Table of Contents

Chapter 2. Structural Studies of Two TCRs Specific for Human Tumour Antigens

Chapter 3. Analysis of TCR-pMHC Complexes

Chapter 5. Development of a Protein Disorder Prediction Tool

The Structural Basis of Immune Receptor Signalling

Rebecca Hamer
St Cross College
Division of Structural Biology

Submitted for the degree of Doctor of Philosophy in Michaelmas 2007

Abstract

This work investigates the mechanisms of binding of T cell receptors (TCRs) to Class I MHC-peptide complexes (pMHC). The structure of a TCR specific for the Melan-A tumour antigen bound to its cognate pMHC was solved to a resolution of 2.5 Å which gives insight into how this TCR could be mutated to optimize binding and subsequently used as a cancer vaccine. Detailed sequence and geometric analyses of all currently available structures of Class I TCR-pMHC complexes revealed that TCRs can bind to pMHC with a range of orientations, yet always focus on the central portion of the peptide and use a specific subset of six residues on the MHC helices for binding. The most striking finding was the use of aromatic residues in the TCR CDR loops to bind to residue Q155 on the MHC α2 helix.

Attempts were also made to express and purify Toll-like receptors (TLRs) with the aim of solving one or more of these structures. However, despite testing of over 50 different constructs from 12 different TLRs or associated proteins, insufficient soluble protein expression was obtained for crystallization trials.

Finally, a protein disorder prediction tool was developed to aid construct design for structural biology studies and improve the chances of obtaining protein crystals. This tool is based on a novel type of neural network and blind tests comparing it to 8 other disorder prediction tools showed it is one of the best in the field. It is freely available at www.strubi.ox.ac.uk/RONN. Analysis of large datasets revealed that the position of order/disorder transitions is quite precisely defined in amino-acid sequences and that transition regions have an amino acid composition distinct from that of bulk ordered and disordered sequences. There is a steady decrease in order-promoting residues on the ordered side of boundaries as well as a weak sequence signal, both of which signify the approaching disorder and may prove useful for improving existing disorder prediction tools.

Acknowledgements

I would like to thank my supervisor, Yvonne Jones, for the opportunity to work in her group on this project and for all her help and support, and the MRC for providing me with funding. Thanks also to Dawn Shepherd and Enzo Cerundolo at the Weatherall Institute of Molecular Medicine for provision of MHC and TCR molecules for crystallization trials and for their expertise in discussions of results.

I am indebted to Robert Esnouf for all his patience, help and support throughout. Many thanks also to Doryen Bubeck, without whom I would never have dared set foot in the lab and who spent many hours teaching me everything I needed to know about protein expression; Radu Aricescu for all his help in the lab and for making me laugh so much; Tom Walter for teaching me how to set up crystallization trials and for all his advice; Karl Harlos for his help in testing crystals in-house; and Christian Siebold for all his patience and help with X-ray data analysis.

Finally, thanks to my family and friends for all their support over the last three years which helped to keep everything in perspective - especially mum and dad for their unending support and interest in my work; Victoria and Phillip for their absolute refusal to talk about science which forced me into other topics of conversation and made me laugh; Laura and Cassi whos energy and ambition spurred me on and who shared in all the ups and downs; and most importantly of all, thanks to Tim, without whom none of this would matter.

Abbreviations

APC	Antigen Presenting Cell
ASU	Asymmetric Unit
BBFNN	Bio-Basis Function Neural Network
BLAST	Basic Local Alignment Tool
BSA	Buried Surface Area
CDR	Complementarity Determining Region
CHO cells	Chinese Hamster Ovary cells
CTL	Cytolytic T Cell
DC	Dendritic cell
DMEM	Dulbecco's Modified Eagles Medium
DTT	Dithiothreitol
EBI	European Bioinformatics Institute
EDTA	Ethylenediamine Tetra-Acetic Acid
ER	Endoplasmic Reticulum
ESRF	European Synchrotron Radiation Facility
EST	Expressed Sequence Tag
FCS	Foetal Calf Serum
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HLA	Human Leukocyte Antigen
Ig	Immunoglobulin
IPTG	Isopropylthio-β-D-galactoside
LB	Luria-Bertani Media

LPS	Lipopolysaccharide
MHC	Major Histocompatibility Complex
MSD	Macromolecular Structure Database
NCBI	National Center for Biotechnology Information
NK	Cells Natural Killer Cells
NMR	Nuclear Magnetic Resonance Spectroscopy
ORF	Open Reading Frame
PAMP	Pathogen-Associated Molecular Pattern
PBL	Peripheral Blood Lymphocyte
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEG	Polyethylene Glycol
PEI	Polyethylenimine
pMHC	Peptide-MHC complex
PRRs	Pattern Recognition Receptors
RONN	Regional Order Neural Network
SDS PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
SFM	*Spodoptera frugiperda* Media
TAP	Transporter Associated with Antigen Processing
TCR	T Cell Receptor
TIL	Tumour Infiltrating Lymphocyte
TLR	Toll-Like Receptor

Chapter 1

Introduction

1 Introduction

1.1 Summary

This thesis reports structural studies and analysis of interactions in the immune system, as well as the development of an *in silico* tool for the prediction of protein disorder.

Chapter one introduces pertinent aspects of the immune system and of protein disorder. A summary of crystallographic theory is included in Appendix A.

1.2 The Immune System

The immune system comprises the cells and molecules which respond to foreign substances. It can be divided into the innate and the adaptive immune systems, although it is now recognised that these two components are intricately linked, with the innate immune system dictating the nature of the adaptive immune response (Fearon and Locksley 1996; Medzhitov and Janeway 1997).

The innate immune system is a non-specific first line of defence against microbes, comprised of physical and chemical barriers (epithelia and antimicrobial substances produced at epithelial surfaces), phagocytic cells (neutrophils and macrophages), NK cells, blood proteins including the complement system and cytokines.

Innate immune recognition relies on a limited number of germline-encoded receptors called pattern recognition receptors (PRRs), such as Toll-like receptors (TLRs). These are found on many cells including macrophages, dendritic cells (DCs), mast cells, neutrophils, eosinophils and NK cells and recognise conserved microbial products called pathogen-associated molecular patterns (PAMPs), for example lipopolysaccharide (LPS) which is a component of gram negative bacterial cell walls (Janeway, Carding et al., 1988; Medzhitov and Janeway 1997).

An immune system based only on innate immunity *i.e.* recognition of PAMPs has some severe limitations: there can only be a limited number of PRRs as they are germline encoded and not clonally distributed, so any one cell can only display a limited number of PRRs. In addition, any new microbial PAMP requires the development of a new PRR, and this system does not allow development of immunological memory (Janeway, Carding et al., 1988).

The adaptive immune system is present in vertebrates only and provides antigen-specific recognition of both microbes and non-microbial foreign substances. Unlike the innate immune system, it has the property of immunological memory whereby exposure to a foreign antigen enhances the ability of the adaptive immune system to respond to that antigen again, usually with a larger and more rapid response.

There are two types of adaptive immune responses - humoral and cellular. Humoral immunity is mediated by antibodies produced by B cells, which can be membrane bound

or soluble. Membrane bound antibodies are the B cell antigen receptor molecules whereas soluble antibodies are found in mucosal secretions and circulate in the blood and bind to extracellular (including membrane-bound) antigens. They can neutralise toxins, promote phagocytosis and trigger the release of inflammatory mediators from leukocytes such as mast cells.

However, intracellular bacteria and viruses are inaccessible to circulating antibodies. Cell mediated immunity therefore allows the destruction of infected cells and hence the removal of reservoirs of infection. The T cells which comprise the cellular immune system have membrane-bound antigen receptors called T cell receptors (TCRs) which recognise antigen presented by infected cells or antigen presenting cells (APCs) in the context of Major Histocompatibility Complex (MHC) molecules, but do not recognise soluble antigen (Zinkernagel and Doherty 1974; Zinkernagel and Doherty 1974; Zinkernagel and Doherty 1997)

αβ T cells can be classified into two major types: CD4+ helper T cells secrete cytokines in response to antigenic stimulation. These stimulate proliferation and differentiation of T cells, B cells, macrophages and other leukocytes. CD8+ cytolytic T cells kill cells presenting foreign antigen.

The basic principles of immunology can be found in Cellular and Molecular Immunology (Abbas 2003).

1.2.1 Antigen Presentation to T Cells of the Adaptive Immune System

There are two classes of MHC molecules which present antigen to the two subsets of T cells. All the MHC genes are located in the MHC loci on chromosome 6 in humans and chromosome 17 in mice. Class I genes consist of Human Leukocyte Antigen (HLA) -A, -B and –C in humans and H-2K, -2D and -2L in mice and class II genes consist of HLA-DR, -DP and –DQ in humans and I-A and I-E in mice (Trowsdale 1995).

Both types of MHC comprise an extracellular peptide-binding groove made up of two α helices with a β-sheet floor, and two immunoglobulin (Ig)-like domains (Bjorkman, Saper et al., 1987; Brown, Jardetzky et al., 1993). Many residues within the peptide binding groove are polymorphic, allowing MHC molecules to bind to a wide range of peptides (Bjorkman, Saper et al., 1987). Molecules encoded by a single MHC allele can bind to 1000-2000 different peptides (Hunt, Henderson et al., 1992; Hunt, Michel et al., 1992).

MHC class I molecules present fragments of intracellular antigens to CD8+ cytotoxic T cells. They consist of an α chain non-covalently associated with a non-MHC-encoded protein called β2-microglobulin (Figure 1). The α chain is divided into three domains, α1, α2 and α3. The carboxy end of the α3 domain has a stretch of approximately 25 hydrophobic amino acids which insert into the cell membrane, followed by about 30 cytoplasmic residues. The peptide binding groove of class I molecules has closed ends, meaning only short peptides of typically 8-11 residues long can be accommodated (Madden, Gorga et al., 1991; Hunt, Henderson et al., 1992; Madden, Gorga et al., 1992).

Six pockets in the peptide binding groove, named A-F, which determine peptide specificity were defined based on the X-ray structure of HLA-A2 (Saper, Bjorkman et al., 1991).

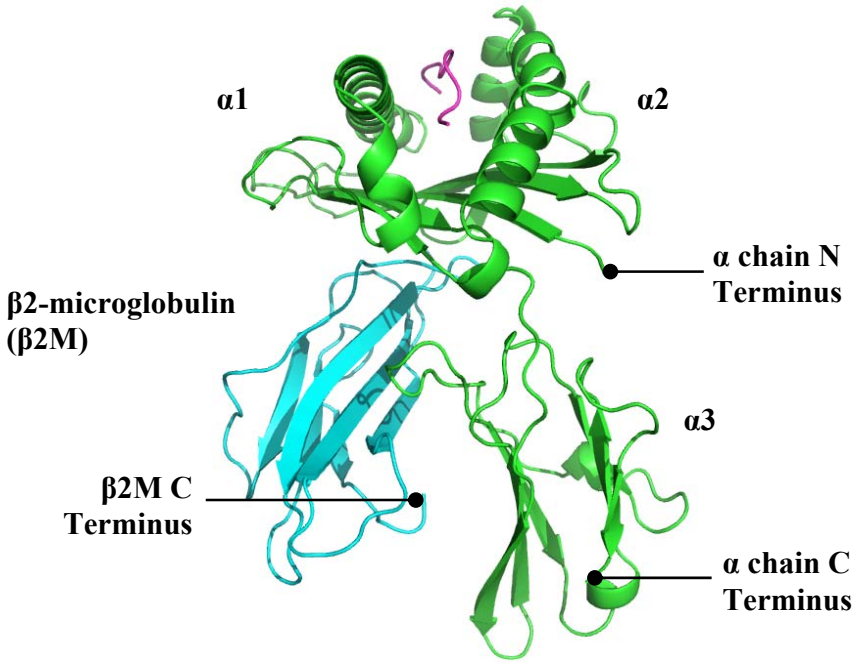


Figure 1. HLA-A2 Class I MHC molecule with tax peptide (coloured magenta) bound in the groove (PDB identifier 1A07). The α chain is coloured green and β2-microgobulin is coloured cyan. This figure was created using Pymol (Delano 2002) as are all structural figures in this thesis.

MHC class II molecules present fragments of extracellular antigens to CD4+ helper T cells and are composed of an α chain and a β chain which are non-covalently associated (Figure 2). The carboxy termini of both these chains have approximately 25-residue long hydrophobic transmembrane regions followed by short cytoplasmic tails. The ends of the peptide-binding groove are open meaning class II molecules can bind longer peptides of 10-30 amino acids (Rudensky, Preston-Hurlburt et al., 1991; Hunt, Michel et al., 1992).

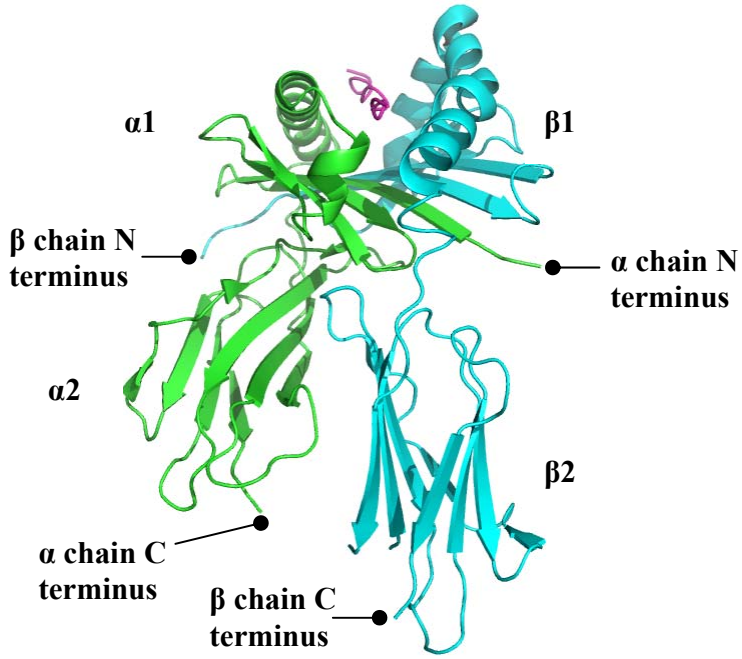


Figure 2. Class II MHC molecule with a peptide (coloured magenta) bound in the groove (PDB identifier 1FYT). The α chain is coloured green and the β chain is coloured cyan.

MHC class I molecules are found on the surface of virtually all nucleated cells whereas MHC class II molecules are expressed only on antigen presenting cells such as B cells, dendritic cells and macrophages. This pattern of MHC expression is closely linked to the functions of the T cells to which they present antigen. MHC class I molecules present antigens derived from intracellular pathogens including viruses. These are recognised by CD8+ T cells which then kill any infected cells. As viruses can infect virtually any nucleated cell, MHC class I molecules must be expressed on all such cells. Class II molecules however, present extracellularly-derived antigens to CD4+ T cells. Naïve CD4+ T cells are only activated when they bind to antigen presented by dendritic cells. Mature CD4+ T cells recognise antigen presented on macrophages that have phagocytosed extracellular microbes and activate them to destroy the pathogen. CD4+ T

cells also activate B cells to produce antibodies that aid elimination of extracellular pathogens.

Most nucleated cells express between 10^4 and 10^6 surface MHC class I molecules and cells which express MHC class II molecules express a similar number, but only 10-100 pMHC complexes per cell are required to evoke effector responses from memory or effector T cells (Germain and Stefanova 1999).

If T cells recognised soluble antigen, they may be prevented from performing their functions of killing infected cells, macrophage activation and B cell activation. In addition, if T cells could bind to antigen that was not in the context of MHC, they would only be capable of recognising those components of bacteria or viruses which were displayed on the cell surface. The mechanism by which peptides are delivered to MHC on the cell surface means many more antigens from infectious agents will be displayed, therefore giving more T cells the opportunity to recognise the infection.

1.2.2 Antigen processing

Class I-associated peptides are generated by the proteolytic degradation of cytosolic proteins which are synthesised within the cell. These may be the products of viruses or intracellular microbes but may also be normal self proteins and are degraded by a large multiprotein enzyme complex called the proteasome (Yewdell and Bennink 2001). The peptides generated are transported from the cytosol into the endoplasmic reticulum (ER) through the transporter associated with antigen processing (TAP), where they associate with newly-synthesised MHC class I molecules and peptide-MHC complexes are transported to the cell surface in exocytic vesicles (Figure 3).

Figure 3. Schematic representation of Class I and Class II antigen processing and presentation pathways. Taken from Cellular and Molecular Immunology (Abbas 2003).

Most class II-associated peptides are derived from extracellular antigens which bind to receptors on antigen presenting cells (APCs) and are internalised by them. Protein antigens are localised in intracellular vesicles called endosomes which contain proteolytic enzymes, whereas microbes are internalised into phagosomes which can fuse with lysosomes and kill the microbe. Within these endocytic vesicles, antigenic proteins are degraded into peptides. Class II MHC molecules are synthesised in the ER but cannot bind peptides within the ER due to the presence of the invariant chain bound in the cleft (Roche and Cresswell 1990; Teyton, O'Sullivan et al., 1990). Exocytic vesicles transport class II molecules out of the ER and fuse with the endocytic vesicles. The invariant chain is then removed from the class II molecules by proteolytic enzymes enabling antigenic peptides to bind in the cleft, and the peptide-MHC complex is transported to the cell surface (Figure 3). Class II antigen processing is reviewed by Trombetta and Mellman (Trombetta and Mellman 2005).

Expression of both class I and class II molecules on the cell surface requires all components of the MHC molecule to be present, as well as a bound peptide (Germain 1994; Trombetta and Mellman 2005).

1.2.3 Antigen Recognition

1.2.3.1 Structure of TCRs

The TCR is a heterodimer consisting of one α and one β chain covalently linked by disulphide bonds, each with a short transmembrane region and a 5-12 residue carboxy terminal cytoplasmic tail. The extracellular domain is structurally analogous to the antigen binding fragment (Fab) of an antibody with each chain consisting of an N-terminal variable Ig-like domain and a C-terminal constant Ig-like domain (Figure 4). The variable domains of the TCR, Vα and Vβ, make up the antigen binding site. Within each V region are 3 hypervariable loops called the complementarity determining regions (CDRs) where the variability between TCRs is concentrated.

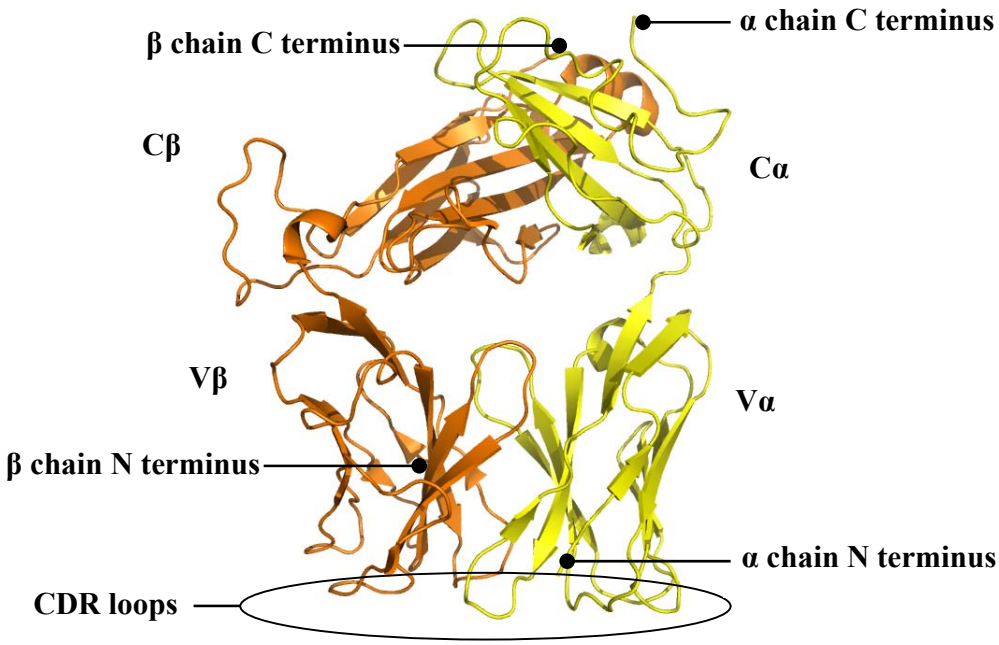


Figure 4. Structure of the 2JM22 TCR (PDB identifier 1OGA) which recognizes a flu matrix peptide bound to HLA-A2. The TCR α chain is shown in yellow and the β chain in orange.

The TCR only functions as part of a complex consisting of the αβ TCR non-covalently linked to the CD3 γ, δ, ε and ζ proteins. The αβ TCR mediates antigen recognition whereas the other components of the complex are critical for TCR signalling. The TCR complex is assembled in the ER during T cell maturation and is only expressed on the cell surface if all components are present (Sussman, Bonifacino et al., 1988; Geisler 1992; Dave, Cao et al., 1997; Haks, Krimpenfort et al., 1998).

CD4 and CD8 are transmembrane glycoproteins associated with the T-cell specific tyrosine kinase $p56^{lck}$ (lck). During T cell activation, the coreceptor must bind to the same MHC molecule as the TCR for optimal signal transduction, and physical association of CD4/CD8 with the TCR is required for fully efficient T cell activation. CD4 and CD8 potentiate the activation signals delivered through the TCR by causing a 30-3000-fold reduction in the ligand density required for T cell activation in different systems. They also appear to be essential for T cell maturation and for activation by the first encounter by antigen (Janeway, Carding et al., 1988).

The first structure of a human TCR was solved in complex with an MHC class I molecule in 1996 (PDB accession 1AO7 (Garboczi, Ghosh et al., 1996)) (Figure 5).

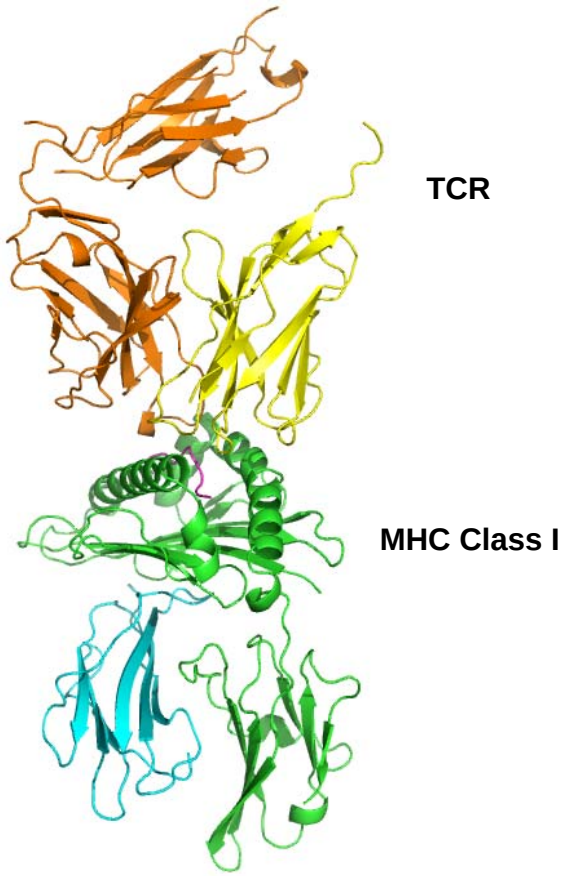


Figure 5. Structure of a class I TCR-pMHC complex (PDB identifier 1A07). The MHC α chain is shown in green, β2-microglobulin in cyan, the TCR α chain in yellow and the β chain in orange. Note that the entire α constant domain of the TCR is missing from this structure.

The structures of CD4 alone (Wang, Yan et al., 1990; Wu, Kwong et al., 1997), CD4 complexed to MHC class II (Wang, Meijers et al., 2001), CD8 alone (Leahy, Axel et al., 1992) and CD8 bound to MHC class I (Gao, Tormo et al., 1997; Kern, Teng et al., 1998; Liu, Xiong et al., 2003) have been determined, and CD8 is known to bind to non-polymorphic residues of the α3 domain of MHC class-I molecules. The structures of mouse CD3εγ (Kim, Sun et al., 2000), human CD3εγ (Kjer-Nielsen, Dunstone et al., 2004) and human CD3εδ (Arnett, Harrison et al., 2004) have also been solved and a model has been proposed for the TCR/CD3εγ/CD3εδ complex (Arnett, Harrison et al., 2004) and extended to include CD8 (Rudolph, Stanfield et al., 2006) (Figure 6).

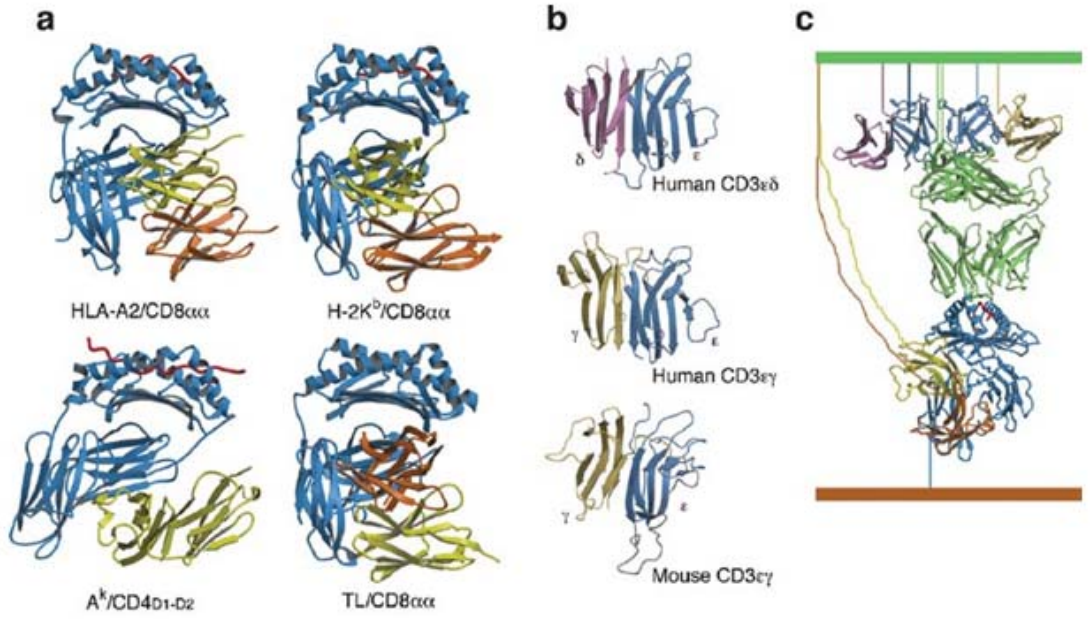


Figure 6. Taken from (Rudolph, Stanfield et al., 2006). CD8 and CD4 coreceptor binding to class I and class II MHC. (*a*) The MHC is colored blue, CD8αα in yellow and orange. It is not yet known which of the domains of CD8αα homodimer correspond to the CD8αβ heterodimer (*upper* and *lower right panel*). The MHC is colored blue, and the CD4 (two N-terminal domains) is colored in yellow (*lower left panel*). (*b*) Human CD3εδ (*top*), human CD3εγ (*middle*), and mouse CD3εγ (*bottom*). In all three panels, the common ε-chain is colored in blue. (*c*) Hypothetical TCR/pMHC/CD3εδ/CD3εγ/CD8 complex. The TCR/pMHC-CD8αα and putative CD8αβ interaction is modeled by superimposing two structures, the HLA-A2/CD8αα complex (1akj) and the TCR A6/HLA-A2/TaxP6A complex (1qrn) on their MHC residues α1–180, with TCR (*green*), MHC (*dark blue*), peptide (*red*) and CD8 (*yellow* and *orange*). The CD3εδ (1xiw, *pink* and *blue*) and CD3εγ (1sy6, *gold* and *blue*) are shown "docked" at the top of the figure, with the common ε-chains colored in blue. This "docking" merely represents placing of the CD3 structures in the vicinity of where they are thought to bind. Lines are drawn in to depict tethers connecting the different subunits to the TCR cell membrane (*top, green*) or the antigen-presenting cell membrane (*brown, bottom*).

1.2.3.2 Generation of TCR diversity

The genes encoding the TCR α and β chains are located at different loci but both the TCR α and β loci include variable (V), joining (J) and constant (C) gene segments, while the β locus also has diversity (D) segments (Arden, Clark et al., 1995; Rowen, Koop et al., 1996). Somatic recombination brings these gene segments together to form a V(D)J gene that codes for the variable region of the TCR. The primary RNA transcript is then

processed to join the V(D)J segment to a C gene segment, forming the mRNA that codes for the entire α or β chain of the TCR. In addition to the diversity generated by the joining of different combinations of V, D and J segments, non-germline 'N' nucleotides are added to or removed from the joints between these segments (junctional diversity). This means that TCRs have the greatest variability at the third complementarity determining region (CDR3), which is formed at the site of V(D)J recombination, with the CDR3β showing most diversity due to the presence of the D segment. Finally, different pairings of α and β chains maximise TCR diversity to give an estimated 10^8 different αβ combinations (Davis and Bjorkman 1988; Arstila, Casrouge et al., 1999; Arstila, Casrouge et al., 2000).

β chain gene rearrangement occurs first. When a functional β chain is expressed on the cell surface in association with an invariant protein called pre-Tα, further rearrangement of the β chain locus is inhibited. This means mature T cells express only one of the two inherited β chain alleles (allelic exclusion). α chain rearrangement then commences, but there is no allelic exclusion meaning α chain rearrangements can occur on both chromosomes. Up to 30% of mature peripheral T cells express 2 different TCRs with different α chains but the same β chain (Casanova, Romero et al., 1991; Padovan, Casorati et al., 1993).

1.2.3.3 Positive and negative selection of T cells

In the bone marrow, immature thymocytes do not express TCRs or the coreceptors CD4 or CD8. They traffic through the blood to the thymus where VDJ rearrangements occur at the β locus and the cells subsequently express both CD4 and CD8 to become double positive thymocytes. Next, α locus recombination occurs and the cells become αβTCR+CD4+CD8+ thymocytes with randomly generated specificities. At this stage, double positive thymocytes expressing functional αβ TCRs which bind to self peptide-MHC with high affinity/avidity undergo negative selection and are deleted by induction of apoptosis. Conversely, thymocytes possessing TCRs which bind with low affinity/avidity are stimulated to survive and mature to single positive thymocytes. If self peptide-self MHC is not recognised at all, the thymocytes die by apoptosis (Starr, Jameson et al., 2003).

The class I or class II MHC restriction of the T cell is fixed by positive selection. For example, if the TCR recognises class I peptide-loaded MHC (pMHC) on a cell and at the same time CD8 interacts with the class I molecules, the T cell will be stimulated to mature into a class I MHC-restricted CD8+ T cell. It is unclear whether there is subsequent random loss of either CD4 or CD8 and only those cells left with the 'correct' coreceptor (CD4 on a class II-restricted cell and CD8 on a class I-restricted cell) will continue to mature, or whether expression of the 'incorrect' coreceptor is actively suppressed.

1.3 Tumour Immunity

Many human tumour-associated antigens which are recognised by cytotoxic T cells (CTL) have been identified (Robbins and Kawakami 1996), showing that an immune response to cancer in humans does exist and raising the possibility of developing clinical cancer vaccines.

One potential vaccine candidate is the NYESO-1 peptide which is expressed by a broad range of tumour types (Chen, Scanlan et al., 1997). It is expressed only in normal testis and ovary but not found in normal adult somatic tissue, and therefore has potential for a vaccine designed to boost CTL responses against many tumours.

Tumour reactive CTLs mainly recognise the peptide consisting of residues 157-165 of NYESO (SLLMWITQC; NYESO9C) (Valmori, Dutoit et al., 2000). However, this peptide binds poorly to HLA-A2, so analogues were generated where the C-terminal cysteine was replaced with either leucine or valine (NYESO9L and NYESO9V) which act as better anchor residues. These analogues do bind to HLA-A2 more efficiently and are recognised by CTL raised against the NYESO9C peptide, but do not themselves induce tumour-specific CTL. In fact, NYESO9L was recognised less efficiently by CTL than NYESO9C (Webb, Dunstone et al., 2004).

The structures of NYESO9C, NYESO9A and NYESO9S bound to HLA-A2 have been solved (Webb, Dunstone et al., 2004) (PDB identifiers 1S9W, 1S9X and 1S9Y

respectively) as well as the structures of a TCR bound to HLA-A2-NYESO9C and to HLA-A2-NYESO9V (Chen, Stewart-Jones et al., 2005) (PDB identifiers 2BNR and 2BNQ respectively). To understand in more detail how TCRs recognise the wild-type antigen, the structure of a different TCR bound to it would be advantageous.

A second potential vaccine target is the melanoma antigen MART-1 (melanoma antigen recognised by T cells) (Kawakami, Eliyahu et al., 1994) or Melan-A (Coulie, Brichard et al., 1994). This is a non-mutated differentiation antigen expressed by normal melanocytes which is recognised by HLA-A2-restricted CTL. HLA-A2 is present in approximately 50% of individuals, up to 1 in 1000 naïve CTL in HLA-$A2^+$ individuals recognise Melan-A (Dutoit, Rubio-Godoy et al., 2002) and the Melan-A antigen is widely expressed on melanomas, making it a good vaccine candidate.

The reason for the large population of naïve, Melan-A-specific CTL in HLA-$A2^+$ individuals is not entirely clear. Large numbers of Melan-A-specific CTL are found in the thymus but the ligand(s) involved in their selection have not been identified (Dutoit, Rubio-Godoy et al., 2002). No Melan-A mRNA expression has been found in the thymus (Coulie, Brichard et al., 1994), but there is the possibility that unknown Melan-A cross-reactive peptides are expressed. The fact that the population of Melan-A-specific T cells in the periphery of healthy HLA-$A2^+$ individuals are naïve, in spite of expression of Melan-A in melanocytes, suggests that tolerance is due to immunological ignorance (Ochsenbein, Klenerman et al., 1999). These CTL can be activated by the presence of melanoma tumours and concomitant inflammation, if sufficient tumour cells reach

draining lymph nodes. However, one study found that although Melan-A-specific T cells circulating in peripheral blood displayed cytotoxic activity, those in melanoma tumour lesions were functionally tolerant (Dutoit, Rubio-Godoy et al., 2002).

The dominantly-recognised epitopes of the Melan-A antigen were found to be a nonapeptide corresponding to residues 27-35 (AAGIGILTV) and a decapeptide corresponding to residues 26-35 (EAAGIGILTV) (Kawakami, Eliyahu et al., 1994). The decapeptide was found to bind better to HLA-A2 and be recognised more efficiently by CTL. However, both peptides form relatively unstable complexes with HLA-A2 and the decapeptide was shown to be no more immunogenic than the nonapeptide (Romero, Gervois et al., 1997). Most peptides which bind to HLA-A2 contain 2 key anchor residues: leucine or methionine in position 2 and valine at position 9 (Falk, Rotzschke et al., 1991), but both the nonapeptide and decapeptide from Melan-A lack the anchor at position 2. Amino acid substitutions were introduced into these peptides with the aim of improving binding to HLA-A2 and recognition by CTL. Some analogues of the nonapeptide bound more efficiently to HLA-A2, but were poorly recognised by CTL. However, one analogue of the decapeptide, ELAGIGILTV (ELA decapeptide), not only bound more stably to HLA-A2 but was also recognised more efficiently by CTL than the natural decapeptide. CTL that recognised the ELA decapeptide largely cross-recognised the wild type nonapeptide and decapeptide (Valmori, Fonteneau et al., 1998; Valmori, Fonteneau et al., 1999).

In 2002 a study involving the adoptive transfer of autologous Melan-A-specific tumour-infiltrating lymphocytes (TILs) into patients with metastatic melanoma led to regression of the melanoma in some patients (Dudley, Wunderlich et al., 2002). Although 5 patients with cancer regression also experienced some autoimmune melanocyte destruction, this work demonstrated that normally expressed self antigens can be useful targets for immunotherapy. In 2006, peripheral blood lymphocytes (PBLs) from patients with melanoma were transduced with a retroviral vector encoding a Melan-A_{27-35}-specific TCR (Morgan, Dudley et al., 2006). Autologous transduced cells were then adoptively transferred into 15 patients, 2 of whom showed full clinical regression of metastatic melanoma lesions. A different approach saw 40 patients vaccinated with the Melan-A wild type or modified decapeptide with adjuvant, but *ex vivo* detectable T cell activation was only found in 7 patients. After vaccination 5 patients had increased frequencies of CTL that bound to tetramers loaded with the ELA decpeptide (Valmori, Dutoit et al., 2000). A subsequent trial on 3 patients vaccinated with the ELA decapeptide showed activation of T cells specific for the wild type tumour antigen (Ayyoub, Zippelius et al., 2003). Clearly this antigen offers great vaccine potential, but improving TCR recognition and binding to it may improve clinical results.

1.4 Thesis Aim 1

One major aim of my thesis work has been to analyse in detail the interaction between TCRs and pMHC and elucidate any binding characteristics. Currently, no definitive 'rules of engagement' have been found that can shed light on why TCRs only bind to pMHC and what determines their specificity. If this can be understood it paves the way for new treatments for diseases such as cancer, as 'optimal' TCRs can be designed for binding to specific antigens. Further to this general analysis I therefore pursued a specific focus on TCR-pMHC interactions of tumour antigens.

1.5 Toll-Like Receptors of the Innate Immune System

A mutant *Drosophila* was described in 1984 and named Toll, meaning 'weird, mad or cool', due to its unusual appearance caused by a mutation in the Toll gene (Anderson and Nusslein-Volhard 1984; Anderson, Jurgens et al., 1985) which was later found to be responsible for dorsal-ventral patterning (Anderson, Bokla et al., 1985).

It was then found that the promoters of the gene for the *Drosophila* antibacterial peptide diptericin contained sequences identical or similar to the mammalian NF-κB and NF-IL6 binding motifs and an interleukin response element (IL-6RE) (Reichhart, Meister et al., 1992), which prompted comparison of insect immunity and the mammalian acute phase response (Ip and Levine 1994; Hoffmann 1995).

Mutations in the Toll signalling pathway were then discovered to cause a dramatic reduction in survival rates in *Drosophila* after fungal or gram positive bacterial infection. It was found that fungal or bacterial infection in *Drosophila* induce signalling via the Toll and immune deficiency (Imd) pathways, which activate the NF-κB homologues Dif, dorsal and relish and lead to production of antimicrobial peptides (Lemaitre, Nicolas et al., 1996). However, *Drosophila* Toll does not directly recognise PAMPs but rather acts as a receptor for spatzle which is a cytokine-like molecule (Levashina, Langley et al., 1999).

Similarity between the cytoplasmic domain of Toll and human interleukin-1 receptor (IL-1R) was discovered in 1991, suggesting that both molecules may be involved in similar pathways (Gay and Keith 1991). However, the first sequence homologous to full-length Toll was not discovered until several years later when it was named Toll/interleukin-1 receptor-like (TIL) and mapped to human chromosome 4 (Taguchi, Mitcham et al., 1996). It is now known as Toll-like receptor 1 (TLR1). Another homologous sequence was later found by Medzhitov *et al.,* by searching the Expressed Sequence Tag (EST) database at the National Center for Biotechnology Information (NCBI) using a sequence profile of the Toll/IL-1R signalling domain. Constitutive expression of this 841-amino acid protein in transfected human cells was shown to cause the activation of NF-κB, and expression of NF-κB-controlled genes for the inflammatory cytokines IL-1, IL-6 and IL-8, as well as expression of the T-cell costimulatory molecule B7.1 (Medzhitov, Preston-Hurlburt et al., 1997).

The NCBI EST database was searched again in 1998 and 5 human Toll-like receptors were found and named TLRs 1-5 (Zernich, Purcell et al., 2004), where the sequence of TLR4 was identical to that identified by Medzhitov *et al.,* above (Medzhitov, Preston-Hurlburt et al., 1997).

TLR6 was discovered in 1999 by the discovery of a murine EST clone resembling human TLR1. A human placenta cDNA library was then screened with the murine TLR6 as a probe to isolate human TLR6 (Takeuchi, Kawai et al., 1999).

TLRs 7, 8 and 9 were then found in 2000 (Chuang and Ulevitch 2000; Valmori, Dutoit et al., 2000) and TLR10 in 2001 (Chuang and Ulevitch 2001).

The completion of the human and mouse genome sequences allowed further searches for TLRs and TLR11 was found. This TLR is expressed only in mice as human TLR11 appears to have stop codons present in the open reading frame (ORF). However, the authors note that this could be a form of genetic polymorphism so further analysis of human TLR11 sequences is needed (Zhang, Zhang et al., 2004).

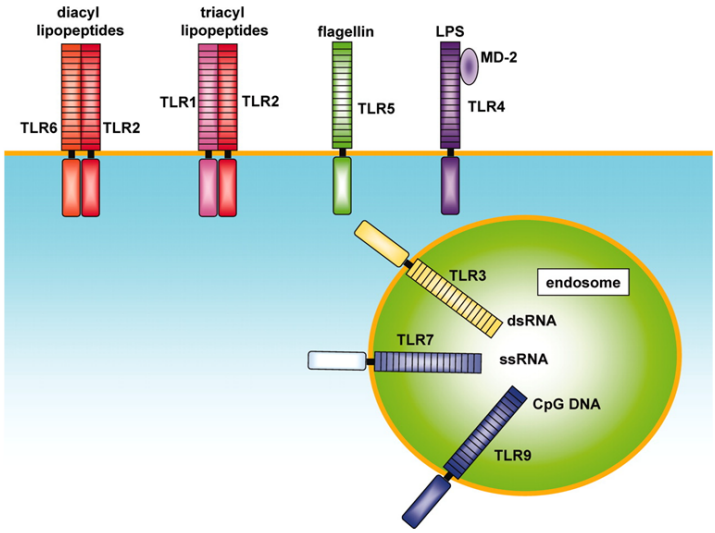


Figure 7. The location of TLRs in the cell and some of their ligands. This figure is taken from a review on TLRs (Takeda and Akira 2005). TLR8 is also located inside the endosome. The TIR domains are shown as unpatterned boxes and the leucine-rich repeats as striped boxes.

The location of TLRs in the cell is shown in Figure 7. TLR1, 2, 4, 5 and 6 are expressed on the cell surface, and TLR2 has been shown to functionally associate with both TLR1

and TLR6 in mice (Takeuchi, Kawai et al., 2001; Takeuchi, Sato et al., 2002). TLR3, 7, 8 and 9 are expressed in intracellular compartments such as endosomes. These cellular localisations reflect the different ligands that the TLRs recognise (Table 1).

Table 1. Confirmed natural ligands for TLRs (adapted from a review on TLR signalling (Akira and Takeda 2004))

Receptor	Confirmed Ligand	Source of ligand
TLR1	Triacyl lipopeptides	Bacteria and mycobacteria
TLR2	Lipoproteins/lipopeptides Peptidoglycan Lipoteichoic acid Lipoarabinomannan Phenol-soluble modulin Glycoinositolphospholipids Glycolipids Porins Atypical lipopolysaccharide Zymosan	Various pathogens Gram positive bacteria Gram positive bacteria Mycobacteria Staphylococcus epidermidis Trypanosoma cruzi Treponema maltophilum Neisseria Leptospira interrogans, Porphyromonas gingivalis Fungi
TLR3	dsRNA	Viruses
TLR4	Lipopolysaccharide (LPS) Taxol	Gram negative bacteria Plants
TLR5	Flagellin	Bacteria
TLR6	Diacyl lipopeptides Lipoteichoic acid Zymosan	Mycoplasma Gram positive bacteria Fungi
TLR7	ssRNA (mouse only)	Viruses
TLR8	ssRNA (human only)	Viruses
TLR9	CpG-containing DNA	Bacteria and viruses
TLR10	unknown	
TLR11	Profilin(Yarovinsky, Zhang et al., 2005) (mouse only)	Toxoplasma gondii

All TLRs found so far have amino-terminal leucine-rich repeats which are responsible for PAMP recognition, and a carboxy-terminal Toll-interleukin-1 (IL-1) receptor (TIR) domain required for intracellular signalling. The TIR domains of TLR1 and TLR2 have been solved (Xu, Tao et al., 2000) (Figure 8), but the only structure of a TLR that has been published so far is that of the extracelluar domain of TLR3 which was solved simultaneously by two groups in 2005 (PDB identifiers 1ZIW and 2A0Z) (Bell, Botos et al., 2005; Choe, Kelker et al., 2005) (Figure 9).

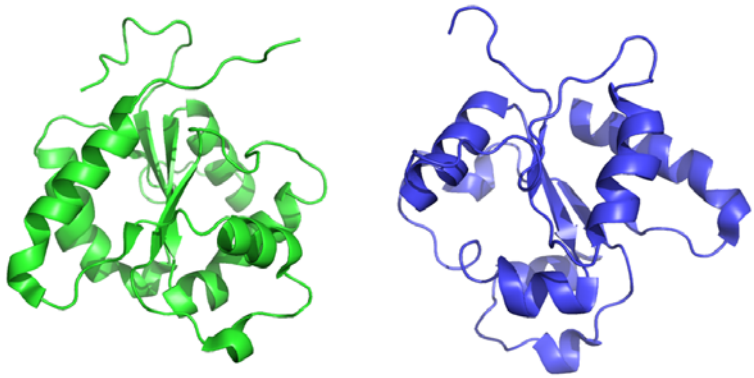


Figure 8. TIR domains of TLR1 (green) and TLR2 (blue) which are responsible for TLR signalling (PDB identifiers 1FYV and 1FYW respectively).

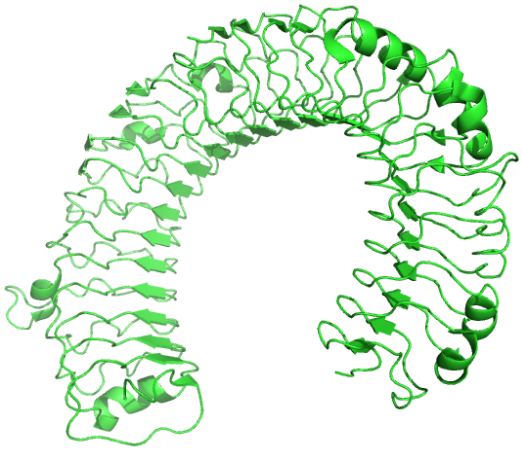


Figure 9. The structure of the ligand-binding domain of TLR3 (PDB identifier 1ZIW). The structure is a horseshoe-shaped solenoid containing 23 leucine-rich repeats. For clarity, glycosylation is not shown.

1.6 Toll-Like Receptors Provide a Crucial Link Between the Innate and Adaptive Immune Systems

Naïve T cells need two signals to be activated: the first signal is a result of TCR recognition of antigen bound to self-MHC on the APC. The second signal is delivered as a consequence of binding of CD28 on the same T cell to the costimulatory molecules CD80 (B7-1) and/or CD86 (B7-2) on the APC (June, Bluestone et al., 1994). This second signal enhances T cell survival, induces production of cytokines such as IL-2 and promotes differentiation of naïve T cells into effector and memory cells. CD28 ligation has also been shown to reduce the number of TCRs on a T cell that need to be triggered to reach the threshold for T cell activation (Viola and Lanzavecchia 1996). This is thought to be achieved by CD28 recruitment of lipid rafts to the site of interaction between T cells and APCs (the immunological synapse), so increasing the local concentration of kinases and segregating phosphorylated substrates from phosphatases (Viola, Schroeder et al., 1999).

Dendritic cells (DCs) are one of the most important types of APC. Immature DCs reside in the peripheral tissues where they actively sample their environment by endocytosis and macropinocytosis. They have high levels of expression of most TLRs so are efficient at recognising pathogens. The DC maturation induced by TLR recognition of PAMPs includes the induction of CD80 and CD86 by the transcription factor NF-κB, production of inflammatory cytokines such as IL-12 and migration of DCs to lymph nodes where they can prime naïve antigen-specific T cells. In addition there is increased expression of

MHC molecules therefore increased antigen presenting ability (Banchereau and Steinman 1998; Reis e Sousa 2001).

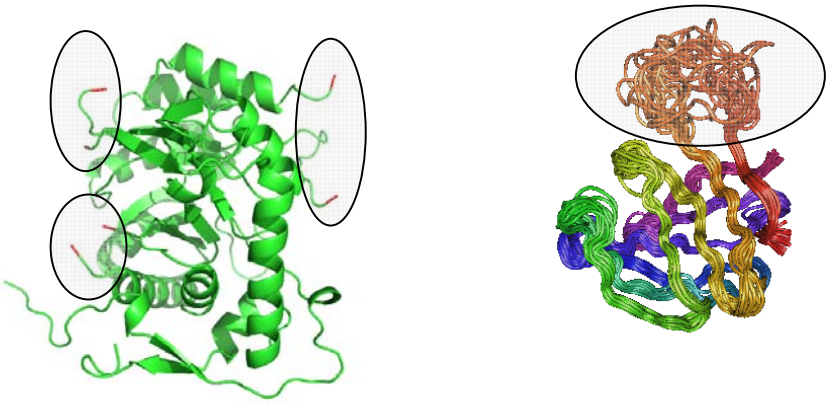
Microbial recognition by TLRs therefore provides a critical link between the innate and adaptive immune systems by stimulating DC maturation.

1.7 Thesis Aim 2

To understand how the body responds to infection or to tumours it is essential to study both the innate and the adaptive immune responses and the ways in which they interact. I therefore sought to analyse TLR-PAMP interactions to complement my work on TCR-pMHC complexes, as understanding of how the innate immune system can trigger the adaptive immune response is likely to lead to improved treatments for infections and tumours.

1.8 Protein Disorder

It has been widely accepted that a protein's primary sequence determines its three-dimensional structure, which in turn determines its function (Anfinsen 1973). However, the subtle interplay of atomic forces, solvent, environment and protein-folding machinery currently renders the *ab initio* prediction of structure from primary sequence alone a remote possibility. Recent studies have added complication by focusing attention on amino-acid sequences that are "disordered", meaning that they show no propensity to form specific three-dimensional structures in their native states. Many proteins contain local regions of such disorder (Figure 10), and some appear to be totally unfolded in their native states (Dunker, Obradovic et al., 2000); it has been predicted that long disordered regions (>30 residues) occur in 2-51% of archaean, 4-45% of eubacterial and 33-67% of eukaryotic proteins, depending on the method used for prediction (Vucetic, Brown et al., 2003; Ward, Sodhi et al., 2004).



a) b)

Figure 10. Regions of disorder in protein structures. a) X-ray crystal structure (PDB identifier 1AK5) with missing electron density circled b) NMR structure (PDB identifier 1A57) with a disordered loop circled.

Studies have shown that many disordered proteins or protein regions are functionally significant and are commonly involved in molecular recognition, often depending on disorder-to-order transitions to enable the natively unfolded proteins to form complexes with their cognate partners. Such molecular interactions can be enzyme/substrate, receptor/ligand, protein/protein, protein/RNA or protein/DNA (Alber, Gilbert et al., 1983; Weinreb, Zhen et al., 1996; Mitchell, Fitzgerald et al., 2006). The transition upon binding can give rise to an interaction having the biologically desirable property of combining high specificity with modest binding affinity (Schulz 1979), thereby avoiding irreversible binding, which is neither suitable nor acceptable for most biological processes. In addition, disordered regions are often able to bind to multiple partners due to their flexibility to adopt different conformations. For example, it has been found that disordered regions play an important role in cell signaling pathways (Frankel and Kim 1991; Weinreb, Zhen et al., 1996).

Accurate recognition of disordered regions in proteins is important to molecular biology for enzyme specificity studies, function recognition and drug design. The detection of such regions is also crucial to structural biology since structure determination by X-ray crystallography and NMR relies on ensembles of (near) identical structures to amplify the experimental signal. At best, disordered regions are invisible to these techniques; at worst they can disrupt the whole experiment by affecting solubility and/or crystallizability (Boutselakis, Dimitropoulos et al., 2003).

One of the most direct ways of identifying disordered regions experimentally is by looking for the amino acids for which there is no electron density in maps obtained by X-ray crystallography (shown in Figure 10a). However, a relatively small fraction of known proteins have had their structures determined and this method has some bias since only those partly structured proteins which have been crystallized can be included. Furthermore, consider a terminal domain of a protein which is believed to be in an expression construct but which is not observed in electron-density maps: (i) this whole domain may be disordered, (ii) a short disordered linker may lead into an ordered domain having no fixed orientation relative to the rest of the protein rendering it invisible in the electron density, or (iii) the protein may have been unexpectedly proteolysed and the domain may be absent from the crystallized entity. In this respect, structures determined by NMR have the advantage since independent domain structures may still be determined for domains connected by an 'invisible' flexible linker region. Classification of disordered regions in entries in the protein data bank (PDB (Berman, Westbrook et al., 2000)) is incomplete; since 1998 it has been possible to annotate unobserved residues using REMARK 465 records but many such remarks have been found to be incomplete and/or inconsistent. However, the development of the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (EBI) (Boutselakis, Dimitropoulos et al., 2003) has meant that residues not observed in a structure are now clearly annotated for most entries in the PDB.

Garner *et al.,* (Garner, Cannon et al., 1998) have shown that disordered regions comprise a sequence-dependent category distinct from that of ordered protein structure, *i.e.* certain

amino acids have a higher frequency of occurrence in disordered regions than in ordered regions and *vice versa*. The aromatic amino acids (tryptophan, tyrosine and phenylalanine) are less likely to appear in long disordered regions, the biological interpretation being that aromatic amino acids have a strong interaction capability which can develop structure and hence inhibit disorder (Burley and Petsko 1985). Cysteine residues have the ability to form disulphide bonds so are also order-promoting, as are the aliphatic amino acids leucine, isoleucine and valine. On the other hand, glutamate, aspartate and lysine are more likely to appear in disordered regions since they are charged and charge imbalance tends to favour disorder. Another residue associated with disorder, serine, increases solubility and provides a flexible locus, two properties inherent to disordered regions.

There are now many tools available for disorder detection. The first widely-available disorder prediction tool was PONDR®, which is now a family of algorithms based on sequence attributes such as amino acid frequencies, hydropathy and flexibility. (Romero, Gervois et al., 1997; Garner, Cannon et al., 1998; Xie, Arnold et al., 1998; Garner, Romero et al., 1999; Li, Romero et al., 1999; Reiser, Darnault et al., 2000; Valmori, Dutoit et al., 2000; Vucetic, Radivojac et al., 2001; Radivojac, Obradovic et al., 2003; Radivojac, Obradovic et al., 2004). These attributes are used to create feed-forward neural networks, but as the rules are derived from prior biological knowledge they therefore rely on the proper determination of an optimised subset of features for the characterization of disorder. Other tools use different methods entirely. For example, FOLDINDEX (Prilusky, Felder et al., 2005) implements the algorithm described by

(Uversky, Gillespie et al., 2000) and predicts disorder based on the average net charge and average hydrophobicity of a protein sequence. IUPred is based on the theory that, in contrast to ordered protein sequences, disordered sequences are unable to form many inter-residue interactions. The algorithm predicts the potential of protein sequences to form such interactions based on the amino acid sequence.

1.9 Thesis Aim 3

The prediction of disorder is now the focus of intensive research and disorder prediction assessment recently formed part of the CASP6 trial (Critical Assessment of Techniques for Protein Structure Prediction; http://predictioncenter.llnl.gov/casp6/Casp6.html), where twenty methods were compared in blind tests. Despite this effort, the accurate prediction of protein disorder remains problematic and new, more accurate tools are badly needed to aid construct design and improve the chances of success of obtaining crystals and subsequently solving crystal structures. I therefore aimed to develop a new protein disorder prediction tool.

1.10 Summary of Thesis Aims

1. To establish any patterns in the binding of CD8+ TCRs to pMHC class I molecules by solving new TCR:pMHC complexes and analysing all available TCR:pMHC complexes using novel bioinformatics methods.

2. To express, purify, crystallize and solve the structures of TLR extracellular domains and compare the structure(s) to that of the extracellular domain of TLR3.

3. To develop a tool to predict intrinsic protein disorder, to aid construct design and maximise the chances of producing protein crystals for X-ray crystallography.

Chapter 2

Structural Studies of Two TCRs Specific for Human Tumour Antigens

2 Structural Studies of Two TCRs Specific for Human Tumour Antigens

2.1 Summary

One potential cancer vaccine candidate is the $NYESO_{157-165}$ peptide (SLLMWITQC; NYESO9C) and the structure of the 1G4 TCR bound to HLA-A2-NYESO9C has been determined (PDB identifier 2BNR) (Chen, Stewart-Jones et al., 2005). The aim was to solve the structure of another TCR bound to this antigen to allow comparison of binding mechanisms used by different TCRs against the same pMHC. However, although the structure of the unliganded 1F2 TCR was solved, no diffracting crystals of the 1F2 TCR bound to pMHC were ever obtained so limited analysis could be done.

The Melan-A antigen is another cancer vaccine candidate, with two dominantly-recognised epitopes; residues 27-35 (AAGIGILTV; nonapeptide) and residues 26-35 (EAAGIGILTV; decapeptide). An analogue of the decapeptide (ELAGIGILTV; ELA decapeptide), binds more strongly to HLA-A2 and is recognised more efficiently by CTL than the wild-type decapeptide. The ELA decapeptide has previously been structurally characterised in complex with HLA-A2 (PDB identifier 1JF1) (Sliz, Michielin et al., 2001) but has not been solved in complex with a TCR. At least 70% of the TCRs that react to Melan-A have the same Vα usage (Vα2.1) (Dutoit, Rubio-Godoy et al., 2002; Trautmann, Labarriere et al., 2002) so one such TCR (2D10) was selected for study. This was crystallized alone and in complex with HLA-A2-ELA, which represents the first

TCR-pMHC complex containing a self peptide tumour antigen. It provides detailed information about how a typical Melan-A-specific TCR binds to pMHC and provides a basis for designing TCR mutations to optimise binding, with the aim of developing more effective cancer vaccines.

2.2 Materials and Methods

All molecular biology and protein production work was done by Dawn Shepherd at the Weatherall Institute of Molecular Medicine, Oxford.

2.2.1 Production of Soluble Heterodimeric TCRs and pMHC Complexes

The generation of soluble TCR heterodimers was based on the procedure described by Stewart-Jones *et al.,* (Stewart-Jones, McMichael et al., 2003). The α and β chains of the Melan-A-specific 2D10 CTL clone and the NYESO-specific 1F2 CTL clone were amplified by PCR from cDNA. The primers used are shown in Table 1.

Table 1. Primers used for cloning of TCRs.

Primer name	Sequence (5'-3')
2D10TCR Vα2.1 forward primer	GGCATATGCAGAAAGAAGTGGAGCAGCATTCTGGACC
2D10TCR Vβ5.1 forward primer	GCCATATGGCGGGCGTGACTCAAACTCCAAGATATCTG
1F2TCR Vα2.1 forward primer	GGCATATGCAGAAAGAAGTGGAGCAGAATTCTGGACC
Vβ2.1 forward primer	GCTACGGCATATGGTGGTGTCTCAACATCCGAGC
Cα reverse primer for 2D10 TCR and 1F2 TCR	CCCGCTCGAGTTTACAACCACCACCGTTTTCTGGGCTGGGGA AGAAGGTGTCTTCTGG
Cβ reverse primer for 2D10 TCR and 1F2 TCR	CGGGGATCCTCGAGCTAGTCACAACCACCACCCCTGTCCTGG TCTGCTCTACCCCAGGCCTCGGCGCTGACG

The reverse primers encode C-terminal constant domain extensions which provide electrostatic complementarity, flexibility and solubility plus additional length to the

regions between the constant domains and the interchain disulphide bridge cysteines to allow the formation of an interchain disulphide bridge during in vitro refolding. The flanking sequences were (PEDTFFPS)PENDGGGCLEKHHHHHH for the α chain and (AEAWGRSR)QDRGGGCD for the β chain.

Residues 1-278 of the HLA-A2 heavy chain were cloned into the pET22b(+) plasmid, expressed in *E. coli* and refolded and purified with the Melan-A peptide ELAGIGILTV or NYESO peptide SLLMWITQC and human β2-microglobulin (β2M) as described previously (Garboczi, Madden et al., 1994). Briefly, PCR products were cloned into pET22b(+) bacterial expression vectors (Novagen). The plasmids containing the TCR α or β chains were transformed into B834-pRARE. The plasmids containing the HLA-A2 and β2M were transformed into *E. coli* strain HMS174. Both were plated on agar containing 100 μg/ml Ampicillin at 37^{o}C overnight. A single colony was picked and grown at 37^{o}C in LB broth containing 100 μg/ml ampicillin. When the bacterial culture was measured to have an optical density (OD^{600}) of 0.6 absorption units the bacteria were induced to express recombinant protein by the addition of isopropylthio-β-D-galactoside (IPTG) to a final concentration of 500 μM and grown for 4 hours. The inclusion bodies were isolated by sonication, washed 3 times in triton buffer (50 mM Tris pH 8, 100 mM NaCl, Triton X100 0.5%, 1 mM DTT, 1 mM EDTA and 0.1% NaN_3), washed once in resuspension buffer (50 mM Tris pH 8, 100 mM NaCl, 1 mM DTT and 1 mM EDTA) and re-solubilized in urea buffer (100 mM Tris pH 8.0, 8 M urea, 2 mM DTT and 1 mM EDTA).

The solubilized TCR α and β chains (30 mg each) were mixed together and diluted into 1 L of refolding buffer (1 M urea, 200 mM L-Arginine HCL, 1 mM EDTA, 150 mM Tris (pH 8.0), 1.5 mM reduced glutathione, 0.15 mM oxidised glutathione) and mixed for 4 days at 4°C. Refolded TCR was concentrated to 10 ml using a 2 L stirred cell and a 10 kD cut off membrane (Amicon) and passed through a Ni^{2+} column. TCR was eluted with 500 mM imidazole buffer (pH 7.5) to separate the αβ and ββ species. The TCR was run on a SD75 gel filtration column and the αβ peak at 65 kD was collected. The α chain C-terminal His_6 tag was removed by incubating the protein (1.2 mg/ml) in 150 mM, 50 mM Tris (pH 7.5) with 80 U/ml carboxypeptidase A beads (Sigma) for 16 hours at 20°C rotating at 20 rpm. The protein was run on an anion exchange column and eluted with a gradient of NaCl. Reduced and non reduced SDS-PAGE analysis confirmed the presence of the interchain disulfide bond and the quality of the purified TCR. The protein was then buffer exchanged into 10 mM NaCl, 10 mM Tris (p H7.0) and concentrated to 10 mg/ml ready for crystallization.

32 mg HLA-A2, 26 mg β2M and 10 mg peptide were added to 1 L refold buffer (100 mM Tris pH 8, 2 mM EDTA, 400 mM L-Arginine-HCl, 5 mM reduced glutathione and 0.5 mM glutathione) and allowed to stir at 4°C for 3 days. Refolded protein was concentrated to 10 ml using a 2 L stirred cell and a 10 kD cut off membrane (Amicon) and run on a SD75 gel filtration column using the running buffer (10 mM NaCl, 10 mM Tris pH 7.0). The peak at 45 kD was collected and concentrated to 10 mg/ml ready for crystallization.

2.2.2 Crystallization

Nanolitre-scale (100 nl + 100 nl), sitting drop, vapour-diffusion crystal trials were set up in 96-well plates using Cartesian robots (Walter, Diprose et al., 2005).

The 1F2 TCR crystallized at room temperature at 10 mg/ml, from 25% PEG3350, 0.1 M HEPES pH 7.5, 0.2 M magnesium chloride.

The 2D10 TCR initially crystallized as small needles at room temperature at 10 mg/ml, from 20% PEG3350, 0.2 M magnesium formate, pH 5.9. A standard optimization screen was set up around this condition (Walter, Diprose et al., 2005) and some large crystals formed in one well which gave some diffraction but not a complete data set. Further optimizations around the condition that gave these crystals were then carried out using an additive screen (Hampton) and diffracting crystals formed in the condition containing 0.02 M sarcosine and a protein:reservoir ratio of 2:1.

The 2D10-HLA-A2-ELA complex crystallized at room temperature from a 1:1 ratio of TCR:pMHC, each at a concentration of 10 mg/ml. Crystals formed in 20% PEG3350, 0.2 M potassium sodium tartrate, pH 7.2.

Examples of the crystals are shown in Figure 1.

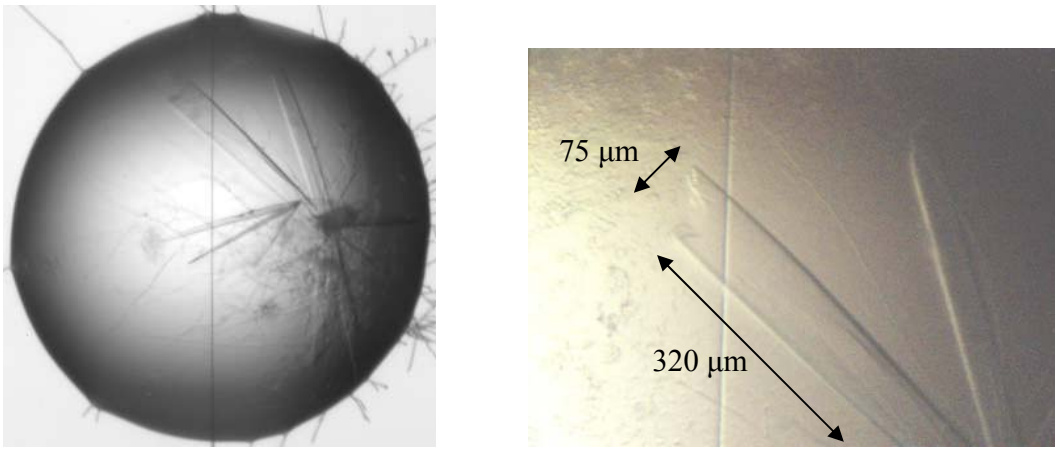


Figure 1a. Crystals of 2D10-A2-ELA which gave diffraction to 2.5 Å

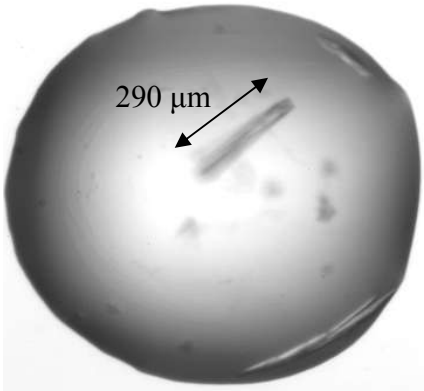


Figure 1b. Crystals of the 1F2TCR which gave diffraction to 2.0 Å

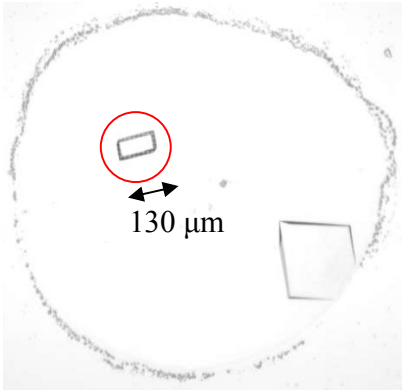


Figure 1c. Crystals of the 2D10 TCR. The circled crystal gave diffraction to 2.3 Å

2.2.3 Data Collection and Analysis

The 1F2 TCR crystal was cryoprotected using Perfluoropolyether PFO-X125/03 and cryo-cooled at 100 K for data collection. Data were collected at the ESRF, Grenoble on beamline ID14 EH2 which has an ADSC Q4R detector, at a wavelength of 0.933 Å.

The 2D10 TCR crystal was cryoprotected using 20% glycerol and cryo-cooled at 100 K for data collection. Data were collected at the ESRF, Grenoble on beamline ID14 EH1 which has an ADSC Q210 detector, at a wavelength of 0.934 Å.

The 2D10-ELA-A2 crystal was cryoprotected using 20% glycerol and cryo-cooled at 100 K for data collection. Data were collected at the ESRF, Grenoble on beamline ID29 which is has an ADSC Q315r detector, at a wavelength of 0.9794 Å.

Data were processed using Denzo and merged with Scalepack (Otwinowski and Minor 1997). Additional data processing statistics were obtained using Xprep. Hydrogen bonds were analysed using HBPLUS (McDonald and Thornton 1994), surface complementarity was calculated using SC (Lawrence and Colman 1993) and buried surface areas were calculated using Naccess.

2.2.3.1 Structure solution of 2D10-ELA-A2 complex

The structure was solved by molecular replacement with Phaser (McCoy, Grosse-Kunstleve et al., 2005) using the structure of a high resolution class I MHC – TCR complex as a model (PDB identifier 1OGA), which was split into two ensembles (pMHC and TCR). CNS (Brunger, Adams et al., 1998) was used for rigid body refinement of the model followed by restrained refinement using Refmac (Murshudov, Vagin et al., 1997). ArpWarp (Perrakis, Harkiolaki et al., 2001) was then used to build the model with the correct amino acid sequences. Iterative cycles of refinement and rebuilding were carried out using Refmac and Coot (Emsley and Cowtan 2004). Waters were added using ArpWarp and manually checked using Coot.

2.2.3.2 Structure solution of the unliganded 2D10 TCR

The structure was solved by molecular replacement with Phaser using the structure of the 2D10 TCR from the 2D10-ELA-A2 complex was used as a model. Many different ensemble combinations were tried before a solution was found using two ensembles, consisting of the variable domain from the 2D10 TCR α chain and the entire 2D10 TCR β chain. This gave good crystallographic packing with no clashes but a poor initial map for the α variable domain. Rigid body refinement using Refmac was attempted to improve this but R factors did not improve significantly. Molrep (Vagin 1997) was then run using the TCR β chain as a fixed input model and the α variable domain as the variable model input. Rigid body refinement followed by restrained refinement using Refmac were then

run on the model output, giving an R value of 0.283 and R_{free} of 0.319. Iterative cycles of refinement and rebuilding were then carried out using Refmac and Coot. Waters were added using ArpWarp and manually checked using Coot.

2.2.3.3 Structure solution of the unliganded 1F2 TCR

The structure was solved by molecular replacement with Phaser using the structure of the A6 TCR from the A6-tax-A2 class I complex as a model (PDB identifier 1QSF). There were two molecules in the asymmetric unit and ArpWarp was used to build into one of these. The model was then superimposed onto the second molecule using SHP (Stuart, Levine et al., 1979). Waters were added using ArpWarp and manually checked using Coot.

Table 2. Crystallographic statistics

Data processing	1F2 TCR	2D10 TCR	HLA-A*0201-ELA-2D10TCR
Space group	$P2_12_12_1$	$P2_12_12$	$P2_1$
Unit cell Dimensions (Å) (a,b,c)	88.3, 91.6, 118.7	96.0, 99.0, 46.3	54.0, 77.3, 117.1
Angles(°) (α,β,γ)	90, 90, 90	90, 90, 90	90, 101.0, 90
Source	ESRF ID14EH2	ESRF ID14EH1	ESRF ID29
Resolution (Å) (highest resolution shell)	30-2.0 (2.10-2.00)*	30-2.3 (2.40-2.30)*	30-2.5 (2.60-2.50)*
Unique reflections	65700	20383	32849
Completeness (%)	99.9 (99.8)*	99.3 (96.6)*	99.1 (94.2)*
I/σ (I)	10.5 (2.2)*	32.35 (8.23)*[†]	7.9 (1.6)*
Rmerge (%)	16.3 (97.7)*	5.5 (26.4)*	18.0 (93.7)*
Refinement			
Resolution range (Å) (highest resolution shell)	30-2.0 (2.10-2.00)*	30-2.3 (2.40-2.30)*	30-2.5 (2.60 – 2.50)*
R_{cryst}	19.2 (20.8)	21.0 (24.7)	20.5 (32.8)
R_{free}	24.2 (27.4)	26.8 (47.1)	26.2 (36.1)
Number of residues in ASU	877	351	822
Number of water molecules	629	153	251
RMS deviation from ideality			
Bond lengths (Å)	0.011	0.013	0.011
Bond angles (°)	1.435	1.334	1.417
Ramachandran Plot (%) (favoured, allowed, generous, disallowed)	(90.2, 9.5, 0.3, 0.0)	(91.7, 8.3, 0.0, 0.0)	(88.3, 11.1, 0.6, 0.0)

[a]$R_{merge}= \Sigma_{hkl}|I<I>|/\Sigma_{hkl}I$ where I is the intensity of unique reflection hkl and <I> is the average over symmetry-related observation of unique reflection hkl.

[b]$R_{cryst} = \Sigma|F_{obs}-F_{calc}|/\Sigma F_{obs}$ where F_{obs} and F_{cal} are the observed and calculated structure factors, respectively.

[c]R_{free} is calculated as for R_{cryst} but using ~5% of reflections sequestered before refinement.

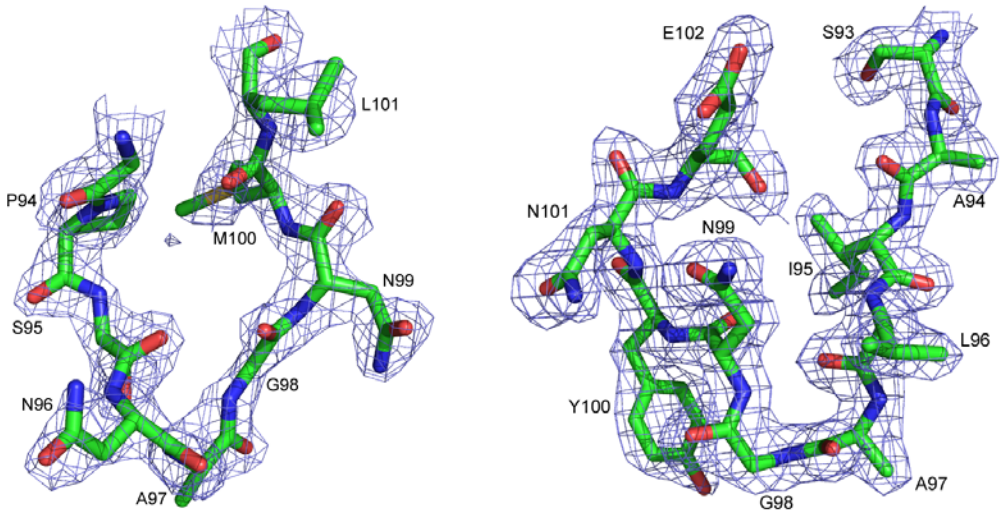
*numbers in parentheses correspond to the outermost shell of data.

[†]No reflections were visible beyond an ice ring, preventing data processing to higher resolution.

2.3 Results and Discussion

2.3.1 The 1F2 TCR Structure

The 1F2 TCR structure was solved to 2.0 Å resolution with two molecules in the asymmetric unit. Crystallographic statistics are shown in Table 2. The electron density around the CDR3 loops of the TCR is shown in Figure 2.



(a) (b)

Figure 2. 2Fo-Fc map at 1.0σ around the CDR3 loops of the 1F2 TCR α chain (a) and β chain (b)

There are 2 molecules in the ASU for the 1F2 TCR. When they are superimposed it is clear they are almost identical (Figure 3) so molecule 1 (chains A and B) was used for further analyses.

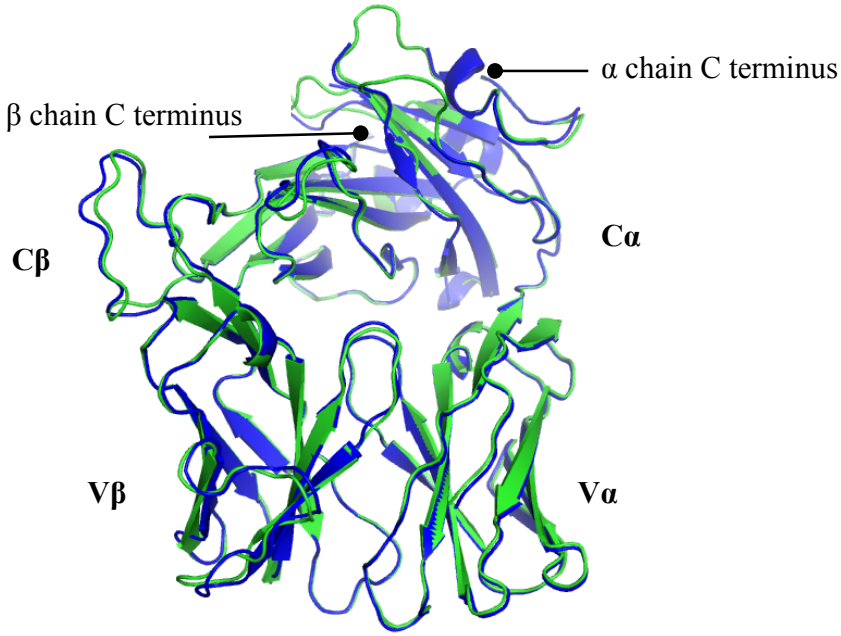


Figure 3. Superimposition of the two molecules of the 1F2TCR in the asymmetric unit. The average RMSD is 0.62 Å.

This is a typical TCR structure consisting of two chains, α and β as described in Chapter 1. Each chain has an N-terminal Ig-like variable domain (Vα and Vβ, 113 and 115 residues long respectively), and a C-terminal Ig-like constant domain (Cα and Cβ, 89 and 128 residues long respectively). In all TCRs, each Ig-like domain consists of two parallel β sheets packed face to face and linked by a disulphide bond into a β sandwhich (Bork, Holm et al., 1994). The β strands of each β sheet are joined by hairpin loops and strands are labeled with the letters A-G in order of their appearance in the amino acid sequence. One β sheet of the each constant domain consists of 4 β strands labeled A, B, D, and E and the other of 2 β strands labeled C, F and G. The first β sheet of each variable domain has two extra β strands labeled C' and C'', with the second β sheet having 4 β strands as for the constant domains.

2.3.2 Comparison of the 1F2 TCR to the 1G4 TCR

The aim of this work was to solve the structure of the 1F2 TCR in complex with HLA-A2-NYESO9C to allow comparison of the binding mechanisms used by different TCRs bound to the same pMHC. Unfortunately, diffracting crystals were only obtained for the unbound 1F2 TCR so limited analysis could be carried out. However, the structure of the unbound 1F2 TCR can be compared to that of the 1G4 TCR bound to HLA-A2-NYESO9C to examine whether the two TCRs could bind in a similar manner.

The 1G4 TCR CDR loops are in very similar positions in the free and bound states (PDB identifiers 2BNU and 2BNR respectively). However, when the bound 1G4 TCR is compared to the 1F2 TCR it can be seen that the CDR3 loops point in opposite directions. Unless the CDR loops of the 1F2 TCR undergo dramatic movements when binding to HLA-A2-NYESO9C, the 1F2 TCR cannot bind this antigen in the same way as the 1G4 TCR as the CDR loops would not fit around the peptide (Figure 4). It is more likely that the 1F2 TCR is shifted/tilted/twisted in a different way to the 1G4 TCR when bound to HLA-A2-NYESO9C. This may explain why the 1F2 TCR binds better to the NYESO9C peptide than the NYESO9V peptide, whereas the 1G4 TCR binds better to the NYESO9V peptide.

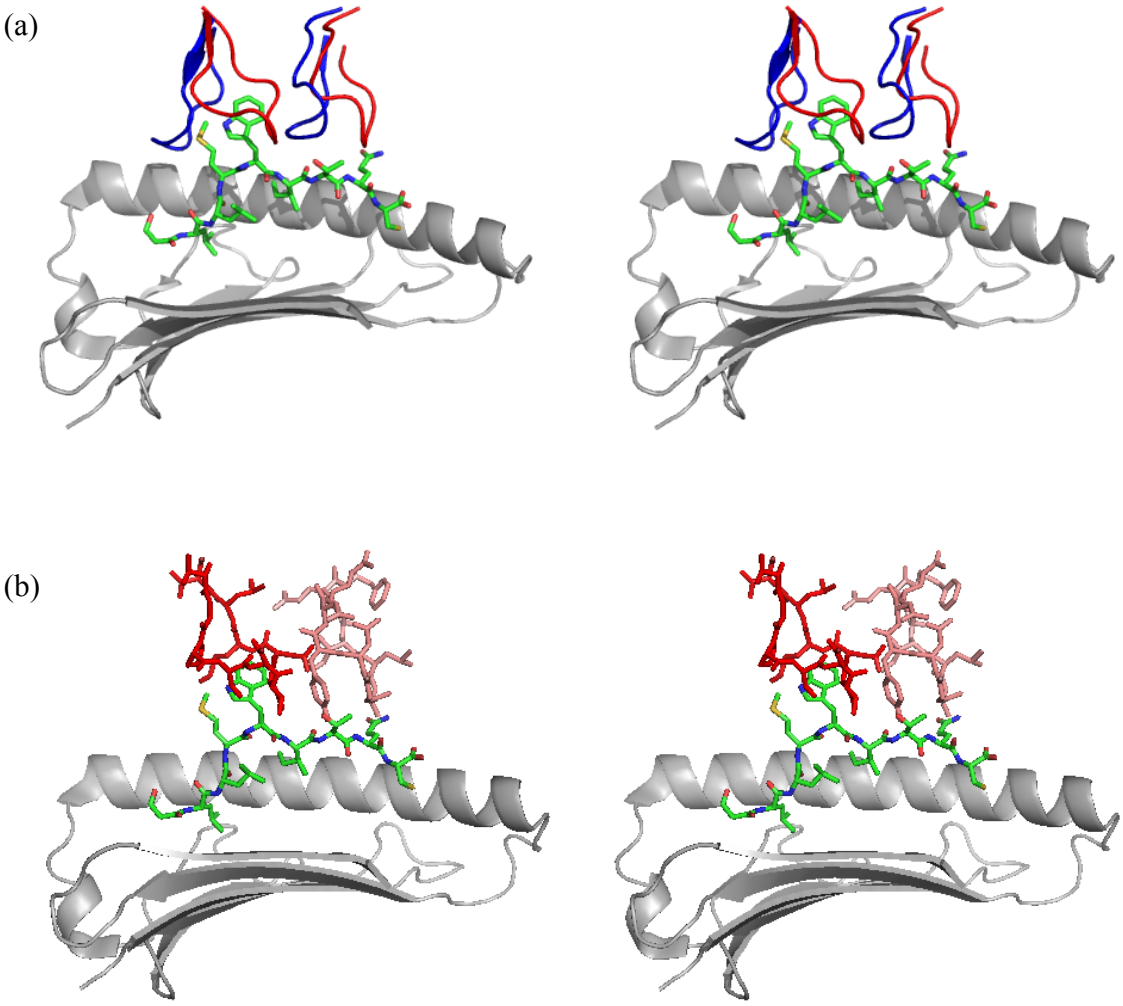


Figure 4. Superimposition of the 1F2 TCR onto the 1G4 TCR bound to HLA-A2-NYESO9C (stereo). The NYESO9C peptide is shown in green bound to HLA-A2 in grey. (a) The position of the 1G4 TCR CDR 3 loops (blue) compared to the position of the 1F2 TCR CDR3 loops (red). CDRα loops are on the left and CDRβ loops are on the right. (b) Stick representation of the 1F2 TCR CDR3 loops showing the clashes with the peptide that would occur if the 1F2 TCR bound to HLA-A2-NYESO in the same orientation as the 1G4 TCR. CDR3α is shown in red and CDR3β in salmon.

2.3.3 The Unbound 2D10TCR Structure

The structure of the unbound 2D10TCR was solved to a resolution of 2.3Å with one molecule per asymmetric unit (Table 2). The α constant domain of the unbound 2D10TCR structure is completely missing, with density only visible up to residue 112 (Figure 5). Examination of the crystal packing shows that there is no room for it to be present in the structure, proving that it has been cleaved off rather than simply being disordered. Cleavage must have occurred within the crystallization drop, as an SDS PAGE gel of the TCR prior to crystallization showed the TCR ran as a heterodimer under non-reducing conditions. This indicates that the α constant domain was present, as the cysteines which form the inter-chain disulphide bond are encoded after the constant domains. It is possible that there was fungus growing in the drop which proteolysed the TCR, although none was visible.

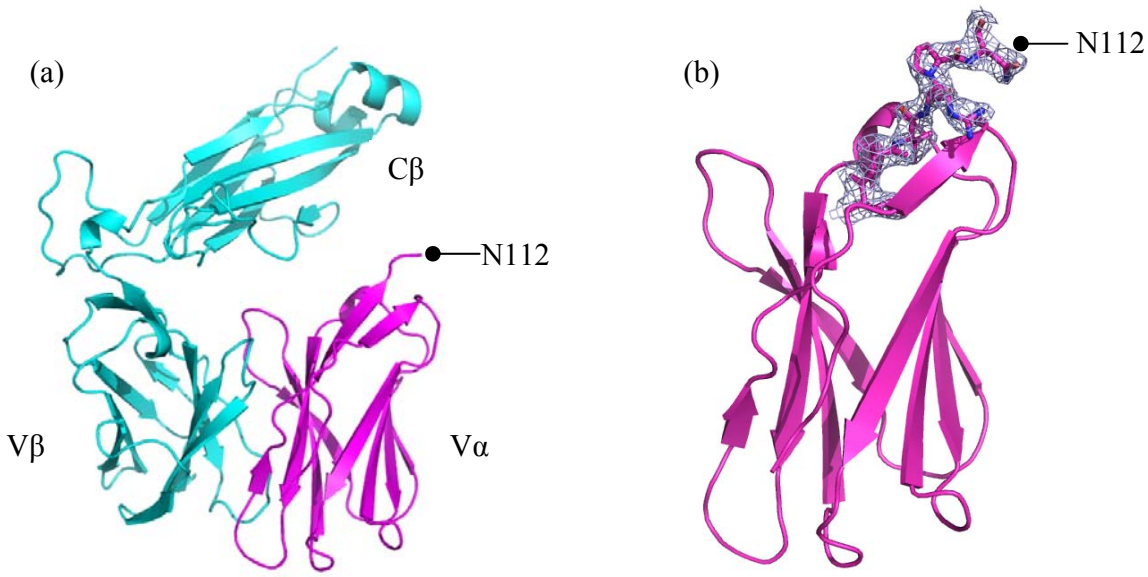


Figure 5. (a) Structure of the unbound 2D10 TCR. The Cα domain had no visible electron density. (b) Electron density around the C-terminus of the Vα domain showing no visible density after residue 112.

The α constant domain is also missing from the A6 TCR in the complex with PDB identifier 1A07 (see Figure 5, Chapter 1), and Rudolph *et al.,* (Rudolph, Huang et al., 2001) crystallized an isolated Vα domain from the murine 2C TCR despite the whole TCR being present in the crystallization mixture.

The electron density around the CDR3 loops of the TCR is shown in Figure 6.

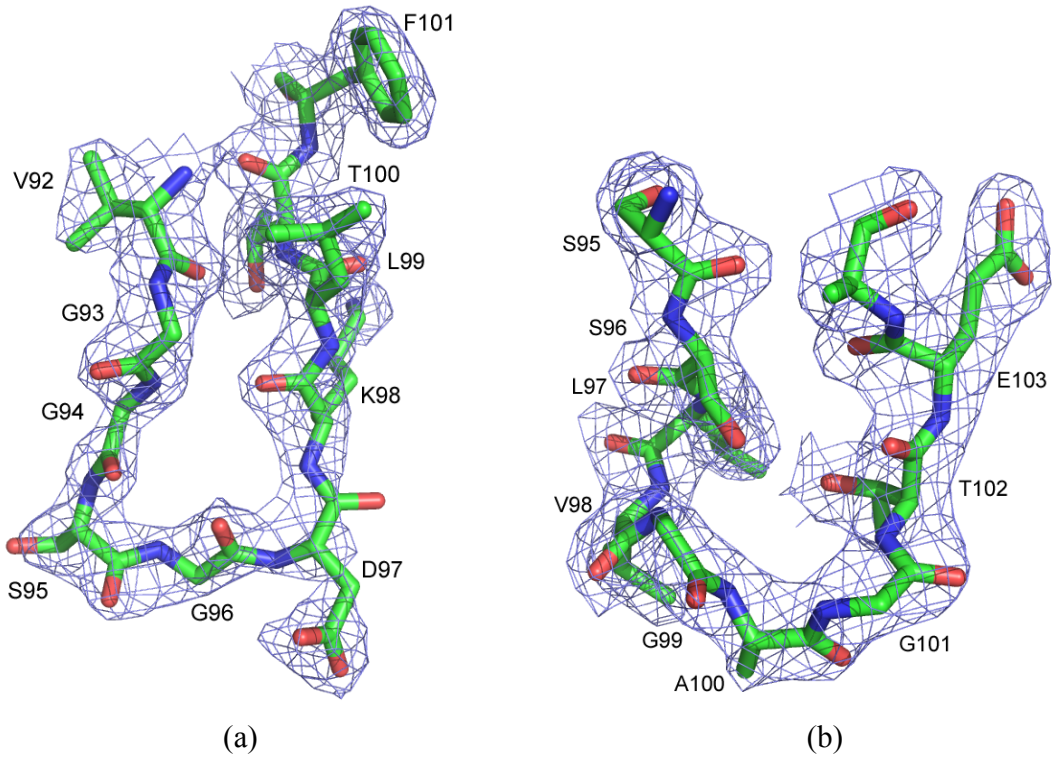


Figure 6. 2Fo-Fc map at 1.0σ around the CDR3 loops of the 2D10 TCR α chain (a) and β chain (b)

2.3.4 The 2D10TCR-ELA-A2 Structure

The structure of the 2D10TCR-ELA-A2 complex was solved to 2.5 Å with one molecule per asymmetric unit, and is fairly typical of class I MHC-TCR complexes. The electron density around the ELA peptide, the CDR3 loop of the TCR α chain and the CDR3 loop of the TCRβ chain is shown in Figures 7, 8(a) and 8(b) respectively. The most distinguishing feature is the shape of the peptide bound in the MHC groove, which bulges sideways towards the α2 helix of the MHC, rather than upwards out of the groove towards the TCR (Figure 9). This matches the conformation of the peptide in the ELA-A2 complex structure (PDB number 1JF1) so the peptide has not been flattened as a result of TCR binding.

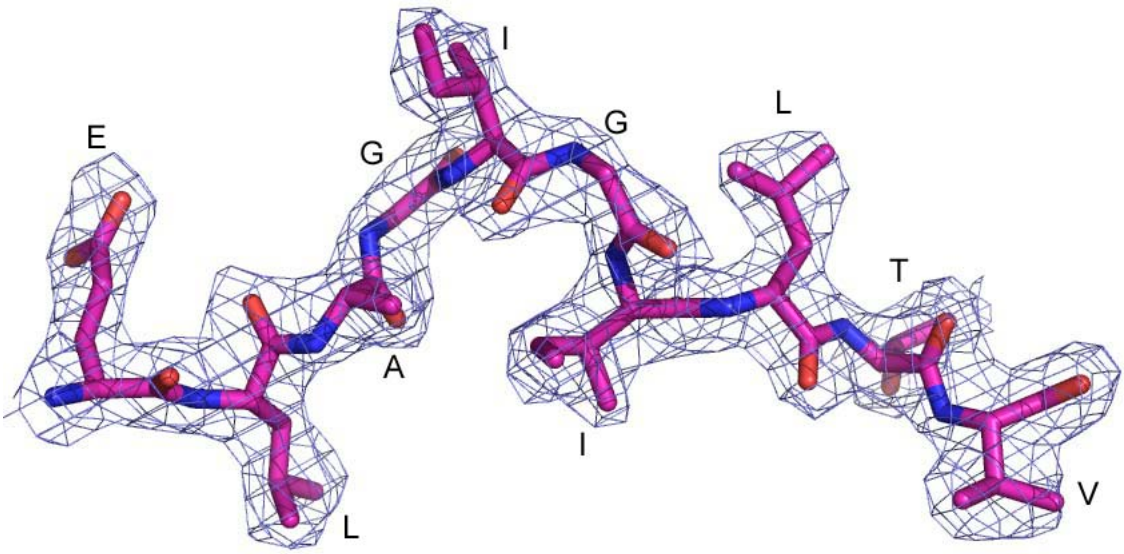


Figure 7. 2Fo-Fc map at 1.0σ around the ELA peptide from the 2D10-ELA-A2 structure (sequence ELAGIGILTV).

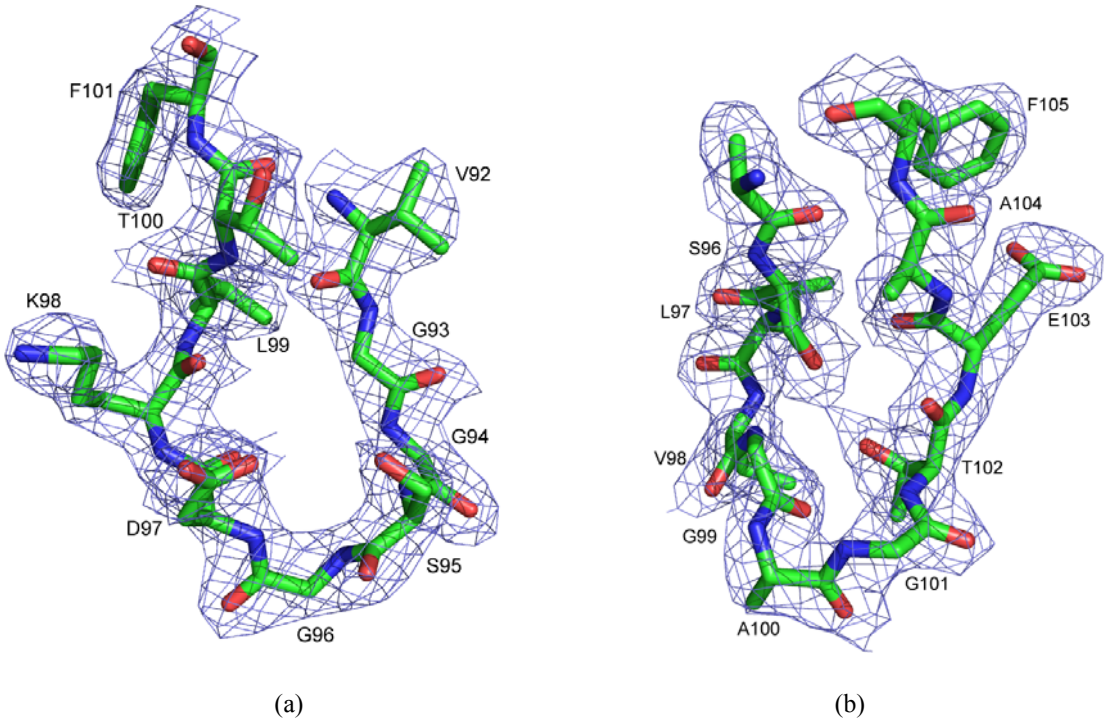


Figure 8. 2Fo-Fc map at 1.0σ around the CDR3 loops of the 2D10 TCR α chain (a) and β chain (b) from the 2D10-ELA-A2 complex structure.

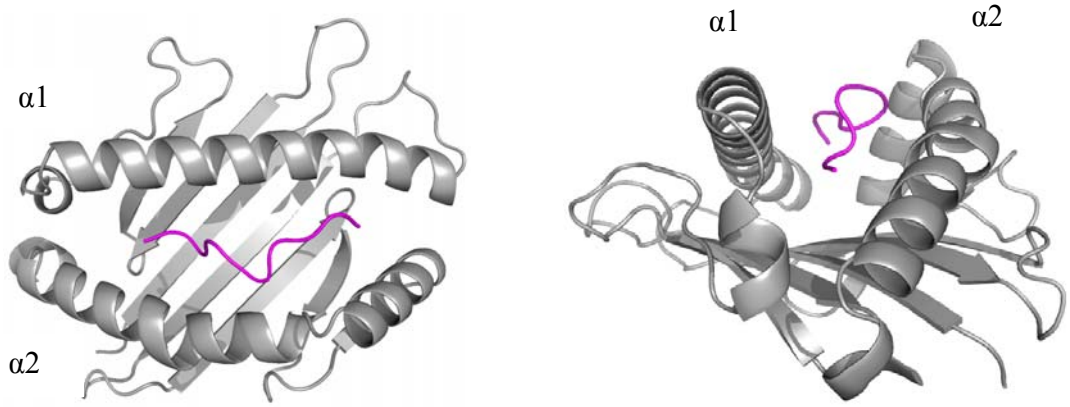


Figure 9. The ELA peptide (mainchain only shown as magenta coil) in the HLA-A2 groove, showing the sideways bulge towards the α2 helix.

The 2D10TCR binds to the pMHC with a twist angle of 33° which is in the normal range for class I complexes (see Chapter 3). The TCR sits directly above the centre of the peptide but is shifted slightly towards the α1 helix of the MHC. Unlike many TCRs bound to class I pMHC, it is barely tilted at all (1.4°) with respect to the plane of the β-sheet of the MHC.

2.3.5 Comparison of Free and Bound 2D10 TCR

When the free TCR is superimposed onto the bound TCR using the TCRβ chain residues, it can be seen that there is a small shift in the position of the Vα domain (Figure 10) but this may be due to the missing Cα domain allowing repositioning of the Vα domain.

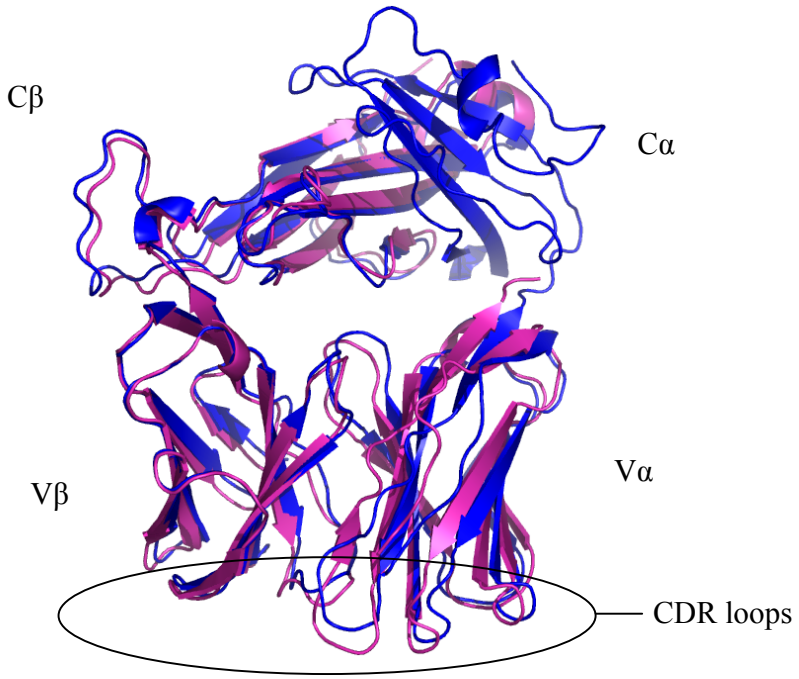


Figure 10. Free TCR (magenta) superimposed onto the bound TCR (blue) via the TCRβ residues.

However, the orientation of the CDR loops is almost identical, indicating the CDR loops do not need to alter structurally to be able to bind to the pMHC. This, together with the very upright orientation of the TCR on the pMHC may explain the fast on and off rates of the TCR. Surface Plasmon resonance measurements were carried out by Dawn Shepherd at the Institute of Molecular Medicine: the K_{off} rate was in the region of 4.2 s^{-1} which is significantly faster than the rates described for other TCR-pMHC complexes (0.01-0.73 s^{-1}). The K_{on} rate was too fast to measure.

If the free and TCR-bound peptide-MHC structures are superimposed using residues 1-180 of chain A of the MHC, it can be seen that there is very little conformational change of the peptide (Figure 11). There is some movement between residues 3 and 4 and between 6 and 9. The maximum difference is a distance of 0.86 Å between the Cα atoms of leucine 8 suggesting that the TCR is 'squashing' the peptide slightly.

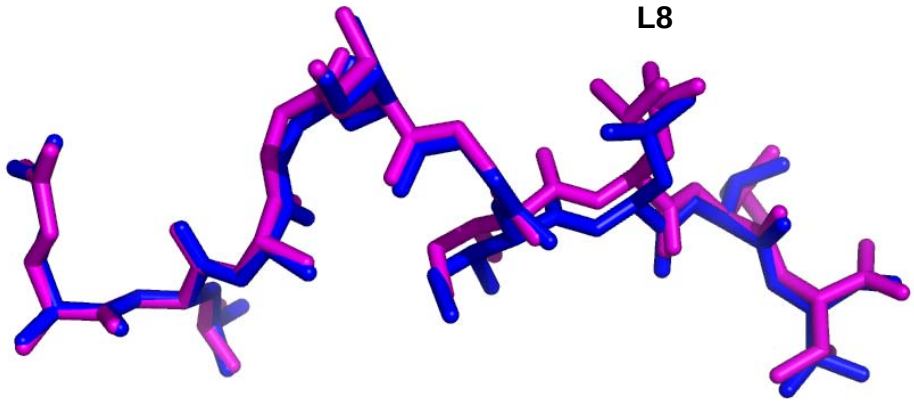


Figure 11. Free ELA peptide (magenta) and ELA peptide bound to TCR (blue). The peptide runs from N- to C-terminus from left to right and leucine 8 is labeled. This residue has 2 alternate conformations in the unbound peptide.

2.3.6 Interactions of the 2D10 TCR with the ELA Peptide-HLA-A2

There are 16 residues on the MHC involved in contacts with the TCR (Table 3 and Figure 12), 9 contacted by TCRα (G62, R65, K66, E154, Q155, A158, Y159, T163 and R170) and 7 by TCRβ (A69, Q72, T73, R75, V76, T80 and A150).

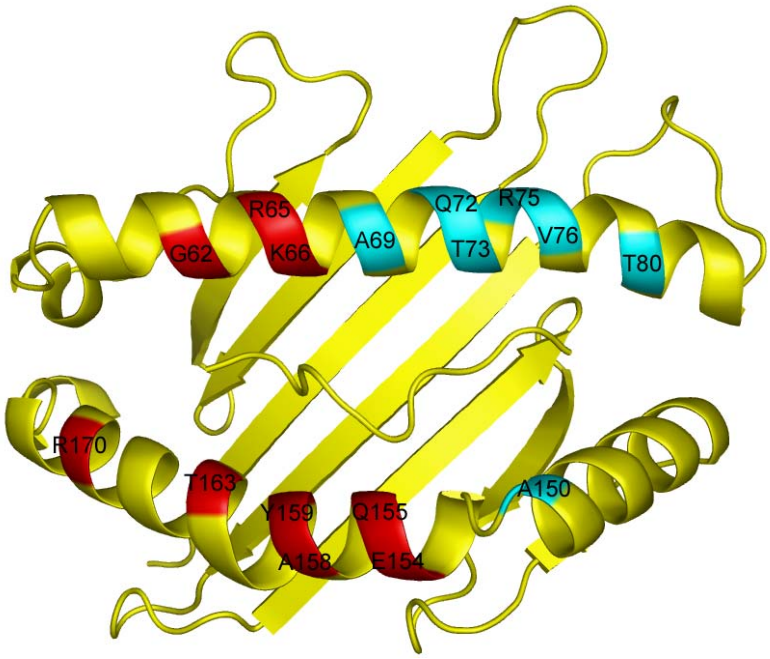


Figure 12. TCR contacts to MHC. TCRα contacts are coloured red and TCRβ contacts are coloured cyan.

Only 5 of these (R65, Q72, R75, E154 and Q155) undergo any movement when comparing the free MHC and MHC bound to the TCR, and movement of E154 and Q155 is very slight (Figure 13). Residues Q72, R75 and E154 are involved in hydrogen bonds with TCR.

Table 3. Contacts at the TCR-pMHC interface. Superscripted designations indicate atoms involved in hydrogen bonding.

	TCR residue	pMHC residue	# contacts
TCR-MHC contacts			
CDR1α	R29	R170	4
	Q32	K66	1
	Q32	Y159	1
	Q32	T163	1
CDR2α	Y52	E154	1
	Y52	Q155	7
	Y52	A158	3
	$K57^{NZ}$	$E154^{OE2}$	5
HV4α	K68	T163	1
CDR3α	G94	K66	1
	S95	G62	1
	S95	R65	1
	S95	K66	2
	G96	R65	3
	G96	K66	2
	D97	R65	2
CDR1β	R32	V76	3
	R32	T80	1
CDR2β	F52	Q72	3
	F52	V76	1
	$S53^{OG}$	$R75^{NH2}$	1
	S53	V76	2
	$R57^{NE}$	$Q72^{OE1}$	6
CDR3β	L97	A150	1
	V98	T73	2
	A100	A69	1
TCR-peptide contacts			
CDR1α	G30	E1	1
	$Q32^{NE2}$	$E1^{OE2}$	4
	$Q32^{NE2}$	$L2^{O}$	3
	Q32	A3	2
	$Q32^{OE1}$	$G4^{N}$	5
	Q32	I5	3
	S33	I5	1
CDR2α	Y52	I5	1
CDR3α	G94	G4	4
CDR1β	R32	T9	6
CDR3β	V98	I7	1
	V98	T9	2
	T102	G4	1

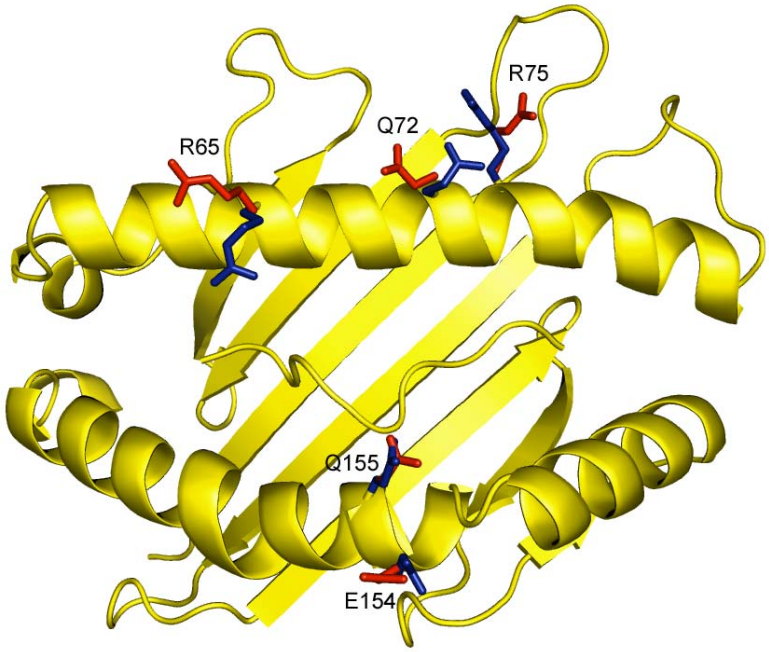


Figure 13. MHC residues involved in TCR contacts which change position when bound to the TCR (R65, Q72, R75, E154, Q155). Positions of free residues are in blue, bound in red.

TCRα dominates contacts to the peptide, mainly using CDR1α. Unusually, the CDR3α and CDR3β loops are used mainly for contacting the MHC rather than the peptide (Figures 14 and 15). This may help to explain the dominant usage of Vα2.1 in TCRs which recognize the Melan-A epitope; Q32 in CDR1α makes extensive contacts to both peptide and MHC (Figure 14) and a Q in this position is only found in Vα2.1, Vα2.2 and Vα2.3. The surrounding residues of the CDR1α loop in Vα2.1 may be the only combination that allows Q32 to sit in the correct orientation to make all these contacts (Figure 15).

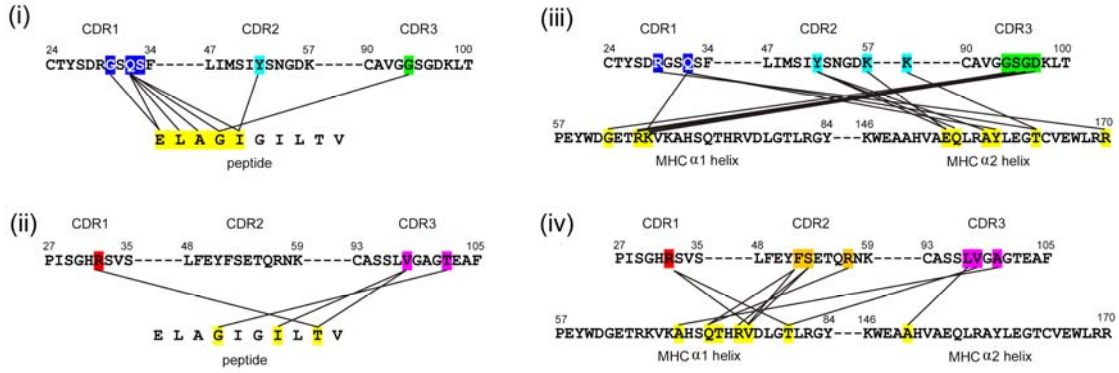


Figure 14. TCR contacts to peptide and MHC. TCR CDR loops are shown above either peptide or MHC sequences. Only MHC residues from the α1 and α2 helices are shown. (a) TCRα contacts to peptide (TCRα residues 30,32,33,52,94), (b) TCRβ contacts to peptide (TCRβ residues 32,98,102), (c) TCRα contacts to MHC (TCRα residues 29,32,52,57,68,94,95,96,97), (d) TCRβ contacts to MHC (TCRβ residues 32,52,53,57,97,98,100).

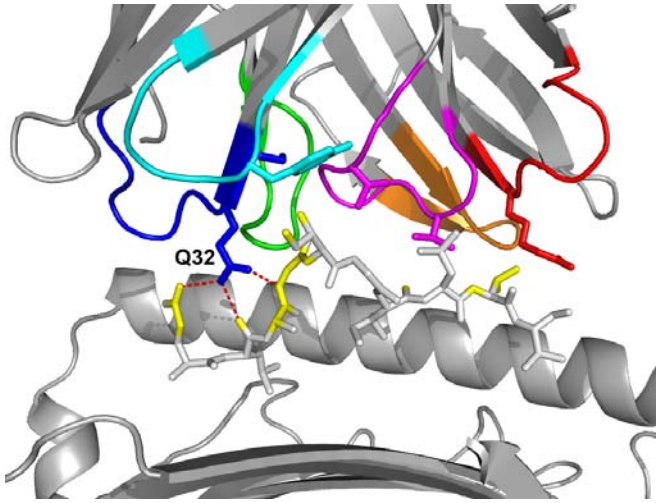


Figure 15. TCR contacts to peptide. The TCR CDR loops are coloured as follows: VαCDR1, blue; VαCDR2, cyan; VαCDR3, green; VβCDR1, red; VβCDR2, orange; VβCDR3, magenta, matching their colouring in Figure 14. TCR residues involved in peptide contacts are shown as sticks. Peptide atoms involved in contacts with TCR are coloured yellow. Residue Q32 in VαCDR1 makes 3 hydrogen bonds with peptide residues E1, L2 and G4, shown as red dotted lines.

Binding of the TCR to the peptide and the MHC is mainly mediated by van der Waals contacts, although there are 3 hydrogen bonds formed between TCRα and the peptide, 1

hydrogen bond between TCRα and the MHC and 2 hydrogen bonds between TCRβ and the MHC (Table 3). Notably, MHC Q155 is only contacted by TCRα. In most other HLA-A2-TCR complexes this residue is contacted by both TCRα and TCRβ.

The total buried surface area between the TCR and the pMHC is ~1900Å^2 which is in the middle of the range for class I complexes (1350-2250 Å^2). TCRα contributes 58% of the total pMHC surface buried by the TCR, with CDR3α contributing 21% of this.

The surface complementarity (sc) between the TCR and the pMHC is 0.65 which is within the normal range for class I complexes. This does not change significantly if water molecules are included in the calculation. The sc between Vα and the peptide is 0.81 compared to 0.64 between Vβ and the peptide, although Vβ has a higher complementarity to the MHC (0.71 for Vβ, 0.63 for Vα). Overall, the sc of Vα to pMHC is 0.69, which is only slightly higher than that of Vβ to pMHC (0.65).

2.3.7 Possible Mutations to Improve Binding

It may be possible to change residues in the CDR3β loop to enable the TCR to bind to peptide residue L8 which is currently not contacted by the TCR at all, despite being fairly prominent (Figure 16).

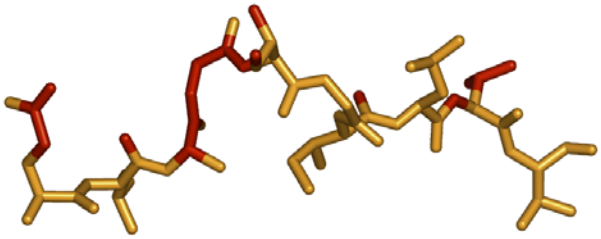


Figure 16. Peptide atoms contacted by TCR are shown in red, atoms not contacted in orange.

Mutating residue L97 of CDR3β to an aromatic residue (phenylalanine, tyrosine or tryptophan) may allow contacts to be made with peptide residues L8 and T9. In addition, residue V98 of CDR3β currently contacts peptide atoms I7 O and T9 N and Cβ. However, mutating this to the polar residue threonine may allow hydrogen bonds to be formed so improving binding of the TCR to the peptide (Figure 17).

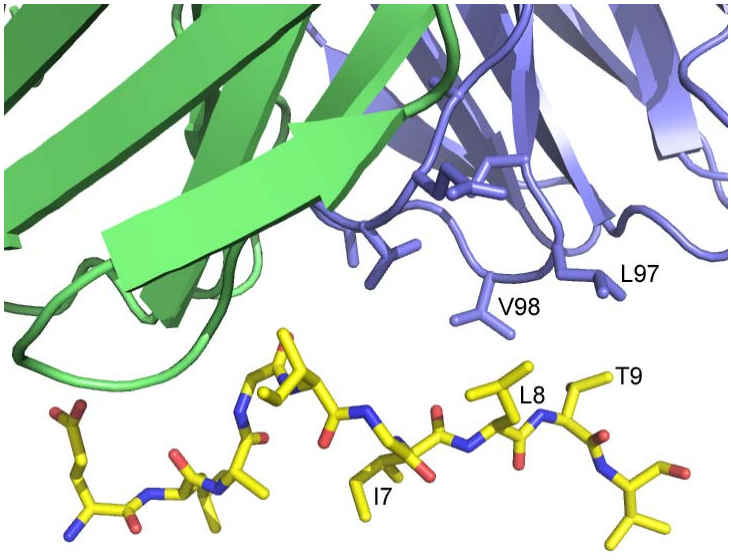


Figure 17. TCR residues V98 and L97 could be mutated to improve TCR binding to the peptide. V98 currently contacts peptide atoms I7 O and T9 N and Cβ, but mutating it to be a threonine may allow hydrogen bonds to be formed. Mutating L97 to be an aromatic residue may allow contacts to peptide L8 and T9.

There may also be ways of engineering the mutated TCR for use as a vaccine, to force mutant pairs of TCR α and β chains to associate together with ease in preference to with wild-type chains. One way of doing this would be to mutate salt bridges between the α and β chains by swapping round the charges on the mutant TCR so only mutant chains could pair together. For example, there are salt bridges between TCRα R124 and TCRβ E130 and between TCRα D138 and TCRβ R196 in the constant domains of the TCR. It may be possible to swap these pairs round (*e.g.* change TCRα R124 to be E and TCRβ E130 to be R) so that the α chain of a mutated TCR could not pair with the β chain of a non-mutated TCR (Figure 18).

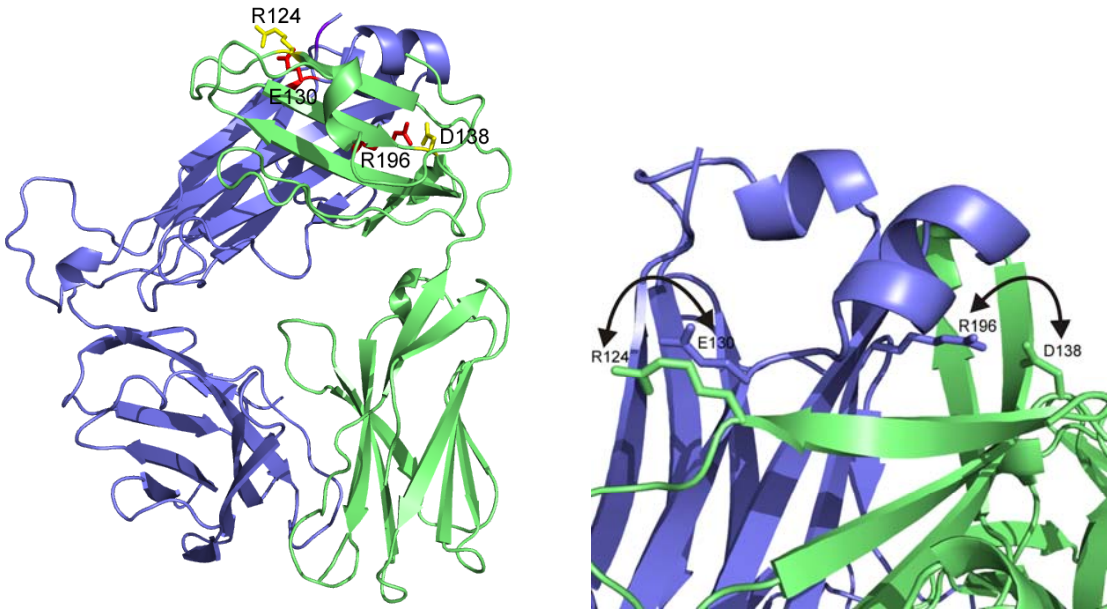


Figure 18. Existing salt bridges between TCRα and TCRβ constant domains which could be reversed to ensure mutant TCR chains pair together. TCRα is shown in green and TCRβ in blue.

2.4 Conclusions

The structure of the 1F2 TCR suggests that this TCR binds to HLA-A2-NYESO9C in a very different orientation to that of the 1G4 TCR, which may explain why the 1F2 TCR binds better to the NYESO9C peptide than the NYESO9V peptide, whereas the 1G4 TCR binds better to the NYESO9V peptide. However, lack of the complex structure of the 1F2TCR bound to HLA-A2-NYESO9C means this is only speculation.

The 2D10TCR binds to HLA-A2-ELA in a very upright orientation and with very little structural alteration of the CDR loops which may explain the very fast on and off rates of this TCR. TCRα dominates contacts to the ELA peptide, mainly using the CDR1α loop, and unusually the CDR3α and CDR3β loops are mainly used for contacting the MHC rather than the peptide. The specific combination of residues in CDR1α, particularly the orientation of Q32 which makes extensive contacts with peptide and MHC, may explain the dominant usage of Vα2.1 in TCRs which recognize the Melan-A epitope.

It may be possible to mutate certain residues in the CDR3β loop of the 2D10TCR to improve binding to the peptide. If experiments prove this to be a successful strategy then salt bridges between the α and β chains of the mutant TCR could be swapped round to prevent mutant TCR chains pairing with native chains, so enhancing the use of the mutated TCR as a cancer vaccine.

Chapter 3

Analysis of TCR-pMHC Complexes

3 Analysis of TCR-pMHC Complexes

3.1 Summary

In 2006 there were 17 TCR-class I pMHC complexes in the Protein Data Bank (PDB) (11 human and 6 mouse), of which 10 (7 human and 3 mouse) were unique *i.e.* not peptide variants of existing structures (Table 1). The Jones group have 3 more unique complexes awaiting publication and in addition I solved the structure of the 2D10TCR-ELA-A2 complex. This database of 14 unique complexes was used to investigate the mechanisms of binding of TCRs to pMHC. In addition, in 2007 the structure of two more class I TCR pMHC complexes were published (Colf, Bankovich et al., 2007; Tynan, Reid et al., 2007) and analysed in the same way.

There are many parameters of TCR-pMHC binding that can be examined. Firstly, the physical positioning of the TCR on the pMHC. There have been several attempts to calculate the twist, tilt and shift of TCRs bound to pMHC (Teng, Smolyar et al., 1998; Rudolph and Wilson 2002; Buslepp, Wang et al., 2003), but each uses a different method making it impossible to accurately compare complexes analysed in different papers. This work gives a comprehensive analysis of all known unique TCR-pMHC complexes and extends previous analyses to include measurements of tilt and shift direction as well as magnitude and measurements of TCR heights.

Secondly, the contact residues used in the TCR-pMHC interface can be investigated. (Reche and Reinherz 2003) aligned all human MHC class I and class II sequences

obtained from the IMTG database (http://imgt.cines.fr/home.html) (Lefranc, Giudicelli et al., 2005) to investigate sequence variability. They combined this with contact maps showing interactions with peptide and TCR, but only had 5 human TCR-pMHC structures available, of which 4 were peptide variants, and 5 mouse complex structures of which only 3 were unique. Here, sequence variability at TCR contact sites on the MHC and also at MHC contact points on the TCR were investigated. Average contact points for each complex were also calculated to examine the 'focus point' of each TCR.

Finally, buried surface areas (BSAs) were calculated for each complex. This demonstrates the variability of the total BSA within the TCR-pMHC complexes as well as the contribution of the peptide, TCRα and TCRβ.

Together, this work represents the most comprehensive analysis of TCR-pMHC structures to date and allows a unique insight into the mechanisms of binding of the TCR to its cognate MHC. The results show that the focus of the interaction is always on the centre of the peptide and the buried surface areas between the TCR and pMHC are very similar for all the complexes, despite considerable variation in tilt, twist and shift of the TCR. Six key residues on the MHC were found to be important in binding of the TCR, which may represent an essential 'footprint' that the TCR must adopt for effective binding. The CDR loops of the TCR are used to contact these six residues, but not by using specific conserved amino acids. However, TCRs generally use at least one aromatic residue in one or more of their CDRα loops to bind MHC Q155.

3.2 Methods

Detailed descriptions of calculations are included in Appendix B and all necessary programs can be found on the CD included with this thesis.

3.2.1 Definitions

β-sheet plane: for the reference complex, 1OGA, the best-fit plane using 29 atoms from the β-sheet floor of the MHC groove was calculated. The RMSD was 1.7Å.

Interface plane: for each complex, atoms in each chain of the TCR and atoms in the pMHC which where within 4Å of each other were obtained, and the average of these was calculated to obtain an 'average contact point'. A plane at this point which was parallel to that of the β-sheet plane is taken to be the 'interface plane'.

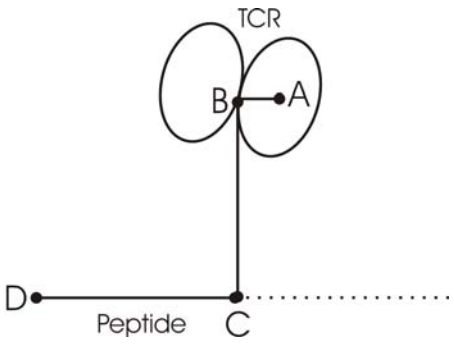
TCR 2-fold axis: this was calculated by superimposing the α chain of the TCR onto the β chain and calculating the rotation matrix. From this, the unit vector of the 2-fold axis between TCRα and TCRβ can be determined. This must point from the TCR towards the MHC in each case for consistent calculations.

All calculations were done by first superimposing each complex onto 1OGA using residues 1-180 of the MHC. The β-sheet plane of 1OGA was used in all calculations of tilt angles and heights of TCRs as this makes the numbers more comparable than if a plane was calculated for each complex individually.

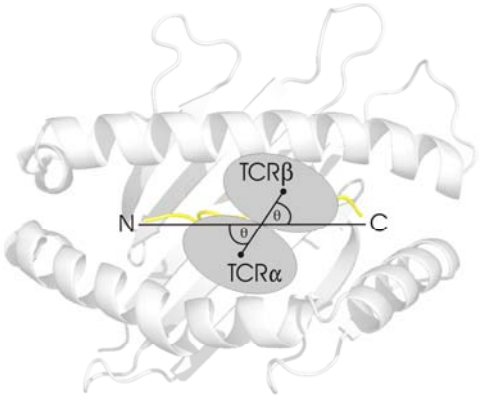
3.2.2 Explanation of Measurements

3.2.2.1 Twist

The twist of a TCR is calculated as a dihedral angle composed of 4 points: A (the midpoint of the disulphide bond in TCRα), B (the midpoint of the line drawn between the midpoint of disulphide bond in TCRα and the midpoint of the disulphide bond in TCRβ), C (the midpoint of the line drawn between the N and C termini of the peptide) and D (the N terminus of the peptide). The angle calculated is between the lines A-B and C-D when looking along the B-C line.

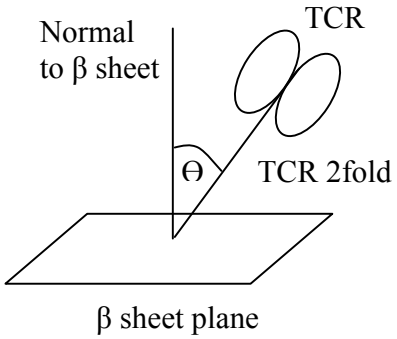


Viewed down the line B-C:

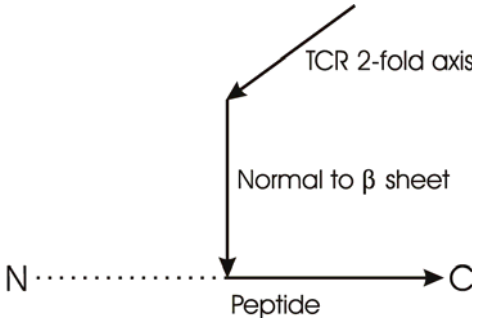


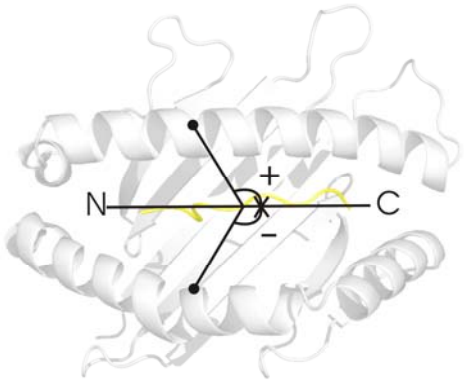
3.2.2.2 Tilt

The tilt angle of a TCR is calculated by taking the absolute value of the dot product of the 2-fold axis of the TCR and the normal to the β-sheet plane.



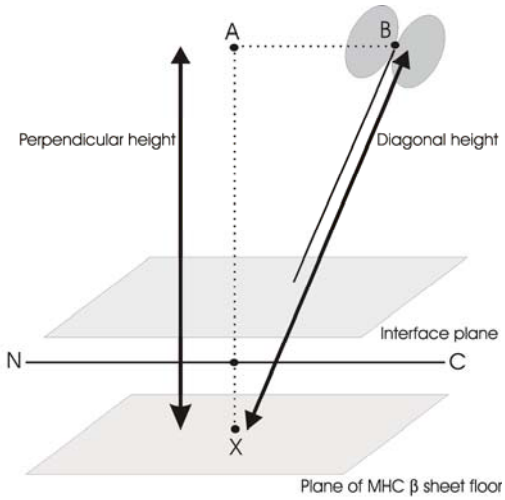
The tilt direction of a TCR is calculated as a dihedral angle using 3 vectors: the TCR 2-fold axis, the normal to the plane of the β-sheet floor of the MHC groove and the vector between the N and C termini of the peptide.





3.2.2.3 Heights

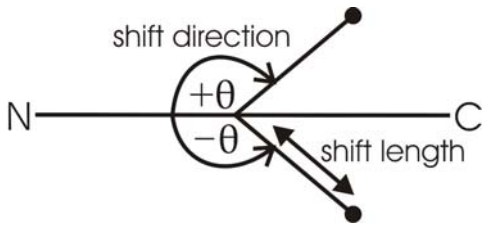
Two heights for a TCR can be calculated: the 'diagonal height', from the β-sheet plane to the midpoint of the line between the disulphide bonds in TCRα and TCRβ, and the 'perpendicular height' as shown below.



X: point on MHC β sheet floor directly below the midpoint of the line drawn between the N and C termini of the peptide

3.2.2.4 Shift

The shift of a TCR is calculated as the distance from the midpoint of the line drawn between the N and C termini of the peptide projected onto the interface plane to the point at which the 2-fold axis of the TCR intersects with the interface plane.



Buried surface areas were calculated using NACESS, a program written by Simon Hubbard.

Sequence logos were made using WebLogo (Crooks, Hon et al., 2004) which is freely available at http://weblogo.berkeley.edu.

Sequence alignments were done using ClustalW (Chenna, Sugawara et al., 2003) which is freely available at www.ebi.ac.uk/Tools/clustalw.

Figures depicting TCR and pMHC were created using Pymol (DeLano Scientific; www.pymol.org) and corelDRAW® 11 (Corel Corporation).

Table 1 shows the database of 14 class I complexes used for analysis.

Table 1. TCR-class I pMHC complexes available in the PDB in 2006.

MHC-peptide-TCR	PDB#	Peptide length	α	β
A2-MP-2JM22	1OGA	9	Vα10.2	Vβ17
A2-tax-A6	1AO7	9	Vα2.3, Jα24	Vβ12.3, Dβ2.1, Jβ2.7, Cβ2
A2-taxP6A-A6	1QRN	9	Vα2.3, Jα24	Vβ12.3, Dβ2.1, Jβ2.7, Cβ2
A2-taxV7R-A6	1QSE	9	Vα2.3, Jα24	Vβ12.3, Dβ2.1, Jβ2.7, Cβ2
A2-taxY8A-A6	1QSF	9	Vα2.3, Jα24	Vβ12.3, Dβ2.1, Jβ2.7, Cβ2
A2-tax-B7	1BD2	9	Vα17.2, Jα54	Vβ12.3, Jβ2.7, Cβ2
A2-p1049-AHIII12.2	1LP9	9	Vα8.3 (TRAV12D-2)	Vβ8.1 (TRBV13-3)
A2-CMV-5TCR	*	9	Vα4.1	Vβ20.1
A2-NYESO9C-1G4	2BNR	9	Vα23	Vβ13
A2-NYESO9V-1G4	2BNQ	9	Vα23	Vβ13
A2-ELA-2D10	*	10	Vα2.1	Vβ5.1
B8-EIY-16TCR	*	8	Vα9.1	Vβ22
B8-EBV-Lc13	1MI5	9	Vα4.1 (TRAV26-2*01)	Vβ6.2 (TRBV7-8*03)
B57-KAF-TCR	*	11	Vα15.1	Vβ17
B35-EBV-SB27	2AK4	13	Vα12S1	Vβ13S3
$H2K^b$-vsv- Bm3.3	1NAM	8	AV17S1 (TRAV16D)	Vβ2S1 (TRVB1)
$H2K^b$-pkb1-Kb5C20	1KJ2	8	Vα2.3, JαA10	Vβ2, Dβ2, Jβ2.3
$H2K^b$-dev8-2C	2CKB	8	Vα3 Jα58 Cα	Vβ8.2 Dβ2 Jβ2.4 Cβ2
$H2K^b$-SIYR-2C	1G6R	8	Vα3 Jα58 Cα	Vβ8.2 Dβ2 Jβ2.4 Cβ2
$H2K^{bm3}$-dev8-2C	1MWA	8	Vα3 Jα58 Cα	Vβ8.2 Dβ2 Jβ2.4 Cβ2
$H2K^b$-pBM1-scBM3.3	1FO0	8	AV17S1 (TRAV16D)	Vβ2S1 (TRBV1)

*Structures solved but not yet published.

Highlighted: the 14 unique structures used for analyses

All complexes are human except for the last set of 6 complexes which are from mouse.

3.3 Results and Discussion

3.3.1 MHC Contact Residues

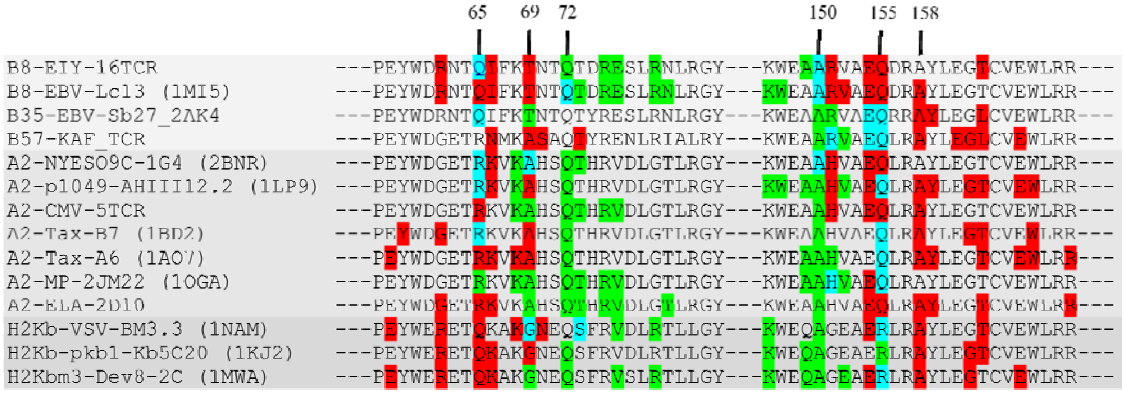


Figure 1: Clustalw alignment of MHC sequences showing residues contacted by the TCR.
Key: red = TCRα contact, green = TCRβ contact, blue = TCRα and TCRβ contact. All contacts are for residues <= 4Å apart. The first block of residues shows the α1 helix, the second block shows the portion of the α2 helix which contacts the TCR.

Analysis of a multiple alignment of the MHC sequences from the 14 complexes (Figure 1) reveals 6 residues on the MHC which are consistently contacted by the TCR in all alleles:

1. 65 (contacted in all complexes except B57-KAF-TCR)
2. 69 (contacted in all complexes)
3. 72 (contacted in all but 3: 2AK4, B57-KAF-TCR and 1NAM)
4. 150 (contacted in all complexes)
5. 155 (contacted in all complexes)
6. 158 (contacted in all but 3: 1MI5, 2BNR and 1OGA)

It can be seen that in most cases (8/14 complexes) all 6 of these residues are contacted and 5 complexes contact 5 of the residues. The minimum number contacted is 4, by B57-KAF-TCR. Figure 2 shows these residues mapped onto the HLA-A2 peptide-binding groove.

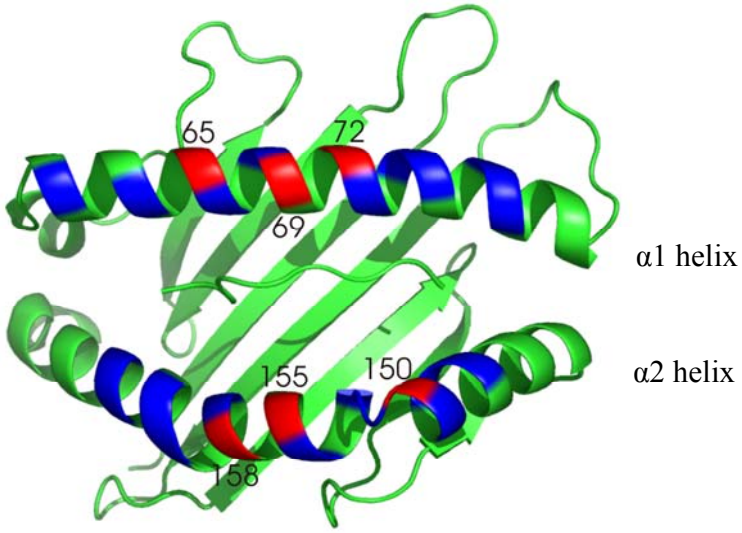


Figure 2. MHC residues contacted by TCRs. The backbone of the HLA-A2 peptide-binding groove is shown, with residues coloured blue to denote often-contacted residues, and red to denote (almost) always-contacted residues (65, 69, 72, 150, 155 and 158).

The alignment in Figure 1 shows that in all complexes, with the exception of 1OGA and B57-KAF-TCR, both the α1 and α2 helices are contacted by both TCR Vα and Vβ domains. 1OGA only has Vβ contacts on the α1 helix and B57-KAF-TCR only has Vα contacts on the α1 helix.

It is not the case that the two TCR domains are specialised to recognise only one of the MHC α helices, as the α2 helix is contacted by both Vα and Vβ domains in all known

structures. Many residues are contacted by Vα in some of the complexes, by Vβ in others and, in some examples, by both Vα and Vβ. Residues on the MHC before 65 and after 155 are only contacted by Vα, and residues after 72 and before 150 are only contacted by Vβ due to the diagonal orientation of all the TCRs.

For 1OGA, there are 6 contacts in total on the α2 helix but all these are between A149 and Q155, whereas all but 1 of the other complexes (2BNR) have contacts extending beyond this point.

There are some residues on the MHC helices which are consistently contacted *only* by one of the TCR V domains. Moreover, some of these residues are HLA allele-specific. For example, in HLA-B8 and H2Kb, residue 62 is an R and is only contacted by Vα. In HLA-A2, this residue is a G and is not contacted at all (with the exception of 1BD2 and A2-ELA-2D10). Residue 79 is R in B8 and H2Kb but a G in A2. It is often contacted in B8 and H2Kb but never contacted in A2. This may represent a pattern that defines the difference between A2 and other alleles *i.e.* if there is an R at these positions it is likely to be used as a contact point on the MHC, but if it is a G it is not. However, complex structures for other HLA-A alleles are not available so this pattern could be A2-specific.

Figure 3 shows sequence logos (Crooks, Hon et al., 2004) of all human HLA alleles and reveals that there is a predominance of Q or R at consistently contacted MHC points. Figure 3(d) shows that for all HLA alleles residue 65 is usually a Q, sometimes R and very occasionally a G (HLA-A alleles only), 69 is either T, A or R and residue 72 is

always Q. Figure 3(h) shows residue 150 is always A (with the exception of a few HLA-A alleles in which it is V – see Figure 3(e)), residue 155 is always Q and residue 158 is nearly always A. Within the HLA-A, B and C groups, the core 'footprint' amino acids have unique combinations of amino acids, with primary polymorphism at positions 65 and 69. Such sets of amino acids may play an important role in defining the restriction of a given TCR to a specific HLA-allele.

As shown in Figures 11-13, 2AK4 is shifted and tilted towards the α2 helix and there are only 2 contacts on the α1 helix: Q65 is contacted by both Vα and Vβ and T69 by Vβ only. However, despite the significant shift and tilt and the 13 residue super-bulged peptide, there are no residues contacted on the α2 helix by 2AK4 that are not contacted in any other complex, and 5 out of the 6 key MHC residues are contacted, the exception being Q72. It has been suggested that the MHC residues contacted by the TCR in this complex represent the minimum requirements for MHC restriction (Tynan, Burrows et al., 2005).

Figure 3a. HLA-A alleles, α-1 helix residues

Figure 3b. HLA-B alleles, α-1 helix residues

Figure 3c. HLA-C alleles, α-1 helix residues

Figure 3d. All alleles, α-1 helix residues

Figure 3e. HLA-A alleles, α-2 helix residues

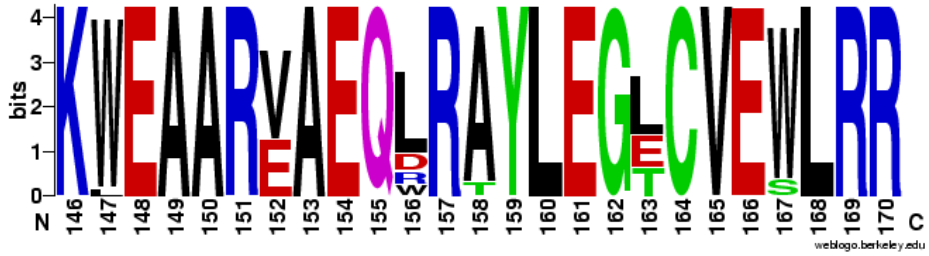


Figure 3f. HLA-B alleles, α-2 helix residues

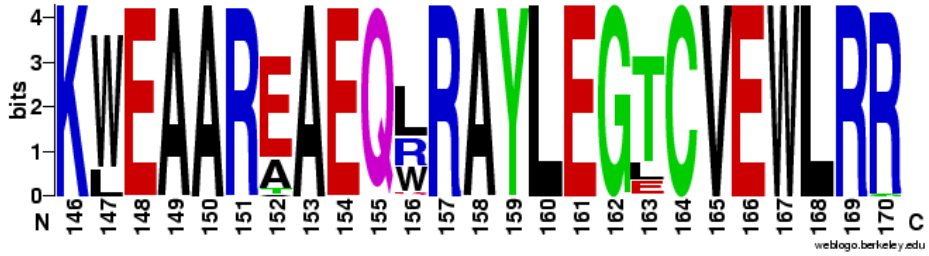


Figure 3g. HLA-C alleles, α-2 helix residues

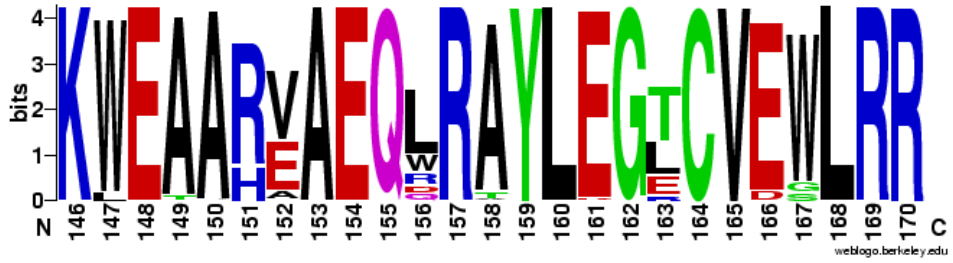


Figure 3h. HLA-C alleles, α-2 helix residues

A recent paper postulates that TCRs which survive negative selection may only retain a few of the conserved interactions with MHC, sufficient for them to bind MHC but not enough for them to be negatively selected (Huseby, White et al., 2005). This leads to two questions:

- Is there a maximum number of contacts which, if exceeded, will lead to negative selection, or is it the specific combination rather than number that is important?
- Is there a core set of residues for each MHC allele that must be contacted for the TCR to recognise and/or bind?

The second question has been partially addressed by one group (Baker, Turner et al., 2001) who showed that mutagenesis of only three amino acids (R65, K66 and A69) affected TCR responses to the Tax-HLA-A2 complex. In this work, each of the 15 amino acids on HLA-A2 that are contacted by the A6 TCR were mutated in turn and the effects on T cell activation measured. R65A and K66A mutations produced a >1000-fold reduction in T cell effector assays and the A69G mutation produced a 10-100-fold reduction. The effects on TCR binding to Tax-HLA-A2 by these three mutants was then assessed and showed that R65A and K66A mutations had large effects on binding kinetics, specifically a reduction of K_{on}. Notably, the authors did not find that a Q155A mutation had any effect on T cell activation. However, each mutation was done singly so perhaps binding to Q155 is more important for T cell positive selection in the thymus than for T cell recognition of MHC in the periphery. Alternatively (or additionally), mutation of Q155 alone may not be sufficient to affect T cell activation. The effects on binding kinetics of this mutation were not analysed. Finally, the authors found that TCRs

specific for the influenza virus $M1_{58-66}$ peptide presented by HLA-A2 are not affected by these three mutations, suggesting that the focus of these TCRs is not on the same part of HLA-A2. They also mention that preliminary results suggest Q155 is important for the recognition of HLA-A2 presenting the M1 peptide. In a later paper, the same group found that most HLA-A2-restricted TCRs are dependent on recognition of K66 and Q155 sidechains (Wang, Turner et al., 2002).

It would be interesting to know what effect on binding would occur if more than one of the other 12 amino acids were mutated at once, to find out if there is a critical 'threshold' for binding and activation of a TCR or if the three residues investigated are sufficient.

Overall, the evidence suggests that different combinations of the 6 key residues on the MHC are important for the binding and activation of different TCRs *i.e.* TCRs specific for different pMHC complexes have different 'footprints' or 'energetic hotspots' on the MHC, which consist of a subset of the 6 key residues plus one or more other residues in the same vicinity. This combination appears to be HLA-A, -B or C-specific, raising the possibility that peripheral TCR repertoire is segmented into allele-specific groups.

3.3.2 TCR Contact Residues

Looking at which TCR residues contact each of the 6 key MHC residues (Figure 4) reveals several previously unrecognised features of TCR binding: TCRα hardly touches MHC 150 but dominates MHC 155 contacts, TCRβ contacts MHC 72, usually using the CDR2β loop, and the TCRβ CDR3 loops nearly always contacts MHC 150. Most importantly, every complex except 1OGA uses at least one aromatic residue (Y, F, or H for B8-EIY-16TCR) to contact Q155. This residue appears to preferably be in the CDRα loops, but 1KJ2 uses a W in CDR3β instead. Note that B8-EIY-16TCR does not have F or Y in any of the CDRα loops which could explain the use of H50. There is no such obvious pattern of use of aromatic residues in TCRβ.

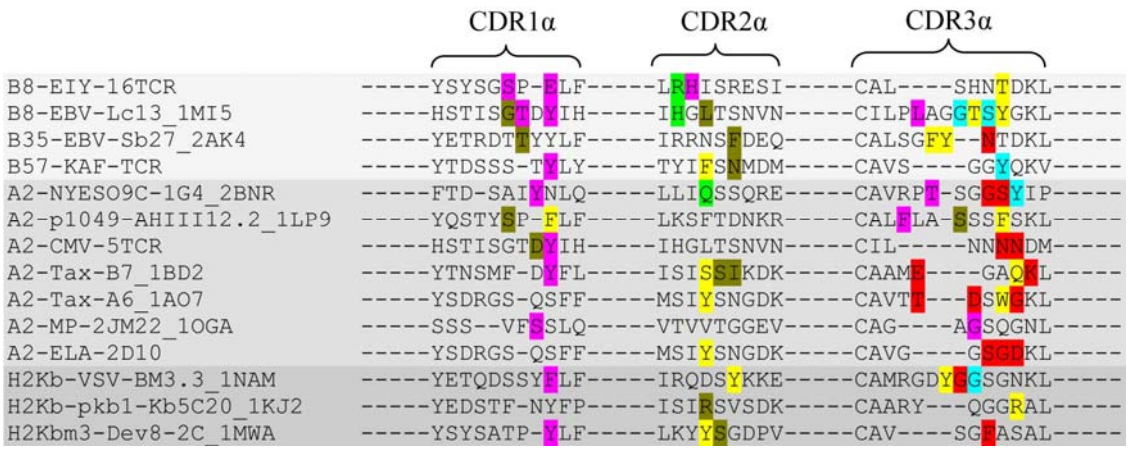


Figure 4a: Clustalw alignment of TCRα sequences showing residues contacting the MHC. All contacts are for residues <= 4Å apart. Key: red = contact to MHC 65, cyan = contact to MHC 69, green = contact to MHC 150, magenta = contact to MHC 155, olive = contact to MHC 158, yellow = contact to more than one MHC residue.

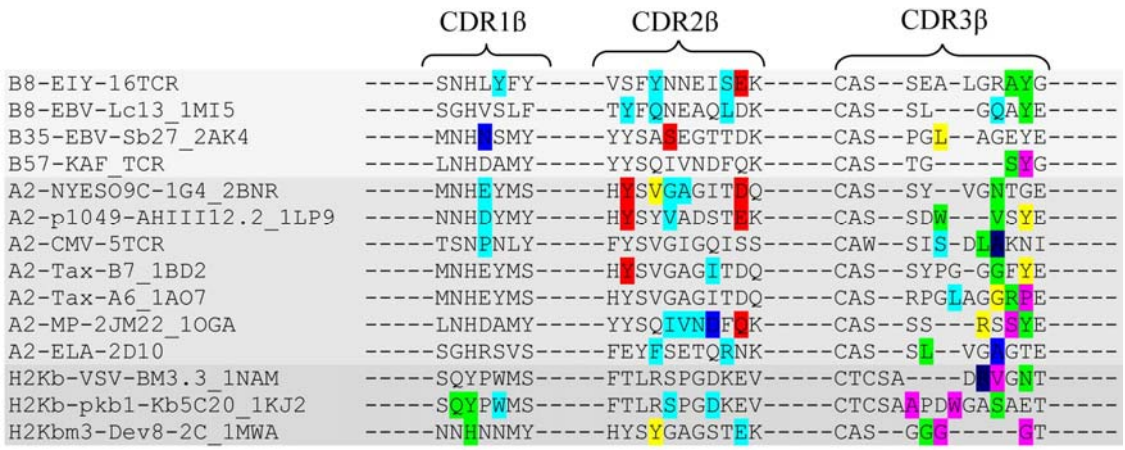


Figure 4b: Clustalw alignment of TCRβ sequences showing residues contacting the MHC. All contacts are for residues <= 4Å apart. Key: red = contact to MHC 65, blue = contact to MHC 69, cyan = contact to MHC 72, green = contact to MHC 150, magenta = contact to MHC 155, yellow = contact to more than one MHC residue, *e.g.* residue L98 of 2AK4 contacts both MHC 150 and 155.

There is evidence that aromatic residues play a key role as 'anchors' in many protein-protein interfaces (Rajamani, Thiel et al., 2004). Most notably, aromatic residues, particularly tyrosines, are found with high frequency in the complementary determining regions of antibodies (Padlan 1990). The authors here suggest that aromatic side chains have a greater potential contribution to the total binding energy which is why they are especially suited for interaction with ligands.

These observations suggest that, in general, it is the TCRα which fixes onto the MHC using residues 65 on the α1 helix and 155 on the α2 helix, and TCRβ adapts and contacts what it can access. It has previously been postulated that Vα is critical for TCR selection in the thymus and in mature T cell activation, with an important but more variable contribution from Vβ (Reinherz, Tan et al., 1999; Hennecke, Carfi et al., 2000; Rudolph and Wilson 2002; Wang and Reinherz 2002). This necessity for Vα to contact specific

MHC residues could also shed light on why the TCR α and β always have the same diagonal orientation, with the α domain situated over α2 helix and the β domain over the α1 helix. It has also been previously observed that the location and size of the TCR Vβ footprint on the MHC varies much more than that of the Vα footprint, suggesting that different TCRs pivot around contacts made by Vα (Hennecke and Wiley 2001). In addition, a two-step mechanism for TCR binding to pMHC has been proposed (Wu, Tuot et al., 2002), where the TCR will initially encounter the MHC helices and dock in a way that is mainly independent of peptide, which then enables peptide sampling by the flexible CDR3 loops.

In summary, the TCR CDR loops are being used to contact some combination of the 6 key residues in the MHC, but not by using conserved TCR residues; the exact way each TCR is contacting the 6 MHC residues is different for each TCR but aromatic residues appear to be important for contacting residue 155 on the MHC α2 helix.

Figure 5 shows that side chains in some positions are more flexible than others *i.e.* they are not confined to one conformation in all complexes. 3 of the key contacted side chains are in fairly fixed positions (69, 150 and 158), whereas residues 65 and 155 are more variable in conformation and residue 72 varies slightly. In mouse alleles, R155 is less flexible than human Q155. It is possible that residues 65 and 155 are more flexible to adapt to the TCR, whereas residues 69, 72, 150 and 158 are more fixed 'anchor' residues.

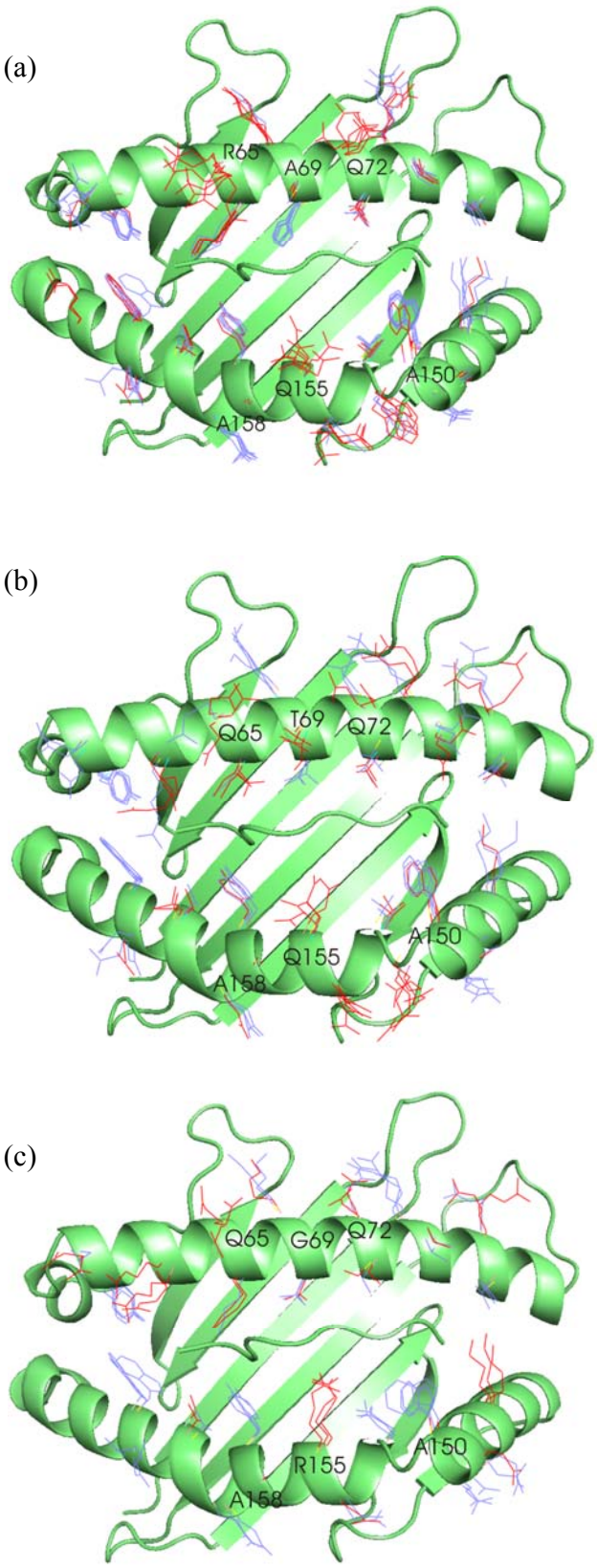


Figure 5. Residues contacted by TCRs on HLA-A alleles (a), HLA-B alleles (b) and mouse H2-K^b alleles (c). In each case, the backbone of HLA-A2 is shown with sidechains at contact positions displayed for each TCR-pMHC complex for the relevant alleles. Sidechains coloured red are contacted, sidechains coloured blue are not contacted.

3.3.3 Average Contact Points on MHC for Each Complex

All residues within 4Å of each other between TCRα and TCRβ and the MHC and between TCRα and TCRβ and the peptide were obtained and an average of all these coordinates was then calculated. Figure 6 shows that this point is remarkably similar for each complex, focussing on the centre of the peptide in each case. This is consistent with previous observations that centrally-located peptide residues are the major determinants of T cell activation (Degano, Garcia et al., 2000).

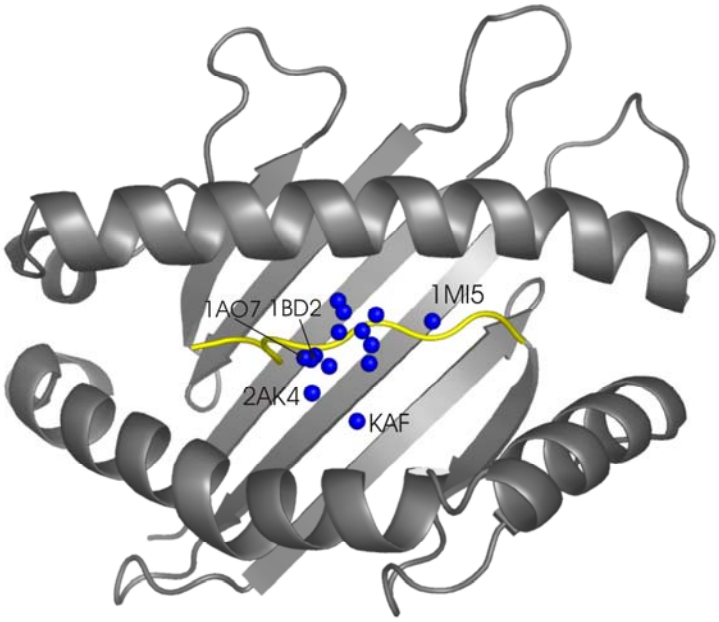


Figure 6. Average contact points for each complex superimposed onto HLA-A2 peptide-binding groove, viewed by looking directly down the normal to the β-sheet plane. KAF refers to the B57-KAF-TCR complex.

1AO7 and 1BD2 are labelled to show that two different TCRs (B7 and A6) which bind to the same pMHC (Tax-HLA-A2) focus on exactly the same point.

In Figure 6, it appears that there are 3 'outliers' which are not on the centre of the peptide. However, this is not the case when the actual peptides for each of these complexes are viewed in the MHC groove rather than the 'representative' MP peptide from 1OGA (Figure 7 (a-c)).

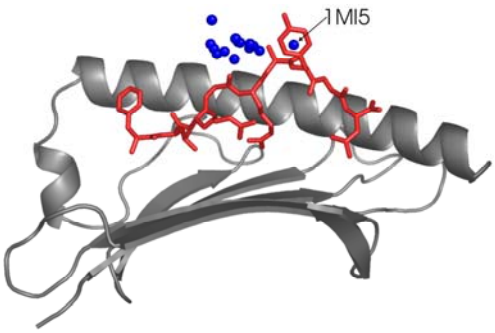


Figure 7(a). 1MI5

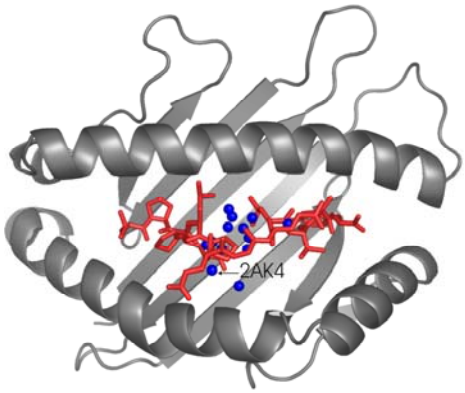


Figure 7(b). 2AK4

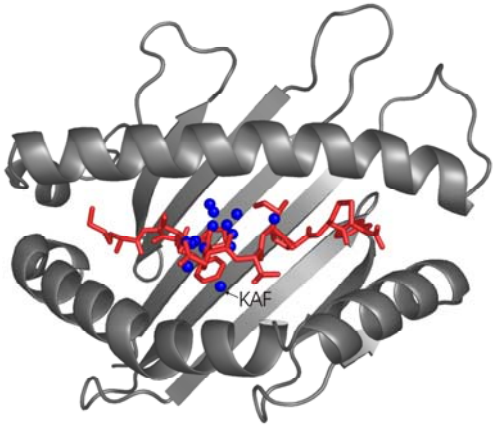


Figure 7(c). B57-KAF-TCR

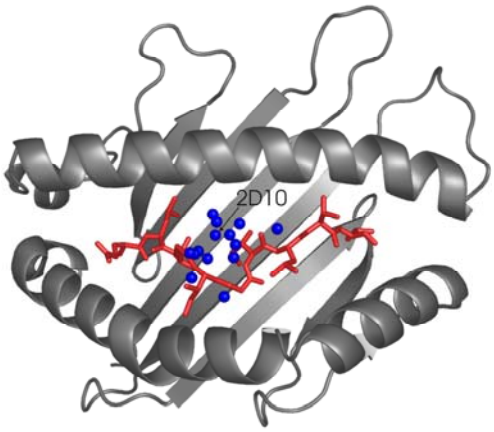


Figure 7(d). A2-ELA-2D10

Figure 7. Average contact points for each complex superimposed onto HLA-A2 peptide-binding groove but with peptides from individual complexes shown in the groove.

It is clear that the TCR tends to focus on the centre of the peptide bulge rather than necessarily the central peptide residues. However, despite the sideways bulge of the ELA

peptide in the A2-ELA-2D10 complex, the average contact point is remarkably central and not shifted towards the α2 helix of the MHC (Figure 7(d)). This may be due to restraints imposed by the peak in the α2 helix of the MHC.

3.3.4 TCR Heights

Table 2 shows various measurements relating to the height of the TCR from the MHC (The diagonal heights take into account the tilt of the TCR. See Methods and Appendix B for an explanation of these figures).

Table 2. Heights of TCR and MHC

TCR-peptide-MHC	PDB#	Perp. height of TCR from βsheet (Å)	Diag. height TCR from βsheet (Å)	β sheet to interface (Å)
A2-MP-2JM22	1OGA	31.7	31.8	16.1
A2-tax-A6	1AO7	32.7	33.0	16.3
A2-tax-B7	1BD2	31.6	32.2	15.8
A2-p1049-AHIII12.2	1LP9	32.6	32.7	15.5
A2-CMV-5TCR	-	31.0	31.7	16.0
A2- NYESO9C-1G4	2BNR	31.7	32.2	16.0
A2-ELA-2D10	-	30.4	30.4	14.5
B8-EIY-16TCR	-	31.8	32.1	15.5
B8-EBV-Lc13	1MI5	30.7	31.7	15.4
B35-13mer-TCR	2AK4	34.0	34.5	18.7
B57-KAF-TCR	-	32.6	32.9	16.9
$H2K^{b}$-VSV-Bm3.3	1NAM	32.3	33.2	15.1
$H2K^{b}$-pkb1-Kb5C20	1KJ2	31.6	31.8	15.3
$H2K^{bm3}$-dev8-2C	1MWA	30.6	30.6	15.7

Figure 8 shows how the average contact points cluster when viewed side-on from two different angles (0˚ and 90˚) making it clear that they are all at very similar heights, despite the variation in peptide bulge. The point for 2AK4 is only slightly raised even though the peptide bulge extends upwards much more than for any other complex.

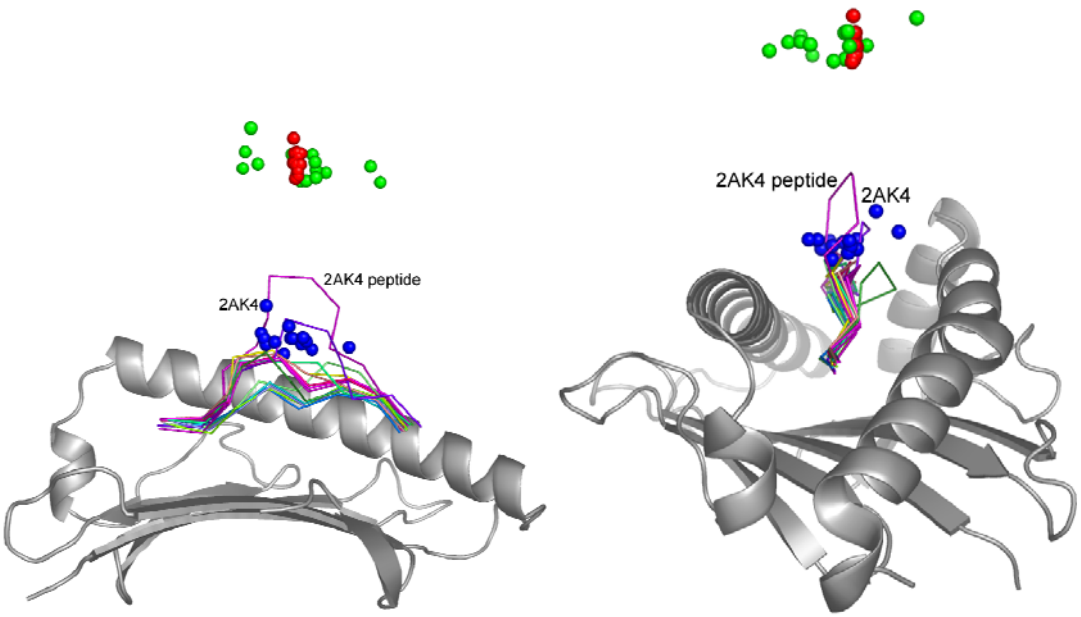


Figure 8. Average contact points (blue dots), perpendicular TCR heights (red dots) and diagonal TCR heights (green dots) viewed at 0˚ and 90˚.

Figure 8 also shows that the perpendicular TCR heights are remarkably similar despite differences in the extent of peptide bulge; the perpendicular heights range from 30.4 Å to 34.0 Å whereas the approximate difference between the midpoint of the flattest peptide and the midpoint of the super-bulged peptide (2AK4) is 11.8 Å (Figure 9).

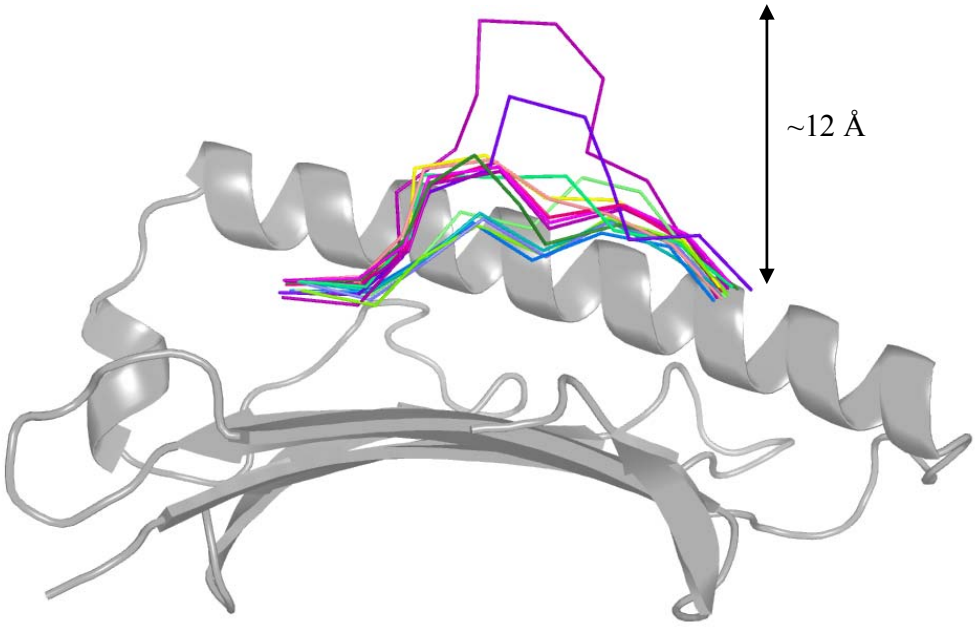


Figure 9. The difference in heights of the different peptides bound to class I MHC molecules is approximately 12 Å. This measurement is between the Cα atom of arginine 5 in the EIY peptide and the Cα atom of glycine 8 in EBV peptide.

3.3.5 Analysis of the Orientation of TCRs in TCR-pMHC Complexes

3.3.5.1 TCR twist

It has been hypothesised that TCRs specific for pMHC class I ligands are permitted a broader range of twists whereas TCRs specific for pMHC class II ligands tend to be more orthogonal (Wang and Reinherz 2002). A theory has also been put forward that the expression of the TCR coreceptors CD4 or CD8 is dependent on the mode of binding of the TCR in the thymus: more orthogonal docking onto pMHC class II will coengage CD4 whereas a diagonal docking mode onto pMHC class I will coengage CD8, and in each

case expression of the other coreceptor is stopped (Reinherz, Tan et al., 1999). However, this work compares all five available class II complex structures to the 14 class I complex structures and shows that class II molecules do not always bind in a more orthogonal way compared to class I molecules (Table 3 and Figure 10).

Table 3. Twist angles of TCR on pMHC for MHC class I complexes

TCR-peptide-MHC	PDB#	Angle calculated (°)	Resolution (Å)
Class I			
A2-MP-2JM22	1OGA	62.5	1.4
A2-tax-A6	1AO7	32.0	2.6
A2-tax-B7	1BD2	46.9	2.5
A2-p1049-AHIII12.2	1LP9	67.0	2.0
A2-CMV-5TCR	-	70.5	2.6
A2- NYESO9C-1G4	2BNR	67.4	1.9
A2-ELA-2D10	-	32.7	2.5
B8-EIY-16TCR	-	48.0	2.3
B8-EBV-Lc13	1MI5	42.6	2.5
B35-13mer-TCR	2AK4	69.2	2.5
B57-KAF-TCR	-	32.9	2.4
$H2K^{b}$-VSV-Bm3.3	1NAM	36.3	2.7
$H2K^{b}$-pkb1-Kb5C20	1KJ2	29.2	2.7
$H2K^{bm3}$-dev8-2C	1MWA	21.1	2.4
Class II			
HLA-DR1-Flu HA1-TCR	1FYT	47.394	2.6
MHC I-AK-CA-D10TCR	1D9K	56.594	3.2
HLA-DR4-Flu HA1-TCR	1J8H	48.454	2.4
MHC I-Au-MBP-TCR172.10	1U3H	46.417	2.4
HLA-DR2b-MBP-TCR	1YMM	90.008	3.5

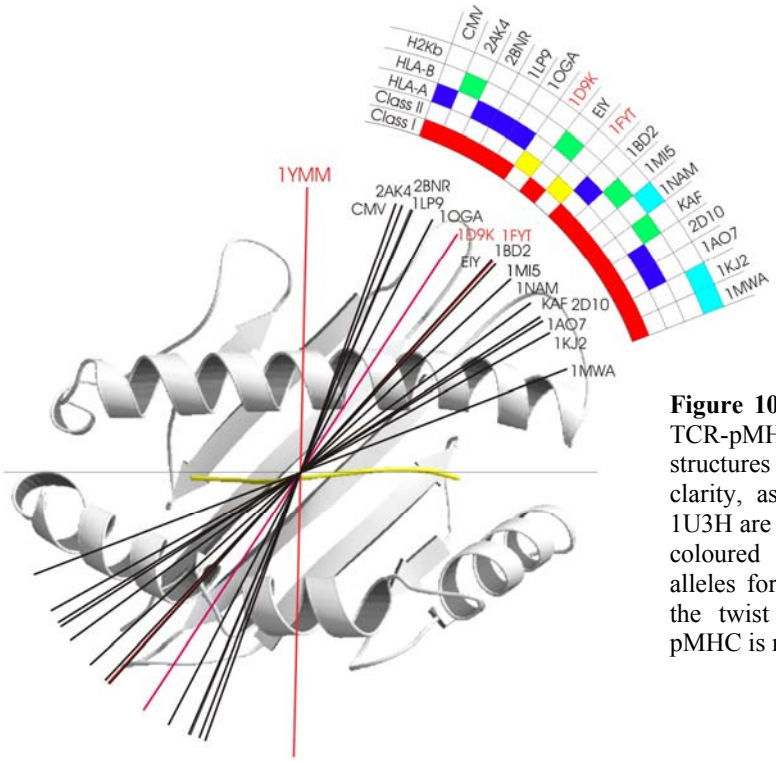


Figure 10. Range of twists of TCRs in TCR-pMHC complexes. Only 3 class II structures are included in this figure for clarity, as twists for 1FYT, 1J8H and 1U3H are almost identical (Table 3). The coloured boxes represent the MHC alleles for each complex and show that the twist of a TCR when bound to pMHC is not dependent on MHC allele.

There is little difference in the range of twist for HLA-A and HLA–B alleles. Mouse alleles appear to have a smaller range than human alleles in that they are less orthogonal (mouse allele twists range from 21.1˚ – 36.3˚ whereas human HLA-A range from 32.0˚ - 70.5˚ and human HLA-B alleles range from 32.9˚ – 69.2˚). There is no clear separation of twist angles for class I and class II molecules, or of class I alleles.

A generally diagonal orientation is probably optimal for TCR recognition of both peptide and MHC. However, TCRs always twist the same way: all TCRs have a twist less than 90˚ and Vα always sits over the α2 helix and Vβ always over the α1 helix. In a database this size, it is unlikely to be due to chance and may be explained by essential contact

residues on the MHC as discussed above *i.e.* it seems that TCRα CDR aromatic residues are used to contact residue 155 on the MHC α2 helix and TCRβ CDR residues are used to contact residue 72 on the MHC α1 helix.

Why do none of the TCRs twist more than 90˚? It is possible that TCRs will simply not bind sufficiently well if the twist if greater than 90˚ because the TCRα would be sitting either over the peak in the α2 helix or over a smaller portion of the MHC α2 helix to the right of this peak.

3.3.5.2 TCR tilt

Table 4. Tilt angles of TCR on pMHC for MHC class I complexes

TCR-peptide-MHC	PDB#	Angle (°)	Direction of Tilt (°)
A2-MP-2JM22	1OGA	18.6	121.1
A2-tax-A6	1AO7	13.2	-168.9
A2-tax-B7	1BD2	16.8	153.5
A2-p1049-AHIII12.2	1LP9	9.8	117.9
A2-CMV-5TCR	-	15.9	94.9
A2- NYESO9C-1G4	2BNR	19.9	77.9
A2-ELA-2D10	-	1.4	-45.1
B8-EIY-16TCR	-	8.1	122.8
B8-EBV-Lc13	1MI5	10.0	54.2
B35-13mer-TCR	2AK4	5.8	-59.8
B57-KAF-TCR	-	3.5	151.9
$H2K^{b}$-VSV-Bm3.3	1NAM	16.7	69.0
$H2K^{b}$-pkb1-Kb5C20	1KJ2	0.7	-175.3
$H2K^{bm3}$-dev8-2C	1MWA	8.6	49.6

It appears that TCRs cannot tilt very much towards the MHC α2 helix (Table 4 and Figure 11). This may indicate that the MHC is to some extent defining the 'rules of

engagement' of the TCR due to the peak in the α2 helix. The exception to this is 2AK4, but this tilt is relatively small (5.8˚) and the TCR is sitting higher up in relation to the pMHC compared to most other complexes, due to the super-bulged peptide. In addition, 2AK4 is shifted towards the N terminus of the peptide (next section).

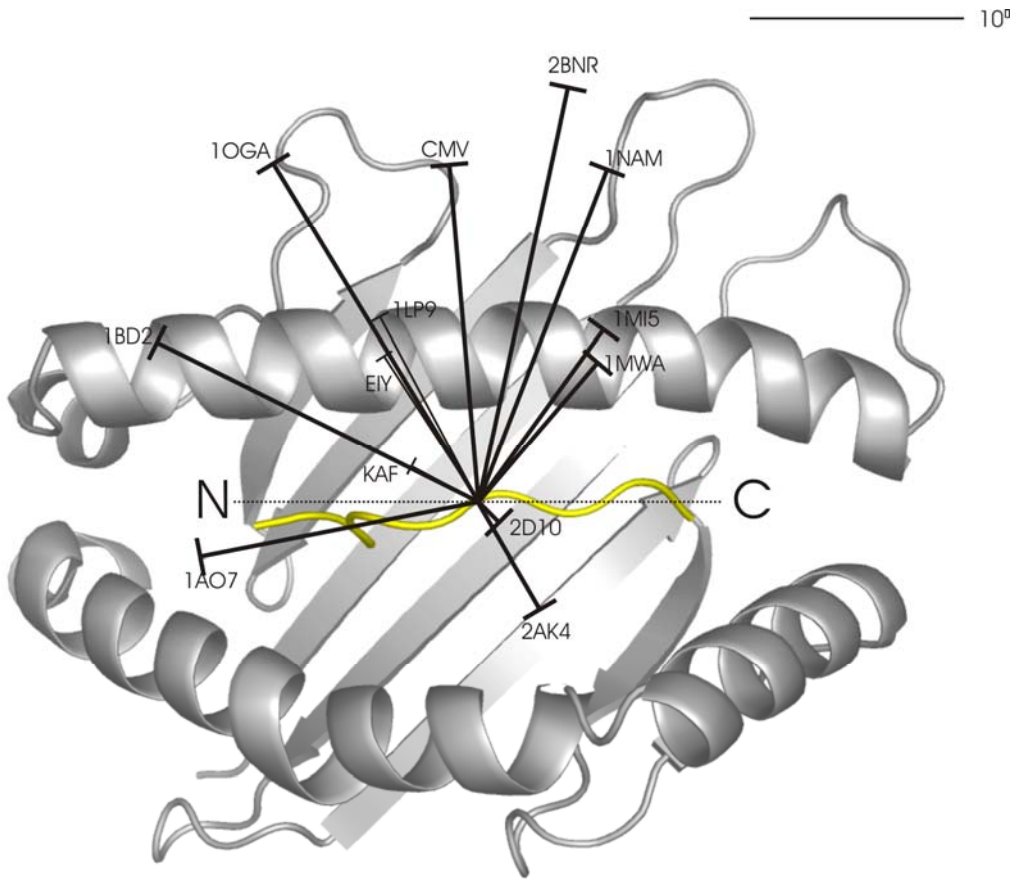


Figure 11. Range of tilts of TCRs in TCR-pMHC complexes. The length of each line is proportional to degree of tilt as shown by the scale bar. 1KJ2 is not shown as the tilt angle is too small (0.7˚).

3.3.5.3 TCR shift

Table 5. Shift of TCR on pMHC for MHC class I complexes

TCR-peptide-MHC	PDB#	Length of shift of TCR (Å)	Angle of shift of TCR (°)
A2-MP-2JM22	1OGA	7.2	-131.2
A2-tax-A6	1AO7	1.6	52.7
A2-tax-B7	1BD2	1.8	49.2
A2-p1049-AHIII12.2	1LP9	2.1	-118.5
A2-CMV-5TCR	-	2.7	82.1
A2- NYESO9C-1G4	2BNR	0.9	-175.8
A2-ELA-2D10	-	1.8	105.1
B8-EIY-16TCR	-	3.5	139.7
B8-EBV-Lc13	1MI5	5.7	164.0
B35-13mer-TCR	2AK4	5.3	-43.6
B57-KAF-TCR	-	4.8	169.8
$H2K^{b}$-VSV-Bm3.3	1NAM	4.2	-177.1
$H2K^{b}$-pkb1-Kb5C20	1KJ2	3.8	6.9
$H2K^{bm3}$-dev8-2C	1MWA	0.9	-92.7

TCRs appear to be able to shift around the surface of the MHC in any direction, but the size of this shift seems to be restricted by the α1 and α2 helices (Table 5 and Figure 12). 1AO7 and 1BD2 have almost identical shift directions and magnitudes and both use Vβ12.3, but this is the only correlation seen between Vα/Vβ usage and TCR twist, tilt or shift so could simply be coincidence.

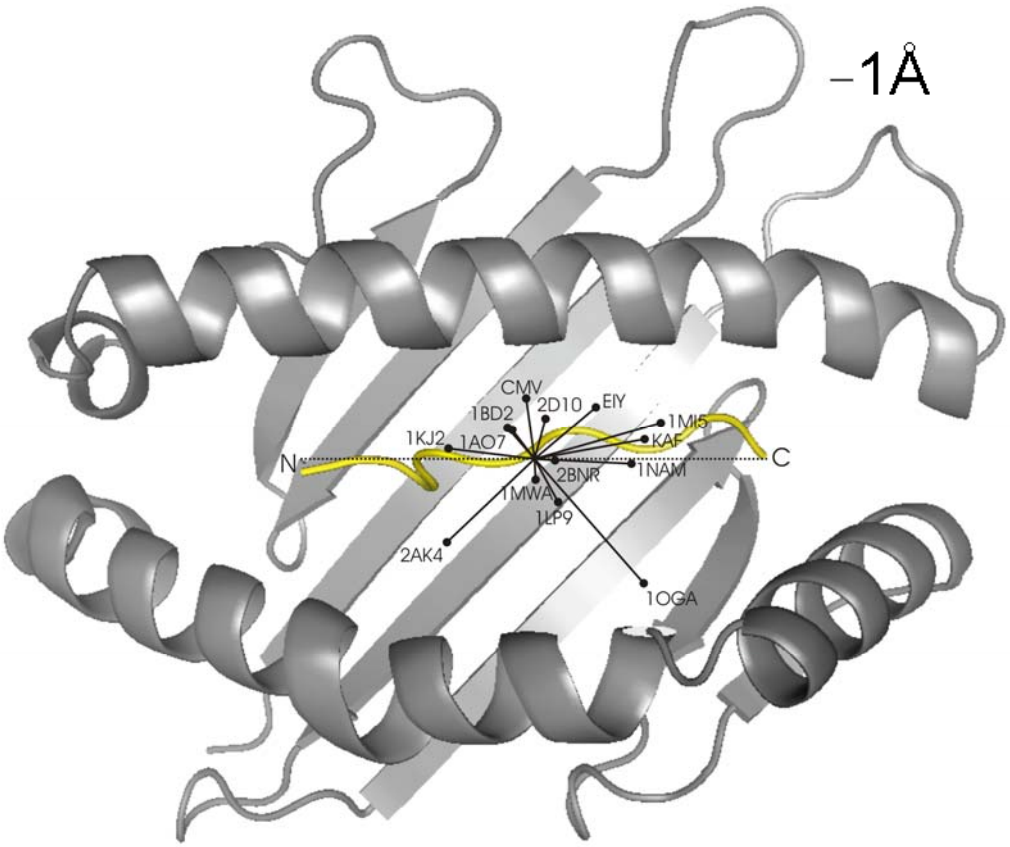


Figure 12. Range of shifts of TCRs in TCR-pMHC complexes.

3.3.5.4 TCR shift plus tilt

Figure 13 illustrates the combination of tilt and shift of TCRs. It is immediately clear that most TCRs are concentrated towards the α1 helix, with the sole exception of 2AK4. This complex is unusual due to a super-bulged peptide which forces the TCR to sit slightly higher than average, possibly explaining why it is able to adopt this more unusual conformation. Note, however, that it is not sitting near the peak in the α2 helix, further supporting the view that this peak restricts the range of orientations of TCRs. In addition,

the 2JM22 TCR is shifted very close to the α2 helix peak, but tilts dramatically away from it.

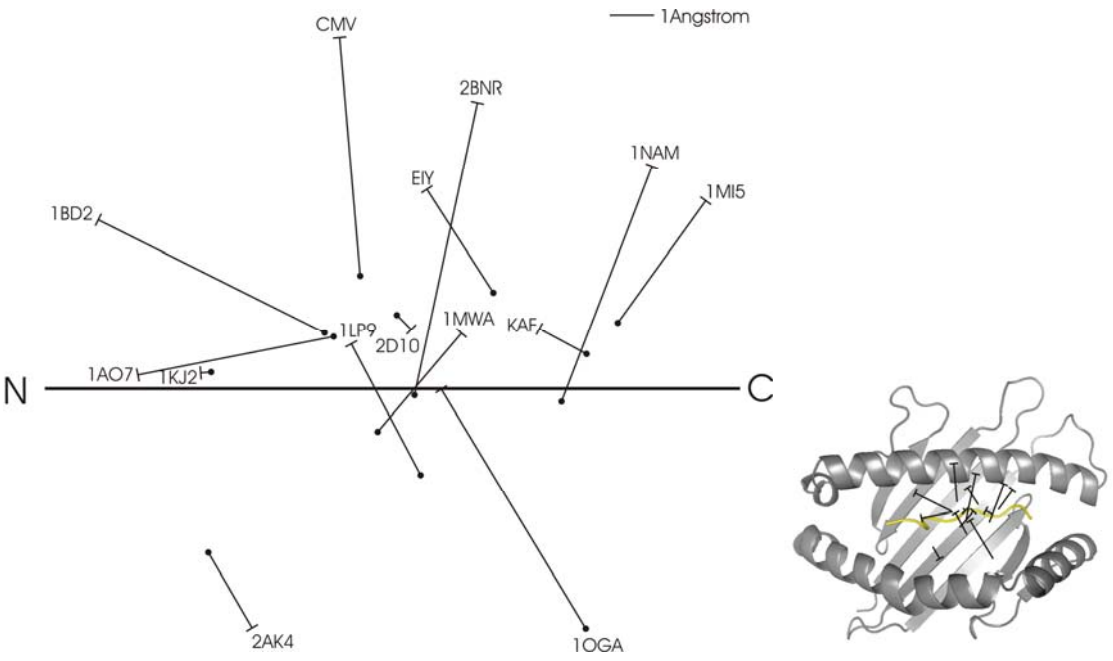


Figure 13. Shifts and tilts of TCR on pMHC for MHC class I complexes. Dots refer to the position the TCR is shifted to and T-bars represent direction and length of tilt. The scale refers to the shift. Inset is a scaled representation of this diagram.

Presumably, the combination of tilt, twist and shift allows the TCR to contact the essential residues on the MHC α1 and α2 helices and may also affect the interaction with CD8. However, there are many unanswered questions: is there an optimal combination of tilt, twist, shift and height of a TCR which would produce optimal signaling, or is it not critical so long as the TCR contacts the essential residues on the MHC? Is the

strength and/or duration of signal dependent on the orientation of the TCR on the MHC or just on the pMHC residues contacted?

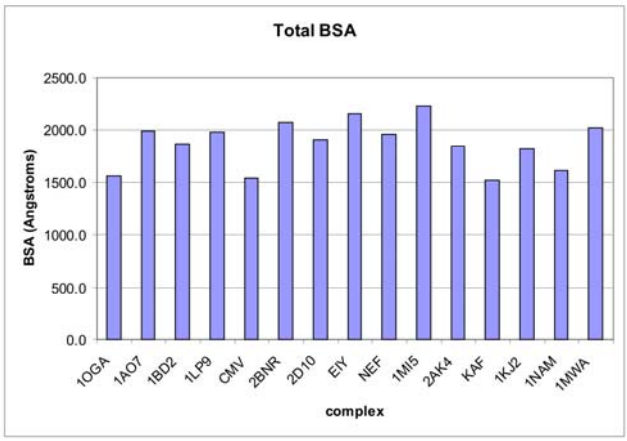
3.3.6 Buried Surface Areas

The total BSA between the TCR and pMHC does not vary as widely as one may expect, given the significant variability in TCR tilt, twist and shift (Table 6 and Figure 14(a)). Figure 14(b) shows that for most complexes, TCRα usually dominates the BSA except for 1OGA, A2-CMV-5TCR, 2BNR and 1KJ2. For 1OGA, TCRβ buries more than double the amount of MHC compared to TCRα.

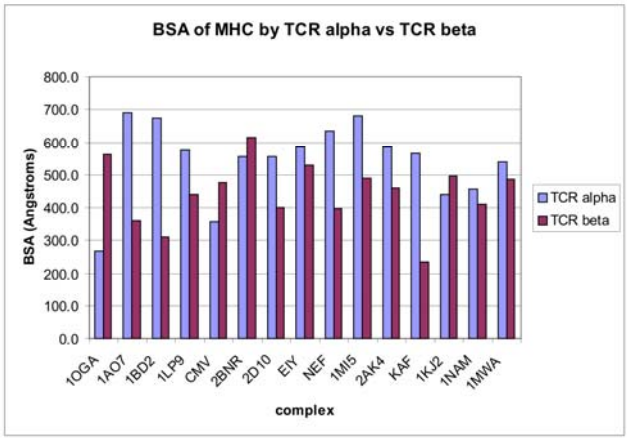
Figure 14(c) shows the contribution of the MHC versus the peptide to the overall BSA of the pMHC. It can be seen that for most complexes, the MHC contributes much more to the BSA of the pMHC than the peptide does. However, for 2AK4 and B57-KAF-TCR, the contributions are much more similar, presumably due to the super-bulged peptides in these complexes. Note that although the peptide in 2AK4 bulges out more than that in B57-KAF-TCR, the contribution of the peptide to the BSA of the pMHC is more for B57-KAF-TCR than for 2AK4.

Table 6. Buried surface areas of class I complexes. All numbers are in Å^2.

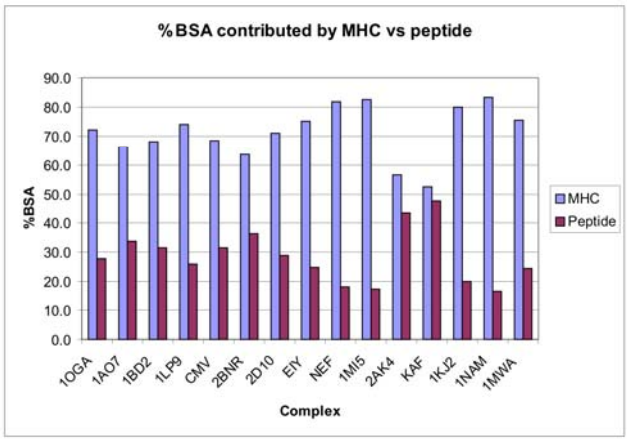
Complex	PDB	Ttotal BSA between TCR and pMHC	BSA of pMHC	BSA of TCR	BSA of pMHC by TCRα	BSA of pMHC by TCRβ	% BSA of pMHC buried by TCRα	% BSA of pMHC buried by TCRβ	%MHC BSA buried by TCR	%pep BSA buried by TCR
2_JM22-MP-A2	1OGA	1564.4	775.0	789.4	268.6	560.5	32.4	67.6	72.1	28.0
A6-tax-A2	1AO7	1990.0	987.8	1002.2	690.7	360.0	65.7	34.3	66.3	33.7
B7-tax-A2	1BD2	1871.4	952.8	918.6	675.3	311.4	68.4	31.6	68.2	31.8
AHIII12.2-p1049-A2	1LP9	1979.8	947.5	1032.3	575.8	440.8	56.6	43.4	73.9	26.1
5TCR-cmv-A2	CMV	1538.7	774.2	764.5	356.7	475.1	42.9	57.1	68.5	31.5
21_1G4_NYESO-A2	2BNR	2077.0	1078.3	998.7	554.7	614.6	47.4	52.6	63.5	36.5
2D10_ELA_A2	2D10	1909.7	933.6	976.1	555.0	400.9	58.1	41.9	71.0	29.0
17_16TCR-EIY-B8	EIY	2161.5	1038.4	1123.1	584.4	529.9	52.4	47.6	75.0	25.0
18_15TCR-nef_B8	NEF	1960.6	962.5	998.1	633.6	398.3	61.4	38.6	81.7	18.3
Lc13-ebv-B8	1MI5	2235.3	1108.5	1126.8	681.1	488.0	58.3	41.7	82.4	17.6
TCR_13mer_B35	2AK4	1848.9	966.1	882.8	586.3	459.3	56.1	43.9	56.5	43.5
TCR-KAF-B57	KAF	1518.6	741.7	776.9	563.8	234.3	70.6	29.4	52.3	47.7
Kb5C20-pkb1-H2K^b	1KJ2	1821.0	895.9	925.1	438.5	494.4	47.0	53.0	80.0	20.1
BM3.3-vsv-H2Kb	1NAM	1612.9	814.3	798.6	456.9	409.7	52.7	47.3	83.13	16.89
2C-dev8-H2K^{bm3}	1MWA	2020.2	1005.7	1014.5	539.0	484.7	52.7	47.3	75.5	24.5



(a)



(b)



(c)

Figure 14. Buried surface areas within TCR-pMHC complexes (Å^2).

3.4 The New TCR-pMHC Complexes

The above analyses were carried out on the class I complexes published in 2007 to see if the conclusions drawn above were valid for these structures.

The first complex is the ELS4 TCR complexed to HLA-B*3501 bound to an 11-residue Epstein-Barr virus (EBV) peptide (PDB identifier 2NX5). The complex of the SB27 TCR complexed to HLA-B*3508 bound to a 13-residue EBV peptide (PDB identifier 2AK4) was included in the database of 14 structures in Table 1.

The second complex is the 2C TCR complexed to its foreign ligand, H-2L^{d} bound to the QL9 peptide (PDB identifier 2OI9). The same TCR bound to its self ligand, H-2Kb bound to the dEV8 peptide, was included in the database of 14 structures in Table 1 (PDB identifier 1MWA).

3.4.1 The HLA-B*3501-EBV-ELS4 complex

Five of the six key residues on the MHC are contacted by the ELS4 TCR, the exception being residue 72. Residue 65 is contacted by both TCRα and TCRβ, residues 69, 150 and 155 by TCRβ, and residue 158 by TCRα. Notably, Q155 is not contacted by any aromatic TCR residues in this complex. However, three different tyrosine residues on TCRβ (Y31, Y48 and Y50) are used to contact MHC Q65, and Y31 is also used to contact MHC T69. Unusually, rather than utilizing CDR loop flexibility to accommodate the bulged peptide, or sitting on top of the bulged peptide as in the 2AK4 complex, the

binding of this TCR to the pMHC flattens the peptide. This may explain why aromatic residues are not essential for binding to MHC Q155 as in all other complexes, conformational changes of the pMHC on TCR binding are minimal.

The binding orientation of the TCR is not unusual, with a twist of 57˚ (the observed range being 21.1˚-70.5˚), and a tilt angle of 9.6˚ (observed range 0.7˚-19.9˚) with a direction angle of 172.1˚, meaning it tilts away from the peak in the MHC α2 helix as for other TCRs. It has a mid-range shift of 2.6 Å, at an angle of -9.9˚, which is also away from the peak in the MHC α2 helix. The perpendicular height of the TCR is 32.3 Å which is typical of class I complexes.

The buried surface area between the TCR and the pMHC is 2107 Å which is towards the top end of the observed range (1518 $Å^2$-2235 $Å^2$). TCRα buries 495 $Å^2$ of the MHC and TCRβ buries 419 $Å^2$, similar values to those for 1NAM. The peptide contributes 25.4% of the BSA of the pMHC and the MHC contributes 74.6% which are typical values for class I complexes.

3.4.2 The H-2L^d-QL9-2C complex

All six key residues on the MHC are contacted by the 2C TCR. Residues 65, 150, 155 and 158 are contacted by TCRα and residues 69 and 72 by TCRβ. MHC residue tyrosine 155 is contacted by the same 2 aromatic TCR residues used in 1MWA (tyrosine 31 in the CDR1α loop and tyrosine 50 in the CDR2α loop).

The twist angle of the 2C TCR in this complex is 45.8˚ which is right in the middle of the observed range for class I complexes, but much larger than for the 21.1˚ angle observed for 1MWA. The tilt angle of the TCR is 7.3˚ with a direction angle of 48.2˚, both of which are very similar to that seen for 1MWA (tilt of 8.6˚ with an angle of 49.6˚). The 2C TCR is shifted by 3.0 Å, which is in the middle of the observed range for class I complexes, at an angle of -148.3˚ which is similar to the shift angle for 1OGA. The perpendicular height of the TCR is 30.7 Å compared to 30.6 Å for 1MWA.

The total buried surface area between the TCR and the pMHC is mid-range at 1714 $Å^2$, with TCRα burying 392 $Å^2$ and TCRβ burying 281 $Å^2$. The peptide contributes 32% of the BSA of the pMHC and the MHC contributes 68% which are normal values for class I complexes.

In summary, the orientation and buried surface areas of the TCRs in these new class I complexes are generally typical of those observed in all the other class I complexes.

3.5 Conclusions

TCRs can bind to MHC with a whole range of tilts, twists and shifts and yet the average contact point is always focused on the centre of the bound peptide, and the TCRs consistently bind to a small set of conserved residues on the MHC helices. These residues always include a subset from the residues 65, 69, 72 on the α1 helix and residues 150, 155 and 158 on the α2 helix. In most cases, all 6 residues are contacted and in all cases, the TCR contacts at least one of these residues from each MHC helix. However, the roles of all of these residues in TCR binding to the MHC and/or TCR activation is not yet clear. Alanine scanning experiments of each of the residues in turn, then in combination with other residues may help to elucidate this. However, it is possible that binding to some of these amino acids is more important during positive selection in the thymus than in T cell binding and activation in the periphery.

The most striking discovery has been the use of tyrosine residues in the TCR CDR loops to contact residue 155 on the MHC. Experiments to establish the effects of mutating these tyrosines on TCR binding and activation would confirm their importance.

Chapter 4

Toll-like Receptors

4 Toll-like Receptors

4.1 Summary

The structure of TLR3 was solved simultaneously by two groups in 2005 (Bell, Botos et al., 2005; Choe, Kelker et al., 2005). I aimed to clone and express as many other TLRs as possible using either a mammalian expression system or a baculovirus expression system, with the goal of solving at least one more TLR structure by X-ray crystallography. However, despite testing over 50 different constructs from 12 different TLRs or associated proteins, I did not get sufficient protein expression to set up crystallization trials. This is perhaps not surprising as no other TLR structures have been published since 2005 despite worldwide effort. However, in June 2007 the X-ray structure of human MD-2 alone and in complex with Lipid IVa was published (Ohto, Fukase et al., 2007) where MD-2 had been expressed in methyltropic yeast *Picia pastoris*.

4.2 Materials and Methods

Protein sequences for all TLRs and associated proteins can be found in Appendix B.

All constructs used are listed in Table 2, and referred to by their construct number in the text.

4.2.1 Bioinformatics

RONN is a freely available tool for the prediction of protein disorder (Yang, Thomson et al., 2005), developed by Rebecca Hamer, Ron Yang and Robert Esnouf (www.strubi.ox.ac.uk/RONN).

OPAL is a freely available tool for carrying out multiple sequence analyses such as prediction of signal peptides, glycosylation sites, transmembrane helices and domain matches. It was written by Rebecca Hamer, and can be accessed at www.oppf.ox.ac.uk/pn/opal/OPAL.php. The particular results quoted are from SignalP (Bendtsen, Nielsen et al., 2004), TMHMM, NetNGlyc and NetOGlyc (Julenius, Molgaard et al., 2005).

Swissprot (O'Donovan, Martin et al., 2002) is available at http://ca.expasy.org/sprot/.

4.2.2 Cloning

Clones were ordered from RZPD (Berlin, Germany, www.rzpd.de) (Table 1) and were delivered as stabs in agar. Stabs were plated out onto LB-agar plates (2.5% LB, 1.5% agar) containing the appropriate antibiotic (ampicillin 100 µg/ml, kanamycin 50 µg/ml, chloramphenicol 27 µg/ml) and incubated overnight at 37˚C. Single colonies were picked and grown at 37˚C overnight in 2.5 ml 2.5% LB media containing the appropriate antibiotic at the concentrations above. DNA was extracted from the bacterial cells using a QIAprep Miniprep kit (Qiagen).

DNA for human RP105 and human MD-1 was also kindly provided by Chris Karp from the Cincinnati Children's Hospital Medical Centre.

Primers were designed as shown in Table 2 and ordered from MWG (www.mwg-biotech.com). Each primer was dissolved in sterile Tris pH8 to give a 1 mM stock. Constructs were amplified using pyrobest DNA polymerase (Takara Bio Inc) using the recommended general reaction mixture. DNA was then purified using a QIAquick gel extraction kit (Qiagen).

Constructs were digested with the appropriate restriction enzymes (Table 2), gel purified using a QIAquick gel extraction kit (Qiagen) then ligated into the appropriate vector using a QuickLigation™ kit (New England Biolabs). The vector pHLsec (Aricescu, Lu et al., 2006) (Figure 1) was used for expression in mammalian 293T cells and pBacPAK9 (Clontech Laboratories) (Figure 2) was used for expression in SF9 insect cells. Certain

constructs were cloned using the In-Fusion™ system (Clontech Laboratories) if restriction sites were present in the construct sequence (Table 2).

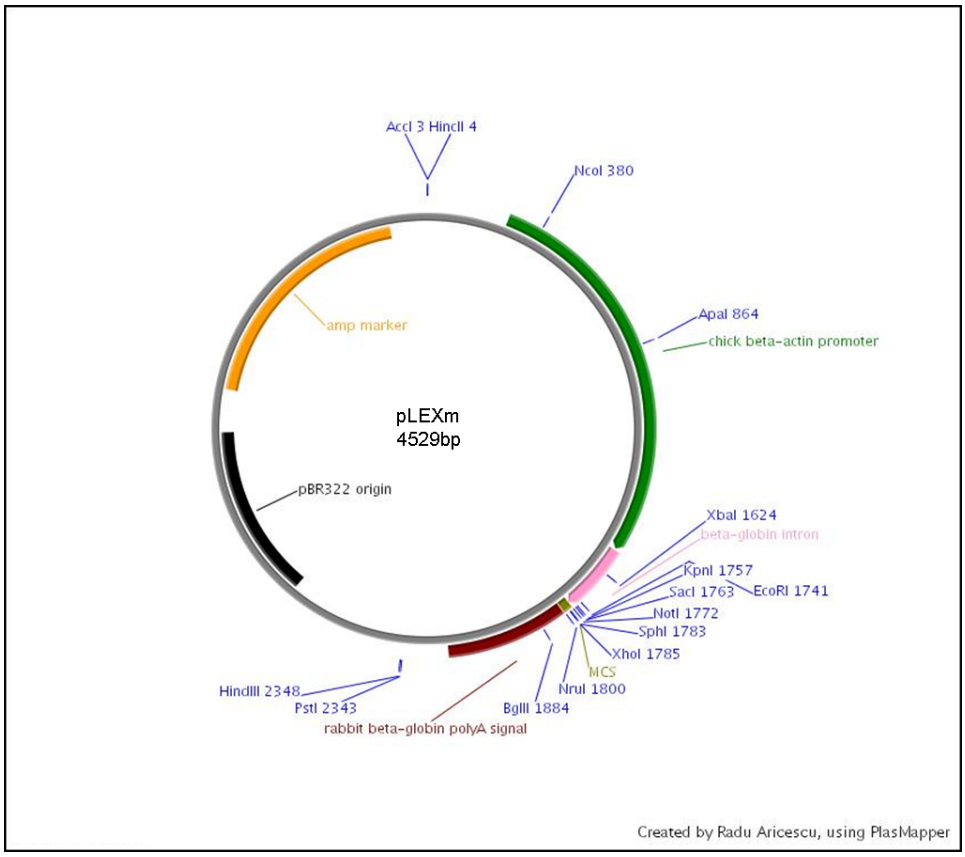


Figure 1a. The pLEXm vector. The pHLsec vector is based on the pLEXm backbone but with a reduced multiple cloning site with a Kozak sequence, a secretion signal sequence and a C-terminal K_His6 tag included (see Figure 1b).

```
EcoRI HindIIIKozak M  G  I  L  P  S  P  G  M  P  A  L  L  S
GAATTCAAGCTTGCCACCATGGGGATCCTTCCCAGCCCTGGGATGCCTGCGCTGCTCTCC

 L  V  S  L  L  S  V  L  L  M  G  C  V  A  E  T  G
CTCGTGAGCCTTCTCTCCGTGCTGCTGATGGGTTGCGTAGCTGAAACCGGT...insert...

 G  T  K  H  H  H  H  H  H  *  *    XhoI
GGTACCAAGCACCACCATCACCATCACTAATGATCACTCGAG
```

Figure 1b. The multiple cloning site from the pHLsec vector. The secretion signal sequence is shown in blue, the AgeI restriction site in pink, the KpnI restriction site in purple and the His tag in orange.

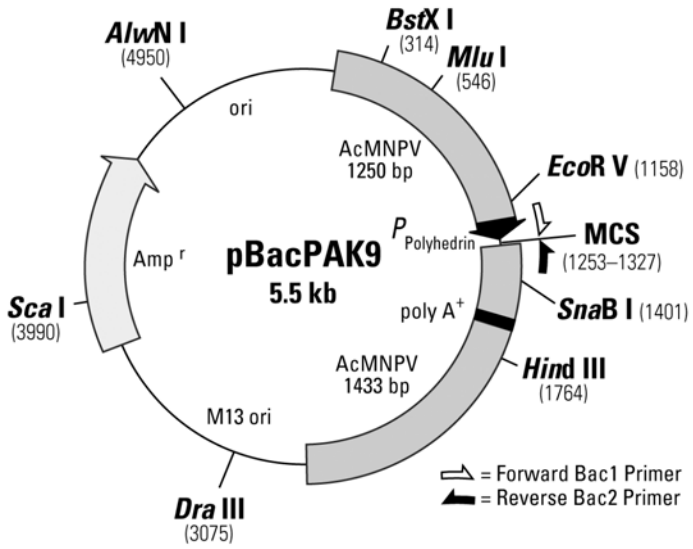


Figure 2a. The pBacPAK9 vector for use in the Baculovirus expression system (Clontech Laboratories).

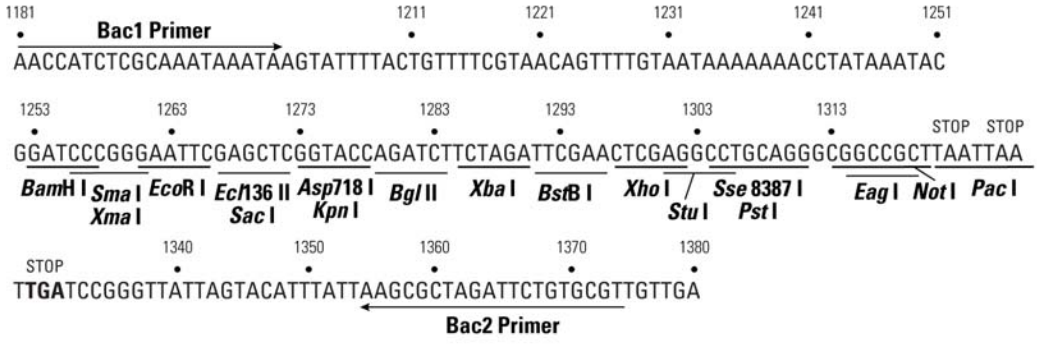


Figure 2b. The pBacPAK9 multiple cloning site.

Competent bacterial cells (Library Efficiency DH5alpha 18263-012, Invitrogen) were transformed using the heat shock method. 44 µl cells were mixed with 3 µl plasmid and left on ice for 30 minutes. They were then heat shocked at 42˚C for 45 seconds and left on ice for a further 2 minutes. 250 µl SOC media (Invitrogen) was added and the cells incubated (shaking) at 37˚C for 1 hour before being plated out onto LB-agar containing 100 µg/ml ampicillin and incubated at 37˚C overnight. Single colonies were picked and incubated in 2.5 ml LB media containing 10 µg/ml ampicillin at 37˚C overnight.

DNA was extracted using a QIAprep Miniprep kit (Qiagen) and a sample digested and run on a gel to check for the presence of the insert. Potentially correct inserts were sequenced verified either at the Sir William Dunn School of Pathology, Oxford or by Agowa, Germany (www.agowa.de). Glycerol stocks of bacteria containing plasmids with the correct insert were made by adding 1 ml bacterial culture to 150 µl sterile glycerol and storing at -80˚C.

To obtain sufficient DNA for transfections, bacteria from glycerol stocks were streaked onto agar plates containing the appropriate antibiotic and grown at 37˚C overnight. Single colonies were then grown for 8 hours in 2.5 ml LB containing the appropriate antibiotic. These cultures were transferred into 500 ml LB-antibiotic and grown at 37˚C overnight. DNA was extracted using a Megaprep kit (Qiagen).

Table 1. Clones ordered from RZPD

TLR	Genbank accession #	MGC #	IMAGE #	Swissprot #	RZPD Clone ID	Antibiotic resistance
TLR1	BC089403	104956	3063206	Q15399	IRAKp961A22199Q	ampicillin
TLR2	BC033756	45301	5213439	O60603	IRATp970B1254D	ampicillin
TLR3	BC096333	117018	40008349	O15455	IRAMp995O012Q	kanamycin
TLR4	BC025294	none	4868078	O00206	IRALp962H0942Q	chloramphenicol
TLR7	BC033651	45096	5582912	Q9NYK1	IRATp970B1256D	ampicillin
TLR8	BC101074	119598	40008787	Q9NR97	IRAMp995I215Q	kanamycin
TLR9	BC032713	45083	5495717	Q9NR96	IRATp970H1255D	ampicillin
TLR10	BC089406	104967	3066588	Q9BXR5	IRAKp961B02199Q	ampicillin
RP105	NM_005582	126233	40034168	Q99467	EX-S0166-B02	ampicillin
MD-1	BC038846	47723	6015123	O95711	IRATp970B0958D	ampicillin
MD-2	BC020690	22424	4767246	Q9Y6Y9	IMAGp958A071629Q2	chloramphenicol
CD14	BC010507	17221	4149008	P08571	IRATp970H0413D	ampicillin

Table 2. Constructs and primers

Template	Construct	Construct number	Fwd primer	Rev primer	Molecular weight (Da)
TLR1	E23-C577	1	CGCACCGGTGAAGAAAGTGAATTTTTAGTTGATAGGTCAAAAAACGG	CGCGGTACCGCAGGATAATTCAGACATGTGAAAGTCC	63085
TLR2	K19-C585	2	CGCACCGGTAAGGAAGAATCCTCCAATCAGGCTTCTCT	CGCGGTACCACATTCCGACACCGAGAGGCGG	64049
TLR2	C30-C585	3	CGCACCGGTTGTGACCGCAATGGTATCTGCAAGG	CGCGGTACCACATTCCGACACCGAGAGGCGG	62888
TLR2	M1-C585	4	cgcAAGCTTGCCACCATGCCACATACTTTGTGGATGGTGTGG	CGCGGTACCACATTCCGACACCGAGAGGCGG	66114
TLR3	K27-A700	5	CGCACCGGTAAGTGCACTGTTAGCCATGAAGTTGCTGA	CTCGAGTGATCATTAGTGATGGTGATGGTGGTGCTTGGCACTGTCTTTGCAAGATGATGTATCAA	76594
TLR3	M1-A700	6	CGCGAATTCGCCACCATGAGACAGACTTTGCCTTGTATCTACTTTTGG	CTCGAGTGATCATTAGTGATGGTGATGGTGGTGCTTGGCACTGTCTTTGCAAGATGATGTATCAA	79470
TLR4	E24-K631	7	CGCACCGGTGAAAGCTGGGAGCCCTGCGTG	CGCGGTACCCTTATTCATCTGACAGGTGATATTCAAACTCAGC	69180
TLR4	M1-K631	8	CGCAAGCTTGCCACCATGATGTCTGCCTCGCGCCTGG	CGCGGTACCCTTATTCATCTGACAGGTGATATTCAAACTCAGC	71586
TLR7	L34-C833	9	CGCACCGGTCTGCCCTGTGATGTCACTCTGGAT	CGCGGTACCACAGGTGTACAGATCCAGGGAGATC	91532
TLR7	C36-C833	10	CGCACCGGTTGTGATGTCACTCTGGATGTTCCAAAGAAC	CGCGGTACCACAGGTGTACAGATCCAGGGAGATC	91322
TLR7	A27-N839	11	CGCACCGGTGCTAGATGGTTTCCTAAAACTCTGCC	CGCGGTACCGTTAGTCAGATCTAACTCACAGGTG	93105
TLR7	A27-V1049	12	CGCACCGGTGCTAGATGGTTTCCTAAAACTCTGCC	CGCGGTACCGACCGTTTCCTTGAACACCTGACTATAG	117838
TLR8	E27-D825	13	CGCACCGGTGAAGAAAATTTTTCTAGAAGCTATCCTTGTG	CGCGGTACCATCTGAAACACAAGTTGTTAGCTCCAGACTCA	91267
TLR8	E27-Y1041	14	CGCACCGGTGAAGAAAATTTTTCTAGAAGCTATCCTTGTG	CGCGGTACCGTATTGCTTAATGGAATCGACATACATATTGTTAT	116907
TLR8	Y34-C822	15	CGCACCGGTTATCCTTGTGATGAGAAAAAGCAAAATGACTCAGTTATT	CGCGGTACCACAAGTTGTTAGCTCCAGACTCACAATACTC	90116
TLR8	C36-C822	16	CGCACCGGTTGTGATGAGAAAAAGCAAAATGACTCAGTTATTGCA	CGCGGTACCACAAGTTGTTAGCTCCAGACTCACAATACTC	89855
TLR8	M1-C822	17	cgcAAGCTTGCCACCATGGAAAACATGTTCCTTCAGTCGTCAATGC	CGCGGTACCACAAGTTGTTAGCTCCAGACTCACAATACTC	93833
TLR9	L26-C818	18	CGCCACCGGCGCTGGGTACCTTGCCTGCCTTC	CGCCTCGAGCTCGAGTGATCATTAGTGATGGTGATGGTGGTGCTTACAGTCCCAGGAGAGGGCCTCA	88266
TLR9	T28-C818	19	CGCCACCGGCGACCTTGCCTGCCTTCCTACCC	CGCGGTACCACAGTCCCAGGAGAGGGCCTCA	88095

TLR9	T28-E1032	20	CGCCACCGGCGACCTTGCCTGCCTTCCTACCC	CGCGGTACCTTCGGCCGTGGGTCCCTGG	113019
TLR9	L26-C809*	21	CGTAGCTGAAACCGGTCTGGGTACCTTGCCTGCCTTC	TGATGGTGGTGCTTGGTACCGCAGAGGCGCAGGTCCT GTG	87233
TLR9	T28-C809*	22	CGTAGCTGAAACCGGTACCTTGCCTGCCTTCCTACCC	TGATGGTGGTGCTTGGTACCGCAGAGGCGCAGGTCCT GTG	87062
TLR9	C35-C809*	23	CGTAGCTGAAACCGGTTGTGAGCTCCAGCCCCACGG	TGATGGTGGTGCTTGGTACCGCAGAGGCGCAGGTCCT GTG	86322
TLR9	L33-C809*	24	CGTAGCTGAAACCGGTCTACCCTGTGAGCTCCAGCCC	TGATGGTGGTGCTTGGTACCGCAGAGGCGCAGGTCCT GTG	86533
TLR9	M1-C809*	25	GTCTCAGGCCGAATTCGCCACCATGGGTTTCTGCCGC AGCGCC	TGATGGTGGTGCTTGGTACCGCAGAGGCGCAGGTCCT GTG	89903
TLR10	D20-C574	26	CGCACCGGTGATGCTCCAGAGCTGCCAGAAGAAA	CGCGGTACCGCAAGATAATTCGTGGAGATGAACGTCT TTTA	64352
TLR10	C34-C574	27	CGCACCGGTTGCTCCAACATGTCTCTAAGAAAGGTTC C	CGCGGTACCGCAAGATAATTCGTGGAGATGAACGTCT TTTA	62726
TLR10	M1-C574	28	cgcAAGCTTGCCACCATGAGACTCATCAGAAACATTT ACATATTTTGTAGTATTG	CGCGGTACCGCAAGATAATTCGTGGAGATGAACGTCT TTTA	66565
RP105	W24-C625	29	CGCACCGGTTGGGATCAGATGTGCATTGAGAAAGAAG C	CGCGGTACCACAGGAAAGCTTGACATCAGATAGCTTA ACT	67375
RP105	M1-C625 (pBacPAK)	30	CGCGGATCCGCCACCATGGCGTTTGACGTCAGCTGCT TC	CGCGGTACCACAGGAAAGCTTGACATCAGATAGCTTA ACT	69915
RP105	M1-C625*	31	GTCTCAGGCCGAATTCGCCACCATGGCGTTTGACGTC AGCTGCTTC	TGATGGTGGTGCTTGGTACCACAGGAAAGCTTGACAT CAGATAGCTTAACT	69915
MD1	G21-S162	32	CGCACCGGTGGCGGCGGTGGGAAAGCC	CGCGGTACCGGAGCACATGATAGTAGCATTGGCAC	15703
MD1	M1-S162	33	cgcGAATTCGCCACCATGAAGGGTTTCACAGCCACTC TCTTC	CGCGGTACCGGAGCACATGATAGTAGCATTGGCAC	17906
MD2	Q19-N160	34	CGCACCGGTCAGAAGCAGTATTGGGTCTGCAACTCAT	CGCGGTACCATTTGAATTAGGTTGGTGTAGGATGACA AACTC	16366
MD2	M1-N160	35	CGCGAATTCGCCACCATGTTACCATTTCTGTTTTTTT CCACCCTGTTTTC	CGCGGTACCATTTGAATTAGGTTGGTGTAGGATGACA AACTC	18446
CD14	P22-L332	36	cgcACCGGTCCAGAACCTTGTGAGCTGGACG	cgcGGTACCCAGGAAGGGATTCCCGTCCAGTG	33666
CD14	M1-L332	37	cgcGAATTCGCCACCATGGAGCGCGCGTCCTGCTTG	cgcGGTACCCAGGAAGGGATTCCCGTCCAGTG	35928

* Primer for use with the In-Fusion cloning system.
Restriction sites and Kozak sequences are underlined: AgeI: ACCGGT, KpnI: GGTACC, SgrAI: CACCGGCG, XhoI: CTCGAG, HindIII: AAGCTT, BamHI: GGATCC, EcoRI: GAATTC, Kozak sequence: GCCACCATG.

4.2.3 Tissue Culture

All transfections were carried out by the department tissue culture technician, Wei-Xian Lu.

4.2.3.1 Transfection of 293T cells

The transfection protocol for 293T cells is described in detail in (Aricescu, Lu et al., 2006). Briefly, cells were maintained in Dulbecco's Modified Eagles Medium (DMEM, Sigma) supplemented with L-glutamine, non-essential amino acids (Invitrogen) and 10% foetal calf serum (FCS, Sigma). For small-scale transfections in 6-well plates, when adherent cells reached 90% confluency, 50 μg of plasmid DNA was added to 5 ml serum-free medium and mixed. 75 μl polyethylenimine (PEI) (1 mg/ml) was added and the solution briefly vortexed then incubated at room temperature for 10 minutes to allow DNA-PEI complex formation. Media from the 6-well plates to be transfected was replaced with media containing 2% serum, then the DNA-PEI complex was added. The dish was briefly rotated to mix and the cells incubated at 37˚C for 4 days. For large-scale transfections, the same protocol was followed with quantities scaled up accordingly but keeping a DNA:PEI ratio of 1:1.5. Cells were incubated in roller bottles for 1 week at 37˚C.

4.2.3.2 Tissue culture of CHO cells

Stable Chinese hamster ovary (CHO) cell lines expressing mouse TLR4 and mouse RP105 and a stable 293T cell line expressing human MD-2 were provided by Chris Karp (Cincinnati Children's Hospital Medical Centre).

CHO cells were maintained in DMEM (Sigma) supplemented with L-glutamine, non-essential amino acids (Invitrogen) and 10% foetal calf serum (FCS, Sigma) and split twice a week. Cells from 160 ml flasks were transferred into 600 ml flasks and grown for 3-4 days, then transferred into roller bottles in 250 ml DMEM. When the cells were confluent the media was removed and replaced with 250 ml UltraCHO media (Biowhittaker) and incubated at 37˚C for 6 days.

4.2.3.3 Transfection of Sf-9 cells

Preparation of P1 stock: $9x10^5$ cells in 2 ml Sf-900 II SFM (Invitrogen) containing the antibiotics penicillin and streptomycin (Sigma) at 0.5 X final concentration were plated into a 6-well dish. 1 µg baculovirus DNA was diluted in 100 µl Sf-900 II SFM without antibiotics. 8 µl pre-mixed Cellfectin® was diluted in 100 µl Sf-900 II SFM without antibiotics then combined with the diluted DNA, mixed gently and incubated for 20 minutes at room temperature. The growth medium was then removed from the cells and the cells washed with Sf-900 II SFM without antibiotics. 0.8 ml Sf-900 II SFM was added to the DNA:cellfectin complex, mixed gently then added to the cells and incubated at 27˚C for 5 hours. The transfection mixture was then removed and replaced with 2 ml

Sf-900 II SFM (Invitrogen) containing antibiotic and incubated at 27˚C for 5 days before harvesting the virus.

The virus was then amplified by preparing a 20 ml culture of SF9 cells grown in Sf-900 II SFM (Invitrogen) to a density of $8x10^5$ cell/ml. 0.2 ml P1 stock was added to this culture and incubated at 27˚C for 7 days. For scale-up, 2 litres of SF9 cells were seeded at a density of $1.5X10^6$ cells/ml. 200ml virus stock was added to this culture and incubated at 27˚C for 72 hours.

4.2.4 Western Blotting

For all cell types, samples of media and cells were taken and tested for protein expression using Western blots.

For His-tagged proteins, Western blots were probed using a pentaHis monoclonal primary antibody (1:1000 dilution, Qiagen) and goat anti-mouse IgG peroxidase-conjugated secondary antibody (1:2000 dilution, Sigma).

For Fc-tagged proteins, Western blots were probed with anti-human IgG Fc-specific peroxidase-conjugate antibody (1:1000 dilution, Sigma).

The signal was visualized by chemiluminescence using the ECL kit (GE Healthcare)

4.2.5 Protein Purification

Media was spun down to pellet any dead cells and Tris pH8 was added to give a final concentration of 10 mM.

All constructs used in 293T cells were His-tagged so media was incubated with Talon® Superflow™ metal affinity resin (Clontech Laboratories) at 16˚C for 30 minutes. Protein was eluted from the beads using 10 mM Tris pH7, 150 mM NaCl, 500 mM imidazole.

All constructs used in CHO cells were Fc-tagged so media was incubated with nProteinA Sepharose™ 4 Fast Flow beads (GE Healthcare) at 16˚C for 2 hours. Protein was cleaved off the beads by incubation with 3C protease at 4˚C overnight, leaving the Fc tag attached to the beads. Liquid was drained from the beads and the 3C protease was removed by incubating with Glutathione Sepharose™ 4 Fast Flow beads (GE Healthcare) at 4˚C for 1 hour. Liquid was drained and concentrated using a 30 000 MW concentrator (Vivascience) and spin filtered using a 0.22 μm filter unit (Millipore) before being loaded onto a size exclusion column. In a variation of this protocol for one small-scale test, Protein G Sepharose™ 4 Fast Flow beads (GE Healthcare) were used. Media was incubated with the beads at room temperature for 1 hour before protein was eluted by boiling for 10 minutes.

4.2.6 Protein Gels

Protein gels containing un-tagged protein samples were stained with SimplyBlue™ SafeStain (Invitrogen).

A 4 - 12% Bis-Tris Gradient gel (Invitrogen) was used where a large range of molecular weights needed to be detected.

4.3 Results and Discussion

4.3.1 Potential Targets for Structural Studies

Human TLRs 1-10 were the primary targets for these structural studies. However, literature searches also revealed the existence of a TLR4 homologue, RP105 (Divanovic, Trompette et al., 2005). This has been identified as a negative regulator of TLR4, with an extracellular leucine-rich repeat domain but no cytosolic TIR domain. The surface expression and signalling of TLR4 is dependent on a protein called MD-2 (Shimazu, Akashi et al., 1999; Visintin, Iliev et al., 2006) and similarly, surface expression of RP105 is dependent on an MD-1. A third protein, CD14, was also found to be important for TLR4 signalling. CD14 is a dimer, each monomer having 11 leucine-rich repeats but no intracellular signalling domain. It binds to LPS and presents it to the TLR4/MD-2 complex; the structure of mouse CD14 has already been solved (PDB identifier 1WWI) (Kim, Lee et al., 2005). Constructs of these proteins were designed in the later stages of the project. Finally, in the case of TLR4 and RP105, mouse clones were also used to see if these gave better expression and/or secretion than their human counterparts.

4.3.2 Bioinformatics Analyses

Before any experimental work was begun, bioinformatics analyses were run on the protein sequences of all the TLRs and associated molecules to ensure they were suitable for structural studies and to aid construct design.

RONN was used to predict protein disorder (Figure 3) (see chapter 5 for details of the development of RONN). No significant regions of disorder were predicted to be present, suggesting these proteins would fold into highly ordered structures.

OPAL was then run on the protein sequences to predict the positions of signal sequences, transmembrane helices and glycosylation sites. This enabled comparison to any annotations in Swissprot and any differences were taken into account when designing constructs (Table 3).

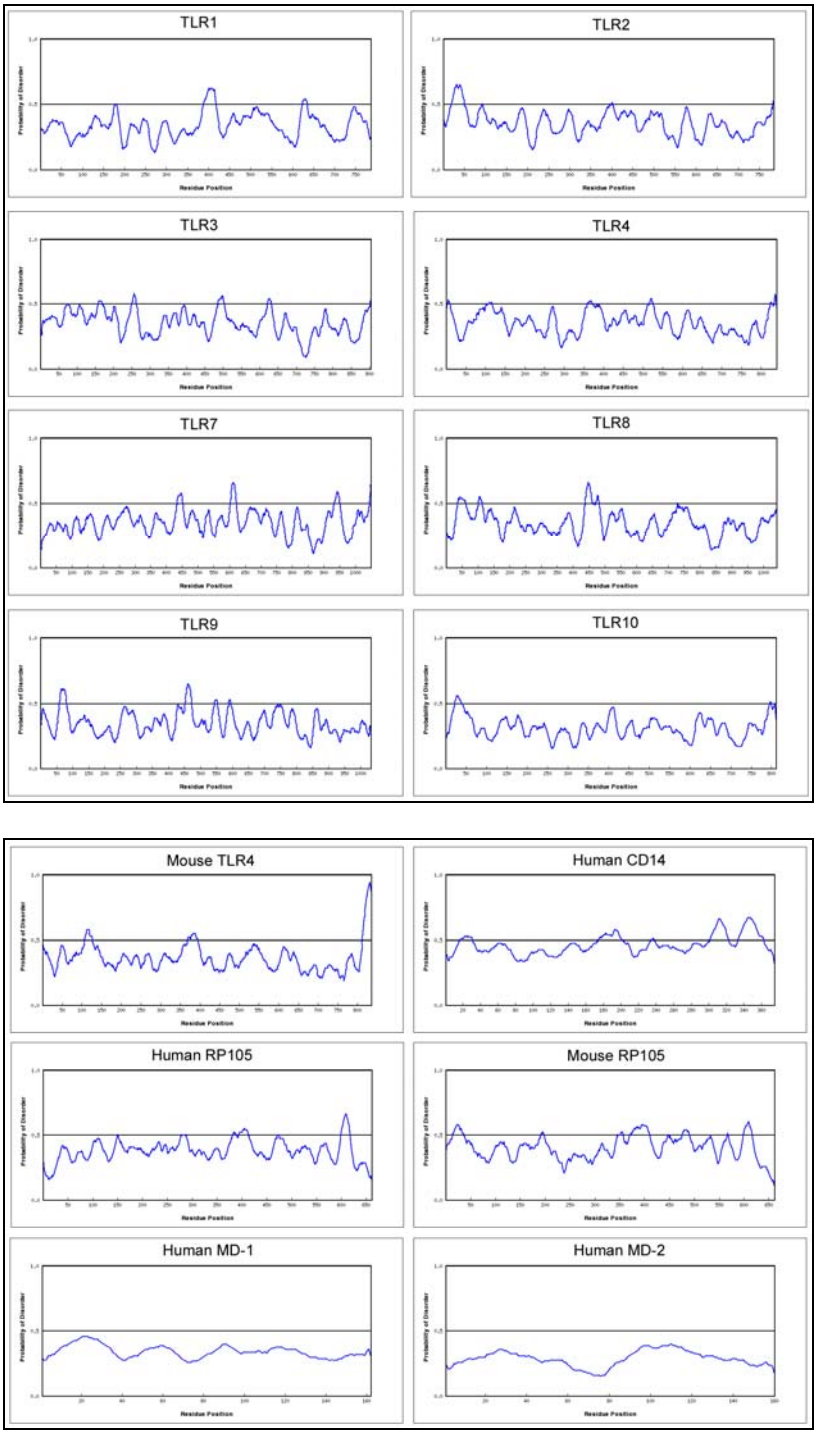


Figure 3. RONN disorder predictions for TLRs and associated molecules. For each graph, the y axis shows the probability of disorder and the x axis shows the residue number. A residue is considered to be disordered if the probability of disorder is above 0.5. Proteins are human unless stated otherwise.

Table 3. Bioinformatics results from OPAL and Swissprot.

TLR	SignalP predicted signal sequence	Swissprot predicted signal sequence	TMHMM predicted transmembrane region	Swissprot predicted transmembrane region	NetNGlyc predicted # N glycosylation sites	NetOGlyc predicted # O glycosylation sites
TLR1	1-22	1-24	582-604	581-601	6	0
TLR2	1-18	1-18	588-610	58-609	4	0
TLR3	1-22	1-23	703-725	705-725	13	0
TLR4	1-23	1-23	634-656	632-652	10	0
TLR7	1-26	1-26	843-865	840-860	14	1
TLR8	1-26	1-26	826-848	828-848	18	0
TLR9	1-27	1-25	819-839	819-839	12	0
TLR10	1-19	1-19	577-599	577-597	8	0
RP105	1-18	1-23	627-649	627-650	9	0
MD-1	1-20	1-20			1	0
MD-2	1-18	1-18			2	0
CD14	1-19	1-19			4	0

4.3.3 Expression of Human TLR7, TLR8 and TLR9

All human TLR sequences except TLR5 and TLR6 were available to buy from RZPD (Table 1). TLR7, 8 and 9 were selected first as these have known ligands which were readily available, meaning there was the potential for co-crystallization trials.

Constructs were initially designed which did not include the signal sequences, as there is a signal sequence present in the pHLsec plasmid, and were either full length or stopped at the residue before the predicted transmembrane region. The signal peptides predicted by SignalP and given by Swissprot had different end points for TLR9, so 2 constructs with different start points were designed. All constructs were cloned into the pHLsec plasmid and small-scale expression tests were carried out in mammalian 293T cells.

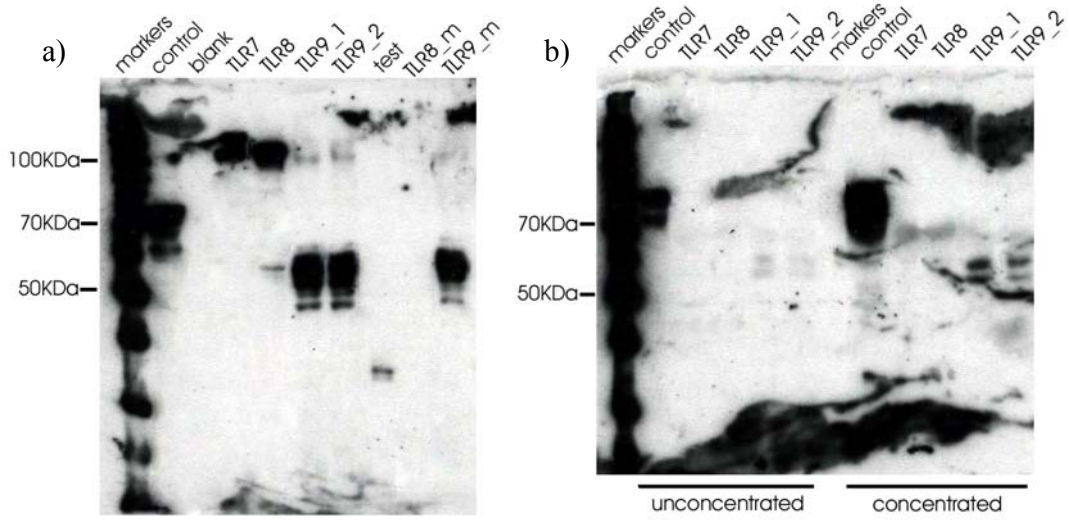


Figure 4. Expression of human TLR7, TLR8 and TLR9 in 293T cells a) expression in the cells b) secretion into the media. The control protein was Nipah virus attachment glycoprotein (74KDa) which is expressed at approximately 1mg/litre. The test lane contains an unrelated protein run for a colleague. TLR7 corresponds to construct 11, TLR8 to construct 13, TLR9_1 to construct 18, TLR9_2 to construct 19, TLR8_m to construct 14 and TLR9_m to construct 20. All samples were run under reducing conditions.

The Western blot is shown in Figure 4 and shows that TLR7 and 8 were expressed in the cells but not secreted into the media. TLR9 should have a molecular weight of 88 KDa so is clearly incorrect. Sequencing revealed that cloning had failed; the pHLsec plasmid used for cloning contained a protein called netrin which should have been replaced by TLR9 but was not in this case. TLR9 was therefore re-cloned into pHLsec and sequenced verified ready for re-testing.

Due to lack of secretion into the media, new constructs for TLR7, TLR8 and TLR9 were designed. These were based on alignments with TLR3 and comparison with the constructs used by the two groups that solved the structure of TLR3. It was observed that the TLR3 construct from the Davies group (Bell, Botos et al., 2005) began one residue before the end of the signal sequence (alanine 22), but the construct used by the Wilson group (Choe, Kelker et al., 2005) started at the residue before the first paired cysteine (lysine 27). Based on this, new constructs starting at both 2 residues before the equivalent cysteine, and at this cysteine, were designed for TLR7, 8 and 9 (extracellular domains only) (Figure 5). Constructs were not designed starting at 1 residue before the first cysteine as this residue is a proline in TLR7, TLR8 and TLR9, which is a poor starting residue.

```
TLR7    MVFPMWTLKRQILILFNIILISKLLGARWFPKTLPCDVTLDVPKNHVIVDCTDKHLTEIP 60
TLR8    MENMFLQSSMLTCIFLLISGSCELCAEENFSRSYPCDE--KKQNDSVIAECSNRRLQEVP 58
TLR9    -MGFCRSALHPLSLLVQAIMLAMTLALGTLPAFLPCELQPHGLVNCNWLFLKSVPHFSMA 59
TLR3    ----MRQTLPCIYFWGGLLPFGMLCASSTT----------KCTVSHEVADCSHLKLTQVP 46
TLR3    ----MRQTLPCIYFWGGLLPFGMLCASSTT----------KCTVSHEVADCSHLKLTQVP 46
```

Figure 5. Alignment of TLR7, 8 and 9 with TLR3 and comparison to the TLR3 constructs used to solve the structure. Signal sequences are highlighted in yellow, the TLR3 construct used by the Wilson group is highlighted in magenta with the paired cysteines in green, and the putatively equivalent cysteines in TLR7, 8 and 9 are highlighted in green. The TLR3 construct used by the Davies group is highlighted in grey.

4.3.4 Expression of Human TLR1, TLR2 and TLR10

In addition, clones for TLR1, TLR2 and TLR10 were ordered from RZPD and extracellular domain constructs were designed which did not include the signal sequence. These were cloned into pHLsec and small-scale expression tests were carried out. TLR1 and TLR2 were transfected together as well as separately as it is known that they associate on the cell surface (Takeda, Kaisho et al., 2003). The Western blot results are shown in Figure 6.

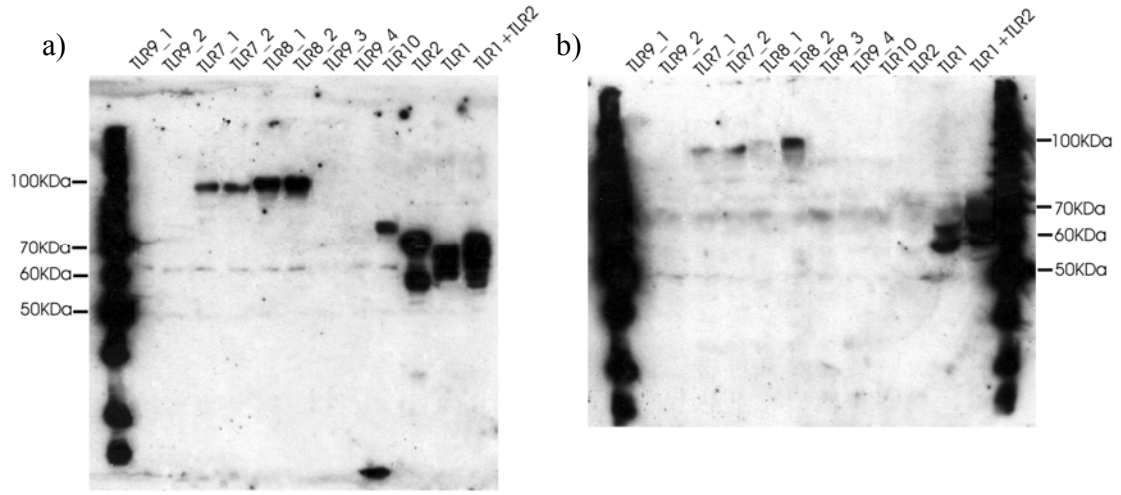


Figure 6. Expression of human TLR1, 2, 7, 8, 9 and 10 in 293T cells a) expression in the cells b) secretion into the media. TLR1 corresponds to construct 1, TLR2 to construct 2, TLR7_1 to construct 9, TLR7_2 to construct 10, TLR8_1 to construct 15, TLR8_2 to construct 16, TLR9_1 to construct 18, TLR9_2 to construct 19, TLR9_3 to construct 24, TLR9_4 to construct 23 and TLR10 to construct 26. All samples were run under reducing conditions.

All constructs of TLR1, 2, 7, 8 and 10 were produced in the cells whereas no TLR9 constructs were expressed at all. It was later realised that the TLR9 constructs were out of frame due to joining of SgrAI and AgeI restriction sites during cloning (although SgrAI gives compatible ends to AgeI, SgrAI has an 8 base restriction site whereas AgeI

has a 6 base restriction site). TLR1 was expressed at a reasonable level in the media but TLR2 and TLR10 were not secreted. All TLR7 and 8 constructs were expressed with low yield in the media, but constructs starting at the first cysteine were secreted with slightly higher yield than those starting 2 residues before the first cysteine.

Small-scale tests for TLR1, TLR7 and TLR8 were then done to see if kifunensine or heparin had any effect on expression and secretion. Kifunensine restricts glycosylation branching and heparin improved expression of a colleague's protein for unknown reasons. However, although the control protein was detected in the media, there was no secretion of any of the TLR proteins (Figure 7), despite the fact that they had been secreted in the previous experiment (Figure 6).

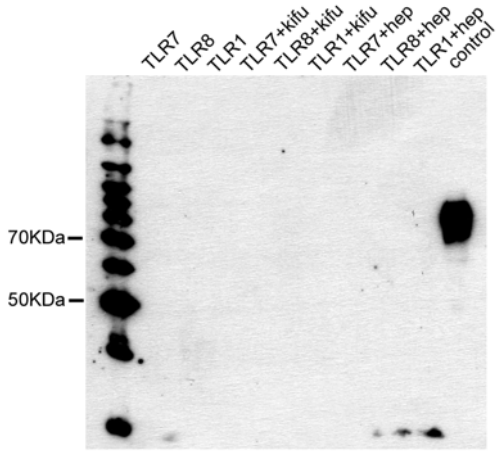


Figure 7. Secretion of human TLR1, 7 and 8 expressed in 293T cells. Kifu=kifunensine, hep=heparin. Control is as for Figure 4. All samples were run under reducing conditions.

Given that TLR1 was secreted into the media in reasonable amounts in Figure 6, a large-scale expression test for TLR1 was carried out. However, although TLR1 was expressed

in the cells it was not secreted into the media (Figure 8). It seems that secretion of TLR1 in Figure 6 was the exception rather than the rule.

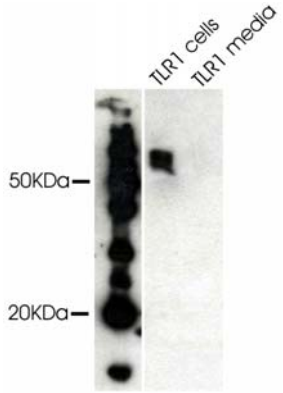


Figure 8. Large-scale expression test in 293T cells of Human TLR1. Samples were run under reducing conditions.

New constructs were then designed for TLR2 and 10 to start at the first paired cysteine, as this had given some secretion for TLR7 and TLR8. TLR1 does not have an equivalent cysteine so no new construct was made. The results of small-scale expression tests were for TLR1, 2, 7, 8 and 10 are shown in Figure 9. All these constructs were expressed in the cells but not secreted into the media.

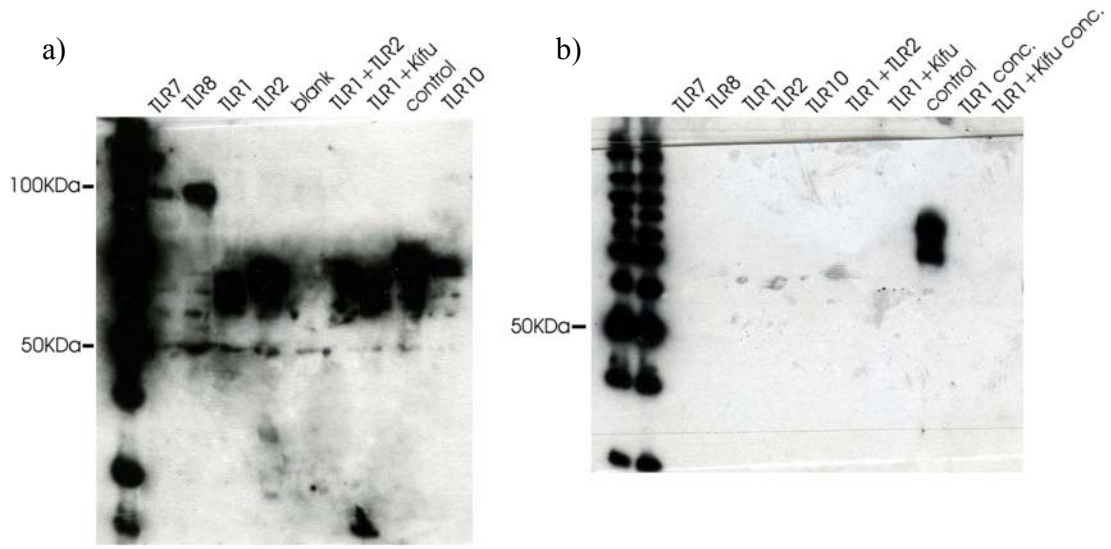


Figure 9. Expression of human TLR1, 2, 7, 8 and 10 in 293T cells a) in the cells b) in the media. Kifu = kifunensine. Control is as for Figure 4. TLR1 corresponds to construct 1, TLR2 to construct 3, TLR7 to construct 9, TLR8 to construct 16 and TLR10 to construct 27. All samples were run under reducing conditions.

4.3.5 Expression of TLR3, TLR4 and TLR Associated Molecules

TLR4, MD-1, MD-2, CD14 and RP105 were ordered from RZPD. TLR4 had not been previously ordered as RZPD stated that it was a 'predicted full length clone' rather than a 'full length cDNA clone'. However, due to the lack of success with the other TLR clones it was thought to be worth trying. As described above, MD-2 is required for surface expression of TLR4, and RP105 is a TLR4 homologue which is dependent on MD-1 for surface expression. Constructs which started after the signal sequence were designed for all of these, and for MD-1, MD-2 and CD14 constructs were also made which included the natural signal sequences of these proteins.

TLR3 was also ordered from RZPD as a control which was expected to be expressed and secreted as the structure has already been solved. This was to check that there was no problem with the expression system being used for TLRs in general. Two constructs were designed for this: the first was identical to that used by the Wilson group (construct 5). The second included the natural signal sequence of TLR3 (construct 6) to test whether the signal sequence in the pHLsec plasmid had any effect on expression and/or secretion.

TLR9 constructs were also redesigned for use with the In-Fusion™ system to circumvent the restriction site problem, with one construct including the natural signal sequence for TLR9.

Results of small-scale expression tests in mammalian 293T cells for all these constructs are shown in Figure 10. In addition to being transfected alone, TLR4 and MD-2, and RP105 and MD-1 were co-transfected to see if this improved expression and/or secretion. The multiple bands seen for some of the proteins are due to different glycosylation states.

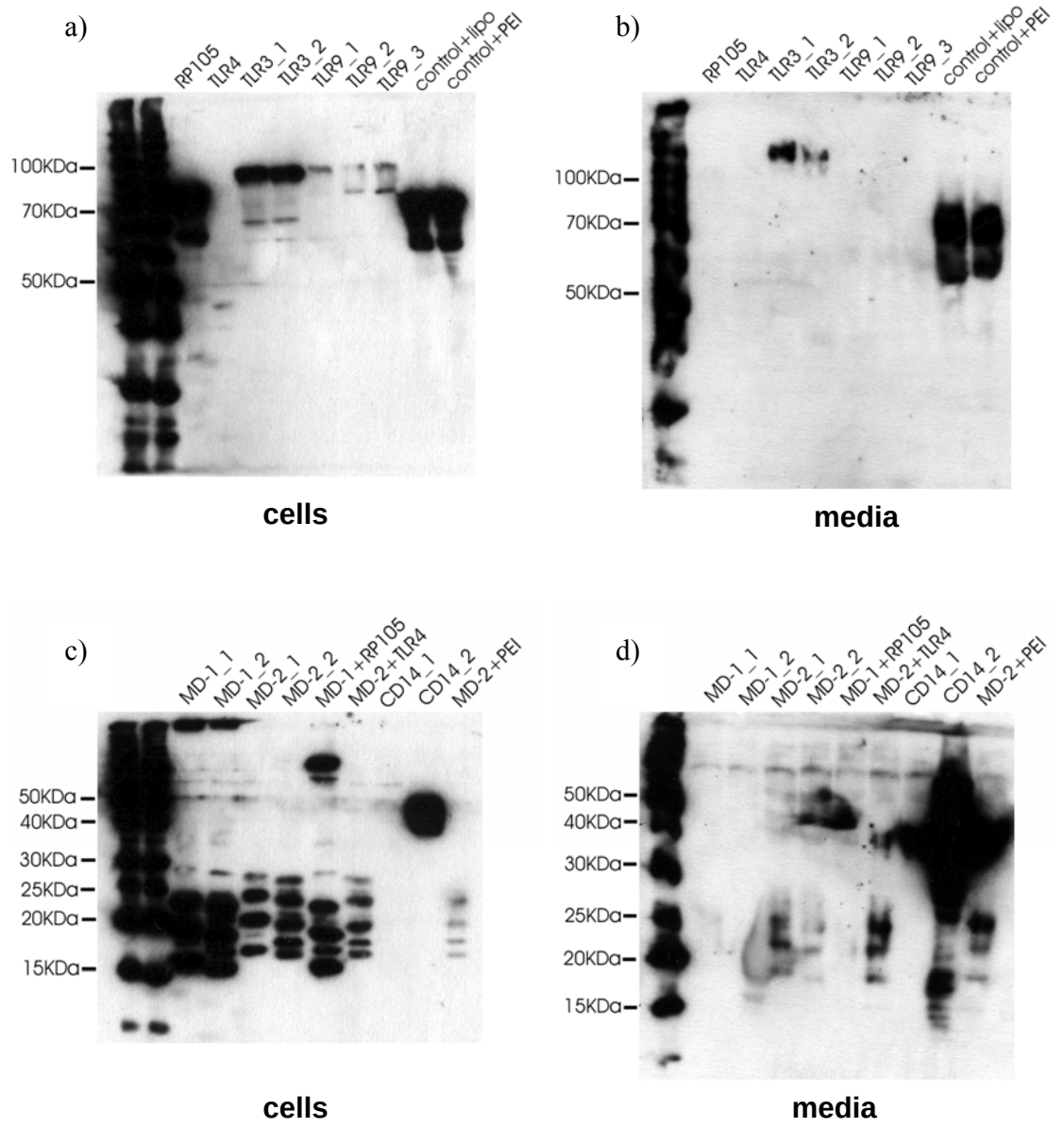


Figure 10. Expression of human TLRs and related proteins in 293T cells. a) expression in cells and b) secretion into the media of the larger proteins run on a 10% SDS gel. c) expression in cells and d) secretion into the media of the smaller proteins run on a 15% SDS gel. The control protein in a) and b) is as in Figure 4. TLR3_1 and TLR3_2 correspond to constructs 5 and 6, TLR9_1, TR9_2 and TLR9_3 to constructs 25, 21 and 22, MD-1_1 and MD-1_2 to constructs 32 and 33, MD-2_1 and MD-2_2 to constructs 34 and 35 and CD14_1 and CD14_2 to constructs 36 and 37. MD-1 construct 33 was used in combination with RP105 and MD-2 construct 35 was used in combination with TLR4 and for test using PEI. PEI=polyethylyeneimine. All other transfections were done using Lipofectamine. All samples were run under reducing conditions.

The upper panels of Figure 10 show the expression (a) and secretion (b) of the larger proteins transfected into 293T cells. Both constructs of TLR3 (one including the natural signal sequence, the other using the signal sequence from the pHLsec plasmid) were expressed well in the cells, as was the control protein and also RP105. The control protein was transfected both with lipofectamine and PEI but this made little difference to expression. TLR9 was expressed in the cells but only in very low amounts. TLR4 was not expressed at all in the cells, but sequencing later showed that the construct ordered from RZPD was missing residues 94-258.

Only TLR3 and the control protein were secreted into the media, and the construct of TLR3 without the natural signal sequence was secreted in a slightly larger amount.

The lower panels show expression (c) and secretion (d) of the smaller proteins transfected into 293T cells. All these proteins were expressed in the cells, with the exception of CD14 without its natural signal sequence. However, this construct was later found to have 2 mutations at bases 95 and 99, which created an AgeI restriction site. AgeI was used to clone the PCR product into pHLsec, so this construct also got cut at this extra site, causing a frame shift and meaning the His tag was out of frame. CD14 with the natural signal sequence was secreted in large amounts. Co-expression of RP105 and MD-1 seemed to decrease expression of both proteins when compared to the expression levels of each protein alone.

MD-1 was not secreted at all, despite good expression levels in the cells, whereas both constructs of MD-2 were secreted in small amounts. The MD-2 construct without the natural signal sequence (construct 34) gave the best secreted yield of protein when both MD-2 constructs were transfected using lipofectamine. However, MD-2 was expressed in the cells in larger amounts when lipofectamine rather than PEI was used in the transfection, but was secreted in larger amounts when PEI was used.

4.3.6 New Human TLR4 DNA Obtained

As the TLR4 clone obtained from RZPD was missing residues 94-258, DNA for TLR4 was obtained from the Miyake group (Japan) (Konno, Wakabayashi et al., 2006) and the same construct was made (construct 7). Small-scale expression tests were carried out, and results are shown in Figure 11. This time, TLR4 was expressed well in the cells, although the addition of MD-2 did not have any impact on expression levels, but neither TLR4 nor MD-2 was secreted into the media.

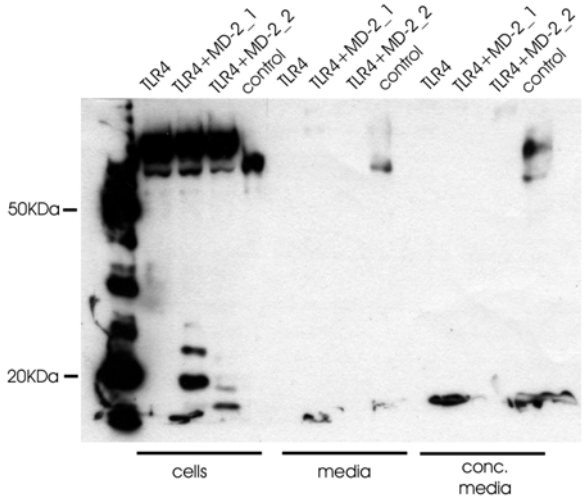


Figure 11. Expression and secretion of human TLR4 from the Miyake group. The control protein is the same as in Figure 4. MD-2_1 and MD-2_2 correspond to constructs 34 and 35 respectively. All samples were run under reducing conditions.

4.3.7 Large-Scale Expression Test of Human MD-2

Given that some secretion of MD-2 construct 34 had been seen in Figure 10, and that construct 34 was expressed better in the cells than construct 35 in Figure 11, a large-scale transfection of MD-2 construct 34 was carried out.

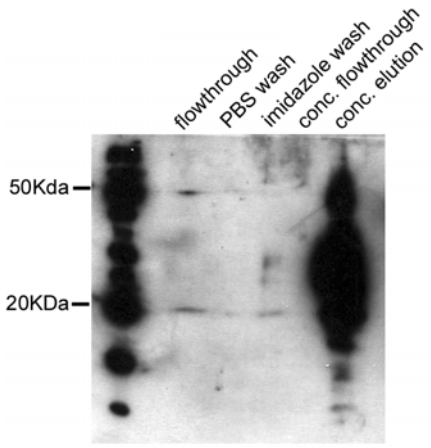


Figure 12. Large-scale expression test of human MD-2 in 293T cells. Flowthrough refers to the media after incubation with Talon® beads. PBS and imidazole were used to wash the beads then protein was eluted with a buffer containing 250mM imidazole. All samples were run under reducing conditions.

The Western blot (Figure 12) shows that MD-2 was secreted into the media with good yield, but no peak was seen when the concentrated elution was run on a size exclusion column. As a result, the experiment was repeated. However, Figure 13 shows that this time MD-2 was produced in the cells but not secreted into the media at all.

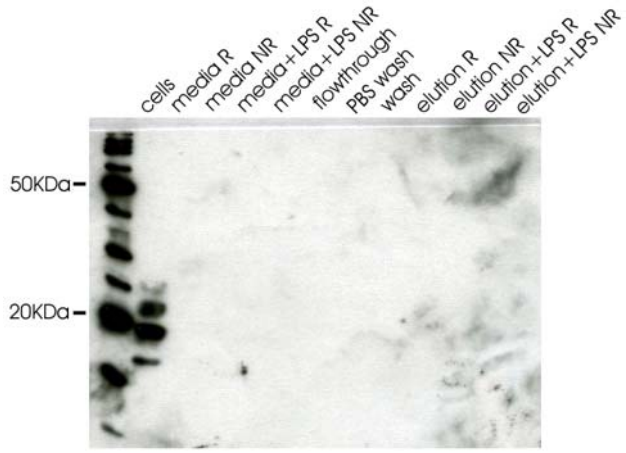


Figure 13. Human MD-2 large-scale expression test repeat in 293T cells. R = reduced, NR = non-reduced. LPS = lipopolysaccharide.

4.3.8 DNA and Stable Cell Lines Obtained for TLR4, RP105, MD-1 and MD-2

DNA for human MD-1, human RP105 and mouse RP105 plus a stable 293T cell line expressing human MD-2 and stable CHO cells lines expressing mouse RP105 and mouse TLR4 were obtained from Chris Karp (Divanovic, Trompette et al., 2005).

The stable cells lines were tested for expression first. Human MD-2 gave no signal on a Western blot either for expression in the cells or in the media. Figure 14 shows the expression of mouse TLR4 and mouse RP105.

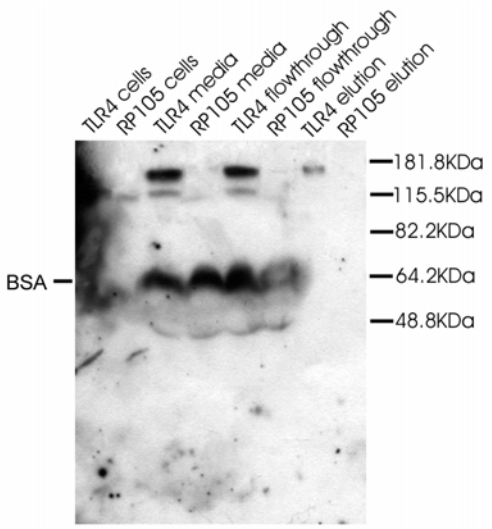


Figure 14. Expression test of mouse TLR4 and mouse RP105 in a CHO stable cell line. Media was incubated with nProteinA beads to bind the Fc-tagged protein. Flowthrough refers to media after incubation with nProteinA beads. Protein was eluted using 0.1M citric acid pH3.0. BSA = bovine serum albumin. Both proteins were Fc-tagged but the standard markers are His-tagged. Only an anti-Fc antibody was used for this western so the markers are not visible, but approximate positions were obtained from the prestained markers on the membrane. All samples were run under reducing conditions.

Mouse RP105 was very poorly expressed in cells and not visible in the media, but mouse TLR4 was expressed in the cells and secreted into the media. However, there is a large proportion of protein left in the flowthrough from the nProteinA beads and little protein visible in the elution. This could be due to poor binding of TLR4 to the nProteinA beads, or to poor elution of the protein from the beads and is investigated below.

4.3.9 Focus on Expression of Mouse TLR4 from the Stable CHO Cell Line

Both nProteinA beads and ProteinG beads bind to IgG Fc-tagged proteins, so binding of mouse TLR4 to both types of beads was compared (Figure 15).

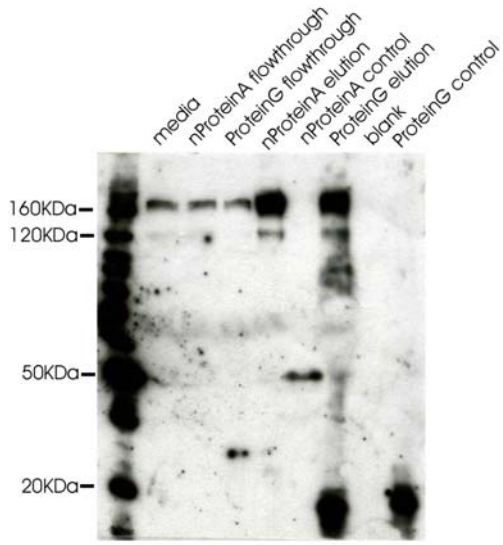


Figure 15. Comparison of mouse TLR4 binding to nProteinA and ProteinG beads. Elution was by incubating the beads at 100˚C for 10 minutes. nProteinA and ProteinG controls are samples of fresh beads which had not been incubated with media. All samples were run in under reducing conditions.

It can be seen that both nProteinA and ProteinG beads bind to mouse TLR4, although despite incubation with the beads for 1 hour at room temperature, protein was left in the flowthrough in both cases. For this experiment, protein was eluted by incubating the beads at 100˚C which gave much better results than using citric acid, but in large-scale tests, the protein would be cleaved off using 3C protease. The elution from nProteinA beads was cleaner than that from ProteinG beads so nProteinA beads were used from this point onwards.

Approximately 2 litres of media from the mouse TLR4 CHO cells was then harvested and incubated with nProteinA beads. Notably, the media smelt bad suggesting it may be infected. In addition, after incubation the beads were clumped together, which had not been observed before.

For this large-scale test, TLR4 was cleaved off the beads using 3C protease. However, no peak was seen on a size exclusion column so samples of media and flowthrough from the nProteinA beads were run on a Western blot (Figure 16). This again shows that after incubation with the nProteinA beads there was still a lot of protein left in the media.

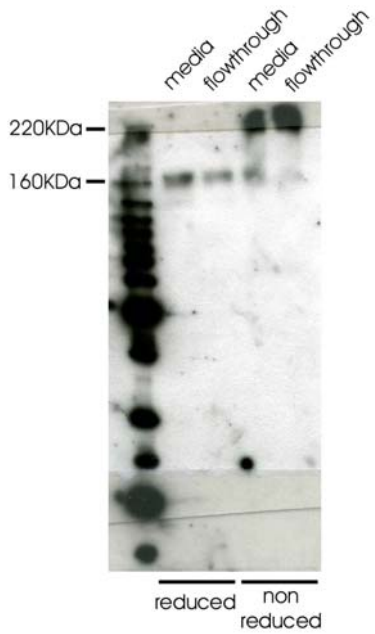


Figure 16. Mouse Fc-tagged TLR4 expression from a stable CHO cell line. Media is UltraCHO and flowthrough refers to the media after incubation with nProteinA beads.

The molecular weight of mouse TLR4 is 95 KDa based on amino acid sequence alone, but according to Naonori Uozumi in the Karp group, the monomeric, glycosylated form of mouse TLR4 runs at greater than 150 KDa on Western blots. This corresponds to the 160 KDa band seen in the reduced samples in Figure 16. However, the non-reduced samples have a band at greater than 220 KDa, showing that oligomerization of TLR4 was occurring due to dimerization of the Fc tag. This may explain why the nProteinA beads

were clumped together. This dimerization may also affect the binding of the Fc tag to the beads and explain why protein was being left in the media, as no reducing agent was added to the media before adding the beads.

A protein gel was also run of samples which no longer had the Fc tag attached (all stages after incubation with the 3C protease) (Figure 17).

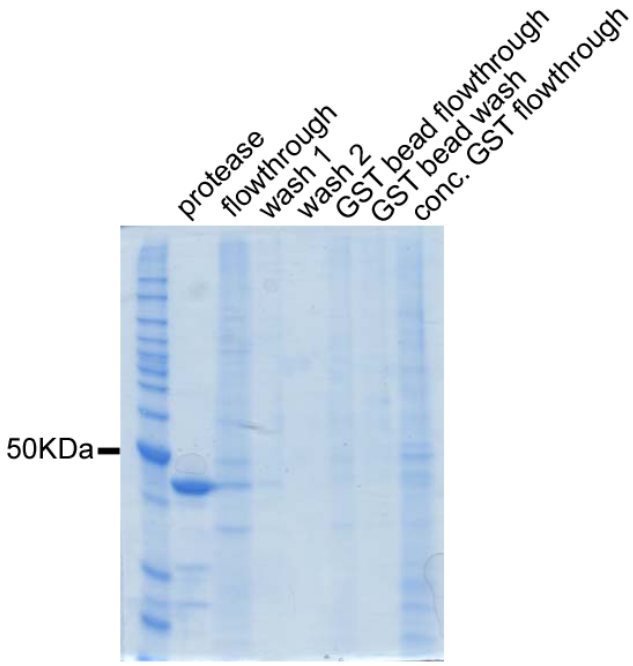


Figure 17. Protein gel of non-Fc tagged samples from mouse TLR4 expressed in a stable CHO cell line. Protease refers to GST-tagged 3C protease. Flowthrough refers to media after incubation with the 3C protease.

There are no bands at the correct molecular weight for monomeric or dimerized mouse TLR4, suggesting either that not enough protein was bound to the nProteinA beads to be detected, and/or that the protease was not efficiently cleaving it from the beads.

To test whether the protease was cleaving the protein from the nProteinA beads and what the optimum incubation time was, the media flowthrough after incubation with the nProteinA beads was re-used. However, despite having been filtered and stored at 4˚C it was slightly cloudy. A sample was taken and the reducing agent DTT added to a final concentration of 1mM, to minimise Fc tag dimerisation. This sample was re-incubated with nProteinA beads at room temperature for 2 hours then the nProteinA beads were incubated with 3C protease. After 1 hour, the liquid was drained from the beads and a sample of beads added to reducing dye and frozen at -80˚C. The liquid was then added back to the beads and incubated for another hour before sampling the beads in the same way. This was repeated once more before leaving the beads with the protease over night.

After incubation with the media, the nProteinA beads should have TLR4 with the Fc tag attached (~160 KDa depending on the glycosylation state). After incubation with the protease, the nProteinA beads should only have the Fc tag attached (50 KDa). The samples were run on a gradient gel due to the wide range of molecular weights to be detected (Figure 18).

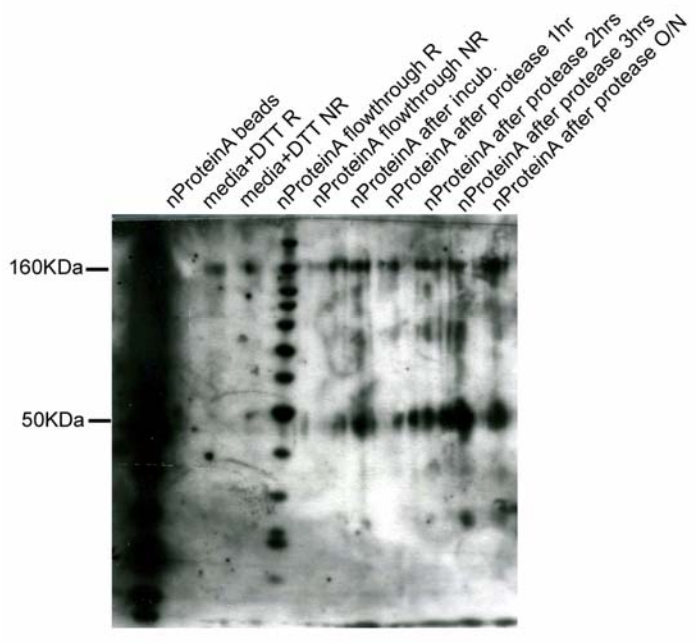


Figure 18. Mouse TLR4 binding to and cleavage from nProteinA beads . All samples were run under reducing conditions unless stated otherwise. R=reduced, NR=non-reduced.

Despite the fact that the blot is difficult to read and apparently has a lane of markers in the middle, it is possible to see that there was some TLR4 in the media (~160 KDa band) and the nProteinA beads did bind some of this. After incubation with the protease, the nProteinA beads show a band at 160 KDa meaning not all protein was cleaved, even after incubation overnight, and a band at 50 KDa which is the Fc tag alone, meaning some protein was cleaved.

A large-scale expression test was then repeated using roller bottles. However, after only 4 days of growth the cells began to detach from the bottles and the media became pale orange (indicating an acidic pH), so media was harvested at this point rather than after a week. No peak was seen a size exclusion column and the Western blot (Figure 19)

detected no protein expression even in the cells. However, this could have been due to the cells being frozen without being put into reducing dye causing the protein to degrade. It is also possible that after only 4 days there was not sufficient protein in the media to detect.

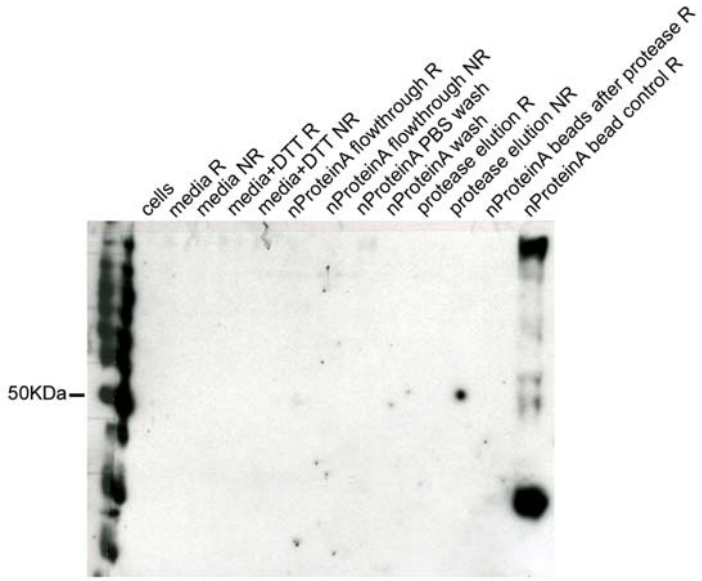


Figure 19. Large-scale expression test of mouse TLR4 expressed in a stable CHO cell line grown in roller bottles.

However, it was necessary to test whether the CHO stable cell line was still expressing TLR4 or whether the gene had been lost, so a small-scale expression test was done (Figure 20).

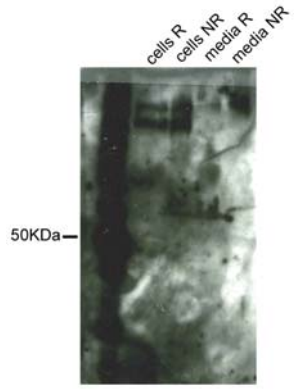


Figure 20. Small-scale expression test of mouse TLR4 expressed in a stable CHO cell line.

Expression of TLR4 was seen in the cells and a small amount was detected in the media, so the stable cell line had not lost the gene.

Due to the fact that the cells in the roller bottles were tending to detach and die after less than a week in UltraCHO media, other media were tested to see if this improved cell growth and boosted expression of TLR4 (Figure 21). However, this showed that the UltraCHO media in fact gave the best expression of mouse TLR4.

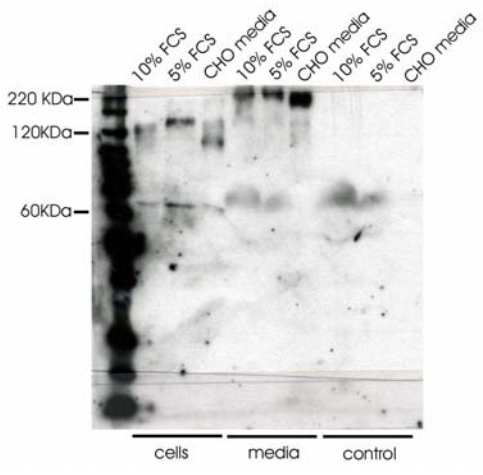


Figure 21. Mouse TLR4 expression from a stable CHO cell line grown in different media.

Given that it had been shown that TLR4 was still being produced by the CHO cells and that the UltraCHO media was the best for expression, another large-scale expression test was carried out in roller bottles, but no peak was seen on a size exclusion column. The Western blot is shown in Figure 22.

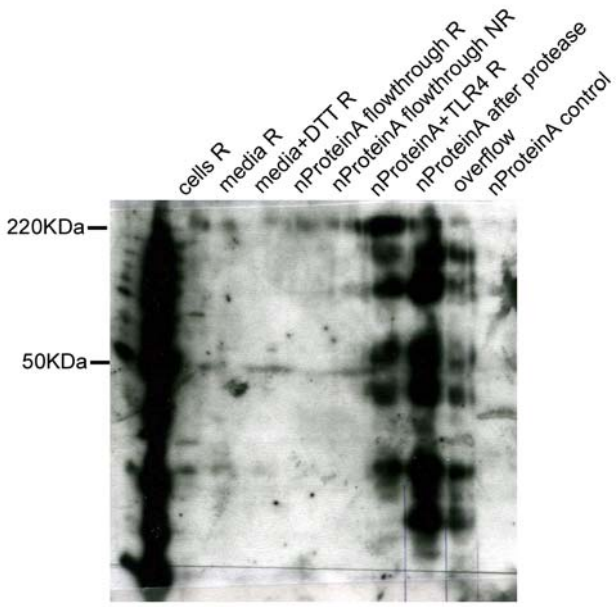


Figure 22. Large-scale expression test of mouse TLR4 from a stable CHO cell line grown in roller bottles. R=reduced, NR=non-reduced.

Very little expression was seen in the cells, even though the cells had been frozen in reducing dye overnight before the gel was run to prevent protein degradation by proteases. Several bands were visible for the lanes containing nProteinA beads. The highest band (~220 KDa) decreases in intensity after the beads had been incubated with protease, and there is a faint band at this molecular weight for the two media lanes. This also corresponds to the molecular weight seen for TLR4 in the media in Figure 21. The identity of the other bands presumably correspond the Fc tag (monomeric form 50 KDa but this can dimerize) and to degradation products of TLR4 or the Fc tag.

Finally, due to the fact that in small-scale expression tests using flasks, TLR4 was being expressed in the cells and the media but in large-scale tests in roller bottles it often was not, it was possible that this stable cell line did not grow well on the plastic used in the roller bottles. Therefore a test was carried out where expression was evaluated in cells grown in a roller bottle, cells grown in a flask and cells grown in a triple flask. The results are shown in Figure 23.

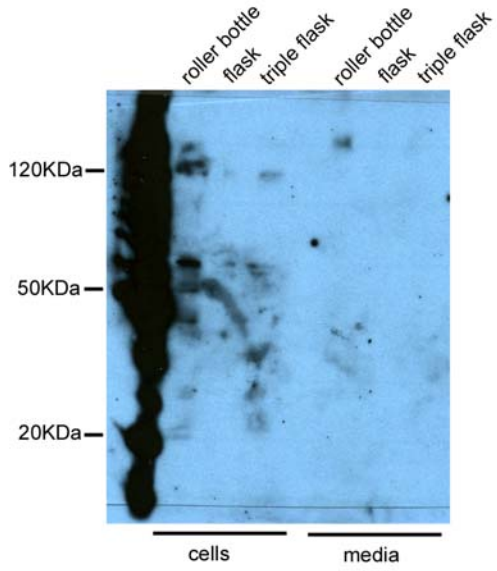


Figure 23. Comparison of expression of mouse TLR4 from a stable CHO cell line grown in roller bottles, flasks and triple flasks.

In this test, more cells were loaded from the roller bottles than from the flasks, as the cells came off the plastic in clumps. However, the Western blot shows that TLR4 is being produced by cells in the roller bottles (as well as in the flasks), but a signal is only visible in the media from roller bottles. This raises the question of why TLR4 expression is only detected in some of the large-scale expression tests but not in others. The only difference

in this blot is that neither the cells nor the media had been frozen before being loaded onto the gel.

4.3.10 Expression of Human RP105 and Human MD-1

Expression tests were also done using constructs 29 (RP105), 32 and 33 (MD-1) made from DNA provided by the Karp group. The results are shown in Figure 24.

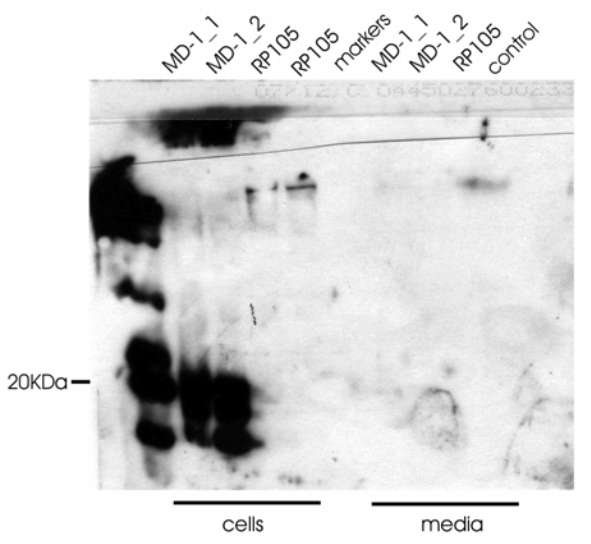


Figure 24. Expression of human MD-1, MD-2 and RP105 in mammalian 293T cells. MD-1_1 refers to construct 33, MD-1_2 to construct 32 and RP105 to construct 29. 2 lanes were loaded with RP105 cells as the first well was blocked. Control is untransfected media.

The markers appear to have degraded and have not run well on this gel. However, it can be seen that MD-1 and MD-2 were expressed in the cells but not secreted, and RP105 had a very low level of expression in the cells and no secretion.

4.3.11 Baculovirus Expression

Due to the limited success in expressing TLRs in mammalian 293T cells, a parallel set of experiments were done in the baculovirus expression system. A small trial was done first to test expression and secretion of human MD-1, MD-2, RP105, TLR3 and TLR4. All these constructs used their natural signal sequences as the pBacPAK9 vector did not have a signal sequence included. Sequencing of constructs cloned into pBacPAK9 showed that MD-1, MD-2, TLR3 and TLR4 were correct, but the results for RP105 were very poor.

The first western blot was very unclear and appeared to have moved during development, but seemed to show that MD-1 and MD-2 were expressed in the cells. No signal was seen for any of the other constructs. This blot was repeated using frozen samples but very little signal was seen, probably due to the fact the samples were not frozen in reducing dye, so may have been proteolytically degraded. Another small-scale expression test was therefore carried out for MD-1 and MD-2 (Figure 25). This shows that both proteins were expressed in the cells but neither was secreted into the media.

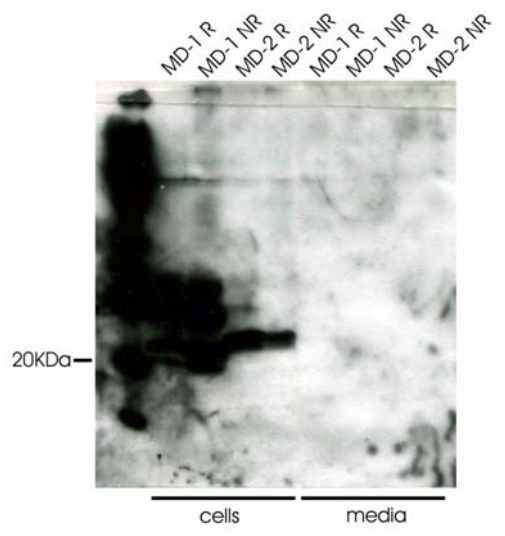


Figure 25. Small-scale expression test of human MD-1 (construct 33) and MD-2 (construct 35) in SF9 insect cells.

Finally, the human RP105 DNA from Chris Karp was used to create construct 30. Sequencing of the construct cloned into pBacPAK9 showed the sequence was correct. Small-scale expression tests were carried out for this construct, and also repeated for MD-1 (construct 33) and MD-2 (construct 35). The results are shown in Figure 26.

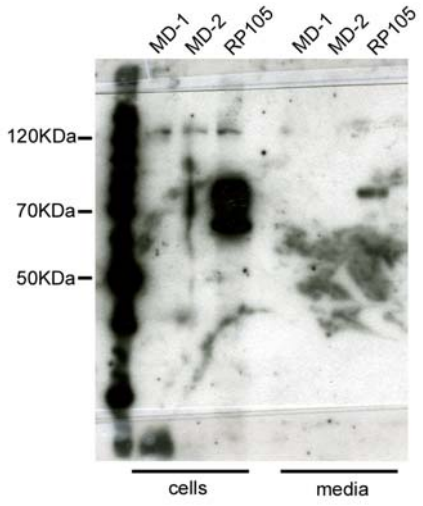


Figure 26. Expression of MD-1, MD-2 and RP105 in insect SF9 cells.

This time neither MD-1 or MD-2 were expressed in the cells, but RP105 was expressed in the cells and secreted in small amounts.

A summary of all the work done is shown in Figure 27.

Figure 27. Summary of work.

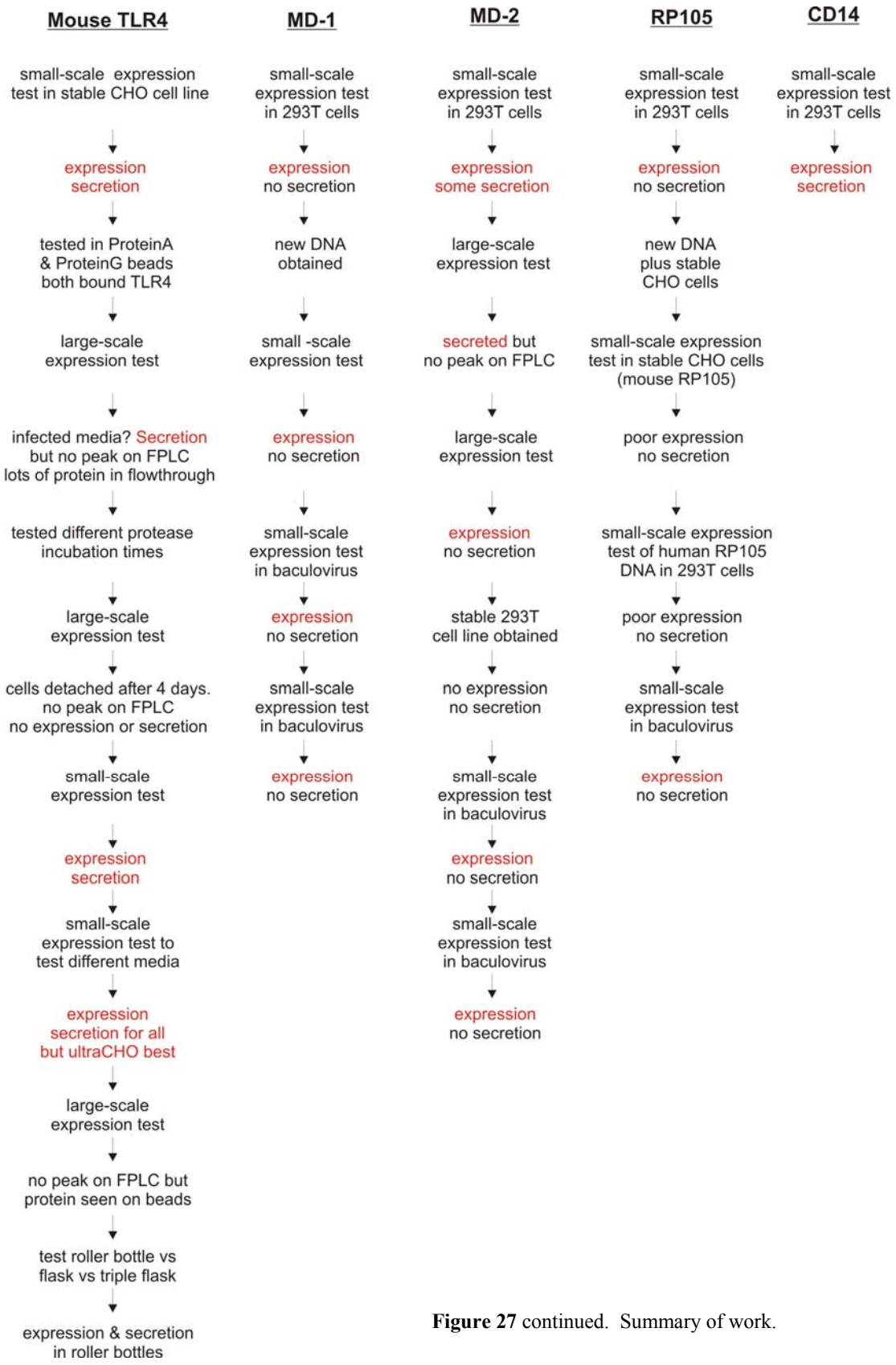


Figure 27 continued. Summary of work.

4.4 Conclusions

Despite many attempts to express human and mouse TLRs and associated proteins, both in mammalian and baculovirus expression systems, no protein was obtained in sufficient amounts to enable crystallization trials to be set up. The proteins for which structures had already been solved (TLR3 and CD14) were used as controls and expression and secretion was seen for both of these. This, combined with the fact that no other TLR structures have been published since 2005, suggests that the proteins that were not secreted may require unknown chaperones to allow proper folding and subsequent secretion from the cells, or that other expression systems may give better results.

The structure of human MD-2 was recently solved (Ohto, Fukase et al., 2007) by expressing the protein in yeast, so this expression system may be worth pursuing for expression of other TLR and related proteins. In addition, the construct for MD-2 used by this group consisted of residues 17-160. This includes two residues from the end of the predicted signal sequence and so differs by two residues compared to the equivalent construct used in this thesis (residues 19-160). However, due to the success of expression and secretion using the longer construct it would be worth attempting to express this construct in the mammalian expression system and trying a similar construct for MD-1.

Chapter 5

Development of a Protein Disorder Prediction Tool

5 Development of a Protein Disorder Prediction Tool

5.1 Summary

A novel approach to protein disorder prediction was developed in 2004 based on an extension of the bio-basis function neural network (BBFNN), a technique originally developed for the detection of protease cleavage sites (Thomson, Hodgman et al., 2003) and written in collaboration with Dr Ron Yang (Department of Computer Science, University of Exeter). An initial implementation demonstrated potential for detecting regions of disorder, but prediction accuracies per residue were low (Thomson and Esnouf, 2004). The work described here substantially improves on this original method, and the resulting tool is termed the Regional Order Neural Network (RONN).

The first version of RONN to be made publicly available (RONNv3) was published in June 2005 (Yang, Thomson et al., 2005) and in blind tests was shown to give the best overall performance when compared to 8 other protein disorder prediction tools. RONN was subsequently re-trained using a larger data set and released in May 2007 (RONNv3.1).

Finally, attempts were made to establish whether there is a signal in the amino acid sequence of a protein which triggers the change from order to disorder or vice versa with the aim of improving the detection of exact transition points between order and disorder.

5.2 Contributions to RONN

Rebecca Hamer and Dr Ron Yang collaborated on the design of the BBFNN, with Rebecca Hamer responsible for the novel use of mutation matrices to extract numerical information from amino acid sequences for input into the neural network. Dr Ron Yang wrote the BBFNN in the programming language C. Rebecca Hamer collected all data sets for training and testing, did all data analysis work and trained and tested RONN. Robert Esnouf gave invaluable advice on data analysis and presentation of figures.

5.3 Methods

5.3.1 The Bio-Basis Function Neural Network

If two proteins have similar biological functions then the primary sequences usually demonstrate a significant similarity which can be detected by sequence alignment techniques, usually using a mutation matrix to score the similarity. Conversely, the biological function of a query sequence is expected to be the same as that of a protein sequence of known function if the alignment score is high enough, with tools like BLAST and its variations commonly used for detecting this similarity (Altschul, Gish et al., 1990; Altschul, Boguski et al., 1994; Altschul, Madden et al., 1997). Here, this expectation is applied to the detection of disordered regions: sequences are compared to a series of sequences of known folding state (ordered or disordered) and the alignment scores against these sequences are used to classify each sequence as ordered or disordered using a suitably trained neural network.

The decision about the likelihood of disorder is based on alignments to an ensemble of sequences of known folding state. Consider a database consisting of sequences corresponding to M ordered and N disordered regions of proteins. How can a decision on the folding state of an unknown protein be based on these $M+N$ alignment scores? Finding the best alignment amongst the ensemble is one option, but this might not be the best way if some sequences are wrongly labelled or if none of the scores is above a threshold value. A natural generalisation is to regard each known sequence as a *prototype* as in prototype theory (Dasarathy 1991; Dasarathy 1994). The recognition of a

property is then based on the statistical relationships between a query and the prototypes and an obvious method of implementation uses neural networks. This idea was originally applied to a sequence-based biological problem, the analysis of protease cleavage rates of specific sub-sequences, and was named the bio-basis function neural network (BBFNN) method (Thomson, Hodgman et al., 2003; Yang and Thomson 2005). A distinguishing feature of this approach is that individual amino acids and sequences are not represented in some arbitrary feature space (such as hydrophobicity, charge *etc.*) according to known properties. Rather, 'distances' (determined by sequence alignment) from a subset of well characterised prototype sequences are calculated and training of the neural network is performed in this 'distance' space (Figure 1).

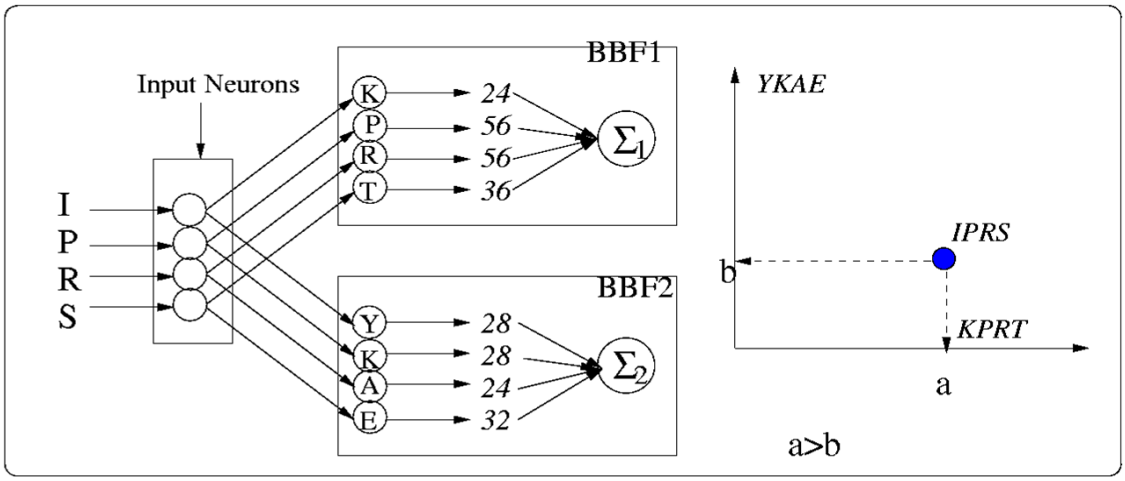


Figure 1. Comparison of an input sequence to prototype sequences and generation of similarity scores. The graph on the right shows that the sequence IPRS is more similar to the sequence KPRT than to YKAE based on the summed similarity scores.

The original BBFNN was developed for analysing sub-sequences of fixed length, transforming a series of alignment scores to a similarity value in a monotonically decreasing way. However, since the length of disordered and ordered regions varies, for this implementation the BBFNN had to be revised by using the concept of non-gapped homology alignment to maximise the alignment score between pairs of sequences (Thomson 2004). In RONN, the prototype sequences can have different lengths, although they must be at least as long as a predefined window size, and sub-sequences for a query sequence (of this window size and centred on each residue in turn) are then aligned to all the prototypes. The resulting homology scores are used for statistical pattern recognition to give a probability of disorder for each query sequence window, and these scores are averaged to give a probability of disorder for each residue in the query sequence (Figure 2).

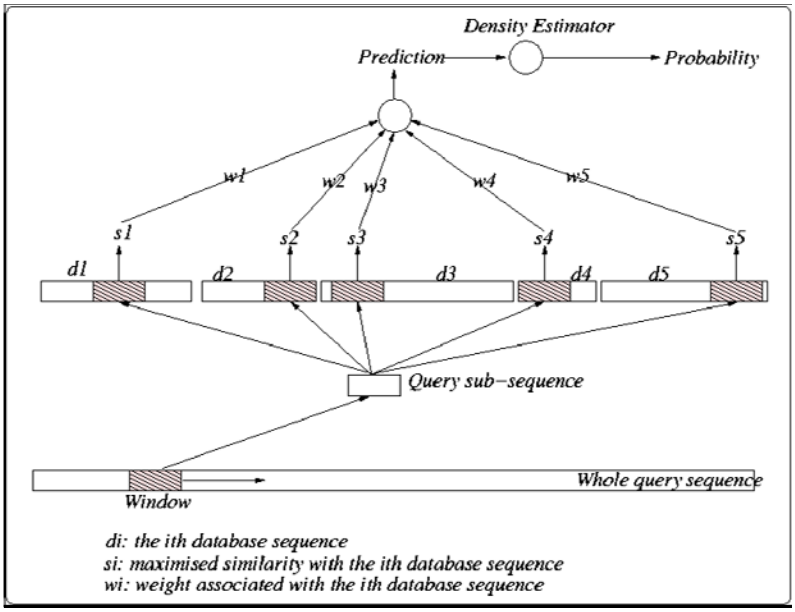


Figure 2. The bio-basis function neural network. A window is slid along the query sequence and compared to each corresponding window in each database sequence (prototype). The window in each prototype with the highest similarity score to the window from the query sequence is used for prediction.

5.3.2 Compilation of Datasets for Training and Testing RONN

To develop disorder prediction techniques using neural networks it is essential to use a large and reliable data set of ordered and disordered sequences for training, validation and testing. Considerable effort was devoted to the compilation of a comprehensive data set based on a complete analysis of entries in the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (Boutselakis, Dimitropoulos et al., 2003).

For each entry in the MSD, the sequence of observed residues derived from ATOM coordinate records in the Protein Data Bank entry (PDB; (Berman, Westbrook et al., 2000)) is aligned both against the complete protein sequence as shown by the SEQRES records in the PDB entry and against the sequence in the corresponding UniProt entry (The Universal Protein Resource (2007)) These two alignments are merged to give the complete residue-level mapping between the sequence of the polypeptide used in the experiment and its UniProt counterpart thereby detecting errors and avoiding any dependence on the PDB REMARK 465 records to identify unobserved residues.

5.3.2.1 Training and testing datasets for RONN version 3

All 25931 entries in the MSD on April 29 2004 were used for analysis. The MSD was not downloadable at this point so these entries were obtained directly from Phil McNeil at the MSD. 7327 entries were found for which there were unobserved regions of at least five consecutive residues. These entries were divided into two sets: those with at least one

disordered region of more than twenty consecutive residues (long set; 1573 entries) and the rest (short set; 5754 entries). Further filters were applied to each set:

- only structures obtained by X-ray crystallography or NMR were included (*i.e.* excluding theoretical models and structures determined by electron diffraction, electron microscopy or unspecified methods).
- multi-component complexes were excluded as disordered regions may have undergone disorder-to-order transitions in these structures.
- only the entry with the highest number was included for entries with the same 3-letter code (*e.g.* 1MOB and 2MOB) since, in general, this entry is the most recent. (Although not always true, this acts as a useful filter for very similar sequences.)
- for entries containing multiple chains (of identical sequence), only residues which were not observed in all chains were used.

After filtering, the long set contained 872 entries from which 105 entries were randomly selected to form the main blind test set. A secondary test set was created by adding randomly selected members of the filtered short set to the main blind set. From the remainder of the long set, sequences for 1691 ordered regions and 983 disordered regions were extracted for training and validation of RONN.

At this point, each of the four data sets (ordered training set, disordered training set, main blind set and secondary blind set) was filtered to remove similar sequences using Cd-Hit (Visintin, Iliev et al., 2006). For a sequence identity cut-off of 90% many sequences were removed but as this cut-off was reduced to about 60% relatively few additional

sequences were excluded. Below about 60% the rate of removal increased again as matches were found between quite distantly related sequences. Thus, a 70% sequence identity cut-off was selected as the best compromise between maximising the number of sequences included and minimising the redundancy in the data. The sizes and compositions of these different data sets are summarized in Table 1.

Table 1. Data sets for training and blind testing of RONNv3

	Training set*	Main blind set*	Secondary set*
Proteins (MSD entries)	—	80	190
Ordered regions	891	—	—
Disordered regions	530	—	—
Ordered residues	170923	29909	64838
Disordered residues	19427	3649	5680
Total residues	190350	33558	70518

* The training set provides prototypes and is used for training and validation. Proteins in the main blind set contain at least one region of disorder of at least 21 consecutive residues. Proteins in the secondary blind set contain at least one region of disorder of at least five consecutive residues.

In addition, a set of disordered sequences where disorder had been established experimentally by methods other than X-ray crystallography was obtained from (Uversky, Gillespie et al., 2000). Of the 91 disordered regions described it was possible to identify precise sequences for 79 which formed the disordered part of the test set. To balance these, a set of 80 protein sequences were added that were annotated as being ordered on the PONDR® web site (www.pondr.com) (retrieved in February 2003). This final blind test set comprised 16568 ordered residues and 14462 disordered residues

5.3.2.2 Training datasets for RONN version 3.1

All 32112 entries in the MSD on 30th May 2006 were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/xml and used for analysis. The same filters were applied to this data as were applied for RONNv3, but here only structures solved by X-ray crystallography were used, and in this version of the MSD obsolete entries which have been replaced with new PDB identifiers have been removed so it was not necessary to only take into account entries with the same 3-letter code. Therefore 25830 structures were selected which were solved by X-ray crystallography with a resolution of 2.5Å or better and were not multi-component complexes. From these protein sequences, 27723 ordered and 10244 disordered sequences were extracted.

The ordered dataset was then filtered to remove redundancy using Cd-Hit with a sequence identity cut-off of 70%, and a minimum sequence length of 19 residues was applied which corresponds to the window size used for training RONN. This left 9495 ordered sequences.

In addition, redundancy was removed from the ordered and disordered datasets using Cd-Hit with a sequence identity cut-off of 100% and a minimum sequence length of 7 residues (the minimum sequence length that Cd-Hit can deal with). This left 15550 ordered sequences and 3883 disordered sequences. BLAST was then used to compare each sequence in this disordered dataset to each sequence in the ordered dataset to find any sequences which were identical in both datasets. A sequence was only considered to be a 'hit' if there was 100% identity to the disordered query sequence, with no gaps in

either sequence. For sequences less than 15 residues long, BLAST was run with the word length (-w) set to 2 and the e value (-e) set to 20000, as advised in the Blast documentation. The results are shown in Table 2.

Table 2. Results of using BLAST to find sequences in the ordered data set which were identical to sequences in the disordered data set. Blast was also used to find disordered sequences which matched sequences in complexes – see section 5.3.2.1.

Minimum disorder sequence length	# disordered sequences	# BLAST hits to ordered sequences from non-complexes	Hits to ordered sequences from complexes
7	3883	483	75
8	3390	385	55
9	3016	321	41
10	2630	257	33
15	1476	93	13
20	902	41	8
25	563	27	4

As these BLAST hits correspond to sequences which were not consistently disordered they were removed from the disordered dataset. The hits were not removed from the ordered dataset as these sequences have the capacity to be ordered in their native states. Cd-Hit was then run on this disordered dataset with a sequence identity cut-off of 70% and a minimum sequence length of 19, to leave 815 disordered regions.

It is necessary to train RONN using an excess of ordered residues as there are many more ordered residues than disordered residues in real protein sequences. However, the 9495 ordered sequences consist of 1653289 residues which is too far in excess of the number

of disordered residues. This would not only cause the training of RONN to take an excessively long time, but would cause accuracies to be poor. For training RONNv3, 530 disordered regions (19427 residues) and 891 ordered regions (170923 residues) were used. This meant 10.2% residues in the training data residues were disordered and 89.6% were ordered. The new training data contained 815 disordered regions (26152 residues) so as a starting point, to obtain a similar ratio of disordered to ordered residues, 1300 ordered sequences (225373 residues) were randomly selected for the training set to give 10.3% disordered residues and 89.8% ordered residues. However, tests showed that increasing the number of ordered sequences to 1500 (268086 residues) improved results in blind tests (data not shown) so the final training dataset consisted of 815 disordered and 1500 ordered sequences (Table 3).

5.3.2.3 Testing dataset for RONNv3.1

A list of 1498 PDB identifiers was obtained using the advanced query page on the PDB website (http://www.pdb.org/pdb/search/advSearch.do). The search criteria were that files had been deposited between 1/9/06 and 21/3/07, were solved by X-ray crystallography with a resolution of 2.5Å or better and similar sequences were removed at 70% identity. From this list, there were 1114 corresponding XML files available from the MSD which were downloaded. Of these, there were 782 sequences which had at least 1 region of disorder of at least one residue long. BLAST was then used to compare these sequences to the whole protein sequences in the training dataset (before ordered and disordered regions were extracted) using a threshold of 90% identity with no gaps

allowed for a hit to be valid. There were 172 hits which were removed, leaving 610 sequences.

The difficulty in preparing a good test set is to ensure a reasonable distribution of lengths of disordered regions to enable rigorous testing of the predictor. Short (<10 residues) regions are very common in protein sequences but longer regions are less common, so randomly selecting a test set from these 610 sequences is not practical as most disordered regions present would be short. To alleviate this problem, all sequences containing at least 1 region of disorder of minimum length 20 residues were selected, plus a random selection of sequences containing at least 1 region of disorder with a minimum length 10 residues. Each of these sequences commonly also has other, shorter disordered regions present. The final non-redundant test set contained 205 whole protein sequences and the distribution of disordered region lengths is shown in Figure 3.

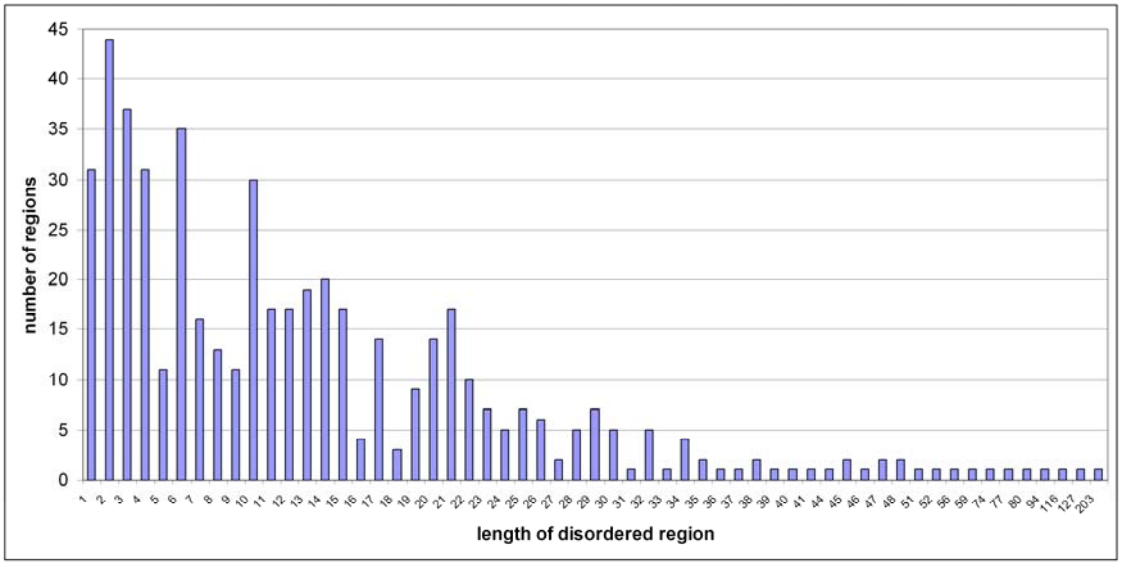


Figure 3. Distribution of disordered region lengths in the 205 blind test sequences.

The training and testing datasets for RONNv3.1 are summarised in Table 3.

Table 3. Summary of the training datasets used for RONNv3.1.

Dataset	# sequences	# residues
Raw data		
Total MSD files	32112	-
Non-complex X-ray structures with resolution 2.5Å or better	25830	-
Ordered regions	27723	-
Disordered regions	10244	-
Ordered regions after sequences 70% identical were removed	9495	1653289
Training data		
Randomly selected ordered regions from the non-redundant ordered dataset	1500	268086
Disordered regions after BLAST hits and sequences 70% identical were removed	815	26152
Blind test set	205	64676

5.3.2.4 Datasets for analysis of order-to-disorder transitions

The same 32112 files from the MSD as were used to train RONNv3.1 were used for analysis. From these files, only those structures solved by X-ray crystallography with a resolution of 2.5 Å or better were extracted, and no complexes were included in the final data set. Cd-hit (Visintin, Iliev et al., 2006) was used to remove redundancy using an identity threshold of 90%. This left 6325 protein sequences for analysis, some of which were fully ordered and some of which contained regions of disorder. From these sequences, the ORDER, DISODER and BOUNDARY datasets were extracted for analysis. The ORDER set is a collection of 11883 fully ordered protein regions and the DISORDER set is a collection of 8210 fully disordered protein regions. The BOUNDARY set consists of 2636 boundaries between order and disorder (1116 order-to-disorder and 1520 disorder-to-order transitions) each of which comprises 5 ordered residues on one side of the boundary and 5 disordered residues on the other. The number of identical transitions in each protein with more than one identical chain was calculated by assessing how many transitions were identical in all the chains of the protein as a fraction of the total number of transitions in all the chains.

The DISPROT dataset contains proteins downloaded from Disprot (Vucetic, Obradovic et al., 2005) (www.disprot.org/downloads.php). Disprot is a curated database of wholly or partially disordered proteins containing 1108 disordered regions in release 3.5. The proportion of disordered regions with minimum length 20 residues for data from Disprot

is 45.8% (91.3% of all disordered residues), compared to 9.9% (37.2% residues) for the data from the MSD.

The average frequencies of amino acids calculated from the genomes of archaea, bacteria and eukaryotes were obtained from published data (Gilis, Massar et al., 2001) but no further information about the composition of this data was available.

The datasets are summarised in Table 4.

Table 4. Datasets used for analysis of order-to-disorder transitions.

	Data Set	**# sequences**	**# residues**
MSD Data	Selected MSD entries	6325	1724094
	Ordered regions (**ORDER set**)	11883	1654549
	Disordered regions (**DISORDER set**)	8210	69545
	Disordered regions >= 20 residues long	810	25877
	Order-to-disorder boundaries	1116	11160
	Disorder-to-order boundaries	1520	15200
	All boundaries (**BOUNDARY set**)	2636	26360
Disprot Release 3.5	Disprot entries	469	214160
	Disordered regions (**DISPROT set**)	1108	59351
	Disordered regions >= 20 residues long	508	54194

5.3.3 Training and Validation of RONN

Step 1: The training sequences are partitioned into ten 'folds'. For each run of ten-fold cross-validation, nine folds are used as prototypes (in some cases only disordered sequences are used as prototypes) and the remaining fold is randomly divided into two parts: one for training to estimate model parameters and one for validation to estimate the probability density functions for both ordered and disordered classes to allow Bayes' rule to be used for decision-making.

Step 2: The BBFNN is trained using the training sequences. A window of predetermined size is used to scan through the training sequences to generate training sub-sequences which are then aligned with each prototype in turn to maximize the non-gapped homology alignment score. The bio-basis function is then used to transform this score into a normalized similarity measurement. Each prototype is considered to be an independent variable and the relationship between these independent variables and the class labels (ordered or disordered) of sub-sequences is assumed to be simple, allowing a linear classifier to be constructed. Parameters weighting the similarity measurements are determined using a pseudo-inverse method (Thomson *et al.*, 2003; Yang and Thomson, 2005).

Step 3: The discrimination threshold between ordered and disordered sequences is optimized using the validation sequences. These sequences are divided into validation sub-sequences in the same way as for training sub-sequences. After input into the model

constructed in step 2, the probability density functions of the outputs are assumed to follow two Gaussian distributions, one for ordered and the other for disordered sequences. A parametric method is used to approximate these two functions. Based on the approximated probability density functions, an optimum threshold for discrimination can be determined using Bayes' rule (Duda *et al.*, 2002). Cost functions are used for final decision making. Let the probabilities of order and disorder generated by the trained model be P_o and P_d, respectively, with associated cost functions C_o and C_d defined such that $C_o + C_d = 1$. The final probability of disorder is then $C_d P_d / (C_o P_o + C_d P_d)$ and is assigned to all the residues within the current window.

Step 4: For each position of the sliding window, all residues within the window are assigned the same probability of disorder. When the window is moved on one residue, another probability of disorder is generated. Thus, for a window size of W residues, every residue will have up to W probabilities of disorder assigned. To arrive at a single estimate for each residue these estimates are simply averaged and the residue is predicted to be disordered if the averaged probability of disorder is >0.5.

Step 5: The training and validation process (steps 2 to 4) is repeated ten times using the non-overlapping folds created in step 1.

5.3.4 Blind Testing of Disorder Prediction Methods

Eleven methods for predicting protein disorder were tested against the blind test sets. Since there was no way of knowing which sequences were used for training each method other than RONN, the random selection procedure of test sequences described above was the most appropriate. The methods tested were: RONN (v3 and v3.1), PONDR® (VL-XT method), FoldIndex, all three variants of DisEMBL 1.4 (coils/loops, hot loops and REMARK 465), GlobPlot 2, DISOPRED2 and PreLink, VSL2 and IUPred (short and long predictors) (Table 5). Except for PONDR® and VSL2, which were licensed and installed locally, all predictions were made by submitting sequences to the public web servers. For each prediction, only the binary classification of each residue as either ordered or disordered was recorded and analyzed (rather than the probability of disorder) since this most closely reflects the use of these tools by structural biologists.

Table 5. Selected disorder prediction programs and their URLs.

Method	URL	Reference
DisEMBL	http://dis.embl.de/	(Linding, Jensen et al., 2003)
DISOPRED2	http://bioinf.cs.ucl.ac.uk/disopred/disopred.html	(Ward, Sodhi et al., 2004)
FoldIndex	http://bip.weizmann.ac.il/fldbin/findex	(Prilusky, Felder et al., 2005)
GlobPlot	http://globplot.embl.de/	(Linding, Russell et al., 2003)
IST-ZORAN/VSL-1	http://www.ist.temple.edu/disprot/predictorVSL1.php	(Vucetic, Obradovic et al., 2005)
IUPRED	http://iupred.enzim.hu/	(Dosztanyi, Csizmok et al., 2005)
VSL2	http://www.ist.temple.edu/disprot/predictorVSL2.php	(Obradovic, Peng et al., 2005; Peng, Radivojac et al., 2006)
PONDR®	http://www.pondr.com/	See text
PreLink	http://genomics.eu.org/PreLink	(Coeytaux and Poupon 2005)

Disopred only allows one sequence at a time to be submitted and the results are only available via email. For this reason, only the smaller blind test set was used to compare results to RONNv3, and Disopred was not compared to RONNv3.1 as the test set used was too large for this to be practical. The results format from the Prelink tool changed after it had been compared to RONNv3 making it impossible to automate blind testing using the large test set, so Prelink was not compared to RONNv3.1.

5.3.5 Measuring the Accuracy of Prediction

Quantifying the accuracy of binary classification procedures (such as the prediction of order or disorder for a residue) is a common problem in computer science and many different measures have been devised. In each case the most appropriate measure will depend, amongst other things, on the relative frequencies of the two classes in the data set and the relative importance of incorrect classification either way. For comparisons against reference data the results can described by four integers:

- ***TN*: true negatives**, *e.g.* the number of correctly classified ordered residues.
- ***FN*: false negatives**, *e.g.* the number of disordered residues incorrectly classified as ordered.
- ***FP*: false positives**, *e.g.* the number of ordered residues incorrectly classified as disordered.
- ***TP*: true positives**, *e.g.* the number of correctly classified disordered residues.

Further measures are derived from these values, for example the sensitivity is TP/(TP+FN); the specificity is TN/(TN+FP); and the precision is TP/(TP+FP). However, none of these measures is adequate in isolation (*e.g.* predicting all residues to be disordered would give sensitivity = 1, but specificity = 0 and precision equal to the fraction of disordered residues in the test set) so more complex measures are required.

Three further measures can be defined:

- Accuracy is (TN+TP)/(TN+FN+FP+TP)
- Matthews' correlation coefficient (Matthews 1975)

  $(TN \cdot TP - FN \cdot FP)/\sqrt{[(FN+TP)(TN+FP)(FP+TP)(FN+TN)]}$
- Probability excess is (TN·TP−FN·FP)/[(FN+TP)(TN+FP)]

The accuracy is heavily affected by the relative frequencies of the two classes (*e.g.* if 90% of residues are ordered then randomly classifying 90% of residues as ordered will give an accuracy of 0.82 for an algorithm that only identifies 10% of disordered residues). The Matthews' correlation coefficient is a better measure since all random algorithms have a coefficient of 0 while a perfect algorithm has a coefficient of 1, but it is still strongly influenced by the relative class frequency in the test set. The probability excess also varies between 0 (for random picking) and 1 (perfect prediction) and in addition is formally independent of the relative class frequency in the test set. This value is also conveniently measured from a plot of sensitivity against specificity, being the distance of a point from the diagonal line corresponding to random algorithms. Indeed, the probability excess can be shown to be simply sensitivity+specificity−1. These

properties make the probability excess the favoured measure of performance and the one used for ranking different algorithms.

5.4 Results

5.4.1 RONNv3

5.4.1.1 Training and Optimizing of RONNv3

Within the framework of the RONN algorithm outlined above several choices remain such as (a) what mutation matrix to use for alignments, (b) what prototypes to use, (c) what window size to use and (d) what choice of cost function to use. All these choices were explored by training and validating RONN and then measuring its performance against the two blind test sets shown in Table 1.

The mutation matrix is central to RONN in that it determines how similar a window of a query sequence is to each of the prototype sequences. In sequence alignment tools like BLAST the mutation matrix measures the effective likelihood of any given mutation occurring by chance during evolution. Such matrices include the Dayhoff and Blosum62 matrices (Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992). For disorder prediction the criterion of similarity might be considered to be the physical similarity between residues, which in turn determines the relative propensities for being in ordered/disordered regions. RONN was tested using the Dayhoff matrix, the Blosum62 matrix and a matrix devised

to encode size, charge and hydrophobicity differences. In all cases the Blosum62 matrix was found to perform best (data not shown) and so was used for all subsequent testing.

The number of prototype sequences used is obviously critical to the success of the algorithm and also affects its execution speed significantly. In general, the more sequences the better the chance of finding a significant match to any given query sequence, but redundancy in prototypes should also be avoided since it could lead to inappropriate weighting. The second question is whether to use just one class of prototypes or a mixture (ordered and disordered). It was found that including ordered prototype sequences did not improve the accuracy of the method (data not shown) and all subsequent testing solely used disordered prototype sequences, which also resulted in significantly faster predictions.

Blind tests for cost functions of 0.5 and 0.6 and for window sizes from 11 to 41 residues were explored systematically (Figure 4). They show that better predictions were consistently obtained with a cost function of 0.6, the optimum window size being around 15 to 21 residues with a window size of 19 residues giving the best prediction overall. Based on these blind tests, RONN version 3 has a sensitivity of 0.603, a specificity of 0.878, an accuracy of 0.849, a Matthews' correlation coefficient of 0.395 and a probability excess of 0.481.

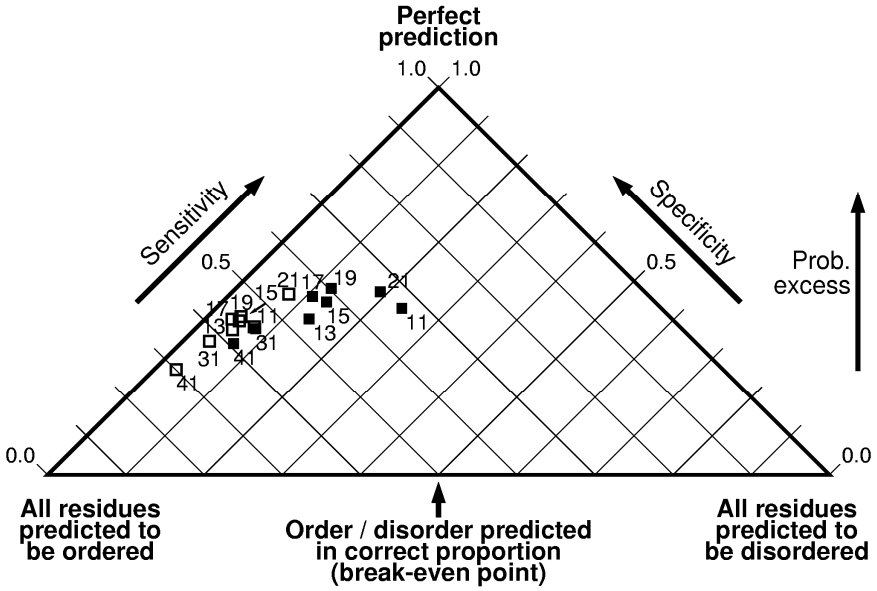


Figure 4. Effect of varying the cost function, C_d, and the window size on RONN predictions for the main blind test set. The plot shows specificity versus sensitivity, but is rotated anticlockwise by 45° to become a plot of probability excess against the ratio of ordered to disordered residues in our prediction. Thus, the better the prediction the higher up the plot it is. Predictions are marked either by empty squares (□; C_d = 0.5) or by filled squares (■; C_d = 0.6) with labels showing the size of the window in residues.

The window size giving the best performance might have been affected by the average size of disordered regions in the main blind test set. This was investigated by further trials using the secondary blind test set (Table 1) that was both substantially larger and also contained proportionately many more short regions of disorder. With this data set the best results were again obtained with a cost function of 0.6 and a 19 residue window, giving a similar result overall although the sensitivity, in particular, was somewhat lower (0.562 compared with 0.603). Thus, this combination of options performed best overall and is the one used in comparisons with other disorder prediction methods below and is also used for the RONN web server.

5.4.1.2 Use of RONN for disorder prediction

In common with other disorder prediction methods, the result of submitting a protein sequence to the RONN web server is a graph of the predicted disorder probability per residue. RONN has been trained and validated to optimize performance based on a probability cutoff of 0.5. However, in some situations it is useful to consider the raw probability predictions when analyzing a sequence, particularly when looking for short regions of disorder such as loops and linkers. Using a 19-residue sliding window the final prediction for each residue is an average of (up to) 19 predictions. Thus, the sensitivity to short disordered regions (<10 residues) forming loops and linkers is somewhat compromised. A different manifestation of the same effect is that RONN has some difficulty in precisely defining the first and last residues of disordered regions, as shown in the plot for PDB entry 1GJJ (Figure 5). As mentioned above, tests with the secondary blind set (Table 1) containing proportionately more short regions showed that the problem is not too severe. Nevertheless, if RONN is being used specifically to search for short regions of disorder then inspection of probability plot may be necessary. The plot for PDB entry 1Q1Q (Figure 5) shows an example of this for a disordered loop from residues 96–102, where, although the predicted probabilities are high for the loop region, they never quite reach the threshold required for a prediction of disorder.

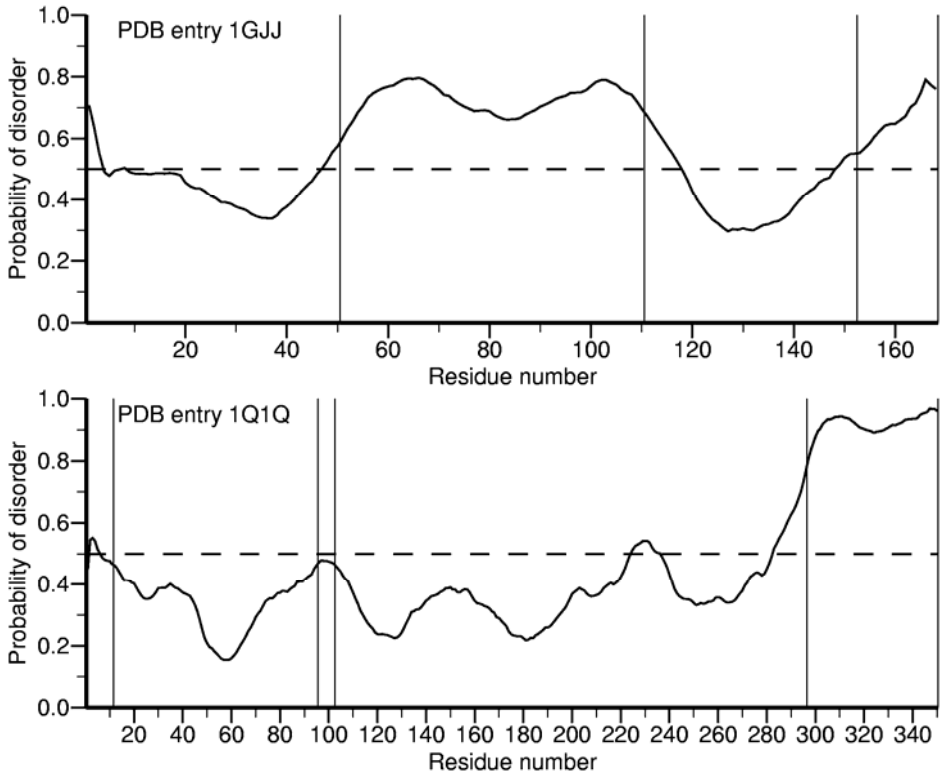


Figure 5. RONN plot of disorder probability per residue for two proteins from the main blind test set (PDB accession codes 1GJJ and 1Q1Q). The horizontal dashed lines mark the threshold for disorder prediction; the shaded regions show the disorder identified from the PDB entries

5.4.1.3 Blind testing of RONNv3 and comparison to other disorder prediction methods

Publicly available web servers for a panel of disorder prediction methods were used to make predictions. The methods tested were RONN, PONDR®, FoldIndex, DisEMBL, GlobPlot, DISOPRED2 and PreLink. The server for DisEMBL provides three choices for prediction. All three were tested independently and these methods are referred to below as DisEMBL (coils) for the definitions based on coils and loops; DisEMBL (hot)

for the definitions based on loops with high *B* factors; and DisEMBL (465) for the definitions based on the residues annotated in the REMARK 465 records. Thus, nine disorder prediction methods were separately analyzed using all 80 proteins in the main blind test data set. (The DISOPRED2 server has a limit of 1000 residues per protein so in this case the results were based on the remaining 77 proteins.). The results can be conveniently summarized by the number of residues in each prediction class (Table 6).

Table 6. Raw results from the blind testing of nine disorder prediction methods against the main blind test set of 80 proteins.

Method	Predicted order (TN)	Missed disorder (FN)	Missed order (FP)	Predicted disorder (TP)
RONN	26275	1449	3634	2200
DISOPRED2[a]	25890	2104	734	1432
PONDR®	24391	1616	5518	2033
DisEMBL (hot)	25121	1854	4788	1795
DisEMBL (465)	29345	2432	564	1217
FoldIndex	24245	1867	5664	1782
PreLink	28312	2786	1597	863
GlobPlot	24255	2292	5654	1357
DisEMBL (coils)	12687	947	17222	2702

[a]Prediction based on 77 proteins (comprising 30160 residues).

From the raw prediction data in Table 6 it is possible to derive all the different performance measures discussed above (Table 7). The order of the algorithms in these tables is determined by the probability excess. The sensitivity and specificity are further used to create a plot of probability excess against the balance of prediction outcomes (Figure 6).

Table 7. Performance measures calculated from the blind testing of nine disorder prediction methods against the main blind test set of 80 proteins.

Method	Sens.	Spec.	Acc.	MCC	Prob. excess
RONN	0.603	0.878	0.849	0.395	0.481
DISOPRED2[a]	0.405	0.972	0.906	0.470	0.377
PONDR®	0.557	0.816	0.787	0.278	0.373
DisEMBL (hot)	0.492	0.840	0.802	0.260	0.332
DisEMBL (465)	0.334	0.981	0.911	0.437	0.315
FoldIndex	0.488	0.811	0.776	0.224	0.299
PreLink	0.237	0.947	0.869	0.219	0.183
GlobPlot	0.372	0.811	0.763	0.140	0.183
DisEMBL (coils)	0.740	0.424	0.459	0.104	0.165

[a]Prediction based on 77 proteins (comprising 30160 residues).

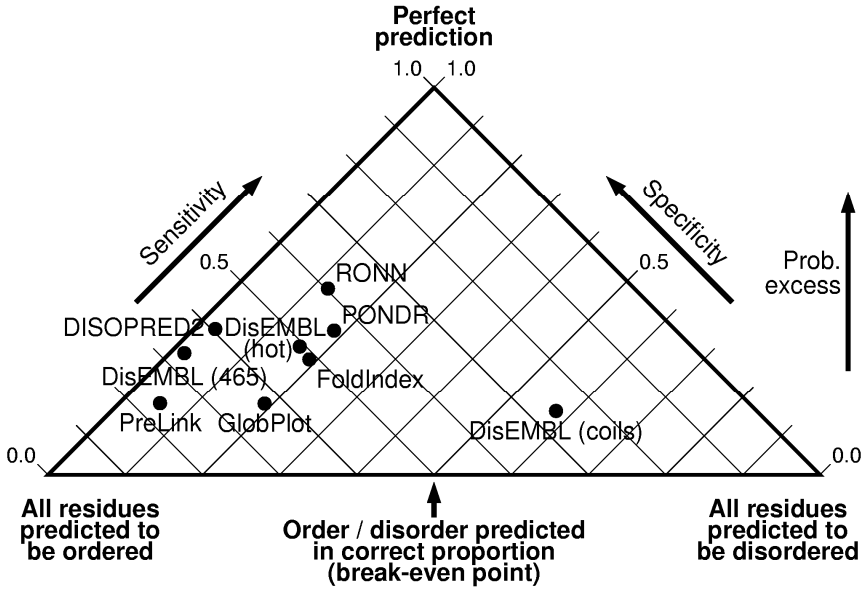


Figure 6. Plot comparing the outcome of the blind testing of nine different disorder prediction algorithms. The detailed form of the plot is described in the legend to Figure 4. Predictions are marked by filled circles (●) with a label giving the method by which that prediction was produced.

The blind tests show that the algorithms perform very differently, roughly dividing into three groups based on specificity. DisEMBL (coils) has an extremely low specificity revealing a significant over-prediction of disordered residues (59% of all residues in the test set are predicted disordered, whereas only 11% are actually annotated as such), *i.e.* the right-hand side of Figure 6.

DISOPRED2, DisEMBL (465) and PreLink form a group of methods with very high specificities, in other words they predict relatively few ordered residues to be disordered. Given the preponderance of ordered residues in the main blind test set, these methods achieve very high accuracy and indeed they all score higher than any of the other algorithms based on this measure (Table 7). However, they achieve this at the expense of missing a significant number of disordered residues (these methods predict 7%, 5% and 7%, respectively, of residues to be disordered rather than 11%; *i.e.* towards the left-hand edge of Figure 6) and in terms of sensitivity these methods perform moderately (in the case of DISOPRED2) to poorly. As discussed above, the Matthews' correlation coefficient is also strongly dependent on relative class frequencies and based on this measure DISOPRED2 and DisEMBL (465) are the best performing algorithms. Despite its under-prediction of disorder DISOPRED2 does give some very good predictions and is ranked second overall based on probability excess, while DisEMBL (465) is ranked fifth due to its missing such a high proportion of disordered residues.

The final group of algorithms is characterized by a better balance between order and disorder prediction (to the left of centre in Figure 6) and the ranking of performance

within the group is RONN (first), PONDR®, DisEMBL (hot), FoldIndex and GlobPlot. Although the balance of these predictor algorithms appears to be better than those of the other methods studied, the error rate is still quite high: a disordered residue has a 37% chance of being predicted disordered with GlobPlot while even with RONN the chance is only 60%. Based on the probability excess RONN is ranked best of all nine methods examined.

The main blind test set contains a higher proportion of disordered residues, thereby reducing the distortion of some performance estimates that arises from the relative class frequencies. Thus, the results presented here are thought to form the largest and most balanced blind test of disorder prediction algorithms in the literature.

Nevertheless, to demonstrate the independence of the training and testing tests the disorder prediction methods were also evaluated against a final test set of sequences compiled from other workers in the field. Many sequences in this set correspond to fully disordered proteins which, by definition, could not be found by analysis of the PDB. Against this panel of 159 sequences RONN again gave the best predictions (Table 8). Several methods showed significantly improved performance, with rules-based methods showing the largest gains. For example, FoldIndex gave predictions second only to those of RONN. The improved performance against these test data can be ascribed to the presence of a significant number of low-complexity disordered sequences which are relatively easy to detect. The main blind test set is believed to offer a more demanding

and realistic test of disorder prediction ability for the important 'real world' application of disorder prediction as an aid to rational construct design for structural analysis.

Table 8. Performance measures calculated from the blind testing of eight disorder prediction methods against the final blind test set of 159 sequences.

Method	Sens.	Spec.	Acc.	MCC	Prob. excess
RONN	0.675	0.888	0.789	0.580	0.563
DISOPRED2[a]	–	–	–	–	–
PONDR®	0.632	0.782	0.712	0.420	0.414
DisEMBL (hot)	0.502	0.749	0.634	0.260	0.251
DisEMBL (465)	0.348	0.978	0.685	0.430	0.327
FoldIndex	0.722	0.815	0.771	0.540	0.536
PreLink	0.319	0.991	0.678	0.430	0.310
GlobPlot	0.308	0.821	0.582	0.151	0.129
DisEMBL (coils)	0.719	0.446	0.573	0.170	0.165

[a]Predictions for DISOPRED2 not done.

As well as providing a stern test for the different algorithms, these blind tests also highlight the inadequacies of the performance measures. Sensitivity and specificity can only properly be considered in combination. Accuracy can be very misleading where the relative class frequencies are very different, and this is also a problem for the Matthews' correlation coefficient as shown by the results for DISOPRED2 and DisEMBL (465).

5.4.2 RONNv3.1

The algorithm for RONNv3.1 is identical to that used for RONNv3 (variations of the algorithm were tested but did not give better results; data not shown). However, it was trained and tested using considerably larger data sets. For all performance tests the blind test set of 205 protein sequences was used (Table 3).

5.4.2.1 BLAST of disordered sequences against ordered sequences

For the 27 disordered regions at least 25 residues long which were identical to ordered sequences (Table 2), all BLAST hits in the ordered database were to different forms of the same protein that contained the disordered sequence. For example, a disordered region in 1CHU (L-aspartate oxidase) had an ordered hit in 1KNR (L-aspartate oxidase R386 mutant).

BLAST was also used to find disordered sequences which were present in ordered sequences from complexes, to see how many sequences became ordered on complex formation (Table 2). Not all of the 75 hits of disordered sequences to ordered sequences from complexes can be assumed to be sequences which become ordered on complex formation. Some may simply be hits to identical sequences in a different protein in the complex. However, all the long hits (minimum length 15) were to identical proteins within complexes. The longest disordered sequence that becomes ordered on complex

formation is 56 residues: 1IK7 (uncomplexed Pelle Death domain, residues 1-56) hits 1D2Z (complex of the death domains of Pelle and Tube, residues 1-56).

5.4.2.2 Performance of RONNv3.1 after training with new data

Table 9 shows the difference the cost function can make to the various performance measures for models trained using the same dataset, all using a window size of 19. Although cost functions of 0.52 and 0.53 gave the same value for the probability excess, 0.53 was chosen for RONNv3.1 as it gave a higher sensitivity.

Table 9. Performance measures of RONNv3.1 with different cost functions.

cost	accuracy	sensitivity	specificity	Prob. excess
0.5	0.852	0.604	0.882	0.486
0.52	0.835	0.627	0.86	0.487
0.53	**0.826**	**0.639**	**0.848**	**0.487**
0.54	0.815	0.651	0.835	0.486
0.55	0.803	0.661	0.82	0.482

When the optimum parameters had been established for RONNv3.1 (trained using 1500 ordered training sequences and 815 disordered sequences and tested using a cost function of 0.53), the performance was compared to that of RONNv3 using the same blind test set of 205 sequences (Table 10).

Table 10. Performance measures of RONNv3.1 compared to RONNv3

Tool	Accuracy	Sensitivity	Specificity	Prob. excess
RONNv3.1	**0.826**	**0.639**	**0.848**	**0.487**
RONNv3	0.841	0.557	0.875	0.432

The sensitivity of RONNv3.1 is much higher than RONNv3 but the specificity is lower. Overall the probability excess is considerably better for RONNv3.1.

The performance of RONNv3.1 was then compared to that of other disorder prediction tools (Table 11 and Figure 7). Based on the probability excess, RONNv3.1 outperformed all the methods except VSL2.

Table 11. Comparison of performance measures for RONNv3.1 to those of other disorder prediction tools.

Tool	Accuracy	Sensitivity	Specificity	Prob excess
RONNv3.1	**0.826**	**0.639**	**0.848**	**0.487**
PONDR	0.760	0.600	0.780	0.380
FOLDINDEX	0.773	0.469	0.810	0.279
GLOBPLOT	0.784	0.369	0.835	0.204
DISEMBL_COILS	0.516	0.770	0.485	0.255
DISEMBL_HOT	0.789	0.556	0.817	0.372
DISEMBL_465	0.916	0.370	0.982	0.352
IUPRED_LONG	0.891	0.411	0.950	0.360
IUPRED_SHORT	0.904	0.517	0.952	0.468
VSL2	**0.810**	**0.742**	**0.819**	**0.560**

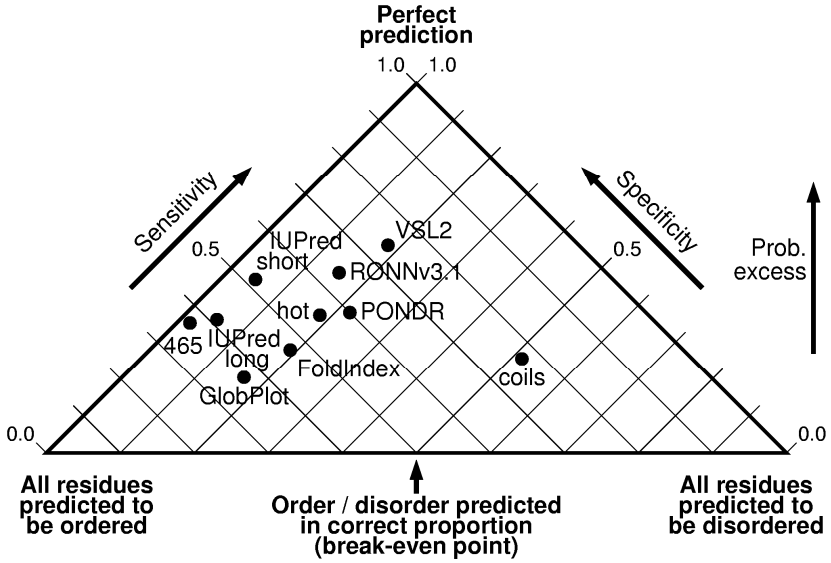


Figure 7. Plot comparing the outcome of the blind testing of nine different disorder prediction algorithms. The detailed form of the plot is described in the legend to Figure 4. Predictions are marked by filled circles (●) with a label giving the method by which that prediction was produced. Hot, 465 and coils refer to the DisEMBL methods.

Many of the blind test proteins contain up to 8 histidine residues at the N- or C-terminus (His tags) to enable the protein to be purified by metal affinity chromatography (Lilius, Persson et al., 1991). These are often disordered in protein structures and should be easily detectable by most algorithms. However, including these in blind testing may artificially increase the performance results. To counter this, all sequences of 6 or more histidine residues found at the start or end of blind test sequences were removed and performance measures recalculated (Table 12).

Table 12. Performance results when all His tags were removed from blind test sequences.

Tool	Accuracy	Sensitivity	Specificity	Prob excess
RONNv3.1	**0.824**	**0.593**	**0.849**	**0.442**
PONDR	0.758	0.550	0.780	0.330
FOLDINDEX	0.776	0.456	0.810	0.266
GLOBPLOT	0.788	0.351	0.835	0.186
DISEMBL_COILS	0.511	0.747	0.485	0.232
DISEMBL_HOT	0.789	0.525	0.817	0.342
DISEMBL_465	0.916	0.298	0.982	0.280
IUPRED_LONG	0.892	0.357	0.950	0.307
IUPRED_SHORT	0.904	0.458	0.952	0.409
VSL2	**0.808**	**0.709**	**0.819**	**0.528**

The probability excess decreased for all tools compared to results in Table 11, but these are perhaps a more reliable measure of the prediction accuracy of each tool.

It was anticipated that if very short regions of disorder (1 or 2 residues long) were discounted (*i.e.* counted as ordered) then RONNv3.1 would perform better, as the window size of 19 used may miss these very small regions. These very short regions

may not actually be disordered, and some may be due to errors in model building by X-ray crystallographers. This theory was tested for RONNv3.1 and the other prediction tools (Table 13).

Table 13. Performance measures for disorder prediction tools if all regions of disorder in the blind test set which were 1 or 2 residues long were counted as ordered

Tool	Accuracy	Sensitivity	Specificity	Prob excess
RONNv3.1	**0.827**	**0.646**	**0.848**	**0.494**
PONDR	0.761	0.602	0.779	0.382
FOLDINDEX	0.774	0.474	0.81	0.284
GLOBPLOT	0.785	0.372	0.835	0.206
DISEMBL_COILS	0.516	0.774	0.485	0.26
DISEMBL_HOT	0.789	0.558	0.816	0.374
DISEMBL_465	0.917	0.374	0.982	0.356
IUPRED_LONG	0.893	0.417	0.950	0.367
IUPRED_SHORT	0.904	0.514	0.950	0.465
VSL2	**0.809**	**0.742**	**0.817**	**0.559**

This improved the probability excess of RONNv3.1 and all other tools except for IUPRED_SHORT. This is unsurprising as IUPRED_SHORT is trained to detect short regions of disorder.

To further investigate the influence of length of disorder on prediction accuracy, graphs were plotted showing the number of correctly predicted residues for disordered regions of various lengths from the blind test set of 205 protein sequences (Figure 8).

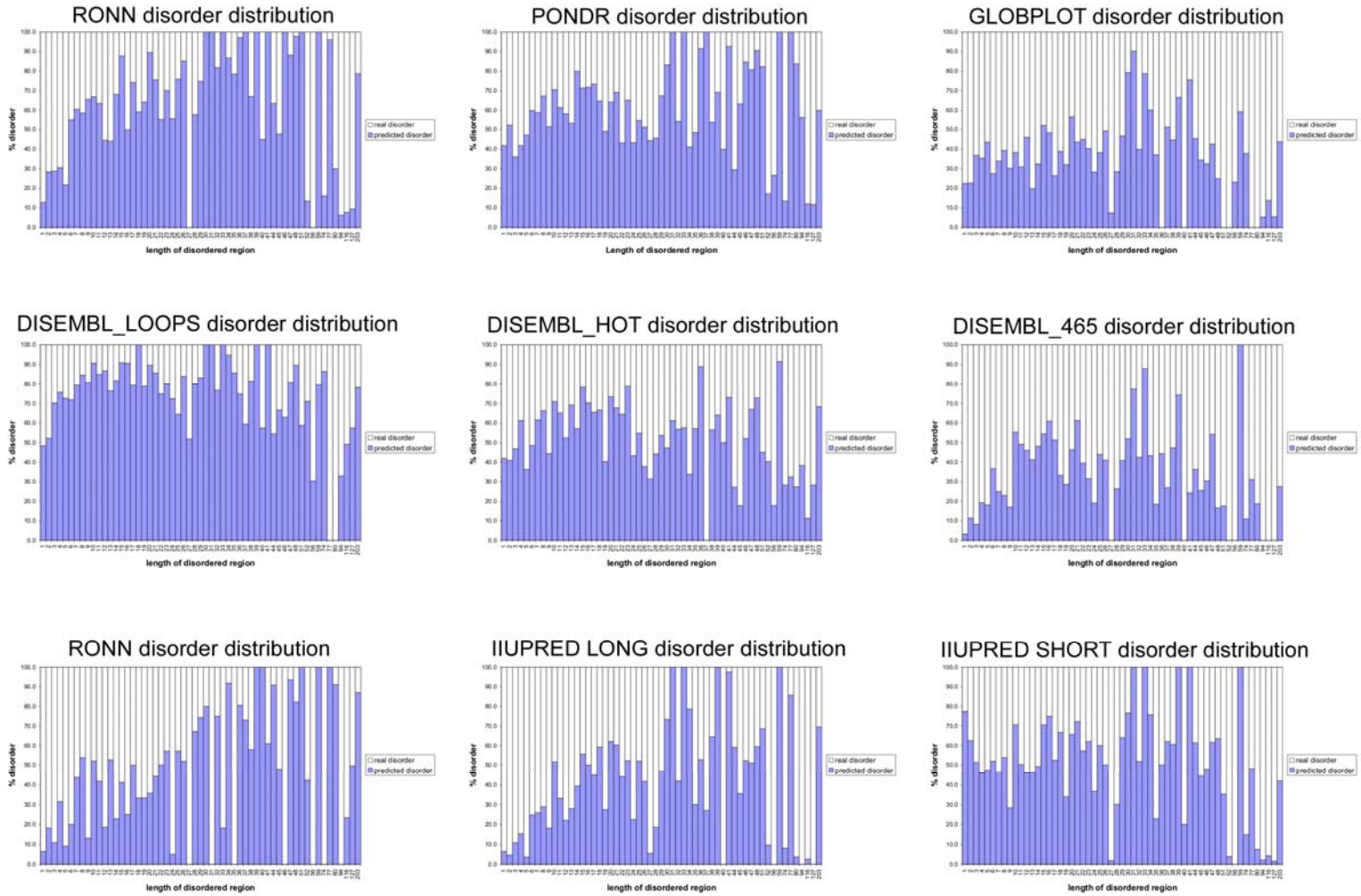


Figure 8. Graphs showing the percentage of correctly predicted disorder for disordered regions of different lengths for RONNv3.1 and other prediction tools. In each case, the x axis shows the length of disordered regions and the y axis shows % disorder. Blue bars represent correctly predicted disorder.

There are two caveats to consider when viewing these graphs; Figure 3 shows the distribution of lengths of disordered regions from this test set. There are many regions of disorder of length between 1 and 15, but the number of regions longer than this decreases rapidly, with only 1 region present for most lengths above 35 residues long. This needs to be borne in mind when viewing Figure 8 as results for shorter regions are averaged over many predictions whereas results for long regions may only be for one prediction. The second matter is that the graphs only show the number of correctly predicted disordered residues, so represent sensitivity but not specificity. In other words, if a tool predicted all residues in a protein to be disordered the corresponding graph would be entirely blue, suggesting very good performance.

Despite these issues the graphs raise some interesting points; RONNv3.1 does generally predict longer regions of disorder with higher sensitivity than for short regions. This is also seen for Foldindex and IUPred Long which is unsurprising given that these tools are designed to detect long regions of disorder.

In addition, RONNv3.1 predicts no disorder for the 2 regions of length 27 or for the region of length 56. This is also true for Disembl_465, Foldindex, IUPred Long and IUPred Short which prompted investigation into these sequences. The 2 regions of length 27 come from the proteins with PDB identifiers 2DCK and 2HCN, but unfortunately the papers associated with these structures are yet to be published. The region of length 56 comes from the protein with PDB identifier 2O1S and the disordered region is from residue 183 to residue 238. However, the paper (Xiang, Usunow et al.,

2007) notes that crystallization of the protein only occurred in the presence of fungus. This caused proteolysis of the protein and it was confirmed by SDS gels that residues 183-238 and 292-317 were cleaved off by the fungal protease so are not disordered.

Due to the fact that many of the other very long disordered regions were poorly predicted by RONNv3.1 and many of the other tools, attempts were then made to verify whether these sequences were truly disordered. All 28 regions of 35 residues long or more were investigated but only 10 proteins had associated published papers. Of these 10 proteins, in 2 cases (including the region of length 56 described above) the authors note that long regions missing in the electron density are due to fungal proteolysis rather than disorder. The region of 203 residues came from the protein 2I7X (residues 423-625) and the authors note that presence of fungus was essential for crystallization and that this whole segment was removed by fungal proteolysis. However, RONNv3.1 predicted most of this region to be disordered, as did most of the other tools. This is not necessarily incorrect as disordered regions will generally be more accessible to proteases, and this protein would not crystallize without this region being removed suggesting it may be disordered.

To summarize, regions annotated as missing from electron density may have been cleaved during crystallization, but this does not necessarily mean that they are not disordered. If a training set of disordered sequences is derived from missing coordinates from X-ray structures, as long as it is sufficiently large then the number of disordered sequences which are actually ordered will be insignificant.

5.4.3 Detection of Order-to-Disorder Transitions by Sequence Analysis

It is widely accepted that the amino acid composition of disordered sequences differs from that of ordered sequences (Tompa 2002). However, what happens to the amino acid sequences immediately surrounding order/disorder transitions has not been systematically studied. Previous work suggested there were compositional differences between boundary and non-boundary data (Radivojac, Obradovic et al., 2003) but the small size of the data set used meant the results were inconclusive. However, by systematically analyzing all known partially ordered structures the aim was to draw more reliable conclusions and determine whether the change of folding state is the result of a gradual change in amino acid composition or is triggered by a specific signal.

5.4.3.1 Propensities for amino acids to be in ordered or disordered regions

Figure 9 shows the propensities for amino acids to be in ordered or disordered regions calculated from the ORDER and DISORDER data sets used in this work (Table 4). This analysis is in line with similar analyses (Vucetic, Brown et al., 2003; Peng, Radivojac et al., 2006; Fuxreiter, Tompa et al., 2007) but the present data sets are much larger than those previously analysed. Charged, polar and flexible residues such as glutamate, lysine, glutamine and especially serine are more common in disordered sequences. In addition, low-complexity proline- and/or glycine-rich regions are unlikely to form stable structures. In contrast, the aromatic amino acids tryptophan, tyrosine and phenylalanine

are more common in ordered regions than in disordered regions. These large side chains can interact strongly, particularly with each other, and so promote and stabilize tertiary structure (Burley and Petsko 1985). Cysteine residues have the ability to form disulphide bonds so are order-promoting, as are the aliphatic amino acids leucine, isoleucine and valine.

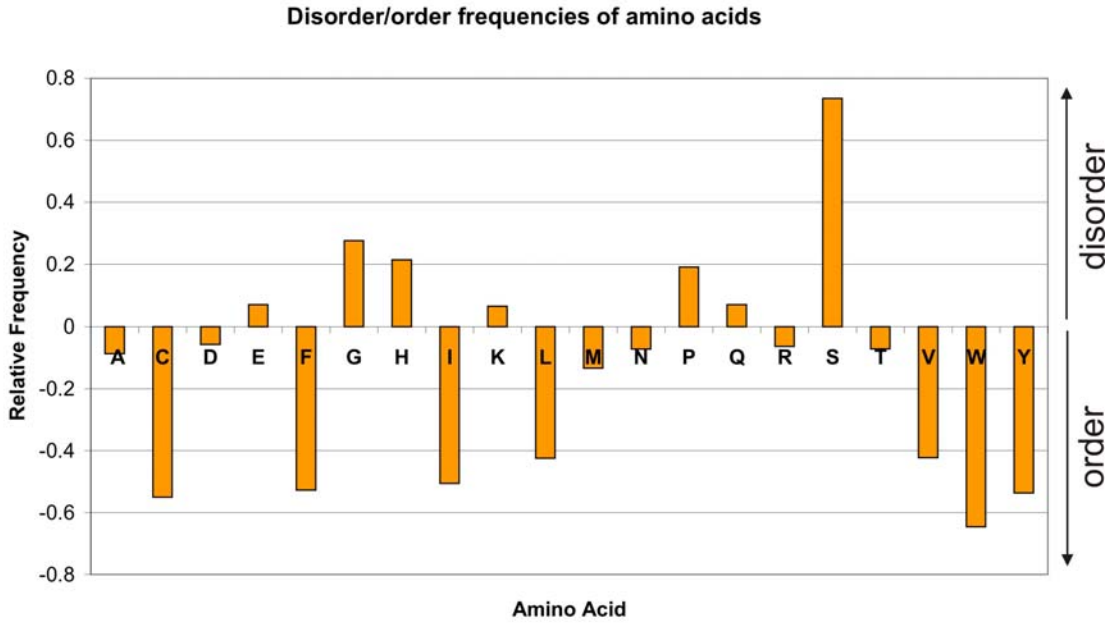


Figure 9. Relative frequencies of amino acids in the ORDER and DISORDER data sets. Each bar represents the ratio (P(D)-P(O))/P(O) where P(D) is the frequency of the amino acid in the DISORDER data set and P(O) is the frequency of the amino acid in the ORDER data set. Positive values indicate that the amino acid occurs more frequently in the DISORDER data set and negative values indicate that the amino acid occurs more frequently in the ORDER data set. The frequency calculation for histidine excludes those histidines identified as being part of artificially engineered purification tags. The frequency calculation for methionine does not include those methionines that are the first residue of a protein, which are commonly unobserved in structures.

5.4.3.2 Comparison of the data sets with full genomes

The amino acid frequencies in genomes were first compared to those in the ORDER, DISORDER and DISPROT data sets (Table 4) to determine whether there were any significant differences. The results are shown in Figure 10.

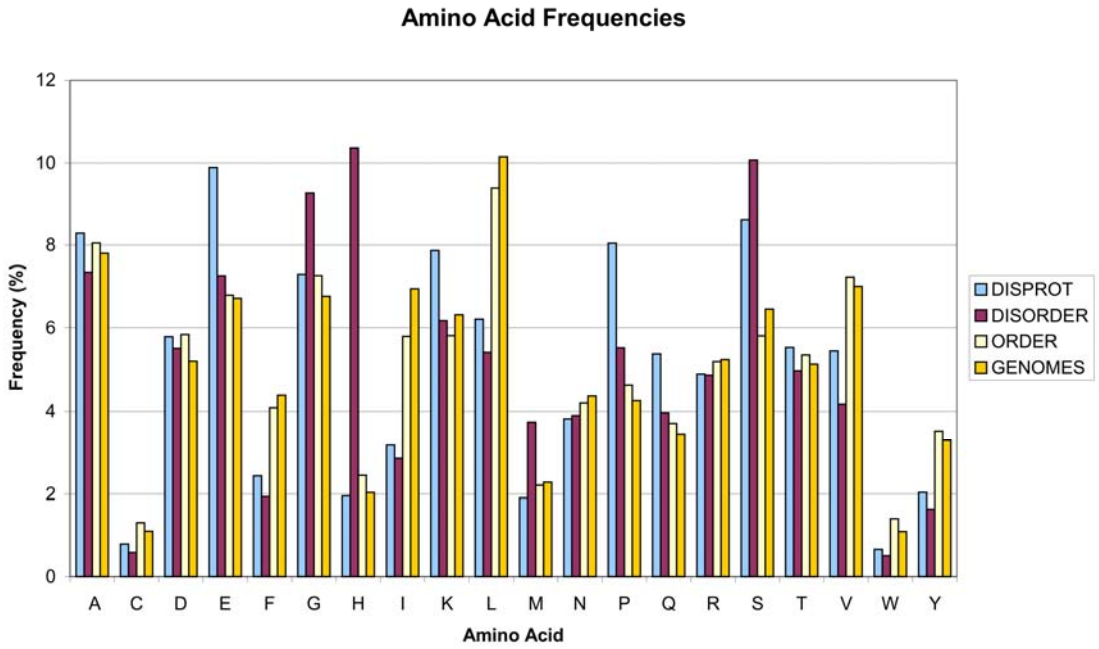


Figure 10. Comparison of amino acid frequencies from the ORDER, DISORDER and DISPROT data sets to average frequencies in genome sequences of archaea, bacteria and eukaryotes.

It can be seen that histidine is far more common in the DISORDER data set, but this is to be expected since many proteins used for structural studies are engineered to have poly-histidine tags at the N- or C-terminus to enable the protein to be purified by metal-affinity chromatography (Lilius, Persson et al., 1991). These tags are not always cleaved off

before crystallization and are usually disordered. In addition, tags may be unexpectedly cleaved by proteolysis and therefore may be erroneously included in the protein sequence but, being unobserved in the protein structure, are included in the DISORDER data. Methionine is similarly common in the DISORDER data: it is the first residue of most proteins but the first few residues of structures are often not clearly defined.

Phenylalanine, isoleucine and leucine are more common in genomes than in ordered sequences. This may be because these residues are common in structured proteins which are under-represented in the PDB entries annotated in the MSD, such as those with transmembrane helices which are difficult to express solubly and crystallize.

Unsurprisingly, the order-promoting residues, cysteine, phenylalanine, isoleucine, leucine, valine, tryptophan and tyrosine are more common in the ORDER data compared to the DISORDER or DISPROT data.

Phenylalanine, isoleucine and leucine have a higher frequency in DISPROT data than in DISORDER data suggesting they may be more tolerated in long disordered sequences than in short ones, as all sequences in the DISORDER data set come from proteins for which a partial crystal structure has been solved.

Glutamate, lysine, proline and glutamine are most frequent in DISPROT sequences, even more so than in sequences from the DISORDER data set, due to their highly disorder-promoting properties and the presence of many wholly/largely disordered proteins in

DISPROT. However, serine and glycine are more common in sequences from the DISORDER set than those from the DISPROT set. This may be due to their high frequency in flexible linker regions in otherwise structured proteins.

5.4.3.3 Analysis of order/disorder transitions

Having examined the bulk properties of the datasets, the disorder/order transitions in the MSD were then examined to look for behaviour distinct from bulk disorder or order. To establish the reliability of the transitions observed in the MSD, each entry which had more than one identical chain was analysed to determine whether the order/disorder transitions in each chain were the same. A total of 5798 files had more than one identical chain and there were a total of 51840 transitions observed in these protein sequences if disorder of any length was considered. 33004 (64%) were exactly the same in *all* chains in a given entry and 38468 (74%) were exact matches or differed by no more than one residue. Given that the ordered/disordered state of short surface loops may be differently affected by crystal packing interactions and differences in substrate/cofactor binding, as well as variation in human interpretation of marginal electron density, this analysis strongly supports the hypothesis that order/disorder transition points are precisely defined.

All the sequences included in the BOUNDARY data set are included within the ORDER and DISORDER data sets. Thus, if there were no signal one would expect that the frequency of each amino acid should be the average of the frequencies in bulk order and

bulk disorder. However, glutamate, lysine, proline, glutamine and arginine are more common at boundaries than in either bulk disorder or bulk order (Figure 11). This suggests that transition regions may have a characteristic signal which may help to define the precise transition boundary.

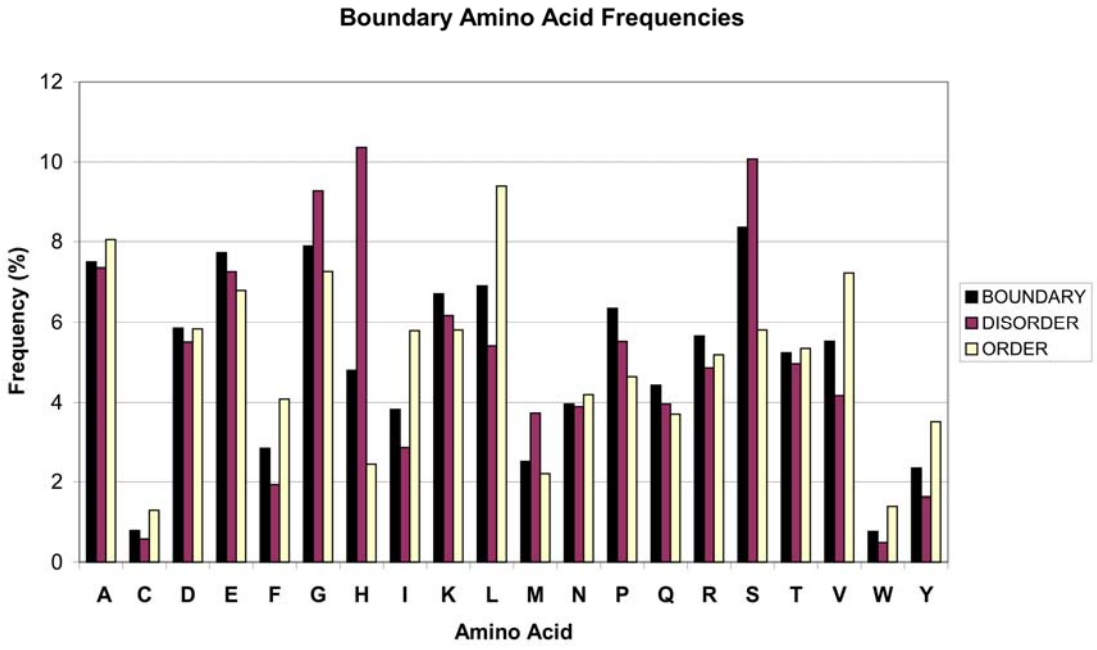


Figure 11. Comparison of amino acid frequencies for the ORDER, DISORDER and BOUNDARY data sets.

This signal may be broadly spread over the transition boundary or else may be made up of changes in amino acid frequency at more specific residue positions relative to the boundary. To investigate this, the position-specific frequencies of each residue in the window of 5 amino acids either side of the boundary in the BOUNDARY data set were

analysed. By moving to a position-specific analysis the signal-to-noise ratio is dramatically reduced but some trends remain quite clear. The order-promoting residues cysteine, isoleucine, leucine, phenylalanine, valine, tryptophan and tyrosine decrease in frequency before the disordered side of the boundary, whereas the frequency of the disorder-promoting residue serine increases on the ordered side of the boundary, as the disorder transition is approached (Figure 12). The figure shows averaged results for order-to-disorder and disorder-to-order transitions to improve the signal-to-noise ratio. These transitions were analysed separately (data not shown) and both showed the same trends. From this analysis we infer that there is little or no directionality in the signal. Furthermore, the analysis was repeated with 10 residues either side of transition boundaries (862 boundaries, data not shown) and the same trends were observed, although as there were fewer boundaries of this size, the noise was increased further. In fact, most of the changes seen in Figure 12 occur between positions -3 on the ordered side of the transition and +1 on the disordered side.

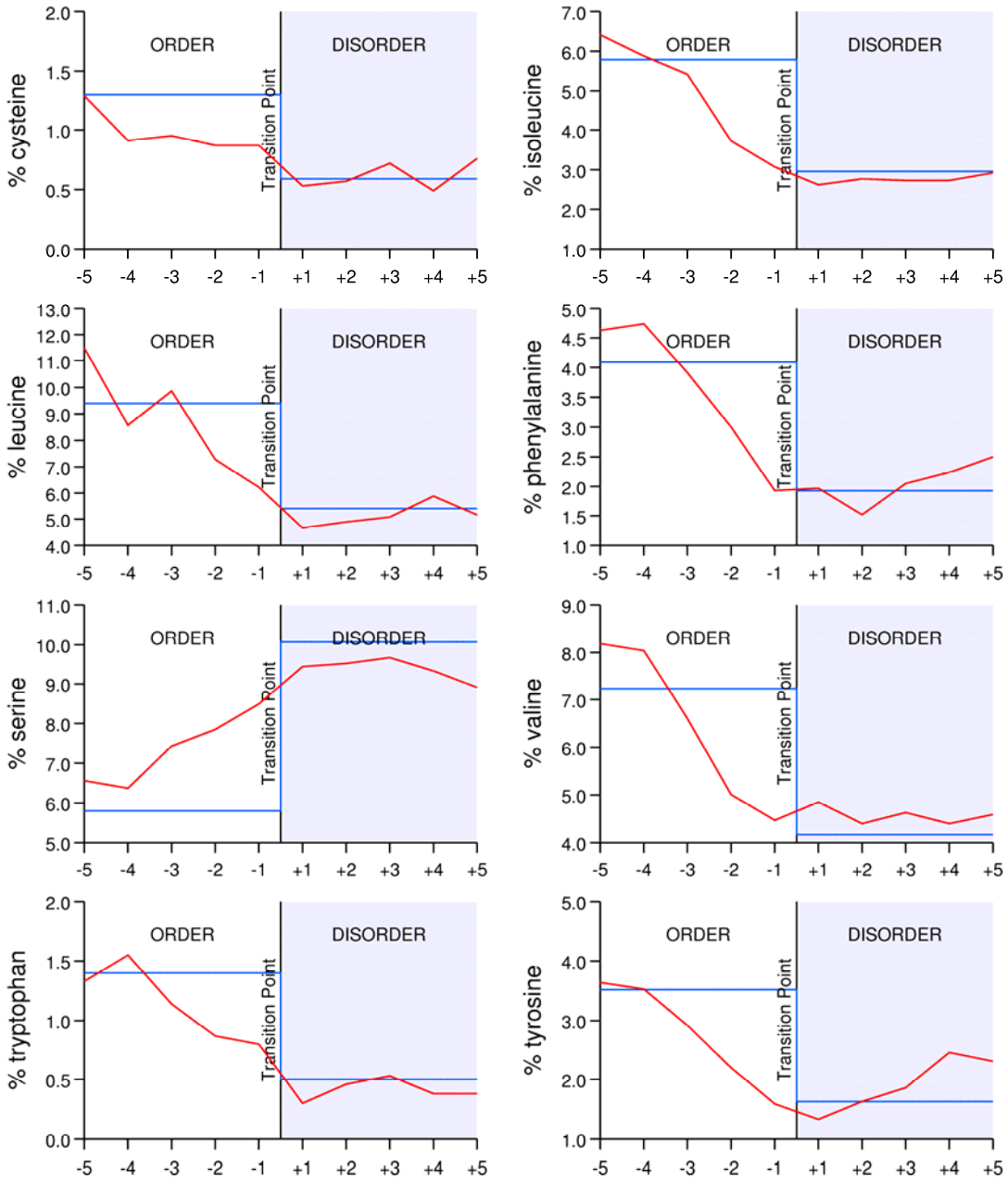


Figure 12. Position-specific amino-acid frequencies at boundaries. These residue frequencies demonstrate a gradual breakdown of the propensity for order before the transition into disorder. The frequency at each position is shown by the red line and the frequency in bulk order/disorder by the blue line.

The non-directionality of the signal is also visible in Figure 13 which shows sequence logos (Crooks, Hon et al., 2004) of boundary regions. Apart from the His tags obvious on the disordered side of disorder-to-order transitions, order-to-disorder and disorder-to-order boundaries look very similar.

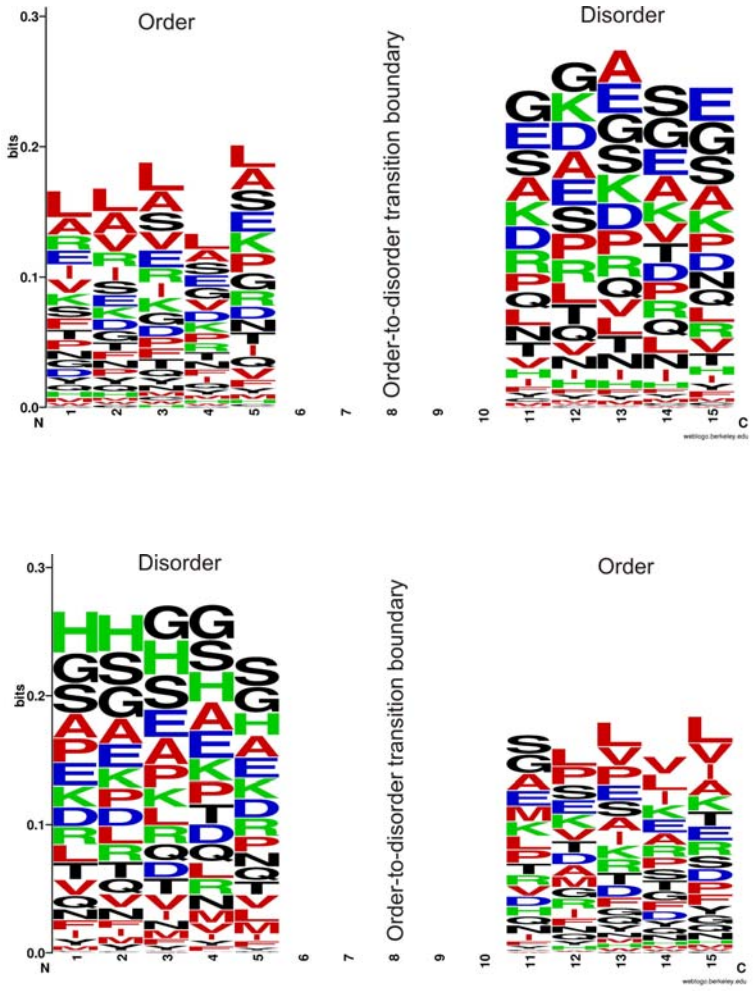


Figure 13. Sequence logos of order-to-disorder and disorder-to-order boundary regions.

5.4.3.4 Signal sequences for order/disorder transitions

The distribution of those amino acids listed above which were over-represented in boundary regions compared to either bulk order or disorder were also analysed in a position-specific manner (Figure 14).

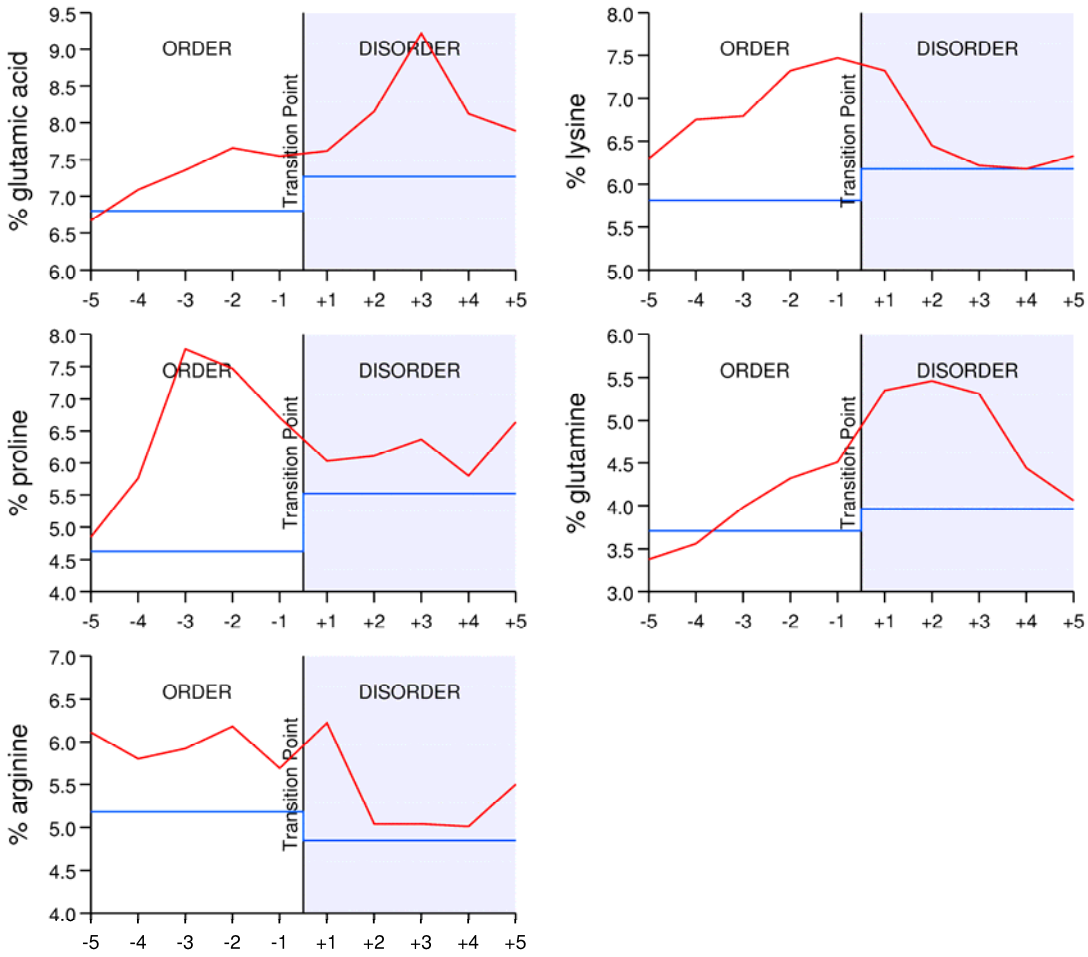


Figure 14. Position-specific amino acid frequencies at boundaries for residues more common at boundaries than in bulk order or disorder. Key: as for Figure 12.

Each amino acid (glutamate, lysine, proline, glutamine and arginine) shows a peak in frequency around certain positions of the boundary, for example at position +3 on the disordered side of the boundary for glutamate and at position -3 on the ordered side of the boundary for proline. It is tempting to speculate that these over-represented amino acids play a role in defining precise transition points between order and disorder. On the ordered side of a transition for example, proline is well known as a breaker of regular secondary structure. The other four amino acids all have long charged/polar side chains which may again play a role in disturbing secondary structure by satisfying hydrogen bonding interactions with nearby main chains.

The occurrence of weak signals superposed on general trends may indicate characteristic signal sequences spanning boundaries between order and disorder which could be exploited by prediction algorithms. In an attempt to identify these sequences, all the 10 residue boundary sequences were fed into a clustering algorithm (TribeMCL (Enright, Van Dongen et al., 2002)). Although some clusters were trivially identified (e.g. poly-histidine, poly-glycine and poly-glutamate on the disorder side of transitions) no useful clustering could be performed which could lead to a useful algorithm. This may suggest (i) a lack of clear multiple-amino acid signal sequences indicating transitions, (ii) limitations of the clustering methodology used, or (iii) simply that our present body of knowledge does not allow us to identify such sequences with any level of accuracy.

5.5 Conclusions

The demand for accurate prediction of natively disordered regions in proteins is being driven by structural genomics initiatives worldwide in an effort to increase success rate, particularly for studies on eukaryotic proteins. Incorporating accurate disorder prediction into the construct design process can reduce the amount of time and resources devoted to non-crystallizable or otherwise badly behaved proteins, allowing researchers in the wet lab to tackle the real technological challenges posed by high-throughput studies. In the current state of the art, no single disorder prediction technique is so accurate that it can be entirely trusted and a strategy of using several well performing methods and looking for common prediction features may be the best way to attempt the reliable identification of regions of disorder in unknown protein sequences.

The RONN algorithm presented here is an application of a revised bio-basis function neural network technique to the prediction of natively disordered regions. The blind tests show that, based on probability excess, RONNv3 was ranked as the best algorithm for disorder prediction with DISO-PRED2 and PONDR® also performing well (Table 6; Figure 6). RONNv3.1 improves on the performance of RONNv3, but a new tool, VSL2, currently gives better results still (Table 10).

RONN still has weaknesses particularly in the detection of short regions of disorder and in defining the first and last residues of disordered regions (Figure 5). Using a shorter window improves the sensitivity, but also increases the noisiness of the prediction and the

overall performance is degraded (Figure 4). A more complex scheme might involve the use of two (or more) window sizes, effectively making multiple predictions and then combining them to return a single overall prediction for the sequence. While such algorithms are necessarily more complex, they could also potentially cope with any dependence of the amino acid composition of disordered regions on their lengths.

This work has also revealed that the position of order/disorder transitions is quite precisely defined in amino-acid sequences. Furthermore, these transition regions do indeed have an amino acid composition distinct from those of bulk ordered and disordered sequences. Across the boundary region there is a steady decrease in order-promoting residues on the ordered side of the boundary signifying the approaching disorder. Superimposed on this is a weak signal of an enrichment of glutamate, lysine, proline, glutamine and arginine residues, apparently at preferred positions relative to the actual boundary. This signal is effectively the same for order-to-disorder and disorder-to-order transitions. The theory that protein folding extends outwards from folding nuclei of secondary structure (Fersht 2000; Dobson 2003) fits well with the presence of such a signal, as disordered regions which are functionally important must be prevented from becoming ordered. Due to the low signal-to-noise ratio of the data, this signature alone is unlikely to be useful for prediction of disorder start and end points, but if incorporated into existing disorder prediction tools it may improve their utility by suggesting precise transition points within an overall tendency to order/disorder.

Chapter 6

Overall Conclusions

6 Overall Conclusions

The structures of the 1F2 TCR, the 2D10 TCR and the 2D10TCR-ELA-A2 complex were solved. Comparison of the structure of the 1F2 TCR to that of the 1G4 TCR, which also binds to the NYESO-A2 pMHC, suggested that these two TCRs bind to the same pMHC in very different ways. However, the structure of the 1F2TCR-NYESO-A2 complex is needed to confirm this. The structures of the 2D10 TCR and 2D10TCR-ELA-A2 complex showed the mechanism of binding of one tumour antigen-specific TCR to its cognate pMHC, and suggested possible mutations that could be introduced into the 2D10 TCR to improve binding which may aid the development of more effective cancer vaccines. The complex structure also added to the overall database of Class I complex structures. Analysis of the binding mechanisms of all the TCR-pMHC complexes in this database showed that TCRs can bind to pMHC with a whole range of orientations, yet always focus on the centre of the peptide and use a specific subset of residues on the MHC helices to bind. The use of aromatic amino acids to bind to one particular MHC residue, Q155, was shown to be a common feature and one which mirrors the use of aromatics by antibody CDR loops.

Attempts were made to express and purify TLRs, with the aim of solving more of the structures of these proteins which form a crucial link between the innate and adaptive immune systems. However, insufficient soluble protein expression was obtained for crystallization trials. No other group has solved a TLR structure since the first was

published in 2005, which suggests we have more to learn about these proteins and their possible chaperones before any more progress can be made.

Finally, a protein disorder prediction tool was developed to aid construct design and improve the chances of obtaining protein crystals. The RONN tool is based on a novel type of neural network, the bio-basis function neural network, and is now one of the best in the field. Work to detect precise order-to-disorder transition points showed that transition regions have a distinct amino acid composition, with a weak signal on the ordered side of the boundary signifying the approaching disorder. Although the signal-to-noise ratio of this data is low, incorporation of it into existing disorder prediction tools may improve their success.

Appendix A

X-ray Crystallography

Appendix A: X-ray Crystallography

An introduction to crystallography can be found in books by David Blow (Blow 2003) and Gale Rhodes (Rhodes 2000).

A.1 Protein Crystallization

For X-ray crystallography, very pure protein samples are needed to set up crystallization trials. The aim is to produce highly ordered crystals which contain many protein molecules in identical orientations, held together by non-covalent interactions. Each molecule will then diffract X-rays in the same way and the diffraction from all the molecules will be combined to produce a strong, detectable diffraction pattern.

Crystallization of protein samples is done by slow, controlled precipitation of the protein from aqueous solution under non-denaturing conditions. Precipitation can be aided by many substances such as salts and polyethylene glycol (PEG). Other important factors affecting precipitation are protein concentration, pH, temperature and ionic strength. Finding the right combination of conditions that enables a protein to crystallize is one of the major bottlenecks in solving protein structures.

The method used for the experiments in this thesis was vapour diffusion using the sitting drop method, where the protein/precipitant solution and a reservoir of precipitant solution are allowed to equilibrate in a closed container.

A.2 Collecting X-ray Data

To collect X-ray diffraction data, a protein crystal is mounted between an X-ray source and a detector. The X-ray beam is diffracted by the crystal onto the detector to produce a diffraction pattern consisting of spots, or reflections. The positions and intensities of these reflections can be measured and this information used to calculate a map of electron density for the protein. However, this is not completely straightforward due to the phase problem, which will be described shortly.

The crystal can be divided into unit cells, which are the smallest volume element that can be used to make up the whole crystal by translation operations only. The crystal is therefore made up of many unit cells stacked together in a three-dimensional array. The spacing of the reflections in the diffraction pattern (the reciprocal lattice) is inversely related to the spacing of the unit cells in the crystalline lattice (the real lattice).

The spots on a diffraction pattern will be in the same positions, *i.e.* the reciprocal lattice will be identical, for all crystals with the same unit cell dimensions and spacegroup exposed to X-rays of the same wavelength. It is the intensities of the spots which give the information about the contents of the unit cell.

A.3 Lattices and Spacegroups

An ideal crystal has lattice symmetry and a unit cell is defined by three lattice translations. The edges of the unit cell correspond to the three lattice translations. The entire lattice is composed of unit cells stacked in three dimensions and in an ideal crystal each unit cell has identical contents.

There are several lattice types. The symbol P describes a primitive lattice which has one lattice point at each corner of the unit cell, called primitive lattice points. All lattice types have these primitive lattice points. The symbol I defines a body-centred (internal) lattice which has an extra lattice point in the centre of the cell. A face-centred lattice, F, has additional points on the centre of each face of the cell and the symbol C means the lattice has an extra point at the centre of one face only.

The only rotation symmetries a lattice can have are 2-fold, 3-fold, 4-fold and 6-fold. There are 7 crystal systems and 14 Bravais lattices which are derived by combining the crystal systems with the lattice types (Table 1).

Table 1. The crystal systems and Bravais lattices. Adapted from the International Tables for Crystallography (1987).

Crystal system	Minimum Symmetry Requirement	Conditions	Bravais Lattice
Triclinic	None	None	P
Monoclinic	1 2-fold axis	$\alpha = \gamma = 90^\circ$ (where b is the unique axis) $\alpha = \beta = 90^\circ$ (where c is the unique axis)	P, C
Orthorhombic	3 perpendicular 2-fold axes	$\alpha = \beta = \gamma = 90^\circ$	P, C, I, F
Tetragonal	1 4-fold axis	$a = b$, $\alpha = \beta = \gamma = 90^\circ$	P, I
Trigonal	1 3-fold axis	$a = b$, $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$ (hexagonal axes) or $a = b = c$, $\alpha = \beta = \gamma$ (rhombohedral axes)	P
Hexagonal	1 6-fold axis	$a = b$, $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$	P
Cubic	4 3-fold axes	$a = b = c$, $\alpha = \beta = \gamma = 90^\circ$	P, I, F

The symmetry of a unit cell is described by its spacegroup, of which there are 230 different possibilities. The spacegroup represents the lattice type and the symmetry operations which can be carried out on the unit cell without changing its appearance.

Symmetry within unit cells is restricted to rotations, translations and screw axes, which are rotations and translations combined. Mirror planes are not found in unit cells of proteins as there is only one enantiomorph of each amino acid in nature.

The smallest unit of the crystal which can generate the whole crystal structure by means of translations and rotations is called the asymmetric unit. This needs no symmetry of its own. In the simplest case for proteins, the asymmetric unit is one protein molecule.

A.4 The Phase Problem

The electron density of a protein can be thought of as a three-dimensional wave repeated in each unit cell, and can therefore be described by a Fourier series as follows:

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l} F_{hkl}\, e^{-2\pi i(hx+ky+lz)}$$

where V is the volume of the unit cell. The indices, *h*, *k* and *l*, show the position of a reflection in the reciprocal space of the diffraction pattern and give the frequencies of each Fourier term in the *x*, *y* and *z* directions respectively.

Each F_{hkl} is a structure factor, itself a Fourier series, describing a single reflection in the diffraction pattern and therefore possessing amplitude, frequency and phase. As each F_{hkl} describes a diffracted X-ray, its frequency is that of the X-ray source. The amplitude is proportional to the square root of the reflection intensity, I_{hkl}. However, the phase, α_{hkl}, is not directly measurable.

Structure factors can be represented as complex vectors, where the length of the vector represents the amplitude of the structure factor. The phase is represented by the angle that the vector makes with the positive real-number axis when the origin of the vector is placed at the point 0+i0. Because the diffractive contributions of atoms (or volume elements) to a single reflection are additive, each contribution can be represented as a

complex vector and the resulting structure factor is the vector sum of all these contributions.

In this way, F_{hkl} can be decomposed into its amplitude, $|\text{Fhkl}|$, and its phase, $\alpha_{hkl:}$

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}\left|F_{hkl}\right|e^{-2\pi i(hx+ky+lz-\alpha_{hkl})}$$

This gives the electron density as a function of the known amplitudes and the unknown phases. Determining these phases is one of the biggest problems in X-ray crystallography and is known as the phase problem.

A.5 Molecular Replacement

The phases from structure factors of a known protein structure can be used to provide initial estimates of the phases for a protein of unknown structure if the known and unknown structures are homologous. This involves finding the position and orientation of the model (known) protein in the unit cell of the new (unknown) protein that will give phases most similar to those of the new protein. Structure factors of the model can then be calculated and these phases used as initial estimates for the phases of the new protein.

Theoretically, it would be possible to search all orientations and positions of the model structure in the unit cell of the new protein, calculate structure factors for each position of

the model and compare the amplitudes with the measured amplitudes obtained from the diffraction intensities of the new protein. The phases for the orientation and position which gives the best match would be calculated and used as starting phases for structure determination of the new protein.

However, the number of orientations and positions to be tested is enormous and this method is often impractical even on the fastest computers. Classically, to solve this problem, the search for the orientation of the model is separated from the search for the position of the model, turning a six-dimensional search into two three-dimensional ones.

A Patterson function is a Fourier series without phases, so can be calculated from the diffraction intensity measurements of the new protein. A Patterson map displays peaks at locations corresponding to vectors between atoms in the unit cell. Intramolecular vectors depend only on the orientation of the molecule, not on its position in the cell so can be used to search for the rotation of the molecule independently of its position in the unit cell (its translation). Intermolecular vectors depend on both orientation and position so can be used to obtain the translation of the molecule in the unit cell.

To find the correct rotation of the model in the unit cell, a Patterson map is calculated for the model protein and rotated over the Patterson map of the new protein, keeping the origins at the same point. For each rotation, the agreement between the two maps is evaluated using a rotation function and the best fit will hopefully give the correct

orientation of the model protein in the unit cell. This involves 'enriching' for intramolecular vectors by using a maximum vector length cutoff.

To obtain the correct translation, the intermolecular vectors in the Patterson map are compared, with the aim of determining the vector between the origin assumed in the model and the position of the actual origin in the new protein structure. This involves enriching for intermolecular vectors by using a minimum vector length cutoff.

The difference between the structure factor amplitudes calculated from the correctly rotated and translated model and the observed structure factor amplitudes can now be used to calculate a difference map. Features of the new structure which are not present in the model structure appear as positive density and features of the model structure that do not appear in the new structure appear as negative density. The model can therefore be corrected to take account of these differences and new, hopefully better phases can then be calculated in a process of iterative refinement.

A.6 Statistics to test the model

The model can be assessed by comparing the measured structure factor amplitudes, $|F_{obs}|$, with the amplitudes calculated from the model, $|F_{calc}|$. The most common way of doing this is using the residual index, or R factor:

$$R = \frac{\sum \| F_{obs} | - | F_{calc} \|}{\sum | F_{obs} |}$$

If the calculated and observed structure factor amplitudes agreed perfectly, R would be 0, whereas if a set of measure amplitudes was compared to a set of random amplitudes R would be about 0.6.

The free R factor, R_{free}, is calculated with a small set of randomly-chosen intensities (*e.g.* 5% of the total number) which are not used for refining at any stage. They are only used to assess the agreement between calculated and observed data by assessing how well the current model predicts this subset of measured intensities. This is different to the R factor, which measures how well the current model predicts the entire data set that produced the model.

Appendix B

Calculation of TCR-pMHC Measurements

Appendix B: Calculation of TCR-pMHC Measurements

B.1 To Calculate the 2-fold Axis of the TCR

After superimposing the variable domains of the TCR, the result is a rotation matrix which can be used to work out the vector through the 2-fold axis of the TCR.

Robert Esnouf provided a rotation matrix 'conversion table' to calculate a rotation of γ around the axis defined by the unit direction vector (a,b,c):

$\cos\gamma + a^2(1-\cos\gamma)$	$-c\sin\gamma + ab(1-\cos\gamma)$	$b\sin\gamma + ac(1-\cos\gamma)$
$c\sin\gamma + ab(1-\cos\gamma)$	$\cos\gamma + b^2(1-\cos\gamma)$	$-a\sin\gamma + bc(1-\cos\gamma)$
$-b\sin\gamma + ac(1-\cos\gamma)$	$a\sin\gamma + bc(1-\cos\gamma)$	$\cos\gamma + c^2(1-\cos\gamma)$

The trace of the rotation matrix can be calculated by adding the diagonals (top left to bottom right):

$$\begin{aligned}\text{Trace} &= \cos\gamma + a^2(1-\cos\gamma) + \cos\gamma + b^2(1-\cos\gamma) + \cos\gamma + c^2(1-\cos\gamma)\\ &= 3\cos\gamma + (a^2 + b^2 + c^2)(1-\cos\gamma)\\ &= 3\cos\gamma + 1 - \cos\gamma\\ &= 2\cos\gamma + 1\end{aligned}$$

This allows the direction cosines to be calculated. The off-diagonals of the matrix can then be used to check whether the values of a, b and c are positive or negative.

B.3 To Calculate the Twist of the TCR in Relation the the pMHC

The twist is calculated as a dihedral angle along three bonds defined by four atoms:

1. The midpoint of the disulphide bond in TCRα
2. The midpoint of the line drawn between the midpoint of disulphide bond in TCRα and the midpoint of the disulphide bond in TCRβ
3. The midpoint of the line drawn between the N and C termini of the peptide
4. The N terminus of the peptide

If VT1, VT2 and VT3 are the bond vectors, then

VN1 = VT1 x VT2

VN2 = VT2 x VT3

To obtain the twist angle, make VN1 and VN2 unit length and dot them to get the cosine of the angle. Give this the sign of the scalar triple product (STP):

STP = VT1 . VT2 x VT3.

B.4 To Calculate the Intersect of the Normal to the β-Sheet Plane with the Interface Plane

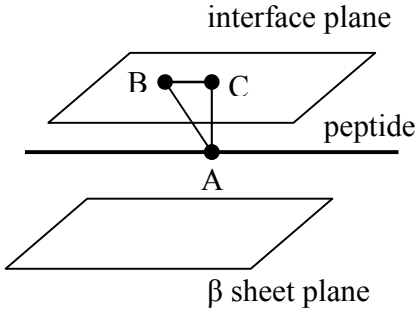
This is used for calculating the TCR height and the height of the interface from the MHC.

A: midpoint of peptide

B: centre of gravity of interface (average coordinate of the contact points)

C: intersect of the normal to the β-sheet plane with the interface plane

$\hat{n}$: normal to the β-sheet plane



$$\left|\overrightarrow{AC}\right| = \overrightarrow{AB}.\hat{n}$$

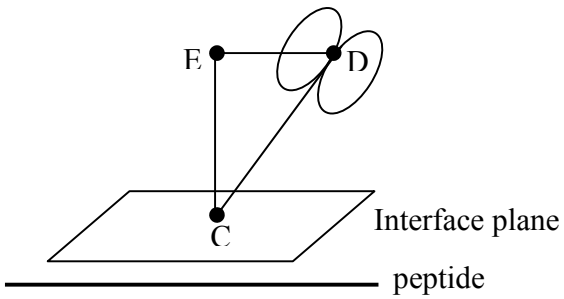
$$\overrightarrow{OC} = \overrightarrow{OA} + (\overrightarrow{AB}.\hat{n})\hat{n}$$

$\left|\overrightarrow{AC}\right|$ is the height of the interface from the pMHC

B.5 To Calculate the Height of the TCR from the Interface.

There are two ways of doing this:

1. Look at the distance between the intersect of the normal to the β sheet and interface plane (C) on the interface plane and the midpoint of the line between the 2 disulphide bonds of the TCR (D) – this takes into account the tilt of the TCR
2. Look at the perpendicular distance of the midpoint of the line between the 2 disulphide bonds (E) from the intersect of the normal to the β sheet and interface plane (C).



For option 1 first calculate the vector $\overrightarrow{CD}$

The length of this vector is the diagonal TCR height.

$\left|\overrightarrow{CD}\right|$ = sqrt ($x^2 + y^2 + z^2$)

For option 2 let E be the point perpendicular to C at the height of D:

$\left|\overrightarrow{CE}\right| = \overrightarrow{CD} . \hat{n}$

B.6 To Calculate the Position of the Peptide N-Terminus When Projected Onto the Interface Plane

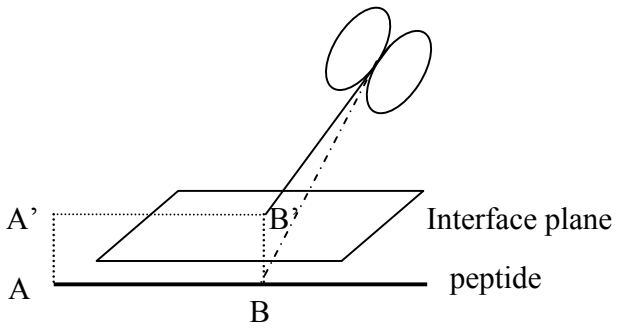
A: N-terminus of the peptide

A': N-terminus of the peptide projected onto the interface plane

B: midpoint of the peptide

B': intersect of the normal to the β sheet and interface plane.

$\hat{n}$: normal to the β-sheet plane



$$\overrightarrow{AA'} = (\overrightarrow{AB'} . \hat{n})\,\hat{n}$$

$$\overrightarrow{OA'} = \overrightarrow{AA'} + \overrightarrow{OA}$$

The position of the C-terminus of the peptide projected onto the interface plane is calculated in the same way. Both are used for calculating the direction of TCR shift (see below)

B.7 To Calculate the Shift of the TCR

B: the midpoint of the interface plane.

C: the point that the 2-fold TCR vector hits the interface plane

D: the midpoint of the line between the disulphide bonds of the TCR.

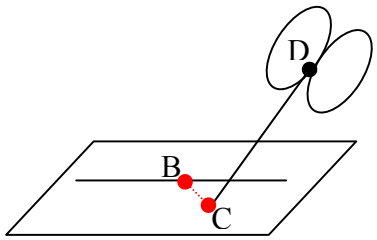
$\hat{v}$: the unit vector between C and D

$\hat{n}$: normal to the β-sheet plane

$\overrightarrow{OD} = \underline{d}$, $\overrightarrow{OB} = \underline{b}$ etc.

$\overrightarrow{BC} . \hat{n} = 0$ as they are perpendicular.

$(\underline{c}\text{-}\underline{b}) . \hat{n} = 0$



To calculate $\hat{v}$, find the length of $\overrightarrow{OD}$ and divide each coordinate of $\overrightarrow{OD}$ by this length.

Equation 1: $\underline{c} = \underline{d} + m\hat{v}$ where m is just a constant and can be negative.

so $(\underline{d} + m\hat{v} - \underline{b}) . \hat{n} = 0$

Expanding this gives

$\underline{d} . \hat{n} + m\hat{v} . \hat{n} - \underline{b} . \hat{n} = 0$

$m = (\underline{b} . \hat{n}) - (\underline{d} . \hat{n}) / \hat{v} . \hat{n}$

$m = (\underline{b} - \underline{d}) . \hat{n} / \hat{v} . \hat{n}$

Put m back into equation 1 to get point C.

Calculate the length of the shift by finding the length of the vector BC.

B.8 To Calculate the Direction of the Shift

Y: N-terminus of the peptide projected onto the interface plane

Z: C-terminus of the peptide projected onto the interface plane

YZ is then in the x direction of the interface plane. The x direction of the shift is therefore $\overrightarrow{BC} . \overline{YZ}$ where $\overline{YZ}$ is the unit vector.

$\cos\Theta = \overrightarrow{BC} . \overline{YZ} / |\overrightarrow{BC}|$

$\sin\Theta = |\overrightarrow{BC} \times \overline{YZ}| / |\overrightarrow{BC}|$

$\text{atan2}(\sin\Theta, \cos\Theta) = \Theta$

Convert to degrees by multiplying by 180/Pi

B.9 Programs to Calculate the TCR-pMHC Measurements

The reference structure was chosen to be 1OGA as this has a high resolution of 1.4Å. To calculate the normal to the β-Sheet plane of 1OGA the fortran program called planes (written by Robert Esnouf) can be used. To compile this program type:

f77 –o planes planes.f gw14.a

Use a PDB file containing ONLY chain A of 1OGA. Only residues in the central 6 β-strands of the β-sheet floor were used to calculate the plane, ignoring the far left and far right strands i.e. for 1OGA the residues used were:

3-13,21-29,31-38,93-104,108-119,121-127

Every other residue was input into the fortran planes program in the format:

CA 20 PRO

where CA is the atom name, 20 is the residue number and PRO the residue name.

When the last residue has been entered type END. The result is the direction cosines of the normal to the plane of best fit for the β-sheet floor of the MHC ($\hat{n}$). For 1OGA:

$\hat{n}$ = (0.73009, 0.34412, 0.59038) RMSE=1.721

Steps to calculate the TCR-pMHC measurements:

1. Superimpose the TCR-MHC complex structure onto that of 1OGA via residues 1-180 of the MHC groove only, for example using SHP (Stuart, Levine et al., 1979).
2. Make PDB files of the variable domains of the superimposed TCR and use SHP to superimpose the alpha variable domain onto the beta variable domain.
3. Check the direction of the TCR 2-fold axis points from the TCR towards the MHC by making sure all the values are positive. If all the values are negative simply change them all to be positive to reverse the vector direction.
4. Run calc_contacts_and_2fold.pl and open up the results.txt file.
5. Run the fortran program calc_TCR_MHC_measurements (from wherever the output file is to be located). This needs some of the results from results.txt as input. It calculates:
    - twist of the TCR
    - projected twist of the TCR
    - tilt of the TCR
    - height of the TCR
    - height of the pMHC
    - shift of the TCR
    - angle of the shift of the TCR

Appendix C

Protein Sequences of TLRs and Associated Proteins

Appendix C. Protein Sequences of TLRs and Associated Proteins

Signal sequences are highlighted in green, transmembrane regions are in red and intracellular domains are in grey.

TLR1 (Human)

```
   1  MTSIFHFAII FMLILQIRIQ LSEESEFLVD RSKNGLIHVP KDLSQKTTIL
  51  NISQNYISEL WTSDILSLSK LRILIISHNR IQYLDISVFK FNQELEYLDL
 101  SHNKLVKISC HPTVNLKHLD LSFNAFDALP ICKEFGNMSQ LKFLGLSTTH
 151  LEKSSVLPIA HLNISKVLLV LGETYGEKED PEGLQDFNTE SLHIVFPTNK
 201  EFHFILDVSV KTVANLELSN IKCVLEDNKC SYFLSILAKL QTNPKLSSLT
 251  LNNIETTWNS FIRILQLVWH TTVWYFSISN VKLQGQLDFR DFDYSGTSLK
 301  ALSIHQVVSD VFGFPQSYIY EIFSNMNIKN FTVSGTRMVH MLCPSKISPF
 351  LNLDFSNNLL TDTVFENCGH LTELETLILQ MNQLKELSKI AEMTTQMKSL
 401  QQLDISQNSV SYDEKKGDCS WTKSLLSLNM SSNILTDTIF RCLPPRIKVL
 451  DLHSNKIKSI PKQVVKLEAL QELNVAFNSL TDLPGCGSFS SLSVLIIDHN
 501  SVSHPSADFF QSCQKMRSIK AGDNPFQCTC ELGEFVKNID QVSSEVLEGW
 551  PDSYKCDYPE SYRGTLLKDF HMSELSCNIT LLIVTIVATM LVLAVTVTSL
 601  CIYLDLPWYL RMVCQWTQTR RRARNIPLEE LQRNLQFHAF ISYSGHDSFW
 651  VKNELLPNLE KEGMQICLHE RNFVPGKSIV ENIITCIEKS YKSIFVLSPN
 701  FVQSEWCHYE LYFAHHNLFH EGSNSLILIL LEPIPQYSIP SSYHKLKSLM
 751  ARRTYLEWPK EKSKRGLFWA NLRAAINIKL TEQAKK
```

TLR2 (Human)

```
   1  MPHTLWMVWV LGVIISLSKE ESSNQASLSC DRNGICKGSS GSLNSIPSGL
  51  TEAVKSLDLS NNRITYISNS DLQRCVNLQA LVLTSNGINT IEEDSFSSLG
 101  SLEHLDLSYN YLSNLSSSWF KPLSSLTFLN LLGNPYKTLG ETSLFSHLTK
 151  LQILRVGNMD TFTKIQRKDF AGLTFLEELE IDASDLQSYE PKSLKSIQNV
 201  SHLILHMKQH ILLLEIFVDV TSSVECLELR DTDLDTFHFS ELSTGETNSL
 251  IKKFTFRNVK ITDESLFQVM KLLNQISGLL ELEFDDCTLN GVGNFRASDN
 301  DRVIDPGKVE TLTIRRLHIP RFYLFYDLST LYSLTERVKR ITVENSKVFL
 351  VPCLLSQHLK SLEYLDLSEN LMVEEYLKNS ACEDAWPSLQ TLILRQNHLA
 401  SLEKTGETLL TLKNLTNIDI SKNSFHSMPE TCQWPEKMKY LNLSSTRIHS
 451  VTGCIPKTLE ILDVSNNNLN LFSLNLPQLK ELYISRNKLM TLPDASLLPM
 501  LLVLKISRNA ITTFSKEQLD SFHTLKTLEA GGNNFICSCE FLSFTQEQQA
 551  LAKVLIDWPA NYLCDSPSHV RGQQVQDVRL SVSECHRTAL VSGMCCALFL
 601  LILLTGVLCH RFHGLWYMKM MWAWLQAKRK PRKAPSRNIC YDAFVSYSER
 651  DAYWVENLMV QELENFNPPF KLCLHKRDFI PGKWIIDNII DSIEKSHKTV
 701  FVLSENFVKS EWCKYELDFS HFRLFDENND AAILILLEPI EKKAIPQRFC
 751  KLRKIMNTKT YLEWPMDEAQ REGFWVNLRA AIKS
```

TLR3 (Human)

```
  1  MRQTLPCIYF WGGLLPFGML CASSTTKCTV SHEVADCSHL KLTQVPDDLP
 51  TNITVLNLTH NQLRRLPAAN FTRYSQLTSL DVGFNTISKL EPELCQKLPM
101  LKVLNLQHNE LSQLSDKTFA FCTNLTELHL MSNSIQKIKN NPFVKQKNLI
151  TLDLSHNGLS STKLGTQVQL ENLQELLLSN NKIQALKSEE LDIFANSSLK
201  KLELSSNQIK EFSPGCFHAI GRLFGLFLNN VQLGPSLTEK LCLELANTSI
251  RNLSLSNSQL STTSNTTFLG LKWTNLTMLD LSYNNLNVVG NDSFAWLPQL
301  EYFFLEYNNI QHLFSHSLHG LFNVRYLNLK RSFTKQSISL ASLPKIDDFS
351  FQWLKCLEHL NMEDNDIPGI KSNMFTGLIN LKYLSLSNSF TSLRTLTNET
401  FVSLAHSPLH ILNLTKNKIS KIESDAFSWL GHLEVLDLGL NEIGQELTGQ
451  EWRGLENIFE IYLSYNKYLQ LTRNSFALVP SLQRLMLRRV ALKNVDSSPS
501  PFQPLRNLTI LDLSNNNIAN INDDMLEGLE KLEILDLQHN NLARLWKHAN
551  PGGPIYFLKG LSHLHILNLE SNGFNEIPVE VFKDLFELKI IDLGLNNLNT
601  LPASVFNNQV SLKSLNLQKN LITSVEKKVF GPAFRNLTEL DMRFNPFDCT
651  CESIAWFVNW INETHTNIPE LSSHYLCNTP PHYHGFPVRL FDTSSCKDSA
701  PFELFFMINT SILLIFIFIV LLIHFEGWRI SFYWNVSVHR VLGFKEIDRQ
751  TEQFEYAAYI IHAYKDKDWV WEHFSSMEKE DQSLKFCLEE RDFEAGVFEL
801  EAIVNSIKRS RKIIFVITHH LLKDPLCKRF KVHHAVQQAI EQNLDSIILV
851  FLEEIPDYKL NHALCLRRGM FKSHCILNWP VQKERIGAFR HKLQVALGSK
901  NSVH
```

TLR4 (Human)

```
  1  MMSASRLAGT LIPAMAFLSC VRPESWEPCV EVVPNITYQC MELNFYKIPD
 51  NLPFSTKNLD LSFNPLRHLG SYSFFSFPEL QVLDLSRCEI QTIEDGAYQS
101  LSHLSTLILT GNPIQSLALG AFSGLSSLQK LVAVETNLAS LENFPIGHLK
151  TLKELNVAHN LIQSFKLPEY FSNLTNLEHL DLSSNKIQSI YCTDLRVLHQ
201  MPLLNLSLDL SLNPMNFIQP GAFKEIRLHK LTLRNNFDSL NVMKTCIQGL
251  AGLEVHRLVL GEFRNEGNLE KFDKSALEGL CNLTIEEFRL AYLDYYLDDI
301  IDLFNCLTNV SSFSLVSVTI ERVKDFSYNF GWQHLELVNC KFGQFPTLKL
351  KSLKRLTFTS NKGGNAFSEV DLPSLEFLDL SRNGLSFKGC CSQSDFGTTS
401  LKYLDLSFNG VITMSSNFLG LEQLEHLDFQ HSNLKQMSEF SVFLSLRNLI
451  YLDISHTHTR VAFNGIFNGL SSLEVLKMAG NSFQENFLPD IFTELRNLTF
501  LDLSQCQLEQ LSPTAFNSLS SLQVLNMSHN NFFSLDTFPY KCLNSLQVLD
551  YSLNHIMTSK KQELQHFPSS LAFLNLTQND FACTCEHQSF LQWIKDQRQL
601  LVEVERMECA TPSDKQGMPV LSLNITCQMN KTIIGVSVLS VLVVSVVAVL
651  VYKFYFHLML LAGCIKYGRG ENIYDAFVIY SSQDEDWVRN ELVKNLEEGV
701  PPFQLCLHYR DFIPGVAIAA NIIHEGFHKS RKVIVVVSQH FIQSRWCIFE
751  YEIAQTWQFL SSRAGIIFIV LQKVEKTLLR QQVELYRLLS RNTYLEWEDS
801  VLGRHIFWRR LRKALLDGKS WNPEGTVGTG CNWQEATSI
```

TLR4 (mouse)

```
   1  MMPPWLLART LIMALFFSCL TPGSLNPCIE VVPNITYQCM DQKLSKVPDD
  51  IPSSTKNIDL SFNPLKILKS YSFSNFSELQ WLDLSRCEIE TIEDKAWHGL
 101  HHLSNLILTG NPIQSFSPGS FSGLTSLENL VAVETKLASL ESFPIGQLIT
 151  LKKLNVAHNF IHSCKLPAYF SNLTNLVHVD LSYNYIQTIT VNDLQFLREN
 201  PQVNLSLDMS LNPIDFIQDQ AFQGIKLHEL TLRGNFNSSN IMKTCLQNLA
 251  GLHVHRLILG EFKDERNLEI FEPSIMEGLC DVTIDEFRLT YTNDFSDDIV
 301  KFHCLANVSA MSLAGVSIKY LEDVPKHFKW QSLSIIRCQL KQFPTLDLPF
 351  LKSLTLTMNK GSISFKKVAL PSLSYLDLSR NALSFSGCCS YSDLGTNSLR
 401  HLDLSFNGAI IMSANFMGLE ELQHLDFQHS TLKRVTEFSA FLSLEKLLYL
 451  DISYTNTKID FDGIFLGLTS LNTLKMAGNS FKDNTLSNVF ANTTNLTFLD
 501  LSKCQLEQIS WGVFDTLHRL QLLNMSHNNL LFLDSSHYNQ LYSLSTLDCS
 551  FNRIETSKGI LQHFPKSLAF FNLTNNSVAC ICEHQKFLQW VKEQKQFLVN
 601  VEQMTCATPV EMNTSLVLDF NNSTCYMYKT IISVSVVSVI VVSTVAFLIY
 651  HFYFHLILIA GCKKYSRGES IYDAFVIYSS QNEDWVRNEL VKNLEEGVPR
 701  FHLCLHYRDF IPGVAIAANI IQEGFHKSRK VIVVVSRHFI QSRWCIFEYE
 751  IAQTWQFLSS RSGIIFIVLE KVEKSLLRQQ VELYRLLSRN TYLEWEDNPL
 801  GRHIFWRRLK NALLDGKASN PEQTAEEEQE TATWT
```

TLR7 (Human)

```
   1  MVFPMWTLKR QILILFNIIL ISKLLGARWF PKTLPCDVTL DVPKNHVIVD
  51  CTDKHLTEIP GGIPTNTTNL TLTINHIPDI SPASFHRLDH LVEIDFRCNC
 101  VPIPLGSKNN MCIKRLQIKP RSFSGLTYLK SLYLDGNQLL EIPQGLPPSL
 151  QLLSLEANNI FSIRKENLTE LANIEILYLG QNCYYRNPCY VSYSIEKDAF
 201  LNLTKLKVLS LKDNNVTAVP TVLPSTLTEL YLYNNMIAKI QEDDFNNLNQ
 251  LQILDLSGNC PRCYNAPFPC APCKNNSPLQ IPVNAFDALT ELKVLRLHSN
 301  SLQHVPPRWF KNINKLQELD LSQNFLAKEI GDAKFLHFLP SLIQLDLSFN
 351  FELQVYRASM NLSQAFSSLK SLKILRIRGY VFKELKSFNL SPLHNLQNLE
 401  VLDLGTNFIK IANLSMFKQF KRLKVIDLSV NKISPSGDSS EVGFCSNART
 451  SVESYEPQVL EQLHYFRYDK YARSCRFKNK EASFMSVNES CYKYGQTLDL
 501  SKNSIFFVKS SDFQHLSFLK CLNLSGNLIS QTLNGSEFQP LAELRYLDFS
 551  NNRLDLLHST AFEELHKLEV LDISSNSHYF QSEGITHMLN FTKNLKVLQK
 601  LMMNDNDISS STSRTMESES LRTLEFRGNH LDVLWREGDN RYLQLFKNLL
 651  KLEELDISKN SLSFLPSGVF DGMPPNLKNL SLAKNGLKSF SWKKLQCLKN
 701  LETLDLSHNQ LTTVPERLSN CSRSLKNLIL KNNQIRSLTK YFLQDAFQLR
 751  YLDLSSNKIQ MIQKTSFPEN VLNNLKMLLL HHNRFLCTCD AVWFVWWVNH
 801  TEVTIPYLAT DVTCVGPGAH KGQSVISLDL YTCELDLTNL ILFSLSISVS
 851  LFLMVMMTAS HLYFWDVWYI YHFCKAKIKG YQRLISPDCC YDAFIVYDTK
 901  DPAVTEWVLA ELVAKLEDPR EKHFNLCLEE RDWLPGQPVL ENLSQSIQLS
 951  KKTVFVMTDK YAKTENFKIA FYLSHQRLMD EKVDVIILIF LEKPFQKSKF
1001  LQLRKRLCGS SVLEWPTNPQ AHPYFWQCLK NALATDNHVA YSQVFKETV
```

TLR8 (Human)

```
   1  MENMFLQSSM LTCIFLLISG SCELCAEENF SRSYPCDEKK QNDSVIAECS
  51  NRRLQEVPQT VGKYVTELDL SDNFITHITN ESFQGLQNLT KINLNHNPNV
 101  QHQNGNPGIQ SNGLNITDGA FLNLKNLREL LLEDNQLPQI PSGLPESLTE
 151  LSLIQNNIYN ITKEGISRLI NLKNLYLAWN CYFNKVCEKT NIEDGVFETL
 201  TNLELLSLSF NSLSHVPPKL PSSLRKLFLS NTQIKYISEE DFKGLINLTL
 251  LDLSGNCPRC FNAPFPCVPC DGGASINIDR FAFQNLTQLR YLNLSSTSLR
 301  KINAAWFKNM PHLKVLDLEF NYLVGEIASG AFLTMLPRLE ILDLSFNYIK
 351  GSYPQHINIS RNFSKLLSLR ALHLRGYVFQ ELREDDFQPL MQLPNLSTIN
 401  LGINFIKQID FKLFQNFSNL EIIYLSENRI SPLVKDTRQS YANSSSFQRH
 451  IRKRRSTDFE FDPHSNFYHF TRPLIKPQCA AYGKALDLSL NSIFFIGPNQ
 501  FENLPDIACL NLSANSNAQV LSGTEFSAIP HVKYLDLTNN RLDFDNASAL
 551  TELSDLEVLD LSYNSHYFRI AGVTHHLEFI QNFTNLKVLN LSHNNIYTLT
 601  DKYNLESKSL VELVFSGNRL DILWNDDDNR YISIFKGLKN LTRLDLSLNR
 651  LKHIPNEAFL NLPASLTELH INDNMLKFFN WTLLQQFPRL ELLDLRGNKL
 701  LFLTDSLSDF TSSLRTLLLS HNRISHLPSG FLSEVSSLKH LDLSSNLLKT
 751  INKSALETKT TTKLSMLELH GNPFECTCDI GDFRRWMDEH LNVKIPRLVD
 801  VICASPGDQR GKSIVSLELT TCVSDVTAVI LFFFTFFITT MVMLAALAHH
 851  LFYWDVWFIY NVCLAKVKGY RSLSTSQTFY DAYISYDTKD ASVTDWVINE
 901  LRYHLEESRD KNVLLCLEER DWDPGLAIID NLMQSINQSK KTVFVLTKKY
 951  AKSWNFKTAF YLALQRLMDE NMDVIIFILL EPVLQHSQYL RLRQRICKSS
1001  ILQWPDNPKA EGLFWQTLRN VVLTENDSRY NNMYVDSIKQ Y
```

TLR9 (Human)

```
   1  MGFCRSALHP LSLLVQAIML AMTLALGTLP AFLPCELQPH GLVNCNWLFL
  51  KSVPHFSMAA PRGNVTSLSL SSNRIHHLHD SDFAHLPSLR HLNLKWNCPP
 101  VGLSPMHFPC HMTIEPSTFL AVPTLEELNL SYNNIMTVPA LPKSLISLSL
 151  SHTNILMLDS ASLAGLHALR FLFMDGNCYY KNPCRQALEV APGALLGLGN
 201  LTHLSLKYNN LTVVPRNLPS SLEYLLLSYN RIVKLAPEDL ANLTALRVLD
 251  VGGNCRRCDH APNPCMECPR HFPQLHPDTF SHLSRLEGLV LKDSSLSWLN
 301  ASWFRGLGNL RVLDLSENFL YKCITKTKAF QGLTQLRKLN LSFNYQKRVS
 351  FAHLSLAPSF GSLVALKELD MHGIFFRSLD ETTLRPLARL PMLQTLRLQM
 401  NFINQAQLGI FRAFPGLRYV DLSDNRISGA SELTATMGEA DGGEKVWLQP
 451  GDLAPAPVDT PSSEDFRPNC STLNFTLDLS RNNLVTVQPE MFAQLSHLQC
 501  LRLSHNCISQ AVNGSQFLPL TGLQVLDLSH NKLDLYHEHS FTELPRLEAL
 551  DLSYNSQPFG MQGVGHNFSF VAHLRTLRHL SLAHNNIHSQ VSQQLCSTSL
 601  RALDFSGNAL GHMWAEGDLY LHFFQGLSGL IWLDLSQNRL HTLLPQTLRN
 651  LPKSLQVLRL RDNYLAFFKW WSLHFLPKLE VLDLAGNQLK ALTNGSLPAG
 701  TRLRRLDVSC NSISFVAPGF FSKAKELREL NLSANALKTV DHSWFGPLAS
 751  ALQILDVSAN PLHCACGAAF MDFLLEVQAA VPGLPSRVKC GSPGQLQGLS
 801  IFAQDLRLCL DEALSWDCFA LSLLAVALGL GVPMLHHLCG WDLWYCFHLC
 851  LAWLPWRGRQ SGRDEDALPY DAFVVFDKTQ SAVADWVYNE LRGQLEECRG
 901  RWALRLCLEE RDWLPGKTLF ENLWASVYGS RKTLFVLAHT DRVSGLLRAS
 951  FLLAQQRLLE DRKDVVVLVI LSPDGRRSRY VRLRQRLCRQ SVLLWPHQPS
1001  GQRSFWAQLG MALTRDNHHF YNRNFCQGPT AE
```

TLR10 (Human)

```
  1  MRLIRNIYIF CSIVMTAEGD APELPEEREL MTNCSNMSLR KVPADLTPAT
 51  TTLDLSYNLL FQLQSSDFHS VSKLRVLILC HNRIQQLDLK TFEFNKELRY
101  LDLSNNRLKS VTWYLLAGLR YLDLSFNDFD TMPICEEAGN MSHLEILGLS
151  GAKIQKSDFQ KIAHLHLNTV FLGFRTLPHY EEGSLPILNT TKLHIVLPMD
201  TNFWVLLRDG IKTSKILEMT NIDGKSQFVS YEMQRNLSLE NAKTSVLLLN
251  KVDLLWDDLF LILQFVWHTS VEHFQIRNVT FGGKAYLDHN SFDYSNTVMR
301  TIKLEHVHFR VFYIQQDKIY LLLTKMDIEN LTISNAQMPH MLFPNYPTKF
351  QYLNFANNIL TDELFKRTIQ LPHLKTLILN GNKLETLSLV SCFANNTPLE
401  HLDLSQNLLQ HKNDENCSWP ETVVNMNLSY NKLSDSVFRC LPKSIQILDL
451  NNNQIQTVPK ETIHLMALRE LNIAFNFLTD LPGCSHFSRL SVLNIEMNFI
501  LSPSLDFVQS CQEVKTLNAG RNPFRCTCEL KNFIQLETYS EVMMVGWSDS
551  YTCEYPLNLR GTRLKDVHLH ELSCNTALLI VTIVVIMLVL GLAVAFCCLH
601  FDLPWYLRML GQCTQTWHRV RKTTQEQLKR NVRFHAFISY SEHDSLWVKN
651  ELIPNLEKED GSILICLYES YFDPGKSISE NIVSFIEKSY KSIFVLSPNF
701  VQNEWCHYEF YFAHHNLFHE NSDHIILILL EPIPFYCIPT RYHKLKALLE
751  KKAYLEWPKD RRKCGLFWAN LRAAINVNVL ATREMYELQT FTELNEESRG
801  STISLMRTDC L
```

RP105 (Human)

```
  1  MAFDVSCFFW VVLFSAGCKV ITSWDQMCIE KEANKTYNCE NLGLSEIPDT
 51  LPNTTEFLEF SFNFLPTIHN RTFSRLMNLT FLDLTRCQIN WIHEDTFQSH
101  HQLSTLVLTG NPLIFMAETS LNGPKSLKHL FLIQTGISNL EFIPVHNLEN
151  LESLYLGSNH ISSIKFPKDF PARNLKVLDF QNNAIHYISR EDMRSLEQAI
201  NLSLNFNGNN VKGIELGAFD STVFQSLNFG GTPNLSVIFN GLQNSTTQSL
251  WLGTFEDIDD EDISSAMLKG LCEMSVESLN LQEHRFSDIS STTFQCFTQL
301  QELDLTATHL KGLPSGMKGL NLLKKLVLSV NHFDQLCQIS AANFPSLTHL
351  YIRGNVKKLH LGVGCLEKLG NLQTLDLSHN DIEASDCCSL QLKNLSHLQT
401  LNLSHNEPLG LQSQAFKECP QLELLDLAFT RLHINAPQSP FQNLHFLQVL
451  NLTYCFLDTS NQHLLAGLPV LRHLNLKGNH FQDGTITKTN LLQTVGSLEV
501  LILSSCGLLS IDQQAFHSLG KMSHVDLSHN SLTCDSIDSL SHLKGIYLNL
551  AANSINIISP RLLPILSQQS TINLSHNPLD CTCSNIHFLT WYKENLHKLE
601  GSEETTCANP PSLRGVKLSD VKLSCGITAI GIFFLIVFLL LLAILLFFAV
651  KYLLRWKYQH I
```

RP105 (mouse)

```
  1  MAPDISCFFL VALFLASCRA TTSSDQKCIE KEVNKTYNCE NLGLNEIPGT
 51  LPNSTECLEF SFNVLPTIQN TTFSRLINLT FLDLTRCQIY WIHEDTFQSQ
101  HRLDTLVLTA NPLIFMAETA LSGPKALKHL FFIQTGISSI DFIPLHNQKT
151  LESLYLGSNH ISSIKLPKGF PTEKLKVLDF QNNAIHYLSK EDMSSLQQAT
201  NLSLNLNGND IAGIELGAFD SAVFQSLNFG GTQNLLVIFK GLKNSTIQSL
251  WLGTFEDMDD EDISPAVFEG LCEMSVESIN LQKHYFFNIS SNTFHCFSGL
301  QELDLTATHL SELPSGLVGL STLKKLVLSA NKFENLCQIS ASNFPSLTHL
351  SIKGNTKRLE LGTGCLENLE NLRELDLSHD DIETSDCCNL QLRNLSHLQS
401  LNLSYNEPLS LKTEAFKECP QLELLDLAFT RLKVKDAQSP FQNLHLLKVL
451  NLSHSLLDIS SEQLFDGLPA LQHLNLQGNH FPKGNIQKTN SLQTLGRLEI
501  LVLSFCDLSS IDQHAFTSLK MMNHVDLSHN RLTSSSIEAL SHLKGIYLNL
551  ASNRISIILP SLLPILSQQR TINLRQNPLD CTCSNIYFLE WYKENMQKLE
601  DTEDTLCENP PLLRGVRLSD VTLSCSMAAV GIFFLIVFLL VFAILLIFAV
651  KYFLRWKYQH I
```

MD-1 (Human)

```
  1  MKGFTATLFL WTLIFPSCSG GGGGKAWPTH VVCSDSGLEV LYQSCDPLQD
 51  FGFSVEKCSK QLKSNINIRF GIILREDIKE LFLDLALMSQ GSSVLNFSYP
101  ICEAALPKFS FCGRRKGEQI YYAGPVNNPE FTIPQGEYQV LLELYTEKRS
151  TVACANATIM CS
```

MD-2 (Human)

```
  1  MLPFLFFSTL FSSIFTEAQK QYWVCNSSDA SISYTYCDKM QYPISINVNP
 51  CIELKGSKGL LHIFYIPRRD LKQLYFNLYI TVNTMNLPKR KEVICRGSDD
101  DYSFCRALKG ETVNTTISFS FKGIKFSKGK YKCVVEAISG SPEEMLFCLE
151  FVILHQPNSN
```

CD14 (Human)

```
  1  MERASCLLLL LLPLVHVSAT TPEPCELDDE DFRCVCNFSE PQPDWSEAFQ
 51  CVSAVEVEIH AGGLNLEPFL KRVDADADPR QYADTVKALR VRRLTVGAAQ
101  VPAQLLVGAL RVLAYSRLKE LTLEDLKITG TMPPLPLEAT GLALSSLRLR
151  NVSWATGRSW LAELQQWLKP GLKVLSIAQA HSPAFSCEQV RAFPALTSLD
201  LSDNPGLGER GLMAALCPHK FPAIQNLALR NTGMETPTGV CAALAAAGVQ
251  PHSLDLSHNS LRATVNPSAP RCMWSSALNS LNLSFAGLEQ VPKGLPAKLR
301  VLDLSCNRLN RAPQPDELPE VDNLTLDGNP FLVPGTALPH EGSMNSGVVP
351  ACARSTLSVG VSGTLVLLQG ARGFA
```

References

References

(1987). International Tables for Crystallography, D. Reidel Publishing Company.

(2007). "The Universal Protein Resource (UniProt)." Nucleic Acids Res **35**(Database issue): D193-7.

Abbas, A. K. L., A. H. (2003). Cellular and Molecular Immunology, Saunders.

Akira, S. and K. Takeda (2004). "Toll-like receptor signalling." Nat Rev Immunol **4**(7): 499-511.

Alber, T., W. A. Gilbert, et al., (1983). "The role of mobility in the substrate binding and catalytic machinery of enzymes." Ciba Found Symp **93**: 4-24.

Altschul, S. F., M. S. Boguski, et al., (1994). "Issues in searching molecular sequence databases." Nat Genet **6**(2): 119-29.

Altschul, S. F., W. Gish, et al., (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al., (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Anderson, K. V., L. Bokla, et al., (1985). "Establishment of dorsal-ventral polarity in the Drosophila embryo: the induction of polarity by the Toll gene product." Cell **42**(3): 791-8.

Anderson, K. V., G. Jurgens, et al., (1985). "Establishment of dorsal-ventral polarity in the Drosophila embryo: genetic studies on the role of the Toll gene product." Cell **42**(3): 779-89.

Anderson, K. V. and C. Nusslein-Volhard (1984). "Information for the dorsal--ventral pattern of the Drosophila embryo is stored as maternal mRNA." Nature **311**(5983): 223-7.

Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains." Science **181**(96): 223-30.

Arden, B., S. P. Clark, et al., (1995). "Human T-cell receptor variable gene segment families." Immunogenetics **42**(6): 455-500.

Aricescu, A. R., W. Lu, et al., (2006). "A time- and cost-efficient system for high-level protein production in mammalian cells." Acta Crystallogr D Biol Crystallogr **62**(Pt 10): 1243-50.

Arnett, K. L., S. C. Harrison, et al., (2004). "Crystal structure of a human CD3-epsilon/delta dimer in complex with a UCHT1 single-chain antibody fragment." Proc Natl Acad Sci U S A **101**(46): 16268-73.

Arstila, T. P., A. Casrouge, et al., (1999). "A direct estimate of the human alphabeta T cell receptor diversity." Science **286**(5441): 958-61.

Arstila, T. P., A. Casrouge, et al., (2000). "Diversity of human alpha beta T cell receptors." Science **288**(5469): 1135.

Ayyoub, M., A. Zippelius, et al., (2003). "Activation of human melanoma reactive CD8+ T cells by vaccination with an immunogenic peptide analog derived from Melan-A/melanoma antigen recognized by T cells-1." Clin Cancer Res **9**(2): 669-77.

Baker, B. M., R. V. Turner, et al., (2001). "Identification of a crucial energetic footprint on the alpha1 helix of human histocompatibility leukocyte antigen (HLA)-A2 that provides functional interactions for recognition by tax peptide/HLA-A2-specific T cell receptors." J Exp Med **193**(5): 551-62.

Banchereau, J. and R. M. Steinman (1998). "Dendritic cells and the control of immunity." Nature **392**(6673): 245-52.

Bell, J. K., I. Botos, et al., (2005). "The molecular structure of the Toll-like receptor 3 ligand-binding domain." Proc Natl Acad Sci U S A **102**(31): 10976-80.

Bendtsen, J. D., H. Nielsen, et al., (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol **340**(4): 783-95.

Berman, H. M., J. Westbrook, et al., (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-42.

Bjorkman, P. J., M. A. Saper, et al., (1987). "The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens." Nature **329**(6139): 512-8.

Bjorkman, P. J., M. A. Saper, et al., (1987). "Structure of the human class I histocompatibility antigen, HLA-A2." Nature **329**(6139): 506-12.

Blow, D. (2003). Outline of crystallography for biologists, Oxford University Press.

Bork, P., L. Holm, et al., (1994). "The immunoglobulin fold. Structural classification, sequence patterns and common core." J Mol Biol **242**(4): 309-20.

Boutselakis, H., D. Dimitropoulos, et al., (2003). "E-MSD: the European Bioinformatics Institute Macromolecular Structure Database." Nucleic Acids Res **31**(1): 458-62.

Brown, J. H., T. S. Jardetzky, et al., (1993). "Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1." Nature **364**(6432): 33-9.

Brunger, A. T., P. D. Adams, et al., (1998). "Crystallography & NMR system: A new software suite for macromolecular structure determination." Acta Crystallogr D Biol Crystallogr **54**(Pt 5): 905-21.

Burley, S. K. and G. A. Petsko (1985). "Aromatic-aromatic interaction: a mechanism of protein structure stabilization." Science **229**(4708): 23-8.

Buslepp, J., H. Wang, et al., (2003). "A correlation between TCR Valpha docking on MHC and CD8 dependence: implications for T cell selection." Immunity **19**(4): 595-606.

Casanova, J. L., P. Romero, et al., (1991). "T cell receptor genes in a series of class I major histocompatibility complex-restricted cytotoxic T lymphocyte clones specific for a Plasmodium berghei nonapeptide: implications for T cell allelic exclusion and antigen-specific repertoire." J Exp Med **174**(6): 1371-83.

Chen, J. L., G. Stewart-Jones, et al., (2005). "Structural and kinetic basis for heightened immunogenicity of T cell vaccines." J Exp Med **201**(8): 1243-55.

Chen, Y. T., M. J. Scanlan, et al., (1997). "A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening." Proc Natl Acad Sci U S A **94**(5): 1914-8.

Chenna, R., H. Sugawara, et al., (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Res **31**(13): 3497-500.

Choe, J., M. S. Kelker, et al., (2005). "Crystal structure of human toll-like receptor 3 (TLR3) ectodomain." Science **309**(5734): 581-5.

Chuang, T. and R. J. Ulevitch (2001). "Identification of hTLR10: a novel human Toll-like receptor preferentially expressed in immune cells." Biochim Biophys Acta **1518**(1-2): 157-61.

Chuang, T. H. and R. J. Ulevitch (2000). "Cloning and characterization of a sub-family of human toll-like receptors: hTLR7, hTLR8 and hTLR9." Eur Cytokine Netw **11**(3): 372-8.

Coeytaux, K. and A. Poupon (2005). "Prediction of unfolded segments in a protein sequence based on amino acid composition." Bioinformatics **21**(9): 1891-900.

Colf, L. A., A. J. Bankovich, et al., (2007). "How a single T cell receptor recognizes both self and foreign MHC." Cell **129**(1): 135-46.

Coulie, P. G., V. Brichard, et al., (1994). "A new gene coding for a differentiation antigen recognized by autologous cytolytic T lymphocytes on HLA-A2 melanomas." J Exp Med **180**(1): 35-42.

Crooks, G. E., G. Hon, et al., (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-90.

Dasarathy, B. V. (1991). "Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques." IEEE Computer Society Press, Los Alamitos, CA.

Dasarathy, B. V. (1994). "Minimal consistent set (MCS) identification for optimal nearest neighbour decision systems design." IEEE Trans. Systems Man Cybernetics **24**: 511-517.

Dave, V. P., Z. Cao, et al., (1997). "CD3 delta deficiency arrests development of the alpha beta but not the gamma delta T cell lineage." Embo J **16**(6): 1360-70.

Davis, M. M. and P. J. Bjorkman (1988). "T-cell antigen receptor genes and T-cell recognition." Nature **334**(6181): 395-402.

Degano, M., K. C. Garcia, et al., (2000). "A functional hot spot for antigen recognition in a superagonist TCR/MHC complex." Immunity **12**(3): 251-61.

Delano, W. L. (2002). The Pymol Molecular Graphics System.

Divanovic, S., A. Trompette, et al., (2005). "Negative regulation of Toll-like receptor 4 signaling by the Toll-like receptor homolog RP105." Nat Immunol **6**(6): 571-8.

Dobson, C. M. (2003). "Protein folding and misfolding." Nature **426**(6968): 884-90.

Dosztanyi, Z., V. Csizmok, et al., (2005). "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content." Bioinformatics **21**(16): 3433-4.

Dudley, M. E., J. R. Wunderlich, et al., (2002). "Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes." Science **298**(5594): 850-4.

Dunker, A. K., Z. Obradovic, et al., (2000). "Intrinsic protein disorder in complete genomes." Genome Inform Ser Workshop Genome Inform **11**: 161-71.

Dutoit, V., V. Rubio-Godoy, et al., (2002). "Degeneracy of antigen recognition as the molecular basis for the high frequency of naive A2/Melan-a peptide multimer(+) CD8(+) T cells in humans." J Exp Med **196**(2): 207-16.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2126-32.

Enright, A. J., S. Van Dongen, et al., (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res **30**(7): 1575-84.

Falk, K., O. Rotzschke, et al., (1991). "Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules." Nature **351**(6324): 290-6.

Fearon, D. T. and R. M. Locksley (1996). "The instructive role of innate immunity in the acquired immune response." Science **272**(5258): 50-3.

Fersht, A. R. (2000). "Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism." Proc Natl Acad Sci U S A **97**(4): 1525-9.

Frankel, A. D. and P. S. Kim (1991). "Modular structure of transcription factors: implications for gene regulation." Cell **65**(5): 717-9.

Fuxreiter, M., P. Tompa, et al., (2007). "Local structural disorder imparts plasticity on linear motifs." Bioinformatics **23**(8): 950-6.

Gao, G. F., J. Tormo, et al., (1997). "Crystal structure of the complex between human CD8alpha(alpha) and HLA-A2." Nature **387**(6633): 630-4.

Garboczi, D. N., P. Ghosh, et al., (1996). "Structure of the complex between human T-cell receptor, viral peptide and HLA-A2." Nature **384**(6605): 134-41.

Garboczi, D. N., D. R. Madden, et al., (1994). "Five viral peptide-HLA-A2 co-crystals. Simultaneous space group determination and X-ray data collection." J Mol Biol **239**(4): 581-7.

Garner, E., P. Cannon, et al., (1998). "Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization." Genome Inform Ser Workshop Genome Inform **9**: 201-213.

Garner, E., P. Romero, et al., (1999). "Predicting Binding Regions within Disordered Proteins." Genome Inform Ser Workshop Genome Inform **10**: 41-50.

Gay, N. J. and F. J. Keith (1991). "Drosophila Toll and IL-1 receptor." Nature **351**(6325): 355-6.

Geisler, C. (1992). "Failure to synthesize the CD3-gamma chain. Consequences for T cell antigen receptor assembly, processing, and expression." J Immunol **148**(8): 2437-45.

Germain, R. N. (1994). "MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation." Cell **76**(2): 287-99.

Germain, R. N. and I. Stefanova (1999). "The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation." Annu Rev Immunol **17**: 467-522.

Gilis, D., S. Massar, et al., (2001). "Optimality of the genetic code with respect to protein stability and amino-acid frequencies." Genome Biol **2**(11): RESEARCH0049.

Haks, M. C., P. Krimpenfort, et al., (1998). "The CD3gamma chain is essential for development of both the TCRalphabeta and TCRgammadelta lineages." Embo J **17**(7): 1871-82.

Hennecke, J., A. Carfi, et al., (2000). "Structure of a covalently stabilized complex of a human alphabeta T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1." Embo J **19**(21): 5611-24.

Hennecke, J. and D. C. Wiley (2001). "T cell receptor-MHC interactions up close." Cell **104**(1): 1-4.

Hoffmann, J. A. (1995). "Innate immunity of insects." Curr Opin Immunol **7**(1): 4-10.

Hunt, D. F., R. A. Henderson, et al., (1992). "Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry." Science **255**(5049): 1261-3.

Hunt, D. F., H. Michel, et al., (1992). "Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad." Science **256**(5065): 1817-20.

Huseby, E. S., J. White, et al., (2005). "How the T cell repertoire becomes peptide and MHC specific." Cell **122**(2): 247-60.

Ip, Y. T. and M. Levine (1994). "Molecular genetics of Drosophila immunity." Curr Opin Genet Dev **4**(5): 672-7.

Janeway, C. A., Jr., S. Carding, et al., (1988). "CD4+ T cells: specificity and function." Immunol Rev **101**: 39-80.

Julenius, K., A. Molgaard, et al., (2005). "Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites." Glycobiology **15**(2): 153-64.

June, C. H., J. A. Bluestone, et al., (1994). "The B7 and CD28 receptor families." Immunol Today **15**(7): 321-31.

Kawakami, Y., S. Eliyahu, et al., (1994). "Cloning of the gene coding for a shared human melanoma antigen recognized by autologous T cells infiltrating into tumor." Proc Natl Acad Sci U S A **91**(9): 3515-9.

Kawakami, Y., S. Eliyahu, et al., (1994). "Identification of the immunodominant peptides of the MART-1 human melanoma antigen recognized by the majority of HLA-A2-restricted tumor infiltrating lymphocytes." J Exp Med **180**(1): 347-52.

Kern, P. S., M. K. Teng, et al., (1998). "Structural basis of CD8 coreceptor function revealed by crystallographic analysis of a murine CD8alphaalpha ectodomain fragment in complex with H-2Kb." Immunity **9**(4): 519-30.

Kim, J. I., C. J. Lee, et al., (2005). "Crystal structure of CD14 and its implications for lipopolysaccharide signaling." J Biol Chem **280**(12): 11347-51.

Kim, K. S., Z. Y. Sun, et al., (2000). "Heterodimeric CD3epsilongamma extracellular domain fragments: production, purification and structural analysis." J Mol Biol **302**(4): 899-916.

Kjer-Nielsen, L., M. A. Dunstone, et al., (2004). "Crystal structure of the human T cell receptor CD3 epsilon gamma heterodimer complexed to the therapeutic mAb OKT3." Proc Natl Acad Sci U S A **101**(20): 7675-80.

Konno, K., Y. Wakabayashi, et al., (2006). "A molecule that is associated with Toll-like receptor 4 and regulates its cell surface expression." Biochem Biophys Res Commun **339**(4): 1076-82.

Lawrence, M. C. and P. M. Colman (1993). "Shape complementarity at protein/protein interfaces." J Mol Biol **234**(4): 946-50.

Leahy, D. J., R. Axel, et al., (1992). "Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2.6 A resolution." Cell **68**(6): 1145-62.

Lefranc, M. P., V. Giudicelli, et al., (2005). "IMGT, the international ImMunoGeneTics information system." Nucleic Acids Res **33**(Database issue): D593-7.

Lemaitre, B., E. Nicolas, et al., (1996). "The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults." Cell **86**(6): 973-83.

Levashina, E. A., E. Langley, et al., (1999). "Constitutive activation of toll-mediated antifungal defense in serpin-deficient Drosophila." Science **285**(5435): 1917-9.

Li, X., P. Romero, et al., (1999). "Predicting Protein Disorder for N-, C-, and Internal Regions." Genome Inform Ser Workshop Genome Inform **10**: 30-40.

Lilius, G., M. Persson, et al., (1991). "Metal affinity precipitation of proteins carrying genetically attached polyhistidine affinity tails." Eur J Biochem **198**(2): 499-504.

Linding, R., L. J. Jensen, et al., (2003). "Protein disorder prediction: implications for structural proteomics." Structure (Camb) **11**(11): 1453-9.

Linding, R., R. B. Russell, et al., (2003). "GlobPlot: Exploring protein sequences for globularity and disorder." Nucleic Acids Res **31**(13): 3701-8.

Liu, Y., Y. Xiong, et al., (2003). "The crystal structure of a TL/CD8alphaalpha complex at 2.1 A resolution: implications for modulation of T cell activation and memory." Immunity **18**(2): 205-15.

Madden, D. R., J. C. Gorga, et al., (1991). "The structure of HLA-B27 reveals nonamer self-peptides bound in an extended conformation." Nature **353**(6342): 321-5.

Madden, D. R., J. C. Gorga, et al., (1992). "The three-dimensional structure of HLA-B27 at 2.1 A resolution suggests a general mechanism for tight peptide binding to MHC." Cell **70**(6): 1035-48.

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." Biochim Biophys Acta **405**(2): 442-51.

McCoy, A. J., R. W. Grosse-Kunstleve, et al., (2005). "Likelihood-enhanced fast translation functions." Acta Crystallogr D Biol Crystallogr **61**(Pt 4): 458-64.

McDonald, I. K. and J. M. Thornton (1994). "Satisfying hydrogen bonding potential in proteins." J Mol Biol **238**(5): 777-93.

Medzhitov, R. and C. A. Janeway, Jr. (1997). "Innate immunity: impact on the adaptive immune response." Curr Opin Immunol **9**(1): 4-9.

Medzhitov, R., P. Preston-Hurlburt, et al., (1997). "A human homologue of the Drosophila Toll protein signals activation of adaptive immunity." Nature **388**(6640): 394-7.

Mitchell, J. A., K. A. Fitzgerald, et al., (2006). "TOLLing away in Brazil." Nat Immunol **7**(7): 675-9.

Morgan, R. A., M. E. Dudley, et al., (2006). "Cancer regression in patients after transfer of genetically engineered lymphocytes." Science **314**(5796): 126-9.

Murshudov, G. N., A. A. Vagin, et al., (1997). "Refinement of macromolecular structures by the maximum-likelihood method." Acta Crystallogr D Biol Crystallogr **53**(Pt 3): 240-55.

O'Donovan, C., M. J. Martin, et al., (2002). "High-quality protein knowledge resource: SWISS-PROT and TrEMBL." Brief Bioinform **3**(3): 275-84.

Obradovic, Z., K. Peng, et al., (2005). "Exploiting heterogeneous sequence properties improves prediction of protein disorder." Proteins **61 Suppl 7**: 176-82.

Ochsenbein, A. F., P. Klenerman, et al., (1999). "Immune surveillance against a solid tumor fails because of immunological ignorance." Proc Natl Acad Sci U S A **96**(5): 2233-8.

Ohto, U., K. Fukase, et al., (2007). "Crystal structures of human MD-2 and its complex with antiendotoxic lipid IVa." Science **316**(5831): 1632-4.

Otwinowski, Z. and W. Minor (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods in Enzymology. J. R. M. S. C.W. Carter. New York, Academic Press. **276:** 307-326.

Padlan, E. A. (1990). "On the nature of antibody combining sites: unusual structural features that may confer on these sites an enhanced capacity for binding ligands." Proteins **7**(2): 112-24.

Padovan, E., G. Casorati, et al., (1993). "Expression of two T cell receptor alpha chains: dual receptor T cells." Science **262**(5132): 422-4.

Peng, K., P. Radivojac, et al., (2006). "Length-dependent prediction of protein intrinsic disorder." BMC Bioinformatics **7**: 208.

Perrakis, A., M. Harkiolaki, et al., (2001). "ARP/wARP and molecular replacement." Acta Crystallogr D Biol Crystallogr **57**(Pt 10): 1445-50.

Prilusky, J., C. E. Felder, et al., (2005). "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded." Bioinformatics **21**(16): 3435-8.

Radivojac, P., Z. Obradovic, et al., (2003). "Prediction of boundaries between intrinsically ordered and disordered protein regions." Pac Symp Biocomput: 216-27.

Radivojac, P., Z. Obradovic, et al., (2004). "Protein flexibility and intrinsic disorder." Protein Sci **13**(1): 71-80.

Rajamani, D., S. Thiel, et al., (2004). "Anchor residues in protein-protein interactions." Proc Natl Acad Sci U S A **101**(31): 11287-92.

Reche, P. A. and E. L. Reinherz (2003). "Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms." J Mol Biol **331**(3): 623-41.

Reichhart, J. M., M. Meister, et al., (1992). "Insect immunity: developmental and inducible activity of the Drosophila diptericin promoter." Embo J **11**(4): 1469-77.

Reinherz, E. L., K. Tan, et al., (1999). "The crystal structure of a T cell receptor in complex with peptide and MHC class II." Science **286**(5446): 1913-21.

Reis e Sousa, C. (2001). "Dendritic cells as sensors of infection." Immunity **14**(5): 495-8.

Reiser, J. B., C. Darnault, et al., (2000). "Crystal structure of a T cell receptor bound to an allogeneic MHC molecule." Nat Immunol **1**(4): 291-7.

Rhodes, G. (2000). Crystallography made crystal clear, Academic Press.

Robbins, P. F. and Y. Kawakami (1996). "Human tumor antigens recognized by T cells." Curr Opin Immunol **8**(5): 628-36.

Roche, P. A. and P. Cresswell (1990). "Invariant chain association with HLA-DR molecules inhibits immunogenic peptide binding." Nature **345**(6276): 615-8.

Romero, P., N. Gervois, et al., (1997). "Cytolytic T lymphocyte recognition of the immunodominant HLA-A*0201-restricted Melan-A/MART-1 antigenic peptide in melanoma." J Immunol **159**(5): 2366-74.

Rowen, L., B. F. Koop, et al., (1996). "The complete 685-kilobase DNA sequence of the human beta T cell receptor locus." Science **272**(5269): 1755-62.

Rudensky, A., P. Preston-Hurlburt, et al., (1991). "Sequence analysis of peptides bound to MHC class II molecules." Nature **353**(6345): 622-7.

Rudolph, M. G., M. Huang, et al., (2001). "Crystal structure of an isolated V(alpha) domain of the 2C T-cell receptor." J Mol Biol **314**(1): 1-8.

Rudolph, M. G., R. L. Stanfield, et al., (2006). "How TCRs bind MHCs, peptides, and coreceptors." Annu Rev Immunol **24**: 419-66.

Rudolph, M. G. and I. A. Wilson (2002). "The specificity of TCR/pMHC interaction." Curr Opin Immunol **14**(1): 52-65.

Saper, M. A., P. J. Bjorkman, et al., (1991). "Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution." J Mol Biol **219**(2): 277-319.

Schulz, G. E. (1979). Nucleotide binding proteins. Molecular Mechanism of Biological Recognition, Elsevier/North-Holland Biomedical Press**:** 79-94.

Shimazu, R., S. Akashi, et al., (1999). "MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4." J Exp Med **189**(11): 1777-82.

Sliz, P., O. Michielin, et al., (2001). "Crystal structures of two closely related but antigenically distinct HLA-A2/melanocyte-melanoma tumor-antigen peptide complexes." J Immunol **167**(6): 3276-84.

Starr, T. K., S. C. Jameson, et al., (2003). "Positive and negative selection of T cells." Annu Rev Immunol **21**: 139-76.

Stewart-Jones, G. B., A. J. McMichael, et al., (2003). "A structural basis for immunodominant human T cell receptor recognition." Nat Immunol **4**(7): 657-63.

Stuart, D. I., M. Levine, et al., (1979). "Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 A." J Mol Biol **134**(1): 109-42.

Sussman, J. J., J. S. Bonifacino, et al., (1988). "Failure to synthesize the T cell CD3-zeta chain: structure and function of a partial T cell receptor complex." Cell **52**(1): 85-95.

Taguchi, T., J. L. Mitcham, et al., (1996). "Chromosomal localization of TIL, a gene encoding a protein related to the Drosophila transmembrane receptor Toll, to human chromosome 4p14." Genomics **32**(3): 486-8.

Takeda, K. and S. Akira (2005). "Toll-like receptors in innate immunity." Int Immunol **17**(1): 1-14.

Takeda, K., T. Kaisho, et al., (2003). "Toll-like receptors." Annu Rev Immunol **21**: 335-76.

Takeuchi, O., T. Kawai, et al., (2001). "Discrimination of bacterial lipoproteins by Toll-like receptor 6." Int Immunol **13**(7): 933-40.

Takeuchi, O., T. Kawai, et al., (1999). "TLR6: A novel member of an expanding toll-like receptor family." Gene **231**(1-2): 59-65.

Takeuchi, O., S. Sato, et al., (2002). "Cutting edge: role of Toll-like receptor 1 in mediating immune response to microbial lipoproteins." J Immunol **169**(1): 10-4.

Teng, M. K., A. Smolyar, et al., (1998). "Identification of a common docking topology with substantial variation among different TCR-peptide-MHC complexes." Curr Biol **8**(7): 409-12.

Teyton, L., D. O'Sullivan, et al., (1990). "Invariant chain distinguishes between the exogenous and endogenous antigen presentation pathways." Nature **348**(6296): 39-44.

Thomson, R., Esnouf, R. (2004). "Prediction of natively disordered regions in proteins using a bio-basis function neural network." Lecture Notes in Computer Science **3177**: 108-116.

Thomson, R., T. C. Hodgman, et al., (2003). "Characterizing proteolytic cleavage site activity using bio-basis function neural networks." Bioinformatics **19**(14): 1741-7.

Tompa, P. (2002). "Intrinsically unstructured proteins." Trends Biochem Sci **27**(10): 527-33.

Trautmann, L., N. Labarriere, et al., (2002). "Dominant TCR V alpha usage by virus and tumor-reactive T cells with wide affinity ranges for their specific antigens." Eur J Immunol **32**(11): 3181-90.

Trombetta, E. S. and I. Mellman (2005). "Cell biology of antigen processing in vitro and in vivo." Annu Rev Immunol **23**: 975-1028.

Trowsdale, J. (1995). ""Both man & bird & beast": comparative organization of MHC genes." Immunogenetics **41**(1): 1-17.

Tynan, F. E., S. R. Burrows, et al., (2005). "T cell receptor recognition of a 'super-bulged' major histocompatibility complex class I-bound peptide." Nat Immunol **6**(11): 1114-22.

Tynan, F. E., H. H. Reid, et al., (2007). "A T cell receptor flattens a bulged antigenic peptide presented by a major histocompatibility complex class I molecule." Nat Immunol **8**(3): 268-76.

Uversky, V. N., J. R. Gillespie, et al., (2000). "Why are "natively unfolded" proteins unstructured under physiologic conditions?" Proteins **41**(3): 415-27.

Vagin, A., Teplyakov, A. (1997). "MOLREP: an automated program for molecular replacement." J. Appl. Cryst. **30**: 1022-1025.

Valmori, D., V. Dutoit, et al., (2000). "Naturally occurring human lymphocyte antigen-A2 restricted CD8+ T-cell response to the cancer testis antigen NY-ESO-1 in melanoma patients." Cancer Res **60**(16): 4499-506.

Valmori, D., J. F. Fonteneau, et al., (1998). "Enhanced generation of specific tumor-reactive CTL in vitro by selected Melan-A/MART-1 immunodominant peptide analogues." J Immunol **160**(4): 1750-8.

Valmori, D., J. F. Fonteneau, et al., (1999). "Optimal activation of tumor-reactive T cells by selected antigenic peptide analogues." Int Immunol **11**(12): 1971-80.

Viola, A. and A. Lanzavecchia (1996). "T cell activation determined by T cell receptor number and tunable thresholds." Science **273**(5271): 104-6.

Viola, A., S. Schroeder, et al., (1999). "T lymphocyte costimulation mediated by reorganization of membrane microdomains." Science **283**(5402): 680-2.

Visintin, A., D. B. Iliev, et al., (2006). "Md-2." Immunobiology **211**(6-8): 437-47.

Vucetic, S., C. J. Brown, et al., (2003). "Flavors of protein disorder." Proteins **52**(4): 573-84.

Vucetic, S., Z. Obradovic, et al., (2005). "DisProt: a database of protein disorder." Bioinformatics **21**(1): 137-40.

Vucetic, S., P. Radivojac, et al., (2001). "Methods for improving protein disorder prediction." Intl Joint INNS-IEEE Conf. Neural Networks **4**: 2718-2723.

Walter, T. S., J. M. Diprose, et al., (2005). "A procedure for setting up high-throughput nanolitre crystallization experiments. Crystallization workflow for initial screening, automated storage, imaging and optimization." Acta Crystallogr D Biol Crystallogr **61**(Pt 6): 651-7.

Wang, J. H., R. Meijers, et al., (2001). "Crystal structure of the human CD4 N-terminal two-domain fragment complexed to a class II MHC molecule." Proc Natl Acad Sci U S A **98**(19): 10799-804.

Wang, J. H. and E. L. Reinherz (2002). "Structural basis of T cell recognition of peptides bound to MHC molecules." Mol Immunol **38**(14): 1039-49.

Wang, J. H., Y. W. Yan, et al., (1990). "Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains." Nature **348**(6300): 411-8.

Wang, Z., R. Turner, et al., (2002). "MHC allele-specific molecular features determine peptide/HLA-A2 conformations that are recognized by HLA-A2-restricted T cell receptors." J Immunol **169**(6): 3146-54.

Ward, J. J., J. S. Sodhi, et al., (2004). "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life." J Mol Biol **337**(3): 635-45.

Webb, A. I., M. A. Dunstone, et al., (2004). "Functional and structural characteristics of NY-ESO-1-related HLA A2-restricted epitopes and the design of a novel immunogenic analogue." J Biol Chem **279**(22): 23438-46.

Weinreb, P. H., W. Zhen, et al., (1996). "NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded." Biochemistry **35**(43): 13709-15.

Wu, H., P. D. Kwong, et al., (1997). "Dimeric association and segmental variability in the structure of human CD4." Nature **387**(6632): 527-30.

Wu, L. C., D. S. Tuot, et al., (2002). "Two-step binding mechanism for T-cell receptor recognition of peptide MHC." Nature **418**(6897): 552-6.

Xiang, S., G. Usunow, et al., (2007). "Crystal structure of 1-deoxy-D-xylulose 5-phosphate synthase, a crucial enzyme for isoprenoids biosynthesis." J Biol Chem **282**(4): 2676-82.

Xie, Q., G. E. Arnold, et al., (1998). "The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder." Genome Inform Ser Workshop Genome Inform **9**: 193-200.

Xu, Y., X. Tao, et al., (2000). "Structural basis for signal transduction by the Toll/interleukin-1 receptor domains." Nature **408**(6808): 111-5.

Yang, Z. R. and R. Thomson (2005). "Bio-basis function neural network for prediction of protease cleavage sites in proteins." IEEE Trans Neural Netw **16**(1): 263-74.

Yang, Z. R., R. Thomson, et al., (2005). "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins." Bioinformatics **21**(16): 3369-76.

Yarovinsky, F., D. Zhang, et al., (2005). "TLR11 activation of dendritic cells by a protozoan profilin-like protein." Science **308**(5728): 1626-9.

Yewdell, J. W. and J. R. Bennink (2001). "Cut and trim: generating MHC class I peptide ligands." Curr Opin Immunol **13**(1): 13-8.

Zernich, D., A. W. Purcell, et al., (2004). "Natural HLA class I polymorphism controls the pathway of antigen presentation and susceptibility to viral evasion." J Exp Med **200**(1): 13-24.

Zhang, D., G. Zhang, et al., (2004). "A toll-like receptor that prevents infection by uropathogenic bacteria." Science **303**(5663): 1522-6.

Zinkernagel, R. M. and P. C. Doherty (1974). "Immunological surveillance against altered self components by sensitised T lymphocytes in lymphocytic choriomeningitis." Nature **251**(5475): 547-8.

Zinkernagel, R. M. and P. C. Doherty (1974). "Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system." Nature **248**(450): 701-2.

Zinkernagel, R. M. and P. C. Doherty (1997). "The discovery of MHC restriction." Immunol Today **18**(1): 14-7.