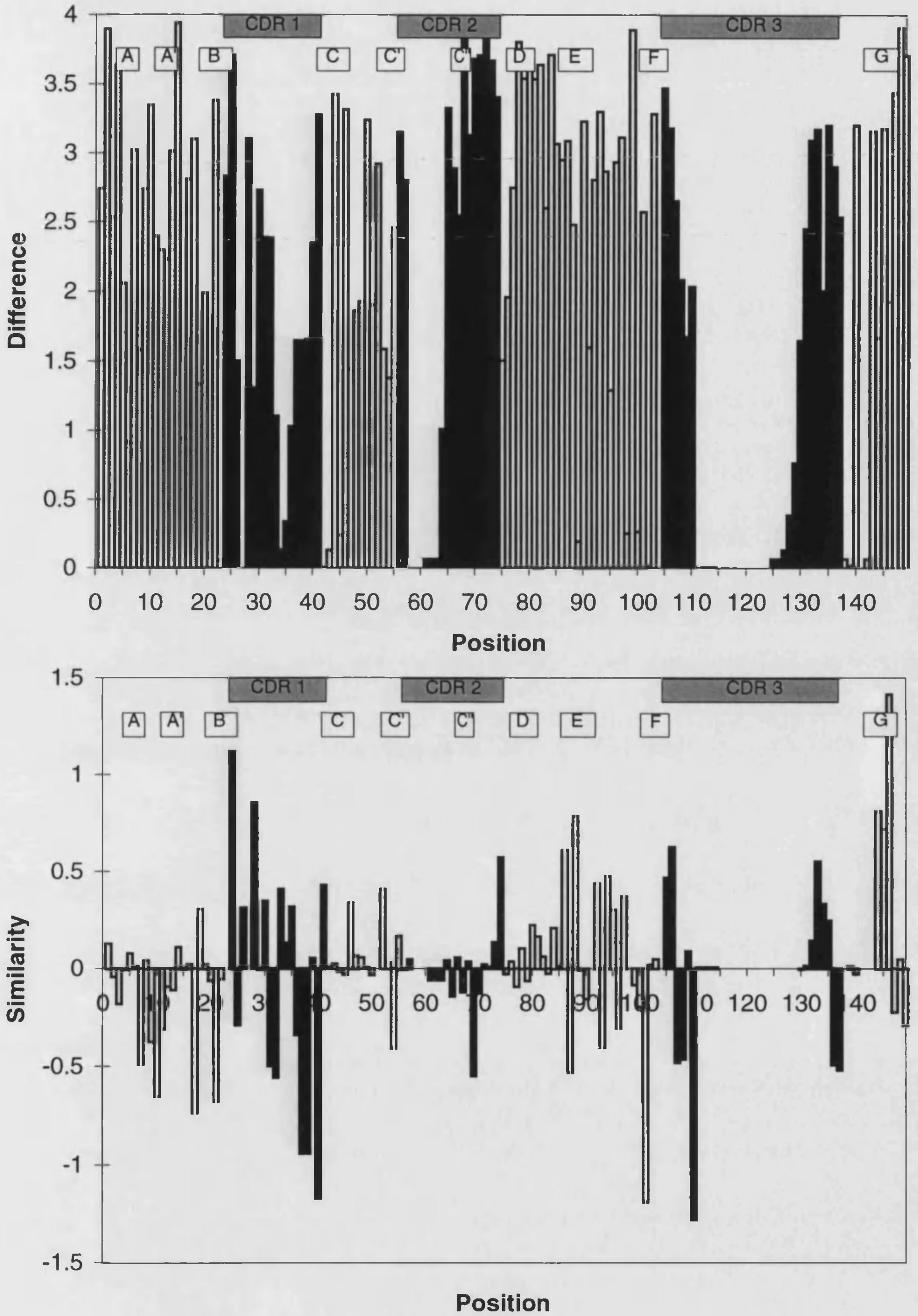


Figure 2.24: Similarity plots for human α chains against antibody κ and λ chains. Negative values indicate greater similarity to antibody κ chains than λ chains and positive values greater similarity to antibody λ chains.

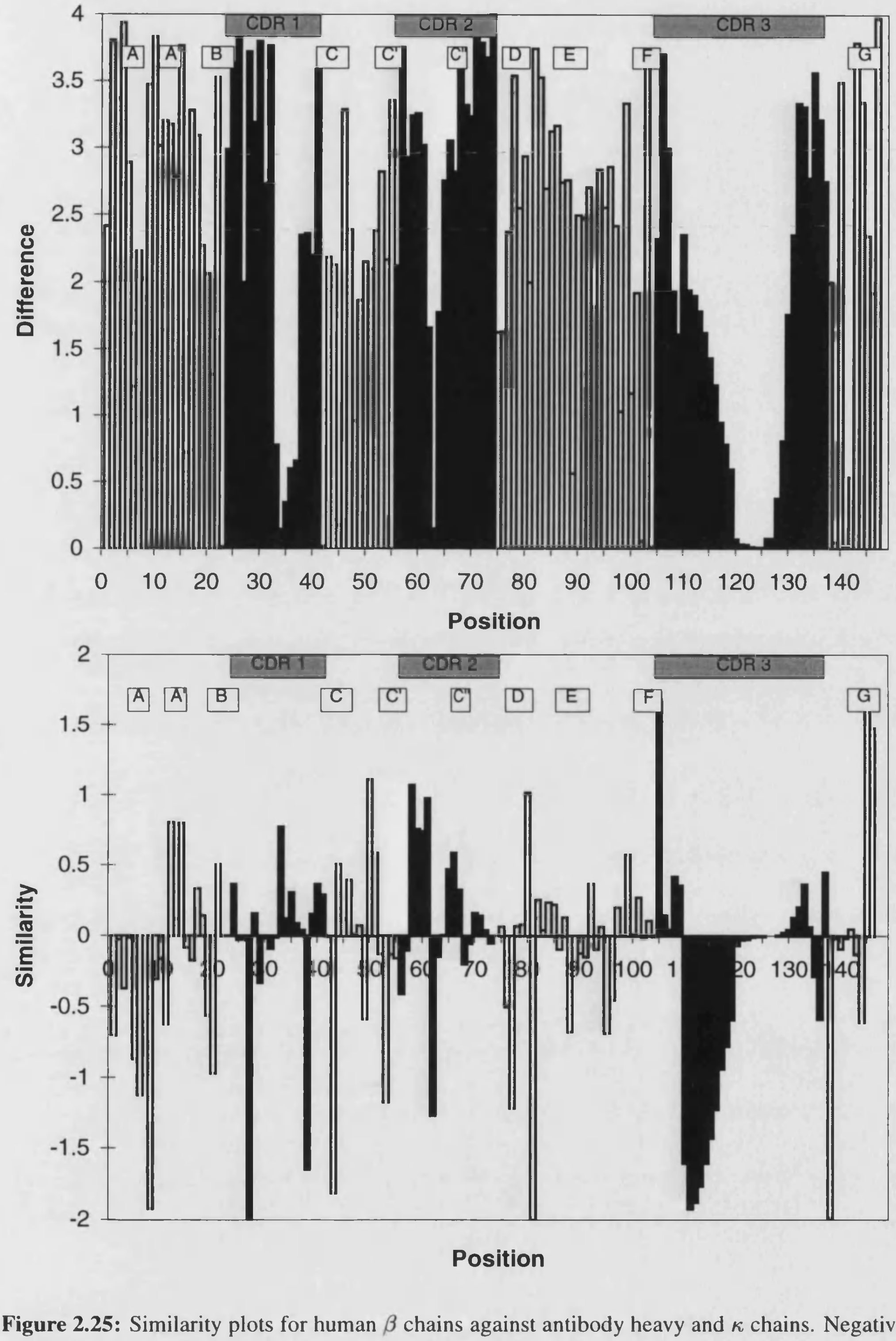


Figure 2.25: Similarity plots for human β chains against antibody heavy and κ chains. Negative values indicate greater similarity to antibody κ chains than heavy chains and positive values greater similarity to antibody heavy chains.

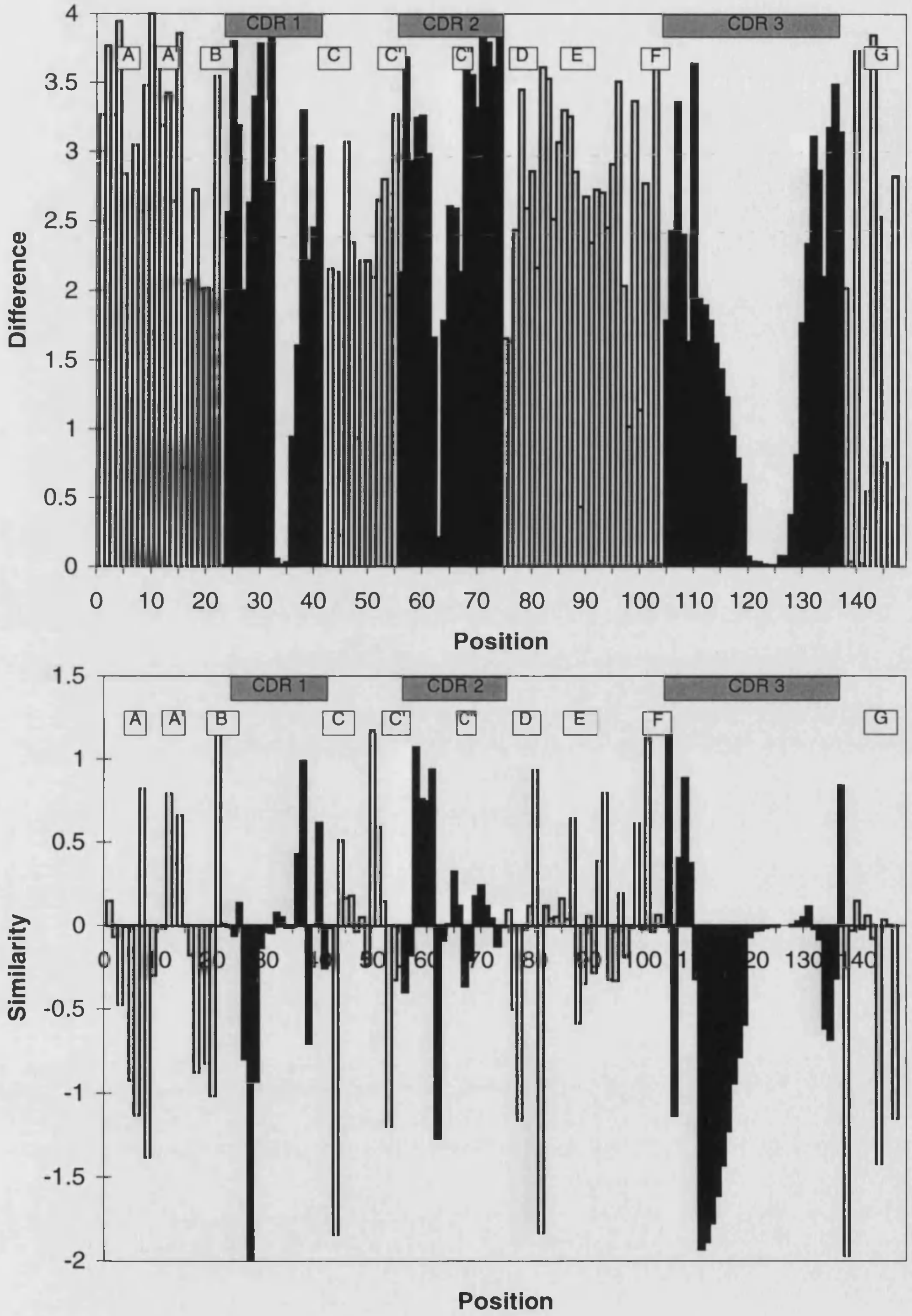
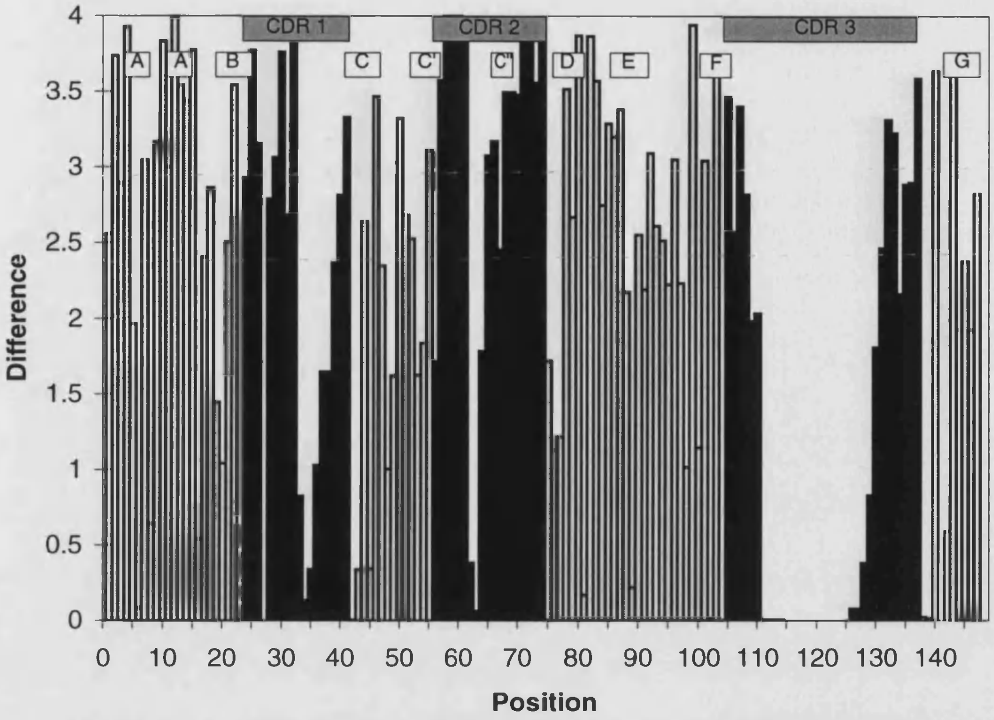


Figure 2.26: Similarity plots for human β chains against antibody heavy and λ chains. Negative values indicate greater similarity to antibody λ chains than heavy chains and positive values greater similarity to antibody heavy chains.



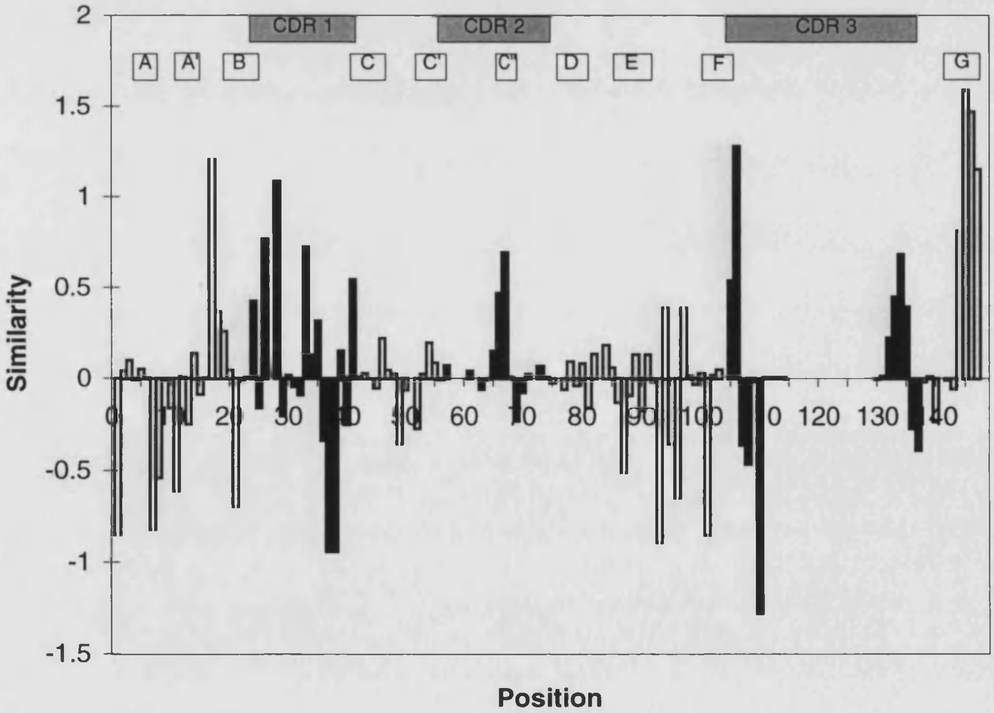


Figure 2.27: Similarity plots for human β chains against antibody κ and λ chains. Negative values indicate greater similarity to antibody κ chains than λ chains and positive values greater similarity to antibody λ chains.

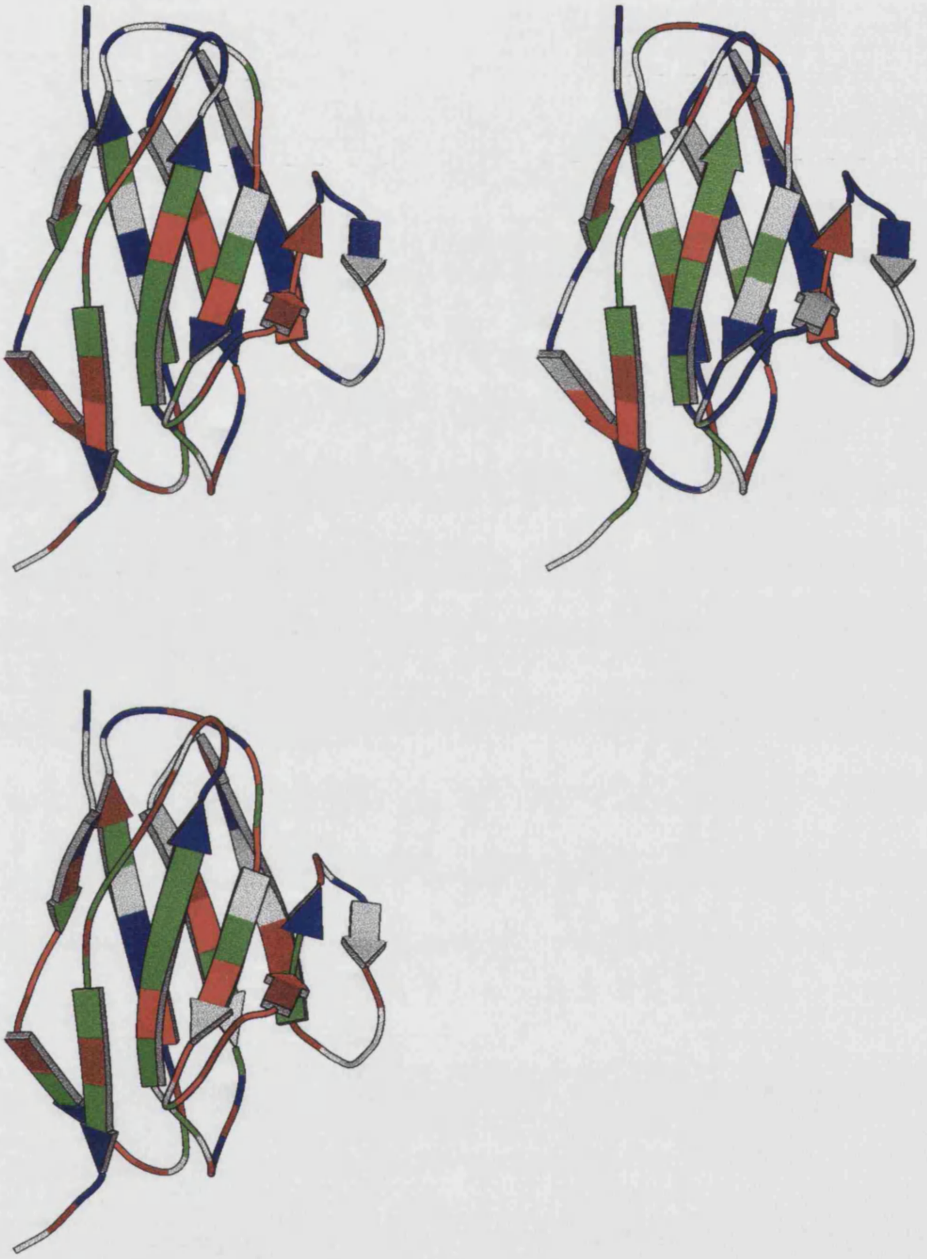


Figure 2.28: The similarity index values for TCR α chain V regions against antibody chain types are plotted onto a schematic of the κ light chain V region of HYHEL-5. The top left plot compares TCR α against Ab κ and Ab γ chains, the top right plot is against Ab λ and Ab γ chains, and the bottom plot against Ab κ and Ab λ chains. Where the value of the difference in frequencies to each chain type was less than 1.0 the residue is coloured green (similar to both). Where the first mentioned antibody sequence type has a difference value less than the other and this value is less than 1.0 the residue is coloured bright red. Where the value is less than 1.5 the residue is coloured dark red. If the second chain type has the lower difference score the residue is coloured bright blue if the value is less than 1.0 and dark blue if the value is less than 1.5. Diagram created with a modified version of molscript [26].

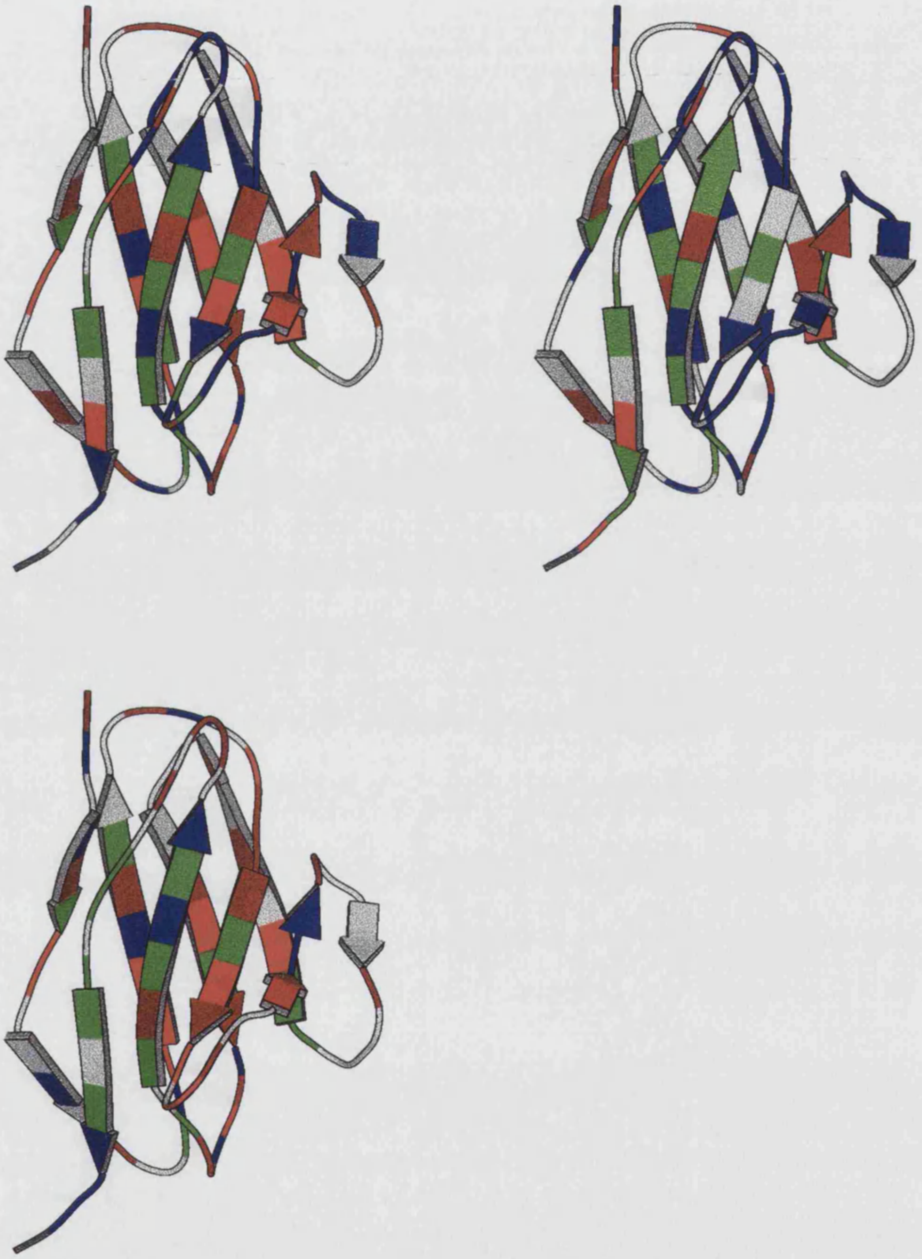


Figure 2.29: The similarity index values for TCR α chain V regions against antibody chain types are plotted onto a schematic of the κ light chain V region of HYHEL-5. The top left plot compares TCR α against Ab κ and Ab γ chains, the top right plot is against Ab λ and Ab γ chains, and the bottom plot against Ab κ and Ab λ chains. Where the value of the difference in frequencies to each chain type was less than 1.0 the residue is coloured green (similar to both). Where the first mentioned antibody sequence type has a difference value less than the other and this value is less than 1.0 the residue is coloured bright red. Where the value is less than 1.5 the residue is coloured dark red. If the second chain type has the lower difference score the residue is coloured bright blue if the value is less than 1.0 and dark blue if the value is less than 1.5. Diagram created with a modified version of molscript [26].

nature of the β-sheet which forms the V domain interface is apparent. The large difference in environment for the D strand of the α chain is highlighted.

2.3.7 Conclusions

In homology modelling the choice of the framework structure is crucial to the structure of the final model [137]. Previous studies on TCR modelling have tended to concentrate on the general similarity of TCR α and β chains to a general immunoglobulin fold. In this chapter the light and heavy chain character of α and β chains have examined.

The results show that both chain types have features characteristic of both light and heavy chains. Heavy chain characteristics include the lengths of the third CDRs of both α and β chains and the conserved alanine at the start of this loop. Also the length of the D-E loop in the α chain and the conservation around this region is heavy chain like. The length of the β chain loop is unlike light or heavy chains. Light chain features include the length of the CDR 1 of both α and β and the CDR 2 of the α chain. The length of the A-A' loop in both α and β is also characteristic of κ light chains and both chain types show conservation of a proline important for the conformation of this loop in κ light chains. Possibly the most persuasive evidence of light chain character is the identity of residues at positions involved in interdomain contacts in light and heavy chains. These positions in both α and β show much higher homology to light chain sequences than to heavy chain sequences.

From these results it is suggested that a light chain dimer may be a better model

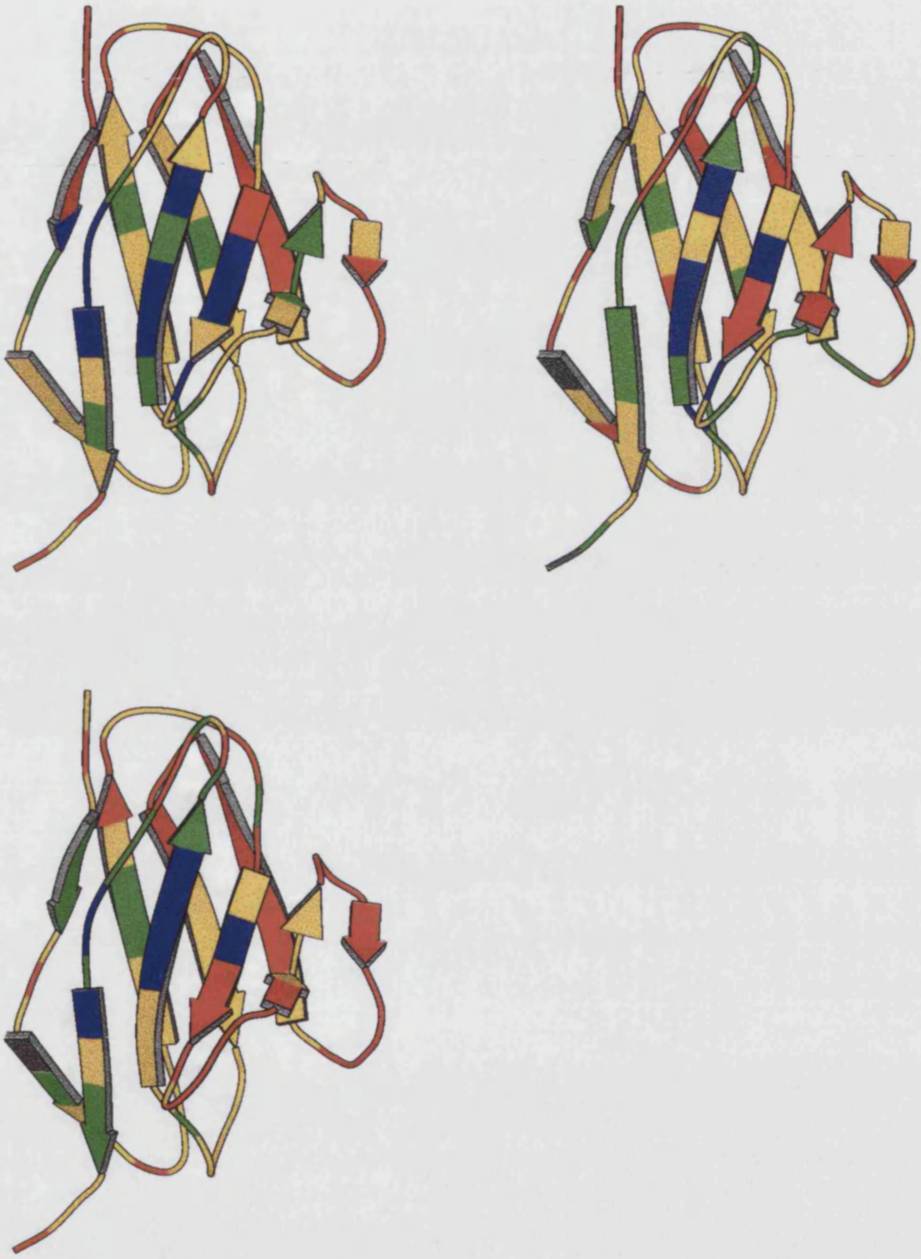


Figure 2.30: Mouse TCR α chains environment comparison to Ab κ (top left), Ab γ (top right) and Ab λ (bottom). The colours indicate the score at each residue position: blue 0 to 0.5 (most similar), green 0.5 to 1.0 , yellow 1.0 to 1.5 and red 1.5 to 2.0 (least similar). Diagram created with a modified version of molscript [26].

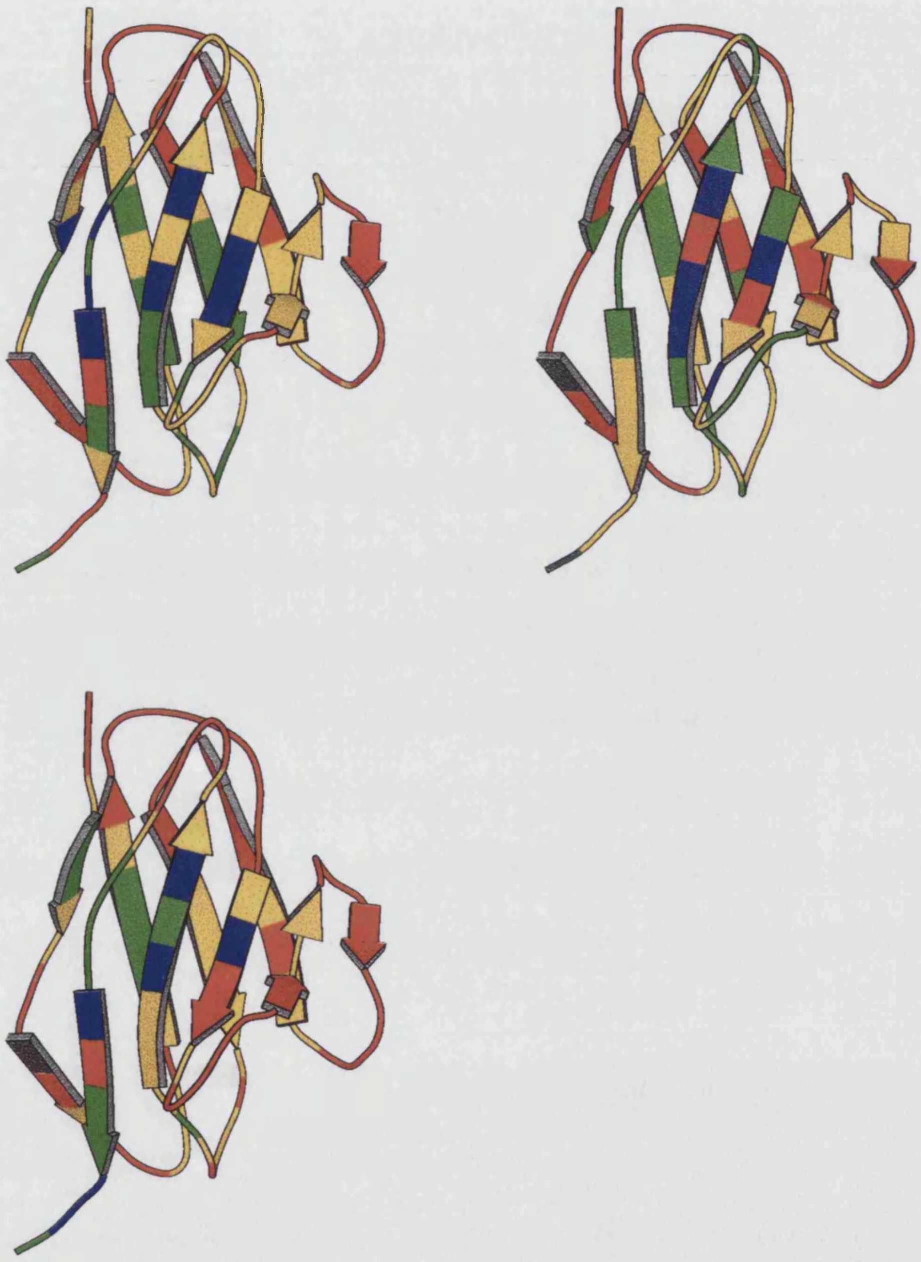


Figure 2.31: Mouse TCR β chains environment comparison to Ab κ (top left), Ab γ (top right) and Ab λ (bottom). The colours indicate the score at each residue position: blue 0 to 0.5 (most similar), green 0.5 to 1.0 , yellow 1.0 to 1.5 and red 1.5 to 2.0 (least similar). Diagram created with a modified version of molscript [26].

of the structure of a TCR than a V_L - V_H dimer. The use of such a framework has been suggested previously [138] though on the basis of very little detailed analysis. The structures of light chain dimers have been compared to antibodies by Novotny and Haber [117]. They tend to associate in a similar manner to the V_L - V_H module, creating a similar domain interface, although the area buried on domain association is less in V_L - V_L by about 3.5 nm^2. The combining site arrangement resembles the V_L - V_H structure.

Chapter 3

Modelling of Antibodies

The high structural homology of all the antibody structures so far determined (see figure 3.2), the large number of sequences that are known and the possible uses to which antibodies could be put [139–141], has led to substantial interest in modelling antibody variable domains.

3.1 Homology Modelling

Evolutionarily related proteins with homologous sequences have been shown to have similar structures. **Homology modelling** uses this fact to generate models of unknown structures. In homology modelling the first step is alignment of sequences [142] of proteins with homologous sequences and superimposition [143] of their structures. This leads to the identification of the **structurally conserved regions** (SCRs) and the **variable regions** (VRs). SCRs are often the buried hydrophobic core residues which are conserved in sequence and structure. The VRs are often the loop regions on the surface of the molecule and are those regions which vary in structure most between homologous proteins.

Modelling the SCRs is often just a case of taking the backbone of the most homologous structure to the desired sequence and replacing the sidechains. The more difficult part of homology modelling is the construction of the VRs. Several methods have been developed. They mainly fall into two categories; database search methods and *ab initio* methods.

3.1.1 Database Searching

Distance constraints are generated from homologous proteins for residues on either side of the loop region. Protein fragments which match the constraints and have the correct number of intervening (loop) residues are selected from a database of protein structures. The fragments selected are fitted onto the anchor regions, sometimes with some minor changes permitted to the conformation of the end residues to produce the best superposition. These conformations are then screened using an energy function to select a final conformation. For short loops the geometric constraints are often sufficient to produce reasonably accurate models, whereas for longer loops it is necessary to include sequence information when searching for fragments.

3.1.2 *Ab initio Methods*

With these methods the generation of all possible loops is achieved by conformational search methods. One such method is used in CONGEN [144]. In this program the conformation of three central residues in a loop are determined by an analytical chain closure algorithm [144–146]. The backbone torsions of the preceding and succeeding residues in the loop are rotated in discrete search steps to search all available conformational space. CONGEN also implements an energy minimisation step for each generated conformation, which allows the search of conformational space between the torsion steps. The problem with the conformational search approach is that the combinatorial explosion which accompanies increasing loop length means that it is only suitable for short loops. Conformational searching has two advantages over random methods such as Monte Carlo simulated annealing [147] or molecular dynamics [148]. First, it is carried out on a regular grid with discrete search steps, unlike dynamics or Monte Carlo methods that sequentially perturb one conformation into another by small increments and hence may sample the same space many times. Secondly, conformational searching does not entail the cost of determining energy derivatives, which are required in molecular dynamics.

3.1.3 Screening Conformations

Both conformational search and database screening normally produce many different possible conformations for the region to be modelled. To determine which one is to be incorporated in the final model it is necessary to have a screening method. Often the energy of each conformation is evaluated using an approriate potential energy fuction. The potential generally includes terms for the bonds, angles, torsions and out of plane angles, as well as the non-bond interactions. Usually, there are also terms for the interactions between the above primary terms, known as cross terms. The equation below is that used in the CVFF [149] forcefield. The first four terms describe the energy required to distort the bonds, valence angles, torsion angles, and out of plane angles. Terms five to nine represent relationships between the first four terms. These are the cross terms. The last two terms describe the van der Waals and electrostatic interactions between atoms. The terms are represented graphically in figure 3.1.

$$
\begin{aligned}
E_{pot} = & \sum_{b} D_b[1 - e^{\alpha(b-b_0)^2}] + \\
& \frac{1}{2}\sum_{\theta} H_\theta(\theta - \theta_0)^2 + \\
& \frac{1}{2}\sum_{\phi} H_\phi[1 + scos(n\phi)] + \\
& \frac{1}{2}\sum_{\chi} H_\chi \chi^2 + \\
& \sum_{b}\sum_{b'} F_{bb'}(b - b_0)(b' - b'_0) + \\
& \sum_{\theta}\sum_{\theta'} F_{\theta\theta'}(\theta - \theta_0)(\theta' - \theta'_0) + \\
& \sum_{b}\sum_{\theta} F_{b\theta}(b - b_0)(\theta - \theta_0) + \\
& \sum_{\phi} F_{\phi\theta\theta'} cos\phi(\theta - \theta_0)(\theta' - \theta'_0) + \\
& \sum_{\chi}\sum_{\chi'} F_{\chi\chi'}\chi\chi' + \\
& \sum \epsilon[(r^*/r)^{12} - 2(r^*/r)^6] + \\
& \sum q_i q_j/\epsilon r_{ij}
\end{aligned}
\tag{3.1}
$$

where b is bond length, θ is valence angle, ϕ is torsion angle and χ is out of plane angle. The variable r is the distance between atoms, q is partial atomic charge and ϵ is the energy of interaction at the most favourable interaction distance r^*. H, F and D are force constants (equation taken from [150]).

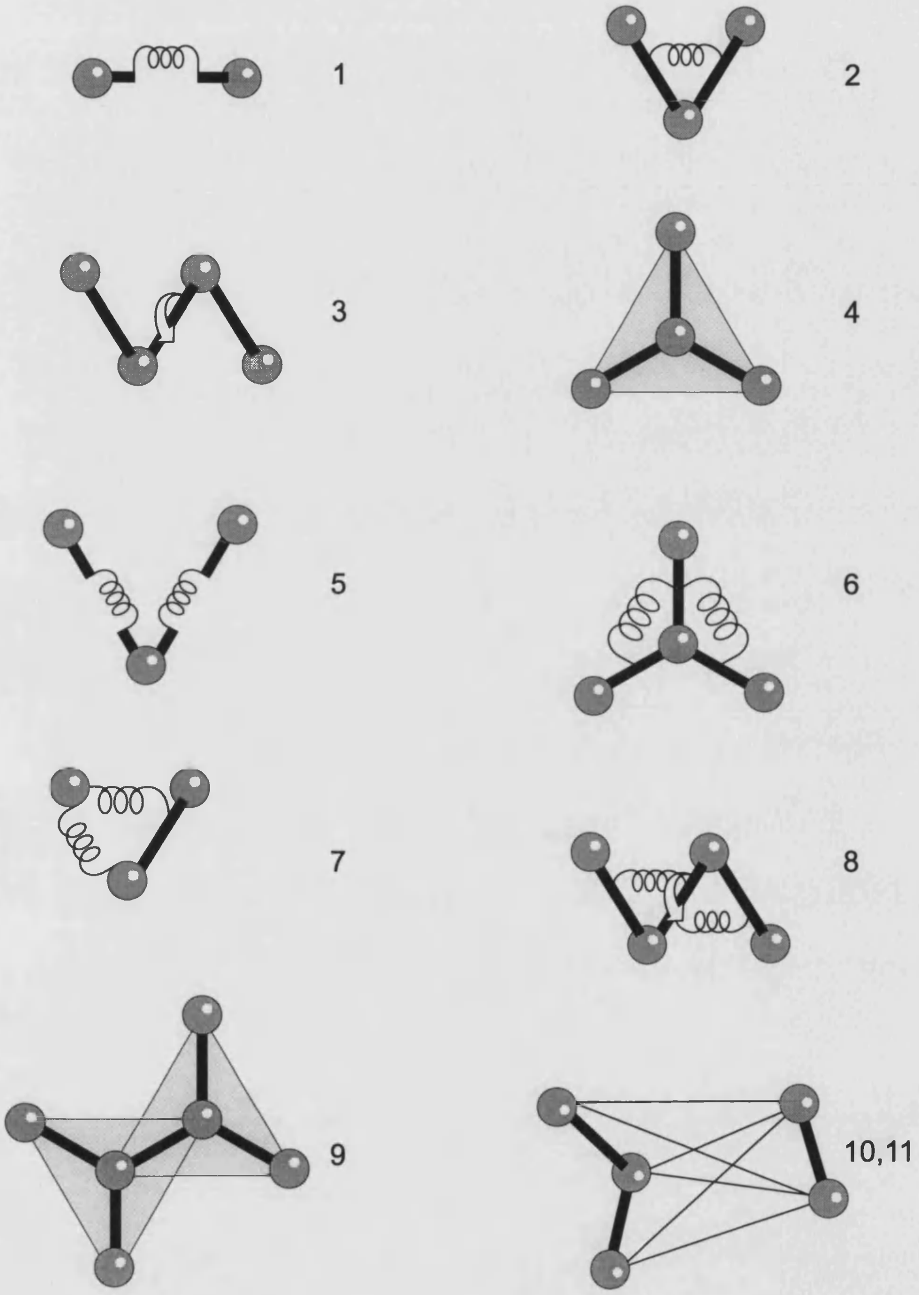


Figure 3.1: A graphical representation of the eleven terms in the CVFF potential described by equation 3.1. The numbers refer to the lines in the equation. Adapted from [150].

3.1.4 Sidechain Building

The two main approaches to sidechain modelling are similar to the two approaches to loop modelling, that is database (or knowledge based) methods and *ab initio* conformational search.

3.1.4.1 Database Methods

Certain conformational preferences have been shown for the sidechains of residues in particular types of secondary structure [151–153]. These preferences are significant in α helices and β sheets, but less conservation of conformation is seen in loop segments. Some preferences have been noted for loop regions [153]. However the information available is for all types of loops and turns collectively, thus giving low confidence when modelling a specific type of loop such as an antibody CDR, which is made up of several different classes of secondary structure.

Ponder and Richards [154, 155] have shown that only a limited set of rotamers are used for each type of sidechain in the core of proteins, and have constructed a library of these conformations. However these rules do not apply to surface residues and do not accurately specify all core sidechains.

In the case of canonical CDR loops database methods usually have a higher confidence than *ab initio* methods [120], but for other CDRs database methods are not suitable. One of the main problems with the above methods is that they do not give much consideration to the local environment of the particular residue, which will be especially important in loop residues.

3.1.4.2 *Ab initio* Methods

Ab initio generation of side chains involves generating conformations and then evaluating them using an objective function. In CONGEN [144] several different methods for generating sidechain conformations are implemented:

- All. Generates all possible conformations, using a series of nested loops, one for each sidechain torsion. This method is impractical for large numbers of sidechains.
- Independent. Each sidechain is build independently of the others. This method produces conformations with many van der Waals clashes.
- Combination. Generates a small number of low energy conformations for each sidechain and then evaluate the energy of all the combinations of these. Again this method is impractical for large numbers of sidechains.
- First. Uses the algorithm described for All, but stops once the first low energy conformation is found.
- Iterative. Cycles through the sidechains in a specified order. For each sidechain finds the lowest energy conformation. Continues modifying each sidechain until the energy converges.

The main problem in sidechain generation is the combinatorial explosion as the number of sidechains to be built increases. Monte Carlo/Metropolis methods [156, 157] have been used to model core sidechains. These methods were adapted for use with surface residues such as those in antibody CDRs by Pedersen [136] in which a solvation term was included to account for the surface exposure of the residues. Sidechain generation by this method is computer intensive and not suitable for generation of sidechains for a large number of loop conformations.

3.2 Homology Modelling of Antibodies

For an antibody Fv domain the SCRs consist of the β-strands, and also all the loops except the six CDRs. Figure 3.2 shows twelve antibody structures superimposed on their most structurally conserved residues. This figure illustrates the high homology of these structures. The CDRs equate to VRs of the antibodies for the purposes of homology modelling.

Modelling of the framework (the β-sheets and non hypervariable loops) is relatively straight forward using the known antibody structures for the backbone structure and a 'maximum overlap' [137, 158] approach to replace residue sidechains differing in the aligned model and structure sequences. The more difficult part in the modelling procedure is building the hypervariable loops. Two approaches have proved successful in modelling these loops, firstly the canonical loop modelling approach [120] and secondly the combined algorithm for modelling antibody loops (CAMAL).

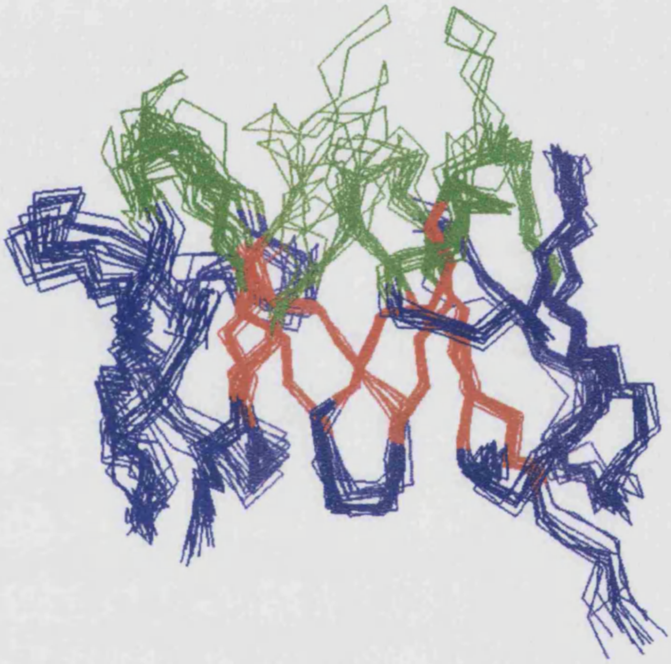


Figure 3.2: Twelve antibody structures superimposed on their conserved β strands. The conserved strands are coloured red, the framework blue and the CDRs green.

3.2.1 Canonical Modelling Method

The canonical loop modelling approach [120] involves identification of similarities between the backbone conformations of the hypervariable regions in the antibodies of known structure. For all of the CDRs except H3, groups of antibody structures with similarly shaped loops have been identified. These "canonical" loops have their structure determined by "key" residues at defined positions in the sequence. These residues are involved in packing, have conserved hydrogen bonds or have unusual torsion angles. The presence of these residues in an antibody sequence to be modelled indicates membership of the canonical class and allows prediction of loop conformation, by substitution of the loop from a known structure in the class.

3.2.2 CAMAL Modelling Method

The other approach is the Combined Algorithm for Modelling Antibody Loops (CAMAL) [5, 6]. The algorithm is termed a "combined" algorithm because a combination of database search methodology and conformational searching are employed to generate conformers for each antibody CDR. The final conformation is chosen from the conformers using energy screening and torsion angle filtering.

The database search utilises distance constraints calculated for the six CDRs in the antibody combining site. These are calculated from the known crystal structures of antibodies. For each CDR in each structure the length from the Cα at either end of the loop to the Cα of each residue within the loop is calculated.

The mean distance and standard deviation are calculated for distances between equivalent positions in all of the structures. The constraints use a range of 3.5 standard deviations either side of the mean distance. A Cα distance database of all the proteins in the Brookhaven protein databank is then searched for fragments of the required length which satisfy the constraints for each loop and these conformations are extracted. Each conformation is fitted onto the antibody structure by translating the N terminus to the position of the framework antibody terminus, rotating the loop to position the C terminus at the C terminus of the framework antibody, and then rotating the loop until its centre of geometry matches that of the structure used as the framework.

The central five residue section from each selected database fragment is then constructed *ab initio* using the conformational search program CONGEN [144]. The algorithm involves rotation of the two end residues of the five residue segment about their torsions. For each conformation of these two residues an analytical chain closure algorithm positions the other three residues [145, 146]. For loops of six and seven residues no backbone conformational search is performed because the database saturates the backbone conformational space. Sidechains are still constructed using CONGEN. For loops of five residues the database search produces too many possible conformations and conformational search of the entire loop is carried out.

The selection of the final conformation occurs in two stages. First each conformation is evaluated using a solvent modified potential excluding the electrostatic and Lennard - Jones attractive potentials. Two different potentials have been used.

Initially the GROMOS potential [159] was used. Since 1992 the EUREKA potential [160] has been used. The final conformation is then picked from the five lowest energy conformations using a torsion angle filter based on the initial database loops.

3.3 AbM: An Algorithm to Model a Complete V Domain

The CAMAL algorithm has been incorporated into a commercial antibody modelling program, AbM (Oxford Molecular Group, plc, UK). The program also implements antibody framework building and canonical loop construction algorithms enabling the user to generate a complete model of an antibody variable domain just by entering the light and heavy chain sequences. The program automatically sets default options for the modelling procedure, although these can be changed. The method is described in the following sections.

3.3.1 Sequence Selection

Sequence selection is the process of choosing sequences from the antibodies of known structure to be used as the basis for modelling the framework regions of the modelled antibody. In AbM the sequence of the light and heavy chain of the unknown are manually aligned against the respective sequences of the antibodies of known structure.

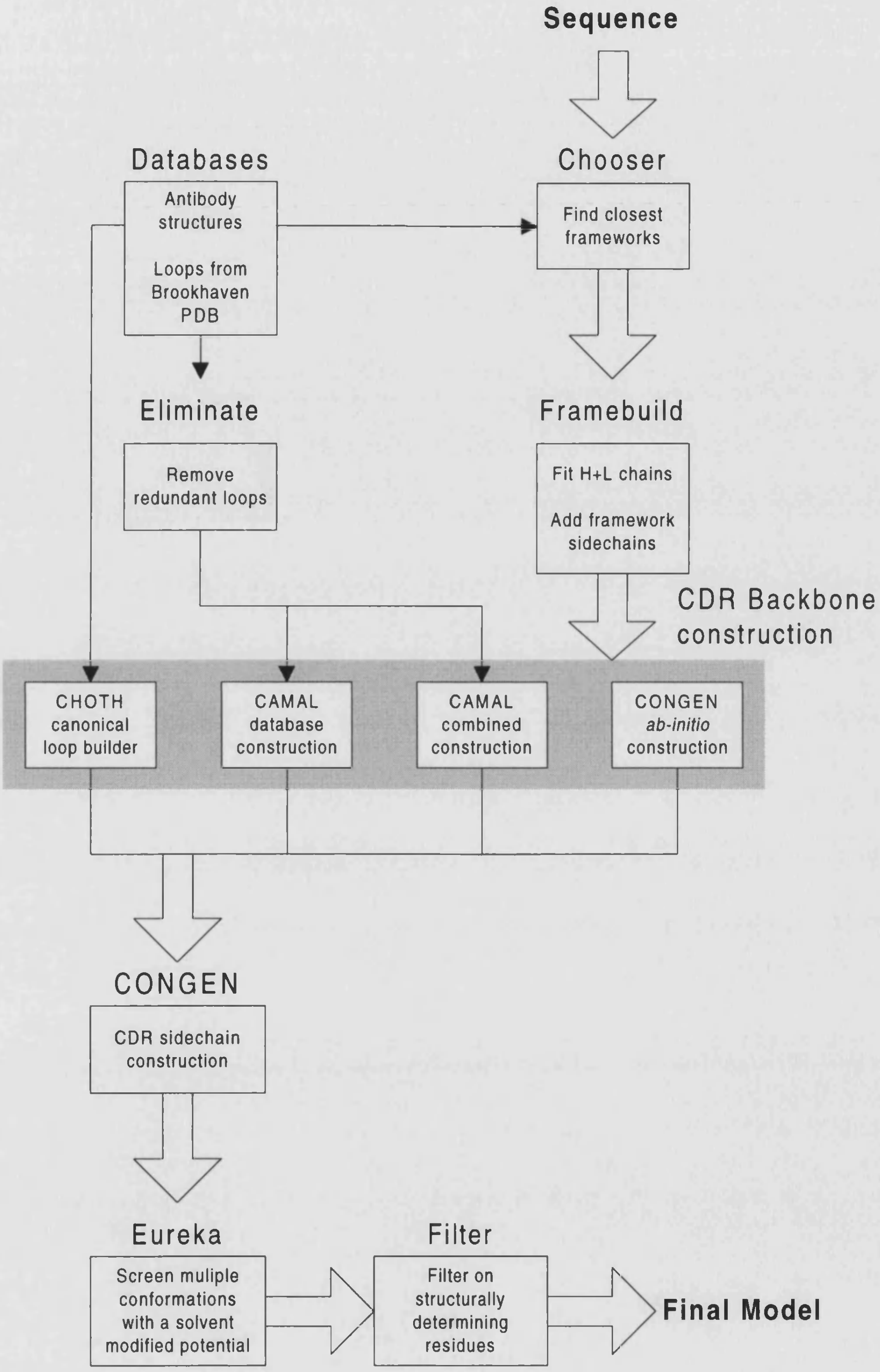

Sequence
Databases
Antibody structures
Loops from Brookhaven PDB
Chooser
Find closest frameworks
Eliminate
Remove redundant loops
Framebuild
Fit H+L chains
Add framework sidechains
CDR Backbone construction
CHOTH canonical loop builder
CAMAL database construction
CAMAL combined construction
CONGEN *ab-initio* construction
CONGEN
CDR sidechain construction
Eureka
Screen muliple conformations with a solvent modified potential
Filter
Filter on structurally determining residues
Final Model

Figure 3.3: Flowchart of the procedures implemented in the AbM program.

3.3.2 Framework Building

The antibody framework strands and conserved loops are built in AbM by the program FRAMEBUILD written by Pedersen [136]. The framework building involves first resequencing the structures chosen for the light and heavy chain V domains to the required sequence for the model, and then fitting the two chains to generate the complete Fv framework model.

Sidechain replacement is performed using a maximum overlap protocol. Where the structure and the model differ in sequence, a sidechain template is fitted onto the backbone atoms with the sidechain torsions adjusted to match those of equivalent torsions in the parent sidechain.

The most structurally conserved strands in the light and heavy chains have previously been identified by Pedersen [136]. These residues were used to generate a mean set of coordinates. By fitting onto these coordinates the two chains are positioned relative to each other.

3.3.3 Use of Canonicals

The program CHOTH, written by the author, implements the canonical loop modelling method. The program takes as its input the sequence of the antibody to be modelled, the names of the structures to be used as the frameworks for the two chains (referred to as the framework structures), and a control file indicating which loops canonical loop modelling is to be attempted for. It also reads the configuration file which indicates which loops have canonical classes defined. For each loop with canonical classes a data file containing sequence patterns for the various classes, in a format similar to the Prosite format, is read in (the format is described in Appendix A). For each class this file also contains the residue range to be replaced (the loop), the length of the loop and two residue ranges which are used for fitting. These pattern files contain the classes described in Chothia's original paper on canonicals [120] as well as the additions made in later papers [121, 122].

Pattern matching code identifies any canonical loops in the sequence to be modelled. Once this has been done it remains to replace the coordinates for each canonical loop in the framework PDB file with those of a loop from a structure which is a member of the same class. The database of antibody structures created for the framework building program described above is used. A sequence file containing the sequences of all the structures in this database is read in, and searched for members of the correct canonical class. The most similar structure in terms of sequence identity is chosen. The loop region and framework fitting regions are extracted from the chosen structure and fitted onto the framework structure. The fitted loop residues are then inserted into the framework structure and a new PDB file written out containing the canonical loops.

Initially the fitting was performed using a similar algorithm to that used in the CAMAL algorithm for fitting database loops onto the framework structure, using the centre of geometry of the framework loop. However this method produced poor RMS values for some modelled loops because the loop from the framework structure had a different take-off angle (angle between the loop and the framework strands on either side) to the canonical loop. Also for CDR H1 the end points of

Structure	CDR Length	Global RMS		Local RMS	
		Ca	Backbone	Backbone	All
gloop2	11	1.103	1.161	0.801	2.898
2hfl	10	1.140	1.150	0.479	0.812
2mcp	17	0.720	0.784	0.546	1.303
2fbj	10	1.681	1.733	0.615	1.052
6fab	11	0.848	0.788	0.631	2.429
1dfb	10	0.871	0.834	0.508	3.256
3hfm	11	0.801	0.775	0.508	2.880
1mam	11	1.251	1.302	0.742	2.812
2fb4	13	0.755	0.780	0.235	2.088
1fdl	11	0.799	0.799	0.267	3.107
1hil	17	1.148	1.151	0.546	1.467
1f19	11	1.161	1.226	0.801	2.885

Table 3.1: RMS deviations for the CDR L1 loops modelled using the canonical modelling method implemented in the program CHOTH. The global RMS values are calculated after fitting the structure and the model on the structurally conserved residues in the interface β strands. The local RMS values are calculated after fitting CDR L1 loop residues of the model with those of the structure. The backbone RMS is for the C, Cα and N atoms of each residue.

the loop and the centre of geometry were almost collinear resulting in inaccuracy of loop placement. The fitting algorithm was therefore changed, using the least squares fitting of structurally conserved regions on either side of the loop to position the canonical loop on the framework structure.

CHOTH does not change the sequence of the canonical loops in the PDB file because this is done in the FRAMEBUILD step. The sidechains are modelled using CONGEN.

Results for modelling of each of the five loops which have canonical classes are shown in tables 3.1, 3.2, 3.3, 3.4 and 3.5.

Structure	CDR Length	Global RMS		Local RMS	
		Ca	Backbone	Backbone	All
gloop2	7	0.631	0.647	0.228	0.654
2hfl	7	0.709	0.712	0.320	1.237
2mcp	7	0.613	0.538	0.436	0.955
4fab	7	0.768	0.792	0.279	1.144
2fbj	7	0.893	0.867	0.320	2.812
6fab	7	0.293	0.304	0.252	1.031
1dfb	7	0.738	0.750	0.228	1.166
3hfm	7	0.978	1.021	0.437	2.064
1mam	7	1.361	1.362	0.252	0.966
1igf	7	0.749	0.763	0.279	1.144
2fb4	7	1.172	1.247	0.966	2.308
1fdl	7	0.944	0.928	0.502	1.611
1hil	7	0.917	0.922	0.436	1.634
1f19	7	0.915	0.966	0.502	1.725

Table 3.2: RMS deviations for the CDR L2 loops modelled using the canonical modelling method implemented in the program CHOTH. The global RMS values are calculated after fitting the structure and the model on the structurally conserved residues in the interface β strands. The local RMS values are calculated after fitting CDR L2 loop residues of the model with those of the structure. The backbone RMS is for the C, Cα and N atoms of each residue.

Structure	CDR Length	Global RMS		Local RMS	
		Ca	Backbone	Backbone	All
gloop2	9	1.003	1.031	0.297	1.104
2mcp	9	0.718	0.739	0.242	1.052
4fab	9	1.231	1.255	0.625	1.900
6fab	9	1.160	1.131	0.930	2.047
3hfm	9	0.426	0.394	0.234	1.463
1mam	9	1.270	1.289	0.260	0.704
1igf	9	0.888	0.877	0.416	1.081
1fdl	9	1.126	1.138	0.652	2.243
1hil	9	1.278	1.288	0.242	1.035
1f19	9	1.382	1.389	0.931	2.055

Table 3.3: RMS deviations for the CDR L3 loops modelled using the canonical modelling method implemented in the program CHOTH. The global RMS values are calculated after fitting the structure and the model on the structurally conserved residues in the interface β strands. The local RMS values are calculated after fitting CDR L3 loop residues of the model with those of the structure. The backbone RMS is for the C, Cα and N atoms of each residue.

3.4 Adaptations to CAMAL

The greater reproducibility, ease of use and flexibility of AbM compared to the initial implementation of CAMAL, together with the increased number of published antibody structures enabled more thorough testing of the algorithm to be performed. The first column of RMS values in table 3.6 shows the modelling results for eight structures using the standard CAMAL algorithm.

The following sections discuss several adaptations to the basic algorithm. The results for modelling after incorporating all of these changes are shown in the second column of RMS values in table 3.6 and will be discussed later.

Structure	CDR Length	Global RMS		Local RMS	
		Ca	Backbone	Backbone	All
gloop2	5	1.884	1.785	0.204	1.664
2hfl	5	1.176	1.261	0.696	2.301
2mcp	5	0.968	1.004	0.275	1.492
4fab	5	0.672	0.721	0.378	1.867
2fbj	5	0.502	0.515	0.142	1.520
6fab	5	1.358	1.341	0.434	2.002
1dfb	5	0.736	0.736	0.206	0.935
3hfm	5	2.037	2.012	0.937	2.098
1mam	5	1.849	1.845	0.275	1.218
1igf	5	1.335	1.310	0.266	1.664
2fb4	5	0.626	0.621	0.096	0.352
1fdl	5	0.869	0.846	0.448	0.794
8fab	5	1.103	1.106	0.096	0.330
1hil	5	0.565	0.563	0.266	1.083
1f19	5	4.444	4.601	0.696	1.280

Table 3.4: RMS deviations for the CDR H1 loops modelled using the canonical modelling method implemented in the program CHOTH. The global RMS values are calculated after fitting the structure and the model on the structurally conserved residues in the interface β strands. The local RMS values are calculated after fitting CDR H1 loop residues of the model with those of the structure. The backbone RMS is for the C, Cα and N atoms of each residue.

Structure	CDR Length	Global RMS		Local RMS	
		Ca	Backbone	Backbone	All
gloop2	10	1.543	1.609	0.624	2.997
2hfl	10	2.202	2.155	0.624	2.455
2mcp	12	2.022	2.014	0.787	1.623
4fab	12	1.922	2.028	0.787	3.455
2fbj	10	0.767	0.779	0.455	3.076
1dfb	10	1.180	1.202	0.455	2.980
3hfm	9	0.914	0.942	0.561	1.406
1igf	10	1.185	1.202	0.284	2.695
2fb4	10	0.708	0.726	0.200	3.352
1fdl	9	1.369	1.413	0.634	3.336
8fab	10	1.183	1.195	0.200	2.848
1hil	10	0.845	0.829	0.284	2.711

Table 3.5: RMS deviations for the CDR H2 loops modelled using the canonical modelling method implemented in the program CHOTH. The global RMS values are calculated after fitting the structure and the model on the structurally conserved residues in the interface β strands. The local RMS values are calculated after fitting CDR H2 loop residues of the model with those of the structure. The backbone RMS is for the C, Cα and N atoms of each residue.

Structure	Loop	Length	Runtime 1 (Hr)	RMS 1	Runtime 2 (Hr)	RMS 2
Gloop2	H2	10	4.75	2.02	Canonical	1.61
	H3	4	0.81	1.38	0.15	1.27
2hfl	H3	7	0.15	8.01	0.15	2.31
2mcp	H3	11	22.4	2.47	18	2.30
4fab	H3	7	0.2	2.47	0.14	2.13
3hfm	H3	5	0.23	3.98	0.10	1.68
2fb4	L1	13	0.63	4.75	Canonical	0.78
	L3	11	13.25	1.74	20.1	1.73
	H3	17	24.10	8.52	18.99	4.22
1fdl	H3	8	1.17	3.87	39.5	2.18
2fbj	H3	9	23.8	3.12	81.2	4.17

Table 3.6: A table of RMS values for the CAMAL modelled CDR loops of eight antibodies modelled in AbM before and after modifications to the method. For the H2 loop of Gloop-2 and L1 loop of 2fb4 the changes to the definition of the canonical loops resulted in their inclusion in canonical classes, and therefore the time required for their generation was minimal.

3.4.1 Database search changes

3.4.1.1 Creation of a Non-redundant Cα Database

The Cα database in the original implementation of CAMAL contained entries for all the structures deposited in Brookhaven database at the time. There are many duplicate structures in the Brookhaven database with differing resolution, as well as entries missing or dubious coordinates. The original algorithm used a program, ELIMINATE, to screen the loops selected in the database search for a loop, from the Cα database, for duplicate entries. However this screen was based on the names of the entries, and often failed to eliminate all duplicate entries. The presence of multiple identical or very similar conformations in the ensemble of conformations chosen in the database search can cause bias in the later filtering stages. To solve this problem, two changes were made.

The Brookhaven entries contained within the Cα distance database were limited, as far as possible, to one of each structure. The annotated PDB file list created by L.L. Walsh [161] was consulted to identify the highest resolution structure with the fewest obvious errors for each protein.

Some proteins are repeated in the non-redundant database, for example if they are in complexes, and some protein families contain highly structurally homologous members. To eliminate duplicates Pedersen clustered the hits from the database search on backbone torsion angles, with one conformation taken from each cluster [136].

3.4.1.2 H3 Constraint Groups

The number of database search constraints that can be generated for a particular CDR is limited by the minimum length of loop found in the antibody structure database. For CDR H3 this is only four residues. Therefore there were only four constraints for H3 using all the structures in the database. It was thought that this small number of distance constraints might not provide a sufficient restriction on the conformational space to be searched for long CDR H3's ($>$ 10 residues).

To test this hypothesis, a set of constraints were created from structures which had long CDR H3 loops (more than ten residues). The H3 loop of the antibody HIL was modelled using two Cα distance databases, one containing HIL coordinates and one without them. The databases were screened using the constraints for long H3's. In the first case the HIL CDR H3 was chosen from the database as the final loop conformation. The middle section, however, which had been constructed with CONGEN was not correct. Without HIL in the database the conformation chosen was again better than using constraints for all antibodies (see table 3.7). Also the run time was reduced from 72 hr to 20 hr on a Hewlett Packard 720 because of the reduction in the number of database conformations extracted with the tighter constraints.

3.4.2 Inclusion of other CDRs while modelling

In the basic CAMAL algorithm each CDR is modelled without any other CDRs present. In the original study the presence of the previously modelled loops was

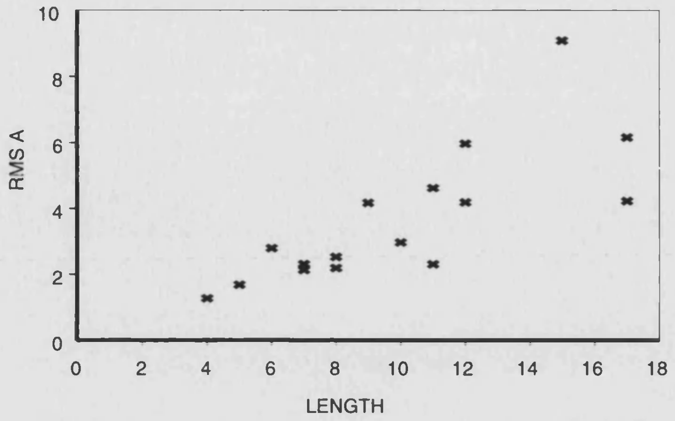


Figure 3.4: RMS versus length plot for modelling H3 CDR.

shown to reduce the accuracy of the later modelled loops.

However in the current study it was found that serious clashes between atoms in two modelled loops often occurred.

Figure 3.4 shows a plot of RMS versus length for CDR H3. There is some correlation between loop length and RMS. For CDR H3s of up to ten residues the loop conformation can be predicted with RMS values equivalent to that of medium resolution X-ray structures. For loops above this length the RMS values for CDR H3 are high. The reasons for this could be that no other CDRs are present when CDR H3 is modelled. The conformational space available to a long loop is large. Without the restrictions defined by the presence of the other loops the H3 loop can sample space not normally available to it giving rise to an inaccurate model.

To investigate this possibility the H3 loop of the antibody HIL [61] was modelled with the backbone atoms of the other loops in place (see table 3.7). In one experiment only the canonical loops were retained (L2, L3, H1 and H2). The

METHOD	CDR H3 Global Backbone RMS (Å)
CDR modelled without any other loops present. Constraints generated using all structures.	5.97
CDR modelled with all other loops present. Constraints generated using all structures.	4.42
CDR modelled with all other loops present. Cα constraints generated for structures with CDR H3s longer than 10 residues. HIL present in Cα database.	2.87
CDR modelled with all other loop backbones present. Cα constraints generated for structures with CDR H3s longer than 10 residues. HIL not in Cα database.	2.68

Table 3.7: The change in RMS deviation for HIL CDR H3 loop as the method has been improved.

loop was modelled better than with all loops absent. However on examination of the model it was noted that the CDR H3 loop had been positioned where CDR L1 would usually be. CDR L1 had been poorly modelled initially but its modelled conformation was included in a second experiment. In this case the RMS of the H3 loop reduced substantially. The region of L1 which contacts H3 is the structurally conserved region at the end of the loop (residues 32-34 (Kabat residue numbering [129])). Therefore although the loop was not modelled well its conformation could be relied on in the region of interest.

3.4.3 Framework take-off angles and H3

CDR H3 occupies a central position in the combining site and therefore has a critical effect on the combining site topology. H3 is defined as a hairpin loop because it inter-connects two β strands. It is very variable in length and therefore there are a large number of possible backbone conformations. There are no assigned canonical structures to CDR H3 because of this high variability. The take-off angle for CDR H3 (see table 3.8) varies by up to 90°. An analysis of take-off angles of all CDRs [162] was performed. The results of the analysis suggested two structural classes of H3 based just on these take-off angles alone. By incorporating additional observations it was possible to define seven classes of CDR H3 (see figure 3.5). The longer CDRs pose more problems for modelling. Therefore for CDR H3 of 13 residues or longer they are all grouped into one class. Most of H3 are less than 12 residues and therefore can be modelled with reasonable accuracy by AbM (see figures 2.12 and 2.11). Canonical structures are assigned to CDRs by finding the most homologous candidate in each class according to the Dayhoff mutation matrix. [163]. However for H3 the approach is to identify the most similar loop within its class of take-off angle (see table 3.8).

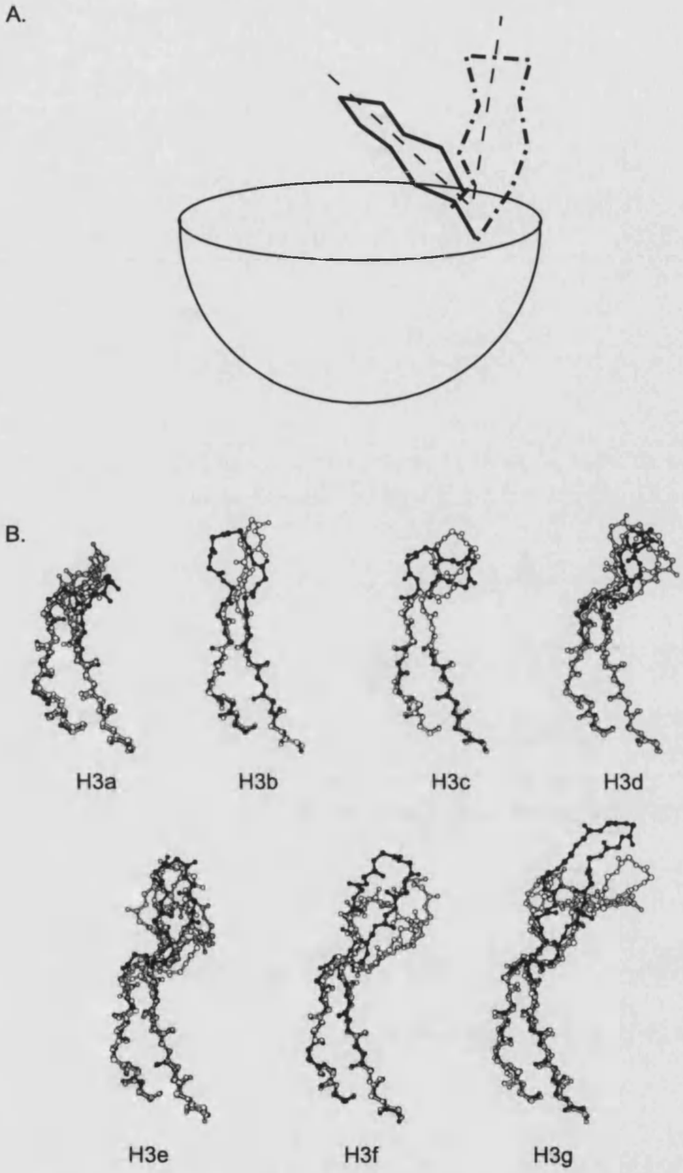


Figure 3.5: (A) A cartoon to illustrate the fact that the differences in CDR take off angle can greatly affect the RMS of the modelled conformation. (B) The seven proposed CDR3 length classes. The definitions for the classes are:-

H3a	loops shorter than 7 residues.
H3b	loops of 7 residues.
H3c-f	loops in which there is an Arg or Lys at position 106 and an Asp, Ala or Gly at position 136 in 8, 9 and 10, 11 or 12 residue loops respectively.
H3g	loops with the motif seen in c-f but of 13 or more residues.

Each class contains structures which differ in their take off angles by less than 35°. The angular difference is calculated as the angle between the planes defined by the N terminal C-α, centre of geometry, and C terminal C-α of the H3 loops for each pair of structures after least squares fitting of the structures on conserved β strands. For class (a) the mean angular difference is 13.6°(5), for class (b) 25.1°(3), for class (c) 7.8°(2), for class (d) 5.9°(6), for class (e) 17.4°(5), for class (f) 12.1°(3), and for class (g) 13.5°(3). The figures in parentheses are the number of structures in each class.

	glb2	3hfm	1baf	2hfl	4fab	d13	1mam	2fbj	3fab	b13i2	2mcp	1hil	8fab	3671	1f19	3d6	2fb4
glb2	0.00	4.60	14.28	85.62	45.32	26.54	27.66	17.02	38.71	30.06	46.06	40.36	27.92	32.89	75.41	44.72	20.32
3hfm	4.60	0.00	10.98	83.86	41.09	22.85	23.18	13.66	35.52	25.80	43.40	36.76	23.88	29.14	72.28	41.76	16.59
1baf	14.28	10.98	0.00	73.30	31.72	12.33	15.94	2.74	24.57	16.84	32.46	26.08	14.15	18.67	61.33	30.78	6.25
2hfl	85.63	83.86	73.30	0.00	56.29	63.90	70.70	70.94	50.77	65.80	41.55	52.03	65.11	59.13	22.71	44.17	69.25
4fab	45.32	41.09	31.72	56.29	0.00	19.88	18.65	29.18	15.90	15.29	21.80	11.09	17.57	14.21	36.51	18.27	25.47
d13	26.54	22.85	12.33	63.90	19.88	0.00	10.00	9.63	13.31	7.08	22.35	13.92	3.62	6.35	49.69	19.91	6.26
1mam	27.66	23.18	15.94	70.70	18.65	10.00	0.00	14.24	20.45	4.92	29.91	18.68	6.80	11.93	53.70	26.80	10.91
2fbj	17.02	13.66	2.74	70.94	29.18	9.63	14.24	0.00	21.86	14.55	29.90	23.34	11.65	15.96	58.63	28.12	3.76
3fab	38.71	35.52	24.57	50.77	15.90	13.31	20.45	21.86	0.00	15.58	9.47	4.81	14.37	8.53	36.77	6.62	19.23
b13i2	30.06	25.80	16.84	65.80	15.29	7.08	4.92	14.55	15.58	0.00	25.03	13.80	3.49	7.11	49.15	21.89	10.81
2mcp	46.06	43.40	32.46	41.55	21.80	22.35	29.91	29.90	9.47	25.03	0.00	12.31	23.78	18.00	29.84	3.61	27.82
1hil	40.36	36.76	26.08	52.03	11.09	13.92	18.68	23.34	4.81	13.80	12.31	0.00	13.71	7.77	35.95	8.77	20.17
8fab	27.92	23.88	14.15	65.11	17.57	3.62	6.80	11.65	14.37	3.49	23.78	13.71	0.00	6.05	49.62	20.94	7.92
3671	32.89	29.14	18.67	59.13	14.21	6.35	11.93	15.96	8.53	7.11	18.00	7.77	6.05	0.00	43.72	15.00	12.57
1f19	75.41	72.28	61.33	22.71	36.51	49.69	53.70	58.63	36.77	49.15	29.84	35.95	49.62	43.72	0.00	30.72	55.85
3d6	44.72	41.76	30.78	44.17	18.27	19.91	26.80	28.12	6.62	21.89	3.61	8.77	20.94	15.00	30.72	0.00	25.70
2fb4	20.32	16.59	6.25	69.25	25.47	6.26	10.91	3.76	19.23	10.81	27.82	20.17	7.92	12.57	55.85	25.70	0.00

Table 3.8: The differences in take-off angle between each pair of structures. The angle calculated is between the centre of geometry of the two framework strands either side of the CDR and the centre of geometry of the loop.

3.4.4 The Problem of Modelling Long H3s: Unsolved or Unsolvable?

The algorithm does not model long CDR H3 loops well. This may be because of problems with the algorithm but to the author's knowledge, no other algorithm can accurately predict long surface loop conformations. This is probably a reflection of the inherent flexibility of these long often highly exposed surface loops. In X-ray crystallographic structures of antibodies the CDR H3 region often has some of the highest B factors, or may even be absent from the density map altogether. It may therefore not be sensible to try to assign a single conformation to such loops, but to define an ensemble of conformations representing the most frequently adopted conformations.

Structure	Length of CDR H3	RMS before changes to method	RMS after changes to method
1dfb	17	6.10	5.44
2f19	15	9.09	7.6

Table 3.9: The RMS deviations of the H3 loops of 2f19 and 1dfb modelled using different versions of AbM.

3.5 Conclusions

Accurate modelling of TCR CDRs will not be possible if antibody CDRs cannot be predicted accurately. To test the ability of the algorithm developed previously in the laboratory to model antibody CDR conformations several antibodies have been modelled. The initial results of this modelling study showed several problems in predicting CDR conformation. CDR H3 presents the greatest problem being most variable in length, sequence and structure. The data show that long loops are not predicted well for CDR H3. The conformational space available to these loops is large. The method did not impose enough restrictions on this space because all the other CDRs are ignored when CDR H3 is constructed. As shown by the results for HIL the predictive accuracy of a relatively long CDR H3 (12 residues) can be substantially improved by including the backbone structures of these loops as alanine residues while modelling H3.

In addition the division of CDR H3 into take off angle and constraint groups based on length has been shown to improve accuracy.

For the other CDRs the use of canonical loops where possible improved the accuracy of their prediction. As the backbone atoms of these loops are included while modelling CDR H3 it is important that they are as accurately modelled as possible.

TCR CDR 3s in both α and β chains show greater sequence homology to CDR H3 than to CDR L3. The lengths of TCR CDR 3s are more conserved in length than H3 with a mean length of about 12 residues. The grouping defined for CDR

H3 conformations determined in this work may help improve accuracy of CDR 3 prediction in TCRs.

Chapter 4

Modelling of T Cell Receptors

In this chapter a molecular modelling study on the TCR is described. In chapter 2 the sequences of TCRs were compared to both immunoglobulin chain types (light and heavy) and evidence was presented that both α and β chains show greater similarity to light chains. Chapter 3 describes some changes made to a program for antibody modelling. These changes were made in parallel with the TCR modelling experiments and were incorporated in the TCR modelling method when they were identified. In this chapter the modelling of several TCRs will be described using a light chain dimer as the structural framework and an algorithm based on the CAMAL algorithm of Martin et. al [6].

A model of CD4 was also produced to confirm that the CAMAL algorithm could be extended from use with antibodies to other immunoglobulins. The modelled TCR sequences for which there were sequence function data were taken from Jorgensen *et al.* [164]. This enabled a validation of the models.

The following introductory sections describe some of the previous experimental structural studies on the TCRs.

4.1 The Sequencing of TCRs

The DNA sequence of a TCR variable region was first isolated in 1984 by Yanagi *et al* [165]. Since then over 500 different TCR mRNAs have been sequenced from a variety of organisms. Since the first sequence was isolated many different TCR α and β chain sequences have been determined using molecular cloning techniques. Anchor polymerase chain reaction has been used and this allows very rapid isolation of an α or β chain gene product using a constant region primer [166].

Also analysis has been performed of the resting levels of different TCR V regions in individuals and the changes in concentrations which occur during disease.

4.2 Mutation studies on TCRs

Patten *et al* [167] attempted to transfer peptide MHC specificity between different TCRs by transfer of putative CDRs. Initially they attempted to transfer CDR3 regions in an attempt to show that these were important for binding to peptide and later they transferred all six putative CDRs on to the TCR framework. Even with all six CDRs plus some of the framework strand regions these experiments failed to transfer specificity, even when the transfer was between TCRs restricted to the same MHC using the same V_α gene segment, and were against similar peptides. The explanation offered drew a parallel with antibody humanisation using CDR transfer where loss of affinity frequently occurs and often requires specific framework mutations back to the original antibody sequence. Therefore, if the transferred CDRs did confer a reduced affinity, since the TCRs in general have low affinities (10^{-6}) compared to antibodies, any further reduction was likely to have made binding undetectable.

Nalefski *et al* [168] mutated two residues in the TCR α chain CDR1 region. The mutated TCR had substantially decreased recognition of the analytic peptide MHC complex. This is most likely to indicate that this region of the molecule is part of the antigen combining site.

Site-directed mutagenesis of a conserved amino acid in the CDR β3 loop, whose presence correlated with a particular cytochrome C specificity, has been shown to affect the fine specificity of that TCR for the antigen [169, 170].

The creation of antibody C region/TCR V region and TCR C region/antibody V region chimeras has provided some information on the characteristics of the V_α-V_β interface. V_HC_α+ $V_\beta C_\beta$ complexes have been produced [171] and it was suggested that the formation of this complex indicated that Vα and V_H domains were similar, with the β being similar to the light chain (because either α or H could act as a partner for β). However this combination is also consistent with a light chain dimer (i.e. both the α and the β being similar to the light chain).

4.3 Structural studies on TCRs

4.3.1 Outline Structure

In 1988 Chothia *et al* produced a paper comparing the sequences of TCRs and antibodies which identified equivalent regions in the sequences of the two types of proteins [127]. They defined forty residues of structural importance in antibody light and heavy chains which are involved in :

- Inter-domain packing between V_L and V_H.
- Inter-domain packing between V_H and C_H.
- Intradomain packing and hydrogen bonding residues maintaining the β-sheet structure.
- Residues with unusual conformations.

The conservation of these residues in T-cell receptor α and β chains was examined. There is high conservation of these residues in T-cell receptors. Further, β sheet framework regions were suggested and a possible arrangement for the antigen binding loops which are like those of antibodies with the third hypervariable loop of each domain sandwiched between the first and second hypervariable loop pairs (see figure 4.1) was proposed.

Novotny *et al.* [172] carried out sequence analysis and predicted secondary structure for the TCR based on antibody structures.

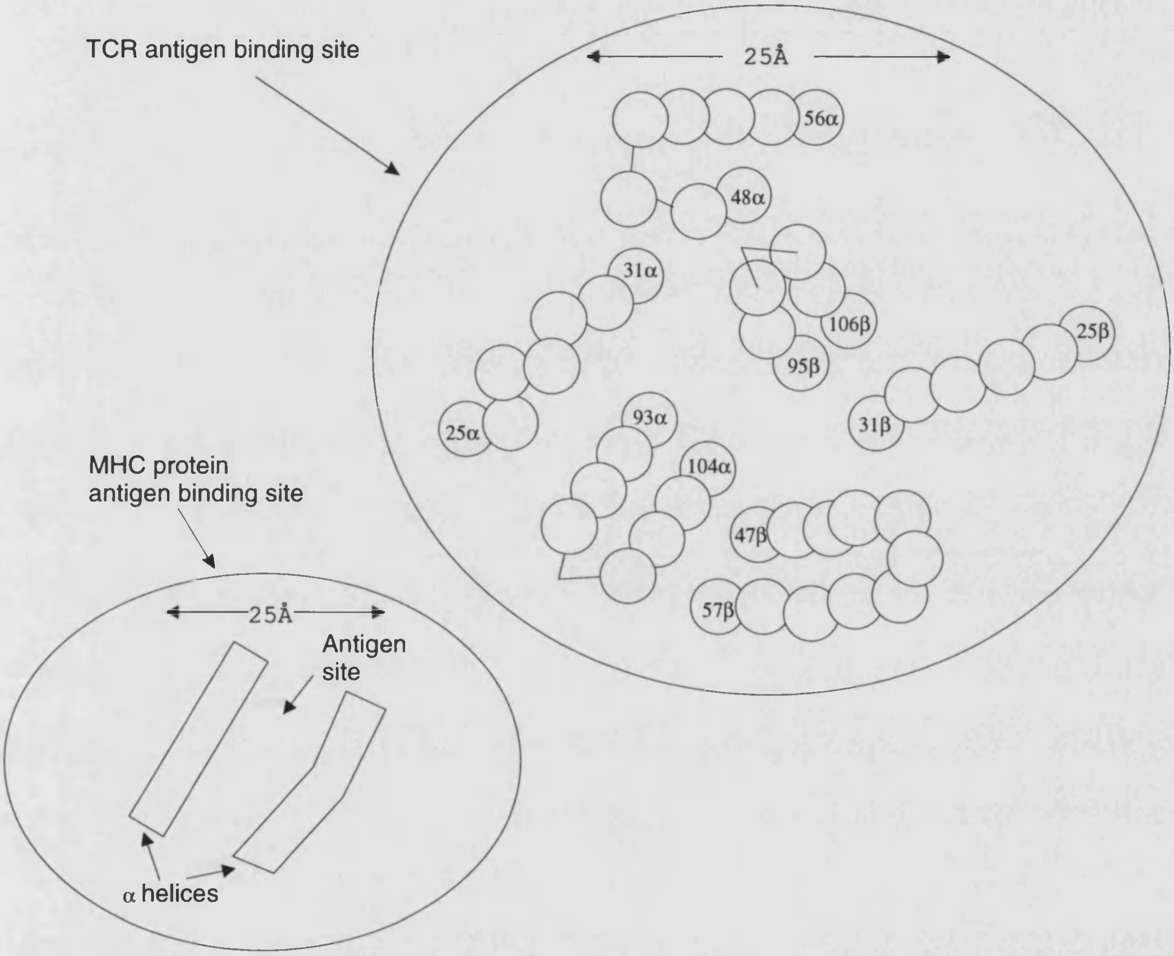


Figure 4.1: This figure is taken from Chothia *et al* [127] and shows the arrangement of combining site loops suggested by them for the T-cell receptor combining site. The receptor is suggested to bind to the MHC with the central loops over the peptide.

4.3.2 Transgenic mice

Jorgensen *et al* [164] used mice transgenic for either TCR α or β chains to map TCR-antigenic peptide contact residues. In their experiments the mice were transgenic for TCR 5C.C7 α or β chain with the chain being expressed on 90% or 98% of the peripheral T lymphocytes. They made substitutions in the peptide which 5C.C7 bound to (MCC 88-103) and injected mice with the changed peptide. They sequenced TCRs produced against the new antigen to characterise the α/β chain which had paired with the transgenic chain. For some changes to the peptide only α chain transgenics responded, indicating the β chain was important for interaction with that peptide residue (because that was the only chain which could change) and *vice versa*. The peptides they used had charge changes and these could be mapped to reciprocal charge changes in the CDR 3 region of the TCR chain. Some of the TCRs produced were cross-reactive with the original peptide and these in general had neutral residues at the position where the charge change occurred. They identified residues in both α and β CDR3s involved in binding the peptide. The interacting residues are shown schematically in figure 4.2.

4.3.3 Cross-reactivity of Antibodies to TCRs and Antibodies

Kaymaz *et al* [173] have shown that there is immunogenic cross-reactivity between human TCR β-chains and human or murine antibody light chains. The best response was to the CDR 1 equivalent region and the third framework region. This is evidence for the structural similarity of TCR chains and antibody chains.

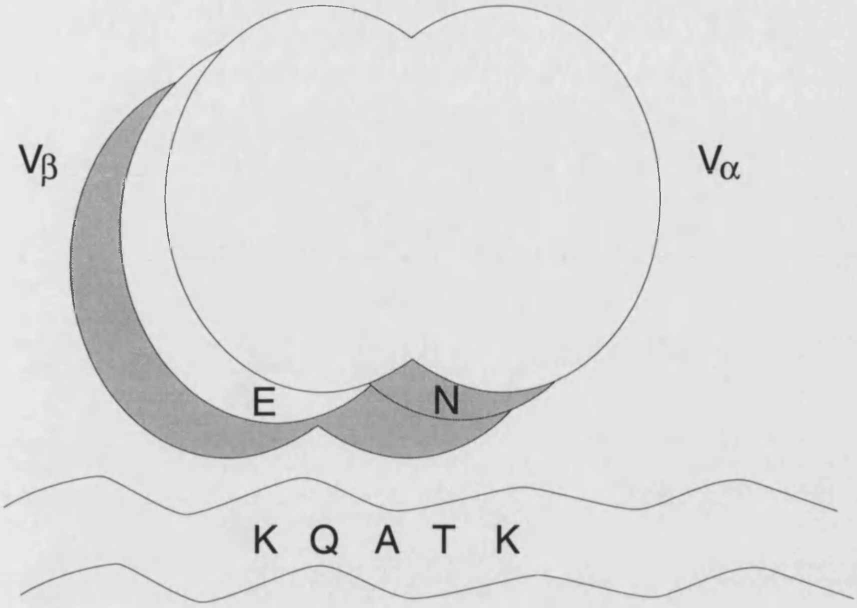


Figure 4.2: Schematic diagram showing mapped TCR-peptide contacts determined using transgenic mice (after Jorgensen *et al.* [164]).

4.3.4 The affinity of the TCR-MHC-peptide Interaction

The affinities of TCRs have proved more difficult to determine than those of antibodies. This has been partly because of the difficulties in producing soluble TCRs.

The first published affinities (K_A), by Matsui *et al* [174], were $4-5\times10^5M^{-1}$. This procedure involved inhibiting the binding of T cells to soluble MHC-peptide complexes using anti-V_β antibodies. Weber used inhibition of T cell activation under antigen limiting conditions by soluble TCR with anti-MHC antibodies. This gave an affinity range of $10^5 - 5 \times 10^6M^{-1}$.

Other studies have used the BIAcore (Pharmacia, Sweden) instrument, in which soluble TCR was bound to a sensor chip and the MHC-peptide complex passed over it [175]. These experiments obtained similar affinities to those described previously. These affinity values were similar to non-somatically mutated IgM antibodies and, as T cell receptors do not undergo somatic mutation, this was considered reasonable.

However, in more recent studies [176,177], affinities of $1 - 2 \times 10^7 M^{-1}$ have been found. Also different responses to different affinity ligands are involved in the selection process in the thymus [178]. High affinity interactions lead to negative selection, while lower affinity ones lead to positive selection. The lowest affinity TCRs (in which no complexes were detectable) underwent no selection. It has been noted that TCRs to allogeneic MHC molecules can have higher affinities than those to syngeneic MHC molecules. This can be explained by the lack of allogeneic molecules in the thymus, thus avoiding the negative selection process [176].

Even these higher TCR affinities are still lower than the affinities of many antibodies which can be as high as $10^{12} M^{-1}$. As a result it has been difficult to reconcile the high sensitivity of TCRs with their low affinity. One explanation invokes serial engagement of multiple TCRs by a single MHC peptide molecule [179], while multimerization of TCRs and MHCs has also been shown to trigger T cells.

4.3.5 Other TCR Molecular Modelling Studies

Novotny *et al.* [180] modelled a single chain TCR V region constructed using an antibody Fv. The CDRs were modelled by conformational search. The model was used successfully to predict which surface hydrophobic sidechains should be mutated to increase solubility.

Buchwalder *et al.* [181] created a model of a fluorescein binding TCR. First, structurally conserved regions were identified in a set of eight antibody structures with peptide or hapten antigen specificities. For each of these the most homologous to the equivalent region in the TCR was identified and used to generate a hybrid framework. A database search method was used to assign coordinates for the CDR regions. The PDB was searched for loops of the correct length which had the best RMS match in the flanking residues. The amino acid sequence was changed to that of the TCR. The final stage was a conformational search on the CDR regions using molecular dynamics. From the model residues were identified which were predicted to be involved in fluorescein binding. Mutation of some of these residues showed that they were involved.

4.4 Modelling Strategy

AbM (CAMAL) was designed for modelling antibodies. As TCRs are members of the immunoglobulin superfamily it was considered valid to attempt to use AbM for modelling TCRs.

As a transitional preliminary step it was decided to model another immunoglobulin superfamily member for which a structure had been determined by X-ray crystallography. The protein selected was CD4 which is generally more similar to an antibody light chain than an antibody heavy chain, as with both TCR chain types.

The CDR2 region of the TCR β is a very different length to the antibody light chain CDR2 and is in fact more similar to antibody heavy chains. If a loop of the length of a β chain CDR2 is positioned using the take off angles from the light chain bad clashes occur. So this segment of the light chain was replaced with the CDR2 region from the heavy chain carrying with it the heavy chain take-off angles. This was first tested in the CD4 modelling which also has a CDR2 loop which is different to the light chain.

In choosing the T-cell receptor to model there were two requirements imposed. Firstly, the sequences of both the α and β chains of a single clone should be known. Secondly, the peptide antigen sequence and MHC specificity should be known. Ideally the MHC restriction should be to an MHC molecule with a solved structure. Choosing TCRs for which some sequence-function data are available would allow comparison of the models with experimental data. A set of sequences with these features was the group sequenced by Jorgensen *et al* against MCC [164]. Once reasonable results had been obtained with these sequences, a second TCR was modelled in order to suggest mutations which would affect binding. This allowed testing of the ability of the modelling procedure to produce models which had predictive value.

4.5 Structural Analysis of Immunoglobulins

When modelling loops in proteins two factors which greatly affect accuracy are the fidelity of placement of the end points and the angle of the loop compared to the framework strands on either side of it. As the antibody modelling studies show, the take-off angle for the CDRs is important for their accurate prediction. Individual TCR CDR loops show different degrees of homology to the antibody heavy and light chains. The take-off angles should therefore be selected from its most appropriate donor. To allow this, CDR segments from the most homologous antibody region were superimposed in the modelling procedure to give the highest probability of a correct take-off angle. The antibody heavy and light chains also have different structural definitions of the start and ends of the CDR regions. For modelling TCRs the end points should be conserved between the two chain types. To determine which regions to superimpose when modelling and the regions to define as the loops, the CDR regions of antibody heavy and light chains and CD4 and CD8 were overlapped using regions on either side of the loop. The overlap ranges were changed until the RMS between the fitting regions and CDR regions reached a minimum. The fitting ranges are shown in table 4.1.

The three figures 4.3, 4.4 and 4.5 show superimpositions of the strands on either side of each of the CDRs and the CDRs in between. CDR 1 is defined differently in light and heavy chains. The regions of conservation between chain types are in the B and C strands. The end point for CDR2 is different in antibody heavy chains and light chains. The nearest conserved piece of structure is the D

Loop	Fitting Range 1	Fitting Range 2
L1	19 - 23	35 - 39
L2	46 - 50	61 - 65
L3	84 - 88	98 - 103
H1	18 - 22	36 - 40
H2	46 - 50	66 - 70
H3	91 - 95	103 - 107

Table 4.1: Fitting ranges for superimposing CDR loops. These ranges define the first structurally conserved regions on either side of the CDR regions that are conserved between chain types (Kabat residue numbering [129]).

strand. CDR H2 is variable in structure up to this point whereas the light chain has a structurally conserved C''-D loop. There is conservation of an RF sequence motif after CDR L2 (95%) and CDR H2 (42%) at the base of the D strand (positions 75 and 76 (residue numbering as in figure 2.2)). The residues contacting this motif are conserved. The high conservation of the motif in TCRs (24% TCR α, 55% TCR β) suggests that it is likely to arrange these regions in a similar way to antibodies. Interestingly mutations in the RF motif from R to G occur in some of the β chain sequences. The arginine interacts with an aspartate at position 98 in the F strand. In some of the β chain sequences the conserved arginine is replaced by a glycine. In these sequences a different motif exists in the contacting residues. The aspartate at 98 is replaced by a glutamine and the residue after the arginine, which is usually phenylalanine, is replaced by a tyrosine. This set of compensatory changes should allow a hydrogen bond to form between the tyrosine and the glutamine, maintaining the structure (see table 4.2).

CD4 and CD8 are also structurally conserved compared to antibodies at the termini of the three CDR regions, indicating that this may be a conserved feature

Position			Frequency
75	76	98	
R	F	D	0.55
G	Y	Q	0.18
K	F	D	0.10
G	Y	P	0.04
N	F	D	0.03
E	S	H	0.02
Q	F	D	0.02
N	L	D	0.02
R	I	D	0.01
R	F	N	0.01
Q	S	D	0.01

Table 4.2: The frequencies of occurrence of the triplets of residues at positions 75, 76 and 98 in the TCR β chain are shown (residue numbering as in figure 2.2).

of the immunoglobulin superfamily. The same difficulty in defining the end of CDR 2 is found and the structurally conserved region is again the start of the D strand (see figure 4.4). The CDR 3 regions are superimposed on the F and G strands, the end points for light and heavy chains being similar.

4.6 Modelling CD4: A non Ab Immunoglobulin

4.6.1 Introduction

CD4 was used as a test case to see whether it was possible to model a non antibody immunoglobulin using the CAMAL algorithm. The aim was to identify which regions might be most difficult to model in T cell receptors and to test the use of a hybrid H/L(heavy/light) framework for modelling.

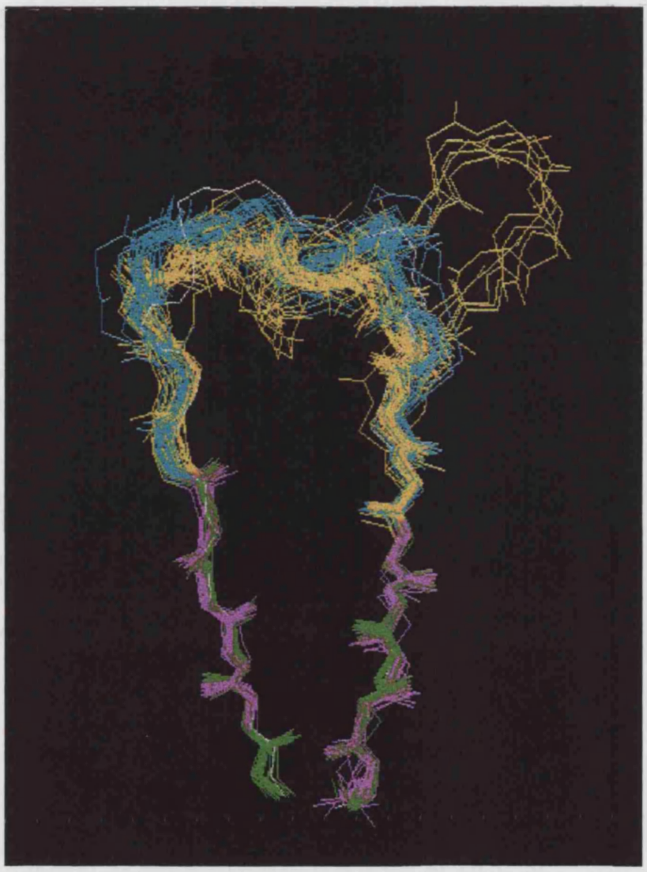


Figure 4.3: Overlapped CDR 1 regions of the structures from the Brookhaven full release and pre-release databases. Light chain framework residues are coloured green, heavy chain framework magenta, light chain CDR region yellow and heavy chain CDR region cyan. CD4 and CD8 chains are coloured white. The chains have been overlapped using five residues either side of the CDR.

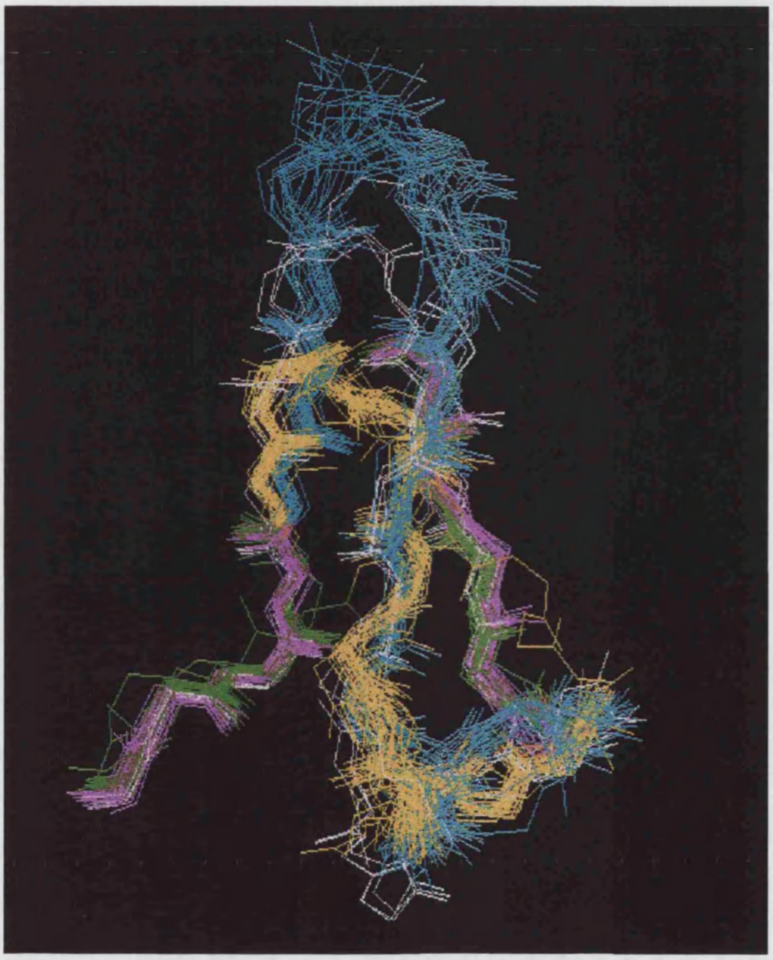


Figure 4.4: Overlapped CDR 2 regions of the structures from the Brookhaven full release and pre-release databases. Light chain framework residues are coloured green, heavy chain framework magenta, light chain CDR region yellow and heavy chain CDR region cyan. CD4 and CD8 chains are coloured white. The chains have been overlapped using five residues either side of the CDR.

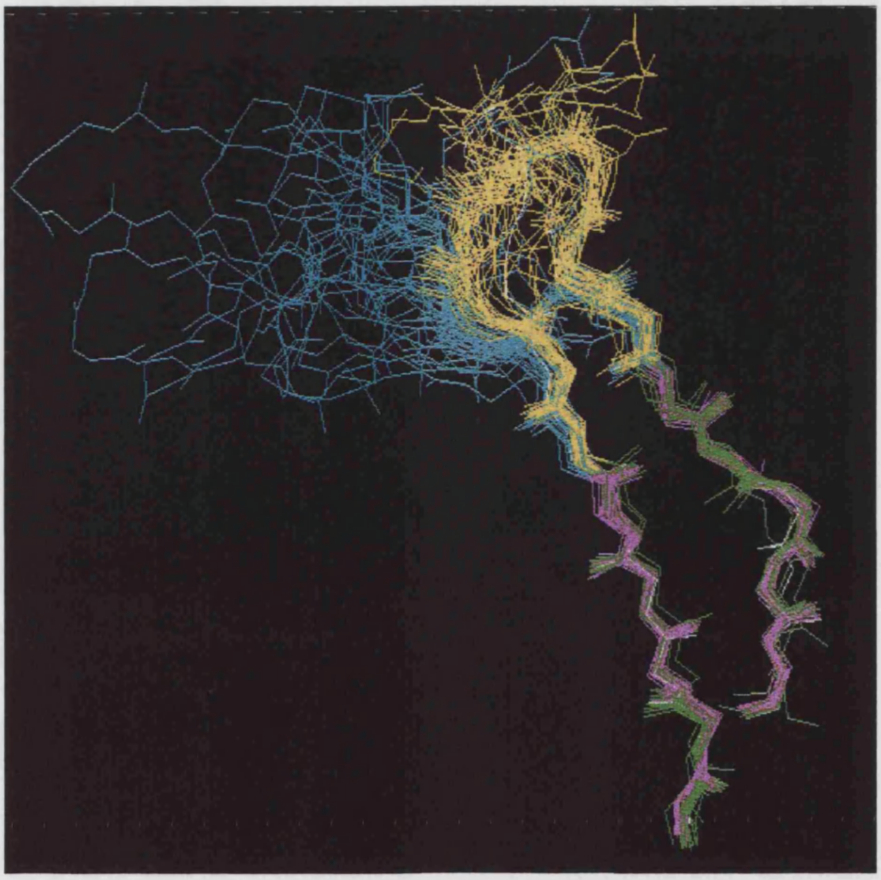


Figure 4.5: Overlapped CDR 3 regions of the structures from the Brookhaven full release and pre-release databases. Light chain framework residues are coloured green, heavy chain framework magenta, light chain CDR region yellow and heavy chain CDR region cyan. CD4 and CD8 are coloured white. The chains have been overlapped using five residues either side of the CDR.

4.6.2 Method for Modelling CD4

One chain of the light chain dimer Rei was used as a framework. The sequence was changed to that of CD4 in the framework regions, using a maximum overlap approach in INSIGHT (Biosym), and this framework energy minimised in DISCOVER (Biosym) to a derivative of 0.1. The loop between the B and C strands was modelled using the loop builder in INSIGHT (Biosym). The CDR2 region of Rei was replaced with that of the most identical heavy chain sequence in this region (R19.9 [59]) by superimposing the C and D strands. The CAMAL algorithm was then applied to the CDR equivalent loops. CDR 1 and CDR 2 equivalent loops were modelled using a combination of database and *ab intio* searching. The constraints used in the database search were those derived from antibody structures. The CDR 3 loop was five residues long and was modelled with CONGEN alone. The D-E equivalent loop of CD4 was modelled using CONGEN in accordance with the original CAMAL algorithm. The sequence alignment of Rei and CD4 is shown in figure 4.6 which also indicates the CDR ranges.

4.6.3 CD4 Modelling: Results

The RMS deviations for the different regions for the model of CD4 are shown in table 4.3. A comparison of the model and the structure is shown in figure 4.7. There are two structures for CD4 domain 1 [43,44]. These agree in the positioning of the CDR 2 and CDR 3 equivalent regions but differ in the position of the CDR 1 equivalent region. The deviation along the chain for the model versus structure

```
                                   CDR 1
             10        20          30          40
REI  DIQMTQSP-SSLSASVGDRVTITC  QASQDIIKTLN  WYQQTPGK
CD4  --------KKVVLGKKGDTVELTC  TASQKKSIQFH  WKNS----
                  CDR 2                D-E
           50          60         70           80
REI   APKLLIY  E-A--SNLQAG-  VPSRFSGSG  SG---  TDYTFTIS
CD4   NQIKILG  NQGSFLTKGPSK  LNDRADSRR  SLWDQ  GNFPLIIK
                    CDR 3
         90          100        110
REI  SLQPEDIATYYC  QQYQSLPYT  FGQGTKLQIT
CD4  NLKIEDSDTYIC  EVE----DQ  KE-EVQLLVF
```

Figure 4.6: Structural alignment of CD4 and REI showing the regions which were modelled as CDR loops.

1cd4.brk is shown in figure 4.9. The CDR 1 and CDR3 regions of the model have low global backbone RMS values of 1.62 Å and 1.31 Å. CDR 2 has an RMS of 2.4 Å which is comparable to that of many of the antibody loops modelled using CAMAL. The conformations of the model CDR equivalent loop regions compared to the structures are shown in figure 4.8. The largest deviations from the crystal structures are in the D-E loop region. This has a different conformation to either light or heavy chains which was not modelled well using CONGEN (RMS of 3.5 Å). In the case of CDR 2 the model was built on a heavy chain CDR 2 which had been spliced into the framework. This shows that reasonable models can be generated using a hybrid H/L framework. The model also indicates how structurally conserved short (10-11 residues) CDR 1 loops are. These regions could be modelled first and included in all further loop modelling.

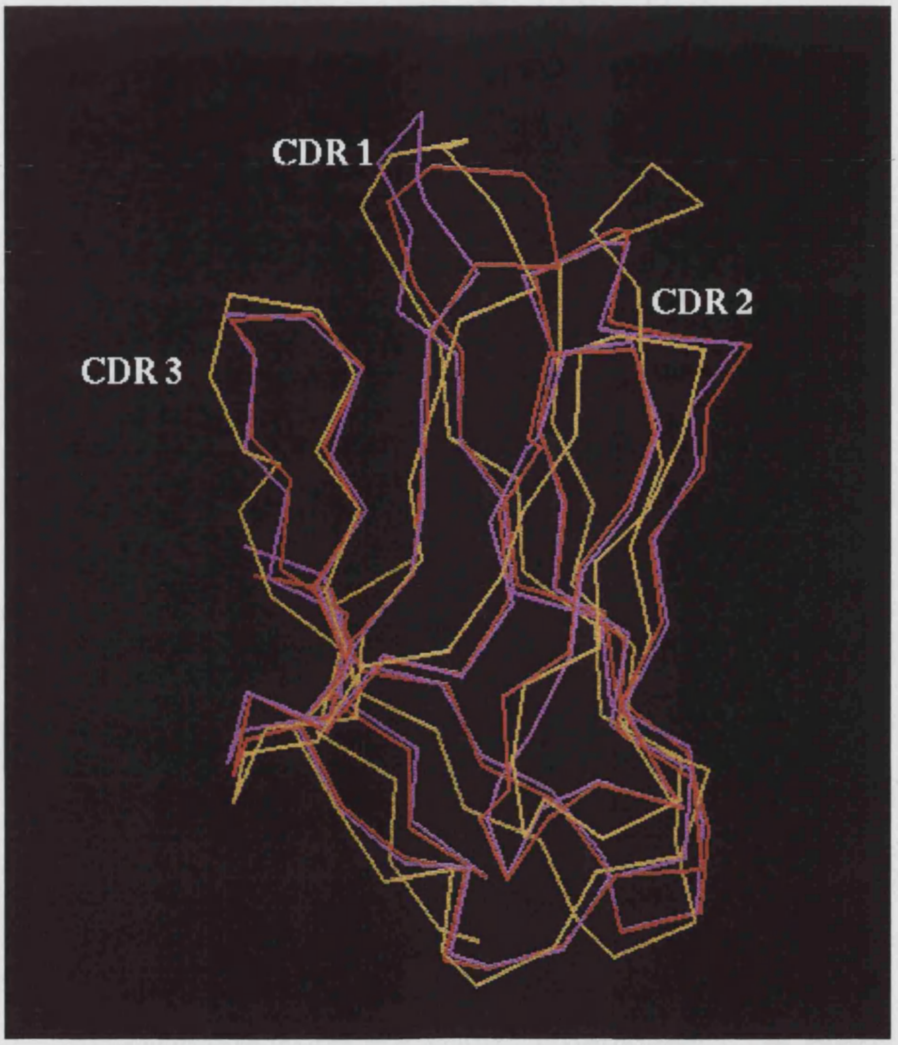


Figure 4.7: The molecular model of CD4 N-terminal domain superimposed on the two structures with Brookhaven accession numbers 1CD4 and 2CD4. The model is coloured yellow, 1CD4 is coloured magenta and 2CD4 is coloured orange.

Loop	Global Backbone RMS (Å)	Local Backbone RMS (Å)
CDR 1 (25 - 35)	1.62	0.96
CDR 2 (51 - 62)	2.40	1.34
CDR 3 (97 - 105)	1.31	0.54

Table 4.3: The global and local backbone RMS deviations for the CDR equivalent loops in CD4 (Brookhaven accession number 1CD4). The residue numbering is as in figure 4.6.

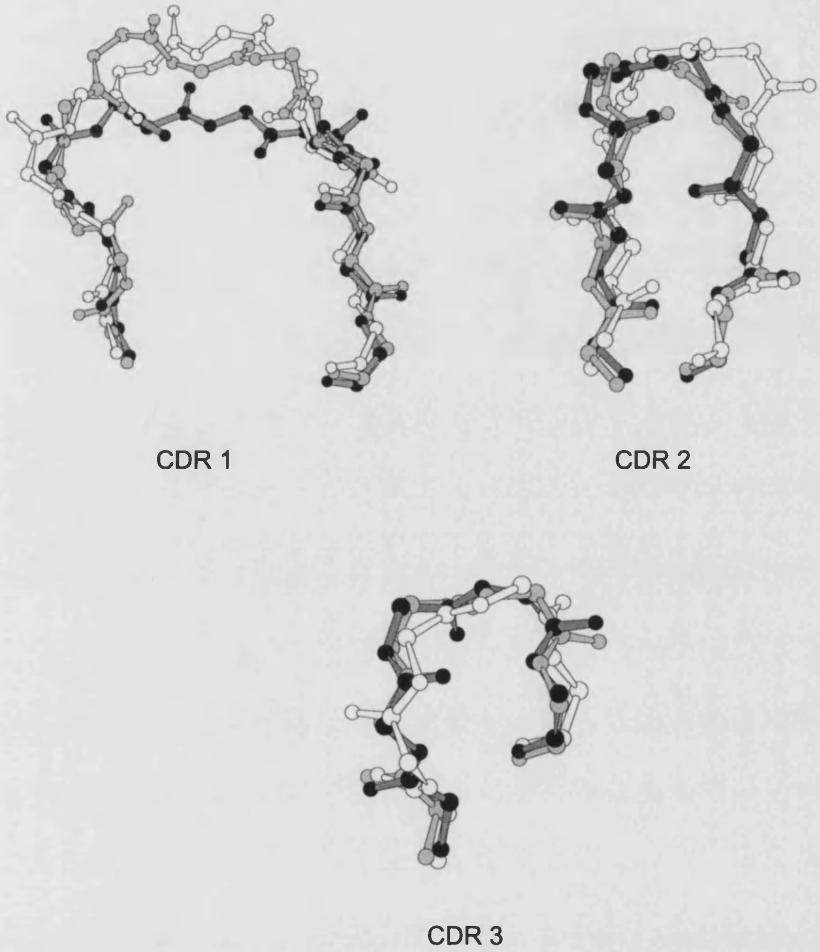


Figure 4.8: A backbone plot of the loop conformations in the model (white) and the two crystal structures of CD4 (Brookhaven accession numbers 1CD4 (dark grey) and 2CD4 (light grey)).

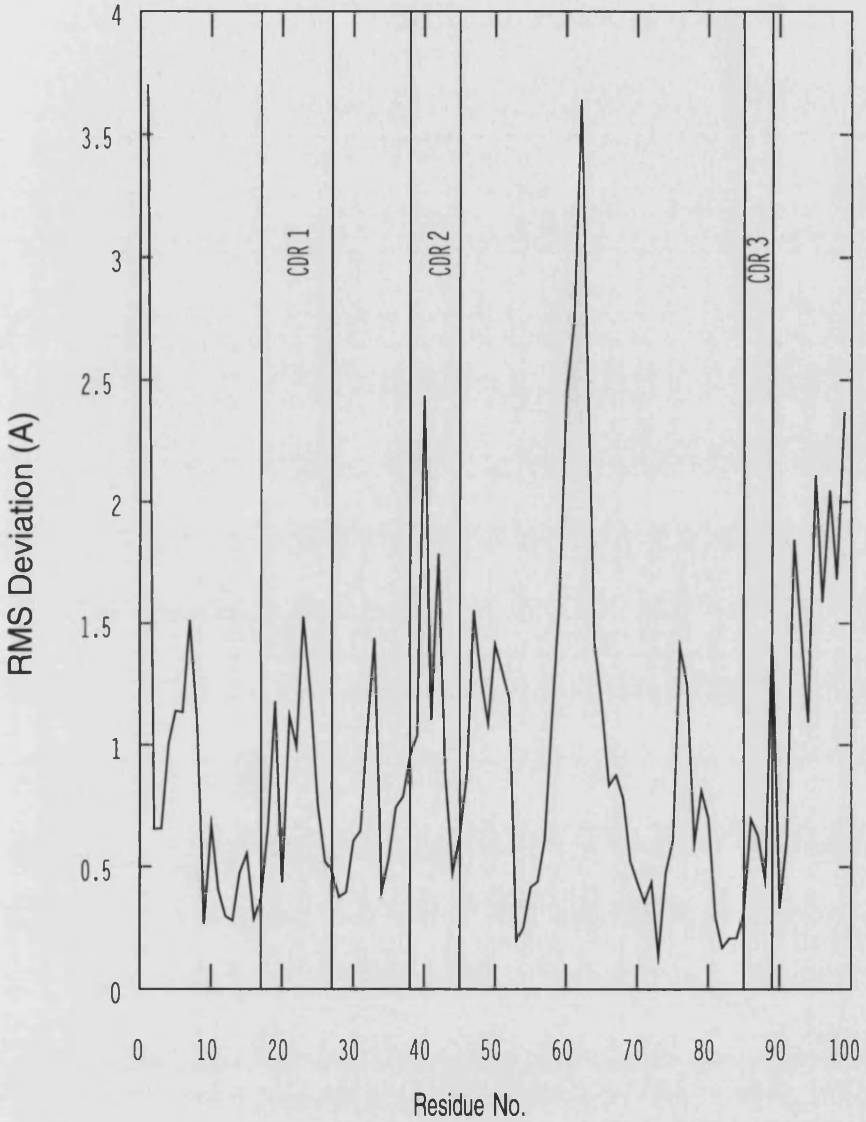


Figure 4.9: A plot of the backbone RMS deviation along the chain for the CD4 model compared to the CD4 structure with the Brookhaven accession number 1CD4. The CDR regions are sectioned off. The high peak after the CDR 2 region shows the poor modelling of the D-E loop. The residue numbering is the same as in 1CD4.

4.6.4 CD4 Modelling: Conclusions

The modelling of CD4 showed that it is possible to model a member of the immunoglobulin superfamily using an antibody V domain, when the sequence alignment is known. Previous modelling of CD4 failed because the alignment was incorrect [44]. In the case of TCRs the alignment against antibodies is easier because of the high conservation of certain key residues such as the cysteines at the start of CDR 1 and CDR 3, the LLIY (residues 53–56 (see figure 2.2 for residue numbering)) motif at the start of CDR 2, an RF (residues 75–76) motif at the end of CDR 2, a WY (residues 42–43) motif at the end of CDR 1 and an FG (residues 138–139) motif after CDR 3. The modelling of CD4 CDR-equivalent regions was quite accurate, CDR 1 being well modelled. CDR 3 was short and modelled well using conformational search. A problem arose with CDR 2. This region has a different conformation to either light or heavy chains. It is more similar to a heavy chain CDR than a light chain CDR and was modelled using a heavy chain CDR spliced into the framework. The general positioning of the loop was accurate. However the precise conformation was not because of contacts with the non CDR D-E loop which was poorly modelled. The rest of the framework was accurately predicted. In TCRs the D-E loop is more variable than antibody D-E loops and CDR2 is more variable in length than the antibodies. These facts suggest that this CDR would be more difficult to model accurately than the other CDRs.

4.7 Modelling TCRs from Transgenic Mice

4.7.1 Introduction

The CD4 modelling exercise showed that it is possible to produce a reasonable model of a non-antibody Ig domain using CAMAL. The next stage was to apply this experience to TCRs. The sequences chosen were those for which some sequence function relationships were known.

4.7.2 Method for Modelling TCRs

TCR frameworks were modelled using the light chain dimer Rei [182]. Where the Rei and TCR sequences differed, the TCR sidechains were inserted using a maximum overlap approach and the framework minimised to a derivative of 0.1 in DISCOVER(Biosym). The CDRs were modelled onto this framework using CAMAL [6]. In the α chain the loop regions modelled were positions 23–33, 49–55 and 91–105; in the β chain the regions were 24–33, 48–59 and 93–107 (Kabat residue numbering [129]).

The sequence alignments of the sequences of 5C.C7 against Rei are shown in figure 4.10, as are the three CDR $\beta 3$ variants of 5C.C7 [164] which have been modelled. For loops of five residues or less, CONGEN [144] was used to generate conformations. The lowest energy conformation was then chosen. For loops of six or seven residues, Cα distance constraints within the loop were used to search the Brookhaven database [183] for structural segments which matched within a tolerance of 3.5 standard deviations. Sidechains were replaced with those of the desired loop sequence in CONGEN and were energy screened using a solvent modified EUREKA potential [160] . A torsional screen using the initial database loops was then performed to choose the final conformation. For loops greater than seven residues a combination of database searching and conformational searching for the central five residues of the loop was performed. The energy screen and torsional screen were then performed on the generated conformations. In all cases each loop was constructed in the absence of the others. Distance constraints for antibody CDR loop regions were used to extract database conformations for the CDRs. For CDR $\alpha 1$ and CDR $\beta 1$ the CDR L1 constraints were used. For CDR $\alpha 2$ the CDR L2 constraints were used. For CDR $\beta 2$ the CDR H2 constraints were used. The range modelled for CDR $\beta 2$ does not include the C"-D loop. This is because the loop was defined for H2 modelling, and also because no database search hits were produced if the longer range including the C"-D loop was included. For both CDR $\alpha 3$ and $\beta 3$ the CDR L3 constraints were used. A five residue segment, predicted to be the D-E loop in both chains, was also modelled using CONGEN (residues 67–71 in α and 70–74 in the β chains (Kabat residue numbering [129])). This region was also ignored when the CDRs were built. The final model was minimised in DISCOVER to a derivative of 0.1.

4.7.3 Results: Comparison to experimental data

Models have been created for four TCR sequences. These are related, having the same Vα and Vβ gene segments and only varying in the joining regions of the α chain. These sequences were taken from Jorgensen's recent TCR MHC mapping

α Chain

```
                                            CDR 1
                10          20            30           40
REI    DIQMTQSPSSLSASVGDRVTITC  QASQDIIKTLN  WYQQTPGK
ALPHA  -DQVEQSPSALSLHEGTGSALRC  NFTTT-MRAVQ  WFRKNSRG
                 CDR 2                    D-E
                50          60          70
REI    APKLLIY  EASNLQAG  VPSRFSGSGS  G--TD  YTFTIS
ALPHA  SLINLFY  LA--SGTK  ENGRLKSAFD  SKERY  STLHIR
                              CDR 3
         80         90            100           110
REI     SLQPEDIATYYC   QQYQ--SLPYT   FGQGTKLQIT
ALPHA   DAQLEDSGTYFC   AAEASNTNKVV   FGQGTILKVY
```

β Chain

```
                                            CDR 1
                10          20            30           40
REI    DIQMTQSPSSLSASVGDRVTITC  QASQDIIKTLN  WYQQTPGK
BETA   -SKVIQTPRYLVKGQGQKAKMRC  IPEKG-HPVVF  WYQQNKNN
                 CDR 2                         D-E
                50          60            70            80
REI    APKLLIY  E-A--SNLQAG-   VPSRFSGSG   SG-TD  YTFTIS
BETA   EFKFLIN  FQNQEVLQQIDM   TEKRFSAEC   PSNSP  CSLEIQ
                          CDR 3
               90          100          110
REI    SLQPEDIATYYC  QQYQ---SLPYT  FGQGTKLQIT
BETA   SSEAGDSALYLC  ASSLNNANSDYT  FGSGTRLLVI
```

α Chain CDR 3 Regions.

Clone	CDR α3 Sequence	Peptide Specificity
5C.C7	C A A E A S N T N K V V F G	MCC (99K)
116	C A A E A S A G N K L T F G	MCC (99K)
202	C A A K S S G S W Q L I F G	99E
226	C A A E P S S G Q K L V F G	Cross Reactive

Figure 4.10: Sequence alignment of the modelled TCR sequences against Rei [57]. The table below the alignment shows the four CDR α3 regions for the clones modelled.

study [164] because of sequence-function data defining the MHC -peptide- TCR interacting residues. The models could therefore be tested for their ability to predict the residue positions important for peptide binding. The TCR models showed that the Jorgensen interacting residues were solvent exposed.

The peptide sequence which the TCRs bound (MCC peptide) was modelled onto the sequence of the peptide in one of the MHC antigen structures (1vab.brk) [25] and the TCR model has been overlaid onto this peptide. The peptide is probably correctly oriented because the TCR contacting residues were exposed and the MHC antigen contacting residues pointed into the MHC antigen. The residues in the TCR α and β chains (residues 93α and 98β (residue numbering as in figure 4.10)), considered to be important for MHC binding, could be oriented to both be less than 5Å away from the peptide contact residues, with the TCR oriented with its two CDR 3's over the peptide and the CDR 1 and 2 regions over the MHC α helices. This distance could be reduced if the side chain conformation at residue 93 was more exposed. In this orientation the CDR 1 and 2 of the α chain contact an MHC helix, as does CDR 2 in the β chain. CDR 1 in the β chain did not contact either peptide or MHC. The D-E loop of the α chain modelled conformation was similar to that of a heavy chain D-E loop. The tip of the D-E loop in the β chain pointed in towards the CDR β1 region.

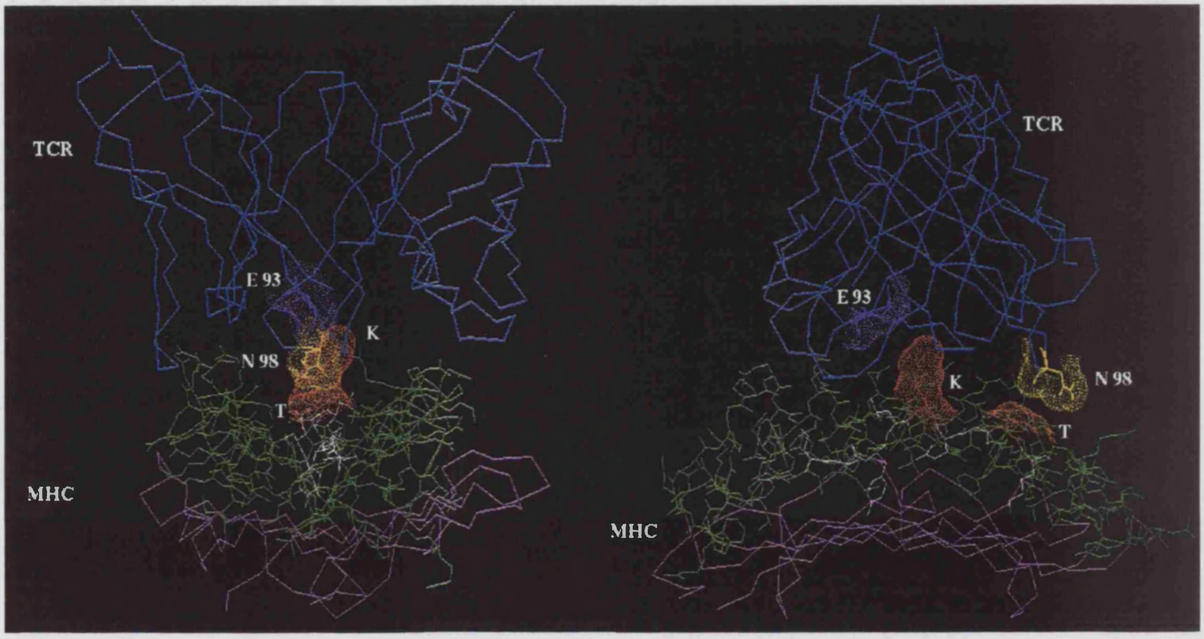


Figure 4.11: Two views of the TCR - peptide - MHC antigen model. The TCR is coloured blue. Only the MHC antigen N- terminal domain is shown. The MHC α helices on either side of the peptide binding cleft are coloured green. The β sheet base of the binding cleft is coloured magenta. The peptide is coloured white. The two TCR residues predicted to be important for TCR peptide interaction are coloured magenta and yellow. The respective residues in the peptide are coloured red and orange. The peptide sequence of the MCC peptide has been modelled onto the peptide from 1vab.brk. The TCR has been positioned so as to minimise the distances between the two contact residues and their peptide partners while avoiding clashes between the TCR and the MHC antigen. The right hand view shows the MHC antigen helices side on while the left view shows them from one end.

4.8 Modelling a TCR to Suggest Mutations

4.8.1 Introduction

HA1.7 is a DR1 restricted, influenza hemagglutinin peptide (307-319, PKYVKQNTLK-LAT) specific, TCR. Wedderburn *et al* (ICRF, London) had expressed this TCR as a chimera with CD3ζ in a rat basophil line and it was therefore amenable to site-directed mutagenesis. The aim of this modelling study was to identify one or more peptide interacting residues in the TCR. These predictions would then be tested by site-directed mutagenesis by Wedderburn *et al.*

4.8.2 Methods

The modelling method for this TCR was essentially similar to the previous modelling. However there are some differences between the automated procedure implemented in AbM and the more manual method described above. Firstly the framework regions were modelled using the program FRAMEBUILD [136]. This positioned the subunits and performed the maximum overlap sidechain replacements. The splicing of the CDR H2 loop into the Rei light chain framework was performed by the canonical loop modelling program CHOTH using the conserved C' and D strands.

α Chain

```
                                  CDR 1
                10        20           30          40
REI    DIQMTQSPSSLSASVGDRVTITC QASQDIIKYLN  WYQQTPGK
ALPHA  DQSVTQLGSHVSVSEGALVLLRC NYSSSVPPYLF  WYVQYPNQ
                  CDR 2                 D-E
                50         60           70
REI    APKLLIY EAS--NLQAG  VPSRFSGSGS  G-TD YTFTIS
ALPHA  GLQLLLK YTSAATLVKG  INGFEAEFKK  SETS FHLTKP
                        CDR 3
       80       90         100         110
REI    SLQPEDIATYYC QQYQS----LPYT  FGQGTKLQIT
ALPHA  SAHMSDAAEYFC AVSESPFGNEKLT  FGTGTRLTII
```

β Chain

```
                                 CDR 1
               10        20           30          40
REI   DIQMTQSPSSLSASVGDRVTITC QASQDIIKYLN  WYQQTPGK
BETA  DVKVTQSSRYLVKRTGEKVFLEC VQDMDH-ENMF  WYRQDPGL
                  CDR 2                 D-E
               50        60             70          80
REI   APKLLIY  EAS---NLQAG   VPSRFSGSG  SG-T  DYTFTIS
BETA  GLRLIYF  SYDVKMKEKGD   IPEGYSVSR  EKKE  RFSLILE
                       CDR 3
               90         100         110
REI   SLQPEDIATYYC QQYQS---LPYT  FGQGTKLQIT
BETA  SASTNQTSMYLC ASSSTGLPYGYT  FGSGTRLTVV
```

Figure 4.12: Sequence alignment of the HA1.7 sequence against Rei used for molecular modelling.

4.8.3 Results: Predictive value

The models of the human TCR HA1.7 were used to predict a peptide contacting residue. The model predicted that a glutamic acid residue in the CDR $\alpha3$ region interacted with lysine 316 of the peptide (see figure 4.13). This residue was backmutated by Wedderburn *et al* [184] and binding was eliminated. This model at least appears to have predicted a residue which may be involved in peptide contacts. Other explanations are possible, for example the single residue change could have altered the conformation of other residues, eliminating their interactions [185].

Wedderburn attempted to make complementary changes in the peptide to restore binding but this did not succeed. Recently it has been shown that the conformation of the MHC helices can change significantly when binding different peptides.

4.9 Conclusions

The models generated were consistent with the mode of binding described by Chothia and Lesk [127] and Davis and Bjorkman [186] for the TCR to the MHC and peptide. In this arrangement CDR $\alpha1$ and CDR $\alpha2$ are supposed to contact one MHC helix, and CDR $\beta1$ and $\beta2$ are supposed to contact the other, with the CDR3 regions contacting the peptide. In the models, no contacts were observed with the CDR $\beta1$ region, which also does not show significantly greater sequence variability than the framework. These two pieces of evidence are therefore con-

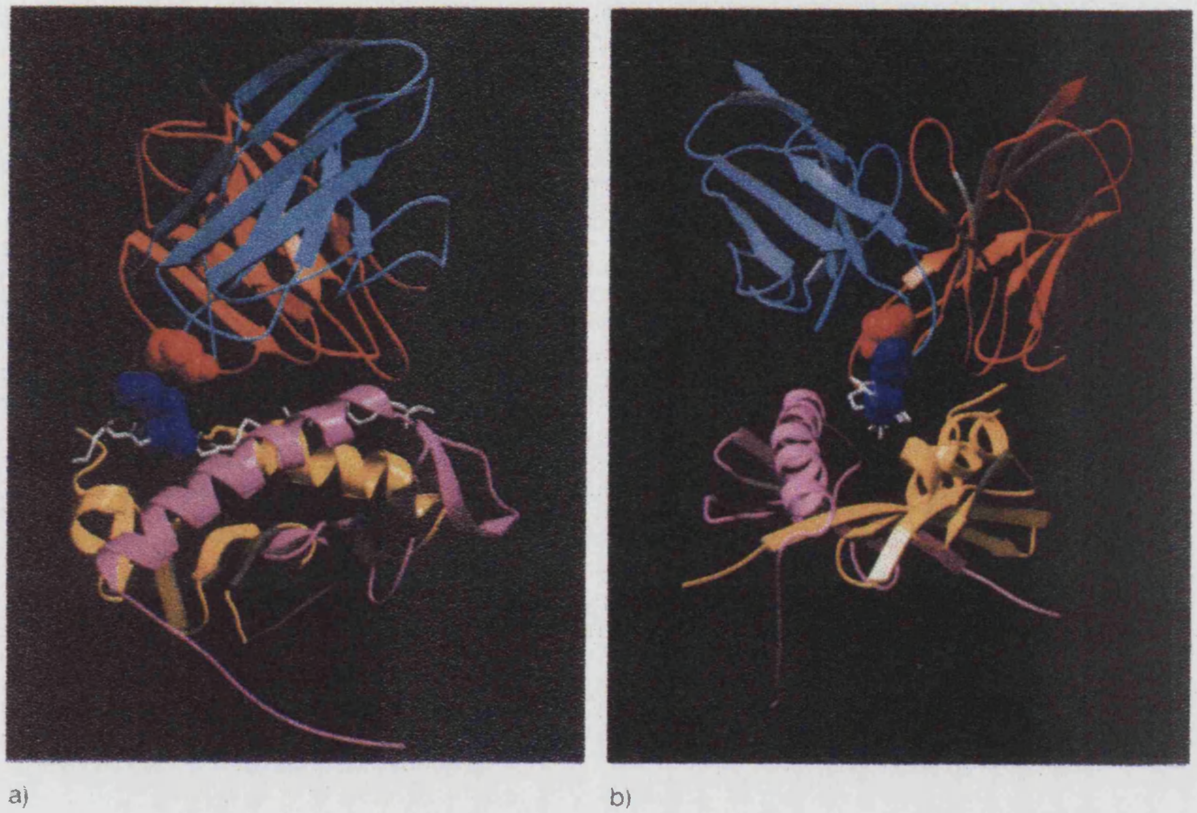


Figure 4.13: Molecular models showing the predicted interaction between TCR HA1.7 and the HA 307-319/DR1 complex. The Vα and Vβ chains are shown in blue and orange, respectively and the MHC α and β chains in pink and yellow, respectively. The peptide is shown in white and is orientated N terminus (right) to C terminus (left) in the two views a) and b). a) side view of the TCR-MHC complex showing the proximity of residue TCR Vα 95E (see figure 4.12 for residue numbering) (space filled red) to peptide residue 316K; b) end view of the TCR-MHC complex obtained by rotation of view a) by 90°. The CDR3 loops of the Vα and Vβ chains can be seen to be located above the peptide, while the CDR1 and CDR2 regions are positioned over the MHC helices. Figures were generated using MOLSCRIPT [26].

sistent with one another.

The models also suggest a reason why there is a more restricted length distribution for the CDR $\alpha 1$ loop compared to the CDR L1 loop of the antibody κ chain. The long CDR L1s have a bulge before the start of the C strand which, in the model of the TCR MHC complex would clash with the MHC.

The structure of a monomeric TCR β chain [3], a homodimer of TCR α chains [4] and, recently, of a mouse MHC class I and a human MHC class I restricted TCR $\alpha\beta$ heterodimer [1,2] have been published. Only the coordinates of the TCR β monomer are at present available from the Brookhaven databank. However, the descriptions of the structures provide some information for comparing with the models.

There follows a description of certain of the features of these structures and comparisons are made to the models of the TCRs.

The crystal structure of the β chain monomers shows higher structural homology to the V_L in its framework regions. The V_α homodimer structure also shows higher similarity, in these regions, to the V_L than the V_H. The light chain dimer framework of the models will reflect this.

The residues involved in the V_α-V_β interface have been identified in the $\alpha\beta$ heterodimer structures. The residues involved are the same as those in antibodies.

The interface in the mouse $\alpha\beta$ heterodimer buries only 1160Å^2 which is at the lower limit of that seen in antibody Fabs. In the human dimer the interface buries about 1575Å^2 which is near the upper limit seen in the Fabs. The models of the TCRs bury around 1500Å^2 which is similar to the human dimer. However in several of the light chain dimers for which structures are available, only about 1100Å^2 are buried.

In the TCR structures the precise conformation and stabilising interactions in the CDR $\alpha 1$ and $\beta 1$ loops is different to that seen in antibodies. However the general conformation is similar. The models are based on light chain constraints for this loop region so the models generated are reasonable.

However a feature unique to the α domain is a strand switch of the C" strand from the inner to the outer sheet. This means that the conformation of the CDR2 region, i.e. the C'-C" loop, the C" strand and the C"-D loop, is significantly different to the light chain CDR2 region. The local conformation of this loop is similar to the CDR H2 region in antibodies. The model uses light chain constraints for this region and is therefore inaccurate.

The conformation of the CDR $\beta 2$ region in the structures was more similar to the CDR H2 region than the CDR L2 region and the local conformation of this loop was also similar to the CDR $\alpha 2$ region. In the models the use of H2 constraints for this region therefore appears to be valid.

The conformations of the D-E loop in the β chain structures have the tip of the loop pointing towards the CDR1 loop. This was seen in the models where the region was modelled using CONGEN. The D-E loop in the α chain folds away from CDR1. The models also display this feature, although the exact conformation is different.

The CDR $\beta 3$ region folds away from the core of the domain in the monomer. However in the two $\alpha\beta$ heterodimers the conformation is much more upright.

Garcia et al [1] suggest that, in the monomer, the absence of the α chain allows the CDR3 to fold away from the domain surface towards the solvent. The light chain dimer framework used means that the take-off angle generated for modelling the CDR $\beta 3$ is very different to that of a CDR H3 loop, but much more similar to that seen in the heterodimer structures. However, without the coordinates it is difficult to make an accurate comparison. The CDR $\alpha 3$ location within the combining site is very similar to that of a light chain CDR3 when an antibody F_V is superimposed.

In the space between the two CDR3 loops, both Vα-Vβ heterodimer structures have a cavity. In the human heterodimer this cavity has been shown to be occupied by an antigenic peptide residue. The models also show a cavity between these two CDRs.

The orientation of the TCR on the MHC in the structures of the complex is very different to that used in the modelling experiments. However the structures do show that the majority of contacts with the peptide are to the CDR3 loops of the TCR. Therefore, although the orientation was wrong, the interactions predicted are probably correct.

Overall, the choice of a light chain dimer probably produced a more accurate model than if a V_LV_H heterodimer had been used.

Chapter 5

Humanisation of Antibodies

5.1 Introduction to humanisation

Antibodies have high specificity and affinity. These properties are just those needed for drug targeting agents, and there has been much research into their use by pharmaceutical companies. However it is difficult to generate human antibodies. Rodent antibodies are often available with the required specificity but these are foreign proteins to the human patient and so produce an immune response eliminating the antibody from the system. To try to solve this problem changes are made to the antibody sequence to reduce the immune response against it. This process is known as humanisation.

The most used form of humanisation is CDR grafting in which the CDR loops of the rodent antibody are grafted onto the framework of a homologous human antibody. The first attempts at this involved grafting the CDRs from a mouse antibody onto the framework sequence of a known human structure (e.g. REI [57] or NEW [50]) [187–190]. However the grafting itself usually failed to produce antibodies which bound antigen. It was often necessary to perform "back mutations" to the mouse framework sequence to produce functional antibodies. Antibody modelling is useful in identifying residues to be backmutated and AbM has been used successfully to predict required back mutations.

As part of the development of a general immunoglobulin modelling algorithm, the differences between mouse and human antibodies were investigated. The results obtained indicate that the number of residues required to produce a given reshaped antibody is surprisingly small.

Two generally applicable methods for reshaping antibodies are described, the first based on CDR grafting using rational rules for generating the sequence, and the second on a new approach which was given the name resurfacing. In the latter approach many fewer changes to the mouse sequence need to be made as only the surface of the antibody is changed to that of a human antibody. The methods have been applied to two antibodies, both of which have been expressed and characterised.

This work has been described in detail in a series of papers [191–193]. Here a brief history of the main parts of the project will be given. A paper is also included describing the results to date for two antibodies humanised by both resurfacing and CDR grafting.

5.2 Creation of sequence database

For both approaches a comprehensive database of antibody sequences is required. Some V_L and V_H sequences are known to be expressed together. Choosing a known V_L-V_H pair has two advantages. Firstly the light and heavy chain in the pair should associate and secondly they will hopefully not generate epitopes which will give rise to a human anti-mouse antibody (HAMA) response.

Aligned human and mouse antibody F_V sequences were extracted from the Kabat database [129]. Clone information defining V_L-V_H pairings was retained.

5.3 Analysis of F_V surface residues

The following protocol was used to identify the surface residues in an antibody. The structures of 12 antibody structures from the Brookhaven database [183] were superimposed. The relative accessibility of each residue was calculated using the DSSP accessibility calculation routine [194]. Using this data for R19.9 [59] a cut-off was defined to identify accessible residues. A value of 30% was chosen [192]. Structurally aligned positions which had an average relative accessibility greater than the cut-off were defined as accessible. There was 98% identity between the antibodies in the positions which were above the cut-off.

The surface positions were then extracted from an alignment of light and heavy chain sequences. The sets of surface residues in mouse and human sequences were compared [192]. The main results from this analysis were, firstly that in no case was a human surface identical to a mouse surface. Secondly the homology between surfaces in the two species was high. Thirdly there were identical surfaces within a species.

5.4 Modelling

Both approaches rely on molecular modelling to identify CDR framework interactions. Models were generated using a modification of the CAMAL procedure [6,7] implemented in the program AbM [195]. The heavy and light framework regions for each model were chosen from the most homologous antibody of known structure. Where possible, the CDRs were modelled using a canonical loop [120]. For CDRs which did not fit any of the canonical classes, a combination of knowledge-based and *ab initio* methods were used to generate possible CDR conformations.

5.5 Method for CDR grafting

The mouse light and heavy chain sequences were aligned against the human sequence database. The human light heavy chain sequence pair with the highest total identity against the mouse sequence was determined. The CDRs in the human sequence were then replaced with those of the mouse antibody, and both the original mouse and initial CDR grafted sequences were modelled. All framework residues within 5Å of the CDRs were examined. Any which might interfere with the CDRs were back mutated. The backmutated sequences were modelled to determine whether the CDR conformation was improved.

5.6 Method for Resurfacing

The mouse light and heavy chain sequences were aligned against the human sequence database. The human light heavy chain sequence pair with the highest total identity against the mouse sequence was determined. The surface residues in the mouse antibody were replaced with the human sequence. Both the original mouse and initial CDR grafted sequences were modelled. All framework residues within 5Å of the CDRs were examined. Back mutations were made at those positions which might interfere with the CDRs. The backmutated sequence was modelled to determine whether the CDR conformation was improved.

5.7 Comparison of Methods

A comparison of the two methods is shown as a flow chart in figure 5.1. CDR grafting involves changing most of the framework to the human sequence whereas resurfacing only involves changing the surface of the antibody to appear human while retaining the mouse core framework (see figure 5.2). Therefore resurfacing should have a better chance of retaining binding while CDR grafting may give lower immunogenicity.

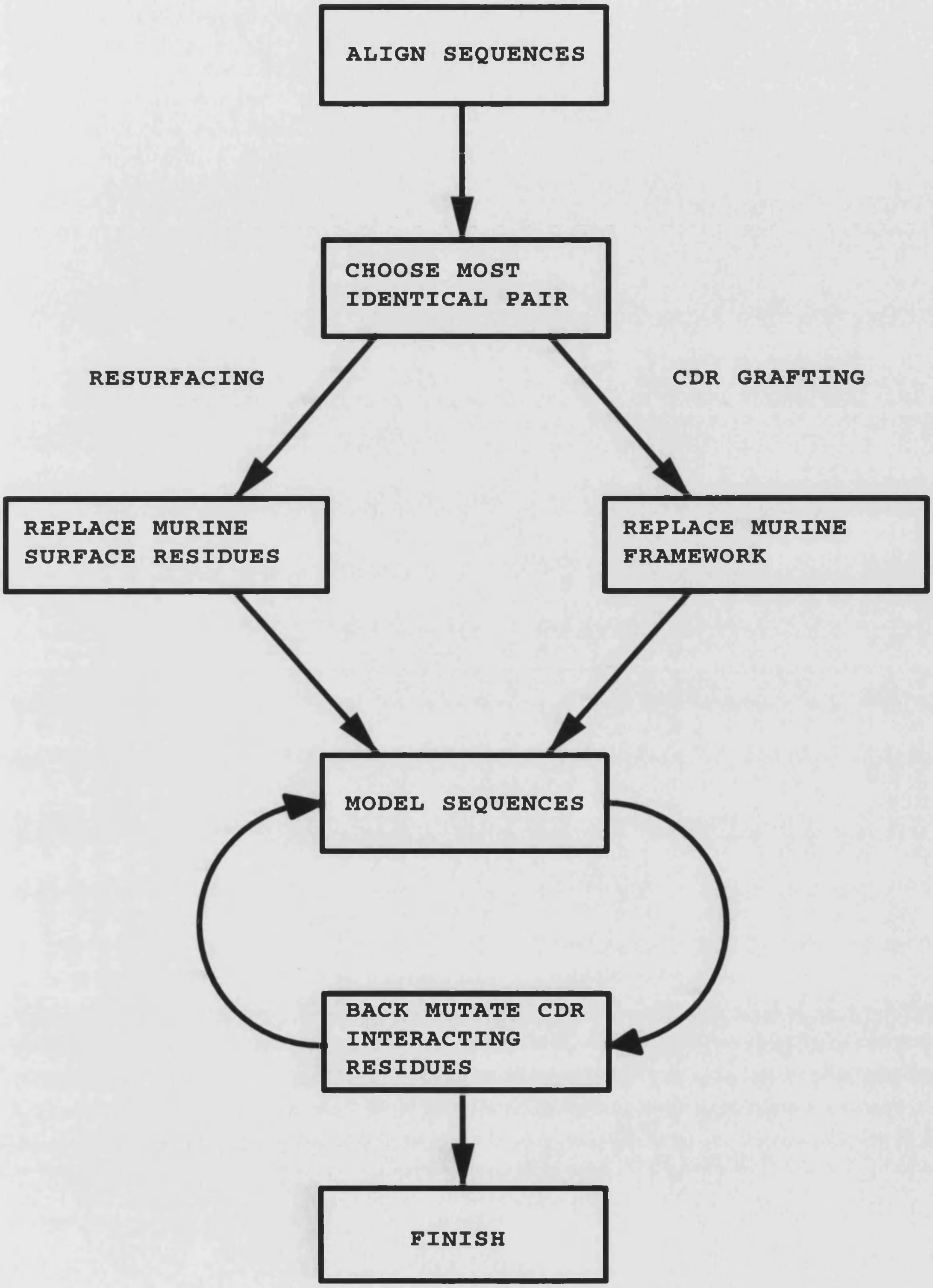

ALIGN SEQUENCES
CHOOSE MOST IDENTICAL PAIR
RESURFACING
CDR GRAFTING
REPLACE MURINE SURFACE RESIDUES
REPLACE MURINE FRAMEWORK
MODEL SEQUENCES
BACK MUTATE CDR INTERACTING RESIDUES
FINISH

Figure 5.1: A flowchart comparing the two humanisation methods.

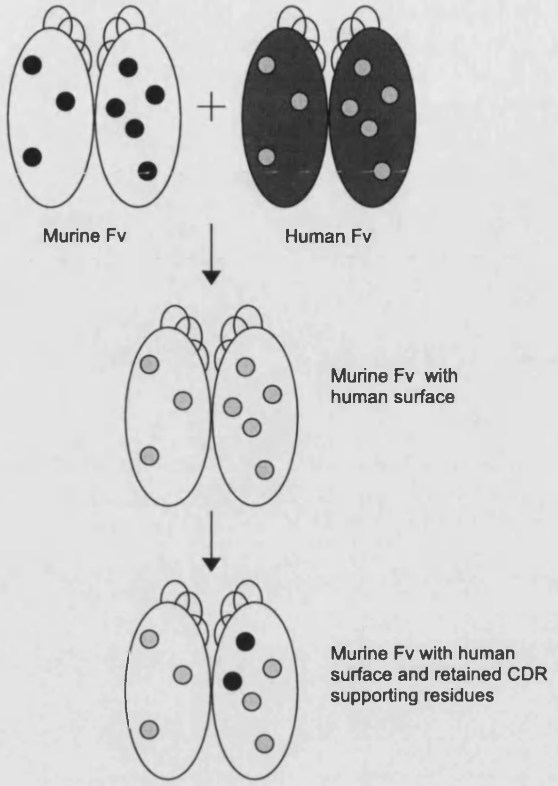

Murine Fv
Human Fv
Murine Fv with human surface
Murine Fv with human surface and retained CDR supporting residues

Figure 5.2: A schematic illustrating the resurfacing method for humanisation. In the first stage, the mouse framework (white) is retained and only the surface residues changed from mouse (black circles) to the closest human pattern (light grey circles). In the second stage, surface residues within 5Å of the CDRs that are predicted by molecular modelling to perturb the CDRs, are replaced with the original mouse residues. This ensures retention of antigen binding.

5.8 Paper describing the humanisation of B4 and N901

The paper included below describes the application of both CDR grafting and resurfacing methods to two antibodies [193]. The theoretical work was done in Bath by Andrew Henry, Jan Pedersen and the author. The experimental work was carried out by Immunogen Inc.

Protein Engineering vol.9 no.10 pp.895–904, 1996

A comparison of two murine monoclonal antibodies humanized by CDR-grafting and variable domain resurfacing

Michael A.Roguska[2], Jan T.Pedersen[1,3], Andrew H.Henry[1,5], Stephen M.J.Searle[1], C.Michelle Roja, Brian Avery, Mary Hoffee, Sherri Cook, John M.Lambert, Walter A.Blättler[6], Anthony R.Rees[1] and Braydon C.Guild[4]

ImmunoGen, Inc., 148 Sidney Street, Cambridge, MA 02139, USA,
[1]Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK
[2]Present address: BASF Bioresearch Corporation, 100 Research Drive, Worcester, MA 20850, USA
[3]Present address: Center for Advanced Research in Biotechnology (CARB), 9600 Gudelsky Drive, Rockville, MD, 20850, USA
[4]Present address: Genome Therapeutic Corp., 100 Beaver Street, Waltham, MA 02154, USA
[5]Present address: Oxford Molecular Limited, The Medawar Centre, Oxford Science Park, Oxford OX4 4GA

[6]To whom correspondence should be addressed

The variable domain resurfacing and CDR-grafting approaches to antibody humanization were compared directly on the two murine monoclonal antibodies N901 (anti-CD56) and anti-B4 (anti-CD19). Resurfacing replaces the set of surface residues of a rodent variable region with a human set of surface residues. The method of CDR-grafting conceptually consists of transferring the CDRs from a rodent antibody onto the Fv framework of a human antibody. Computer-aided molecular modeling was used to design the initial CDR-grafted and resurfaced versions of these two antibodies. The initial versions of resurfaced N901 and resurfaced anti-B4 maintained the full binding affinity of the original murine parent antibodies and further refinements to these versions described herein generated five new resurfaced antibodies that contain fewer murine residues at surface positions, four of which also have the full parental binding affinity. A mutational study of three surface positions within 5 Å of the CDRs of resurfaced anti-B4 revealed a remarkable ability of the resurfaced antibodies to maintain binding affinity despite dramatic changes of charges near their antigen recognition surfaces, suggesting that the resurfacing approach can be used with a high degree of confidence to design humanized antibodies that maintain the full parental binding affinity. By comparison CDR-grafted anti-B4 antibodies with parental affinity were produced only after seventeen versions were attempted using two different strategies for selecting the human acceptor frameworks. For both the CDR-grafted anti-B4 and N901 antibodies, full restoration of antigen binding affinity was achieved when the most identical human acceptor frameworks were selected. The CDR-grafted anti-B4 antibodies that maintained high affinity binding for CD19 had more murine residues at surface positions than any of the three versions of the resurfaced anti-B4 antibody. This observation suggests that the resurfacing approach can be used to produce humanized antibodies with reduced antigenic potential relative to their corresponding CDR-grafted versions.

Key words: antibody/immunotherapy/reshaping/resurfacing

Introduction

The method of immunoglobulin variable domain resurfacing (Pedersen *et al.*, 1994; Roguska *et al.*, 1994) attempts to reduce the immunogenicity of murine antibodies while maintaining their affinity and specificity by humanizing only the surface accessible residues located at conserved positions of the Fv framework. Differences in the presentation of surface residues in a small number of murine and human antibody variable regions have been described by Padlan (1991), and a statistical analysis of a large database of murine and human immunoglobulin sequences revealed that the variable domains of murine and human antibodies have distinct sets of surface residues that mirror the V-gene families (Kabat *et al.*, 1991; Pedersen *et al.*, 1994). No mouse framework displays the exact pattern of surface residues found in any human framework, but generally only a small number of changes are required to convert the set of surface residues in a murine Fv to that of the most identical surface pattern found on a human Fv. Thus, a general algorithm for the humanization of murine Fvs by variable domain resurfacing was developed and successfully applied to two murine monoclonal antibodies, N901 and anti-B4 (Pedersen *et al.*, 1994; Roguska *et al.*, 1994).

In its simplest form, the method of CDR-grafting for humanizing antibodies consists of grafting the CDRs from a rodent antibody onto the Fv frameworks of a human antibody. Constructs made in this way have the advantage of containing a minimal number of murine residues in the variable region. However the successful design of high-affinity CDR-grafted antibodies usually requires that key murine residues be substituted into the human acceptor framework to preserve the CDR conformations (reviewed in Winter and Harris, 1993). Although progress has been made towards identifying such framework residues either experimentally (Foote and Winter, 1992) or by computational methods (Queen *et al.*, 1989), this process is generally unique for each reshaped antibody and can therefore be difficult to predict (Kolbinger *et al.*, 1993). For many CDR-grafted antibodies, the percentages of murine residues in the human framework may not differ significantly from those found for corresponding resurfaced versions.

The resurfacing algorithm as previously described preserved murine residues at surface positions if they were located within 5 Å of a CDR in order to maintain CDR conformations. A mutational analysis of the surface positions within 5 Å of the CDRs of the resurfaced antibodies N901 and anti-B4 has now led to resurfaced antibodies containing fewer murine surface residues. These versions are compared with the corresponding CDR-grafted N901 and anti-B4 antibodies and with the published sequences of other CDR-grafted antibodies.

Materials and methods

Murine monoclonal antibodies and cell lines

Anti-B4 and N901 are murine monoclonal antibodies specific for the CD19 antigen found on human B cells (IgG1, κ; Nadler

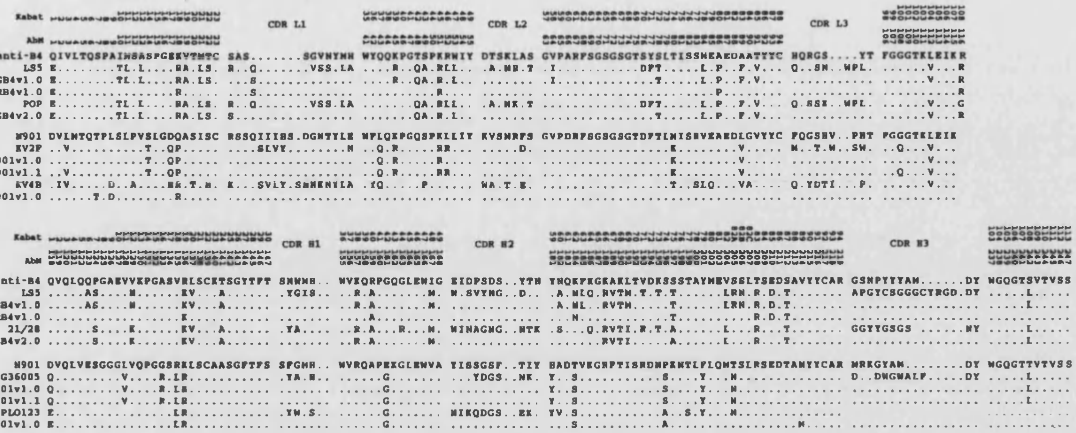

Kabat
AbM
CDR L1
CDR L2
CDR L3
anti-B4 QIVLTQSPAIMSASPGEKVTMTC SAS.......SGVNYMH WYQQKPGTSPKRWIY DTSKLAS GVPARFSGSGSGTSYSLTISSMEAEDAATYYC HQRGS....YT FGGGTKLEIKR
LS5 E........TL.L....RA.LS. R..Q.......VSS.LA ....R..QA.RLL.. .A.NR.T .I...........DFT.....L.P..F.V... Q..SN..WPL. ......V...R
GB4v1.0 E........TL.L....RA.LS. ...S............. ....R..QA.R.L.. .... .I.............T.....L.P..F.V... ....V...R
RB4v1.0 E...............R..... ...S............. ....R.... ....P...... ....R
POP E........TL.L....RA.LS. R..Q.......VSS.LA .......QA.RLL.. .A.NK.T ..............DFT.....L.P..F.V... Q.SSE..WPL. ......V...G
GB4v2.0 E........TL.L....RA.LS. S................ .......QA.R.L.. ....T.....L.P..F.V... ....V...R
N901 DVLMTQTPLSLPVSLGDQASISC RSSQIIIHS.DGNTYLE WFLQKPGQSPKLLIY KVSNRFS GVPDRFSGSGSGTDFTLKISRVEAEDLGVYYC FQGSHV..PHT FGGGTKLEIK
KV2F ..V..........T..QP..... ....SLVY........N ..Q.R.....RR... .....D. ................K........V..... M..T.W..SW. ..Q...V...
GN901v1.0 ..............T..QP..... ....Q.R.....R.... ....K........V..... ....V...
GN901v1.1 ..V..........T..QP..... ....Q.R.....RR... ....K........V..... ....Q...V...
KV4B .IV.....D..A....ER.T.N. K...SVLY.SNNKNYLA .YQ.....P...... WA.T.E. ................T..SLQ...VA.... Q.YDTI...P. ......V...
RN901v1.0 ......T.D........R..... ....
Kabat
AbM
CDR H1
CDR H2
CDR H3
anti-B4 QVQLQQPGAEVVKPGASVKLSCKTSGYTFT SNWMH.. WVKQRPGQGLEWIG EIDPSDS..YTN YNQKFKGKAKLTVDKSSSTAYMEVSSLTSEDSAVYYCAR GSNPYYYAM.....DY WGQGTSVTVSS
LS5 .....AS....M......KV...A...... .YGIS.. ..R.A.......M. W.SVYNG..D.. .A.MLQ.RVTM.T.T.T......LRN.R.D.T....... APGYCSGGGCYRGD.DY .....L.....
GB4v1.0 .....AS....M......KV...A...... ....R.A.......M. .... .A.ML..RVTM....T......LRN.R.D.T....... ....L.....
RB4v1.0 ....K........... ....A......... ....N...........T..........R.D.T....... ....
21/28 ......S....K......KV...A...... .YA.... ..R.A...R...M. WINAGNG..NTK .S...Q.RVTI.R.T.A.......L...R...T....... GGYYGSGS ......NY .....L.....
GB4v2.0 ......S....K......KV...A...... ....R.A.......M. ....RVTI.....A.......L...R...T....... ....L.....
N901 DVQLVESGGGLVQPGGSRKLSCAASGFTFS SFGMH.. WVRQAPEKGLEWVA YISSGSF..TIY HADTVKGRFTISRDNPKNTLFLQMTSLRSEDTAMYYCAR MRKGYAM........DY WGQGTTVTVSS
G36005 Q.........V....R.LR........... .YA.N.. ......G....... ...YDGS..NK. Y..S...........S....Y...N............. D..DNGWALF.....DY .....L.....
GN901v1.0 Q.........V....R.LR........... ....G....... .... Y..S...........S....Y...N............. ....L.....
GN901v1.1 Q.........V....R.LR........... ....G....... .... Y..S...........S....Y...N............. ....L.....
PLO123 E.................LR........... .YW.S.. ......G....... NIKQDGS..EK. YV.S...........A...S.Y...N........... ....
RN901v1.0 E.................LR........... ....G....... ....S...........A................M...... ....

Fig. 1. V_L (upper) and V_H (lower) amino acid sequence alignments of the murine (anti-B4, N901), human (LS5, POP, 21/28, KV4B, KV2F, PLO128, and G36005) CDR-grafted (GB4*v1.0*, GB4*v2.0*, GN901*v1.0*) and resurfaced (RB4*v1.0*, RN901*v1.0*) variable regions. A period indicates identity to the murine residue. Sequences are numbered according to the AbM program (Antibody Modeling program, Oxford Molecular Group, UK) and the AbM numbering is used in the text. The antibody numbering according to Kabat (Kabat *et al.*, 1991) is included for comparison.

et al., 1983), and the CD56 antigen found on human natural killer cells (IgG1, κ; Griffin *et al.*, 1983; Nitta *et al.*, 1989), respectively. The human cell lines, SW-2, developed by Smith *et al.*, (1989), and the human lymphoblastoid line, Namalwa, (A.T.C.C. CRL 1432) were used as the CD56-positive and CD19-positive cell lines, respectively.

Molecular modeling

Models of the murine and humanized Fvs were generated as described previously (Roguska *et al.*, 1994) using the methods of Martin *et al.* (1989, 1991) and recent modifications (Pedersen *et al.*, 1992; Rees *et al.*, 1995), which are encoded in the program AbM (Oxford Molecular Group, UK). The structures of the most identical sequences in the database of crystal structures were used to model the light and heavy chain variable regions. The V_H and V_L domains were paired by a least-squares fit onto the most structurally conserved strands of the Fv β-barrel, and then the framework side chains were introduced using a maximum overlap procedure. CDR backbone conformations were constructed using canonical loop structures where possible (Chothia and Lesk, 1987; Chothia *et al.*, 1989), otherwise they were built using the CAMAL algorithm (Martin *et al.*, 1989, 1991; Pedersen *et al.*, 1992), which combines a C^{α} search of the Brookhaven Protein Databank with an *ab initio* conformational search, followed by a series of screening steps. Framework residues in the murine and humanized anti-B4 and N901 models were then examined by a 5 Å proximity procedure (Roguska *et al.*, 1994). This method analyzes framework CDR interactions that have the potential to contribute to antigen binding. If a framework residue was located within 5 Å of a CDR and there was a change in size, charge, hydrophobicity, or potential to form hydrogen bonds that could disturb a CDR conformation, the human residue was considered a candidate for replacement by the amino acid from the murine Fv.

Construction of humanized V-region genes

The resurfaced and CDR-grafted V_H and V_L genes were constructed by extension and amplification of four overlapping oligonucleotides (120–130 nucleotides) comprising alternating strands of the full-length genes (Daugherty *et al.*, 1991). For N901, the assembled V_H and V_L genes encoded the following sequences in order (5′–3′): a *Hin*dIII cloning site, the consensus Kozak sequence (5′-GCCGCCACC-3′) (Kozak, 1989), an immunoglobulin signal sequence (Jones *et al.*, 1986), an intron, the humanized V_H or V_L coding region, a 3′ non-coding sequence including a splice site, and a *Bam*HI cloning site. For anti-B4, the assembled V_H and V_L genes differ from the above only by the elimination of an intron between the immunoglobulin signal sequence and the V_H or V_L coding region. The assembled humanized variable-region genes were subcloned into immunoglobulin expression vectors derived from HCMV-V_LLys-K_R (Maeda *et al.*, 1991), containing either the human constant κ exon or human-γ1 constant region gene. Expression vectors carrying the resurfaced and CDR-grafted anti-B4 variable-region genes differ from the N901 expression vectors by the addition of the human *DHFR* cDNA positioned at the 5′ end of the cytomegalovirus (CMV) enhancer in reverse orientation with respect to the CMV promoter. Mutant V_H and V_L variable-region genes were constructed by PCR mutagenesis (Ho *et al.*, 1989) or by site-directed mutagenesis according to Kunkel *et al.* (1987) of the CDR-grafted and resurfaced frameworks.

Expression and purification of recombinant antibodies

Three days prior to transfection, COS cells were plated on 15 cm tissue culture dishes at 2×10^6 cells/30 ml/plate in DMEM supplemented with 10% bovine calf serum (total of 12 plates per experiment) which were incubated at 37°C in 7% CO_2. After 3 days, each dish was washed twice with 10 ml HBS (137 mM NaCl, 5 mM KCl, 0.7 mM Na_2HPO_4, 6 mM dextrose, 20 mM HEPES, pH 7.05), and then treated with 5 ml trypsin solution (Gibco) for 5 min at 37°C, followed by vigorous pipetting to detach and disaggregate the cells. The cell suspension was then added to 5 ml SFM (Hybridoma Serum Free Medium, Gibco). Cells were pelleted by centrifugation for 5 min at 300 *g* and supernatants removed by aspiration. Cell pellets (each about 1×10^7 cells) were washed twice in 10 ml HBS, resuspended in 0.8 ml HBS containing 40 μg each of

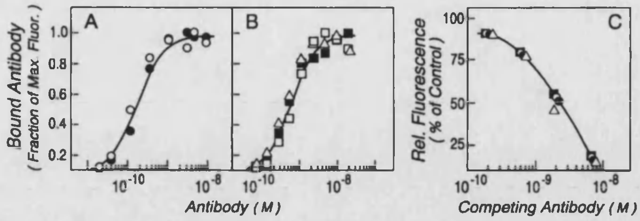


Fig. 2. Comparison of the binding of resurfaced, CDR-grafted, and murine N901 antibodies. Recombinant and murine antibodies were compared as follows: (**A**, **B**) Binding of antibodies to SW-2 cells as measured by indirect immunofluorescence. (**C**) Competition binding assay measuring the ability of resurfaced, CDR-grafted, and murine N901 to compete with fluorescein-labeled murine N901 for binding to SW-2 cells. The symbols used in the figure are as follows: (○) murine N901, (●) GN901*v1.0*, (△) GN901*v1.1*, (□) RN901*v1.0*, (■) RN901*v1.1*.

the V_H and V_L immunoglobulin expression plasmids, and then transferred to 0.4 cm electroporation cuvettes. Cuvettes containing the mixture of COS cells and expression plasmids were chilled on ice for 10 min, and then pulsed at 230 V and 960 µF in a Bio-Rad electroporation apparatus, followed by incubation on ice for 10 min. Following electroporation, cells from individual cuvettes were added to 40 ml of SFM, plated in a 15 cm tissue culture dish, and incubated at 37°C in 7% CO_2. The supernatant was harvested from the transfected cells 3 days after transfection. Pooled cell supernatants (approximately 450 ml total) were centrifuged for 20 min at 4500 *g*, filtered through a 0.22 µm filter and concentrated to a final volume of 30 ml using a stirred cell apparatus fitted with an Amicon YM10 membrane. Tris–HCl (pH 7.4) was added to a final concentration of 0.1 M and the resulting solution was applied to a column of immobilized protein A (Pierce Chemical Co., Rockford, IL.). The column was washed with 0.1 M Tris–HCl buffer, pH 7.4, containing 0.15 M NaCl. The bound antibody was eluted with 100 mM acetic acid containing 150 mM NaCl, and then dialyzed against 10 mM sodium phosphate, 150 mM NaCl, pH 7.4. Protein concentration was determined by measuring absorbance at 280 nm and assuming $E^{1\%}{}_{1\text{cm}}$ of 14.0 for IgG. Yields of antibodies were typically 1–2 mg from 0.5 l of cell culture supernatant. Antibodies produced in this manner are free from any contaminating bovine immunoglobulin as determined by isoelectric focusing gels (Roguska *et al*., 1994), thus allowing accurate concentration measurements. The purity and integrity of the purified products were confirmed by polyacrylamide gel electrophoresis under reducing and non-reducing conditions. To prevent any non-specific losses due to adsorption, bovine serum albumin at a final concentration of 1 mg/ml was added to the purified antibody for storage prior to binding studies.

Competition binding assay

Competition binding assays were performed as described previously (Lambert *et al*., 1991). Briefly, Namalwa cells (CD19 antigen-positive, 3×10^5 cells/well) or SW2 cells (CD56 antigen-positive, 5×10^5 cells /well) in AB buffer [2.5% pooled human AB serum (Whittaker) in minimal essential medium (Whittaker)] were plated at 25 µl per well in 96-well dishes. Twenty-five microliters of AB buffer containing fluorescein-labeled anti-B4 (4–6 nM) or N901 (1–3 nM) antibody mixed with various concentrations of competing murine or recombinant antibody were added to each well. Positive controls lacked competing antibody while negative controls lacked fluorescein-labeled anti-B4 or N901. Plates were incubated for 30 min at 4°C. The cells were washed once with AB buffer (175 µl/well) and fixed with 1% formaldehyde in 10 mM potassium phosphate, 150 mM NaCl, pH 7.2. The labeled cells were analyzed on a FACScan flow cytometer (Becton-Dickinson) set on linear fluorescence. The binding of fluorescently-labeled antibody was expressed as a percentage of the fluorescence of the positive control.

Indirect binding assay

Namalwa cells (for anti-B4) or SW2 cells (for N901) were mixed with carrier 4.5 µm polystyrene microspheres (Polysciences, Inc.) and washed twice in 10 ml NGS buffer (2.5% goat serum, 2 mM HEPES, pH 7.2, in minimal essential medium) at 4°C and resuspended in NGS buffer at 2×10^5 cells/ml. Cells (2×10^4/well) and microspheres (1.8×10^6/well) were plated at 100 µl/well in 96-well dishes. Various concentrations of primary murine or recombinant human IgG_1 antibodies were added to each well in 100 µl of NGS buffer. One hundred microlitres of NGS buffer lacking primary antibody were added to control wells. The plates were incubated at 4°C for 2 h, and then washed twice with 200 µl NGS buffer. Fluorescein-labeled secondary antibodies (goat anti-human or goat anti-mouse) were added in 100 µl NGS buffer per well and plates were incubated in the dark for 1 h at 4°C. The cells were again washed twice with NGS buffer (200 µl), fixed with 1% formaldehyde in 10 mM potassium phosphate, 150 mM NaCl, pH 7.2, and analyzed on a FACScan flow cytometer (Becton-Dickinson). The binding of antibody to cells is expressed as the fraction of maximal fluorescence.

Results

Humanization of murine monoclonal antibody N901 by CDR grafting

The human frameworks chosen to accept the N901 CDRs, KV2F V_L (Klobeck *et al*., 1985) and G36005 V_H (Schroeder *et al*., 1990), were selected based on their highest sequence identity to the N901 V_L and V_H framework sequences (Figure 1). The V_L amino acids of N901 and KV2F were the same at 66 of 80 (84%) framework positions, and the V_H amino acids of N901 and G36005 were identical at 80 of 94 (86%) framework positions. Modeling identified the two human Fv residues Val3 and Arg52 as falling within 5 Å of a CDR and having the potential to disturb CDR conformation. Val3 is adjacent to Arg24 and could alter the structure of CDR L1. Arg52 is located between Glu40 in CDR L1 and Met228 in CDR H3 and could influence the conformation of either CDR. Therefore, both of these residues were replaced with Leu, the corresponding residue at both positions in the murine N901 V_L. A third residue in the human V_L, Gln108, had the potential of interfering

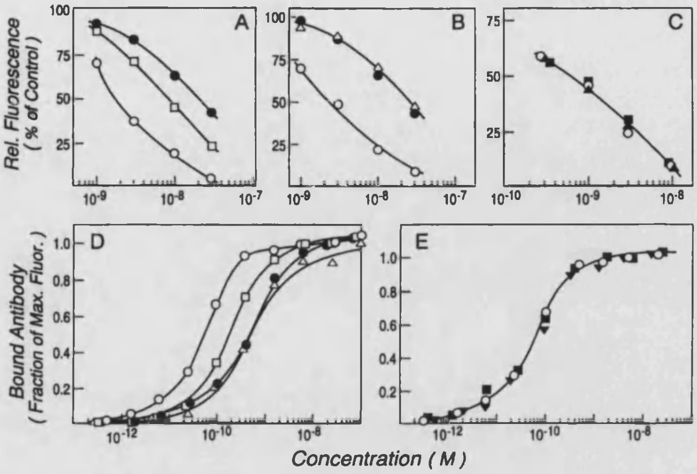


Fig. 3. Comparison of the binding of CDR-grafted anti-B4 antibodies. Antibodies were compared as follows: (**A**, **B**, **C**) Competition binding assay measuring the ability of CDR-grafted and murine anti-B4 to compete with fluorescein-labeled murine anti-B4 for binding to Namalwa cells. (**D**, **E**) Binding of CDR-grafted and murine anti-B4 antibodies to Namalwa cells as measured by indirect immunofluorescence. The symbols used in the Figure are as follows: (○) murine anti-B4, (●) GB4*v1.0*, (□) GB4*v1.1*, (△) GB4*v2.0*, (◆) GB4*v2.1*, (■) GB4*v2.3*.

with the docking of the KV2F V_L and G36005 V_H, and was replaced with the murine Gly. A CDR grafted N901 antibody incorporating these substitutions (GN901*v1.0*, Table 1A) was made and its binding to SW-2 cells was assessed by an indirect immunofluorescence assay and a competition binding assay. In both assays, the binding of the CDR-grafted GN901*v1.0* antibody was indistinguishable from that of the parent murine N901 antibody (Figure 2A,C).

In order to minimize the antigenic potential of the CDR grafted N901 antibody, we wished to test whether it was possible to reduce the overall occurrence of murine residues in the human frameworks while maintaining the binding affinity of the murine antibody. Although the assumption made in our analysis is that there may be interactions between residues that are predicted by modeling to be within 5 Å of one another, the influence of such interactions between CDR and framework residues on antigen binding is difficult to predict and must be experimentally determined (Glaser *et al*, 1992; Wilson and Stanfield, 1993). Therefore, antibody version GN901*v1.1* was constructed, which restored the original human residues Val, Arg, and Gln at the three framework positions 3, 52, and 108. GN901*v1.1* thus contained no murine residues in the human frameworks. When this antibody was analyzed for binding on SW-2 cells, it was also found to be indistinguishable from that of the parent murine N901 antibody (Figure 2B,C). Thus, N901 is a rare example of a murine antibody that was humanized by simply transferring its CDRs onto the most identical human variable-region frameworks while retaining the binding affinity of the original murine monoclonal antibody. This is likely a consequence of the high sequence identity between the murine and human framework sequences.

Humanization of murine monoclonal antibody anti-B4 by CDR grafting

Selection of most identical native human Fv. In designing the CDR-grafted anti-B4 antibody, two methods for selecting the human acceptor frameworks were considered. For a design based on using the most identical human V_H and V_L frameworks, the strategy that had been applied successfully for the CDR grafting of the N901 antibody, the most identical human frameworks found in the database (Figure 1) were the human V_H 21/28 (Dersimonian *et al.*, 1987) and human V_L POP (Spatz *et al.*, 1990). For an alternative design based on using the most identical clonally derived V_H–V_L pair, the strategy used successfully for the resurfacing of anti-B4, the most identical human Fv was LS5 (Silberstein *et al.*, 1989). The advantage of the latter strategy is that it ensures proper association of the V_H and V_L subunits without the need for evaluation by molecular modeling of the quality of V_H–V_L docking, and that it reduces the risk of generating potentially immunogenic neoepitopes that might be presented by the pairing of non-native heavy and light chains. While the advantage of the former strategy is that V_H and V_L sequences of higher homology to the murine sequences can be selected, the overall framework homologies of the two human sequences with respect to anti-B4 were quite similar. The LS5 V_H and 21/28 V_H were found to be 71% and 77% identical to the anti-B4 VH, and the LS5 V_L and POP V_L were found to be 70% and 73% identical to the anti-B4 V_L, respectively (Figure 1). Models of the murine and humanized anti-B4 Fvs were constructed and analysed by the 5 Å proximity procedure as described. Six LS5 human framework residues that had the potential to alter a CDR conformation were identified and replaced with the corresponding murine residue from the parent anti-B4 (Table IIA). This version, GB4*v1.0*, was made and analysed for binding to antigen-positive Namalwa cells (Figure 3A,D). In two different assays, binding affinity was reduced by about a factor of 10 relative to the parent murine anti-B4. In an attempt to improve the binding of GB4*v1.0*, nine subsequent antibodies were derived from GB4*v1.0* by substituting the appropriate murine residues at different LS5 framework positions (Table IIA, footnotes). Only one version, GB4*v1.1*, showed a significant improvement in binding (three-fold) relative to the GB4*v1.0* antibody (Figure 3A,D). GB4*v1.1* was

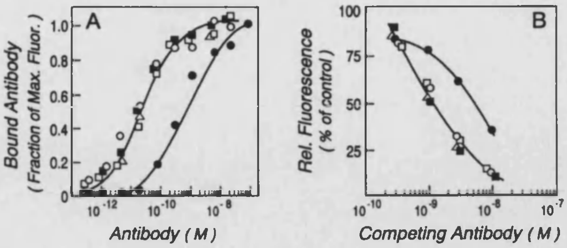


Fig. 4. Comparison of binding of resurfaced and murine antibodies. (**A**) Binding of resurfaced and murine anti-B4 antibodies to Namalwa cells as measured by indirect immunofluorescence. (**B**) Competition binding assay measuring the ability of resurfaced and murine anti-B4 to compete with fluorescein-labeled murine anti-B4 for binding to Namalwa cells. The symbols used in the figure are as follows: (○) murine anti-B4, (●) RB4*v1.1*, (□) RB4*v1.2*, (■) RB4*v1.3*, (△) RB4*v1.4*.

Table I. Humanized versions of N901[a]

A	CDR grafted N901: GN901				Relative binding affinity[b]
	V_L			V_H	CDR-grafted version / Murine N901
Residue number	3*	52	108		
KV2F V_L + G36005 V_H	*V*	*R*	*Q*		
GN901*v1.0*	L	L	G		1.0
GN901*v1.1*					1.0
B	Resurfaced N901: RN901				
	V_L			V_H	Resurfaced version / Murine N901
Residue number	3*				
KV4B V_L + PLO123 V_H	*V*				
RN901*v1.0*	L				1.0
RN901*v1.1*					1.0

[a]The human sequences are identified in italics. The various versions of GN901 or RN901 have human amino acids in all framework positions except where a murine residue is indicated by non-italicized type. A blank space for the various versions indicates identity to the human residue. An asterisk (*) identifies residues located at surface positions according to the criteria of Pedersen *et al.* (1994).

[b]Ratio of the concentration of humanized N901 required for 50% inhibition of binding of fluorescein-labeled murine N901 to that required by unlabeled murine N901.

constructed by introducing murine residues at V_H positions 182, 184, and 185. These residues are included in CDR H2 according to Kabat *et al.* (1991), although in our approach they are modeled as framework residues (Pedersen *et al.*, 1994; Roguska *et al.*, 1994). While little evidence exists for the direct interaction of residues in this region with antigens, the amino acids positions 182, 184, and 185 are highly conserved, which suggested that they might serve a CDR-scaffold function. However, even though the binding affinity of GB4*v1.1* was increased relative to GB4*v1.0*, it was still less than that of the murine anti-B4. Further attempts to improve the affinity of the CDR-grafted anti-B4 by further substitutions of murine residues in the LS5 frameworks were unsuccessful (Table IIA).

Selection of most identical human V_H and V_L frameworks. In this strategy, the most identical human V_H, 21/28, and most identical human V_L, POP, were selected as frameworks to accept the anti-B4 CDRs. An Fv model of the anti-B4 CDRs grafted onto the 21/28 and POP frameworks was generated and potential CDR-framework interactions were identified using the 5 Å proximity procedure. A CDR-grafted version GB4*v2.0* was produced that contained murine residues substituted at nine human framework positions (Table IIA). However, the binding of GB4*v2.0* was similar to that of GB4*v1.0*, ~10-fold lower than that of murine anti-B4 (Figure 3B,D).

In an effort to improve binding affinity, the Fv models were re-examined. The 5 Å proximity procedure was broadened to 6 Å in an effort to identify residues which, if mutated to the corresponding murine residue, might improve the binding affinity of GB4*v2.0*. By this criterion, an additional six murine residues were substituted into the human frameworks (Table IIA). Two other substitutions were introduced at positions 112 and 242 based on the possible effects of these residues on the variable-constant region packing which may influence CDR conformations. Finally, Pro124 found in murine anti-B4 is a very unusual residue at this position (except for murine subgroup IIB; Kabat *et al.*, 1991), and proline was not found at this position in any known human antibody, leading to speculation that the unusual placement of a proline here might contribute to the binding of anti-B4. A new version was made, GB4*v2.1* (Table IIA), that now incorporated a total of 18 murine framework residues, and when tested in the two binding assays it was found to be indistinguishable from the parent murine anti-B4 antibody (Figure 3C,E).

To further dissect the contributions of the nine additional substitutions to the binding affinity of GB4*v2.1* relative to GB4*v2.0*, GB4*v2.2* and GB4*v2.3* were generated by co-expressing the light chain of GB4*v2.0* and the heavy chain of GB4*v2.1*, and by co-expressing the light chain of GB4*v2.1* with the heavy chain of GB4*v2.0*, respectively (Table IIA). Binding tests revealed that the binding of GB4*v2.3*, like GB4*v2.1*, was equal to that of murine anti-B4, while the binding of the GB4*v2.2* version was about one-tenth of that for anti-B4 (Figure 3C,E). Thus, only the three additional changes in the V_L framework of GB4*v2.0* described above were necessary to produce the ten-fold improvement in binding affinity needed to restore the full affinity of murine anti-B4. The additional six substitutions in the 21/28 V_H framework were neither necessary nor deleterious to antigen binding. The rationalization of these results will be discussed later.

Contributions to binding affinity by residues located at surface positions. We wished to test whether the design of

Table II. Relative binding of different versions of humanized anti-B4[a]

A																								Relative binding affinity[b]
	CDR-grafted anti-B4: GB4																							CRD-grafted version / murine anti-B4
	V_L								V_H															
Residue Number	51*	52	53	76*	77	78	112	116	124	141	163	167	182	184	185	186*	189	190	191	193	195*	204	242	
LS5 $V_L + V_H$	*R*	*L*	*L*	*D*	*F*	*T*	*V*	*R*	*S*	*A*	*G*	*M*	*A*	*N*	*L*	*Q*	*V*	*T*	*M*	*T*	*T*	*L*	*L*	
GB4*v1.0*		R		S	Y											K				V	K			0.10
GB4*v1.1*		R		S	Y								N	K	F	K				V	K			0.33
[d]GB4*v1.2*		R		S	Y								N		F	K				V	K			0.15
[d]GB4*v1.3*		R		S	Y								N	K		K				V	K			n.d.[c]
[d]GB4*v1.4*		R		S	Y									K		K				V	K			n.d.
[e]GB4*v1.5*		R		S	Y					T						K				V	K			n.d.
[e]GB4*v1.6*		R		S	Y					T			N	K	F	K				V	K			0.22
[e]GB4*v1.7*		R		S	Y					T			N		F	K				V	K			n.d.
[f]GB4*v1.8*		R		S	Y											K	A			V	K	V		n.d.
[g]GB4*v1.9*	K	R	W	S	Y											K				V	K			0.20
Residue Number	51*	52	53	76*	77	78	112	116	124	141	163	167	182	184	185	186*	189	190	191	193	195*	204	242	
POP V_L + *21/28* V_H	*R*	*L*	*L*	*D*	*F*	*T*	*V*	*G*	*S*	*A*	*R*	*M*	*S*	*K*	*F*	*Q*	*V*	*T*	*I*	*R*	*T*	*L*	*L*	
GB4*v2.0*		R		S	Y			R			G		N			K				V	K			0.11
GB4*v2.1*		R	W	S	Y	S	L	R	P	T	G	I	N			K		K	L	V	K		S	1.0
GB4*v2.2*		R		S	Y			R	P	T	G	I	N			K		K	L	V	K		S	0.12
GB4*v2.3*		R	W	S	Y	S	L	R			G		N			K				V	K			1.0
GB4*v2.4*		R	W		Y	S	L	R			G		N							V	K			0.40
GB4*v2.5*		R	W		Y	S	L	R			G		N			K				V	K			0.43
GB4*v2.6*		R	W	S	Y	S	L	R			G		N							V	K			0.45

B				
	Resurfaced anti-B4: RB4			resurfaced version / murine anti-B4
	V_L	V_H		
Residue Number	76*	186*	195*	
LS5 $V_L + V_H$	*D*	*Q*	*T*	
RB4*v1.0*	S	K	K	1.0
RB4*v1.1*				0.2
RB4*v1.2*			K	1.0
RB4*v1.3*		K		1.0
RB4*v1.4*	S			1.0

[a]The human sequences are identified in italics. The various versions of GB4 or RB4 have human amino acids in all framework positions except where a murine residue is indicated by non-italicized type. A blank space for the various versions indicates identity to the human residue. An asterisk (*) identifies residues located at surface positions according to the criterion of Pedersen *et al.* (1994).

[b]Ratio of the concentration of humanized anti-B4 required for 50% inhibition of binding of fluorescein-labeled murine anti-B4 to that required by unlabeled murine anti-B4.

[c]n.d. Humanized anti-B4 showed no detectable binding in the competition binding assay (50% displacement of labeled anti-B4 at greater than 20-fold higher concentration than murine anti-B4).

[d]Permutations of GB4*v1.1*.as described in the text.

[e]Ala141 is located in the N-terminal region bordering CDR H1. It is not considered part of H1 based on the sequence variability definition (Kabat *et al.*, 1991), but structurally it forms part of the hypervariable loop and has been shown previously to influence the binding affinity of two antibodies (Woodle *et al.*, 1992; Presta *et al.*, 1993). The murine Thr141 was introduced into GB4*v1.0*, GB4*v1.1* and GB4*v1.2*, creating GB4*v1.5*, GB4*v1.6* and GB4*v1.7*

[f]Saul and Poljak (1993) found that amino acid sequence variations at V_H residues 126 and 189 defined structural framework patterns in the V_H region based on conformational differences in the main polypeptide chain and side-chains of residues 126, 135, 189, and 204 (AbM numbering) and suggested that these structural patterns may directly affect the conformation of some residues of CDR H2. Retention of the murine residue at position 189 was essential for maintenance of the murine binding activity in 4 variants of humanized MaE11 (Presta *et al.*, 1993). Introduction of the murine residue at this position also contributed to the successful CDR grafting of a murine anti-IgE antibody (Kolbinger *et al.*, 1993). GB4*v1.8* incorporated the murine Ala189 and Val204 into GB4*v1.0*.

[g]The murine Arg52 had been introduced into GB4*v1.0* because of potential interactions with CDR H3. The sequence of the murine anti-B4 at positions Lys51-Arg52-Trp53 is highly conserved in antibodies of the Kabat murine subgroup VI (Kabat *et al.*, 1991). Thus, two changes were made to GB4*v1.0* to generate GB4*v1.9* in order to test whether the introduction of all three murine residues would restore parental binding affinity to GB4*v1.0*.

the CDR-grafted anti-B4 version GB4*v2.3* could be improved by applying the principles of resurfacing, which are based on the premise that the immunogenicity of a foreign protein originates with the surface residues (Pedersen *et al.*, 1994). The GB4*v2.3* human frameworks contained a total of 12 murine residues, three of which were located at solvent-accessible positions, Ser76, Lys186, and Lys195 (Pedersen *et al.*, 1994; Roguska *et al.*, 1994). We constructed and tested antibodies GB4*v2.4*, GB4*v2.5* and GB4*v2.6* (see Table IIA), to examine the contributions of the human Asp76 and Gln186 either singly or in combination to the binding affinity GB4*v2.3*. However, the three new antibodies displayed a two- to three-fold decrease in binding relative to GB4*v2.3* and murine anti-B4, indicating that the murine residues at these surface positions contribute, albeit in a small way, to maintaining the proper conformations of the CDR-grafted anti-B4 CDRs. These results are in contrast to those obtained with the resurfaced anti-B4 (see below), where the residues

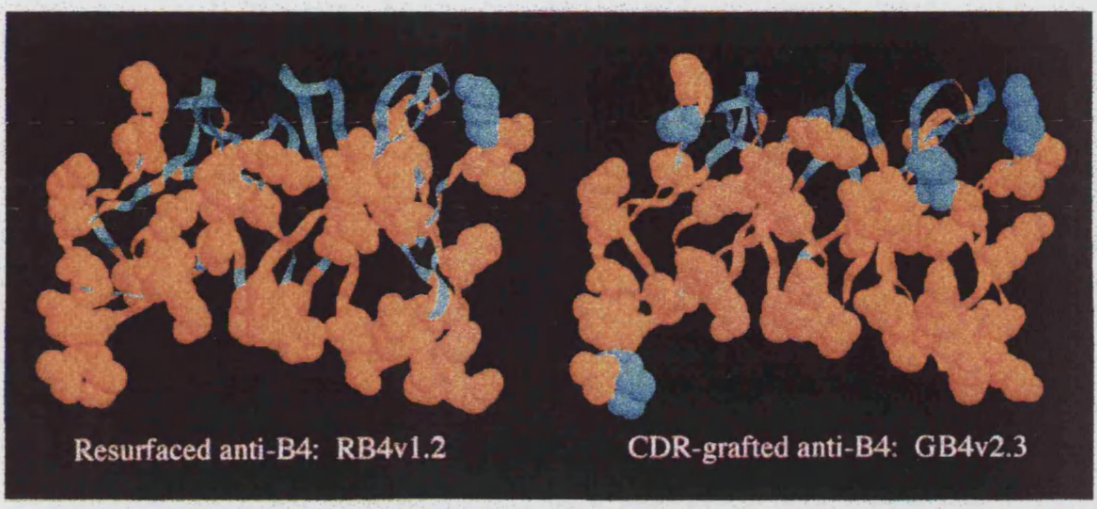


Fig. 5. Comparison of the optimal resurfaced and CDR-grafted versions of anti-B4. Surface residues are space-filled, core and CDR residues are shown as ribbons. For all positions (surface, core CDRs) murine residues are colored blue and human residues are colored yellow.

at these positions could be humanized with no loss of binding.

Refinement of resurfaced N901

We have previously described the modeling, construction, expression, purification, and binding characteristics of a resurfaced N901 antibody (Pedersen *et al.*, 1994; Roguska *et al.*, 1994), herein called RN901*v1.0* (Figure 1). Ideally, all of the surface-exposed amino acid residues in the Fv framework of a resurfaced antibody form a human surface pattern. However, some framework surface residues of the murine antibody may be necessary to preserve the proper conformation of the CDRs or may be in direct contact with the antigen. Previously, when the RN901*v1.0* Fv was modeled and compared with the model of murine Fv N901, the retention of only one murine surface amino acid was thought necessary, namely Leu at position 3 of V_L. A second version of resurfaced N901, called RN901*v1.1*, was constructed to test this prediction by replacing the murine Leu3 with the human Val3 (Table I). The apparent binding affinity of RN901*v1.1* for cell surface antigen was indistinguishable from that of the murine N901 in the two different binding assays (Figure 2B,C). Therefore the set of surface residues on the murine N901 Fv could be replaced entirely by the most identical set of human Fv surface residues without affecting the binding affinity of the antibody.

Refinement of resurfaced anti-B4

The modeling analysis of the resurfaced anti-B4 antibody described previously (Roguska *et al.*, 1994), herein called RB4*v1.0* (Figure 1), had identified three residues (Ser76, Lys186, and Lys195) at surface framework positions that could possibly interfere with CDR conformations according to the 5 Å proximity procedure. To analyze their contributions to antigen binding, they were systematically replaced with the human residues Asp76, Gln186, and Thr195 and the effect of these substitutions on binding affinity was determined (Table IIB, Figure 4). The introduction of all three human residues onto the RB4*v1.0* surface (version RB4*v1.1*) resulted in a lower binding affinity relative to RB4*v1.0* and murine anti-B4 (Figure 4A,B). Subsequently, three versions with single mutations (RB4*v1.2*, RB4*v1.3*, and RB4*v1.4*) were tested, (Table IIB). The binding affinity of each of these versions was equivalent to that of murine anti-B4 (Figure 4A,B). Thus, the full murine binding affinity could be maintained in the resurfaced anti-B4 by the retention of a single murine surface residue. Paradoxically, it appears that any two of the three positions can be substituted. Position 76 is located in a serine and threonine rich region of the antibody surface and is surrounded by five polar side chains (Ser24, Ser71, Ser73, Ser75, Thr78). The mutations Lys186Gln and Lys195Thr are both charge changes in or near the recognition surface of the antibody, and while these are not CDR residues, it is possible they may be directly involved in antigen binding.

To evaluate which of the three versions of resurfaced anti-B4 may be optimal for clinical development, the Kabat database (Kabat *et al.*, 1991) was examined to determine human residue preferences at positions 76, 186, and 195. At position 76 of V_L, no human κ V_L sequences had a Ser at position 76, while Asp was seen in 101 out of 135 human κ V_L sequences. At position 186 of V_H in Kabat human subgroup I (the LS5 V_H is a member of subgroup I), Gln was found in 44 of 49 sequences while Lys appeared only once. However, at position 195 there was no marked preference for either the human Thr or the murine Lys in the 49 human sequences, with Thr present in 21 sequences and Lys in 11 sequences. We concluded that the surface pattern of the resurfaced version RB4*v1.2* antibody was closest to the surface pattern of a human antibody.

Models of the optimal CDR-grafted and resurfaced antibodies were then compared (Figure 5). Both Fv sequences contain a similar percentage of total murine framework and CDR residues, 39% for RB4*v1.2* and 28% for GB4*v2.3*. However, RB4*v1.2* and GB4*v2.3* contain one and three murine residues at framework surface positions, respectively. While such comparisons cannot be used to predict the immunogenicity of antibodies, they do suggest that the resurfacing strategy can be used to design humanized antibodies that are no more likely to be immunogenic than a corresponding CDR-grafted version.

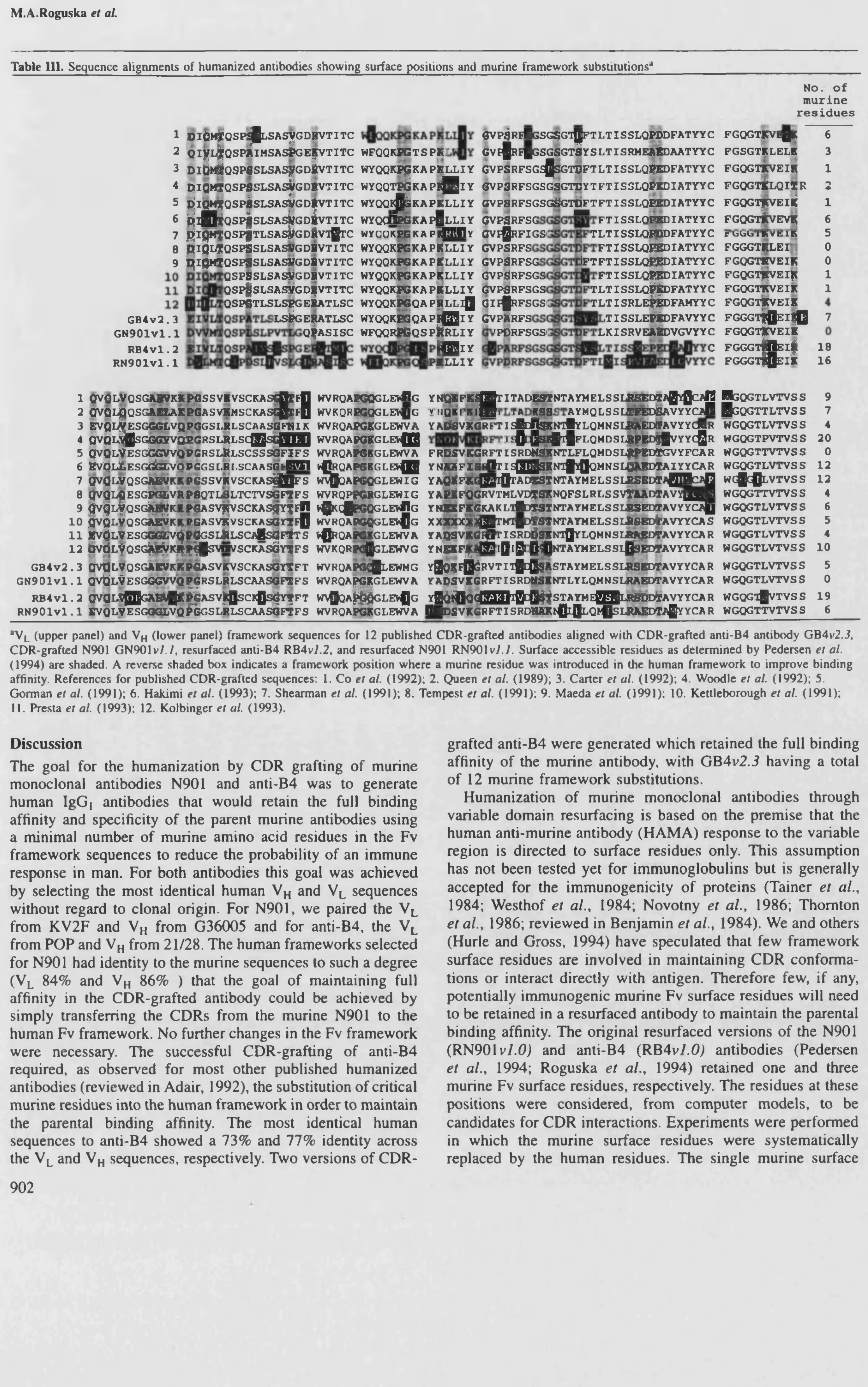


Table III. Sequence alignments of humanized antibodies showing surface positions and murine framework substitutions[a]

					No. of murine residues
1	DIQMTQSPSSLSASVGDRVTITC	WYQQKPGKAPKLLIY	GVPSRFSGSGSGTDFTLTISSLQPDDFATYYC	FGQGTKVEIK	6
2	QIVLTQSPAIMSASPGEKVTITC	WFQQKPGTSPKLWIY	GVPVRFSGSGSGTSYSLTISRMEAEDAATYYC	FGSGTKLELK	3
3	DIQMTQSPSSLSASVGDRVTITC	WYQQKPGKAPKLLIY	GVPSRFSGSRSGTDFTLTISSLQPEDFATYYC	FGQGTKVEIK	1
4	DIQMTQSPSSLSASVGDRVTITC	WYQQTPGKAPKLLIY	GVPSRFSGSGSGTDYTFTISSLQPEDIATYYC	FGQGTKLQITR	2
5	DIQMTQSPSSLSASVGDRVTITC	WYQQKPGKAPKLLIY	GVPSRFSGSGSGTDFTFTISSLQPEDIATYYC	FGQGTKVEIK	1
6	DIQLTQSPSSLSASVGDRVTITC	WYQQKPGKAPKLLIY	GVPSRFSGSGSGTSYTFTISSLQPEDIATYYC	FGQGTKVEVK	6
7	DIQMTQSPSTLSASVGDRVTITC	WYQQKPGKAPKRWIY	GVPARFIGSGSGTEFTLTISSLQPDDFATYYC	FGGGTKVEIK	5
8	DIQLTQSPSSLSASVGDRVTITC	WYQQKPGKAPKLLIY	GVPSRFSGSGSGTDFTFTISSLQPEDIATYYC	FGGGTKLEI	0
9	DIQMTQSPSSLSASVGDRVTITC	WYQQKPGKAPKLLIY	GVPSRFSGSGSGTDFTFTISSLQPEDIATYYC	FGQGTKVEIK	0
10	DIQMTQSPSSLSASVGDRVTITC	WYQQKPGKAPKLLIY	GVPSRFSGSGSGTDYTFTISSLQPEDIATYYC	FGQGTKVEIK	1
11	DIQLTQSPSSLSASVGDRVTITC	WYQQKPGKAPKLLIY	GVPSRFSGSGSGTDFTLTISSLQPEDFATYYC	FGQGTKVEIK	1
12	DIVLTQSPGTLSLSPGERATLSC	WYQQKPGQAPRLLIY	GIPDRFSGSGSGTDFTLTISRLEPEDFAMYYC	FGQGTKVEIK	4
GB4v2.3	EIVLTQSPATLSLSPGERATLSC	WYQQKPGQAPRRWIY	GVPARFSGSGSGTSYSLTISSLEPEDFAVYYC	FGGGTKLEIKR	7
GN901v1.1	DVVMTQSPLSLPVTLGQPASISC	WFQQRPGQSPRRLIY	GVPDRFSGSGSGTDFTLKISRVEAEDVGVYYC	FGQGTKVEIK	0
RB4v1.2	EIVLTQSPAIMSASPGERVTMTC	WYQQKPGTSPRRWIY	GVPARFSGSGSGTSYSLTISSMEPEDAATYYC	FGGGTKLEIK	18
RN901v1.1	DVLMTQTPDSLAVSLGERASISC	WFLQKPGQSPKLLIY	GVPDRFSGSGSGTDFTLKISRVEAEDLGVYYC	FGGGTKLEIK	16
1	QVQLVQSGAEVKKPGSSVKVSCKASGYTFT	WVRQAPGQGLEWIG	YNQKFKSKATITADESTNTAYMELSSLRSEDTAVYYCAR	WGQGTLVTVSS	9
2	QVQLQQSGAELAKPGASVKMSCKASGYTFT	WVKQRPGQGLEWIG	YNQKFKDKATLTADKSSSTAYMQLSSLTFEDSAVYYCAR	WGQGTTLTVSS	7
3	EVQLVESGGGLVQPGGSLRLSCAASGFNIK	WVRQAPGKGLEWVA	YADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSR	WGQGTLVTVSS	4
4	QVQLVQSGGGVVQPGRSLRLSCKASGYIFT	WVRQAPGKGLEWIG	YNQKVKDRFTISIDKSKNTAFLQMDSLRPEDTAVYYCAR	WGQGTPVTVSS	20
5	QVQLVESGGGVVQPGRSLRLSCSSSGFIFS	WVRQAPGKGLEWVA	FRDSVKGRFTISRDNSKNTLFLQMDSLRPEDTGVYFCAR	WGQGTTVTVSS	0
6	EVQLLESGGGLVQPGGSLRLSCAASGFSVT	WIRQAPGKGLEWIG	YNAAFISRLTISKDNSKNTYYLQMNSLQAEDTAIYYCAR	WGQGTLVTVSS	12
7	QVQLVQSGAEVKKPGSSVKVSCKASGYTFS	WVKQAPGQGLEWIG	YAQKFKGKATITADESTNTAYMELSSLRSEDTAVHYCAR	WGQGTLVTVSS	12
8	QVQLQESGPGLVRPSQTLSLTCTVSGFTFS	WVRQPPGRGLEWIG	YAPKFQGRVTMLVDTSKNQFSLRLSSVTAADTAVYYCAS	WGQGTTVTVSS	4
9	QVQLVQSGAEVKKPGASVKVSCKASGYTFI	WVKQAPGQGLEWIG	YNEKFKGKAKLTVDTSTNTAYMELSSLRSEDTAVYYCAT	WGQGTLVTVSS	6
10	QVQLVQSGAEVKKPGASVKVSCKASGYTFI	WVRQAPGQGLEWIG	XXXXXXXKATMTVDTSTNTAYMELSSLRSEDTAVYYCAS	WGQGTLVTVSS	5
11	EVQLVESGGGLVQPGGSLRLSCAASGFTTS	WIRQAPGKGLEWVA	YADSVKGRFTISRDDSKNTLYLQMNSLRAEDTAVYYCAR	WGQGTLVTVSS	4
12	QVQLVQSGAEVKKPGSSVKVSCKASGYTFS	WVKQRPGQGLEWVG	YNEKFKAKATITADESTNTAYMELSSLRSEDTAVYYCAR	WGQGTLVTVSS	10
GB4v2.3	QVQLVQSGAEVKKPGASVKVSCKASGYTFT	WVRQAPGQGLEWMG	YNQKFKGRVTITADKSASTAYMELSSLRSEDTAVYYCAR	WGQGTLVTVSS	5
GN901v1.1	QVQLVESGGGVVQPGRSLRLSCAASGFTFS	WVRQAPGKGLEWVA	YADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR	WGQGTLVTVSS	0
RB4v1.2	QVQLVQPGAEVVKPGASVKLSCKTSGYTFT	WVKQAPGQGLEWIG	YNQNFQGKAKLTVDKSTSTAYMEVSSLRSDDTAVYYCAR	WGQGTTVTVSS	19
RN901v1.1	EVQLVESGGGLVQPGGSLRLSCAASGFTFS	WVRQAPGKGLEWVA	YADSVKGRFTISRDNAKNTLYLQMNSLRAEDTAVYYCAR	WGQGTTVTVSS	6

[a] V_L (upper panel) and V_H (lower panel) framework sequences for 12 published CDR-grafted antibodies aligned with CDR-grafted anti-B4 antibody GB4*v2.3*, CDR-grafted N901 GN901*v1.1*, resurfaced anti-B4 RB4*v1.2*, and resurfaced N901 RN901*v1.1*. Surface accessible residues as determined by Pedersen *et al.* (1994) are shaded. A reverse shaded box indicates a framework position where a murine residue was introduced in the human framework to improve binding affinity. References for published CDR-grafted sequences: 1. Co *et al.* (1992); 2. Queen *et al.* (1989); 3. Carter *et al.* (1992); 4. Woodle *et al.* (1992); 5. Gorman *et al.* (1991); 6. Hakimi *et al.* (1993); 7. Shearman *et al.* (1991); 8. Tempest *et al.* (1991); 9. Maeda *et al.* (1991); 10. Kettleborough *et al.* (1991); 11. Presta *et al.* (1993); 12. Kolbinger *et al.* (1993).

Discussion

The goal for the humanization by CDR grafting of murine monoclonal antibodies N901 and anti-B4 was to generate human IgG_1 antibodies that would retain the full binding affinity and specificity of the parent murine antibodies using a minimal number of murine amino acid residues in the Fv framework sequences to reduce the probability of an immune response in man. For both antibodies this goal was achieved by selecting the most identical human V_H and V_L sequences without regard to clonal origin. For N901, we paired the V_L from KV2F and V_H from G36005 and for anti-B4, the V_L from POP and V_H from 21/28. The human frameworks selected for N901 had identity to the murine sequences to such a degree (V_L 84% and V_H 86%) that the goal of maintaining full affinity in the CDR-grafted antibody could be achieved by simply transferring the CDRs from the murine N901 to the human Fv framework. No further changes in the Fv framework were necessary. The successful CDR-grafting of anti-B4 required, as observed for most other published humanized antibodies (reviewed in Adair, 1992), the substitution of critical murine residues into the human framework in order to maintain the parental binding affinity. The most identical human sequences to anti-B4 showed a 73% and 77% identity across the V_L and V_H sequences, respectively. Two versions of CDR-grafted anti-B4 were generated which retained the full binding affinity of the murine antibody, with GB4*v2.3* having a total of 12 murine framework substitutions.

Humanization of murine monoclonal antibodies through variable domain resurfacing is based on the premise that the human anti-murine antibody (HAMA) response to the variable region is directed to surface residues only. This assumption has not been tested yet for immunoglobulins but is generally accepted for the immunogenicity of proteins (Tainer *et al.*, 1984; Westhof *et al.*, 1984; Novotny *et al.*, 1986; Thornton *et al.*, 1986; reviewed in Benjamin *et al.*, 1984). We and others (Hurle and Gross, 1994) have speculated that few framework surface residues are involved in maintaining CDR conformations or interact directly with antigen. Therefore few, if any, potentially immunogenic murine Fv surface residues will need to be retained in a resurfaced antibody to maintain the parental binding affinity. The original resurfaced versions of the N901 (RN901*v1.0*) and anti-B4 (RB4*v1.0*) antibodies (Pedersen *et al.*, 1994; Roguska *et al.*, 1994) retained one and three murine Fv surface residues, respectively. The residues at these positions were considered, from computer models, to be candidates for CDR interactions. Experiments were performed in which the murine surface residues were systematically replaced by the human residues. The single murine surface residue in N901 could be replaced without affecting the binding affinity, thereby creating a resurfaced N901 Fv with a completely human framework surface. For anti-B4, fully humanizing the surface resulted in a lower binding affinity relative to the murine antibody. However, full binding could be restored by the reintroduction of a single murine residue. Interestingly, with respect to affinity, it was irrelevant which of the three sequence positions was occupied by a murine residue. We propose that in such instances selection should be made according to the residue preference for the human antibodies listed in the Kabat database. Of the constructions made, RB4*v1.2* had a surface pattern most closely resembling that of the human antibody. These results lend support to the notion that few Fv surface residues are important for maintaining binding affinity.

In seeking further supportive evidence for this idea, we also analyzed the sequence positions of murine residues that were retained for the purpose of maintaining binding affinity in Fv sequences of published CDR-grafted antibodies. Table III lists twelve published examples of V_H and V_L framework sequences of CDR-grafted antibodies known to us, together with the resurfaced and CDR-grafted anti-B4 and N901 sequences. The surface residue positions and the murine amino acid residues are highlighted. While the published framework sequences of CDR-grafted antibodies contain between 0 and 22 murine residues (Table III), five Fv sequences contain no murine residues on the surface and four sequences have only one murine surface residue. This analysis suggests again that few surface residues contribute to binding and indicates that humanization by variable domain resurfacing has a high probability of maintaining the original affinity of the murine antibody.

Variable domain resurfacing maintains the core murine residues of the Fv sequences, and it could be argued that these foreign sequences, despite being internal, could elicit an immune response. As noted above, CDR-grafted antibodies also generally contain some murine framework amino acid residues and rather few are located at surface positions. Indeed, in the framework sequences of the CDR-grafted antibodies listed in Table III, between 0 and 16 murine residues are part of the Fv core. The two resurfaced antibodies shown here, RB4*v1.2* and RN901*v1.1*, have 36 and 22 murine core residues, respectively. While these numbers are about two- to three-fold larger than for CDR-grafted antibodies, this difference may not be significant for the human immune system (see discussion, Pedersen *et al.*, 1994). Furthermore, considering that eight of the 12 published CDR-grafted sequences listed in Table III did not maintain the full binding of their parent antibodies, the number of murine residues quoted for these CDR-grafted versions would be expected to increase with additional attempts at improving affinity, thereby further diminishing the differences between CDR-grafted and resurfaced antibodies.

CDR-grafted and resurfaced versions of N901 were generated (GN901*v1.1* and RN901*v1.1*) that had no murine residues at surface framework positions. Because the CDR-grafted version contains no murine core framework residues as well, a conservative approach would be to choose this antibody over the resurfaced version for a therapeutic application. For anti-B4, however, the CDR-grafted and resurfaced versions differed in the degree to which their surfaces could be humanized with retention of full binding affinity. The best CDR-grafted version, GB4*v2.3*, required three murine residues at surface positions to maintain binding, while the best resurfaced versions needed only one surface murine residue. Thus, even though the resurfaced version of anti-B4 RB4*v1.2* has 36 murine residues in the Fv core, it is conceivable that it would be less immunogenic than the CDR-grafted version with nine murine residues in the Fv core because it has a pattern of surface residues that is more identical to a human surface pattern. This argument ignores, of course, the possibiltiy of immunogenic T-cell epitopes, the importance of which in antibody immunogenicity is unknown. While this hypothesis can only be proven by clinical trials, this analysis would predict that similarly low immunogenicity might be expected for Fv resurfaced antibodies as has been observed for CDR-grafted antibodies (see unpublished proceedings of the *Fifth Annual IBC Conference on Antibody Engineering*, 7–9 December, 1994, San Diego, USA, and the unpublished proceedings of the *Tenth International Conference on Monoclonal Antibody Immunoconjugates for Cancer*, 9–11 March, 1995, San Diego, USA; Anasetti *et al.*, 1994; Caron *et al.*, 1994).

Acknowledgements

We thank B.Bourrett, K.Mattocks, and R.Steeves for expert technical assistance.

References

Adair,J.R. (1992) *Immunol. Rev.*, **130**, 5–40.

Anasetti,C. *et al.* (1994) *Blood*, **84**, 1320–1327.

Benjamin,D.C. *et al.* (1984) *Ann. Rev. Immunol.*, **2**, 67–101.

Caron,P.C. *et al.* (1994) *Blood*, **83**, 1760–1768.

Carter,P., Presta,L., Gorman,C.M., Ridgway,J.B.B., Henner,D., Wong,W-L.T., Rowland,A.M., Kotts,C., Carver,M.E., and Shepard,H.M. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 4285–4289.

Chothia and Lesk, (1987) *J. Mol. Biol.*, **196**, 901–917.

Chothia,C., *et al.* (1989) *Nature*, **342**, 877–883.

Co,M-S., Avdalovic,N.M., Caron,C.P., Avdalovic,M.V., Scheinberg,D.A. and Queen,C. (1992) *J. Immunol.*, **148**, 1149–1154.

Daugherty,B.L., DeMartino,J.A., Law,M.-F., Kawka,D.W., Singer,I.I. and Mark,G.E. (1991) *Nucleic Acids Res.*, **19**, 2471–2476.

Dersimonian,H., Schwartz,R.S., Barrett,K.J. and Stoller,B.D. (1987) *J. Immunol.*, **139**, 2496–2501.

Foote,J. and Winter,G. (1992) *J. Mol. Biol.*, **224**, 487–499.

Glaser,S.M., Vásquez,M., Payne,P.W. and Schneider,W.P. (1992) *J. Immunol.*, **149**, 2607–2614.

Gorman,S.D., Clark,M.R., Routledge,E.G., Cobbold,S.P. and Waldmann,H. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 4181–4185.

Griffin,J.D., Hercend,T., Beveridge,R. and Schlossman,S.F. (1983) *J. Immunol.*, **130**, 2947–2951.

Hakimi,J. *et al.* (1993) *J. Immunol.*, **151**, 1075–1085.

Ho,S.N., Hunt,H.D., Horton,R.M., Pullen,J.K. and Pease,L.R. (1989), *Gene*, **77**, 51–59.

Hurle,M.R. and Gross,M. (1994) *Curr. Opin. Biotechnol.*, **5**, 428–433.

Jones,P.T., Dear,P.H., Foote,J., Neuberger,M.S. and Winter,G. (1986) *Nature*, **321**, 522–525.

Kabat,E.A., Wu,T.T., Perry,H.M., Gottesman,K.S. and Foeller,C (1991) *Sequences of Proteins of Immunological Interest*, 4th edn., US Department of Health and Human Services, NIH, Washington, DC.

Kettleborough,C.A., Saldanha,J., Heath,V.J., Morrison,C.J. and Bendig,M.M. (1991) *Protein Engng*, **4**, 773–783.

Klobeck,H.G., Bornkamm,G.W., Combriato,G., Mocikat,R., Pohlenz,H.D., and Zachau,H.G. (1985) *Nucleic Acids Res.*, **13**, 6515–6529.

Kolbinger,F., Saldanha,J., Hardman,N. and Bendig,M. (1993) *Protein Engng*, **6**, 971–980.

Kozak,M. (1989) *J. Cell Biol.*, **108**, 229–241.

Kunkel,T.A., Roberts,J.D. and Zakour,R.A. ((1987) *Methods Enzymol.*, **154**, 367–382.

Lambert,J.M., Goldmacher,V.S., Collinson,A.R., Nadler,L.M. and Blättler,W.A. (1991) *Cancer Res.*, **51**, 6236–6242.

Martin,A.C.R., Cheetham,J.C. and Rees,A.R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9268–9272.

Maeda,H., Matsushita,S., Eda,Y., Kimachi,K., Tokiyoshi,S. and Bendig,M.M. (1991) *Hum. Antibod. Hybridomas*, **2**, 124–134.

Martin,A.C.R., Cheetham,J.C. and Rees,A.R. (1991) *Methods Enzymol.*, **203**, 121–153.

Nadler,L.M., Anderson,K.C., Marti,G., Bates,M.P., Park,E., Daley,J.F. and Schlossman,S.F. (1983) *J. Immunol.*, **131**, 244–250.

Nitta,T., Yagata,H., Sato,K. and Okumura,K. (1989) *J. Exp. Med.*, **170**, 1757–1761.

Novotny,J., Handschumacher,M., Haber,E., Bruccoleri,R.E., Carlson,W.B., Fanning,D.W., Smith,J.A. and Rose,G.D. (1986) *Proc. Natl Acad. Sci USA*, **83**, 226–230.

Padlan,E.A. (1991) *Mol. Immunol.*, **28**, 489–498.

Pedersen,J.T., Searle,S.M., Henry,H.A. and Rees,A.R. (1992) *Immunomethods*, **1**, 126–136.

Pedersen,J.T., Henry,H.A., Searle,S.J., Guild,B.C., Roguska,M.A. and Rees,A.R. (1994) *J. Mol. Biol.*, **235**, 959–973.

Presta,L.G., Lahr,S.J., Shields,R.L., Porter,J.P., Gorman,C.M. and Fendly,B.M. (1993) *J. Immunol.*, **151**, 2623–2632.

Queen,C., Schneider,W.P., Selick,H.E., Payne,P.W., Landolfi,N.F., Duncan,J.F., Avdalovic,N.M., Levitt,M., Junghans,R.P. and Waldman,T.A. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 10029–10033.

Rees,A.R., *et al.*, (1995) in *Antibody Engineering, 2nd edn.* Borrebaeck,C.A., ed. Oxford University Press, 3–57.

Roguska,M.A., Pedersen,J.T., Keddy,C.A., Henry,A.H., Searle,S.J., Lambert,J.M., Goldmacher,V.S., Blättler,W.A., Rees,A.R. and Guild,B.C. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 969–973.

Saul,F.A. and Poljak,R.J. (1993) *J. Mol. Biol.*, **230**, 15–20.

Schroeder, H.W.,Jr and Wang,J.Y. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 6146–6150.

Shearman,C.W., Pollock,D., White,G., Hehir,K., Moore,G.P., Kanzy,E.J. and Kurrle,R. (1991) *J. Immunol.*, **147**, 4366–4373.

Silberstein,L.E., Liwin,S. and Carmack,C.E. (1989) *J. Exp. Med.*, **169**, 1631–1643.

Smith,A., Waibel,R., Westera,G., Martin,A., Zimmerman,A.T. and Stahel,R.A. (1989) *Br. J. Cancer*, **59**, 174–178.

Spatz,L.A., Wong,K.K., Williams,M., Desai,R., Golier,J., Berman,J.E., Alt,F.W. and Latov,N. (1990) *J. Immunol.*, **144**, 2821–2828.

Tainer,J.A., Getzoff,E.D., Alexander,H., Houghten,R.A., Olsen,A.J. and Lerner,R.A. (1984) *Nature*, **312**, 127–133.

Tempest,P.R., Bremmer,P., Lambert,M., Taylor,G., Furze,J.M., Carr,F.J. and Harris,W.J. (1991) *Bio/Technology*, **9**, 266–271.

Thornton,J.M., Edwards,M.S., Taylor,W.R. and Barlow,D.J. (1986) *EMBO* **5**, 409–413.

Westhof,E., Altshuh,D., Moras,D., Bloomer,D., Modragon,A.C., Klug,A. and van Regenmortel,M.H.V. (1984) *Nature*, **311**, 123–126.

Wilson,I.A. and Stanfield,R.L., (1993) *Curr. Opin. Struct. Biol.*, **3**, 113–118.

Winter,G. and Harris,W.J. (1993) *Trends Pharm. Sci.*, **14**, 139–143.

Woodle,E.S., Thistlewaite,J.R., Jolliffe,L.K., Zivin,R.A., Collins,A., Adair,J.R., Bodmer,M., Athwal,D., Alegre,M-L. and Bluestone,J.A. (1992) *J. Immunol.*, **148**, 2756–2763.

Received February 14, 1996; revised July 17, 1996; accepted July 25, 1996

5.9 Conclusions

The two reshaping methods described represent two different strategies to solve the same problem. Essentially a resurfaced antibody is murine with human surface residues while a CDR grafted antibody is human with a murine combining site.

The methods have both been successful in generating reshaped antibodies which bind antigen with the same affinity as the original antibody. The resurfacing method is less prone to affinity loss since the initial murine framework-CDR combination is retained. However the immunogenicity of the antibodies is not yet known.

These data may also indicate why some murine antibody subgroups are more immunogenic than others. Within some subgroups few mutations (2-3 residues) are required to generate a human surface and within others there is a more distant relation between the V_L/V_H surfaces (8-10 residues).

5.10 The TCR Epitope Problem

One problem which antibody resurfacing does not address is the removal of TCR epitopes, that is the processed peptides presented to T cells by MHC, from the antibody sequences. There are a very large number of possible peptide sequences and mouse antibody sequences are likely to have unique epitopes which will mark them as foreign proteins.

5.10.1 An Approach to TCR Epitope Removal

The body is tolerant to all the antibodies it produces. Many of these antibodies are similar in sequence to mouse antibodies. However any differences will result in foreign peptides which could act as antigenic peptides, which when associated with MHC would activate the T Cell response. To limit this response modifications could be made to the mouse antibody to make it contain only linear combinations of residues which also occur in human sequences, thus removing the foreign peptides. It is still necessary to retain binding and reduce B cell immunogenicity as well.

For a resurfacing approach the two aims would be to maintain the human surface residues, and to substitute the fewest possible residues to remove peptide epitopes, so that the mouse core structure and CDR loop topology were maintained.

Each chain in an antibody F_V domain is over 100 residues in length. The length of an antigenic peptide is 10-15 residues for an MHC class II restricted T cell. The number of possible combinations of residues is therefore very large, and searching through all the possible combinations would not be practical. However a Monte Carlo/Metropolis method would be able to sample some of the sequence phase space. In such an approach individual random changes would be made to the sequence. After each change, a score would be generated, which would include terms for similarity to the mouse sequence and number of peptides which matched human fragments. If the score was improved after a particular change it would be retained. If the score was worse, the change could still be retained. The probability of retaining a worse score would depend on a temperature factor. A fixed number of steps would be performed at each temperature, after which it would be reduced. The whole process would then be repeated at a series of decreasing temperatures. In theory, the high temperature steps would allow the exploration of a large amount of sequence space, and the gradual reduction in temperature would trap the sequence at a maximum score.

A program has been written which attempts to remove T cell epitopes using the above ideas. The residues at the surface positions and in the CDRs are fixed. A database of human fragments of a specified length at each position is created. The number of steps at each temperature, the staring temperature, the final temperature and the temperature step size are input. The sequence is scored using the following equation:

$$E = NMatchFrag + (IdentScore \times LenFrag) \tag{5.1}$$

The equation used to calculate the probability of retaining a worse score is:

$$P(E) = exp(dE/Temp) \tag{5.2}$$

The residue patterns required for binding to some MHC class I molecules are known [23]. However the MHC class II molecules do not have the same restrictive binding pockets as the MHC class I molecules. Molecular dynamics has been used to attempt to predict potential T cell epitopes [196]. If it is reliable this information could be used to reduce the number of peptide sequences which need to be examined. Work on this project is ongoing.

Chapter 6

General Discussion

6.1 Reasons for Modelling TCRs

The sequences of proteins are more easily and more rapidly determinable (using the techniques of molecular biology) than are their structures (using X-ray crystallography). In the case of TCRs the first sequence was elucidated in 1984. Since that time many hundreds of different TCR sequences have been determined. However it is only in the last two years that the structure of a TCR was first solved. Many groups around the world had been trying to accomplish this, but the TCR proved difficult to solubilise and crystallise. The initial structure determined was only of a β chain [3], which had to be mutated to remove oligosacharide binding sites. Later an α chain dimer and very recently a TCR/MHC complex have been solved.

Although modelling is never completely accurate, the immunoglobulin superfamily is probably one of the most amenable to it. The large number of sequences and the increasing number of structures of antibodies and other immunoglobulins such as CD4 and CD8, provides a useful database of information to start modelling other members of the family with reasonable accuracy. Techniques already exist, and modelling has been used successfully to predict mutations in humanisation, and in other antibody modification experiments.

6.2 TCR Modelling

In Chapter 2, both TCR α and β chains were shown to be more similar to κ light chains than to antibody heavy chains. In Chapter 4 the similarities found between the predicted models and the X-ray structures showed that the use of a light chain dimer framework was reasonable. The models also correlated well with the two structures in the existence of a cavity between the two CDR3 regions of the TCR. This was shown to be involved in peptide binding in the structure of the TCR MHC peptide complex. However CDR $\alpha 2$ was poorly modelled. The conformation of the region is unique to this TCR chain type. It is therefore not surprising that it was inaccurately modelled. However it was encouraging that the sequence analyses through the sequence similarity and environment scores, as well as the CD4 modelling study, indicated that the region would be difficult to model.

The orientation of the TCR in the model complex was not correct. With hindsight the diagonal orientation shown in both structures so far determined does produce a much closer interaction between the TCR and the MHC. No previous modelling study on the TCR correctly predicted the orientation.

6.3 Future of TCR Modelling

In the future, modelling studies on TCRs will obviously be able to use a TCR framework. Also it is possible that canonical loop conformations will exist for at least the CDR1 and CDR2 loops of both TCR chains. It has been suggested that the larger overall variability between TCR sequences will mean that the canonical classes will probably be restricted to particular sub families. Canonical classes exist for CDR L3 in antibodies. Given the greater light chain character of both TCR chain types, it may be that there are canonical classes for both CDR $\alpha3$ and CDR $\beta3$. However more structures will need to be determined before this can be known. The greater possibility for variability in this region in TCRs compared to antibodies may argue against the existence of canonicals for either CDR3 region.

6.4 Antibody Modelling

The TCR modelling algorithm is based on antibody modelling and, during the development of the algorithm, many antibody models were created. It became apparent that the CAMAL algorithm did not always produce accurate results. Modifications to the algorithm, described in Chapter 3, increased the accuracy of modelling. One change which improved accuracy of CDR L1-L3 and H1-H2 was the use of canonical loops where possible. This produced more consistent results than the CAMAL algorithm. CDR H3 is the most variable in structure and gave the most variability in accuracy of modelling. Two modifications; firstly inclusion of the backbone (including C_β carbon) for the previously modelled loops and, secondly, use of broad structure classes for CDR H3 to restrict the choice of take-off angle, have improved accuracy for H3. The exclusion of other CDRs while modelling each loop was based on evidence that the inclusion of all atoms of previously modelled loops could produce a cumulative error. The original antibody used to develop CAMAL, Gloop 2, has a very short CDR3 loop. It may be that inclusion of CDRs is not necessary in the case of short loops, but, for modelling longer CDR loops, their presence is required to restrict the conformational space available to the loop. The broad classes defined for CDR H3 are only designed to restrict the take-off angle. The conformations of the loops within a class are not necessarily similar. It is still not possible to model CDR H3 loops of more than 12 residues with any accuracy.

6.5 Humanisation

The resurfacing method for humanising antibodies requires fewer changes to the sequence than the CDR grafting. In the two antibodies for which this method has been used, N901 and B4, full binding affinity was retained, with the backmutation of only 0 and 1 residues respectively required. It seems likely that the chances of success will be greater using this method rather than CDR grafting. However it still remains to be determined whether resurfaced antibodies will be non immunogenic in humans. These studies are being undertaken at present by ImmunoGen Inc. (USA).

6.6 Evolutionary Relationships

Immunoglobulins are present in all extant (still surviving) jawed vertebrates but are absent from invertebrates. This indicates that they evolved more than 500 million years ago. The two largest protein groups in the Ig superfamily are the TCRs and the antibodies. The evolutionary relationships among these rearranging receptors is still uncertain. It has been suggested that the rearranging receptors may evolve at a faster rate than other proteins because they have to adapt to match the contra-evolving (evolving to try to evade the immune response) bacterial and viral antigens. This may make it more difficult to determine relationships between them.

The previous chapters contain a presentation of sequence and structural analysis of the TCR and possible structural relationship of the TCR to antibodies. One puzzling feature of this work was the discovery that both α and β TCR chains were more similar to the antibody light chain. It would be interesting as future work to try to incorporate this into the scheme of immunoglobulin evolution.

Bibliography

[1] Garcia, K.C., Degano, M., Stanfield, R.L., Brunmark, A., Jackson, M.R., Petersen, P.A., Teyton, L., and Wilson, I.A. An α β T cell receptor structure at 2.5Å and its orientation in the TCR-MHC complex. *Science*, 274:209–219, 1996.

[2] Garboczi, D.N., Ghosh, P., Utz, U., Qing, R.F., Biddison, W.E., and Wiley, D.C. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature (London)*, 384:134–141, 1996.

[3] Bentley, G.A., Boulot, G., Karjalainen, K., and Mariuzza, R.A. Crystal structure of the β chain of a T cell antigen receptor. *Science*, 267:1984–1987, 1995.

[4] Fields, B.A., Ober, B., Malchiodi, E.L., Lebedeva, M.I., Braden, B.C., Ysern, X., Kim, J., Shao, X., Ward, E.S., and Mariuzza, R.A. Crystal structure of the Vα domain of a T cell antigen receptor. *Science*, 270:1821–1824, 1995.

[5] Martin, A.C.R., Cheetham, J.C., and Rees, A.R. Molecular modeling of antibody combining sites. *Meth. Enz.*, 203:121–153, 1991.

[6] Martin, A.C.R., Cheetham, J.C., and Rees, A.R. Modelling antibody hypervariable loops: A "combined algorithm". *Proc. Natl. Acad. Sci. USA*, 86:9268–9272, 1989.

[7] Pedersen, J., Searle, S., Henry, A., and Rees, A.R. Antibody modelling: Beyond homology. *Immunomethods*, 1:126–136, 1992.

[8] Short, N. Immune cells and their interactions. *Nature (London)*, 372:217, 1994.

[9] Gumperz, J.E. and Parham, P. The enigma of the natural killer cell. *Nature (London)*, 378:245–248, 1995.

[10] Abbas, A.K., Lichtman, A.H., and Pober, J.S. *Cellular and Molecular Immunology*. W.B.Saunders Company, Philadelphia, second edition, 1991.

[11] Janeway Jr., C.A. How the immune system recognizes invaders. *Sci. Am.*, 269:73–79, 1993.

[12] Theofilopoulos, A.N. The basis of autoimmunity: Part II genetic predisposition. *Immun. Tod.*, 16:150–159, 1995.

[13] Theofilopoulos, A.N. The basis of autoimmunity: Part I. Mechanisms of aberrant self-recognition. *Immun. Tod.*, 16:90–98, 1995.

[14] Sutton, B.J. and Gould, H.J. The human IgE network. *Nature (London)*, 366:421–428, 1993.

[15] Sanz, I. Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *J. Immunol.*, 147:1720–1729, 1991.

[16] Jorgensen, J.L., Reay, P.A., Ehrlich, E.W., and Davis, M.M. Molecular components of T-cell recognition. *Annu. Rev. Immunol.*, 10:835–873, 1992.

[17] Rock, E.P. and Davis, M.M. Structural aspects and chemistry of T cell receptor recognition of antigen-MHC complexes. *Acc. Chem. Res.*, 26:435–441, 1993.

[18] Pleiman, C.M., D'Ambrosio, D., and Cambier, J.C. The B-cell antigen receptor complex: structure and signal transduction. *Immunology Today*, 15:393–399, 1994.

[19] Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennet, W.S., Strominger, J.L., and Wiley, D.C. Structure of the human class I histocompatibility antigen HLA-A2. *Nature (London)*, 329:506, 1987.

[20] Garrett, T.P.J., Saper, M.A., Bjorkman, P.J., Strominger, J.L., and Wiley, D.C. Specificity pockets for the side chains of peptide antigens in HLA-A$W68. *Nature (London)*, 342:692–700, 1989.

[21] Young, A.C.M., Nathenson, S.G., and Sacchettini, J.C. Structural studies of class I major histocompatibility complex proteins: Insights into antigen presentation. *Faseb J.*, 9:26–36, 1995.

[22] Brown, J.H., Jardetzsky, T.S., Gorga, J.C., Stern, L.J., Urban, R.G., Strominger, J.L., and Wiley, D.C. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature (London)*, 364:33–39, 1993.

[23] Madden, D.R. The three-dimensional structure of peptide-MHC complexes. *Annu. Rev. Immunol.*, 13:587–622, 1995.

[24] Schafer, P.H. and Pierce, S.K. Evidence for dimers of MHC class II molecules in B lymphocytes and their role in low affinity T cell responses. *Immunity*, 1:699–707, 1994.

[25] Fremont, D.H., Matsumura, M., Stura, E.A., Petersen, P.A., and Wilson, I.A. Crystal structures of two viral peptides in complex with murine MHC class I H-2K^b. *Science*, 257:919–927, 1992.

[26] Kraulis, Per J. Molscript: a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography*, 24:946–950, 1991.

[27] Germain, R.N. Antigen processing and CD4+ T-cell depletion in AIDS. *Cell*, 54:441–444, 1988.

[28] Leahy, D.J. A structural view of CD4 and CD8. *Faseb J.*, 9:17–25, 1995.

[29] Saito, H., Kranz, D.M., Takagaki, Y., Hayday, A.C., Eisen, H.N., and Tonegawa, S. A third rearrangement and expressed gene in a clone of cytotoxic T lymphocytes. *Nature (London)*, 312:36–40, 1984.

[30] Brenner, M.B., Mclean, J., Dialynas, D., Strominger, J.L., Smith, J.A., and Owen, F.L. Identification of a putative second T cell receptor. *Nature (London)*, 322:145–49, 1986.

[31] Haas, W., Pereira, P., and Tonegawa, S. γ/δ cells. *Annu. Rev. Immunol.*, 11:637–685, 1993.

[32] Weintraub, B.C., Jackson, M.R., and Hedrick, S.M. $\gamma\delta$ T cells can recognize nonclassical MHC in the absence of conventional antigenic peptides. *J. Immunol.*, 153:3051–3058, 1994.

[33] O'Brien, R.L., Happ, M.P., Dallas, A., Palmer, E., Kubo, R., and Born, W.K. Stimulation of a major set of lymphocyte expressing T cell receptor $\gamma\delta$ by antigen derived from *mycobacterium tuberculosis*. *Cell*, 57:667–674, 1989.

[34] Janeway, C.A., Jones, B., and Hayday, A.C. Specificity and function of cells bearing $\gamma\delta$ T cell receptors. *Immunology Today*, 9:73–76, 1988.

[35] Fraser, J.D., Straus, D., and Weiss, A. Signal transduction events leading to T-cell lymphokine gene expression. *Immun. Tod.*, 14:357–362, 1993.

[36] Malissen, B. and Schmitt-Verhulst, A.M. Transmembrane signalling through the T-cell-receptor-CD3 complex. *Curr. Op. Immun.*, 5:324–333, 1993.

[37] Male, D., Champion, B., Cooke, A., and Owen, M. *Advanced Immunology*. Gower Medical Publishing, second edition, 1991.

[38] Denzin, L.K. and Cresswell, P. HLA-DM induces clip dissociation from MHC class II $\alpha\beta$ dimers and facilitates peptide loading. *Cell*, 82:155–165, 1995.

[39] Van Oss, C.J. and Van Regenmortel, M.H.V. *Immunochemistry.* Marcel Dekker Inc., first edition, 1994.

[40] Williams, A.F. and Barclay, A.N. The immunoglobulin superfamily domain for cell surface recognition. *Annu. Rev. Immunol.*, 6:381–405, 1988.

[41] Murphy, P.M. Molecular mimicry and the generation of host defense protein diversity. *Cell*, 72:823–826, 1993.

[42] Poljak, R.J., Amzel, L.M., Avery, H., Chen, B.L., Phizackerley, R.P., and Saul, F. Three dimensional structure of the F_{ab}-fragment of a human immunoglobulin at 2.8 Å resolution. *Proc. Natl. Acad. Sci. USA*, 70:3305–3310, 1973.

[43] Wang, J., Yan, Y., Garret, T.P.J., Liu, J., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E.L., and Harrison, S.C. Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature (London)*, 348:411–418, 1990.

[44] Ryu, S., Kwong, P.D., Truneh, A., Porter, T.G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N., Axel, R., Sweet, R.W., and Hendrickson, W.A. Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature (London)*, 348:419–425, 1990.

[45] Leahy, D.J., Axel, R., and Hendrickson, W.A. Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2.6 Å resolution. *Cell*, 68:1145–1162, 1992.

[46] Lesk, A.M. and Chothia, C. Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. *J. Mol. Biol.*, 160:325–342, 1982.

[47] Sheriff, S., Silverton, E.W., Padlan, E.A., Cohen, G.H., Smith-Gill, S.J., Finzel, B.C., and Davies, D.R. Three-dimensional structure of an antibody-antigen complex. *Proc. Natl. Acad. Sci. USA*, 84:8075–8079, 1987.

[48] Padlan, E.A., Silverton, E.W., Sheriff, S., Cohen, G.H., Smith-Gill, S.J., and Davies, D.R. Structure of antibody-antigen complex : crystal structure of the HyHEL-10 F_{ab}-lysozyme complex. *Proc. Natl. Acad. Sci. USA*, 86:5938–5942, 1989.

[49] Mainhart, C.R., Potter, M., and Feldmann, R.J. A refined model for the variable domains (Fv) of the J539 β (1,6)-d-galactan-binding immunoglobulin. *Mol. Immunol.*, 21:469–478, 1984.

[50] Saul, F.A., Amzel, L.M., and Poljak, R.J. The preliminary refinement and structural analysis of the Fab fragment from human immunoglobulin New at 2.0Å resolution. *J. Biol. Chem.*, 253:585–597, 1978.

[51] Herron, J.N., He, X., Mason, M.L., Voss Jnr., E.W., and Edmundson, A.B. Three dimensional structure of a fluorescein-F_{ab} complex crystallized in

2-methyl-2,4-pentanediol. *Proteins: Struct., Funct., Genet.*, 5:271–276, 1989.

[52] Rose, D.R., Strong, R.K., Margolis, M.N., Gefter, M.L., and Petsko, G.A. Crystal structure of the antigen-binding fragment of the murine anti-arsonate monoclonal antibody 36-71 at 2.9Å resolution. *Proc. Natl. Acad. Sci. USA*, 87:338–342, 1990.

[53] Rudikoff, S., Satow, Y., Padlan, E.A., Davies, D.R., and Potter, M. κ chain structure from a crystallized murine Fab′: The role of the joining segment in hapten binding. *Mol. Immunol.*, 18:705–711, 1981.

[54] Ely, K.R., Herron, J.N., Harker, M., and Edmunson, A.B. Three-dimensional structure of a light chain dimer crystallised in water. Conformational flexibility of a molecule in two crystal forms. *J. Mol. Biol.*, 210:601–615, 1989.

[55] Ely, K.R., Herron, J.N., and Edmundson, A.B. Three-dimensional structure of a hybrid light chain dimer. Protein engineering of a binding cavity. *Mol. Immunol.*, 27:101–114, 1990.

[56] Furey, W.J., Wang, B.C., Yoo, C.S., and Sax, M. Structure of a novel Bence-Jones protein (RHE) fragment at 1.9Å resolution. *J. Mol. Biol.*, 167:661, 1983.

[57] Palm, W. and Hilschmann, N. Die Primärstruktur einer kristallinen monoklonalen immunoglobulin-L-Kette vom κ-Typ, Subgruppe I (Bence-Jones-

Protein Rei), Isolierung und Charakterisierung der tryptischen Peptide; die vollständige Aminosäuresequenz des Proteins. *Hoppe-Seyler's Z. Physiol. Chem.*, 356:167–191, 1975.

[58] Marquart, M., Deisenhofer, J., and Huber, R. Crystallographic refinement and atomic models of the intact immunoglobulin molecule KOL and its antigen-binding fragment at 3.0Å and 1.9Å resolution. *J. Mol. Biol.*, 141:369–391, 1980.

[59] Lascombe, M.B., Alzari, P.M., Boulot, G., Salujian, P., Tougard, P., Berek, C., Haba, S., Rosen, E.M., Nisonof, A., and Poljak, R.J. Three-dimensional structure of Fab R19.9, a monoclonal murine antibody specific for the p-azo-benzene-arsonate group. *Proc. Natl. Acad. Sci. USA*, 86:607, 1989.

[60] Amit, A.G., Mariuzza, R.A., Phillips, S.E.V., and Poljak, R.J. The three-dimensional structure of an antigen-antibody complex at 2.8Å resolution. *Science*, 233:747–753, 1986.

[61] Saul, F.A. and Poljak, R.J. Crystal structure of the F_{ab} fragment from the human myeloma immunoglobulin IGG HIL at 1.8Å resolution. *To be published*, 1993.

[62] Brünger, A.T., Leahy, D.J., Hynes, T.R., and Fox, R.O. 2.9 Å resolution structure of an anti-dinitrophenyl-spin-label monoclonal antibody F_{ab} fragment with bound hapten. *J. Mol. Biol.*, 221:239–256, 1991.

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

[64] Stanfield, R.L., Fieser, T.M., Lerner, R.A., and Wilson, I.A. Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 Å. *Science*, 248:712–719, 1990.

[65] Grunow, R., Jahn, S., Porstman, T., Kiessig, T., Steinkeller, H., Steindl, F., Mattanovich, D., Gurtler, L., Deinhardt, F., Katinger, H., and Baehr von R. The high efficiency, human B cell immortalizing heteromyeloma CB-F7. *J. Immunol. Meth.*, 106:257–265, 1988.

[66] Fan, Z.C., Shan, L., Guddat, L.W., He, X.M., Gray, W.R., Raison, R.L., and Edmunson, A.B. Three dimensional structure of an Fv from a human IGM immunoglobulin. *J. Mol. Biol.*, 228:188–207, 1992.

[67] Tormo, J., Stadler, E., Skern, T., Auer, H., Kanzler, O., Betzel, C., Blaas, D., and Fita, I. Three dimensional structure of the F_{ab} fragment of a neutralizing antibody to human rhinovirus serotype 2. *Protein Sci.*, 1:1154–1161, 1992.

[68] Tulip, W.R., Varghese, J.N., Laver, W.G., Webster, R.G., and Colman, P.M. Refined crystal structure of the influenza virus N9 neuraminidase/NC41 F_{ab} complex. *J. Mol. Biol.*, 227:122–148, 1992.

[69] Jeffrey, P.D., Strong, R.K., Sieker, L.C., Chang, C.Y., Campbell, R.L., Petsko, G.A., Haber, E., Margolies, M.N., and Sheriff, S. 26-10 F_{ab}-digoxin complex - affinity and specificity due to surface complimentarity. *Proc. Natl. Acad. Sci. USA*, 90:10310–10314, 1993.

[70] Rini, J.M., Stanfield, R.L., Stura, E.A., Salinas, P.A., Profy, A.T., and Wilson, I.A. Crystal structure of a human immunodeficiency virus type 1 neutralizing antibody, 50.1, in complex with its V3 loop peptide antigen. *Proc. Natl. Acad. Sci. USA*, 90:6325–6329, 1993.

[71] Ghiara, J.B., Stura, E.A., Stanfield, R.L., Profy, A.T., and Wilson, I.A. Crystal structure of the principal neutalizing site of HIV-1. *Science*, 264:82, 1994.

[72] Brady, R.L., Edwards, D.J., Hubbard, R.E., Jiang, J.S., Lange, G., Roberts, S.M., Todd, R.J., Adair, J.R., Emtage, J.S., King, D.J., and Low, D.C. Crystal structure of a chimeric F_{ab}' fragment of an antibody binding tumour cells. *J. Mol. Biol.*, 227:253–264, 1992.

[73] Schormann, N., Murrell, J.R., Liepnieks, J.J., and Benson, M.D. Tertiary structure of an amyloid immunoglobulin light chain protein: A proposed model for amyloid fibril formation. *Proc. Natl. Acad. Sci. USA*, 92:9490–9494, 1995.

[74] Herron, J.N., He, X.M., Ballard, D.W., Blier, P.R., Pace, P.E., Bothwell, A.L.M., Voss Jr., E., and Edmundson, A.B. An autoantibody to single-

stranded DNA: Comparison of the three-dimensional structures of the unliganded F_{ab} and a deoxynucleotide-F_{ab} complex. *Proteins: Struct., Funct., Genet.*, 11:159–175, 1991.

[75] Zhou, G.W., Guo, J., Huang, W., Scanlan, T.S., and Fletterick, R.J. Crystal structure of a catalytic antibody with a serine protease active site. *Science*, 265:1059–1064, 1994.

[76] Lescar, J., Pellegrini, M., Souchon, H., Tello, D., Poljak, R., Peterson, N., Greene, M., and Alzari, P. Crystal structure of a cross-reaction complex between F_{ab} F9.13.7 and guinea-fowl lysozyme. *J. Biol. Chem.*, 270:18067–18076, 1995.

[77] Eigenbrot, C., Gonzalez, T., Mayeda, J., Carter, P., Werther, W., Hotaling, T., Fox, J., and Kessler, J. X-ray structures of fragments from binding and nonbinding versions of a humanized anti-CD18 antibody - structural indications of the key role of V_H residues 59 to 65. *Proteins: Struct., Funct., Genet.*, 18:49–62, 1994.

[78] Haynes, M.R., Stura, E.A., Hilvert, D., and Wilson, I.A. Routes to catalysis: Structure of a catalytic antibody and comparison with its natural counterpart. *Science*, 263:646–652, 1994.

[79] Liu, H., Smith, T.J., Lee, W., Mosser, A.G., Rueckert, R.R., Olson, N.H., Cheng, R.H., and Baker, T.S. Structure determination of an F_{ab} fragment

that neutralizes human rhinovirus and analysis of the F_{ab} virus complex. *J. Mol. Biol.*, 240:127–137, 1994.

[80] Wien, M.W., Filman, D.J., Stura, E.A., Guillot, S., Delpeyroux, F., Crainic, R., and Hogle, J.M. Three-dimensional structure of the complex between the F_{ab} fragment of a neutralizing antibody for type 1 poliovirus and its viral epitope. *Nat.Struct.Biol*, 2:232–243, 1995.

[81] Churchill, M.E.A., Stura, E., Pinilla, C., Appel, J.R., Houghten, R.A., Kono, D.H., Balderas, R.S., Fieser, G.G., Schulze-Gahmen, U., and Wilson, I.A. Crystal structure of a peptide complex of anti-influenza peptide antibody F_{ab} 26/9: Comparison of 2 different antibodies bound to the same peptide antigen. *J. Mol. Biol.*, 241:534–556, 1994.

[82] Eigenbrot, C., Randal, M., Presta, L., Carter, P., and Kossiakoff, A.A. X-ray structures of the antigen-binding domains from three variant of humanized anti-p185(her2) antibody 4D5 and comparison with molecular modeling. *J. Mol. Biol.*, 229:969–995, 1993.

[83] Bizebard, T., Gigant, B., Rigolet, P., Rasmussen, B., Diat, O., Bosecke, P., Wharton, S.A., Skehel, J.J., and Knossow, M. Structure of influenza virus haemagglutinin complexed with a neutralizing antibody. *Nature (London)*, 376:92–94, 1995.

[84] Derrick, J.P. and Wigley, D.B. The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure

alone and in a complex with F$_{ab}$. *J. Mol. Biol.*, 243:906–918, 1994.

[85] Altschuh, D., Vix, O., Rees, B., and Thierry, J.C. A conformation of cyclosporin A in aqueous environment revealed by the X-ray structure of a cyclosporin-F$_{ab}$ complex. *Science*, 256:92–94, 1992.

[86] Love, R.A., Villafranca, J.E., Aust, R.M., Nakamura, K.K., Jue, R.A., Major, J.G.J., Radhakrishnan, R., and Butler, W.F. How the anti-(metal chelate) antibody CHA255 is specific for the metal ion of its antigen: X-ray structures for two F$_{ab}$/hapten complexes with different metals in the chelate. *Biochemistry*, 32:10950–10959, 1993.

[87] Essen, L.O. and Skerra, A. The de novo design of an antibody combining site: Crystallographic analysis of the V$_L$ domain confirms the structural model. *J. Mol. Biol.*, 238:226–244, 1994.

[88] Prasad, L., Sharma, S., Vandonselaar, M., Quail, J.W., Lee, J.S., Waygood, E.B., Wilson, K.S., Dauter, Z., and Delbaere, L.T.J. Evaluation of mutagenesis for epitope mapping: Structure of an antibody/protein antigen complex. *J. Biol. Chem.*, 268:10705–10708, 1993.

[89] Chitarra, V., Alzari, P.M., Bentley, G.A., Bhat, T.N., Eisele, J.L., Houdusse, A., Lescar, J., Souchon, H., and Poljak, R.J. Three-dimensional structure of a heteroclitic antigen-antibody cross-reaction complex. *Proc. Natl. Acad. Sci. USA*, 90:7711–7715, 1993.

[90] Perisic, O., Webb, P.A., Holliger, P., Winter, G., and Williams, R. The structure of a bivalent diabody. *Structure*, 2:1217, 1994.

[91] Braden, B.C., Souchon, H., Eisele, J.L., Bentley, G.A., Bhat, T.N., Navaza, J., and Poljak, R.J. Three-dimensional structures of the free and the antigen-complexed F_{ab} from monoclonal anti-lysozyme antibody D44.1. *J. Mol. Biol.*, 243:767–781, 1994.

[92] Tulip, W.R., Harley, V.R., Webster, R.G., and Novotny, J. N9 neuraminidase complexes with antibodies NC41 and NC10: Empirical free-energy calculations capture specificity trends observed with mutant binding data. *Biochemistry*, 33:7986–7997, 1994.

[93] Bossart-Whitaker, P., Chang, C.Y., Novotny, J., Benjamin, D.C., and Sheriff, S. The crystal structure of antibody N10-staphylococcal nuclease complex at 2.9Å resolution. *J. Mol. Biol.*, 253:559–575, 1995.

[94] Kodandapani, R., Veerapandian, B., Kunicki, T.J., and Ely, K.R. Crystal structure of the OPG2 F_{ab}: An antireceptor antibody that mimics an rgd cell adhesion site. *J. Biol. Chem.*, 270:2268–2273, 1995.

[95] Jedrzejas, M.J., Miglietta, J., Griffin, J.A., and Luo, M. Structures of monoclonal anti-ICAM-1 antibody R6.6 fragment at 2.8Å resolution. *To be published*, 1995.

[96] Shoham, M. Crystal structure of an anticholera toxin peptide complex at 2.3Å. *J. Mol. Biol.*, 232:1169–1175, 1993.

[97] Bhat, T.N., Bentley, G.A., Fischmann, T.O., Boulot, G., and Poljak, R.J. Small rearrangements in structures of F_V and F_{ab} fragments of an antibody D1.3 on antigen binding. *Nature (London)*, 347:483–485, 1990.

[98] Huang, D.B., Chang, C.H., Ainsworth, C., Brünger, A.T., Eulitz, M., Solomon, A., Stevens, F.J., and Schiffer, M. Comparison of crystal structures of two homologous proteins: Structural origin of altered domain interactions in immunoglobulin light chain dimers. *Biochemistry*, 33:14848–14857, 1994.

[99] Guddat, L.W., Shan, L., Anchin, J.M., Linthicum, D.S., and Edmunson, A.B. Local and transmitted conformational changes on complexation of an anti-sweetener F_{ab}. *J. Mol. Biol.*, 236:247–274, 1994.

[100] Arevalo, J.H., Taussig, M.J., and Wilson, I.A. Molecular basis of crossreactivity and the limits of antibody-antigen complementarity. *Nature (London)*, 365:859–863, 1993.

[101] Lascombe, M.B., Alzari, P.M., Poljak, R.J., and Nisonoff, A. Three-dimensional structure of two crystal forms of F_{ab} R19.9 from a monoclonal anti-arsonate antibody. *Proc. Natl. Acad. Sci. USA*, 89:9429–9433, 1992.

[102] Charbonnier, J.B., Carpenter, E., Gigant, B., Golinelli-Pimpaneau, B., Eshhar, Z., Green, B.S., and Knossow, M. Crystal structure the complex of a catalytic antibody F_{ab} fragment with a transition state analogue: Structural

similarities in esterase-like catalytic antibodies. *Proc. Natl. Acad. Sci. USA*, 92:11721–11725, 1995.

[103] Chang, C.H., Short, M.T., Westholm, F.A., Stevens, F.J., Wang, B.C., Furey, W.J., Solomon, A., and Schiffer, M. Novel arrangement of immunoglobulin variable domains: X-ray crystallographic analysis of the λ-chain dimer Bence-Jones protein LOC. *Biochemistry*, 24:4890, 1985.

[104] Novotny, J., Bruccoleri, R., Newell, J., Murphy, D., Haber, E., and Karplus, M. Molecular anatomy of the antibody binding site. *J. Biol. Chem.*, 258:14433–14437, 1983.

[105] Toyonaga, B., Yoshikai, Y., Vadasz, V., Chin, B., and Mak, T.W. Organization and sequences of the diversity, joining, and constant region genes of the human T-cell receptor β chain. *Proc. Natl. Acad. Sci. USA*, 82:8624–8628, 1985.

[106] Satyanarayana, K., Hata, S., Devlin, P., Roncarolo, M.G., Vries, J.E.D., Spits, H., Strominger, J.L., and Krangel, M.S. Genomic organization of the human T-cell antigen-receptor α/δ locus. *Proc. Natl. Acad. Sci. USA*, 85:8166–8170, 1988.

[107] Winoto, A., Mjolsness, S., and Hood, L. Genomic organization of the genes encoding mouse T-cell receptor α-chain. *Nature (London)*, 316:832–836, 1985.

[108] Davis, M.M. T cell receptor gene diversity and selection. *Annu. Rev. Biochem.*, 59:475–496, 1990.

[109] Fink, P.J., Matis, L.A., McElligott, D.L., Bookman, M., and Hedrick, S.M. Correlations between T-cell specificity and the structure of the antigen receptor. *Nature (London)*, 321:219–236, 1986.

[110] Hochgeschwender, U., Simon, H., Weltzien, H.U., Bartels, F., Becker, A., and Epplen, J.T. Dominance of one T-cell receptor in the H-2K^b/TNP response. *Nature (London)*, 326:307–309, 1987.

[111] Wade, T., Bill, J., Marrack, P.C., Palmer, E., and Kappler, J.W. Molecular basis for the nonexpression of Vβ17 in some strains of mice. *J. Immunol.*, 141:2165–2167, 1988.

[112] Tan, K., Datlof, B.M., Gilmore, J.A., Kronman, A.C., Lee, J.H., Maxam, A.M., and Rao, A. The T cell receptor Vα3 gene segment is associated with reactivity to p-azonbenzenearsonate. *Cell*, 54:247–261, 1988.

[113] Goverman, J., Minard, K., Shastri, N., Hunkapiller, T., Hansburg, D., Sercarz, E., and Hood, L. Rearranged β T cell receptor genes in a helper T cell clone specific for lysosyme: No correlation between Vβ and MHC restriction. *Cell*, 40:859–867, 1985.

[114] Wilson, I.A., Rini, J.M., Fremont, D.H., Fieser, G.G., and Stura, E.A. X-ray crystallographic analysis of free and antigen-complexed F$_{ab}$ fragments

to investigate structural basis of immune recognition. *Meth. Enz.*, 203:153–176, 1991.

[115] Davies, D.R. and Chacko, S. Antibody structure. *Acc. Chem. Res.*, 26:241–427, 1993.

[116] Davies, D.R., Padlan, E.A., and Sheriff, S. Antibody-antigen complexes. *Annu. Rev. Biochem.*, 59:439–473, 1990.

[117] Novotny, J. and Haber, E. Structural invariants of antigen binding: Comparisons of immunoglobulin V_L-V_H and V_L-V_L domain dimers. *Proc. Natl. Acad. Sci. USA*, 82:4592–4596, 1985.

[118] Chothia, C., Novotny, J., Bruccoleri, R., and Karplus, M. Domain association in immunoglobulin molecules. *J. Mol. Biol.*, 186:651–663, 1985.

[119] Chothia, C., Lesk, A.M., Levitt, M., Amit, A.G., Mariuzza, R.A., Phillips, S.E.V., and Poljak, R.J. The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. *Science*, 233:755–758, 1986.

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

[121] Chothia, C., Lesk, A.M., Gherardi, E., Tomlinson, I.M., Walter, G., Marks, J.D., Llewelyn, M.B., and Winter, G. Structural repertoire of the human V_H segments. *J. Mol. Biol.*, 227:799–817, 1992.

[122] Wu, S. and Cygler, M. Conformation of complementarity determining region L1 loop in murine IgG λ light chains extends the repertoire of canonical forms. *J. Mol. Biol.*, 229:597–601, 1993.

[123] Padlan, E.A. On the nature of antibody combining sites: Unusual structural features that may confer on these sites an enhanced capacity for binding ligands. *Proteins: Struct., Funct., Genet.*, 7:112–124, 1990.

[124] Williams, A.F., Strominger, J.L., Bell, J.I., Mak, T.W., Kappler, J., Marrack, P., Arden, B., Lefranc, M.P., Hood, L., Tonegawa, S., Davis, M.M., and Kazatchkine, M.D. WHO-IUIS nomenclature sub-committee on TCR designation: Nomenclature for T-cell receptor (TCR) gene segments of the immune system. *Immunogenetics*, 42:451–453, 1995.

[125] Arden, B., Clark, S.P., Kabelitz, D., and Mak, T.W. Human T-cell receptor variable gene segment families. *Immunogenetics*, 42:455–500, 1995.

[126] Arden, B., Clark, S.P., Kabelitz, D., and Mak, T.W. Mouse T-cell receptor variable gene segment families. *Immunogenetics*, 42:501–530, 1995.

[127] Chothia, C., Bosswell, D.R., and Lesk, A.M. The outline structure of the T-cell $\alpha\beta$ receptor. *EMBO J.*, 7:3745–3755, 1988.

[128] Jores, R., Alzari, P.M., and Meo, T. Resolution of hypervariable regions in T-cell receptor β chains by a modified Wu-Kabat index of amino acid diversity. *Proc. Natl. Acad. Sci. USA*, 87:9138–9142, 1990.

[129] Kabat, E.A., Wu, T.T., Reid-Miller, M., Perry, H.M., and Gottesman, K.S. *Sequences of Proteins of Immunological Interest.* U.S. Department of Health and Human Services, Fifth edition, 1991.

[130] Searle, S.M.J. MOL: A protein analysis program. Unpublished.

[131] Searle, S.M.J. SR: A sequence analysis program. Unpublished.

[132] Barton, G.J. and Sternberg, M.J.E. A strategy for the rapid multiple alignment of protein sequences. *J. Mol. Biol.*, 198:327–337, 1987.

[133] Thompson, J.D., Higgins, D.G., and Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nuc. Ac. Res.*, 22:4673–4680, 1994.

[134] Genetics Computer Group. *Program Manual for the Wisconsin Package.* Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711, eighth edition, 1994.

[135] Wu, T.T. and Kabat, E.A. An analysis of the sequences of variable regions of Bence Jones proteins and myeloma light chains and the implications for antibody complementarity. *J. Exp. Med.*, 132:211–250, 1970.

[136] Pedersen, J.T. *Molecular Modelling of Antibody Combining Sites.* D. Phil. Thesis, University of Bath, 1993.

[137] Sutcliffe, M.J., Haneef, I., Carney, D., and Blundell, T.L. Knowledge based modelling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.*, 1:377–384, 1987.

[138] Becker, D.M., Patten, P., Chien, Y., Yokota, T., Eshhar, Z., Giedlin, M., Gascoigne, N.R.J., Goodnow, C., Wolf, R., Arai, K., and Davis, M.M. Variability and repertoire size of T-cell receptor Vα gene segments. *Nature (London)*, 317:430–434, 1985.

[139] Reid, E., Cook, G.M.W., and Morre, D.J., editors. *Investigation and Exploitation of Antibody Combining Sites.* Plenum, 1984.

[140] Vitetta, E.S., Fulton, R.J., May, R.D., Till, M., and Uhr, J.W. Redesigning nature's poisons to create anti-tumor reagents. *Science*, 238:1098–1104, 1987.

[141] Shultz, P.G. Catalytic antibodies. *Angewandte Chemie*, 28:1283–1295, 1989.

[142] Needleman, S.B. and Wunch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[143] McLachlan, A.D. Gene in the structural evolution of chymotrypsin. *J. Mol. Biol.*, 128:49–79, 1979.

[144] Brucoleri, R.E. and Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, 26:137–168, 1987.

[145] Palmer, K.A. and Scheraga, H.A. Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid-geometry approximation. I. Chain closure through a limited search of "loop" conformations. *J. Comp. Chem.*, 12:505–526, 1991.

[146] Go, N. and Sheraga, H.A. Ring closure and local conformational deformations of chain molecules. *Macromolecules*, 3:178–187, 1970.

[147] Higo, J., Collura, V., and Garnier, J. Development of an extended simulated annealing method: Application to the modelling of complementary determining regions of immunoglobulins. *Biopolymers*, 32:33–43, 1992.

[148] Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L., and Levinthal, C. Predicting antibody hypervariable loop conformations II: Minimisation and molecular dynamics studies of McPC603 from many randomly generated loop conformations. *Proteins: Struct., Funct., Genet.*, 1:342–362, 1986.

[149] Lifson, S., Hagler, A., and Dauber, P. Consistent force field studies of intermolecular forces in hydrogen bonded crystals. 1. carboxylic acids, amides, and the C=0...H-hydrogen bonds. *J. Am. Chem. Soc.*, 101:55–111, 1979.

[150] Discover manual. Biosym Technologies Inc.

[151] Summers, N.L., Carlson, W.D., and Karplus, M. Analysis of side-chain orientations in homologous proteins. *J. Mol. Biol.*, 196:175–198, 1987.

[152] McGregor, M.J., Islam, S.A., and Sternberg, M.J.E. Analysis of the relationship between side-chain conformation and secondary structure in globular-proteins. *J. Mol. Biol.*, 198:295–310, 1987.

[153] Sutcliffe, M.J., Hayes, F.R.F., and Blundell, T.L. Knowledge based modelling of homologous proteins, part II: Rules for the conformations of substituted sidechains. *Protein Eng.*, 1:385–392, 1987.

[154] Ponder, J. and Richards, F. Internal packing and protein structural classes. *Cold Spring Harbor Quant. Symp. Biochem.*, 52:421–428, 1987.

[155] Ponder, J. and Richards, F. Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193:775–791, 1987.

[156] Lee, C. and Levitt, M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature (London)*, 352:448–451, 1991.

[157] Lee, C. and Subbiah, S. Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.*, 217:373–388, 1991.

[158] Padlan, E.A., Davies, D.R., Pecht, I., Givol, D., and Wright, C. Model building studies of antigen-binding sites: The hapten-binding site of

MOPC-315. *Cold Spring Harbor Quant. Symp. Biochem.*, 41:627–637, 1976.

[159] Åqvist, J., Van Gunsteren, W.F., Leifonmark, M., and Tapia, O. A molecular-dynamics study of the C-terminal fragment of the L7/L12 ribosomal-protein - secondary structure motion in a 150 picosecond trajectory. *J. Mol. Biol.*, 183:461–477, 1985.

[160] Dauber-Osguthorpe, P., Roberts, V.A., Osguthorpe, D.J., Wolff, J., Genest, M., and Hagler, A.T. Structure and energetics of ligand binding to proteins: Escherichia coli dihydrofolate reductase-trimethoprim, a drug-receptor system. *Proteins: Struct., Funct., Genet.*, 4:31–47, 1988.

[161] Walsh, L.L. Annotated PDB file listing. 1993. personnal communication.

[162] Borrebaeck, C.A.K. *Antibody Engineering*. Oxford University Press, second edition, 1995.

[163] Dayhoff, M.O., Barker, W.C., and Hunt, L.T. Establishing homologies in protein sequences. *Meth. Enz.*, 91:524–545, 1983.

[164] Jorgensen, J.L., Esser, U., de St. Groth, B.F., Reay, P.A., and Davis, M.M. Mapping T-cell receptor-peptide contacts by variant peptide immunization of single-chain transgenics. *Nature (London)*, 355:224–230, 1992.

[165] Yanagi, Y., Yoshikai, Y., Leggett, K., Clark, S.P., Aleksander, I., and Mak, T.W. A human T cell-specific cDNA clone encodes a protein having exten-

sive homology to immunoglobulin chains. *Nature (London)*, 308:145–149, 1984.

[166] Loh, E.Y., Elliot, J.F., Cwirla, S., Lanier, L.L., and Davis, M.M. Polymerase chain reaction with single-sided specificity - analysis of T-cell receptor δ-chain. *Science*, 243:243–247, 1989.

[167] Patten, P.A., Rock, E.P., Sonoda, T., de St. Groth, B.F., Jorgensen, J.L., and Davis, M.M. Transfer of putative complementarity-determining region loops of T cell receptor V domains confers toxin reactivity but not peptide/MHC specificity. *J. Immunol.*, 150:2281–2294, 1993.

[168] Nalefski, E.A., Wong, J.G.P., and Rao, A. Amino acid substitutions in the first complementarity-determining region of a murine T-cell receptor α chain affect antigen-major histocompatibility complex recognition. *J. Biol. Chem.*, 265:8842–8846, 1990.

[169] Engel, I. and Hedrick, S.M. Site-directed mutations in the VDJ junctional region of a T cell receptor β chain cause changes in antgenic peptide recognition. *Cell*, 54:473–484, 1988.

[170] Hedrick, S.M., Engel, I., McElligott, D.L., Fink, P.J., Hsu, M., Hansburg, D., and Matis, L.A. Selection of amino acid sequences in the β chain of the T cell antigen receptor. *Science*, 239:1541–1544, 1988.

[171] Gross, G. and Eshhar, Z. Endowing T cells with antibody specificity using chimeric T cell receptors. *Faseb J.*, 6:3370–3378, 1992.

[172] Novotny, J., Tonegawa, S., Saito, H., Kranz, D.M., and Eisen, H.N. Secondary, tertiary, and quaternary structure of T-cell-specific immunoglobulin-like polypeptide chains. *Proc. Natl. Acad. Sci. USA*, 83:742–746, 1986.

[173] Kaymaz, H., Dedeoglu, F., Schluter, S.F., Edmundson, A.B., and Marchalonis, J.J. Reactions of anti-immunoglobulin sera with synthetic T cell receptor peptides: Implications for the three-dimensional structure and function of the TCR β chain. *Int. Immunol.*, 5:491–502, 1993.

[174] Matsui, K., Boniface, J.J., Steffner, P., Reay, P.A., and Davis, M.M. Kinetics of T-cell receptor binding to peptide/I-E^k complexes: Correlation of the dissociation rate with T-cell responsiveness. *Proc. Natl. Acad. Sci. USA*, 91:12862–12866, 1994.

[175] Malmqvist, M. Biospsecific interaction analysis using biosensor technology. *Nature (London)*, 361:186–187, 1993.

[176] Sykulev, Y., Brunmark, A., Tsomides, T.J., Kageyama, S., Jackson, M., Peterson, P.A., and Eisen, H.N. High-affinity reactions between antigen-specific T-cell receptors and peptides associated with allogeneic and syngeneic major histocompatibility complex class I proteins. *Proc. Natl. Acad. Sci. USA*, 91:11487–11491, 1994.

[177] Alam, S.M., Travers, P.J., Wung, J.L., Nasholds, W., Redpath, S., Jameson, S.C., and Gascoigne, N.R.J. T-cell-receptor affinity and thymocyte positive

selection. *Nature (London)*, 381:616–620, 1996.

[178] Margulies, D.H. An affinity for learning. *Nature (London)*, 381:558–559, 1996.

[179] Davis, M.M. Serial engagement proposed. *Nature (London)*, 375:104, 1995.

[180] Novotny, J., Ganja, R.K., Smiley, S.T., Hussey, R.E., Luther, M.A., Recny, M.A., Silicano, R.F., and Reinherz, E.L. A soluble, single chain T-cell receptor fragment endowed with antigen-combining properties. *Proc. Natl. Acad. Sci. USA*, 88:8646–8650, 1991.

[181] BuchWalder, A., Krangel, M.S., Hao, P., and Diamond, D.J. Immunochemical and molecular analysis of antigen binding to lipid anchored and soluble forms of an MHC independent human α/β T cell receptor. *Mol. Immunol.*, 31:857–872, 1994.

[182] Epp, O., Lattman, E.E., Schiffer, M., Huber, R., and Palm, W. The molecular structure of a dimer composed of the variable portions of the Bence-Jones protein Rei refined at 2.0 Å resolution. *Biochemistry*, 14:4963–4975, 1975.

[183] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The protein databank: A computer based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

[184] Wedderburn, L.R., O'Hehir, R.E., Hewitt, C.R.A., Lamb, J.R., and Owen, M.J. *In vivo* clonal dominance and limited T-cell receptor usage in human CD4$^+$ T-cell recognition of house dust mite allergens. *Proc. Natl. Acad. Sci. USA*, 90:8214–8218, 1993.

[185] Wedderburn, L.R., Searle, S.J.M., Rees, A.R., Lamb, J.R., and Owen, M.J. Mapping T cell recognition: The identification of a T cell receptor residue critical to the specific interaction with an influenza hemagglutinin peptide. *Eur. J. Immun.*, 25:1654–1662, 1995.

[186] Davis, M.M. and Bjorkman, P.J. T-cell antigen receptor genes and T-cell recognition. *Nature (London)*, 334:395–402, 1988.

[187] Hale, G., Dyer, M.J.S., Clark, M.R., and Waldmann, H. Development and clinical-experience with humanized monoclonal-antibodies. *Developments in Biotherapy*, 1(1):195–199, 1991.

[188] Verhoeyen, M.E., Saunders, J.A., Broderick, E.L., Eida, S.J., and Badley, R.A. Reshaping human monoclonal-antibodies for imaging and therapy. *Disease Markers*, 9(3):4, 1991.

[189] Kyle, V., Roddy, J., Hale, G., Hazleman, B.L., and Waldmann, H. Humanized monoclonal-antibody treatment in rheumatoid-arthritis. *Journal of Rheumatology*, 18(11):1737–1738, 1991.

[190] Crowe, J.S., Hall, V.S., Smith, M.A., Cooper, H.J., and Tite, J.P. Humanized monoclonal antibody campath-1H - myeloma cell expression of

genomic constructs nucleotide sequence of cDNA constructs and comparison of effector mechanisms of myeloma and chinese-hamster ovary cell-derived material. *Clinical and Experimental Immunology*, 87(1):105–110, 1992.

[191] Roguska, M.A., Pedersen, J.T., Keddy, C.A., Henry, A.H., Searle, S.J., Lambert, J.M., Goldmacher, V.S., Blattler, W.A., Rees, A.R., and Guild, B.C. Humanization of murine monoclonal antibodies through variable domain resurfacing. *Proc. Natl. Acad. Sci. USA*, 91:969–973, 1994.

[192] Pedersen, J.T., Henry, A.H., Searle, S.J., Guild, B.C., Roguska, M., and Rees, A.R. Comparison of surface accessible residues in human and murine immunoglobulin F_v domains. Implication for humanization of murine antibodies. *J. Mol. Biol.*, 235:959–973, 1994.

[193] Roguska, M.A., Pedersen, J.T., Henry, A.H., Searle, S.M.J., Roja, C.M., Avery, B., Hoffee, M., Cook, S., Lambert, J.M., Blattler, W.A., Rees, A.R., and Guild, B.C. A comparison of two murine monoclonal antibodies humanized by CDR-grafting and variable domain resurfacing. *Protein Eng.*, 9:895–904, 1996.

[194] Kabsch, W. and Sander, C. Dictionary of protein secondary structure. *Biopolymers*, 22:2577–2637, 1983.

[195] AbM. An automated immunoglobulin modelling program. Oxford Molecular Ltd., The Magdalen Centre, Oxford Science Park, Sandford-on-Thames,

Oxford, U.K.

[196] Rognan, D., Scapozza, L., Folkers, G., and Daser, A. Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry*, 33:11476–11485, 1994.

Appendix A

Programs Used and Written

A.1 ALSCRIPT

Produces colour sequence alignment postscript files. It was written by Geoff Barton (gjb@bioch.ox.ac.uk) and is available from him.

A.2 AMPS

Geoff Barton's multiple sequence alignment program. It is available from him (gjb@bioch.ox.ac.uk).

A.3 AbM

The commercial version of the CAMAL antibody modelling program suite originally written by Andrew C.R. Martin [6], and modified by Jan Pedersen and the author. It is maintained by, and available from, Oxford Molecular Ltd.

AbM has a curses menu interface in which the sequences are entered and modelling options set. The interface program writes a command file for the builder program. This program does some of the processing steps itself, while others are performed by separate programs, which are run by the builder. These programs are:-

- FRAMEBUILD. Builds the framework regions.

- CHOTH. Models the canonical loops and replaces the CDR H3 loop with one from the same H3 takeoff angle class.

- CONGEN. Generates loop conformations and adds sidechains to the CDR residues.

- EUREKA. Perform the energy screen of the generated loop conformations.

A.4 TCRM

A version of AbM modified by the author for use in some of the modelling described in this thesis. This had greater flexibility than AbM in areas such as loop range definition, choice of framework and selection of modelling method. Many of the flexibility enhancements in TCRM have since been added to AbM.

A.5 INSIGHT

A commercial molecular display program published by Biosym Ltd.

A.6 DISCOVER

A commercial molecular dynamics/minimisation program published by Biosym Ltd.

A.7 MASE

A multiple alignment sequence editor written by Don Faulkner. The program is available from mbcrr.harvard.edu. This program has been modified by the author to allow NBRF format file reading and writing, and some bugs have been fixed.

A.8 MULFIT

This is a multiple structure fitting program written by Jan Pedersen (jan@iris8.carb.nist.gov).

A.9 McSIDE

A protein structural analysis and sidechain generation program (using Monte Carlo techniques) written by Jan Pedersen which also includes code for producing postscript ball and stick plots.

A.10 MOLSCRIPT

This program written by Per Kraulis produces protein schematic pictures. It is available from him (pjk@ciclid.csb.ki.se). Modifications were made by the author to molscript (v1.4) to allow individual secondary structure elements to have more than one colour, with the colour being specified by the atom colour of the Cα atom of each residue.

A.11 RASTER3D and IMAGEMAGICK

These programs were used to create and display rendered versions of the MOLSCRIPT plots. RASTER3D was originally written by David J. Bacon and Wayne F. Anderson with modifications by Mark Israel, Stephen Samuel, Michael Murphy, Albert Berghuis and Ethan A Merritt (merritt@u.washington.edu). It is available from ftp.bmsc.washington.edu. IMAGEMAGICK is written by John Cristy (cristy@dupont.com) and available as:

ftp://ftp.x.org/contrib/applications/ImageMagick/ImageMagick-3.6.1.tar.gz.

A.12 GRASP

GRASP generates pictures of molecular surfaces. It is available from Anthony Nicholls (grasp@cumbih.bioc.columbia.edu).

A.13 MOL

MOL is a program written by the author of this thesis to analyse sequence structure, and structure structure relationships in proteins. It includes code to perform:

- Environmental analysis.
- Kabat fixlen format reading. The program performs the following checks on the data:
  - Format of references.
  - Inserted residues are in correct place and have valid three letter code.
  - Line lengths are correct for SEQTPA entries.
  - Looks for missing AAIN (inserted residue) entries by scanning the comments for certain phrases.
- Kabat database search for strings. This can be used to pull out the various domains e.g. T cell receptor beta chain V regions, or to do more specific searches of the Kabat database.
- Protein energy calculations (using vff potential).
- Protein accessibility calculations.
- Alignment of two protein sequence groups.
- Theoretical Removal of TCR Epitopes.
- Basic molecular display .
- Atom selection.
- Structure RMS calculations.
- Structural fitting of proteins.

A.14 SR (QUALIS)

SR (which has also been called QUALIS in publications) is a sequence analysis program written by the author of this thesis to overcome some of the problems experienced with the available sequence analysis programs:

- It uses dynamic memory allocation to allow handling of the large numbers of sequences that immunoglobulin sequence analysis entails.
- It allows multiple chains from a single clone to be associated with one another, which enables the analysis of α / β chain pairs.
- Sequence searches, alignments, translations and variability, frequency and homology calculations can all be performed in a single program.
- Command files allow a degree of automation of repetitive tasks. Searches on properties such as specific sequence patterns, title, journal etc. can be performed to create subsets. A prosite pattern search routine can identify sequence motifs in the sequences.

The program can read NBRF, PIR, GCG, EMBL, Kabat FixLen and AMPS block file formats. It can write AMPS input, NBRF and PIR formats.

Sequence patterns, specified in the prosite format are used to search for specific sequences. Files containing prosite motifs can also be read. Sets of sequences can be scanned against these motif databases to identify any motifs they contain.

Sequence pairs can be identified in SR. The program compares clone name entries and creates links between entries with matching names. New comments are added to the entries indicating the entry ID of the paired sequence, and the pair can be written out in a modified NBRF format in which multiple sequences may be specified for each entry.

If the sequence data originated from the Kabat database and contains residues which do not fit in the standard alignment, SR can expand the alignment to include all the inserted residues by determining the maximum number of inserted residues at each position and then adding gaps to maintain alignment of the other sequences.

The DNA translation function reads GCG format translation table. If the sequences were read from a GCG or EMBL format file, features from the entries in the file can be used to identify the regions to translate. If a CDS feature exists for an entry, this feature is selected by default for translation. If no features exist for an entry or if the user prefers, s/he can specify the regions to translate as sequence positions. A logfile is maintained indicating which features or ranges were translated for each sequence in the list.

A.15 Conversion Utilities

A.15.1 RENAT and RENUMPDB

These were used to renumber PDB files. RENAT was written by the author and RENUMPDB by Jan T. Pederson.

A.15.2 TOUDB

This program was used to renumber the final model PDB file from AbM or TCRM to the standardised numbering scheme used in those programs. It was written by the author.

A.15.3 RDDUMP

This was used to convert from Kabat dump format to NBRF sequence format. It was written by the author.

A.15.4 READSEQ

This was used to convert between various sequence formats. This program was written by D. G. Gilbert (gilbert@bio.indiana.edu) and is available from ftp.bio.indiana.edu.

A.16 CHOTH Canonical Definition File Format

An example of a Canonical Definition File is shown in figure A.1. Comment lines begin with a #. The first non comment line identifies which loop the classes refer to. This has the keyword LOOP followed by the loop identifier. Following this line are multi-line records, one for each class. Each record begins with a line starting with the keyword CLASS and end with a line starting with the keyword ENDCLASS. Each class contains the following fields:

LENGTH <int> The length of the loop in the range defined by the RANGE field.

RANGE 2× <int> The range of residues to calculate length over.

POS <int> The position in the sequence to start pattern matching.

PAT <string> The prosite pattern to match.

FRAME 4× <int> The two framework regions to use for fitting

NTER <int> The residue number of the residue before the first residue to replace

CTER <int> The residue number of the residue after the last residue to replace

CHAIN <char> The chain identifier for the chain containing the loop

A.17 Machines used

Most of the programs were run on either Silicon Graphics Personal Irises and Indigos under Irix 3.3, 4.05 and 5.2, or on Hewlett Packard 720 and 735 workstations under HPUX 9. Some of the later work was done on a Pentium Pro PC under Linux.

A.18 Thesis Preparation

The thesis was written in LaTeX 2e. Many of the schematic diagrams were created in CorelDraw! v4 and v7 and Visio 4.0. The colour molecule figures were created as screen dumps on an SG. All figures were included as encapsulated Postscript files. Dvips was used to convert to a single final Postscript file, which was printed using a Hewlett Packard 5M Postscript laser printer, except the colour pages which were printed on a Tektronix Phaser 340 (my thanks go to David Osguthorpe for allowing me to use this) and an Epson Colour 800. The humanisation paper was scanned in using a Microtek E6 page scanner.

```
#L3 canonical classes S.M.J.Searle 22/7/92
LOOP L3
#class one
CLASS 1
LENGTH 9
RANGE 95 105
POS 94
PAT <C(1)-x(1)-[QNH]-x(6)-P.
FRAME 90 94 106 110
NTER 94
CTER 106
CHAIN 1
ENDCLASS
#class two
CLASS 2
LENGTH 9
RANGE 95 105
POS 94
PAT <C(1)-x(1)-Q-x(5)-P-P.
FRAME 90 94 106 110
NTER 94
CTER 106
CHAIN 1
ENDCLASS
#class three
CLASS 3
LENGTH 8
RANGE 95 105
POS 94
PAT <C(1)-x(1)-Q-x(6)-P.
FRAME 90 94 106 110
NTER 94
CTER 106
CHAIN 1
ENDCLASS
#end of L3 classes (3)
```

Figure A.1: The canonical definition file used for the CDR L3 loop of antibodies.

Appendix B

Data on Sequences and Structures

B.1 Numbering Schemes

Alignment in Thesis	Kabat Heavy	Kabat Kappa	Kabat Lambda	Kabat Alpha	Kabat Beta
1	1	1	1	0	1
2	2	2	2	1	2
3	3	3	3	2	3
4	4	4	4	3	4
5	5	5	5	4	5
6	6	6	6	5	6
7	7	7	7	6	7
8	8	8	8	7	8
9	9	9	9	8	9
10	-	10	10	9	10
11	10	11	11	10	11
12	11	12	12	11	12
13	12	13	13	12	13
14	13	14	14	13	14
15	14	15	15	14	15
16	15	16	16	15	16
17	16	17	17	16	17
18	17	18	18	17	18

Alignment in Thesis	Kabat Heavy	Kabat Kappa	Kabat Lambda	Kabat Alpha	Kabat Beta
19	18	19	19	18	19
20	19	20	20	19	20
21	20	21	21	20	21
22	21	22	22	21	22
23	22	23	23	22	23
24	23	24	24	23	24
25	24	25	25	24	25
26	25	26	26	25	26
27	26	-	-	-	-
28	27	27	27	26	27
29	28	D	D	27	28
30	29	D	D	28	29
31	30	D	D	29	30
32	31	D	D	30	30A
33	32	D	D	30A	-
34	-	D	D	D	-
35	-	D	D	D	-
36	-	D	D	D	-
37	-	D	D	D	-
38	D	D	D	D	-
39	D	32	32	D	31
40	D	33	33	32	32
41	D	34	34	33	33
42	36	35	35	34	34
43	37	36	36	35	35
44	38	37	37	36	36
45	39	38	38	37	37
46	40	39	39	38	38
47	41	40	40	39	39
48	42	41	41	40	40
49	43	42	42	41	41
50	44	43	43	42	42
51	45	44	44	43	43
52	46	45	45	44	44
53	47	46	46	45	45
54	48	47	47	46	46
55	49	48	48	47	47
56	50	49	49	48	48
57	52	50	50	49	49
58	52A	-	-	-	50
59	D	-	-	-	51
60	D	-	-	-	52
61	D	-	-	-	53
62	D	-	-	-	D
63	D	-	-	-	D
64	D	-	-	50	D
65	56	51	51	51	D
66	57	52	52	52	D

Alignment in Thesis	Kabat Heavy	Kabat Kappa	Kabat Lambda	Kabat Alpha	Kabat Beta
67	58	53	53	53	D
68	59	54	54	54	D
69	60	55	55	55	D
70	61	56	56	56	D
71	62	57	57	57	D
72	63	58	58	58	D
73	64	59	59	59	D
74	65	60	60	60	D
75	66	61	61	61	D
76	67	62	62	62	65
77	68	63	63	63	66
78	69	64	64	64	67
79	70	65	65	65	68
80	71	66	66	66	69
81	72	-	-	67	-
82	73	-	-	68	70
83	74	67	67	69	71
84	75	68	68	70	72
85	76	69	69	71	73
86	77	70	70	72	74
87	78	71	71	73	75
88	79	72	72	74	76
89	80	73	73	75	77
90	81	74	74	76	78
91	82	75	75	77	79
92	82A	76	76	78	80
93	82B	77	77	79	81
94	82C	78	78	80	82
95	83	79	79	81	83
96	84	80	80	82	84
97	85	81	81	83	85
98	86	82	82	84	86
99	87	83	83	85	87
100	88	84	84	86	88
101	89	85	85	87	89
102	90	86	86	88	90
103	91	87	87	89	91
104	92	88	88	90	92
105	93	89	89	91	93
106	94	90	90	D	94
107	95	91	91	D	96
108	96	92	92	D	97
109	D	93	93	D	D
110	D	D	D	D	D
111	D	D	D	D	D
112	D	D	D	D	D
113	D	D	D	D	D
114	D	D	D	D	D

Alignment in Thesis	Kabat Heavy	Kabat Kappa	Kabat Lambda	Kabat Alpha	Kabat Beta
115	D	D	D	D	D
116	D	D	D	D	D
117	D	D	D	D	D
118	D	D	D	D	D
119	D	D	D	D	D
120	D	D	D	D	D
121	D	D	D	D	D
122	D	D	D	D	D
123	D	D	D	D	D
124	D	D	D	D	D
125	D	D	D	D	D
126	D	D	D	D	D
127	D	D	D	D	D
128	D	D	D	D	D
129	D	D	D	D	D
130	D	D	D	D	D
131	D	D	D	D	D
132	D	D	D	D	D
133	D	D	D	D	D
134	D	D	D	102	D
135	D	D	D	103	105
136	D	96	96	104	106
137	D	97	97	105	107
138	103	98	98	106	108
139	104	99	99	107	109
140	105	100	100	108	110
141	106	101	101	109	111
142	107	102	102	110	112
143	108	103	103	111	113
144	109	104	104	112	114
145	110	105	105	113	115
146	111	106	106	114	116
147	112	107	D	115	116A
148	113	108	D	116	N
149	N	109	D	116A	N
150	N	N	D	N	N

Table B.1: The table shows the various Kabat numbering schemes and the scheme used in the multiple chain type alignment created in this thesis (see figures 2.2 and 2.3). There are some regions which are aligned differently in the alignments, and therefore it is not possible to match position numbers between them (these occur in the three CDR regions). These are indicated by a D.

B.2 IUIS Classification of Kabat Entries

This section contains six multipage tables of TCR sequences (tables B.2–B.7). The sequences are divided into their official class designations. The tables identify which sequences in the Kabat database are in each of the officially designated classes, and which of these sequences were used in the analysis. Inconsistencies between the Kabat sequences and those in the official designation papers [125, 126] are also noted.

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV1S1	008449 !	TT11'CL	In Paper.	Complete	Complete	X01133	??????
	008450	A10'CL		Complete	None	??????	??????
	008451	Ts3-3a		Complete	Complete	??????	??????
AV1S2	015626 P	TA31	In Paper.	Complete-1	Joining Errors	X02928	??????
	008460 !	NA5&10'CL	In Paper.	Complete	Complete	M27352	??????
	008458	TA84'CL		Complete	Complete	??????	??????
	008459	TA46'CL		Complete	Complete	??????	??????
AV1S3	008452	5H'CL	In Paper. 0 aa/1 nuc diff - Explained	Complete	None	X02833	??????
	008453 X !	5/10-20D'CL	In Paper.	Complete-1	Complete	X05733	009096
	008457	M14T Va1 sterile'CL		Partial	None	??????	??????
AV1S4	008445 X !	MOUSE E1'CL	In Paper.	Complete	Complete		009163
AV1S5	008455 X !	C5'CL	In Paper.	Complete	Complete		009145
AV1S6	008454 X !	C11'CL	In Paper. 1 aa/1 nuc diff - Unexplained: M26423 has correct sequence	Partial	Complete	M26423	009168
AV1S7	015643 P	MOUSE Va1-Ja6.19 (J6.19 in paper)	In Paper.	Complete	Joining Errors	M38104	??????
AV1S8	008441 X !	No. 8	In Paper.	Complete	Complete	X56701	009130
	008444	BVI/5.a1'CL (BV1/5.a1 in paper)	In Paper. 0 aa/1 nuc diff - Explained	Complete	Complete	M31647	??????
	008442	B10-4'CL	0 aa/1 nuc diff	Complete	Partial	??????	??????
	008443	B615-1'CL	0 aa/3 nuc diff	Complete	Partial	??????	??????
AV1S?	008479 !	14.12'CL	Closest to AV1S2 (9 aa/15 nuc diff)	Partial	Partial	??????	??????
	008480 !	14.16'CL	Closest to AV1S2 (9 aa/15 nuc diff)	Partial	Partial	??????	??????
	008456 !	BDC2.5'CL	Closest to AV1S5 (7 aa/12 nuc diff)	Complete	None	??????	??????
	021464 !	Va1-NEW'CL	Closest to AV1S6 (14 aa/35 nuc diff)	Partial	None	??????	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV2S1	008430 !	TA39'CL	In Paper.	Complete	Complete	X02929	??????
AV2S2	008428 !	Va2.2(pFGB)'CL	In Paper.	Complete	None		??????
AV2S3	008421 !	PL23.1'CL	In Paper.	Complete-2	Complete	M21857	??????
	008422	PL51.1.1'CL	Identical to 8421	Complete-2	Complete	??????	??????
	008424	PL127.8'CL	Identical to 8421	Complete-2	Complete	??????	??????
	008425	PL172.10'CL	Identical to 8421	Complete-2	Complete	??????	??????
	008426	PL183.2'CL	Identical to 8421	Complete-2	Complete	??????	??????
	008423	PL83.12'CL	Identical to 8421	Complete-2	Complete	??????	??????
AV2S4	025908 !	P14A.2'CL (P14A.1 in paper (P14A.1 is AV18S1))	In Paper.	Complete	Complete	X06771	??????
AV2S5	008432 !	8I'CL	In Paper.	Complete-2	Complete		??????
	008433	4I'CL		Complete-2	Complete	??????	??????
	008434	Ts3-9a1		Complete-2	Complete	??????	??????
	019355	L-9.w7'CL		Complete-2	Partial	??????	??????
	021435	Tcra V2.1'CL		Complete-2	None	??????	??????
AV2S6	008427 X !	9C127'CL	In Paper.	Complete-2	Complete		009226
AV2S7	008429 !	KB5 (Va2-JaA10'CL)	In Paper.	Complete	Complete	M60999	??????
DV2S8	010137 δ	MOUSE Vd8'CL (41BNT-117 in paper)	In Paper.	Partial	None		??????
	019409 !	Va2.4'CL		Complete	None	??????	??????
AV2S9	??????	LD3	In Paper.	Complete	Complete	M34196	??????
	008431 X !	10I'CL		Complete	Complete	??????	009223
	021439	V2.5'CL		Complete	None	??????	??????
AV2S?	019411 !	Va2.6'CL	Closest to AV2S1 (4 aa/7 nuc)	Complete-1	None	??????	??????
	021436	Tcra V2.2'CL	Closest to AV2S2 (1 aa/1 nuc diff)	Complete	None	??????	??????
	008439	M14T Va2 sterile'CL	Closest to AV2S4 (2 aa/2 nuc diff). Pseudogene?	Partial	None	??????	??????
	021437	Tcra V2.3'CL	Closest to AV2S4 (2 aa/2 nuc diff). Pseudogene?	Complete-2	None	??????	??????
	021440	Tcra V2.6'CL	Closest to AV2S6 (1 aa/3 nuc diff)	Complete-2	None	??????	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	019410 !	Va2.5'CL	Closest to DV2S8 (3 aa/5 nuc diff)	Complete	None	??????	??????
	021438 !	Tcra V2.4'CL	Closest to DV2S8 (3 aa/5 nuc diff). Similar to 019357	Complete	None	??????	??????
	019357 !	L-9.10'CL	Closest to DV2S8 (3 aa/6 nuc diff). Similar to 021438 (1 nuc del)	Complete	Partial	??????	??????
	019359 !	H-10.B5'CL	Closest to DV2S8 (3 aa/7 nuc diff)	Complete	Partial	??????	??????
	019361 !	H-16.B6'CL	Closest to DV2S8 (2 aa/5 nuc diff)	Complete	Partial	??????	??????
AV3S1	??????	λ2.2	In Paper.	Complete	None	X02857	??????
	008399 !	pHDS58'CL	In Paper.	Complete	Complete		??????
AV3S2	008401 X !	MOUSE C9'CL	In Paper.	Complete	Complete	X05734	009154
	008404	K1-2'CL		Partial	Complete	??????	??????
	008405	KL'CL		Partial	Complete	??????	??????
	008403	K1-3'CL		Partial	Complete	??????	??????
AV3S3	008400 X !	8/10-2'CL	In Paper.	Complete	Complete	X05732	009138
AV3S4	008406 X !	AF.3.G7'CL	In Paper.	Complete	Complete	M16678	009153
AV3S5	008397 X !	AR-5'CL	In Paper.	Complete	Complete	M21202	009162
	008398	Ts3-8a		Complete	Complete	??????	??????
AV3S6	??????	P2111	In Paper.	Complete	None	M33586	??????
	008402 !	M14T		Complete	None	??????	??????
AV3S7	?????? !	LD1'CL	In Paper. (025412 matches name but is fragment)	Complete	Partial	M34194	??????
AV3S8	??????	PJF1A	In Paper.	Complete	None	M76612	??????
	008407 !	Va3.2'CL		Complete-1	None	??????	??????
AV4S1	008409 !	TA65'CL	In Paper.	Complete	Complete	X02932	??????
	008411	50.1 ALPHA'CL	0 aa/1 nuc diff - Explained	Complete	Complete	??????	??????
	008410	PL214.12'CL	1 aa/2 nuc diff (CT to TC swap)	Complete	Complete	??????	??????
AV4S2	015630 P	MD13	In Paper.	Complete	Joining Errors	X02969	??????
	008414 !	TA28'CL	In Paper.	Partial	Complete	X02931	??????
	??????	MA16	In Paper.	????????	????????		??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV4S3	008419 !	112-2 GENOMIC'CL	In Paper.	Complete	None	X05187?	??????
	008578	112-2'CL	In Paper.	Partial	Complete	X05187?	??????
	??????	AP11.2'CL	In Paper.	????????	????????		??????
AV4S4	008417	Va4.4'CL	In Paper.	Complete	None	M38678	??????
	027522 !	Va4.4-Ja24'CL	0 aa/1 nuc diff	Complete	Complete	??????	??????
AV4S5	008486	BDFLAII (BDFLII in paper)	In Paper. Pseudogene	Complete	Joining Errors	X03669	??????
AV4S6	008420 X !	DA.33.C2'CL	In Paper.	Complete	Complete	M16675	009085
AV4S7	008416 X !	BB1.D5'CL	In Paper.	Complete	Complete	M16676	009160
DV4S8	010146 δ	δ2.3	In Paper.	Partial	Complete	X13316	??????
AV4S9	008412	PJR-25'CL	In Paper.	Complete	Partial	M21205	??????
	?????? X !	MT1-14	In Paper.	Complete	Complete	M34198	M34199
AV4S10	008418 !	M14T allele 1'CL (A26 in paper)	In Paper. 1 extra aa	Complete	Complete	M22660	??????
AV4S11	008413 X !	F5'CL	In Paper. Many aa/many nuc diff - Unexplained: X14387 has correct sequence	Complete	Complete	X14387	009116
AV4S12	?????? !	HL228	In Paper.	Complete	Complete	M61133	??????
AV4S?	008485 !	YLA6'CL	Closest to AV4S11 (9 aa/25 nuc diff)	Complete	None	??????	??????
	008415 !	8.2'CL	Closest to AV4S9 (2 aa/4 nuc diff)	Partial	Complete	??????	??????
AV5S1	008446 !	TA72'CL	In Paper.	Complete	Complete	X02933	??????
AV5S2	008484 !	MDA'CL	In Paper.	Complete	Complete	X02967	??????
	015627 P	TA80	In Paper. 2 aa/6 nuc diff - Unexplained	Partial	Joining Errors	X02939	??????
AV5S3	??????	MA25	In Paper.	????????	????????		??????
AV6S1	008494 !	TA1'CL	In Paper.	Complete-1	Complete	X02934	??????
DV6S2	025902 !	TCRAV6'CL (AV6 in paper)	In Paper.	Complete	Complete	M94080	??????
	010147 δ	Z68'CL	In Paper.	Partial	Partial	M37279	??????
ADV7S1	015628 P	TA27	In Paper.	Complete	Joining Errors	X02935	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	??????	R16.14	In Paper.	????????	????????		??????
	??????	Vδ6.3	In Paper.	????????	????????		??????
ADV7S2	??????	A38	In Paper.	????????	????????		??????
	015637 P	Va7.2 (Vα7.2 in paper)	In Paper.	Complete	None	M38679	??????
	008435 !	TCDa3’CL	In Paper.	Complete	Complete	M37597	??????
	??????	R16.8	In Paper.	????????	????????		??????
	015722 P	M23’CL	In Paper. 0 aa/1 nuc diff - Explained	Complete	Joining Errors		??????
	010135 δ	Y93A’CL	In Paper. 0 aa/1 nuc diff - Explained	Complete	Complete	M26447	??????
	010133 δ	TCDd2’CL		Complete	Partial	??????	??????
	008436	VCa3’CL		Complete	Partial	??????	??????
DV7S3	015718 P	2B4.Exp	In Paper.	Complete	Joining Errors		??????
DV7S4	015725 P	Z49	In Paper.	Complete	Joining Errors	M37287	??????
DV7S5	010132 δ	Z53’CL	In Paper.	Complete	Partial	M37285	??????
DV7S6	010134 δ	TCDδ1	In Paper.	Complete	Partial	M37600	??????
	010136 δ	T195/BW	In Paper.	Complete	Complete	M26448	??????
AV8S1	008394 !	TA61’CL	In Paper.	Partial	Complete	X02936	??????
AV8S2	008386 !	P71’CL	In Paper.	Complete	Complete	X02970	??????
	015393 P	BVI/5.a2	In Paper.	Complete	Joining Errors	M31649	??????
	008388	Cw3/1.1A’CL	0 aa/1 nuc diff. 1 extra aa	Complete	Complete	??????	??????
AV8S3	008381	F3.4’CL	In Paper.	Complete	None	X06306	??????
	008379	3F9-ALPHA7’CL (3F9 in paper)	In Paper. 0 aa/1 nuc diff. - Explained. 1 extra aa	Complete	Complete	M15063	??????
	008380 !	M14T allele 2’CL	1 extra aa	Complete	Complete	??????	??????
	008393 X	LB2’CL	1 aa/1 nuc diff (at N terminus)	Partial	Complete	??????	009156
AV8S4	008390 !	Va8.4’CL	In Paper	Complete	None	M38680	??????
AV8S5	008383	F3.2’CL	In Paper.	Complete	None	X04332	??????
	008382 X !	5/10-20K’CL		Complete	Complete	??????	009150
AV8S6	008387 !	F3.3’CL	In Paper.	Complete	None	X06305	??????
AV8S7	008389 !	F3.5’CL	In Paper.	Complete	None	X06307	??????
AV8S8	008391	F3.6’CL	In Paper.	Complete	None	X06308	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	021351 X !	N15'CL		Complete	Complete	??????	021353
AV8S9	008385 X !	ZZ38'CL	In Paper.	Complete	Complete	M16677	009120
AV8S10	?????? !	p114	In Paper. Pseudogene	Complete	None	X17181	??????
AV8S11	?????? !	p011	In Paper. Pseudogene	Complete	None	X17173	??????
AV8S12	?????? !	MT1-33	In Paper.	Complete	Complete	M34206	??????
AV8S13	?????? ! 008392	MT1-7 4C1'CL	In Paper.	Complete Partial	Complete Complete	M34208 ??????	?????? ??????
AV8S14	?????? !	P1F12C4	In Paper.	Complete	Complete	M34210	??????
AV8S15	015672 P	GHY-ALPHA-2 (HY-A2 in paper)	In Paper.	Complete	Joining Errors	X60320	??????
AV8S?	008384 ! 024704 ! 008395 ! 008396	F3-20'CL CAS 20'CL M14T-6'CL M14T-7'CL	Closest to AV8S5 (1 aa/1 nuc diff) Closest to AV8S5 (2 aa/4 nuc diff) Closest to AV8S8 (1 aa/2 nuc diff) Closest to AV8S8 (1 aa/2 nuc diff)	Complete Complete Partial Partial	Complete Complete Complete Complete	?????? ?????? ?????? ??????	?????? ?????? ?????? ??????
AV9S1	??????	BM1037	In Paper.	????????	????????		??????
AV9S2	008489 !	GHY-ALPHA-1'CL(HY-A1 in paper)	In Paper.	Complete	Complete	X60319	??????
AV9S?	019366 !	H-12.C4'CL	Closest to AV9S2 (2 aa/3 nuc diff)	Complete	Partial	??????	??????
AV10S1	008473 ! 008475	Va10-Ja26'CL FN1-18'CL	In Paper. 1 aa/2 nuc diff (At N terminus)	Complete Partial	Complete Complete	M38102 ??????	?????? ??????
AV10S2	008472 X !	D6'CL	In Paper. 1 extra aa	Complete	Complete	M20875	009109
AV10S3	008474 X !	1F8'CL	In Paper. 0 aa/2 nuc diff - Unexplained: M20876 has correct sequence. 1 extra aa	Complete	Complete	M20876	009110
AV10S4	??????	P022	In Paper.	????????	????????		??????
AV10S5	?????? !	P102	In Paper.	Complete	None	X17176	??????
AV10S6	?????? 019364 !	P109s H-16.B11'CL	In Paper.	Complete Complete	None Partial	X17178 ??????	?????? ??????
DV10S7	010145 δ	KN25-D4'CL	In Paper. 1 extra aa	Complete	Partial	M26299	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV10S8	?????? !	P1D3A6'CL	In Paper. (025411 matches name but is fragment)	Complete	Complete	M34212	??????
AV10S9	?????? !	7/6AH1'CL	In Paper. (025415 matches name but is fragment)	Partial	Complete	M34216	??????
AV10S?	019362 !	L-9.2'CL	Closest to AV10S1 (10 aa/15 nuc diff)	Complete	Partial	??????	??????
AV11S1	008464 X	B10'CL	In Paper. 1 extra aa	Complete	Complete	X03860	009123
	008462 X !	C.F6'CL	1 extra aa	Complete	Complete	??????	009100
	008463 X	5C.C7'CL	1 extra aa	Complete	Complete	??????	009101
	008465 X	4.C3'CL	1 extra aa	Complete	Complete	??????	009124
AV11S2	008466 !	MOUSE 2B4'CL	In Paper. 1 extra aa	Complete	Complete	X02968	??????
AV11S3	008468 !	Va11-Ja39'CL(Vα11.3 in paper)	In Paper. 1 extra aa	Complete	Complete	M22603	??????
AV11S4	008467 !	NA3'CL	In Paper. 1 extra aa	Complete	Complete	M27351	??????
ADV11S5	008471	11.1(a)'CL (Vα11.1a in paper)	In Paper.	Complete	None	M73263	??????
	010153 δ	dG8'CL	In Paper.	Partial	Complete	X14095	??????
	025197 !	Va11.1-Ja17'CL		Complete	Complete	??????	??????
AV11S6	008470 !	11.1(d)'CL (Vα11.1d in paper)	In Paper. 1 extra aa	Complete	None	M73264	??????
AV11S7	008469 !	11.3(a)'CL (Vα11.3a in paper)	In Paper. 1 extra aa	Complete	None	M73265	??????
AV11S?	008490 !	Bm2T.3.1'CL	Closest to AV11S2 (Half is homologous, half isn't)	Complete	Complete	M87848	??????
	008491 !	BO4H.9.1'CL	Closest to AV11S3 (1 aa/7 nuc diff)	Complete	Complete	??????	??????
	026331 !	231F1'CL	Closest to AV11S3 (2 aa/8 nuc diff). 1 extra aa	Complete	Complete	??????	??????
AV12S1	008408 !	BDFLAI'CL (BDFLI in paper)	In Paper.	Complete	Complete	X03668	??????
AV13S1	008437 !	Va13.1'CL	In Paper.	Complete	None	M38681	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV13S2	008438 !	MOUSE Va13.2'CL	In Paper. 0 aa/1 nuc diff - Unexplained	Partial	Complete		??????
AV14S1	008477	14.1'CL (Va14.1 in paper)	In Paper.	Complete-2	None	D90229	??????
	008476 !	MOUSE T-s'CL		Complete	Complete	??????	??????
AV14S2	008478	14.2'CL (Va14.2 in paper)	In Paper.	Complete	None	D90230	??????
	025879 !	MTs79.1	2 extra aa	Complete	Complete	??????	??????
AV15S1	008448 !	SJL-HE-1.1'CL	In Paper.	Complete	Complete	X57397	??????
	008447	T2.5-5'CL	In Paper. 1 aa/1 nuc missing - Unexplained	Complete	Complete		??????
	008482	BDC4.12'CL		Partial	None	??????	??????
AV15S?	008689 !	57.7'CL	Closest to AV15S1 (4 aa/9 nuc diff)	Partial	Complete	??????	??????
	023479 !	RF33.21-Va'CL	Closest to AV15S1 (1 aa/2 nuc diff)	Complete	Complete	??????	??????
AV16S1P	015641 P	MOUSE 58a-b-BWaB (BW.B in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	Joining Errors	X51643	??????
AV16S2P	025900 P	AV16.1	In Paper. Many aa/0 nuc diff - Differently translated	Complete	Joining Errors	M94080	??????
AV16S?	008440 !	C10'CL	Closest to AV16S1P (2 aa/5 nuc diff). Functional	Partial	Complete	??????	??????
AV17S1	008492 !	42H11'CL	In Paper. 1 extra aa	Complete	Complete	M16118	??????
ADV17S2	008493 !	5.3.18'CL	In Paper. 0 aa/1 nuc diff - Explained: M16119 has correct sequence. 1 extra aa	Complete	Complete	M16119	??????
	??????	PCDS81	In Paper.	Complete	None	X17226	??????
AV17S3	??????	BM.B	In Paper.	????????	????????		??????
AV17S?	019421 !	NY-CTL'CL	Closest to AV17S2 (9 aa/13 nuc diff)	Complete	None	??????	??????
	022133 !	A2G10'CL	Closest to AV17S2 (5 aa/8 nuc diff)	Complete	Partial	??????	??????
	019368 !	H-15.E3'CL	Closest to AV17S1 (3 aa/4 nuc diff)	Complete	Partial	??????	??????
AV18S1	025906 P	P14A.1	In Paper. (See also AV2S4)	Partial	Joining Errors	X06773	??????
AV18S2	??????	BM.A	In Paper.	????????	????????		??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV18S?	008487 !	17.2’CL (17.A2)	Closest to AV18S2 (17 aa/35 nuc diff)	Complete	Complete	??????	??????
	008488 !	23.32’CL (23.A1)	Closest to AV18S2 (17 aa/35 nuc diff)	Complete	Complete	??????	??????
AV19S1	008461 !	VaA10-JaA10’CL	In Paper.	Complete	Complete	M22604	??????
	008481	9.4’CL	1 aa/1 nuc missing.	Partial	Partial	??????	??????
AV20S1	008483 !	Va5T’CL	In Paper.	Complete	None		??????
DV101S1	010150	717.7D’CL (7-17.1 in paper)	In Paper.	Complete	Complete	M23545	??????
	010148	M21	In Paper.	Complete	Partial	X63934	??????
	010149	d7.1 (δ7.1 in paper)	In Paper.	Complete	Complete	X13314	??????
	010151	T245/BW	In Paper. 1 aa/1 nuc diff - Unexplained: M26449 has correct sequence	Complete	Complete	M26449	??????
	010152	M16’CL		Partial	Partial	??????	??????
DV102S1	025905	TCRDV2’CL (DV2 in paper)	In Paper.	Complete	None	M94080	??????
	015721 P	M11	In Paper.	Complete	Joining Errors		??????
DV104S1	010130	Z10	In Paper.	Complete	Partial	M32780	??????
	010131	KN12-D1	In Paper.	Complete	Partial		??????
	010129	DN-4’CL		Complete	Complete	??????	??????
DV105S1	010143	Z72	In Paper.	Partial	Partial	M37282	??????
	??????	*No name*	In Paper.	Complete	None	M64239	??????
	010140	MOUSE Vd5 GERMLINE’CL (glVδ5 in paper)	In Paper. 0 aa/1 nuc diff - Explained	Complete	None	M23382	??????
	010139	MOUSE d7.3’CL (glδ7.3 in paper)	In Paper. 0 aa/1 nuc diff - Explained	Complete	Complete	M23095	??????
	010144	Z35’CL		Partial	Partial	??????	??????
	010138	NYD4’CL		Complete	Complete	??????	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	010142	Z44'CL	1 aa/1 nuc diff	Partial	Partial	??????	??????

Table B.2: The table shows the Kabat entry IDs of members of the officially designated classes for mouse α and δ chains [126]. A group of six question marks indicates data which is not known or not available. In the Kabat ID column there are also flags indicating whether the sequence has a known pair (X), whether it was used in the analysis (!), and if it is defined as a psuedogene in the Kabat database (P). Where sequences did not match the sequences or names of members of any of the existing classes, the closest matching class was identified and a temporary class created with a name ending in a question mark. For example if a sequence was similar to AV5S1 the temporary class would be called AV5S?.

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV1S1	008117 !	HAP10'CL	In Paper	Complete	Complete	X04949/M13722	??????
	008265	HAP60'CL	In Paper	Partial	None	M13723	??????
	026043	U8'CL	2 aa/6 nuc diff (Nter)	Partial	Complete	??????	??????
AV1S2A1N1T	008107	PY14'CL	In Paper	Complete	Complete	M12423	??????
	008110	HAVT18'CL	In Paper	Complete	Complete	M27368	??????
AV1S2A1N2T	??????	Vα1.2	In Paper	Complete	Complete	X02592/M12959	??????
	008255	AA27'CL	In Paper	Partial	Complete	M17666	??????
	008109	AB22'CL	In Paper	Complete	Complete	M17646	??????
AV1S2A1N3T	??????	UBα14/4	In Paper	Complete	Complete	X63455	??????
	008111 X !	HA1.7'CL		Complete	Complete	??????	008921
AV1S2A2T	008108 !	pJM3E11'CL	In Paper. 1 aa/1 nuc diff - Unexplained: M12959 has correct seq	Complete	Complete	M12959	??????
AV1S2A3T	008112	WADM31F'CL	In Paper.	Complete	None	D13077	??????
	008113 !	AB18'CL	In Paper. 1 aa/2 nuc diff - Explained	Complete	Complete	M17647	??????
AV1S2A4T	015395 P !	AP511	In Paper	Partial	Joining Error	M17665	??????
AV1S2A5T	008251 !	pHaT'CL (pHαT3 in paper)	In Paper. 10 extra aa	Partial	Complete	K02777	??????
AV1S3A1T	008118 !	PY14.2'CL	In Paper.	Complete	None	X02850	??????
	??????	A87'CL	In Paper.	Partial	Complete	J03597	??????
	008256	AT5B1'CL		Partial	Partial	??????	??????
	008260	AT5B1'CL		Partial	Complete	??????	??????
AV1S3A2T	025946 X !	WM (WM-3'CL in paper)	In Paper	Complete	Complete	M86361	025948
	008231	WADM07B'CL	In Paper	Complete	None	D13070	??????
	008119	AA17'CL	In Paper. 12 aa/3 nuc diffs - 11 aa Explained	Complete	Complete	M17649	??????
	??????	*No name*	In Paper	Complete	Complete	Z26593	??????
	??????	40	In Paper	Partial	Complete	M87871	??????
	??????	28	In Paper	Partial	Complete	M87869	??????
AV1S4A1N1T	008116 !	R10'CL	In Paper	Complete	Complete	M35617	??????
	008115	IGRa08'CL	In Paper	Complete	None	X58769	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV1S4A1N2T	008258	AE11'CL	In Paper	Partial	Complete	M17668	??????
	008230	Va1.n1'CL	In Paper. 1 aa/4 nuc diff - Explained. 0 aa/2 nuc diff - Unexplained	Complete	None	L06885	??????
AV1S5	008114 !	AE24A'CL	In Paper	Complete	Complete	M17650	??????
AV1S?	008247 !	AB17'CL	Closest to AV1S2A1N2T (5 aa/8 nuc diff (1 del res))	Partial	Complete	??????	??????
	008253 !	AB28'CL	Closest to AV1S2A1N2T (1 aa/1 nuc diff)	Partial	Complete	??????	??????
AV2S1A1	008244 !	HAVT06'CL	In Paper	Partial	Complete	M27369	??????
	??????	AV2S1*01	In Paper	????????	????????		??????
AV2S1A2	?????? ! X	8B3	In Paper	Complete	Complete	M81774	M81773
	008144	IGa09'CL	In Paper	Complete-2	None	X58746	??????
	008142	AF110'CL	In Paper	Partial	Complete	M17652	??????
	008143	AC112'CL	In Paper	Partial	Complete	M17653	??????
	008145	AV2S1'CL (AV2S1*02/03 in paper)	In Paper	Partial	None	L11159	??????
AV2S1A3T	008242	HAP26'CL	In Paper	Partial	Complete	X04940/M13724	??????
	008243 !	HAP71'CL	In Paper	Partial	Complete	X04946/M13725	??????
	??????	A13ct7	In Paper	Partial	None	S60781	??????
AV2S2A1T	008128	AG110'CL	In Paper	Complete	Complete	M17655	??????
	008245	AD17'CL	In Paper	Partial	Complete	M17669	??????
	008127	AA13'CL	In Paper	Complete	Complete	M17654	??????
	??????	pV.1	In Paper	Complete	None	X06193	??????
	008130 !	p24.1'CL	In Paper	Complete	Complete	X06192	??????
AV2S2A2T	008129 !	AC17'CL	In Paper	Complete	Complete	M17656	??????
AV2S3A1T	008131 !	AA25'CL	In Paper	Complete	Complete	M17657	??????
AV2S3A2T	008204	WADM31J'CL	In Paper	Complete	None	D13078	??????
	008290 X !	S14.50'CL		Partial	Complete	??????	008889
AV2S?	023270 X !	XPZ10'CL	Closest to AV2S3A2T (3 aa/3 nuc diffs)	Partial	Complete	??????	023276

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV3S1	008152	HAP05'CL	In Paper.	Complete	Complete	X04948/M13726	??????
	008151	HAP44'CL	In Paper.	Complete	Complete	X04955/M13727	??????
	027015 X !	18039 D6'CL		Complete	Complete	??????	027023
	027016	18039 G14'CL		Complete	Complete	??????	??????
AV3S?	026048 !	P9'CL	Closest to AV3S1 (1 aa/3 nuc diff)	Partial	Complete	??????	??????
	008207	3.1'CL	Closest to AV3S1 (0 aa/2 nuc diff)	Complete	Complete	??????	??????
AV4S1	008181 !	HAP08'CL	In Paper.	Complete	Complete	X04937/M13728	??????
	??????	DD1	In Paper.	Complete	Complete	L26451	??????
	??????	AS1	In Paper.	Complete	Complete	L29035	??????
AV4S2A1T	008176 !	HAVT01'CL	In Paper.	Complete	Complete	M27372	??????
	008175	HAVT27'CL	In Paper.	Complete	Complete	M27371	??????
AV4S2A2T	008177 !	HAVT33'CL	In Paper.	Complete	Complete	M27370	??????
	008179	A-55'CL	In Paper.	Complete	Complete	M18460	??????
AV4S2A3T	008178 !	Va4.n1'CL	In Paper.	Complete	None	L06886	??????
AV4S?	008377 !	AV4S1'CL	Closest to AV4S1 (2 aa/2 nuc diffs)	Partial	None	??????	??????
AV5S1	008153	IGRa10'CL	In Paper.	Complete	None	X58747	??????
	008246 !	HAP35'CL	In Paper.	Partial	Complete	X04953/M13729	??????
	008226	WADM06D'CL	In Paper.	Complete	None	D13069	??????
AV5S?	023271 !	XPD25'CL	Closest to AV5S1 (2 aa/3 nuc diff)	Partial	Complete	??????	??????
ADV6S1A1N1	?????? !	GERM V	In Paper.	Complete	None	M21626	??????
	008238	TCRAV06S1*01'CL	In Paper. 14 aa/1 nuc diff - Unexplained (G inserted). Doesn't match class (2 real nuc diffs)	Partial	None	L09757	??????
ADV6S1A1N2T	009773	DS6'CL δ	In Paper.	Complete	Complete	M21624	??????
	009966	KT06A'CL δ	In Paper. 22 aa/1 nuc diff - Explained (C inserted)	Partial	Complete		??????
AV6S1A2N1	008239	TCRAV06S1*02'CL	In Paper. 14 aa/1 nuc diff - Unexplained (G inserted)	Partial	None	L09758	??????
AV6S1A2N2	?????? !	*No name*	In Paper.	Complete	Complete	Z14996	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008237	Vaw33.n1 (Vαw31n in paper)	In Paper.	Complete	None	L06884	??????
	008240	TCRAV06S1*03'CL	In Paper. 14 aa/1 nuc diff - Unexplained (G inserted)	Partial	None	L10122	??????
	008172	HAP01'CL	In Paper. Many diffs to paper - Explained (but 027014 has same seq. Why?)	Complete	Complete	M13730	??????
	??????	VA6.2	In Paper.	Complete	None	S51029	??????
	??????	17A2	In Paper.	Partial	Complete	M87870	??????
AV6S?	008238	TCRAV06S1*01'CL	In class ADV6S1A1N1 in paper but doesn't match it. Closest to AV6S1A2N2 but 1 nuc diff	Partial	None	L09757	??????
	027014 X !	18030 B31'CL	Same as 8172!	Complete	Complete	??????	027022
AV7S1A1	??????	*No name*	In Paper.	Complete	Complete	D21847	??????
	?????? P	$\lambda S\alpha 9$	In Paper.	Complete	Joining Error	M12070	??????
	008139 !	XS9'CL	In Paper.	Complete	Complete	M16746	??????
	015394 P	SUP-T1	In Paper.	Complete	Joining Error	X03751	??????
	??????	AV7S1*01/03/04	In Paper.	????????	????????		??????
	026155	TcHST2'CL		Complete	Complete	??????	??????
AV7S1A2	008138 !	HAP21'CL	In Paper.	Complete	Complete	X04939/M13731	??????
	008213	AV7S1'CL (AV7S1*02/05 in paper)	In Paper.	Partial	None	L11161	??????
	026045	US3'CL		Partial	Complete	??????	??????
AV7S2	008257	HAP12'CL	In Paper.	Partial	None	X04938/M13732	??????
	008134 !	IGRa11'CL	In Paper.	Complete	None	X58744	??????
AV7S?	025436 !	CF8'CL	Closest to AV7S2 (2 aa/2 nuc diffs)	Complete	Complete	??????	??????
AV8S1A1	008146 !	HAP41'CL	In Paper.	Complete	Complete	X04954/M13733	??????
	008147	HAP17'CL	In Paper.	Complete	Complete	X04950/M13734	??????
	008148	HAP49'CL	In Paper.	Complete	Complete	X04944/M13735	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	??????	AV8S1*01/03/04	In Paper. 1 aa/1 nuc diff - Unexplained	Partial	None	L11162	??????
	??????	AV8S1*1	In Paper.	????????	????????		??????
	008378	AV8S1'CL	1 aa/1 nuc diff	Partial	None	??????	??????
AV8S1A2	008150	WADM35A'CL	In Paper.	Complete	None	D13079	??????
	??????	AV8S1*02	In Paper.	????????	????????		??????
	??????	AV8S1*2	In Paper.	Complete	None	M99570	??????
	008149 X !	S14.4'CL	In Paper. 1 aa/1 nuc diff - Unexplained. Same as AV8S1A1	Complete	Complete	M97714	008869
	023272 !	XPR04'CL	1 aa/1 nuc diff	Complete	Complete	??????	??????
AV8S2A1N1T	008164 !	HAP50'CL	In Paper.	Complete	Complete	X04956/M13736	??????
	008165	HAVT24'CL	In Paper.	Complete	Complete	M27373	??????
	008167	8.2'CL		Complete	Complete	??????	??????
	023273	XPZ10-V8S2J11'CL		Complete	Complete	??????	??????
AV8S2A1N2T	008166	AG212'CL	In Paper.	Complete	Complete	M17658	??????
AV9S1	008122 !	HAP36'CL	In Paper.	Complete	Complete	X04942/M13737	??????
	?????? P	*No name*	In Paper.	Partial	Joining Error	M90479	??????
	023274	XPF10-V9S1J53'CL		Complete	Complete	??????	??????
AV10S1A1	008206 !	TCRAV10S2*01'CL	In Paper	Complete	None	L09760	??????
	??????	22DAG	In Paper	Partial	Complete	M87868	??????
AV10S1A2	008211 !	WADM22B	In Paper	Complete	None	D13075	??????
	008205	TCRAV10S2*02'CL	In Paper	Complete	None	L09759	??????
AV10S1A3T	008154 !	HAP58'CL	In Paper	Complete	Complete	X04957/M13738	??????
AV11S1A1T	008160	HAP02'CL	In Paper	Complete	Complete	X04936/M13739	??????
	008281 P	HAP28'CL	In Paper. Lot missing res within sequence	Fragment	Partial	X04952/M13740	??????
	008162	HAP29'CL	In Paper	Complete	Complete	X04941/M13741	??????
	008161	HAP32'CL	In Paper	Complete	Complete	M13742	??????
	023269 X !	XPE15'CL	2 extra aa	Complete	Complete	??????	023275
AV11S1A2T	008163 !	AB19'CL	In Paper	Complete-1	Complete	M17659	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV11S?	008196	Va11.n1'CL	N terminus looks wrong (3 nuc repeat+4 aa/7 nuc diff)	Complete	None	??????	??????
AV12S1	??????	HTA73	In Paper	Complete	None	X70310	??????
	?????? δ	pGA-5	In Paper	Complete	Partial	X01403	??????
	008254	AC25'CL	In Paper	Partial	Complete	M17670	??????
	??????	64.8P	In Paper	Partial	Complete	M87866	??????
	??????	E4	In Paper	Complete	Complete	Z46641	??????
	008123 X !	HBP-MLT'CL		Complete	Complete	??????	008988
AV13S1	008200 X !	S26.2'CL	In Paper	Complete	Complete	M97722	008904
	008197	HAVT15'CL	In Paper	Complete	Complete	M27374	??????
	008198 X	AL62.24'CL		Complete	Complete	??????	008959
	008199 X	S25.13'CL		Complete	Complete	??????	008891
ADV14S1	025191 δ	DV8-E6'CL (E6 in paper)	In Paper	Complete	Partial	Z46644	??????
	008233	HUMAN HAVT20'CL (HAVT20 in paper)	In Paper (see also 8235)	Complete	Complete	M27375	??????
	008235	HAVT20'CL	In Paper (see also 8233). 10 aa/2 nuc diffs - Unexplained (Bad alignment for 9 aa diffs)	Complete	None	M27375	??????
	?????? !	115E15	In Paper	Complete	Complete	Z29614	??????
AV14S2A1T	025886 !	Va14.2-AL4.1'CL	In Paper	Complete	Complete	M64355	??????
	008234	14.1'CL (Va14.1 in paper)	In Paper. 11 aa/5 nuc diffs - Partially explained	Complete	Partial	X58157	??????
AV14S2A2N1T	008241 !	WADM20F'CL	In Paper	Complete	None	D13074	??????
AV14S2A2N2T	008236	Va14.n1'CL (Va14.n in paper)	In Paper.	Complete	None	L06880	??????
AV14S2A3T	?????? !	*No name*	In Paper	Complete	Complete	M95394	??????
AV14S?	008124 !	14.2'CL	Closest to ADV14S1 (1 aa/2 nuc diff. 8235 has same aa diff)	Partial	None	??????	??????
	026052	U39'CL	Ambiguous - matches all AV14S2 classes	Partial	Complete	??????	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV15S1	??????	*No name*	In Paper	Complete	Complete	S60795	??????
	008170	HAVT31'CL	In Paper	Complete	Complete	M27376	??????
	008249	Pt'CL (λP13 in paper)	In Paper. 1 aa/2 nuc diff - Explained	Complete	Complete	X05771	??????
	??????	2DLE	In Paper	Partial	Complete	M90482	??????
	008295 X	DE49'CL	In Paper	Partial	Complete	Z22965	009066
	008171	AL8'CL	2 extra aa	Complete	Complete	??????	??????
	026746 X !	37'CL		Complete	Complete	??????	026751
	026747 X	43'CL		Complete	Complete	??????	026752
AV16S1A1T	008120 !	AG21'CL	In Paper.	Complete	Complete	M17651	??????
AV16S1A2PT	015396 P	HAVT32'CL	In Paper	Complete	Joining Errors	M27377	??????
ADV17S1A1T	008157 !	AB11'CL	In Paper.	Complete-1	Complete	M17660	??????
	??????	K15A	In Paper.	Partial	None	M22936	??????
	??????	KT05E	In Paper.	????????	????????		??????
	??????	AV17.1a	In Paper.	Partial	None	D13071	??????
	008248	Va131.1'CL	Also matches ADV17S1A2N1T	Partial	None	??????	??????
AV17S1A2N1T	008220 X	AL61.102'CL	In Paper.	Complete	Complete	M97704	008868
	008221 X !	S14.6'CL		Complete	Complete	??????	008890
	008248	Va131.1'CL	Also matches ADV17S1A1T	Partial	None	??????	??????
AV17S1A2N2T	??????	HTA61	In Paper.	Complete	None	X70309	??????
	008232 X	RFL3.8'CL	In Paper.	Complete	Complete	M77498	008953
AV18S1	008158 !	AB21'CL	In Paper	Complete-1	Complete	M17661	??????
AV19S1	008156	AC24'CL	In Paper	Complete-1	Complete	M17662	??????
	008201 X !	S30.10'CL	In Paper	Complete-1	Complete	M97724	008954
AV20S1	008180 !	AE212'CL	In Paper. 1 extra aa	Complete	Complete	M17663	??????
AV21S1A1N1	??????	Vα21a	In Paper	????????	????????		??????
	008168	AF211'CL	In Paper. 3 aa/6 nuc diff - Partially explained	Complete	Complete	M17664	??????
	008227	WADM24B'CL	In Paper.	Complete	None	D13076	??????
	008300 X	DE5'CL	In Paper.	Fragment	Complete	Z23047	009071
ADV21S1A1N2	??????	Vα21b	In Paper	????????	????????		??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008169 X !	L17Ti'CL	In Paper	Complete	Complete	M15565	008839
	009774 δ	KT08A'CL	In Paper	Complete	Complete	X14548/Y00793	??????
AV21S1A2PT	??????	Vα21c	In Paper	????????	????????		??????
AV21S?	026046	U64'CL	Closest to ADV21S1A1N2 (0 aa/1 nuc diff)	Partial	Complete	??????	??????
	026050	P26'CL	Closest to ADV21S1A1N2 (3 aa/6 nuc diff (1 del nuc causes 3 aa diff)).	Partial	Complete	??????	??????
AV22S1A1N1T	008121	IGRa12'CL	In Paper	Complete	None	X58745	??????
	008252	AC9'CL	In Paper. 4 aa/7 nuc diffs - Unexplained	Partial	Complete	M17671	??????
	008223 X !	AL61.270'CL	In Paper	Complete	Complete	M97706	008878
	008224 X	S14.107'CL		Complete	Complete	??????	008848
AV22S1A1N2T	008229	Va22.n1'CL	In Paper	Complete	None	L06881	??????
	026748	NP-7'CL		Complete	Complete	??????	??????
AV22S1A2N1T	008228 !	WADM13D'CL	In Paper	Complete-1	None	D13072	??????
AV22S1A2N2T	008225	Va22.n2'CL	In Paper	Complete-1	None	L06882	??????
AV22S?	023337 !	ICHV10'CL	Closest to AV22S1A1N1T (2 aa/3 nuc diffs)	Complete	Complete	??????	??????
	008222 !	L90'CL	Closest to AV22S1A1N1T (1 aa/1 nuc diff)	Complete	Complete	??????	??????
	026041 !	TcPUN'CL	Closest to AV22S1A1N1T (3 aa/7 nuc diff). V similar to 008252	Partial	Complete	??????	??????
AV23S1	008133 !	IGRa01'CL	In Paper	Complete	None	X58736	??????
	008259	AD210'CL	In Paper	Fragment	Complete		??????
AV23S?	008132 X !	UA-S2'CL	Closest to AV23S1 (1 aa/1 nuc diff)	Complete	Complete	??????	008829
	008214	J33'CL	Closest to AV23S1 (1 aa/1 nuc diff)	Complete	Complete	??????	??????
	027018	LWF A20,C8-1'CL	Closest to AV23S1 (1 aa/1 nuc diff)	Complete	Complete	??????	??????
	008215	K64'CL	Closest to AV23S1 (1 aa/2 nuc diff)	Complete	Complete	??????	??????
	008216	KSR2'CL	Closest to AV23S1 (2 aa/3 nuc diff)	Complete	Complete	??????	??????
	008217	L113'CL	Closest to AV23S1 (2 aa/2 nuc diff)	Complete	Complete	??????	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008218	K42'CL	Closest to AV23S1 (3 aa/4 nuc diff)	Complete	Complete	??????	??????
	008219	L97'CL	Closest to AV23S1 (3 aa/3 nuc diff)	Complete	Complete	??????	??????
AV24S1	008135	IGRa02'CL	In Paper	Complete	None	X58737	??????
	008250	Linv (λL3 in paper)	In Paper.	Partial	Complete	X05770	??????
	027019 !	LWF A20,C8-2'CL		Complete	Complete	??????	??????
AV25S1	008141 !	IGRa03'CL	In Paper	Complete	None	X58738	??????
AV26S1	008137	IGRa04'CL	In Paper. 1 aa/1 nuc diff - Unexplained: X58739 has correct sequence	Complete	None	X58739	??????
	008136 !	62.119'CL		Complete	Complete	??????	??????
AV27S1	008155	IGRa05'CL	In Paper. 1 aa/1 nuc diff - Unexplained	Complete	None	X58740	??????
	?????? !	H'CL	In Paper. 1 aa/1 nuc diff - Unexplained (same seq as 8155)	Partial	Complete	M23431	??????
AV28S1A1T	008140 !	IGRa06'CL	In Paper.	Complete-1	None	X58767	??????
AV28S1A2T	?????? !	IGRa15'CL	In Paper.	Complete	Complete	X61070	??????
DV28S1A3T	025190 δ	DV7-E2'CL (E2 in paper)	In Paper.	Complete-1	Partial	Z46643	??????
AV29S1A1T	008209 !	IGRa07'CL	In Paper.	Complete	None	X58768	??????
AV29S1A2T	008210 !	Vaw29.n1'CL (Vaw29.n in paper)	In Paper.	Complete??	None	L06883	??????
	??????	*No name*	In Paper	????????	????????		??????
AV29S1A3T	?????? !	MT-ALL	In Paper.	Complete	Complete	S63879	??????
	??????	E5	In Paper.	Complete	Partial	Z46642	??????
AV30S1A1T	008203	VA30'CL	In Paper.	Complete-1	None	X68696/S59345	??????
	??????	A152	In Paper.	????????	????????		??????
	??????	*No name*	In Paper	Complete	None	D16586	??????
	008202 X !	S14.11'CL	In Paper.	Complete-1	Complete	M97712	008903
	??????	139DRD	In Paper.	Partial	Complete	M87865	??????
	024068	A28-761'CL		Complete-1	Complete	??????	??????
AV30S1A2T	?????? !	A17ct3	In Paper.	Complete	None	S60789	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
AV30S1A3T	?????? !	HTA129	In Paper.	Complete	None	X70305	??????
AV31S1	??????	*No name*	In Paper	Complete	None	X73521	??????
	008195 !	VA31'CL	In Paper.	Partial	None	X68697/S59347	??????
	??????	*No name*	In Paper	Complete-2	None	X70306	??????
	??????	A32ct7	In Paper.	Complete-2	None	S60792/D16585	??????
AV32S1	025885 !	Va30-KT2'CL	In Paper.	Complete	Complete	M64350	??????
	008208	VA32'CL	In Paper.	Complete	None	X68698/S59349	??????
	008212	WADM15B'CL	In Paper.	Complete	None	D13075	??????
	??????	Vαw32	In Paper.	Complete	None	X70307	??????
	??????	NA20	In Paper.	Complete	Fragment	S50881	??????
DV101S1	009768	HUMAN Vd1'CL (K15A in paper)	In Paper.	Complete	None	M22198	??????
	009771	KT10E	In Paper.	Complete	Complete	X14545	??????
	009772	HUMAN GROUP O'CL (O-240/47 in paper)	In Paper. 1 aa/1 nuc diff - Unexplained: M18414 has correct sequence	Complete	Complete	M18414	??????
	??????	D105	In Paper.	Partial	Complete	X15021	??????
	009767	Pr81	In Paper.	Complete	Complete	X06557/Y00289	??????
	009769	HUMAN F6C7'CL		Complete	Complete	??????	010528
	009769	HUMAN IDP2'CL		Complete	Complete	??????	??????
DV102S1A1T	009785	TRDV2'CL (λLY67Vδ2 in paper)	In Paper.	Complete-1	None	X15207	??????
	009781	LB117'CL (LB117δ1.7 in paper)	In Paper. 1 extra aa	Complete	Complete	X13950	??????
	009783	LB207'CL (LB207δ in paper)	In Paper. 1 extra aa	Complete	Complete	X13952	??????
	009782	LB210'CL (LB210δ in paper)	In Paper. 1 extra aa	Complete	Complete	X13951	??????
	009784	KT19E'CL	In Paper. 1 extra aa	Complete	Complete	X14546	??????
DV102S1A2T	??????	X13	In Paper.	Complete	Complete	X53849	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	??????	PT11	In Paper.	Complete	Partial	M21784	??????
	009780	KT14E	In Paper. 1 extra aa	Complete	Complete	X14547	??????
	??????	*No name*	In Paper.	Complete	Complete	X72501	??????
	009778	G6'CL	1 extra aa	Complete	Complete	??????	??????
	009779	AB12'CL	1 extra aa	Complete	Complete	??????	010527
	009964	VTC'CL	1 extra aa	Partial	Complete	??????	??????
DV103S1A1T	009775	Vd2'CL (λP11 in paper)	In Paper.	Complete	None	M23326	??????
	009776	WM14'CL	In Paper.	Complete	Complete	X13954	??????
	009777	KT041	In Paper.	Complete	Complete	X14544	??????
	??????	HCδ4	In Paper.	????????	????????	M94081	??????
DV103S1A2T	??????	GLVδ3	In Paper.	Complete	Fragment	X15261	??????

Table B.3: The table shows the Kabat entry IDs of members of the officially designated classes for human α and δ chains [125]. A group of six question marks indicates data which is not known or not available. In the Kabat ID column there are also flags indicating whether the sequence has a known pair (X), whether it was used in the analysis (!), and if it is defined as a psuedogene in the Kabat database (P). Where sequences did not match the sequences or names of members of any of the existing classes, the closest matching class was identified and a temporary class created with a name ending in a question mark. For example if a sequence was similar to AV5S1 the temporary class would be called AV5S?.

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV1S1A1N1	??????	H18	In Paper	Complete	None	L36092	??????
	008866	HBVT73'CL	In Paper	Complete	Complete	M27381	??????
	008860	PL5.2'CL	In Paper	Complete	Complete	M13836	??????
	008861	PL5.6'CL	In Paper	Complete	Complete	M13837	??????
	008862	PL6.1'CL	In Paper	Complete	Complete	M13838	??????
	008863	PL6.4'CL	In Paper	Complete	Complete	M13839	??????
	008864	SUP-T1'CL	In Paper	Complete	Complete	M16834	??????
	??????	308G	In Paper	Complete-3	None	M27912	??????
	??????	Vβ1.2	In Paper	Complete	Complete	X74841	??????
	026751 X !	37'CL		Complete	Complete	??????	026746
	026752 X	43'CL		Complete	Complete	??????	026747
BV1S1A1N2T	008865	HBVT96'CL	In Paper	Complete	Complete	M27380	??????
BV1S1A2	008867 !	VB1 VARIANT'CL (308C in paper)	In Paper. 1 aa/1 nuc diff - Unexplained: M27904 has correct sequence	Complete	None	M27904	??????
BV2S1A1	008887	PUCM4-4 (C-BETA-2)'CL	In Paper.	Complete	Complete	M12886	??????
	008886	PL6.21'CL	In Paper.	Complete	Complete	M13842	??????
	008885	PL2.13'CL	In Paper.	Complete	Complete	M13840	??????
	??????	MT1-1G	In Paper.	Complete	None	M11955	??????
	008898	ph7'CL	In Paper. 2 aa/3 nuc diffs - Unexplained. Matches BV2S1A3N2T	Partial	Complete	M15222	??????
	008888	HT1.9'CL	In Paper.	Partial	None	X57603	??????
	??????	BV2.1a	In Paper.	Complete	None	D13082	??????
	??????	B16	In Paper.	Complete	Partial	S50221	??????
	008890 X !	S14.6'CL	In Paper.	Complete	Complete	M97719	008221
	??????	C215	In Paper.	Complete	None	L36092	??????
	008889 X	S14.50'CL		Complete	Complete	??????	008290
BV2S1A2	008893	WBDP25G'CL	In Paper.	Complete	None	D13087	??????
	008892	TCRBV2S1*2'CL	In Paper.	Complete	None	X72719	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008891 X !	S25.13'CL	In Paper.	Complete	Complete	M97721	008199
BV2S1A3N1	008896	HT120'CL	In Paper.	Complete	None	X57604	??????
BV2S1A3N2T	008895 !	ph34'CL	In Paper.	Complete	Complete	M14263	??????
	008898	ph7'CL	Classed as BV2S1A1 in paper. Matches this class	Partial	Complete	??????	??????
BV2S1A3N3T	008897	WBDM11D'CL	In Paper.	Complete	None	D13088	??????
BV2S1A4T	008894 !	MT1-1	In Paper. 1 aa/2 nuc diff - Unexplained (AC to CA swap): M11954 has correct sequence	Complete	Complete	M11954	??????
BV2S1A5T	?????? !	4.49	In Paper.	Complete	Complete	X74852	??????
BV2S2A1O	008968	ORBV2S1*2'CL (ORBV2S2*1 in paper)	In Paper.	Complete	None	X72718	??????
	??????	HVB22.1	In Paper. 3 aa/3 nuc diff - Unexplained. Matches BV2S2A2O	Complete	None	L05149	??????
BV2S2A2O	008899	ORBV2S2*1'CL (ORBV2S2*2 in paper)	In Paper. 1 aa/1 nuc diff - Unexplained	Complete	None	X72717	??????
	??????	HVB22.1	Classed as BV2S2A1O in Paper. Matches this class	Complete	None	L05149	??????
BV2S2A3OT	008900	HT22G'CL	In Paper. 2 aa/2 nuc diffs - Unexplained: X57605 has correct sequence	Partial	None	X57605	??????
BV3S1	??????	X11	In Paper.	Complete	None	L36092	??????
	008920	HT12'CL	In Paper.	Complete	None	X57610	??????
	008919	HBVT22'CL	In Paper.	Complete	Complete	M27382	??????
	008918	PL4.4'CL	In Paper.	Complete	Complete	M13843	??????
	??????	Vb3.1	In Paper.	Complete	Complete	X74846	??????
	008922	DT259'CL	In Paper. 2 aa/3 nuc diff - Explained	Complete	Complete	X04929	??????
	009066 X	DE49'CL	In Paper.	Fragment	Complete	Z22967	008295
	008921 X !	HA1.7'CL	In Paper.	Complete	Complete	X63456	008111
	008923	B2-59'CL	In Paper.	Partial	Complete	M18462	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV4S1A1T	??????	111E15	In Paper.	Complete	Complete	Z29580	??????
	008901	PL2.14'CL	In Paper.	Complete	Complete	M13846	??????
	008907	2G2'CL (2G2β in paper)	In Paper.	Partial	Complete	M13553	??????
	008902	DT110'CL	In Paper.	Complete	Complete	X04921	??????
	008906	B1-75'CL	In Paper.	Complete	Complete	M18461	??????
	008904 X !	S26.2'CL	In Paper.	Complete	Complete	M97723	008200
	??????	X6A	In Paper.	Complete	None	L36092	??????
	008903 X	S14.11'CL		Complete	Complete	??????	008202
	027022 X	18030 B31'CL		Complete	Complete	??????	027014
	027023 X	18030 D6'CL		Complete	Complete	??????	027015
	009015	102DRF	Ambiguous: Also matches BV4S1A2T	Partial	Complete	??????	??????
BV4S1A2T	009007 !	HBP48'CL	In Paper.	Partial	Complete	X04926	??????
	009015	102DRF	Ambiguous: Also matches BV4S1A1T	Partial	Complete	??????	??????
BV4S1A3T	008905 !	PL5.7'CL	In Paper.	Complete	Complete	M13847	??????
BV4S2O	?????? !	H28.1	In Paper.	Complete	None	L05150	??????
BV5S1A1T	??????	K56	In Paper.	Complete	None	L36092	??????
	008870	HBP51'CL	In Paper.	Complete	Complete	X04927	??????
	008871	PL7.16'CL	In Paper.	Partial	Complete	M13849	??????
	008994	PL4.16'CL	In Paper.	Partial	Complete	M13848	??????
	008868 X !	AL61.102'CL	In Paper.	Complete	Complete	M97705	008220
	008869 X	S14.4'CL	In Paper.	Complete	Complete	M97715	008149
BV5S1A2T	008872 !	ph24'CL	In Paper.	Complete	Complete	M14271	??????
BV5S2	??????	A27	In Paper.	Complete	None	L36092	??????
	008877	IGRb09'CL	In Paper.	Complete	None	X58802	??????
	008977 !	PL2.5'CL	In Paper.	Partial	Complete	M13850	??????
	020388	BV5S2		Partial	None	??????	??????
BV5S3A1T	008874	HT415.9'CL	In Paper.	Complete	None	X57611	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008873	12A1'CL	In Paper. 3 aa/9 nuc diff - Unexplained	Partial	Complete	M13551	??????
	??????	12A1'CL	In Paper. 3 aa/9 nuc diff - Unexplained	Partial	Complete	M14299	??????
	??????	HBP-β2	In Paper. 3 aa/1 nuc diff (1 nuc insert) - Unexplained	Partial	Complete	X01411/K02780	??????
	025948 X !	WM'CL (WM-3 in paper)	In Paper.	Complete	Complete	M86362	025946
BV5S3A2T	??????	A27	In Paper.	Complete	None	L36092	??????
	008875 !	HT415.3'CL	In Paper.	Complete	None	X57612	??????
BV5S3A3T	008876 !	IGRb08'CL	In Paper.	Complete	None	X55801	??????
	020389	BV5S3'CL		Partial	None	??????	??????
BV5S4A1T	008979 !	IGRb06'CL	In Paper.	Partial	None	X58803	??????
BV5S4A2T	008959 X !	AL62.24'CL	In Paper.	Complete	Complete	M97709	008198
	??????	A14	In Paper.	Complete	None	L36092	??????
BV5S5P	??????	9	In Paper.	Complete	None	X61439	??????
	??????	H18	In Paper.	Complete	None	L36092	??????
	020390	BV5S5'CL		Partial	None	??????	??????
BV5S6A1T	008879 !	HT415'CL	In Paper.	Complete	None	X57615	??????
BV5S6A2T	008996 !	IGRb07'CL	In Paper.	Partial	None	X58804	??????
BV5S6A3N1T	008882	R1F3'CL	In Paper.	Partial	None	S50547	??????
	008883	R2B6-5'CL		Partial	None	??????	??????
	020391	BV5S6'CL	Ambiguous: Also matches BV5S6A3N2T	Partial	None	??????	??????
BV5S6A3N2T	??????	X1A	In Paper.	Complete	None	L36092	??????
	008878 X !	AL61.270'CL	In Paper.	Complete	Complete	M97707	008223
	??????	8.19	In Paper.	Complete	Complete	X74847	??????
	008965	CH35B'CL (*No name in paper)*	In Paper.	Complete	None	M73465	??????
	??????	*No name*	In Paper.	Complete	Complete	X68527	??????
	?????? P	*No name*	In Paper.	Complete	Joining Error?	X56142	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	020391	BV5S6'CL	Ambiguous: Also matches BV5S6A3N1T	Partial	None	??????	??????
BV5S7P	??????	A14	In Paper.	Complete	None	L36092	??????
	020392	BV5S7'CL (H139.1 in paper)	In Paper. 2 aa/6 nuc diffs - Unexplained: L26226 has correct sequence	Partial	None	L26226	??????
BV5S?	020387 !	BV5S1	Closest to BV5S1A1T (1 aa/1 nuc diff)	Partial	None	??????	??????
BV6S1A1N1	008834 !	HBP50'CL	In Paper.	Complete	Complete	X04934	??????
	??????	12.1	In Paper.	???????	????????		??????
	??????	9	In Paper.	Complete	None	X61440	??????
	??????	H18	In Paper.	Complete	None	L36092	??????
BV6S1A1N2T	??????	1.26	In Paper.	Complete	Complete	X74843	??????
BV6S1A2P	008957	6.1B (3.2 in paper)	In Paper.	Complete	Partial	M97943	??????
BV6S1A3T	015402 P	4D1	In Paper. 1 extra aa.	Partial	Joining Error	M13550	??????
	008988 X !	HBP-MLT'CL (4D1 in paper)	In paper. 1 extra aa. 0 aa/1 nuc diff - Unexplained: M12883 has correct sequence	Partial	Complete	M12883	008123
	009010	PL4.14'CL	In Paper.	Partial	Complete	M13852	??????
	009022	HBP04'CL	In Paper.	Partial	Complete	X04922	??????
BV6S2A1N1T	008851	ATL12-2 (ATL12-2G in paper)	In Paper.	Complete	Complete	M11953	??????
	008853 !	HBVT23'CL	In Paper.	Complete	Complete	M27383	??????
	009009	PL5.10'CL	In Paper. 1 aa/3 nuc diff - Unexplained	Partial	Complete	M13854	??????
	008852	ph5'CL	In Paper.	Complete	Complete	M14260	??????
	??????	Vβ6.3a	In Paper.	???????	????????		??????
	??????	A14	In Paper.	Complete	None	L36092	??????
BV6S2A1N2T	008854	HBVT10'CL	In Paper.	Complete	Complete	M27384	??????
BV6S2A2T	?????? !	4-1	In Paper.	Complete	None	X61441	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV6S3A1N1T	??????	A27	In Paper.	Complete	None	L36092	??????
	008986 !	HBP25'CL	In Paper. 2 aa/2 nuc diff - Unexplained: X04931 has correct sequence	Partial	Complete	X04931	??????
	??????	vb6.13b	In Paper.	Partial	None	L14480	??????
BV6S3A1N2T	008858	IGRb11'CL	In Paper.	Complete	None	X58806	??????
BV6S4A1	??????	A14	In Paper.	Complete	None	L36092	??????
	008838	IGRb10'CL	In Paper.	Complete	None	X58805	??????
	??????	Vβ6.9a	In Paper.	????????	????????		??????
	008837 !	HBVT11'CL	In Paper. 1 aa/3 nuc diff - Unexplained: M27386 has correct sequence	Complete	Complete	M27386	??????
BV6S4A2	008840	WBDM28A'CL	In Paper.	Complete	None	D13085	??????
	?????? !	IW28	In Paper.	Complete	Complete	X64742	??????
BV6S4A3T	008839 X !	L17Ti'CL (L17 in paper)	In Paper.	Complete	Complete	M15564	008169
	008841	L17'CL	In Paper.	Complete-2	Complete	M13552	??????
	??????	H7.1	In Paper. 0 aa/1 nuc diff - Unexplained	Partial	None	U03115	??????
BV6S4A4T	008842 !	ph22'CL	In Paper.	Complete	Complete	M14261	??????
BV6S4A5N1T	008836	HBVT116'CL	In Paper. 1 aa/3 nuc diff - Unexplained: M27385 has correct sequence	Complete	Complete	M27385	??????
BV6S4A5N2T	?????? !	1.40	In Paper.	Complete	Complete	X74844	??????
BV6S4A6T	?????? !	*No name*	In Paper.	Partial	Complete	L14854	??????
BV6S5A1N1	008830	ph16'CL	In Paper.	Complete	Complete	M14262	??????
	008831	ph79'CL	In Paper.	Complete	Complete	M15221	??????
	008992	IGRb12'CL	In Paper. 9 aa/14 nuc diff - Explained	Partial	None	X58807	??????
	024899	Vb6.7a'CL (5-2 in paper)	In Paper.	Partial	Complete	X61442	??????
	008829 X !	UA-S2'CL	In Paper.	Complete	Complete	M24089	008132

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	009013	OT-1 (OT-1/2 in paper)	In Paper.	Partial	Complete		??????
	008991	HUMAN OT-2'CL (OT-1/2 in paper)	In Paper.	Partial	None		??????
	008990	PCR-1'CL (PCR-1/2 in paper)	In Paper.	Partial	Complete		??????
	??????	H137	In Paper.	Complete	None	L36092	??????
BV6S5A1N2T	008832	HVBT45'CL (HBVT45 in paper)	In Paper.	Complete	Complete	M27387	??????
BV6S5A2	?????? !	HVB15	In Paper.	Complete	None	L36190	??????
	024900	Vb6.7b'CL (GL-PA in paper)	In Paper.	Partial	Complete	X61443	??????
BV6S6A1T	008859 !	HT147'CL	In Paper.	Complete	None	X57607	??????
BV6S6A2T	?????? !	A212	In Paper.	Complete	None	L36092	??????
	??????	Vβ6.14b	In Paper.	Partial	None	L14483	??????
BV6S7P	015596 P	TCRB-V6.10 (11 in paper)	In Paper.	Complete	None	X61444	??????
	??????	X21B	In Paper.	Complete	None	L36092	??????
BV6S8A1T	??????	Vβ6.11a	In Paper.	????????	????????		??????
BV6S8A2T	?????? !	X1A	In Paper.	Complete	None	L36092	??????
	??????	Vβ6.11c	In Paper.	Partial	None	L14432	??????
BV6S9P	??????	X1A	In Paper. Many diffs - Explained	Complete	None	L36092	??????
	??????	Vβ6.12a	In Paper. Many diffs - Explained	Partial	None	M97503	??????
BV6S?	024348 P	TCRBV6S1*3p'CL	Closest to BV6S1A2P (1 aa/1 nuc diff)	Complete	None	??????	??????
BV7S1A1N1T	008880	PL4.9'CL	In Paper. 1 aa/1 nuc diff - Unexplained: M13855 has correct sequence	Partial	Complete	M13855	??????
BV7S1A1N2T	??????	K56	In Paper.	Complete	None	L36092	??????
	008963	IGRb19'CL	In Paper.	Complete	None	X58813	??????
	008964	HT459'CL	In Paper.	Complete	None	X57728	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	?????? !	Vβ7.1	In Paper. 1 aa/2 nuc diff - Unexplained	Complete	Complete	X74842	??????
BV7S2A1N1T	008981	HT267’CL	In Paper.	Partial	None	X57618	??????
	009014	PL4.19’CL	In Paper. 5 aa/15 nuc diff - Unexplained	Partial	Complete	M13856	??????
	008955	vb7.n1’CL	In Paper. Many diffs - Explained	Complete	None	L06887	??????
	021892 !	BV7S3’CL	Ambiguous: Also matches BV7S2A1N4T	Complete	None	??????	??????
BV7S2A1N2T	008980	HT267.1’CL	In Paper. 0 aa/1 nuc diff - Unexplained	Partial	None	X57616	??????
BV7S2A1N3T	008958	vb7.n2’CL	In Paper.	Complete	None	L06888	??????
BV7S2A1N4T	??????	G54	In Paper.	Complete	None	L36092	??????
	021892	BV7S3’CL	Ambiguous: Also matches BV7S1A2N1T	Complete	None	??????	??????
BV7S2A2T	008881 !	IGRb18’CL	In Paper.	Complete	None	X58812	??????
BV7S3A1T	008825 !	IGRb17’CL	In Paper.	Complete	None	X58811	??????
BV7S3A2T	??????	X21B	In Paper.	Complete	None	L36092	??????
	008982 !	HT267.2’CL	In Paper.	Complete	None	X57617	??????
	??????	BV7.3b	In Paper. 1 aa/2 nuc diff - Unexplained	Complete	None	D13083	??????
	021891	BV7S2’CL		Complete	None	??????	??????
BV8S1	008822	VB8.1’CL (M18H7.1B7 in paper)	In Paper.	Complete	None	X07192	??????
	008820	YT35’CL	In Paper.	Complete	Complete	K01571/X00437	??????
	008989	JM ’CL (4D8 in paper)	In Paper.	Partial	Complete	K02885/X01417	??????
	??????	*No name*	In Paper.	Partial	Complete	X01417	??????
	008821 !	ph11’CL	In Paper.	Complete	Complete	M14265	??????
	??????	H7.1	In Paper.	Complete	None	U03115	??????
BV8S2A1T	??????	H7.1	In Paper.	Complete	None	U03115	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008816	BM3-2’CL (M3-2 in paper)	In Paper.	Complete	None	K02546	??????
	008815	VB8.2’CL (p8H7.1B5 in paper)	In Paper.	Complete	None	X07222	??????
	008993	HBP41’CL	In Paper.	Partial	None	X04925	??????
	008824	PL3.3’CL	In Paper.	Partial	Complete	M13858/M16307	??????
	?????? ! X	8B3	In Paper.	Complete	Complete	M81773	M81774
BV8S2A2N1T	008819 !	ph8’CL	In Paper.	Complete	Complete	M14264	??????
	008817	HT2.12’CL	In Paper. 1 aa/1 nuc diff - Unexplained	Complete	None	X57619	??????
BV8S2A2N2T	008818	HT242’CL	In Paper. 1 aa/1 nuc diff - Unexplained (same diff as 8817)	Complete	None	X57720	??????
BV8S3	008823 !	VB8.3’CL	In Paper.	Complete	None	X07223	??????
	??????	H130.1	In Paper.	Complete	None	U03115	??????
BV8S4P	015399 P	VB8.4 (pBH9.1R3 in paper)	In Paper.	Complete	None	X07224	??????
BV8S5P	015400 P	VB8.5 (M18VB8.5 in paper)	In Paper.	Complete	None	X06936	??????
	??????	*No name*	In Paper. Many diff - Unexplained	Partial	Partial	M13576	??????
	??????	G15	In Paper.	Complete	None	L36092	??????
BV8S?	009038 !	CH1B’CL	Closest to BV8S3 (1 aa/1 nuc diff)	Partial	None	??????	??????
BV9S1A1T	??????	K56	In Paper.	Complete	None	L36092	??????
	009071 X	DE5’CL	In Paper.	Fragment	Complete	Z23044	008300
	008961	HT307’CL	In Paper.	Complete	None	X57614	??????
	008960	IGRb20’CL	In Paper.	Complete	None	X58814	??????
	008978 !	PL2.6’CL	In Paper. 2 aa/2 nuc diff - Unexplained	Partial	Complete	M13859	??????
	021893	BV9S1’CL		Complete	None	??????	??????
BV9S1A2T	008962 !	Vb9.n1’CL (Vb9.n in paper)	In Paper.	Complete	None	L06889	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV9S2A1PT	015518 P	H307.1 (HT307.1 in paper)	In Paper.	Complete	None	X57608	??????
	??????	VW114	In Paper. 1 aa/1 nuc diff - Explained	Complete	None	M33240	??????
BV9S2A2PT	??????	X21B	In Paper.	Complete	None	L36092	??????
BV9S?	021894 P	BV9S2	Closest to BV9S2A2PT (0 aa/1 nuc diff)	Complete	None	??????	??????
BV10S1P	??????	C215	In Paper	Complete	None	L36092	??????
	008850	PL3.9	In Paper. 1 aa/1 nuc diff - Unexplained (Could be functional gene)	Complete	Complete	M13860/M16309	??????
	015401 P	ATL12-1	In Paper. Many aa/0 nuc diff - Translated differently in Kabat	Complete	None	M11956	??????
BV10S2O	??????	HVB26.1	In Paper	Complete	None	L05151	??????
BV10S?	008850	PL3.9	Classed as BV10S1P (1 aa/1 nuc diff - Unexplained (Could be functional gene)	Complete	Complete	M13860/M16309	??????
BV11S1A1T	??????	G1	In Paper.	Complete	None	L36092	??????
	008937 !	PL3.12'CL	In Paper	Complete	Complete	M13861	??????
	008936	ph15'CL	In Paper	Complete	Complete	M14266	??????
BV11S1A2T	?????? !	1.3	In Paper. 11 extra aa (may be wrong)	Complete	Complete	X74845	??????
BV11S2OP	??????	HVB36.1	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	L05152	??????
BV12S1A1N1	??????	Allele 1	In Paper. Genbank code not found	????????	????????	S47256	??????
BV12S1A1N2	??????	H7.1	In Paper	Complete	None	U03115	??????
	015403 P	ph27	In Paper	Complete	Joining Errors	M14268	??????
	008984 !	HBP54	In Paper.	Partial	Complete	X04935	??????
	020393	BV12S2'CL		Partial	None	??????	??????
BV12S1A1N3	008983	PL4.2'CL	In Paper. 5 aa/2 nuc diff (insert/delete pair) - Explained.	Partial	Complete	M13862	??????
BV12S1A1N4	??????	Allele 4	In Paper.	????????	????????		??????
BV12S2A1T	??????	H18	In Paper.	Complete	None	L36092	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008938	IGRb13'CL	In Paper.	Complete	None	X58808	??????
	?????? !	GM2.11	In Paper	Complete	Complete	M64352	??????
BV12S2A2T	020395 !	BV12S4'CL (H18.1 in paper)	In Paper.	Partial	None	L26230	??????
BV12S2A3T	008956 !	WBDM21C'CL	In Paper.	Complete	None	D13084	??????
BV12S3	??????	G15	In Paper	Complete	None	L36092	??????
	009006	HT96'CL	In Paper.	Partial	None	X57609	??????
	??????	*No name*	In Paper	????????	????????		??????
	027026 !	LWF A20,C8'CL		Complete	Complete	??????	??????
BV12S?	020394 !	BV12S3'CL	Closest to BV12S3 (2 aa/2 nuc diffs)	Partial	None	??????	??????
BV13S1	??????	X1A	In Paper	Complete	None	L36092	??????
	008930 !	PL4.24'CL	In Paper.	Complete	Complete	M13863	??????
	008929	HBP34	In Paper	Complete	Complete	X04932	??????
	??????	*No name*	In Paper	Complete	Complete	Z26594	??????
	020511	BV13S1'CL		Partial	None	??????	??????
BV13S2A1T	??????	5-2'CL	In Paper.	Complete	None	X61445	??????
	008999	PL5.3'CL	In Paper.	Partial	Complete	M13864	??????
	??????	HVB15.2	In Paper	Partial	None	L26229	??????
	008953 X !	RFL3.8 (RFL3.8b in paper)	In Paper	Complete	Complete	M77498	008232
	??????	X21B	In Paper	Complete	None	L36092	??????
	??????	G54	In Paper	Complete	None	L36092	??????
	020512	BV13S2'CL		Partial	None	??????	??????
BV13S2A2PT	008939	CEM-VB1'CL (CEM.1 in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	M13575	??????
	015405 P	HPB-2	In Paper. Many aa/0 nuc diff - Differently Translated	Complete	None	M31347	??????
BV13S2A3PT	??????	pMF β4.3	In Paper.	????????	????????		??????
BV13S3	??????	11	In Paper	Complete	None	X61446	??????
	008940	IGRb14'CL	In Paper.	Complete	None	X58809	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008952	Vβ13.n1'CL	In Paper	Complete	None	L06890	??????
	?????? !	Vβ13.3	In Paper	Complete	Complete	X74850	??????
	??????	K56	In Paper	Complete	None	L36092	??????
	020513	BV13S3'CL		Partial	None	??????	??????
BV13S4	?????? !	4-1	In Paper.	Complete	None	X61447	??????
	??????	A14	In Paper.	Complete	None	L36092	??????
	020514	BV13S4'CL		Partial	None	??????	??????
BV13S5	??????	9	In Paper.	Complete	None	X61653	??????
	008941 !	IGRb15'CL	In Paper.	Complete	None	X58810	??????
	008951	Vb13.n2'CL	In Paper. 1 aa/2 nuc diff - Unexplained (GC to CG swap)	Complete	None	L06891	??????
	??????	H18	In Paper.	Complete	None	L36092	??????
	020515	BV13S5'CL		Partial	None	??????	??????
BV13S6A1N1T	008932 !	HT165'CL	In Paper.	Complete	None	X57721	??????
BV13S6A1N2T	008931	IGRb16'CL	In Paper.	Complete	None	X58815	??????
	009001	17A2'CL	Ambiguous: Also matches BV13S6A2T	Partial	Complete	??????	??????
BV13S6A2T	??????	X1A	In Paper.	????????	????????	L36092	??????
	008934	HT165.2'CL	In Paper.	Complete	None	X57606	??????
	008935 !	A2'CL (A2β in paper)	In Paper.	Complete	Complete	S60794	??????
	020516	BV13S6'CL		Partial	None	??????	??????
	009001	17A2'CL	Ambiguous: Also matches BV13S6A1N2T	Partial	Complete	??????	??????
BV13S6A3T	?????? !	3.1	In Paper.	Complete	Complete	X74848	??????
BV13S6A4T	008933 !	Vb13.n3'CL	In Paper.	Complete	None	L06892	??????
BV13S7	?????? !	A212	In Paper.	Complete	None	L36092	??????
	020517	BV13S7'CL (H127 in paper)	In Paper. 1 aa/1 nuc diff - Unexplained	Partial	None	L26228	??????
BV13S8P	??????	A27	In Paper.	Complete	None	L36092	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	020518	BV13S8'CL (H127 in paper)	In Paper.	Partial	None	L26227	??????
BV13S?	008998	G36'CL	Closest to BV13S3 (0 aa/1 nuc diff)	Partial	Complete	??????	??????
	025888	Vb13.6-BR5.11'CL	Closest to BV13S6A2T (3 aa/5 nuc diff)	Partial	None	??????	??????
BV14S1	??????	C21	In Paper.	Complete	None	L36092	??????
	008924 !	ph21'CL	In Paper.	Complete	Complete	M14267	??????
	008925	PL8.1'CL	In Paper.	Complete	Complete	M16314/M13865	??????
	008985	HBP55'CL	In Paper.	Partial	Complete	X04928	??????
	??????	8.9	In Paper.	Complete-2	Partial	M17200	??????
	008997	67DRF'CL		Partial	Complete	??????	??????
	009000	40'CL		Partial	Complete	??????	??????
	009003	212DRD'CL		Partial	Complete	??????	??????
BV15S1	008926	ATL2-1 (ATL2-1G in paper)	In Paper.	Complete	Complete	M11951	??????
	008927 !	ph32'CL	In Paper.	Complete	Complete	M14269	??????
	??????	G1	In Paper.	Complete	None	L36092	??????
	009002	5A2'CL		Partial	Complete	??????	??????
BV15S2OP	??????	V11B	In Paper.	Complete	None	L05153	??????
	??????	HT-9	In Paper.	????????	????????		??????
BV16S1A1N1	008826	VB16'CL (*No Name in paper)*	In Paper.	Complete	None	X06154	??????
	008987 !	HBP42'CL	In Paper.	Partial	Complete	X04933	??????
	008828	HT370'CL	In Paper.	Complete	None	X57723	??????
	??????	H130.1	In Paper.	Complete	None	U03115	??????
BV16S1A1N2	008827	HT219'CL	In Paper.	Complete	None	X57722	??????
BV17S1A1T	??????	C215	In Paper.	Complete	None	L36092	??????
	008928 !	HBVT02'CL	In Paper.	Complete	Complete	M27388	??????
	023276 X	XPZ10'CL	Ambiguous: Also matches BV17S1A3T	Partial	Complete	??????	023270

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV17S1A2T	008954 X !	S30.10'CL	In Paper.	Complete	Complete	M97725	008201
BV17S1A3T	?????? !	*No name*	In Paper. 1 aa/1 nuc diff - Unexplained	Complete	None	L19936	??????
	023276	XPZ10'CL	Ambiguous: Also matches BV17S1A1T	Partial	Complete	??????	??????
BV17S?	024070 !	B17-438'CL	Closest to BV17S1A1T (1 aa/1 nuc diff)	Complete	Complete	??????	??????
BV18S1	??????	A16	In Paper.	Complete	None	L36092	??????
	008847	HBVT56'CL	In Paper.	Complete	Complete	M27389	??????
	008845	p29'CL	In Paper.	Complete	Complete	M14270	??????
	008846	p26'CL	In Paper.	Complete	Complete	M15223	??????
	008848 X !	S14.107'CL	In Paper.	Complete	Complete	M97711	008224
BV19S1P	??????	C215	In Paper.	Complete	None	L36092	??????
	008849	HBVT72'CL	In Paper.	Complete	Complete	M27390	??????
	020519	BV19S1		Partial	None	??????	??????
BV19S2O	020520	BV19S2(O)'CL	In Paper.	Partial	None	L26225	??????
BV20S1A1N1	008946 !	Vb18(A)'CL (H29 in paper)	In Paper. 1 extra aa	Complete	None	Z13967	??????
BV20S1A1N2	008947	Vb18(B)'CL	In Paper.	Complete	None		??????
	??????	BV20.1a	In Paper.	Complete	None	D13086	??????
	??????	BV20S1	In Paper.	Complete	None	L36092	??????
	008949	WBDM30A'CL	1 extra aa	Complete	None	??????	??????
BV20S1A1N3T	008948	Vb20.n1'CL	In Paper.	Complete	None	L06893	??????
BV20S1A2P	015595 P	Vb18(C) (Vβ18C in paper)	In Paper.	Complete	None		??????
	??????	Allele 1	In Paper.	????????	????????		??????
BV20S1A3T	008950 !	HUT'CL (HUT102β in paper)	In Paper.	Complete	Complete	M13554	??????
BV21S1	008844 !	Vb21.1'CL (BV21.1 or H18.1 in paper)	In Paper.	Complete	None	M33233	??????
	??????	B17ct7	In Paper.	Complete-2	None	D16584	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV21S2A1N1	??????	Vβ21.4a	In Paper.	????????	????????		??????
BV21S2A1N2T	008966 !	IGRb02'CL	In Paper.	Complete	None	X58797	??????
BV21S2A2	??????	H7.1	In Paper.	Complete	None	U03115	??????
	?????? !	IW6	In Paper.	Complete	Complete	X56665	??????
	??????	H7.1	In Paper. 6 aa/8 nuc diff - Explained	Complete	None	M33234	??????
	??????	H12.18	In Paper.	Complete	None	L36092	??????
	008855	Vb21.2'CL	7 aa/9 nuc diff. Could be M33234 (H7.1)	Complete	None	??????	??????
BV21S2A3T	008843 !	Vb21'CL	In Paper.	Partial	Complete	M62377	??????
BV21S3A1T	008856 !	Vb21.3'CL (BV21.3 in paper)	In Paper.	Complete	None	M33235	??????
BV21S3A2N1T	008857	IGRb01'CL	In Paper.	Complete	None	X58796	??????
	??????	IW10	In Paper.	????????	????????		??????
BV21S3A2N2T	??????	BV21S3	In Paper.	Complete	None	L36092	??????
	?????? !	DD11	In Paper. (009069 is fragment of this seq)	Complete	Complete	Z23042	??????
BV22S1A1T	008917 !	Vb23'CL	In Paper.	Complete	None	M62379	??????
BV22S1A2N1T	??????	K26	In Paper.	Complete	None	L36092	??????
	008915	IGRb03'CL	In Paper.	Complete	None	X58798	??????
	008916	HT2.10'CL	In Paper.	Complete	None	X57727	??????
BV22S1A2N2T	?????? !	*No name*	In Paper.	Complete	Complete	M64351	??????
BV22S?	024355 !	newmal87pro'CL	Closest to BV22S1A2N1T (3 aa/3 nuc diff)	Complete	None	??????	??????
BV23S1A1T	008914 !	Vb22'CL	In Paper.	Complete	None	M62378	??????
BV23S1A2T	??????	H7.1	In Paper.	Complete	None	U03115	??????
	008913 !	HT183'CL	In Paper.	Complete	None	X57613	??????
	008912	IGRb04'CL	In Paper.	Complete	None	X58799	??????
	??????	IW22	In Paper.	????????	????????		??????
BV24S1A1T	008910 !	Vb24'CL	In Paper.	Complete	None	M62376	??????
BV24S1A2T	008908	IGRb05'CL	In Paper.	Complete	None	X58800	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	008909	HT77'CL	In Paper.	Complete	None	X57725	??????
	008967	CH18B'CL *(No name in paper)*	In Paper.	Complete	None	M73464	??????
	023275 X !	XPE15'CL		Partial	Complete	??????	023269
BV24S1A3T	?????? !	H130.1	In Paper.	Complete	None	U03115	??????
BV25S1A1T	022634 !	BV25S1'CL (HVB30.A in paper)	In Paper. 1 aa/1 nuc diff - Unexplained: L26231 has correct sequence	Complete	None	L26231	??????
BV25S1A2PT	??????	H130.1	In Paper.	Complete	None	U03115	??????
BV25S1A3T	?????? !	HsVB25	In Paper.	Complete	None	L26054	??????
BV26S1P	??????	H130.1	In Paper.	Complete	None	U03115	??????
BV27S1P	??????	K26	In Paper.	Complete	None	L36092	??????
BV28S1P	??????	C68	In Paper.	Complete	None	L36092	??????
BV29S1P	??????	C215	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	L36092	??????
BV30S1A1PT	??????	H137	In Paper.	????????	????????	L36092	??????
BV30S1A2PT	??????	HVB15	In Paper.	????????	????????	L36190	??????
BV31S1P	??????	H137	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	L36092	??????
BV32S1P	??????	H18	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	L36092	??????
BV33S1P	??????	C68	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	L36092	??????
BV34S1P	??????	C21	In Paper.	????????	????????	L36092	??????

Table B.4: The table shows the Kabat entry IDs of members of the officially designated classes for human beta chains [125]. A group of six question marks indicates data which is not known or not available. In the Kabat ID column there are also flags indicating whether the sequence has a known pair (X), whether it was used in the analysis (!), and if it is defined as a psuedogene in the Kabat database (P). Where sequences did not match the sequences or names of members of any of the existing classes, the closest matching class was identified and a temporary class created with a name ending in a question mark. For example if a sequence was similar to BV5S1 the temporary class would be called BV5S?.

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV1S1A1	009111	BW14'CL	In Paper.	Complete	Complete	M20177	??????
	009084	1.9.2'CL	In Paper.	Complete	Complete	X02778	??????
	015688 P	Vb1-Db2.1-Jb2.4 (37.2A10 in paper)	In Paper.	Complete	Joining Errors	M22606	??????
	009083	A20.2.15'CL	In Paper.	Complete	Complete	M11456	??????
	009087	BW5147'CL	In Paper. 1 aa/1 nuc diff - Explained (Variant 1)	Complete	Complete	X02779	??????
	009086 !	VB11'CL	In Paper. 1 aa/1 nuc diff - Explained (Variant 1)	Complete	Partial	M13676	??????
	009082	86T1'CL	In Paper. 0 aa/1 nuc diff - Explained (Variant 2)	Complete	Complete	X00438	??????
	009109 X	D6'CL	In Paper. 0 aa/1 nuc diff - Explained (Variant 2)	Complete	Complete	M20877	008472
	009110 X !	1F8'CL	In Paper. 0 aa/1 nuc diff - Explained (Variant 2)	Complete	Complete	M20878	008474
	009085 X !	DA.33.C2'CL		Complete	Complete	??????	008420
BV1S1A2	009112	VB1a'CL (SWR-1 in paper)	In Paper.	Complete	None		??????
BV2S1	009159	AR1'CL	In Paper. 0 aa/1 nuc diff - Unexplained: X02780 has correct sequence	Complete	Complete	X02780	??????
	009165	VB6'CL	In Paper.	Partial	Partial	M13671	??????
	009161	BDFLBI (BDFL1 in paper)	In Paper.	Complete	Complete	X03670	??????
	009164	MOUSE Vb2-Db1.1-Jb1.2'CL (18.2A10 in paper)	In Paper.	Partial	Complete	M22605	??????
	009162 X !	AR-5'CL	In Paper.	Complete	Complete	M21203	008397
	??????	MT1-6	In Paper.	Partial	Partial	M34203	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	009163 X !	E1'CL	In Paper. 1 aa/2 nuc diff - Explained (Variant)	Complete	Complete	X01642	008445
	009160 X	BB1.D5'CL		Complete	Complete	??????	008416
	009287	18N.30'CL		Partial	Complete	??????	??????
BV3S1A1	009102	3H.25'CL	In Paper.	Complete	Complete	M12415	??????
	009099	2B4#71'CL (B.6(2B4) in paper)	In Paper.	Complete	Complete	K02548	??????
	009101 X	5C.C7'CL	In Paper.	Complete	Complete	X03863	008463
	009100 X !	C.F6'CL	In Paper.	Complete	Complete	X03862	008462
	009103	V3.1b'CL		Complete	None	??????	??????
	009286	A4.A1'CL	Ambiguous: Also matches BV3S1A2	Partial	Complete	??????	??????
BV3S1A2	009104 !	V3.1a'CL (SWR-3 in paper?)	In Paper.	Complete	None		??????
	009286	A4.A1'CL	Ambiguous: Also matches BV3S1A1	Partial	Complete	??????	??????
BV4S1	009089 !	TB3'CL	In Paper.	Complete	Complete	X02781	??????
	009088	VB9'CL	In Paper.	Complete	Partial	M13674	??????
	?????? X	MT1-14	In Paper.	Complete	Partial	M34199	M34198
	??????	52H10F11'CL	In Paper. (009271 is fragment of this sequence)	Complete	Complete	M16121	??????
	??????	42H11"CL	In Paper. (009270 is fragment of this sequence)	Complete	Complete	M16122	??????
	009090	BB02"CL	In Paper.	Complete-1	Complete	X54322	??????
BV5S1	009092	VB5.1'CL	In Paper.	Complete	None	M15613	??????
	009093 !	NB3'CL	In Paper.	Complete	Complete	M27349	??????
	009091	VB8'CL	In Paper.	Complete	Partial	M13673	??????
	015695 P	LH8-2	In Paper.	Complete	Joining Errors	M21674	??????
	009094 !	TB21'CL	In Paper. 1 aa/1 nuc diff - Explained (Variant)	Complete	Complete	X02782	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	025314	9429A-5-1'CL		Partial	Complete	??????	??????
	025315	9429A-5-22'CL		Partial	Complete	??????	??????
	025317	9429B-5-1.24'CL		Partial	Complete	??????	??????
	025318	9429B-5-25'CL		Partial	Complete	??????	??????
	025319	9429B-5-26'CL		Partial	Complete	??????	??????
	025320	B6-5-1'CL		Partial	Complete	??????	??????
	025321	B6-5-2'CL		Partial	Complete	??????	??????
	025322	B6-5-3'CL		Partial	Complete	??????	??????
	025323	B6-5-4'CL		Partial	Complete	??????	??????
	025324	B6-5-5'CL		Partial	Complete	??????	??????
	025325	B6-5-6'CL		Partial	Complete	??????	??????
	025326	B6-5-7'CL	1 aa/1 nuc diff	Partial	Complete	??????	??????
	025327	B6-5-8'CL		Partial	Complete	??????	??????
	025328	B6-5-9'CL		Partial	Complete	??????	??????
	025329	B6-5-10'CL		Partial	Complete	??????	??????
BV5S2A1	009095	VB5.2'CL	In Paper.	Complete	None	M15614	??????
	015694 P	BW12	In Paper.	Complete	Joining Errors	M20136	??????
	009096 X !	5/10-20D	Class BV5S2A2 in paper but matches this class	Complete	Complete	X05737	008453
	009097	LH8-1	Class BV5S2A2 in paper but matches this class	Complete	Complete	M21673	??????
	021353 X !	N15'CL		Complete	Complete	??????	021351
BV5S2A2	009098 !	NZW22s'CL	In Paper.	Complete	Complete	M30881	??????
	009096 X	5/10-20D'CL	In Paper. 1 aa/1 nuc diff - Unexplained (see BV5S2A1)	Complete	Complete	X05737	008453
	009097	LH8-1'CL	In Paper. 1 aa/1 nuc diff - Unexplained (see BV5S2A1)	Complete	Complete	M21673	??????
BV5S3P	015691 P	VB5.3 (Vβ5.3P in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	M15615	??????
BV6S1A1	009154 X !	C9'CL	In Paper.	Complete	Complete	X05738	008401

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	009151	VB1'CL	In Paper.	Complete	Partial	M10093	??????
	009152	M3'CL	In Paper.	Complete	Complete	M12434	??????
	??????	5.3.18'CL	In Paper. (009272 is fragment of this sequence)	Complete	Complete	M16120	??????
	009158	p3F9'CL (3F9 in paper)	In Paper.	Partial	Complete	M29841	??????
	009156 X !	LB2'CL	In Paper. 1 aa/1 nuc diff - Explained (Variant)	Complete	Complete	X01643	008393
	009153 X	AF.3.G7'CL		Complete	Complete	??????	008406
	009155	V6b'CL		Complete	None	??????	??????
BV6S1A2	009157	V6a'CL (SWR-6 in paper)	In Paper.	Complete	None		??????
BV7S1	009149 !	2CB[pHDS11]'CL (pHDS11 in paper)	In Paper.	Complete	Complete	X00696	??????
	009150 X	5/10-20K'CL	In Paper.	Partial	Complete	X05735	008382
BV8S1	009144	VB8.1'CL	In Paper.	Complete	None	M15616	??????
	009145 X !	C5'CL	In Paper.	Complete	Complete	X01641	008455
	009143	TB12'CL	In Paper.	Complete	Complete	X02783	??????
	025910	P14B.1'CL	In Paper.	Complete	Complete	X06772	??????
	009226 X	9C127'CL	Text says vb8.1	Fragment	Complete	??????	008427
BV8S2A1	009136	VB8.2'CL	In Paper.	Complete	Complete	M15617	??????
	009135	TB2'CL	In Paper.	Complete	Complete	X02784	??????
	009134	VB4'CL	In Paper.	Complete	Partial	M13669	??????
	009137	B6.2.16	In Paper.	Complete	Partial	M19404	??????
	??????	V2.1	In Paper.	Partial	Fragment	M34205	??????
	009139	3A9'CL	In Paper. 1 aa/1 nuc diff - Unexplained (N terminus)	Complete	Complete	M26417	??????
	009141 !	NB1'CL	In Paper. 2 aa/2 nuc diff - Explained (Variant)	Complete	Complete	M27350	??????
	009138 X !	8/10-2'CL		Complete	Complete	??????	008400
BV8S2A2	009140 !	C127'CL	In Paper.	Complete	None		??????
BV8S2A3	009142 !	ER34'CL	In Paper.	Complete	None		??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV8S3	009147	VB8.3'CL	In Paper. 1 aa/1 nuc diff - Unexplained	Complete	None	M15618	??????
	009146	TB23'CL	In Paper.	Complete	None	X02785	??????
	?????? !	2B11	In Paper.	Complete	Complete	M34219	??????
	009148	MOUSE MDA 'CL		Partial	Partial	??????	??????
	009223 X	MOUSE 10I'CL	Text says vb8.3	Fragment	Complete	??????	008431
	009227	MOUSE 8F8.10'CL	Text says vb8.3	Fragment	Complete	??????	??????
BV8S?	023723 !	38CH Vb8.2'CL	Closest to BV8S2A1 (1 aa/1 nuc diff)	Complete	Complete	??????	??????
BV9S1	009133	VB2'CL	In Paper.	Complete	None	M13677	??????
	027524 !	Vb9-Db1.1-Jb2.1'CL		Complete	Complete	??????	??????
BV10S1A1	??????	Vβ10-8	In Paper. (some uncertain nucleotides)	Complete	None	X56725	??????
	009130 X !	No.8	In Paper.	Complete	Complete	X56702	008441
	009132	VB3'CL	In Paper. 1 aa/3 nuc diff - Unexplained (N terminal)	Partial	Partial	M13678	??????
	??????	MT1-27	In Paper.	Partial	Complete	M34201	??????
	009128	Cw3/1.1'CL		Complete	Complete	??????	??????
	009127	V10B'CL		Complete	None	??????	??????
BV10S1A2	009131 !	V10A'CL (SWR-1 in paper)	In Paper.	Complete	None		??????
	019735	BUB/BnJ'CL		Partial	None	??????	??????
BV10S?	009129 !	50.1 BETA'CL	Closest to BV10S1A1 (1 aa/4 nuc diff)	Complete	Complete	??????	??????
BV11S1	009114	VAK (LVAK in paper)	In Paper.	Complete	None	N00046	??????
	009113	VB5'CL	In Paper.	Complete	Partial	M13670	??????
	009115 !	F3-42 (cF3-42 in paper)	In Paper.	Complete	Complete	X04331	??????
	009117	AK1'CL	In Paper. 2 aa/3 nuc diff - Unexplained	Complete	Complete	M15459	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	009116 X !	F5'CL	In Paper. 1 aa/2 nuc diff - Unexplained (CT to TC swap)	Complete	Complete	X14388	008413
BV12S1T	009122	NZW8'CL	In Paper. 2 extra aa	Partial	Complete	M30880	??????
	015689 P	VB7 (Vβ7 in paper)	In Paper. 1 aa/5 nuc diff	Partial	Joining Errors	M13672	??????
	009120 X !	ZZ38'CL	47 extra aa (complete v region)	Complete	Complete	??????	008385
BV13S1	009118 !	BVI/5.b11'CL	In Paper.	Complete	Complete	M31648	??????
	??????	Vβ10	In Paper. 3 aa/7 nuc diff - Unexplained: Pseudogene in M13675	????????	????????	M13675	??????
	009119	Vb12H6 [Vb13]'CL (Vβ12H6 in paper)	In Paper.	Complete	Complete	M25913	??????
BV14S1	009167 !	VB14-J6.19'CL (VB14GL in paper?)	In Paper.	Complete	Complete	X03277	??????
	009166	SJL33'CL	In Paper.	Complete	Partial	M11858	??????
	009168 X !	C11'CL	In Paper. 1 aa/1 nuc diff - Explained (Variant)	Complete	Complete	M26418	008454
BV15S1A1	?????? !	FN1-18	In Paper.	Complete	Complete	X04047	??????
BV15S1A2	009121 !	SJL73'CL	In Paper.	Complete	Partial		??????
BV16S1A1	009124 X !	4.C3'CL	In Paper.	Complete	Complete	X03865	008465
	009123 X	B10'CL	In Paper.	Complete	Complete	X03864	008464
BV16S1A2	009125 !	BDFLBII'CL (BDFLI in paper)	In Paper.	Complete	Partial	X03671	??????
	009126	SJL4	In Paper. 1 aa/2 nuc diff - Unexplained: M11860 has correct sequence	Complete	Partial	M11860	??????
BV17S1A1	009105 !	VB17A'CL	In Paper.	Complete-2	Complete	M16203	??????
BV17S1A2P	015692 P	VB17B	In Paper.	Complete	None	M22007	??????
BV17S1A3	009106 !	VB17A2'CL	In Paper.	Complete-2	None	M61184	??????
BV18S1	009169 !	Vb18'CL	In Paper.	Complete	None	X16695	??????
	??????	pM1pr2	In Paper. 4 aa/10 nuc diff - Unexplained (N terminal diffs)	Partial	Complete	M14294	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
BV19S1A1	??????	Vβ19a	In Paper.	????????	????????		??????
BV19S1A2P	009108	VbN3(Vb19)'CL (N3 in paper)	In Paper.	Complete	None	X16691	??????
BV20S1	009107 !	VB20'CL (K9 in paper)	In Paper.	Partial	None	X59150	??????
BV21S1P	015696 P	MOUSE VbN1 (N1 in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	X16689	??????
BV22S1P	015697 P	MOUSE VbN2 (N2 in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	X16690	??????
BV23S1P	015698 P	MOUSE VbN5 (N5 in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	X16692	??????
BV24S1P	015699 P	MOUSE VbN8 (N8 in paper)	In Paper.	Complete	None	X16693	??????
BV25S1P	015700 P	MOUSE VbN9 (N9 in paper)	In Paper.	Partial	None	X16694	??????

Table B.5: The table shows the Kabat entry IDs of members of the officially designated classes for mouse β chains [126]. A group of six question marks indicates data which is not known or not available. In the Kabat ID column there are also flags indicating whether the sequence has a known pair (X), whether it was used in the analysis (!), and if it is defined as a psuedogene in the Kabat database (P). Where sequences did not match the sequences or names of members of any of the existing classes, the closest matching class was identified and a temporary class created with a name ending in a question mark. For example if a sequence was similar to BV5S1 the temporary class would be called BV5S?.

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
GV1S1P	015409 P	V1.GL'CL (λSH4 in paper)	In Paper.	Complete	None	M12949	??????
GV1S2A1T	010516	V2.GL'CL (λSH4 in paper)	In Paper.	Complete-2	None	M13429	??????
	015417 P	HGP01	In Paper. Many aa/Many nuc diffs - Unexplained	Complete	None	M27338	??????
	010518	V2-JP1'CL (*No name in paper)*	In Paper.	Complete	Complete	S72759	??????
GV1S2A2T	010517	HGT26	In Paper.	Complete-1	Partial	M27337	??????
GV1S3A1N1T	010507	HUMAN V3.GL'CL (λSH4 in paper)	In Paper.	Complete	Partial?	M13430	??????
	010508	MOLT-13 (k in paper)	In Paper.	Complete	Partial	Y00790	??????
	010506	LSG12'CL(λSγ12 in paper)	In Paper.	Complete	Partial	M13824/X03437	??????
	015416 P	HGP03	In Paper.	Complete	Joining Errors	M27336	??????
	010510	V3-J1rs'CL (Vγ3 in paper)	In Paper. 2 extra aa	Complete	Partial?	S72844	??????
	010723	G3 (Vγ3 in paper)	In Paper	Partial	Partial	S60175	??????
	010722	E103'CL		Partial	Partial	??????	??????
GV1S3A1N2T	010511	pTγ1/2	In Paper.	Complete-4	Partial	M13231	??????
	010509	Vg1.1'CL (pgVγ1.1 in paper)	In Paper.	Complete	Partial	X04038	??????
GV1S4A1N1T	010515	V4.R'CL (λS6 in paper)	In Paper.	Complete	None	M13584	??????
GV1S4A1N2T	??????	601	In Paper.	Complete	Partial	M36285/X13354	??????
GV1S5	010512	Vg5'CL (B27 in paper)	In Paper.	Complete	None	M36286/X13355	??????
	010513	Vg5'CL (Vγ5 in paper)	In Paper. 0 aa/2 nuc - Unexplained: Y00482 has correct sequence	Complete	Complete	Y00482	??????
	010514	GC12-GH3'CL (GH3 in paper)	In Paper.	Complete	Complete	X15018/Y00814	??????
GV1S5P	015410 P	V5.GL'CL (λSH7 in paper)	In Paper.	Complete	None	M13431	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
GV1S6P	015411 P	V6.GL'CL (λSH7 in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	None	M13432	??????
GV1S7P	015412 P	V7.GL'CL (λSH7 in paper)	In Paper. Several aa/0 nuc diff - Differently translated	Complete	None	M13433	??????
GV1S8	010519	V8.GL'CL (λK20 in paper)	In Paper.	Complete	None	M13434	??????
	010520	pM17c64'CL	In Paper.	Complete	Complete	X06774	??????
	010521	Pg1'CL (Pγ1 in paper)	In Paper.	Complete-4	Complete	M30894	??????
GV2S1A1	010522	HGP02	In Paper.	Partial	Complete	M27335	??????
	010525	Vg9(A6)'CL (λA6 in paper)	In Paper.	Partial	None	X08086	??????
	010526	PBLC1.15	In Paper.	Complete	Partial	M16768	??????
	010529	IDP2.11	In Paper. 2 aa/3 nuc diff - Unexplained	Partial	Complete	M16804	??????
	??????	*No name*	In Paper.	Complete	Complete	X72500	??????
	010527	AB12'CL		Complete	Complete	??????	009779
	010528	F6C7'CL		Complete	Complete	??????	009769
GV2S1A2	010523	LKG20'CL (λKγ20 in paper)	In Paper.	Partial	Complete	X03436	??????
	010524	Vg9(K20)'CL		Partial	None	??????	??????
GV3S1P	??????	9-4	In Paper. 2 aa/0 nuc diff - Explained	Complete	Complete	S60779	??????
	015422 P	HUMAN Vg10-Jg2 (λR12 in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	NA	X05503	??????
	010721	HGP06	In Paper.	Partial	Complete	M27343	??????
	019510 P	TRGV10 (Vγ10 in paper)	In Paper. Many aa/0 nuc diff - Differently translated	Complete	NA	X74774	??????
	026351 P	H TCRG-V10		Complete	NA	??????	??????
GV4S1P	015424 P	HUMAN V11 (λJM15 in paper)	In Paper.	Complete-5	None	X07207	??????
	??????	5A7	In Paper.	Complete	Complete	S60780	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	019509 P	Vγ11	In Paper. Many aa/0 nuc diff - Differently translated	Complete	Joining Errors	X74775	??????
GV5S1P	015425 P	HUMAN VA (λA6 in paper)	In Paper.	Complete	None	X07208	??????
GV6S1P	015426 P	HUMAN VB (λJM15 in paper)	In Paper. Many aa/0 nuc diff- Differently translated	Complete	None	X07209	??????

Table B.6: The table shows the Kabat entry IDs of members of the officially designated classes for human γ chains [125]. A group of six question marks indicates data which is not known or not available. In the Kabat ID column there are also flags indicating whether the sequence has a known pair (X), whether it was used in the analysis (!), and if it is defined as a psuedogene in the Kabat database (P). Where sequences did not match the sequences or names of members of any of the existing classes, the closest matching class was identified and a temporary class created with a name ending in a question mark. For example if a sequence was similar to GV5S1 the temporary class would be called GV5S?.

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
GV1S1	010791	V3'CL (TC-13 in paper)	In Paper.	Complete	Partial	M13337	??????
	??????	FT6	In Paper.	Partial	None	X04396	??????
	??????	pAA42	In Paper.	Partial	None	X62546	??????
	025688	Vg5 B'CL (gB-V5 in paper)	In Paper. 3 extra aa	Complete	Partial?	Z48592	??????
GV2S1A1	010792	MOUSE V4'CL (TC-11 in paper)	In Paper.	Complete	Partial	M13338	??????
	??????	13	In Paper. Only 2 aa in M13339!	Fragment	None	M13339	??????
GV2S1A2	025689	Vg6 B'CL (gB-V6 in paper)	In Paper.	Complete	None	Z48593	??????
GV3S1A1	010788	MOUSE V2'CL (TC-17 in paper)	In Paper.	Complete	Partial	M13336	??????
	015839 P	3F9-GAMMA6 (3F9-γ6 in paper)	In Paper.	Partial	Joining Errors	X03984	??????
	010789	MNG1'CL	In Paper.	Complete	Complete		??????
	010790	MNG7'CL	In Paper.	Complete	Complete		??????
	??????	pAA21	In Paper.	Partial	None	X62545	??????
	?????? P	FT2	In Paper. 1 aa/1 nuc diff - Explained (Variant). Pseudogene	Complete	Joining Errors	X04315	??????
GV3S1A2	??????	G3	In Paper.	Complete	Joining Errors?	Z12299	??????
	??????	8.2	In Paper. Pseudogene	Complete	Joining Errors	Z22841	??????
	??????	pD17γ4	In Paper. 4 aa/8 nuc diff - Explained (Variant).	Complete-6	Joining Errors?	M26765	??????
GV3S1A3	025691 P	Vg4 B'CL (gB-V4 in paper)	In Paper.	Complete	None	Z48591	??????
GV4S1A1	??????	BW3.8.1	In Paper.	Complete	Joining Errors	X05501	??????
	025690	Vg7 B'CL (gB-V7 in paper)	In Paper.	Complete	None	Z48594	??????
GV4S1A2	??????	*No name*	In Paper.	Complete	None	M71214	??????
	??????	pAA22	In Paper.	Partial	None	X62544	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	025684	Vg7 A'CL (gA-V7 in paper)	In Paper.	Complete	None	Z49051	??????
GV5S1A1	010786	V108B'CL (V10.8B in paper)	In Paper.	Complete	None	M12832	??????
	??????	14.0	In Paper. 1 aa/3 nuc diff - Explained (Variant) (could be GA to AG swap)	Complete	Joining Errors	M26763	??????
GV5S1A2	??????	14.9	In Paper.	Complete	None	Z22847	??????
	??????	5/10-13γ1.2	In Paper. 2 aa/2 nuc diff - Explained (Variant)	Complete	Complete	X03802	??????
GV5S1A3	010785	TCRG-V1'CL (Vγ1 in paper)	In Paper. 3 extra aa	Complete	None	M77017	??????
GV5S1A4	025685	Vg1 B'CL (gB-V1 in paper) 0 aa/1 nuc diff - Unexplained	In Paper.	Complete	None	Z48588	??????
	??????	Vγ1	In Paper.	????????	????????		??????
GV5S2A1	010781	V108A (V10.8A in paper)	In Paper.	Complete	None	M12831	??????
	010780	MOUSE 2CA[pHDS4/pHDS203]'CL (pHDS4/203 in paper)	In Paper.	Complete	Complete	X00697	??????
	015835 P	pHDS34	In Paper. Many aa/0 nuc diff - Differently translated	Complete	Joining Errors	K02899	??????
	010782	DFL12	In Paper.	Complete	Complete	K02900	??????
	??????	FT12	In Paper.	Complete	Joining Errors	X04397	??????
	015836 P	MOUSE 8/10-2 (8/10-2γ1.1 in paper)	In Paper.	Complete	Joining Errors	X03801	??????
	010783	TCRG-V2'CL (Vγ2 in paper)	In Paper.	Complete	None	M77018	??????
	??????	10.4	In Paper.	Complete	Partial	Z22846	??????
	015837 P	MOUSE 3F9-GAMMA4 (3F9-γ4 in paper)	In Paper.	Complete	Joining Errors	X03983	??????

Family	Kabat ID	Clone name	Comments	V region	J Region	GENBank/EMBL accession number	PairCode
	015838 P	MOUSE 3F9-GAMMA7 (3F9-γ7 in paper)	In Paper. 0 aa/1 nuc diff - Explained (Variant)	Complete	Joining Errors	X03985	??????
	010784	MNG8	In Paper. 1 aa/3 nuc diff - Explained (Variant)	Complete	Complete		??????
GV5S2A2	025686	Vg2 B'CL (gB-V2 in paper)	In Paper.	Complete	None	Z48589	??????
	??????	Vγ2	In Paper.	????????	????????		??????
GV5S3A1	??????	4.5	In Paper.	Complete	Joining Errors	M26764	??????
	010936	V5.7'CL	In Paper.	Partial	None	M12833	??????
GV5S3A2	025687	Vg3 B'CL (gB-V3 in paper)	In Paper.	Complete	None	Z48590	??????
	??????	Vγ3	In Paper.	????????	????????		??????
GV5S?	010787	5/10-12'CL	Closest to GV5S1A1 (1 aa/2 nuc diff)	Complete	Complete	??????	??????
	010973	MNG9'CL	Closest to GV5S2A1 (4 aa/13 nuc diff)	Partial	Complete	??????	??????

Table B.7: The table shows the Kabat entry IDs of members of the officially designated classes for mouse γ chains [126]. A group of six question marks indicates data which is not known or not available. In the Kabat ID column there are also flags indicating whether the sequence has a known pair (X), whether it was used in the analysis (!), and if it is defined as a psuedogene in the Kabat database (P). Where sequences did not match the sequences or names of members of any of the existing classes, the closest matching class was identified and a temporary class created with a name ending in a question mark. For example if a sequence was similar to GV5S1 the temporary class would be called GV5S?.