Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Design, optimisation and functional relevance of *in silico* tools for the study of coeliac disease

A thesis submitted for the degree of Doctor of Philosophy

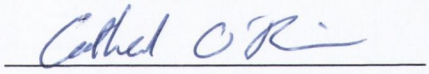
By

Cathal O'Brien, BSc. (Hons), PgDip

Declaration

I declare that this thesis is my own work and has not been submitted for any degree at this or any other university and except where otherwise stated; it is entirely my own work.

I agree that the library of the University of Dublin, Trinity College may lend or copy this thesis on request



Cathal O'Brien

Summary

Peptide-MHC interaction is a crucial pre-requisite for the recognition of antigen by T cells. It has long been recognised that peptides with a high affinity for MHC molecules are more likely to elicit a pronounced T cell response. In order to reduce the experimental burden in studies of immunogenic protein fragments, many researchers have developed methods to predict the affinity of a peptide for a chosen MHC molecule. It was elected to develop such methods for the prediction of peptide binding to the coeliac disease associated MHC class II molecule HLA-DQ2.

The selected approach necessitated the initial creation and assessment of three different affinity prediction algorithms. The three algorithms were based on different statistical methods i.e. Gribskov's profile analysis (GPA), hidden Markov models (HMMs) and quantitative structure activity relationship (QSAR). This comparison demonstrated that each algorithm has individual strengths and weaknesses. However, it was recognised that when the GPA algorithm was re-optimised to use a chemical similarity matrix, it was particularly advantageous, especially when the number of available data was low. Alternatively, the QSAR based algorithm generally performed better with larger datasets, especially if the training dataset was composed of substituted analogues or contained a large number of medium to low affinity data. Accordingly, it was suggested that the choice of appropriate algorithm should be determined after examination of the characteristics of the training dataset.

The relationship between peptide length and MHC affinity was subsequently investigated. A new model of peptide MHC class II interaction was also proposed and tested. The results of our *in silico* analysis clearly demonstrated that, within limits, the elongation of a peptide can markedly increase affinity for MHC class II molecules. It was proposed that the mechanism for this effect may be cooperative binding from an increased number of binding registers created as a result of

peptide elongation. This proposal was tested but not found to be any more appropriate than the current 'single binding register' model.
The results of these analyses mean that the length of peptides in training and test sets should be controlled for and recognised when deriving models of peptide-MHC class II interaction.

The results of the previous analyses were used as the basis for an original computational resource for screening prolamins for potential HLA-DQ2 dependent immunogenicity. The resource merged information available from a novel HLA-DQ2 binding motif with current knowledge of tTG mediated deamidation and proline dependence. Validation of the predictive abilities of the resource showed it to be quite capable of identifying characterised epitopes. The resource was then used to screen a dataset of gliadins and gliadin-like prolamins. The results of the analysis confirmed the likely antigenicity of characterised immunogenic proteins in addition to highlighting the potential immunogenicity of lesser studied prolamins such as the omega gliadins. The implemented resource should prove useful in future studies of prolamin immunogenicity and the development of a 'coeliac-friendly' strain of bread wheat.

Table of Contents

For my Mam and Dad

Acknowledgements

To my family I owe the greatest debt, for always being there and supporting my protracted stint in third level education. I owe a debt of gratitude to my parents, Pat and Margaret; my sisters, Catriona and Sinead; my younger brothers, Patrick and William and my big brother Owen 'Fonz' O'Brien who consistently managed to drag me to the pub over the last few years.

To Professor Conleth Feighery ('the boss'), I extend my most sincere gratitude. I could not have hoped for a better or more amicable supervisor and mentor. Many thanks also to Darren Flower for critical reading of the manuscript, and Declan Gasparro for running plates on the AutoDelfia®.

Thanks to my friends and colleagues in the coeliac disease research group: Bashir, Big Al, Christian, Jean, Mohamed, Sarah, Sharon, Stacey and Suzanne. Thanks also to everyone in the diagnostic immunology lab for fostering my initial interest in immunology, and everyone in DIT for the helpful and constructive advice.

Cheers also to Greg and Shane for coffee, beers and general tomfoolery.

And Melissa; you accept me just as I am, and support me in everything I do. I am truly lucky.

Communications from this thesis

Journal articles in preparation

- On the nature of peptide-MHC class II affinity: Empirical dependence on peptide length. O'Brien C, Flower DR, Feighery C
- Enhanced profile based prediction of peptide-MHC binding: Optimally selected amino acid indices can significantly improve prediction accuracy. O'Brien C, Flower DR, Feighery C

Oral Presentations

- • A comparative analysis of quantitative peptide-MHC binding predictions. O'Brien C, Feighery C
  - *International Summer School on Mathematical Modelling of Biological Function*. Bremen, 2005
- • Peptide length markedly influences peptide affinity for MHC class II molecules. O'Brien C, Flower DR, Feighery C
  - *Annual meeting of the Irish Society for Immunology, DCU, 2007*

Poster Presentations

- • Implementation of an *in silico* screening resource for characterisation of immunogenicity in coeliac disease associated prolamin proteins. O'Brien C, Flower DR, Feighery C
  - *Annual meeting of the Irish Society for Immunology*, DCU 2007
- • A QSAR model of peptide binding to the coeliac disease associated molecule HLA-DQ2. O'Brien C, Flower DR, Feighery C.
  - *XIIth International Coeliac Disease Symposium*, New York, 2006
- • Class II MHC-Peptide affinity can be directly linked to peptide length. Implications for Coeliac disease and its associated pathogenic peptides.
  - *XIIth International Coeliac Disease Symposium*, New York, 2006
- • Computational prediction of peptide-MHC affinity by Quantitative Structure Activity Relationship (QSAR).
  - *Annual meeting of the Irish Society for Immunology*, Croke Park, 2005

List of Abbreviations

ANN	Artificial Neural Network
APC	Antigen Presenting Cell
BCR	B Cell Receptor
CDR3	Complementarity Determining Region 3
ER	Endoplasmic Reticulum
ERAAP	Endoplasmic Reticulum Aminopeptidase associated with Antigen Processing
GPA	Gribskov's Profile Analysis
HLA	Human Leukocyte Antigen
HMM	Hidden Markov Model
IBS	Independent Binding of Side Chains
$-\ln IC_{50}$	The negative logarithm of the observed IC_{50} value
MBR	Multiple Binding Register
MHC	Major Histocompatability Comples
ORF	Open Reading Frame
PLS	Partial Least Squares
PFR	Peptide Flanking Regions
QSAR	Quantitative Structure Activity Relationship
SBR	Single Binding Register
TAP1	Transporter Associated with Antigen Processing 1
TAP2	Transporter Associated with Antigen Processing 2
TCR	T Cell Receptor
TLR4	Toll Like Receptor 4
tTG	Tissue Transglutaminase

1 Introduction

1.1 Evolution and genomic organisation of the MHC

The human MHC, which is one of the best characterised immunologically relevant regions of the genome, is located on the short arm of chromosome 6 and spans 3.6Mbp. The murine MHC (denoted H2) is located on chromosome 17 (Kumanovics et al., 2003). The MHC is known to be the most gene dense region of the human genome and codes for over 260 genes in the MHC and extended MHC regions (MHC Sequencing Consortium, 1999). Additionally, it codes many of the human genome's most polymorphic proteins (Trowsdale and Parham, 2004). The classical MHC can be divided into three main regions; the class I and class II regions contain the genes encoding the class I and II molecules respectively, located between these regions is the class III region which encodes a combination of immune and non-immune genes (Kumanovics et al., 2003). Outside of the classical MHC regions lie the extended MHC regions known as the extended MHC class I and extended MHC class II, so named, because of their proximity to the classical class I and class II regions respectively (Roitt and Delves, 2001). Different subsections of the MHC contain differing proportions of immune related genes; for instance, the class II region only codes for immune related genes whereas the extended class II region codes for non-immune related genes, the one exception being the Ring3 gene (MHC Sequencing Consortium, 1999). Interestingly, while the TAP1 and TAP2 proteins which are involved in MHC class I antigen presentation, are encoded within the MHC they do not localise to the class I region; instead, they are encoded within the class II region. Furthermore, tapasin, which is also involved in MHC class I mediated antigen presentation, is encoded by the extended class II region (Kumanovics et al., 2003). The distributed nature of genes involved in singular processes serves to illustrate the complex control mechanisms governing the expression of MHC proteins.

The acquired immune system - of which the MHC is a component - is thought to have evolved after the jawed/jawless vertebrate split (Danchin et al., 2004). This theory is strengthened by a lack of an acquired immune system in all species besides the jawed vertebrates and a failure to identify key molecules of the acquired immune response at the molecular level in species lacking an

acquired immune response. A study by Trowsdale showed that among the jawed vertebrates, the MHC region was quite stable throughout evolution, with all species analysed possessing genes coding for both MHC class I and MHC class II molecules. Additionally, across all species examined, the MHC class II molecules retained a similar structure-function relationship i.e. a two subunit with a polymorphic binding groove (Trowsdale, 1995). The evolution of the MHC is also subject to selection, and although the processes responsible for this selection have not been identified, the maintenance of diversity has been suggested to occur as a result of parasitic infection and mate choice (Piertney and Oliver, 2006).

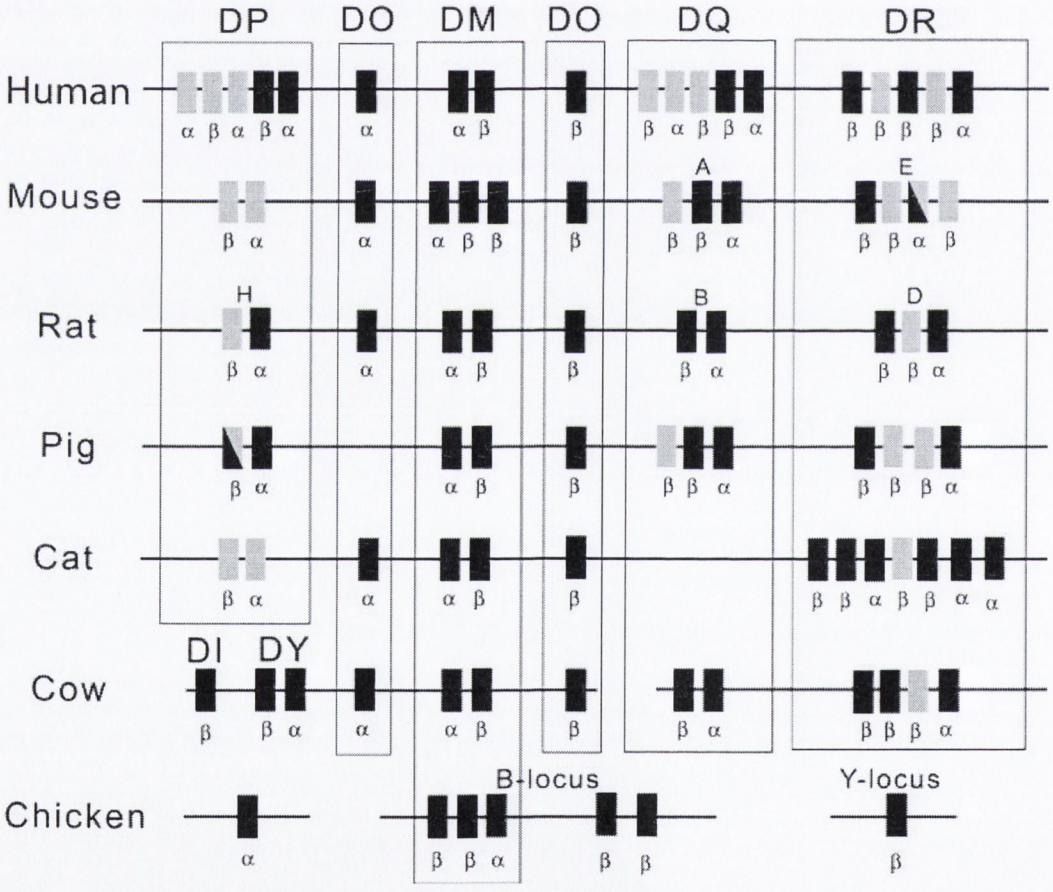


Figure 1-1 Interspecies distribution of MHC class II genes.

Distribution of MHC class II genes across a number of species (Kumanovics et al., 2003). Genes are coded in black, pseudogenes in grey; the alpha subunit of the murine IE molecules is inactive in approximately half of the inbred and wild strains. The class II groups evident throughout the mammalian species are not present in non-mammals such as the chicken.

Although it is quite well conserved across a broad range of species from humans to sharks, the MHC does display many inter-species differences. For example, the human MHC has a number of gene duplications not observed in the mouse (Figure 1-1). When compared with the mouse, humans show a remarkable difference in expression of the various MHC class II subtypes e.g. in over half of the characterised mouse strains, expression of the HLA-DR and –DP orthologs has not been observed (Kumanovics et al., 2003). However, the mouse is thought to make up for this lack of diversity by encoding over 30 different MHC class I molecules, which is a considerable repertoire considering that there are only 8 MHC class I molecules recognised in humans (HLA-A, -B, -C, -E, -F, -G, MICA and MICB).

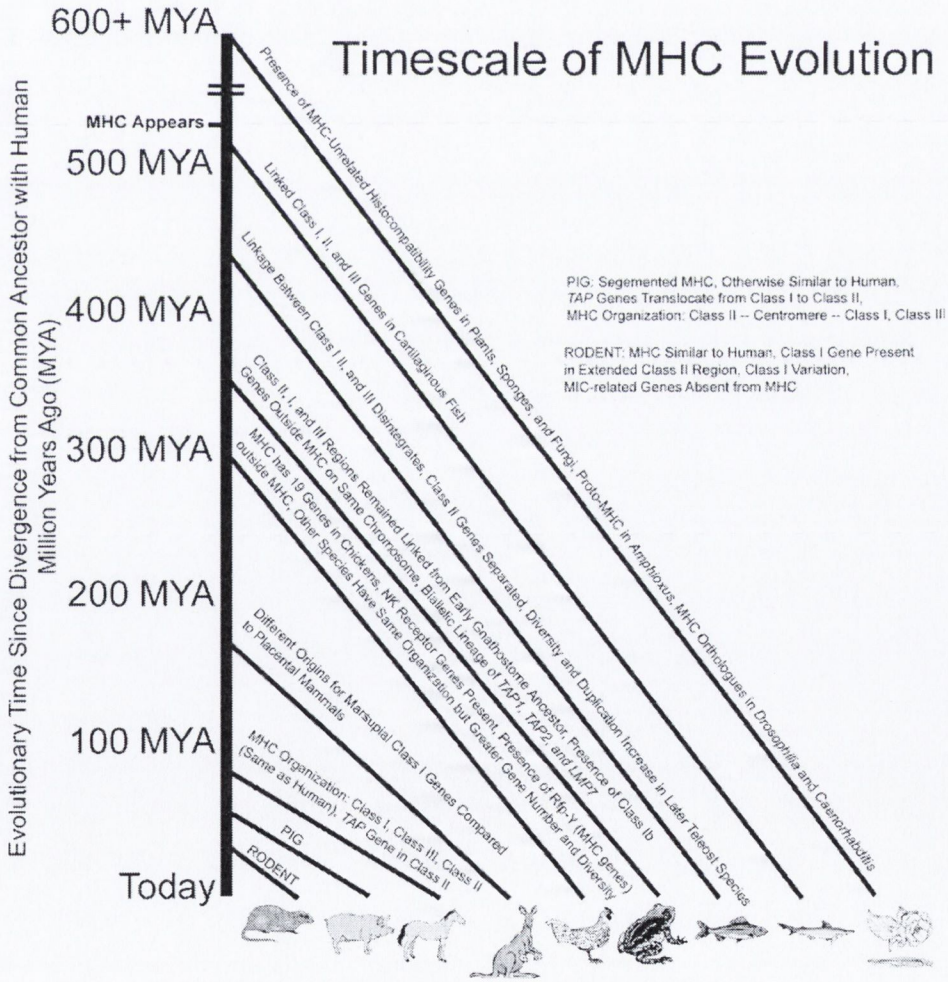


Figure 1-2 Timescale of MHC evolution

A timescale of the evolution of the MHC region as determined from MHC region sequences from a range of species (Kelley et al., 2005).

1.2 Classical MHC molecules bind antigen for recognition by T cells

1.2.1 Structure of the classical MHC molecules

The classical MHC molecules are broadly divided into two groups. These groups are denoted class I and class II and contain variable numbers of alleles, all of which are capable of binding antigen for presentation. The MHC class I molecules are comprised of a single polymorphic chain coupled to a single β2 microglobulin protein. From Figure 1-3 we can see that the β2 microglobulin protein does not form part of the peptide binding cleft/groove. From this observation it can be inferred that the peptide binding specificity of a given class I molecules is determined by its polymorphic α chain. In contrast, the MHC class II molecules are composed of two polymorphic subunits an α and a β chain. The β chain is in general the most polymorphic of the two and this is especially true for the HLA-DR molecule as the α chain is thought to be monomorphic. Thus, for HLA-DR, differences in peptide binding specificity result exclusively from differences in the alleles coding the β chain. This also helps explain the almost absolute requirement for a bulky hydrophobic residue at position p1 for the majority of characterised HLA-DR binding motifs, as the binding pocket at p1 is formed by the α chain.

As can also be seen from Figure 1-3 the binding grooves of MHC class I and MHC class II differ somewhat. In MHC class I the α-chain forms the entirety of the binding groove – a stark contrast to the MHC class II proteins, for which both subunits contribute an approximate equal amount to the binding groove. For both classes the groove is formed by a beta pleated sheet bordered by two antiparallel alpha helices. A notable difference between the two classes however, is the open-ended groove of the MHC class II molecules compared to the more restricted MHC class I. This structural feature which was initially observed using X-ray crystallographic studies, has been used to explain the observed differences in peptide length accommodated by the two classes (Jardetzky et al., 1994, Stern et al., 1994, Dessen et al., 1997). That is, the class I molecules can typically accommodate peptides of ten residues or less, whereas the MHC class II molecules can accommodate much larger peptides.

This feature of MHC class II is due to its open ended binding groove which allows residues not contained within the central region to extend out of the groove.

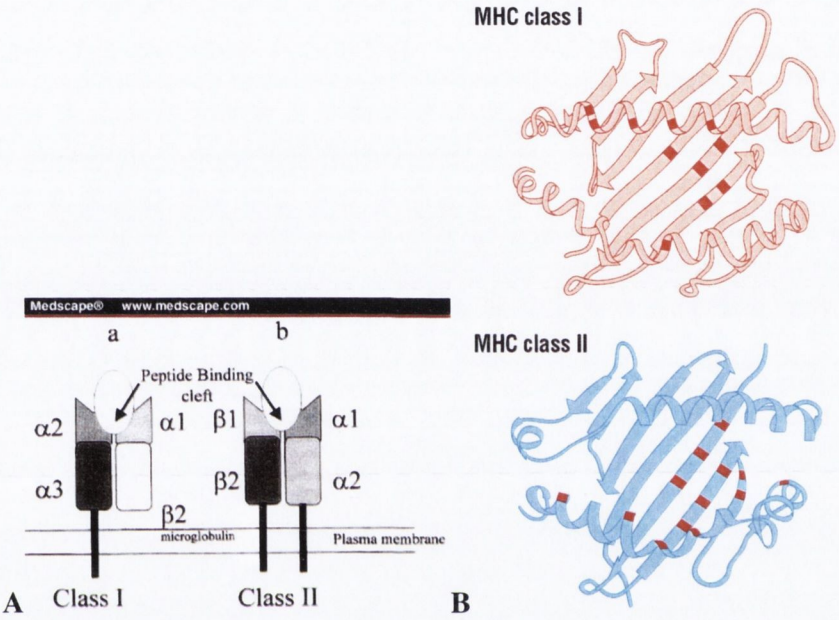


Figure 1-3 Structure and distribution of polymorphisms on the classical MHC molecules
Class I molecules composed of a polymorphic α chain coupled a single β2-microglobulin molecule, class II molecules are composed of two chains α and β both of which can be polymorphic (**A**) (adapted from www.medscape.com). The distribution of polymorphisms across the two classes of MHC molecules is also shown ((**B**), (DeFranco et al., 2007))

The first X-ray crystallographic structures of the MHC class I and MHC class II molecules pointed to different mechanisms for antigen binding. Initially studies by Falk *et al* demonstrated that MHC class I alleles exhibited different peptide binding motifs requiring a restricted set of residues at particular locations in the peptide binding groove. (Falk et al., 1991).Examination of the crystal structures of a number of peptide MHC class I complexes allowed the delineation of a common binding mechanism for the class I molecules. This mechanism demonstrated a network of hydrogen bonds interacting with the peptide amino and carboxy termini, while the peptide binding specificity was provided by the interaction between the peptide side chains and the polymorphic 'pockets' in the binding groove (Madden, 1995). Stern *et al*

carried out an analysis of an influenza derived peptide complexed with HLA-DR1 and found that peptide bound in an extended conformation with five binding pockets engaging peptide side chains while the flanking regions extended out of the binding groove (Stern et al., 1994).

1.2.2 Antigen presentation by MHC class I molecules

The differing roles of MHC class I and MHC class II in antigen presentation make even greater sense when one considers that each molecule locates antigen through a separate pathway. To begin, MHC class I molecules present antigenic peptides to CD8+ T cells. CD8+ (cytotoxic) T cells which recognise and respond to such antigen typically induce cytolysis or apoptosis in the host cell. As cytotoxic T cells are known to provide protection against viral infection and tumour development, it stands to reason that the cytotoxic T cell response must have evolved to detect peptide fragments endogenously produced by the host cell. This logic is supported by the well established finding that MHC class I are expressed by practically all nucleated cells in mammals. Additionally, such cells also produce cellular machinery with which to present such endogenously produced peptides in the context of MHC class I.
In order to be presented as peptide fragments of the appropriate length, in the context of MHC class I, cytosolic proteins must first be degraded. The best characterised mechanism by which proteins are degraded is the ubiquitin-proteasome pathway which targets proteins for proteolysis by tagging lysine residues with ubiquitin protein (Ciechanover, 1994). The link between proteasomal degradation and MHC class I epitope presentation was initially confirmed when it was demonstrated that the use of proteasome inhibitors could prevent degradation of proteins and presentation of MHC class I bound peptides from *Listeria monocytogenes* (Sijts et al., 1996). Subsequently, a more thorough understanding of the peptide presentation machinery was developed, and the tight interconnection between MHC molecules and peptide presentation was characterised.
After synthesis and folding of the MHC class I heavy chain, it binds to β2-microglobulin and is incorporated into the aptly named peptide-loading complex, a complex of proteins which also includes the transporter associated with antigen processing (TAP1 and TAP2) in addition to tapasin, calreticulin

and ERp57 (Cresswell et al., 2005). Cytosolic peptides are then transferred into the endoplasmic reticulum via TAP for loading into the peptide binding groove. For the majority of naturally processed peptides, trimming by the endoplasmic reticulum aminopeptidase associated with antigen processing (ERAAP) is necessary to ensure formation of viable peptide-MHC class I complexes (Hammer et al., 2006). At this stage of processing, as both peptide and MHC class I are present at relatively high concentrations and in close proximity, competition for MHC class I binding is likely to occur between peptides. Presumably, peptides with higher affinity for particular MHC class I alleles and/or a greater molar concentration would have the advantage and thus bind a greater proportion of MHC class I. Once a stable peptide-MHC class I complex is formed it is transported to the cell surface for possible recognition by cytotoxic T cells.

1.2.3 Antigen presentation by MHC class II molecules

In stark contrast to cytotoxic T cells, helper T cells recognise exogenous antigen presented by professional antigen presenting cells in the context of MHC class II. For this reason, the immune system has evolved a separate pathway for acquisition and presentation of antigen by MHC class II. Typically, extracellular antigens are internalised by APCs via endocytosis or macropinocytosis before being digested in endosomes. As part of a more specific mechanism, antigen can also be taken up by receptor mediated phagocytosis and subsequently degraded in the acidic phagosomes. Endogenous proteins may also end up in endosomes where they are digested in the same way as their exogenous counterparts (Robinson and Delvig, 2002).

Much like MHC class I molecules, MHC class II molecules are assembled in the ER prior to being loaded with peptides from degraded proteins. However, MHC class II proteins are initially transported to the golgi, complexed with the invariant chain and loaded into endosomes and phagosomes. When loading of peptide is to take place, HLA-DM within the endosome/phagosome initiates the removal of the class II invariant chain peptide (CLIP) from the binding groove and the ensuing loading of the antigenic peptide. Interestingly, the fact that the HLA-DM orthologs are conserved across all jawed vertebrates so far analysed serves to illustrate their importance in peptide loading (Figure 1-1).

Following binding, peptide-MHC class II complexes are transferred to the outer cell membrane for presentation to helper T cells.

1.2.4 Non classical modes of antigen presentation

In addition to the classical mechanisms underlying antigen presentation by MHC, additional means exist by which antigens can be presented to effector lymphocytes. Cross-presentation is the process by which APCs present endogenous antigen to cytotoxic T cells in the context of MHC class I and although the underlying mechanism is only recently being heavily investigated, its existence and biological relevance is well established (Basta and Alatery, 2007). Additionally, the MHC class II molecules are not confined to presentation of exogenous antigen and presentation of endogenous antigen in the context of MHC class II has also been well described in the literature (Robinson and Delvig, 2002). A subset of the MHC class I proteins (denoted class Ib) has also been categorised which binds non-traditional ligands. A traditional example of the class Ib molecules is CD1 which can present lipid antigens to T cells (Rodgers and Cook, 2005). Thus, while the classical MHC molecules form the basis of what is arguably the most biologically significant role encoded by the MHC, a great number of separate, albeit less well characterised, mechanisms of antigen presentation exist to extend the scope of the immune response.

1.2.5 Recognition of the antigen-MHC complex by T cells

The cellular distribution of the different classes of MHC molecules, in addition to their unique structural features, makes them suitable for recognition by different types of T cells. A dogmatic aspect of T cell-MHC interaction is the class restriction imposed by the CD4 and CD8 molecules. That is, in order for cognate recognition of a peptide-MHC complex by a T cell, the T cell receptor must not only successfully engage peptide-MHC, but a secondary T cell molecule must also engage the MHC molecule in a peptide independent manner. The secondary molecule harbours specificity for the class of MHC molecule with which the T cell must interact; CD4 molecules show a singular

specificity for MHC class II whereas the CD8 molecules found on cytotoxic T cells are specific for the MHC class I proteins (Figure 1-4). This interaction of specific T cell subsets with select classes of MHC molecules may also partially explain their function. Dogmatically, cytotoxic T cells attempt to recognise viral and tumour antigens which may potentially originate in any nucleated cell, hence MHC class I enjoys an almost ubiquitous expression. In contrast, the MHC class II molecules are predominantly expressed on professional antigen presenting cells (APCs) which can engulf pathogens and present exogenously derived antigens to CD4+ T helper cells.

The interaction of the TCR with the peptide MHC complex has been characterised in a number of X-ray crystallographic studies. Teng *et al* reported on a common docking motif present in the crystal structures of MHC class I molecules using both murine and human models. Their analysis pointed to a 'diagonal' docking motif in which the TCR Vα region overlayed the MHC α2 helix and the Vβ region overlayed the MHC α1 helix (Teng et al., 1998). Characterisation of the first TCR-peptide-MHC class II structure by Reinherz *et al* furthered the understanding of this interaction using a similar approach (Reinherz et al., 1999). Somewhat surprisingly, the authors of this study found the TCR docking mode to be more orthogonal than diagonal (Figure 1-4). Hennecke *et al* furthered this understanding by demonstrating that the specific regions (known as CDR3) of the TCR α and β chains seemed to align over the residue at p5 in the binding groove suggesting that the TCR may be more tolerant of changes at the peptide termini (Hennecke et al., 2000, Hennecke and Wiley, 2002). However, a separate study by De Oliveira *et al* demonstrated a dependence on residues other than p5 (De Oliveira et al., 2000), which would suggest the TCR-peptide-MHC class II interaction to be relatively heterogenous.

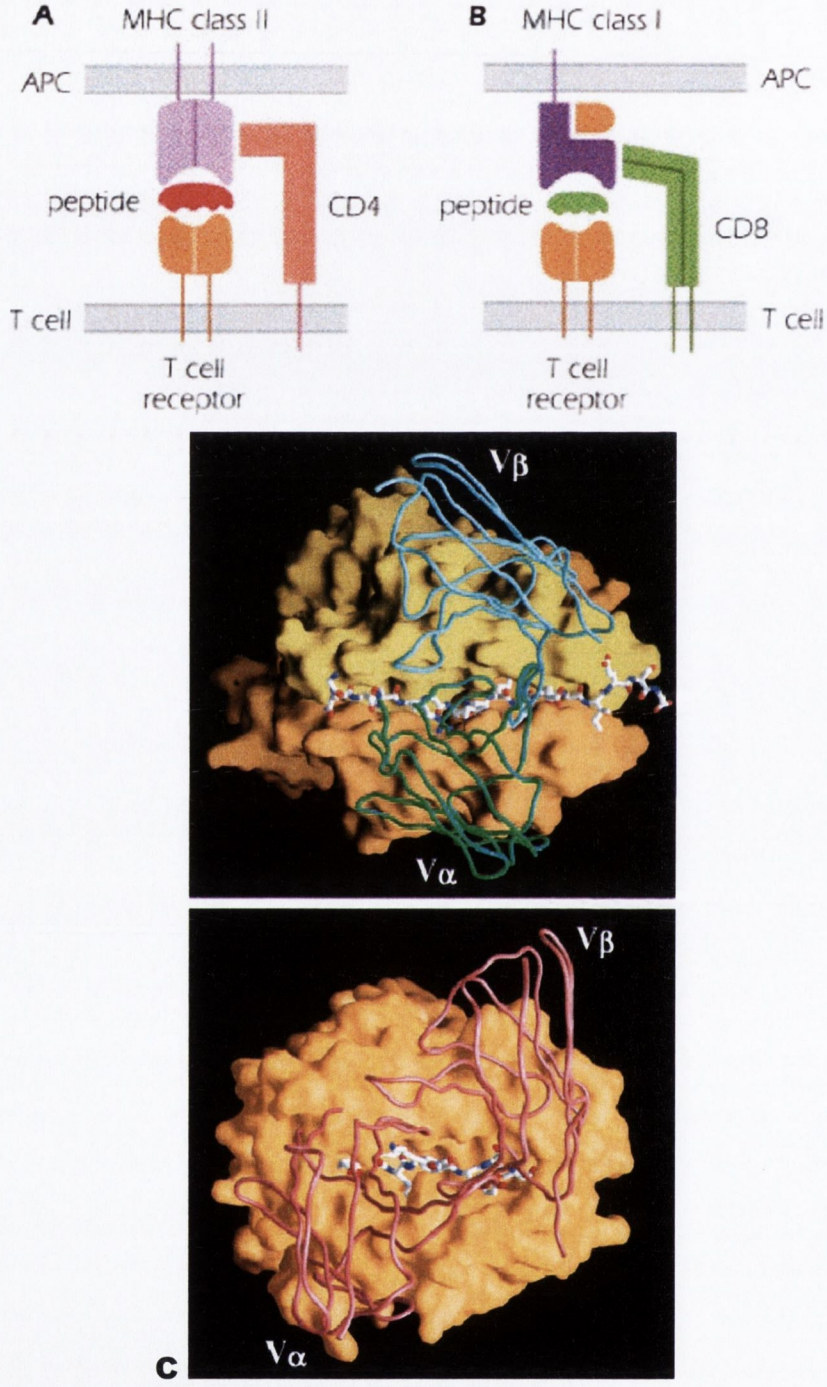


Figure 1-4 The peptide-MHC-TCR interaction differs between MHC molecule classes

(A, B) different MHC molecules require different 'co-receptors' located on the T cell surface for T cell activation (image from www2.hawaii.edu). (C) The TCR-peptide-MHC class I interface (lower frame) is more diagonal than the orthogonal interface observed for MHC class II (Reinherz et al., 1999)

1.3 MHC linked disease

1.3.1 Autoimmune disease

A great number of studies of autoimmune conditions in mouse models have provided compelling evidence to support the central role of T cells in autoimmunity. This finding is greater contextualised by subsequent risk assessment studies which have identified key MHC alleles as being associated with specific autoimmune conditions (Kuby, 1994). It follows from these findings that individuals capable of eliciting such an immune response must possess MHC molecules and T cell receptors compatible with presentation of pathogenic sequences. Thus, many disease association studies have demonstrated the association of certain autoimmune diseases with specific MHC alleles (Table 1-1). However, none of these diseases can be considered to be of Mendelian inheritance as possession of a HLA allele alone is insufficient to initiate the disease process. Epidemiological studies with identical twins have demonstrated that while one twin may develop an autoimmune disease, the other may not, indicating an environmental component to disease development.

Functional studies of MHC allele restricted T cell responses have also provided compelling evidence for the role of specific MHC molecules in autoimmune disease. For example, transgenic rats expressing the human HLA-B27 and B2M genes spontaneously develop a human spondyloarthropathy-like inflammatory disease. A separate experiment with transgenic mice expressing human HLA-DQ8 also revealed the spontaneous development of diabetes in transgenic versus wild type mice (Klein and Sato, 2000). Klein *et al* suggest that although the exact mechanisms underlying autoimmunity are unknown, reactivity to self antigens is maintained as low affinity binders can escape negative selection in the thymus. Activation of these self reactive T cells has been postulated to occur through molecular mimicry or an otherwise aberrant immune response (Kuby, 1994). However, there are currently no studies which can provide conclusive evidence in support of a single mechanism underlying the pathogenesis of any autoimmune process.

Disease	HLA Allele	Relative Risk
Ankylosing spondylitis	B27	90
Goodpasture's Syndrome	DR2	16
Graves' Disease	B8/DR3	3-4
Insulin-dependent diabetes mellitus	DR4/DR3 DR3/DQ8	20 100
Juvenile rheumatoid arthritis	B27/DR5	4
Multiple sclerosis	DR2	5
Myasthenia gravis	DR3	10
Pernicious anaemia	DR5	5
Psoriatic arthritis (central)	B27	11
Reiter's syndrome	B27	37
Rheumatoid arthritis	DR4	10
Systemic lupus erythmatosus	DR3	5
Ulcerative colitis	B5	4

Table 1-1 HLA alleles associated with increased risk for various autoimmune diseases

Adapted from Kuby (Kuby, 1994)

An understanding of linkage disequilibrium (LD) has also been instrumental in our evolved understanding of autoimmunity and MHC-linked disease. As mentioned above (section 1.1), the MHC represents a highly dense region of gene expression. This level of density makes the region more prone to linkage disequilibrium as a result of an increased likelihood of inheritance of haplotypic block. This LD resulted in a number of risk factor assessment studies initially attributing risk to MHC class I molecules, whereas, subsequent analysis demonstrated a greater link to MHC class II molecules (Kuby, 1994). It has also been demonstrated that apparent association of certain diseases with MHC molecules may occur as a result of LD; this was postulated to be the case for narcolepsy (See below, section 1.3.2).

1.3.2 MHC linked non-autoimmune disease

Linkage of non-autoimmune disease to MHC associated genes is not without precedent. Narcolepsy is a well characterised MHC linked disease being linked to the HLA-DQ locus in humans, although the exact role of the HLA-DQ molecules in the disease is ill-understood. It has been argued that the association with the MHC region may actually be masking the true disease causing gene through LD. For example, HCRTR2 is a gene associated with narcolepsy, a chronic and debilitating sleep disorder and has a demonstrable role in the disorder in both canines and mice. It is thought to be strongly over represented in tandem with the disease associated HLA-DQB1*0602 and DQA1*0102 alleles due to their close proximity within the genome. Thus, while initial disease causality may be attributed to the MHC alleles, it could easily be either the MHC genes or the HCRTR2 gene which actually cause the disease as all three alleles are inherited together (Klein and Sato, 2000).

An additional and unsurprising link has also been suggested between HLA genes and human infectious disease. Susceptibility to infectious human infection diseases such as leprosy and tuberculosis have been suggested to have a HLA-DR linked component. Progression of HIV to AIDS is thought to be delayed by interaction with a select subset of HLA-B alleles (Martin et al., 2002). Additionally, a protective effect from malaria infection has been shown to occur as a result of inheriting the HLA-B*53 or HLA-DRB1*1302 genes (Hill, 2006).

A tenuous classification of coeliac disease would be as a hypersensitivity disorder with an autoimmune component as opposed to a classical autoimmune disorder. However, among immunologically mediated MHC linked disease, it is among the best understood having crucial environmental and genetic components identified which exemplifies its role as a model MHC-linked immunological disorder.

1.4 Coeliac Disease

1.4.1 Coeliac disease - a well characterised inflammatory disorder

Coeliac disease, an inflammatory disease of the upper small intestine, is active only while those affected ingest gluten-containing foods such as wheat and rye, only to resolve completely with adherence to a gluten free diet. The well characterised inflammatory process is typified by duodenal villous atrophy and crypt cell hyperplasia. As this inflammatory damage occurs at a site crucial to absorption of nutrients, the knock-on effects of this inflammation –namely malabsorption– affect not only the small intestine but many other body systems too. A summary of this is presented below (Table 1.1). The range of clinical presentations in coeliac disease is quite broad and varies depending on the age of presentation and severity of the inflammation, with iron deficiency anaemia, now the most common presentation (Farrell and Kelly, 2002). The gold standard for identification of coeliac disease is still intestinal biopsy (Farrell and Kelly, 2002) which typically shows loss of absorptive villi and crypt cell hyperplasia (Dieterich et al., 1997).

Organ affected	Observable Effects
GIT	Diarrhoea, abdominal distension, loss of appetite, secondary lactose intolerance
Bone	Decreased bone density, osteoporosis, rickets
Muscle	Myopathy
Nervous system	Difficulty learning, irritability, fatigue
Blood	Anaemia, iron, folate or rarely Vit B_{12} deficiency
Dermatological	Dermatitis herpetiformis

Table 1-2 Effects of coeliac disease on body systems

Adapted from Feighery, C (Feighery, 1999)

Used as a model for the study of MHC linked disease, CD provides a number of advantages, most notably that of having crucial genetic and environmental factors identified. These factors make it ideal for studying the complex interaction between genes and environment. Environmentally, symptoms are exacerbated by the ingestion of gluten containing foods. This link was first made by the Dutch paediatrician Dicke, who, in 1950, noted that children suffering from coeliac disease showed an improvement in symptoms during the food shortages of World War Two only to relapse upon reintroduction of normal food supplies. The genetic link in coeliac disease is shown through family and twin studies. Firstly, family studies show that around 10% of first degree relatives of patients with confirmed coeliac disease are also positively diagnosed. Secondly, identical twin studies show concordance rates of between 60% and 70%. Both of these factors indicate a strong genetic link (Feighery, 1999). The best recognised and validated association in coeliac disease is with HLA-DQ2 (Sollid et al., 1989), and to a lesser extent HLA-DQ8 (Lundin et al., 1994). It is currently thought the approximately 95% of coeliac disease patients express HLA-DQ2, whereas the majority of the remainder express HLA_DQ8. It has also been demonstrated that a gene dosage effect may exist which would place those homozygous for HLA-DQ2 expression at greater risk of disease development (Vader et al., 2003b). Genome wide association studies have also identified a number of other possible contributory factors, including genes encoding cytotoxic T lymphocyte antigen 4 (CTLA4), the T helper 2 (T_H2) cytokine cluster, CD14 and Tim. The latter three of these functions have been implicated in numerous other hypersensitivity disorders (Sollid, 2002). More recently, linkage studies have highlighted putative associations with the genomic regions bordering the genes encoding IL2 and IL21 (van Heel et al., 2007).

A feature of CD that seems to fuel a lot of scientific dialogue is the actual mechanism by which gliadin induces intestinal inflammation in active coeliac disease. One cannot doubt that epithelial cells incur damage during the active stage of the disease. However, one must puzzle at the exact means through which this damage occurs. Firstly, it is well established that gliadin specific T cell clones cultured from intestinal biopsies of those with active CD exhibit a T_H1 type cytokine profile (Olaussen et al., 2002). This finding would suggest

that gliadin reactivity is not an example of an IgE mediated food allergy. Additionally, a T cell response to a HLA-A2 restricted epitope has been reported (Gianfrani et al., 2003), although the *in vivo* significance of this finding has yet to be determined. Intraepithelial lymphocytes may also be a cause of damage as shown in some studies (Ciccocioppo, 2000). In addition epithelial cells may act as antigen presenting cells (APCs) however, and present antigen in the context of MHC class II molecules which present exogenously acquired antigen. Exogenously acquired antigen such as gliadin may then be presented to T helper (T_H) cells in the lamina propria which can secrete cytokines in response to this stimulus. The fact that cytokines can affect the growth and differentiation of epithelial cells makes this a possible mechanism in the formation of the coeliac lesion. The finding of anti-tTG antibodies in almost all coeliac patients also proves a possible mechanism for lesion formation as these antibodies may inhibit tTG function (Dieterich et al., 1997). A study of this hypothesis by Byrne *et al* demonstrated the efficacy of coeliac disease associated IgA antibodies in inhibition of tTG function (Byrne et al., 2007). tTG function may be critical for mucosal integrity and therefore this arresting of function could lead to the characteristic villous atrophy (Griffin et al., 2002).

More recently, a putative role for an innate response in coeliac disease was suggested by Maiuri *et al* (Maiuri et al., 2003). The authors of this study found that a segment of a gamma gliadin (denoted p31-43) could induce T cell mediated inflammation when incubated with explant duodenal tissue. However, the authors found that this response was dependent on what appeared to be activation of cells of the innate immune system induced by a separate peptide derived from the same protein. These observations were specific to biopsies cultured from coeliac disease patients. This article gave rise to much speculation as to the significance of such inflammation in coeliac disease pathophysiology. A number of studies focussed on the induction of IL-15 production which could in turn stimulate NKG2D mediated epithelial cell killing (Stepniak and Koning, 2006). In spite of these findings, however, an innate receptor for segments of the gliadin protein has yet to be characterised. Given that the purported initial inflammation in coeliac disease is thought to be innate, yet specific to coeliac disease, identification of such a receptor would

allow the elucidation of the role of the innate immune response in coeliac disease. A separate study found a level of cross-reactivity to exist between anti-tTG antibodies and TLR4 in coeliac disease (Zanoni et al., 2006). The antibody mediated activation of monocytes may partly explain the activation of components of the classical immune response observed by Maiuri *et al.*

1.4.2 Autoantigens and serology in coeliac disease

As previously stated, coeliac patients are exquisitely sensitive to gluten. Studies have shown that the protein to which coeliacs are most sensitive is a protein called gliadin – the alcohol soluble component of gluten – and, that coeliac patients typically produce antibodies directed against gliadin. However, in addition to antibodies directed against this external antigen, many coeliac patients also show the presence of antibodies which react with endomysium of monkey oesophagus or human umbilical cord using indirect immunofluorescence (Chorzelski et al., 1984). This fact was, and still is, used in diagnosis through serological tests, namely indirect immunofluorescence testing for anti-endomysium antibodies. However, a search persisted to identify the exact protein against which these antibodies were directed, and in 1997, Dietrich *et al* identified this protein as tissue transglutaminase (tTG) (Dieterich et al., 1997, Chorzelski et al., 1984). Tissue transglutaminase (tTG) is one of a group of proteins known as transglutaminases, which catalyse transamidation reactions and exhibit variable tissue distributions and functions. The functions of transglutaminases in physiological systems vary with the type (six being recognised at present) from blood coagulation to maintenance of connective tissue integrity and induction of apoptosis (Griffin et al., 2002). The reactions catalysed by transglutaminases result in cross-linking of proteins hence their being dubbed 'biological glue' by Griffin *et al.* The most evident example being the active form of factor XIII which is involved in the stabilisation of fibrin clots. The ubiquitous tissue transglutaminase (tTG) or Type II transglutaminase is the target antigen of anti-tTG antibodies. tTG is controlled by its own set of controlling mechanisms and several transcriptional activators have been described including vitamin D, steroid hormones, and cytokines (Griffin et al., 2002), and it has been shown that it can activate transforming growth factor beta (TGF-β) (Dieterich et al., 1998). The importance of tTG has

yet to be fully elucidated but studies in knockout mice have shown little early side effects, the most notable complications being a near diabetic response to glucose overload, and reportedly slowed dermal wound healing which is probably due to changes in cell motility and cytoskeletal changes (Griffin et al., 2002). In addition to its numerous intracellular substrates, numerous extracellular substrates have also been recognised including collagen, fibronectin and importantly for coeliac disease, the exogenous wheat protein gliadin. In fact, a stronger link to coeliac disease was provided by the role of tTG mediated peptide deamidation in T cell epitope recognition (Arentz-Hansen et al., 2000b). Tissue transglutaminase although mainly confined to the intracellular compartment is released during tissue damage with the possible role of maintaining tissue integrity. It was postulated by Dietrich et al (1997) that the transglutaminases were released in response to gluten-induced cell damage whereupon they reacted with gliadin to form gliadin-gliadin or gliadin-tTG complexes. The gliadin-tTG complexes then provide a means for building a host humoral immune response against self-tTG, as explained below (section 1.4.3)

1.4.3 Auto-antibody production in coeliac disease

As mentioned previously antibodies to tTG are a diagnostic serological marker for coeliac disease. In order for these antibodies to be produced a B cell specific for tTG must first engage the antigen with its B cell receptor (BCR) internalise it and present processed peptide fragments to T helper cells in the context of its MHC class II molecules. Upon recognising the foreign peptide the T helper cell through CD40L may activate the B cell and depending on the cytokine milieu to which the B cell is exposed, the result may be production of antibody, heavy chain class switch, or suppression of the B cell response. However, B cells are not necessarily confined to presenting the antigen that they have bound by their BCR through their MHC molecule. Therefore, it may be assumed that any non-self molecule presented to a pro-inflammatory T cell of a given antigen specificity interacting with a B cell may induce an active response. This active response may therefore be induced against a self-antigen by the T helper cell as long as its first encounter with that peptide sequence was in a pro-inflammatory state. As mentioned above, this mechanism was

postulated by Sollid as a possible mechainism for inducing the production of antibodies against self-tTG (Sollid et al., 1997). Although, the exact role of the anti-tTG antibody immune response in coeliac disease is still debated, some studies have provided evidence to suggest an ability to alter normal function by tTG inhibition (Byrne et al., 2007) and TLR4 ligation (Zanoni et al., 2006), as mentioned above.

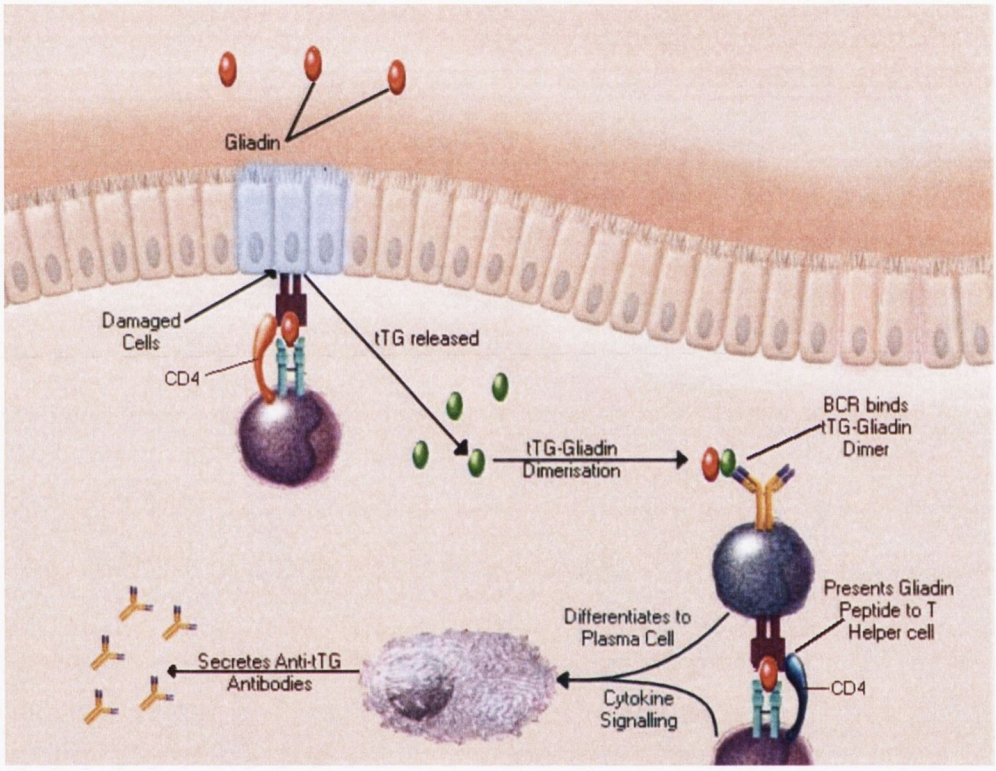


Figure 1-5: Model of gliadin induced tTG reactivity

1.5 Aim

Given the well characterised nature of coeliac disease as an MHC class II dependent disorder it is considered that the optimal alternative to the traditional gluten free diet would need to exploit the interaction of gliadin peptides with the disease associated MHC molecules. One approach would be to block interaction of disease associated peptides with MHC class II molecules using blockers. A separate approach would involve the identification of wheat strains in which the gliadin and glutenin peptides would be minimally antigenic. These minimally antigenic strains of wheat could then form the basis for the creation of a suitable strain of wheat for bread manufacture, yet lacking sequences capable of triggering the disease. As the cost and time involved in testing many of the available strains of wheat for immunogenicity would be quite considerable, we proposed that computational scanning of gliadin proteins would reduce this experimental burden significantly. Thus, it was elected to design and implement a screening resource which could be made publicly available for such a purpose. However, in order to create such a resource it was necessary to first create a model of peptide-HLA-DQ2 binding. This particular step necessitated the evaluation and optimisation of a number of binding affinity prediction algorithms to identify the optimal algorithm for creation of our binding model. It was also necessary to evaluate the contribution of peptide length to MHC affinity as we hypothesised that this could drastically affect the performance of peptide-MHC class II affinity prediction. Accordingly, our main goals were as follows:

- Evaluate and optimise for the task of quantitative epitope prediction, currently established algorithms.
- Establish the impact of peptide length on reported MHC class II affinity and if possible ascertain a mechanism by which any observed effect might occur.
- Create and evaluate the efficacy of a resource for the screening of prolamins for potentially immunogenic sequences. This step would necessitate the creation of a HLA-DQ2 binding model and generation of additional HLA-DQ2 binding data to supplement the currently available data.

2 Implementation and evaluation of tools for the quantitative prediction of peptide MHC interaction

2.1 Introduction

2.1.1 Bioinformatics – a 21st century tool for 21st century biological research

The volume of articles returned when using the word 'bioinformatics' in a standard pubmed search is astounding, at over fifteen thousand articles of which over 2,500 are review articles. Bioinformatics as a word can convey different meanings to different types of researchers. Many evolutionary biologists would find it difficult to test any hypothesis without it, while molecular biologists may only ever need to run a single BLAST search to characterise a protein of interest. The sheer heterogeneity of computational methods used to analyse a plethora of biological and experimental data seems to defy an exact yet all-encompassing definition. However, authors do attempt to clarify the exact meaning of the term: one such article by Luscombe *et al* stated that "Bioinformatics is conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organise the information associated with these molecules, on a large-scale" (Luscombe et al., 2001). This definition by Luscombe is quite broad and, as such, includes a great deal of what many would call bioinformatics. One textbook by Mount (Mount, 2004) contains a varied description of topics from simple BLAST searching, to phylogenetic tree reconstruction and microarray analysis.

Prototypic tools such as BLAST (Altschul et al., 1990) lead the field in terms of publicly visible bioinformatics applications, being cited over 21,000 times (citation count from the ISI web of knowledge). The BLAST suite of programs are frequently used to search a nucleic acid or protein sequence database to locate proteins that display possible homology with a query protein. Other tools such as Clustal (Higgins and Sharp, 1988) and PHYLIP (Felsenstein, 1996) are frequently utilised for multiple sequence alignment and phylogenetic tree reconstruction respectively. In many cases, such tools have found their way into common molecular biology being commonly utilised for contextual

analysis of protein and nucleic acid sequences. In one such example, Duffy *et al* utilised a phylogenetic tree to visually represent the relationship between the ADAMS family of protein sequences to supplement their review on the topic (Duffy et al., 2003).

The increased number of sequenced genomes as represented on the Genomes On Line Database (GOLD) was over 500 as of April 2007 with over 2000 ongoing sequencing projects. This explosion of data has lead to an increased amount of bioinformatics analyses taking place at every stage, from assembling contigs to annotating the resultant open reading frames (ORFs). The availability of genome data has allowed the study of evolution on a hitherto impossible level and researchers have been able to research various stages of genomic evolution from gene loss and gain, to genome duplication (McLysaght et al., 2002, McLysaght et al., 2003). Analysis of genomic data from a variety of mammalian species permitted Lynn *et al* to uncover evidence for amino acids under positive selection in alpha-defensins, allowing the identification of sites with possible critical relevance for immune function (Lynn et al., 2004).

The use of the word bioinformatics has also been extended to tools used to analyse data from high content experiments such as microarray analysis using either graphical or scripted tools. The analysis of the vast quantities of data produced by such studies encompasses a broad range of tools, well known examples of which would be Bioconductor (Reimers and Carey, 2006) and Metacore® for contextual pathway analysis (Ekins et al., 2007).

2.1.2 Immunoinformatics - a 21st century tool for 21st century immunology

.

A pubmed search of the word immunoinformatics is practically dwarfed by the sheer number of articles tagged with the word bioinformatics. Of the 37 articles which were classified under the term immunoinformatics, 10 were reviews and almost all of the remainder involved T-cell epitope or MHC binding prediction. These simple findings alone tell us a number of things about

immunoinformatics. Firstly, it is a subdivision of bioinformatics that is in its infancy. Secondly, to a great number of people immunoinformatics can mean only one thing – computational vaccine design, based on T and (occasionally) B cell epitope prediction. The first pubmed indexed article to contain the word immunoinformatics is by Segel (Segel, 2001) and describes the exchange of information between cells of the immune system. However, this would appear to be the only such usage as the next article to contain the word immunoinformatics is a review of computational vaccinology (Flower, 2003a).

One of the most recent reviews on the topic of Immunoinformatics by Korber *et al* provides an exhaustive description of the computational resources available (Korber et al., 2006). In the article Korber *et al* describe nearly 25 web-based tools relevant to T cell epitope prediction and 15 publicly available databases of immunologically relevant data. The sheer breadth of available epitope prediction tools is a reflection of the lack of a standard approach amongst researchers in the area. However, this lack of standardisation is quite common in newly developing fields where optimal approaches have yet to be defined.

In another recent review by De Groot, the discussion of immunoinformatics draws from a number of different sources of data and analysis tools in a succinct prediction of the future of epitope design (De Groot and Berzofsky, 2004). The proposed methodologies which would be used to develop vaccines for future studies would incorporate genomic data from species selected as being potential causes of future epidemics or pandemics. This genomic data could then be utilised to preferentially target proteins from virulent strains of bacteria while minimising self reactivity and possibly incorporating gene expression study data to further refine the vaccine design process. One such example is outlined in another article by De Groot (De Groot and Rappuoli, 2004) where the open reading frames (ORFs) common to H. Pylori and a variety of common bacteria are iteratively depleted to provide a study set of H. Pylori specific genes. This overlap between immunoinformatics and traditional bioinformatics in the form of genomics has been termed immunomics (Brusic,

2003). Being more systems oriented, immunomics draws its tools from a variety of disciplines, one of which is immunoinformatics.

Possibly the best recognised single resource available to computational immunologists at present is the IMGT - the ImMunoGeneTics information system (Lefranc et al., 1995). This online resource, like many others in immunoinformatics is centered on the principal molecules of the adaptive immune response i.e. MHC, BCR and TCR. The IMGT system contains both databases and analysis tools with particular focus on the molecules themselves – as opposed to their interacting ligands. The systems analysis tools allow scientists integrate the available sequence data with current genomic knowledge in addition to structural data and are widely used by researchers worldwide (Lefranc, 2005). Of particular interest in the approach of the IMGT researchers is their use of a standardised system for curating data IMGT-ONTOLOGY (Giudicelli and Lefranc, 1999), in addition to a markup language designed specifically for the IMGT system (Lefranc, 2005). As a model immunome research resource, it is quite formidable and projects such as this are likely to form the backbone of future immunome research.

As the uptake of bioinformatics analyses in immunology increases, it is likely that the usage of computational techniques will increase and expand into more novel applications. For example, the use of microarray expression studies has been suggested as a novel method for classifying T cell responses (Hyatt et al., 2006), and mathematical and theoretical modelling are also beginning to become more utilised tools in immunological investigation (Chakraborty et al., 2003, Perelson, 2002). Additionally, the integration of epitope prediction software and microarray DNA expression studies has been reported (Sturniolo et al., 1999) which represents a significant merging of computational biology with the vaccine design process.

2.1.3 Epitope Prediction

Some of the earliest utilised predictors of antigenicity were based on general predictors of sequence property such as those implemented using amino acid

property indices. However, such general measures of sequence property were found to be only marginally better predictors than random for T cell epitopes (Deavin et al., 1996) and B-cell epitopes (Blythe and Flower, 2005). The move away from such analyses for T-cell epitope prediction had commenced even before the work of Deavin *et al* when a study by Sette *et al* demonstrated that many MHC binding epitopes could be identified by restricted sequence motifs specific to unrelated sequences known to bind particular murine MHC class II molecules (Sette et al., 1989). This allele-specific motif property of MHC molecules was subsequently strengthened by the study by Falk *et al*, which also showed such motifs to be present in peptides eluted from MHC class I molecules (Falk et al., 1991).

More modern approaches to epitope prediction have also strived to exploit the large volumes of data and knowledge available on peptide MHC interactions and T cell epitope mapping studies. A variety of methods have been applied to the analysis of allele specific peptide interactions such as profile motifs (Reche et al., 2002), average relative binding (Bui et al., 2005), 2D QSAR (Doytchinova et al., 2002, Doytchinova, 2003 #8), 3D QSAR (Doytchinova and Flower, 2002), hidden markov models (HMMs) (Brusic et al., 2002, Mamitsuka, 1998), artificial neural networks (ANNs) (Buus et al., 2003, Honeyman et al., 1998), threading (Jojic et al., 2006), genetic algorithms (Del Carpio et al., 2002), virtual binding pockets (Sturniolo et al., 1999), or more atomistic approaches (Rognan et al., 1999, Logean et al., 2001, Meng et al., 2000). Additionally, some researchers such as Bhasin *et al* have attempted to directly classify cytotolytic T cell epitopes while bypassing the peptide-MHC interaction using an artificial neural network (Bhasin and Raghava, 2004). All of these approaches have varied degrees of availability with some implemented on publicly available web servers, freely available programs, commercial software, or, not available to the public (Schirle et al., 2001). Moreover, the use of many methods to go beyond simple peptide-MHC interaction predictions into the realm of accurately predicted novel T cell responses is quite limited.

Motifs and quantitative matrices generally require the assumption of independent binding of side chains (IBS). The IBS hypothesis develops the concept that for any given residue binding within the cleft of a MHC molecule,

the contribution of each amino acid side chain to binding is independent to the next. The necessity to account for peptide side chain interactions has been reported as both effective (Doytchinova et al., 2002) and only marginally better (Peters et al., 2003). However, the ability of researchers to model such interactions is more likely a function of dataset size. Profile based methods such as RANKPEP (Reche et al., 2002) rely on tools whose traditional roots are in protein domain profiling (Gribskov et al., 1987, Henikoff and Henikoff, 1996). However, through statistical evaluation Reche *et al* were able to demonstrate the applicability of profile analysis to the field of epitope prediction. Del Carpio *et al* tried to further refine the use of profile methodologies by implementing a genetic algorithm to define an optimal substitution matrix for use in a modified form of Gribskov's profile analysis (Del Carpio et al., 2002). This matrix was evolved through a feedback system to render the method capable of outputting quantitative predictions of MHC affinity. Bui *et al* also extended the abilities of matrix based methods in their automated generation of MHC binding matrices by incorporating a routine to convert the scores generated by their matrices to IC_{50} values (Bui et al., 2005).

Methods such as 2D QSAR take the concept of profile based methodologies one step further by deriving an equation/PSSM, which will, like the method of Del Carpio *et al* output a prediction of the *in vitro* binding affinity of a peptide for a specific MHC allele. However, unlike the previously outlined methods the 2D QSAR based additive method defined by Doytchinova *et al* is readily extensible and can be extended to account for interactions between adjacent side chains which may relate to the peptides conformation both *in vivo* and *in vitro.*

More traditional machine learning techniques such as artificial neural networks (ANNs) and Hidden Markov Models (HMMs) have been successfully implemented in MHC binding and T cell epitope prediction (Bhasin and Raghava, 2004, Brusic et al., 2002, Buus et al., 2003, Honeyman et al., 1998, Mamitsuka, 1998). These methods can take into account the interactions between residues of the binding peptide in addition to modelling nonlinear data, and are tolerant of experimental error making them well suited to analysis

of biological data (Brusic et al., 2004). However, a problem inherent to the construction of ANNs is the ability to over-train the model (Doytchinova and Flower, 2002). Nontheless, other authors suggest that with a suitable level of expertise one should be able to adjust the implemented ANN to account for such problems (Livingstone et al., 1997). The use of HMMs is also not surprising, as is the case with profile based methodologies, they are heavily rooted in protein domain analysis which would seem to translate quite well to MHC binding affinity prediction. Much like ANNs, HMMs can take into account the interactions between individual residues and have been recognised as valuable epitope prediction tools (Brusic et al., 2004).

The unfortunate downside to methods such as 2D QSAR, ANNs and HMMs is their inability to integrate biochemically relevant data as found in substitution and similarity matrices. Given the small quantities of data used to build models for the initial RANKPEP server (Reche et al., 2002), this biochemically relevant information would surely have added to the information content of the resultant PSSMs. Thus, more advanced methodologies such as 2D QSAR, ANNS and HMMs require a greater critical mass of data in order to build a robust model, as they assume no relationship –chemical or otherwise- between different residues.

2.1.4 MHC class I epitope prediction

The rather strict constraints of MHC peptide length make MHC class I epitope prediction a more attractive first stop when developing and implementing epitope prediction algorithms. As an example, both the additive and motif based methods implemented on the MHCPred and RANKPEP servers, respectively, were initially established as a resource for MHC class I epitope prediction (Doytchinova et al., 2002, Reche et al., 2002). Only after the methods had been demonstrated to operate correctly for MHC class I were they adapted for class II molecules (Doytchinova and Flower, 2003, Reche et al., 2004). For the purposes of model building, peptides of a single length, binding to the same MHC class I are assumed to occupy the binding groove in the same mode (Doytchinova et al., 2002, Reche et al., 2002, Bui et al., 2005, Brusic et

al., 2002, Sette et al., 1989, Peters et al., 2003). This approach for analysing 'pre-aligned' sequences has allowed the design of algorithms to predict binding of peptides of a set length to a single MHC molecule. It is the success of nonameric models that has undoubtedly lead to the transfer of these algorithms to the study of MHC class II binding, given the nonameric binding core of the MHC class II molecules.

As indicated above, a vast array of computational methods have been applied to the prediction of MHC class I binding and epitope responses. Some such algorithms have been used to mine out the first characterised CD8 T cell restricted response in coeliac disease (Gianfrani et al., 2003) using the methods described by Gulokota *et al* (Gulukota et al., 1997). Additionally, the online epitope prediction servers RANKPEP, SYFPEITHI, and BIMAS were all shown to be capable of prediction of a novel subdominant epitope of respiratory syncytial virus infection (Lee et al., 2007). However, the ability of such models to predict novel epitopes from sequences of unknown antigenic status was challenged by Pelte *et al* (Pelte et al., 2004). On the other hand, the validity of these findings has endured some criticism due to the authors' suboptimal choice of readout (Khanna, 2004).

2.1.5 Predicting proteasomal cleavage

More recently, for those wishing to computationally design candidate subunit vaccines, the integration of multiple tools for the single purpose of epitope prediction has become a valid prospect (De Groot and Berzofsky, 2004). One tool which has been added to the computational immunologist's repertoire is that of prediction of proteasomal cleavage sites (De Groot and Berzofsky, 2004, Flower, 2003b). The ability to combine the prediction of proteasomal cleavage sites with affinity predictions would provide an attractive option given its role in the selection of epitopes for presentation. Currently, a number of approaches have been taken to predict proteasomal cleavage sites using 20S proteasome degredation data (Nussbaum et al., 2001, Holzhutter et al., 1999), or a combination of both degredation and eluted peptide data (Kesmir et al., 2002).

2.1.6 MHC class II epitope prediction

The use of MHC binding predictions is hampered by the heterogenic length of characterised epitopes (Flower, 2003b). The main obstacle presented by such a dataset is the location of the principal binding determinant. As a result of the elucidation of the antigen complexed with MHC class II the principal binding determinant is assumed to be nonameric. To this end many available models for the prediction of peptide-MHC interactions are - at the most basic level - a model linking aligned core regions to the overall binding ability of the full length parent peptide while ignoring the contributions of flanking regions. Given the current state of knowledge in respect of peptide-MHC binding, one would not consider any of these model building assumptions to be without justification (Stern et al., 1994, Dessen et al., 1997, Kim et al., 2004). However, it would appear that assumptions may be over simplifications as MHC class II prediction has yet to live up to the abilities of MHC class I predictions (Flower, 2003b).

The majority of current MHC class II prediction methods are - to some degree - based upon extensions of MHC class I prediction algorithms. Reche *et al* use the MEME motif discovery program to form nonameric block alignments of allele specific MHC class II binding peptides (Reche et al., 2004). These block alignments are then used to build allele-specific profiles using the Henikoff approach as outlined in their previous report on MHC I prediction (Reche et al., 2002). Sturniolo *et al* used a virtual matrix approach for the design of their TEPITOPE software. Much like approach adopted by Reche *et al,* this required a prealigned set of sequences for definition of the initial binding pocket profiles (Sturniolo et al., 1999). Similarly, Doytchinova *et al* extended the additive method to the analysis of MHC class II molecules which was subsequently extended to the analysis of MHC class II supertypes (Doytchinova and Flower, 2003, Doytchinova and Flower, 2005). In contrast to Reche *et al*, however, the method outlined by Doytchinova *et al* was based on a novel iterative algorithm which constructed nonameric binding models which were then used to infer the location of the core binding region for the subsequent iteration. Honeyman *et al* derived qualitative binding models for

HLA-DR4 using an Artificial Neural Network and as was the case with Reche *et al* these peptides required prealignment (Honeyman et al., 1998).

The above methods for MHC class II prediction are all basic extensions of class I prediction algorithms. These methods incorporate approximations which may explain their shortcomings when compared to class I prediction routines. In a great number of cases the core regions are not experimentally determined but only inferred. Additionally, the impact of flanking regions on the binding affinity is also omitted (see also chapter 3). The development of more atomistic approaches will undoubtedly lead to such approximations becoming unnecessary as modelling of an increasing number of crucial variables will become a realistic option.

The sheer variety of methods available for the assessment and analysis of peptide-MHC interactions reflects not only the varied expertise and background of researchers in the area but also the data upon which the model is built. For example those researchers, who have available to them quantitative peptide-MHC binding data, are likely use such a resource to create a quantitative binding model. On the other hand, those working primarily with binary data are limited to qualitative models.

Choice of the optimal algorithm for a given dataset can be heavily dictated by the quantity of data, with small datasets requiring the use of motifs or atomistic approaches (Flower, 2003b, Brusic et al., 2004). As one increases dataset size more complex statistical models such as QSAR can be implemented, with quite large volumes of data required for optimal implementation of machine learning approaches such as HMMs and ANNs (Brusic et al., 2004).

2.1.7 Applicability of epitope prediction to *in vivo* antigenicity

The ability of any method to accurately predict *in vitro* and *in vivo* phenomena can only be properly assessed by actual novel prediction. Doytchinova *et al* have been quite successful in their prediction of novel binding affinities which could be experimentally verified using *in vitro* binding assays (Doytchinova et

al., 2004). Their findings demonstrated that their additive method was of use not only for locating putative epitopes but also for the design of high affinity peptides and modification of the binding properties of known epitopes. In a contrasting study by Andersen *et al,* it was reported that epitope predictions from the BIMAS and Eppredict server failed to correlate with *in vitro* binding determinations. However, the authors of this report compared the predictive abilities across numerous HLA alleles, whereas Doytchinova et al worked primarily on HLA-A*0201. The fact that HLA-A*0201 is well studied and the subject of many immunoinformatic evaluations, coupled with the quantitative nature of the additive method serves to increase one's confidence in the former report. Additionally, the usage of other epitope prediction methods to direct hypothesis driven research has also been reported; for example Gross et al used the TEPITOPE method to identify human leukocyte function-associated antigen-1 as a candidate autoantigen in treatment resistant Lyme arthritis (Gross et al., 1998). Additionally, Gianfrani *et al* successfully used MHC class I epitope prediction algorithms to characterise the first instance of CD8+ T cell reactivity to gliadin in coeliac disease (Gianfrani et al., 2003).

2.2 Materials and Methods

2.2.1 MHC Binding Data

Data containing information about peptides with characterised binding affinities for selected MHC molecules were obtained from the AntiJen database. These data were processed to remove any undesirable bias that may manifest when used in combination with binding prediction algorithms. The processed data were then used to form further discrete datasets tailored for particular tasks.

2.2.2 A set of quantitative binding data for the MHC allele A-0201

Data were complied from the AntiJen database by firstly obtaining all data for nonameric peptides known to bind the human MHC class I allele A-0201; this allele was chosen due to its widespread use in MHC binding predictions and the availably of binding data. Data were then normalised so as to provide the largest possible dataset of IC_{50} values under standardised conditions. Data were processed to ensure homogeneity of pH, molar concentration of peptides, indicator peptide sequence and concentration. IC_{50} values were converted to the negative of the natural logarithm of the peptide concentration expressed in Moles (the $-\ln IC_{50}$). Where duplicate sequences were found the following steps were taken:

a) For values that agreed to within 1.5 (-lnIC50), the average of the values was taken

b) Where the values were not with the above margin of agreement, the sequences and their associated data were removed from the set.

A second set of data was also compiled to act as a test set for method benchmarking in the same way as outlined above. This set was formulated by using peptides tested against the second most abundant indicator peptide. Data from this validation set were also compared to the training sets to ensure no duplicate entries existed between the two. If duplicates were found they were

removed from the training set in order to preserve a validation set of reasonable size.

2.2.3 Modelling of a set of theoretical nonameric peptide sequences for QSAR algorithm testing

The theoretical peptides and associated affinities were created by generating a matrix of 20x9 randomised values. This matrix corresponded to the theoretical contributions of twenty amino acids to each position in the binding groove. A series of randomly generated peptide sequences was then generated and the affinities of these random peptides generated using a custom python script to sum the contributions of each amino acid at each position in the random peptide as determined from our matrix of random numbers. Data were then divided into discrete datasets based on the required usage e.g. for one application a dataset of high affinity binders was produced using a purpose written computer program to randomly choose data of greater than a preselected cut-off value.

2.2.4 A set of quantitative peptide binding data for the MHC II allele DRB 0401

A dataset of peptides binding to the MHC class II molecule DR4 (DRB-0401) was compiled in a similar manner to the class I dataset above. Briefly, data were selected to provide the largest dataset of peptides assayed under homogeneous experimental conditions. As with our class I dataset, above, the molecule DR4 (DRB-0401) was chosen because of its well studied nature and data availability.

As with the MHC class I dataset generation above, a validation peptide set was also generated. This set of validation data were formulated from data of peptides tested against the second most abundant indicator peptide. IC_{50} values were converted to the negative of the natural logarithm of the peptide concentration expressed in Moles (the $-\ln IC_{50}$). Where duplicate sequences were found the following steps were taken:

a) For values that agreed to within 1.5 ($-\ln IC_{50}$) of one another, the average of the values was taken

b) Where the values were not within the above margin of agreement, the sequences and their associated data were removed from the set.

2.2.5 Filtering the MHC class I dataset to remove artificial and redundant sequences.

Data which originate from single residue substitutions of known binders, in addition to purely artificial sequences such as poly-alanine are common components of many MHC peptide binding databases. However, for profile methodologies, these data are assumed to be unsuitable because of their ability to skew the resultant profile. Therefore, a python script was written which would iteratively eliminate data for any sequence which had greater than a specified number of residues in common with any other sequence in the dataset. In order to determine the appropriate similarity cut-off the algorithm was run for a series of cut-offs and the appropriate value determined as a compromise between reducing peptide information bias while maintaining an adequate dataset size. In order to give a visual interpretation of the effect of redundant sequence removal on the dataset composition, heatmaps and dendrograms were plotted for data both pre- and post- processing. Heatmaps and clustering of the raw and cleaned datasets were implemented in R using a python generated matrix of sequence similarities. The heatmap function in R was used as follows: heatmap (<data_matrix>, scale='none')

2.2.6 Selection of amino acid similarity matrices

Ten matrices were chosen to represent different measures of amino acid similarity. These matrices were downloaded from the AAIndex website (http://www.genome.ad.jp/dbget/AAindex/list_of_matrices). The selected matrices were provided in the format shown in Figure 2.1 below. The data from these matrices was then converted into flat text files in the form of a 20x20 matrix of values to allow for easier manipulation using scripts. The AAIndex matrix IDs and their single line descriptions are presented in Table

2-1 (below). The matrices were chosen semi-randomly in order to represent a broad range of those present in the AAIndex resource.

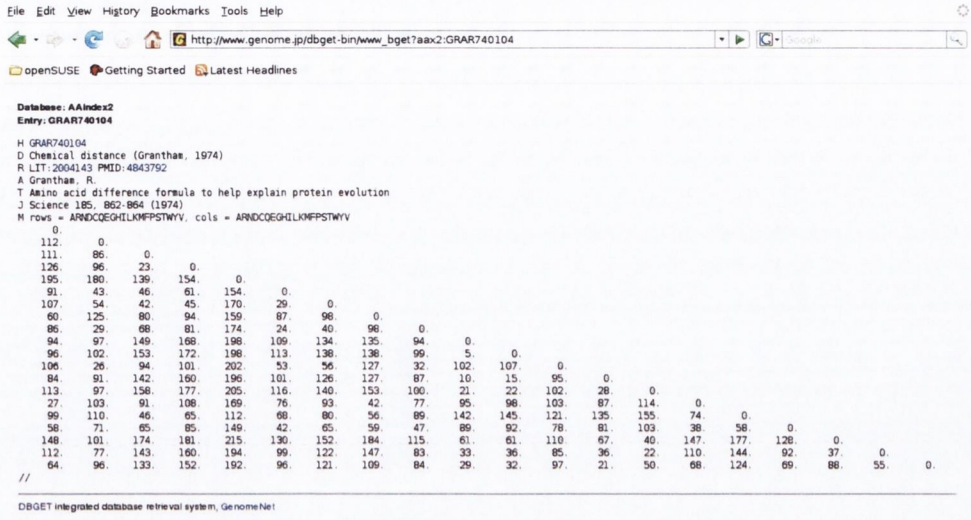


Figure 2-1 an example amino acid similarity matrix as viewed on the AAindex web resource

ID	Single Line Description	Ref
HENS920102	BLOSUM62 substitution matrix	Henikoff et al 1992
ALTS910101	The PAM-120 matrix	Altschul, 1991
MCLA720101	Chemical Similarity Matrix	Mc Lachlan, 1972
GEOD900101	Hydrophobicity scoring matrix	George et al., 1990
GRAR740104	Chemical distance	Grantham, 1974
LINK010101	Substitution matrices from a neural network	Lin et al., 2001
RIER950101	Hydrophobicity scoring matrix	Riek et al., 1995
TUDE900101	Isomorphicity of replacements	Tudos et al., 1990
WEIL970101	WAC matrix constructed from amino acid comparative profiles	Wei et al., 1997
WEIL970102	Difference matrix obtained by subtracting the BLOSUM62 from the WAC matrix	Wei et al., 1997

Table 2-1 Amino acid substitution matrices obtained from the AAindex web resource

Hereafter, the HENS920102 and ALTS910101 matrices will be referred to by their more common names i.e. BLOSUM62 and PAM120 respectively.

2.2.7 Examining the level of redundancy between the selected matrices

A heatmap and dendrogram was plotted to examine the level of redundancy between the different matrices. The plot was created using the 'heatmap.2' function from the 'gplots' package in R. The R command as issued from python was as follows:

```
r.heatmap_2 (<matrix_of_values>, scale='column',
col=r.redgreen (75), key=True, trace="none", cexRow=0.3,
cexCol=0.8, labCol=column_headers)
```

The scale setting was set to 'column' to account for the differences in absolute values between the different matrices.

2.2.8 Implementation of published epitope binding prediction algorithms

2.2.8.1 Gribskov's Profile Analysis

Profile analysis was performed as per the method defined by Gribskov *et al* (Gribskov et al., 1987). The defined method was implemented as a Python algorithm, outputting a matrix of m×n values, where m corresponds to one of each of the twenty amino acids and n corresponds to each of the relative positions in the epitope binding groove. Ordinarily, this was created as a 20×9 matrix to account for the contribution of each of the twenty amino acids at each of the 9 relative binding positions in the MHC binding groove (as it was unnecessary to account for gaps in the sequences). Each number in the matrix can be referred to as the position specific weight.

The equation initially outlined by Gribskov is as follows:

$$M(p,a) = \sum_{b=1}^{20} W(p,b) \times Y(a,b)$$

Where;

M(p,a) is the score for amino acid *a* at relative position *p*,

W(p,b) is the weight for amino acid *b* at relative position *p,* and

Y(a,b) is the amino acid similarity as calculated using Dayhoff's distance matrix.

The PAM120 matrix was used as the default matrix for the algorithm. The position specific weight $W(p,b)$ is calculated by dividing the frequency of occurrence of an amino acid b at position p divided by the number of amino acids at that point in the alignment – this was determined to be a superior method to normalising by the frequency of occurrence of a given amino acid (Section 2.3.6). The algorithm was implemented as a python script which has the capacity to build a Gribskov-type position specific scoring matrix (PSSM) for aligned epitope sequences using a training dataset. This profile was used to calculate a score for a query peptide by summing the score for each amino acid at each position of the test peptide with its counterpart in the PSSM. Hereafter the Gribskov's profile analysis (GPA) based algorithm will be referred to as the GPA algorithm.

2.2.8.2 Hidden Markov Models

Hidden Markov models of aligned sequences were created using the freely available HMMER suite of programs (Eddy, 1998). A python script was created from which the various programs were controlled. The resultant outputs from these programs were stored in intermediary files and parsed using a custom python script. The model was built using the 'hmmbuild' program with the 'amino' and 'hand' options specified. The model was then calibrated using the 'hmmcalibrate' program. Epitope scoring was done using the 'hmmsearch' program. The score for each nonamer as output by the 'hmmsearch' program was used as the epitope affinity measure.

2.2.8.3 The Additive Method

The Additive method was essentially performed as described by Doytchinova et al (Doytchinova et al., 2002) with two main exceptions; firstly, the number of latent variables was fixed throughout the analysis (beginning with a default value of four); secondly, the algorithm was implemented using open source as opposed to proprietary software. The algorithm was implemented using a combination of two languages: Python was used for data processing and R was

used for statistical analysis of prepared data. The two languages were linked using the rPy interface and partial least squares models were generated using the pls.pcr library for R.

The algorithm initially created a model using a set of training data. These training data were converted to binary bitstrings of 20 * *n* amino acids in length, where *n* represents the number of relative binding positions being modelled and 20 represents one bit for each of the twenty standard amino acids. All of the models used in this thesis were built using nonameric sequences resulting in all bitstrings being of an equal 180 bits in length (20 amino acids * 9 relative binding positions). These bitstrings were then compiled to form a matrix of 180 x *m* values, and their associated affinities used to form a separate matrix of 1 x *m* values where *m* represents the number of sequences in the training set. The derived model was assessed, where appropriate, using the Q^2 value which is similar to the R^2 value but generated using cross validation data.

In specified instances, certain data were subject to automated removal if they were found to be in strong disagreement with the remainder of the dataset by use of a pruning algorithm (Doytchinova et al., 2002). This dataset 'pruning' was performed iteratively, and on each iteration removed a single epitope with the largest observed residual value (i.e. difference between observed and predicted affinity) from the training set. Dataset 'pruning' was implemented in the QSAR methodology outlined by Doytchinova *et al* (Doytchinova et al., 2002). The algorithm was designed to cease pruning on one of the following conditions being true:

a) The greatest observed residual is below a pre-selected value (default value of 2), or
b) 5% of values have been removed from the dataset.

A final model was generated after completion of pruning and this was used to predict the *in vitro* –lnIC50 values for our test epitopes. Hereafter the additive method will be referred to as the QSAR algorithm in order to best reflect its statistical basis.

2.2.9 A direct comparison of three epitope prediction algorithms

The three algorithms outlined above (Gribskov, HMM and QSAR) were each used to predict the affinity of peptides in the HLA-A2 test set. Three separate models were built with each algorithm using the three versions of the HLA-A2 dataset i.e. the raw dataset and both of the cleaned datasets using cut-off values of 3 and 4 (Section 2.2.5). For the Gribskov method, the default substitution matrix was PAM120. Pruning was also implemented for the Additive method as outlined in the results (Section 2.3.5).

The predicted and experimentally determined affinities for each peptide in the training dataset were then compared using the Pearson correlation statistic. The list of pruned peptides (generated using the QSAR pruning algorithm) was visually examined for patterns that might account for an uncontrolled source of systematic variation in the training data.

2.2.10 Optimising the GPA-algorithm for epitope prediction

2.2.10.1 Choosing the optimal averaging denominator for calculation of the position specific weight

The first step in optimising the Gribskov method was the determination of the optimal averaging denominator i.e. whether to divide by the number of residues in a column or the number of occurrences of a particular residue in the training set. This step was performed by generating two different binding models for the same allele and testing their ability to make predictions that correlated with those found experimentally.

2.2.10.2 Optimal Dataset Size

Correlation coefficients were compared for models built using both the raw dataset and both cleaned datasets for HLA-A*0201.

2.2.10.3 Selecting the best substitution matrix for epitope prediction

The second approach was to determine the optimal amino acid substitution matrix for the profile method. For this purpose a selection of ten different substitution matrices was selected and downloaded as outlined above (section 2.2.6). Each of these matrices was then used in place of the default PAM120 matrix. A leave-one-out cross validation approach was used to choose the optimal matrix and was implemented as part of a custom Python script. The Python script operated by sequentially removing a single epitope from the training data and generating a binding model using the training set of n-1 peptides. The affinity of the 'left-out' peptide was then predicted using the binding model created in its absence. This proceeded until a separate model and associated prediction was made for each epitope in the training set, thus providing a predicted and *in vitro* measure of affinity for each epitope in the training set.

2.2.11 Optimising the QSAR method for epitope prediction

2.2.11.1 Optimal Dataset Size

Correlation coefficients were compared for models built using both the raw dataset and both cleaned datasets for HLA-A2. These results were compared using both leave-one-out-cross-validation (LOOCV) and the Pearson correlation coefficient for predicted affinities of validation data.

2.2.11.2 The effect of dataset composition on QSAR model predictivity

In order to assess the impact of dataset composition on the generated binding model, modelled MHC class I binding data were used (Section 2.2.3). The affinity measurements were used to generate a series of training datasets of peptides with differing affinity composition e.g. a dataset of high affinity peptides only. These different training sets were used to generate different binding models. The ability of each model to predict affinity was measured

using the Pearson correlation coefficient for a validation set of 1000 modelled peptides of mixed affinity.

2.2.11.3 Selecting the optimal number of latent variables for model building

A single dataset of mixed affinity modelled MHC class I binding data was created as outlined in section 2.2.3 (n=180). The NORMINV function of Microsoft Excel was used introduce random variation into the modelled peptide affinities. The coefficients of variation for the modelled random error ranged from 0-30% and increased in 5% increments. Binding models were constructed using a series of 1 to 10 latent variables. The correlation coefficients were calculated using predicted vs. actual affinity for a series of 1000 modelled MHC class I binding peptides without any introduced error. The calculation was performed 4 times and the average R^2 values analysed.

2.2.12 Reanalysis of the three binding algorithms using optimisations

The three prediction algorithms used for the comparison, above, were retested using the same datasets as before (Section 2.2.9). In contrast to the previous comparison however, method optimisations based on results from the optimisation steps were implemented (sections 2.2.10 and 2.2.11, above). The Gribskov analysis was implemented using column based averaging and the MCLA7201010 chemical similarity matrix. The hidden markov model algorithm was implemented without any further optimisation, and, the additive method was implemented using three latent variables as opposed to the initial default value of four. The pruning algorithm (section 2.2.8.3) was also implemented as a part of the QSAR algorithm where indicated.

2.2.13 Method Comparison for core region determination of MHC class II binding peptides

Unlike, MHC class I, MHC class II do not impose constraints on the length of peptides. Therefore, the variable length of MHC class II molecules necessitates pre-processing of the data to predict which region of the peptide is the likely principal binding determinant. In order to determine the best method for nonameric core prediction, three methods were assessed:

a) Clustalw, a standard tool in protein sequence alignment more commonly used for evolutionarily related (namely divergent) sequences
b) Motifs: a DR4 binding motif was used to predict the nonameric core using a simple scoring scheme. The motif used for this study was a modified form of that used by Boots *et al* (Boots et al., 1997).
c) An iterative self consistent (ISC) algorithm which iteratively builds successive binding models based on the core regions it predicts and uses these to refine its binding and core predictions. This approach was initially defined by Doytchinova *et al* (Doytchinova and Flower, 2003).

2.2.13.1 Alignment using Clustalw

All Clustalw runs were performed locally using version 1.82 of the software (Chenna et al., 2003, Higgins and Sharp, 1988). Data were prepared for alignment by converting to FASTA format using an in-house python script. All data were analysed using default parameters with one exception i.e. gap creation and extension penalties were increased to the maximum permissible levels to account for the ungapped nature of MHC binding peptides.

Over-representation of sequences in the dataset has the capacity to skew the alignment and affect the information content of the plot. In order to overcome this problem a separate algorithm was implemented to reduce any redundancies in the data by extracting data which had a common wordmatch of greater than 5 amino acids.

All alignments were visually inspected but not edited. The information content and summary of each alignment was obtained by using sequence logos (Schneider and Stephens, 1990). Sequence Logos were rendered using the WebLOGO service – (http://weblogo.berkeley.edu/logo.cgi).

2.2.13.2 Alignment using motifs

A motif representing the binding preferences of pockets in the HLA-DR4 binding groove was obtained from the SYFPEITHI database (Rammensee et al., 1999). A simple motif search using regular expressions failed to adequately identify the core regions of our known DR4 binders. To counteract this lack of sensitivity the motif was implemented as a position specific binary scoring matrix (see Table 2-2, below). Using this scoring scheme a python script was implemented which selected for the most appropriate (i.e. highest scoring) nonameric core from each peptide in our dataset. Knowledge of 'unpreferred' residues was omitted from the scoring scheme as it was not available as part of the original motif.

	1	2	3	4	5	6	7	8	9
A	0	0	0	1	0	0	1	0	1
R	0	0	0	0	0	1	1	0	1
N	0	0	0	0	0	1	1	0	1
D	0	0	0	1	0	0	1	0	1
C	0	0	0	0	0	0	1	0	1
Q	0	0	0	0	0	1	1	0	1
E	0	0	0	1	0	0	1	0	1
G	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	1	1	0	1
I	1	0	0	1	0	0	1	0	1
L	1	0	0	1	0	0	1	0	1
K	0	0	0	0	0	0	1	0	1
M	1	0	0	0	0	0	1	0	1
F	1	0	0	0	0	0	0	0	0
P	0	0	0	1	0	0	0	0	0
S	0	0	0	0	0	1	1	0	1
T	0	0	0	0	0	1	1	0	1
W	1	0	0	1	0	0	0	0	0
Y	1	0	0	0	0	0	1	0	1
V	1	0	0	1	0	0	1	0	1

Table 2-2 the binary matrix used to implement a semi-quantitative version of the binding motif obtained from the SYFPEITHI database.

Highlighted in blue are those residues that are favoured at a given position. Highlighted in red are those that are know to be selected against but are not accounted for by the motif or our binary position specific scoring matrix.

As the output from the matrix based scoring scheme produced data with a degree of overlap, the data were processed to remove excessively redundant data. This was performed as for the data filtering for our class I molecules (Section 2.2.5, above); some manual editing was necessary to remove duplicate sequences. The information content of the data was visually assessed as with our Clustalw alignment by using the WebLOGO service. The level of

redundant information in our data was assessed using heatmap and dendrogram plots as outlined above (Section 2.2.5).

2.2.14 Affinity prediction using motif based nonameric core alignments

Motif based alignments were used to form a basic training set of nonameric sequences with associated affinities. The affinities associated with each nonameric sub-sequence were those of the full length, parent sequence. This approach assumed a primary role for the nonameric core in the determination of the binding affinity and a negligible contribution from residues outside of the nonameric core. Using this training set, binding models were built using the same optimised methods as for the class I molecules, above (Section 2.2.12).
The only data available to validate the model were peptides of varying lengths, all of which were greater than nine amino acids in length. To circumvent this limitation a highest scoring sub-sequence approach was taken. Using this approach each peptide was used as a base to form a series of (n-8) nonameric sub-sequences, where n is the length of the parent sequence. Each of these nonamers was evaluated against the binding model and the highest predicted value for any nonameric sub-sequence chosen as the binding affinity for the parent sequence.

2.2.15 Affinity prediction using the Iterative Self Consistent Algorithm

The iterative self consistent algorithm was implemented quite similarly to the algorithm outlined by Doytchinova et al (Doytchinova and Flower, 2003). The in-house implementation differed from the original method in the following ways:

a) The implementation used open source (Python, R and rPy) as opposed to commercial software (SYBYL6.7)
b) A fixed number of latent variables (3) was used throughout the iterative procedure.
c) Epitope data were not pre-selected based on length.

d) Core data were not confined to having a bulky hydrophobic residue at position 1

Briefly, for each peptide in the training set, a series of overlapping nonameric sub-sequences was generated. Each sub-sequence was linked to the binding affinity of the parent sequence. These sub-sequences and associated affinities were then used to build an initial binding model. This initial model was then used to predict the affinities of each nonameric sub-sequence from each peptide in the training set. From this list of predicted affinities one nonameric sub-sequence was chosen to represent the principal binding determinant of the peptide. In one scenario, the highest scoring nonameric sub-sequence was chosen as the principal determinant, and in another, the closest predicted affinity to the in vitro score was chosen. The chosen nonameric cores coupled with the associated parent sequence affinity were then used to generate a new training set of nonameric peptides. This list was then used to build a second model and the prediction of principal determinants reiterated as before. The new list of nonameric cores was used to generate another list and proceed again with model building. This step was repeated until the model was seen to obey one of two criteria;

a) the model was seen to 'converge' i.e. the predicted nonameric cores for a given iteration were identical to those from the previous iteration, or,

b) Model predictions fell into a repeating pattern of nonameric core selection by virtue of which, convergence is rendered impossible. In this case the model with the highest Q^2 value was chosen as the final model.

As with section 2.2.9, the validation of the model was performed using a set of known HLA-DR4 binding peptides. And as with the previous analysis (Section 2.2.14) the same methods were used to account for varying length of validation peptides i.e. a highest scoring sub-sequence approach was taken. Using this approach each peptide was used as a base to form a series of (n-8) nonameric sub-sequences; where n is the length of the parent sequence. Each of these nonamers was evaluated against the binding model and the highest predicted value for any nonameric sub-sequence chosen as the predicted binding affinity for the parent sequence.

2.3 Results

2.3.1 A set of quantitative binding data for the MHC allele A-0201

The dataset from our initial processing comprised 349 nonameric sequences of known binding affinity for HLA-A2. Peptides in the set were preselected on the basis of being tested against the most widely used indicator peptide sequence (FLYSDYFPSV). The initial dataset was further trimmed to remove potential sources of bias resulting from differing pH or peptide concentrations. This process generated a final set of 349 peptides. A second peptide dataset, tested against a separate indicator peptide, was generated in the same manner to act as a test set (n=34).

2.3.2 Modelled MHC class I binding data for algorithm testing

The modelled class I data were used to generate four training sets of 180 theoretical peptides with each comprised of high (>17), medium (14-17), low (<14) or mixed affinities. Additionally, a test set of 1000 peptides of mixed affinities was also generated to validate the predictive abilities of the algorithms.

2.3.3 A quantitative set of binding data for the MHC class II molecule HLA-DR4 (DRB1*0401)

The largest dataset of values tested against the same indicator peptide (YARFQSQTTLKQKT) were filtered and processed to generate a list of peptides plus associated affinities generated under the same experimental conditions. This processing generated a dataset of 134 distinct peptides with distinct radiometrically determined IC_{50} values.

The validation set – formed from data generated against the second most abundant peptide – comprised 42 peptides with fluorometrically determined IC_{50} values. The differences between a fluorometric and radiometric assay

based dataset for validation were thought to be minimal considering the minor differences that exist between the assays.

2.3.4 Filtering the MHC class I dataset to remove artificial and redundant sequences

The initial heatmap and dendrogram of the HLA-A2 dataset (Figure 2-2) shows a number of clusters of highly similar sequences. In order to provide a data-subset that was presumed to be more compatible with the profile based methodologies, it was necessary to remove these clusters. However, in order to ensure an adequate supply of data for model building, removal of too much data needed to be avoided. In order to make the best compromise between these two conflicting aims the data cleaning process was implemented for a variety of cut-offs. The effect of this test is shown in Figure 2-3, which demonstrates the two reciprocal effects of altering the cut-off. In the first of the two graphs, we can see that, while very few sequences may have a large number of residues in common, most if not all have at least one residue in common with at least one other member of the dataset. The reciprocal of this graph is also included for illustration purposes and shows the effect of removing data points with a set number of position specific residues in common with at least one other data point.

As a result of analysing the impact of common residue cut-off on dataset size, it was decided to use two cut-offs which resulted in removal of peptides with greater than 5 or 4 residues common to specific positions. A heatmap and dendrogram was then generated to demonstrate the effect of this pruning on the data. The resultant graph (Figure 2-4) clearly shows that a number of clusters of highly similar sequences have been removed, generating two cleaner - albeit smaller - datasets.

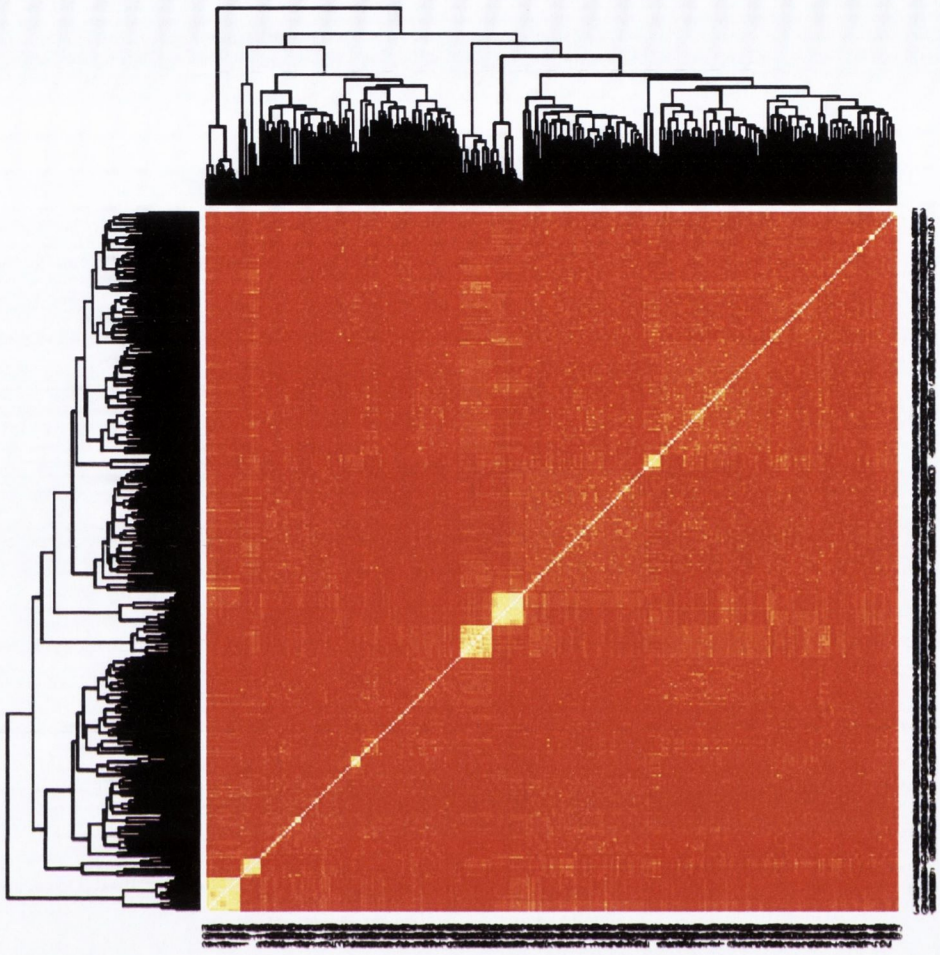
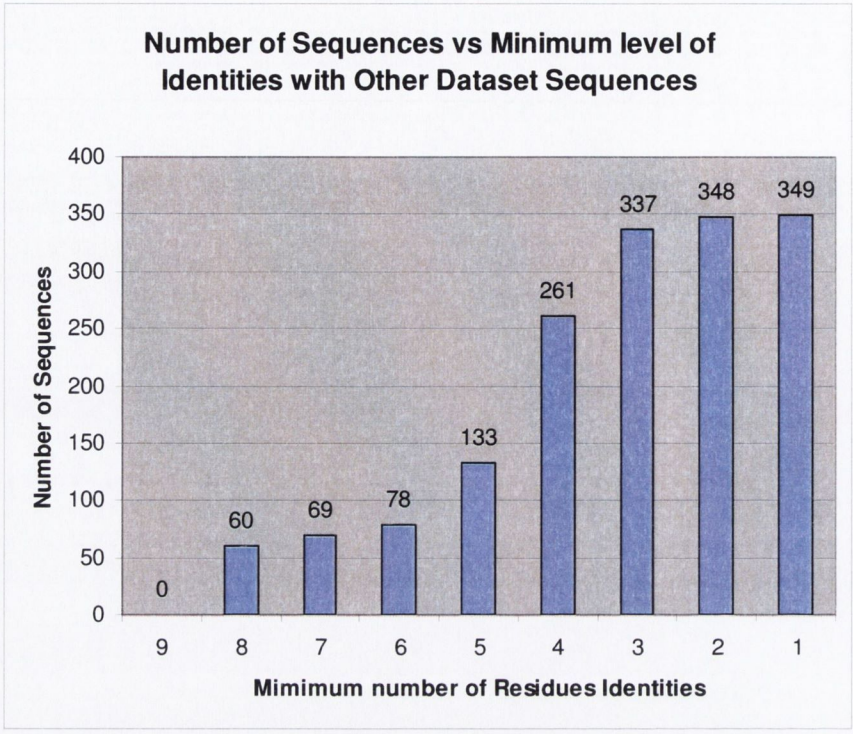
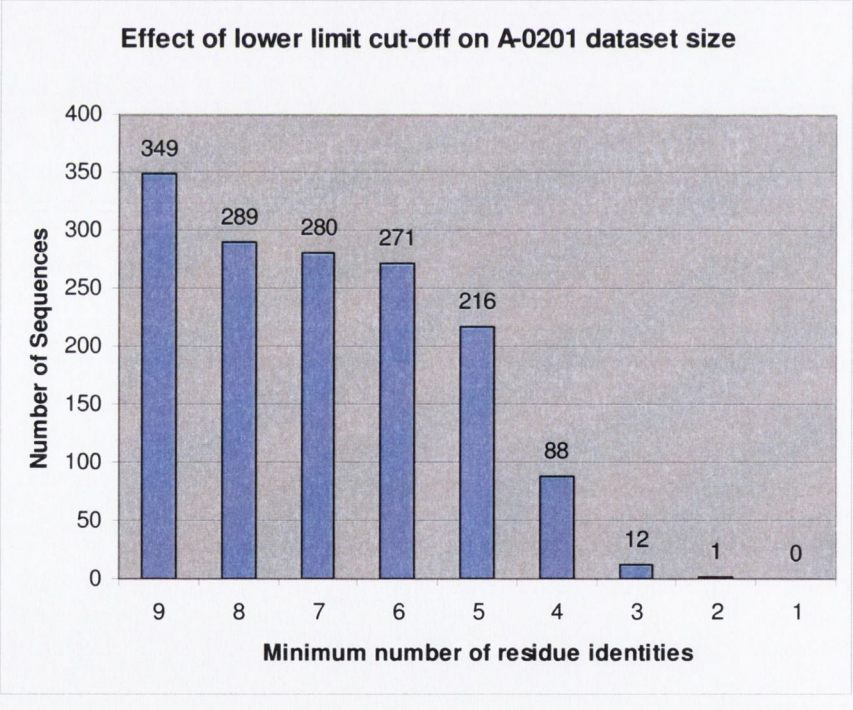


Figure 2-2: A heatmap and dendrogram of sequence similarities pre-processing.
This figure shows the similarity between sequences in the unprocessed HLA-A2 dataset. The data composition of this chart shows sequences with greater similarity in lighter shades. Thus, one can see identical sequences which are shown in white (the diagonal line bisecting the diagram). Also apparent, is the effect of clustering the data, the clustering places closely related groups of sequences in close proximity to one another.



A



B

Figure 2-3 Demonstration of the effect of cut-off stringency on dataset size.

The above figures show the number of sequences in the initial dataset with greater than a specified number of position specific residue identities with one or more other sequences in the dataset (A), the effect of removing such data is also illustrated (B).

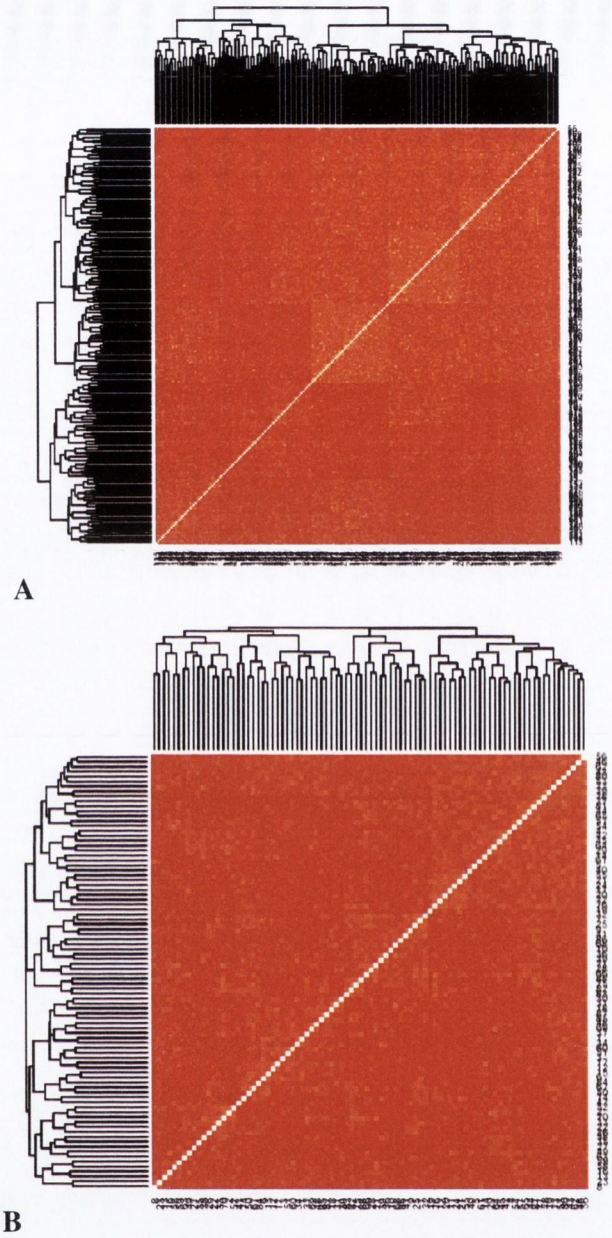


Figure 2-4 Heatmap and dendrogram of sequence similarities after data cleaning

The above heatmap and dendrogram plots depict the levels of common sequence identity between peptides in the HLA-A2 dataset after implementing data cleaning. The two plots illustrate the commonalities between data processed using a cut-off of either 4 (A) of 3 (B) sequence identities. It can be seen that using a cut-off of 4 has lead to a dramatic reduction in closely related sequences (A) but a level of close grouping is still evident throughout the set. In contrast, reducing the cut-off to 3 has removed all but the most subtle of sequence similarity (B).

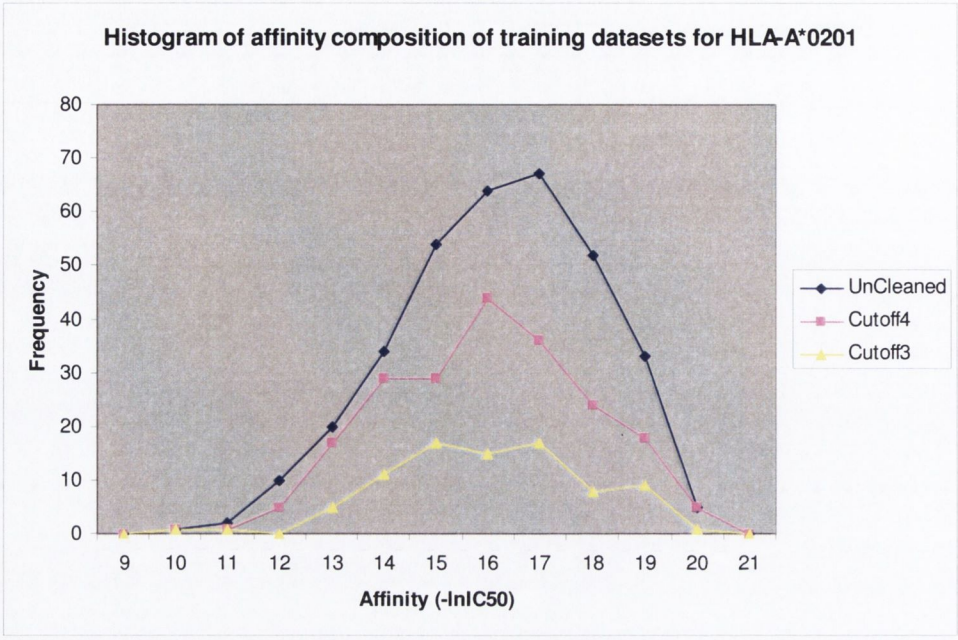


Figure 2-5 Histogram of affinity measurements of data in both cleaned and raw datasets

The above histogram illustrates the composition of the HLA-A2 training dataset before and after implementation of the data cleaning algorithm. While the data cleaning does reduce the number of training sequences, it does not radically alter the affinity distribution.

2.3.5 Direct comparison of predictivity for three MHC binding prediction algorithms

The comparison of the three epitope prediction algorithms demonstrated a marked ability for each algorithm to predict the affinities of peptides from the validation set which correlated with the experimentally derived values. This ability varied depending on the dataset composition, which varied in terms of amino acid bias and size. To assess the impact of data overlap between training and validation sets the analyses were also performed using the datasets as they appeared before removal of data common to both. Comparison of the prediction correlation coefficients before (Table 2-3), and after (Table 2-4), removal of common data show the different effects such a training set bias could have on the model building process.

The pruning element of the additive method was also selected in certain analyses to analyse its impact on the predictions. The pruned data and the training set after pruning are visualised using sequence Logos (Figure 2-6,

below), which seem to show no obvious difference between the amino acid content of either pruned or training data.

		Algorithm			
Dataset	**n**	**Gribskov**	**HMM**	**QSAR**	**Pruning**
Cleaned - Cut-off: 3	88	0.559	0.617	0.435	N
Cleaned - Cut-off: 4	216	0.614	0.554	0.672	N
Raw Dataset	349	0.626	0.640	0.697	N
Raw Dataset	349	-	-	0.708	Y

Table 2-3 Pearson correlation coefficients for experimental and predicted HLA-A2 affinities for a series of epitope prediction algorithms

The QSAR model would appear to show the best predictivity when coupled with the largest dataset and when incorporating the outlier-pruning algorithm.

		Algorithm			
Dataset	**n**	**Gribskov**	**HMM**	**QSAR**	**Pruning**
Cleaned - Cut-off: 3	85	0.522	0.516	0.242	N
Cleaned - Cut-off: 3	208	0.600	0.524	0.556	N
Raw Dataset	341	0.620	0.632	0.621	N
Raw Dataset	341	-	-	0.610	Y

Table 2-4 Table showing the correlation coefficients of the different methods after removal of data common to both the training and validation sets

Removal of data common to the training and test sets results in a reduced level of predictivity for each algorithm when compared with Table 2-3. The HMM algorithm coupled with the raw dataset would now appear to be the best method.

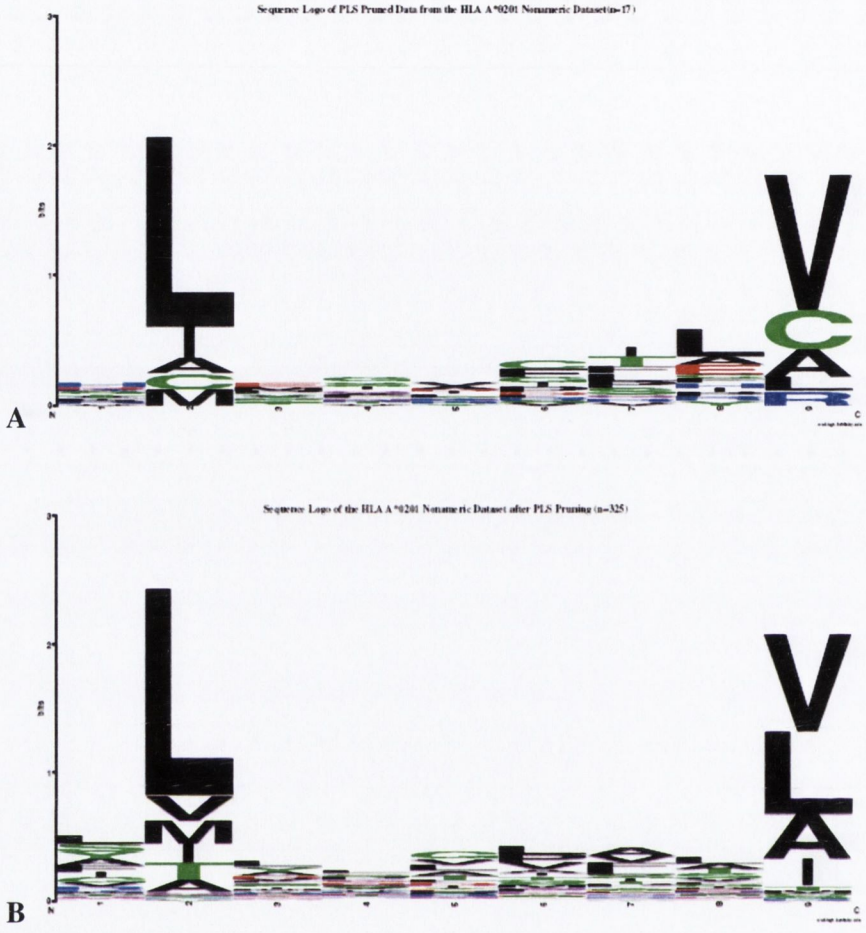


Figure 2-6 Sequence Logos demonstrating the effects of dataset pruning

The above Logos show the information content of the data pruned from the HLA-A0201 training set (A) and the information content of the dataset after pruning (B). In these Logos the greater the height of a particular letter the greater the representation of that residue at a particular position relative to background.

2.3.6 Comparison of row and column normalisations for the Gribskov algorithm

Comparison of row and column normalisation methods for Gribskov's profile analysis shows that for every variation of the dataset, the column normalised training data gave better predictions (Table 2-5). This strongly suggested column normalisation to be the better normalisation method for epitope prediction.

	Normalisation	
Dataset	**Column**	**Row**
Cleaned - Cut-off: 3	0.522	0.145
Cleaned - Cut-off: 4	0.600	0.149
Raw Dataset	0.620	0.200

Table 2-5 Pearson correlation coefficients of predicted affinities for peptides in the validation set binding to HLA-A2

The Pearson correlation coefficients for results generated with models formed using either 'Row' or 'Column' normalisations are shown.

2.3.7 Comparison of Amino Acid substitution matrices for use in profile based epitope prediction

Ten amino acid similarity/distance matrices were downloaded from the AAindex webpage for use in profile based algorithms. The cluster and heatmap analysis of these is shown below (Figure 2-7). Just above the horizontal axis of the plot one can note the clustering of the different similarity/substitution matrices. The grouping of the PAM120 and BLOSUM62 matrices demonstrates the redundancy present amongst a number of the matrices – especially between data which are used for a similar purpose such as the PAM and BLOSUM series of matrices. Overall however, the chosen matrices exhibit a degree of difference that makes them suitable for comparative study.

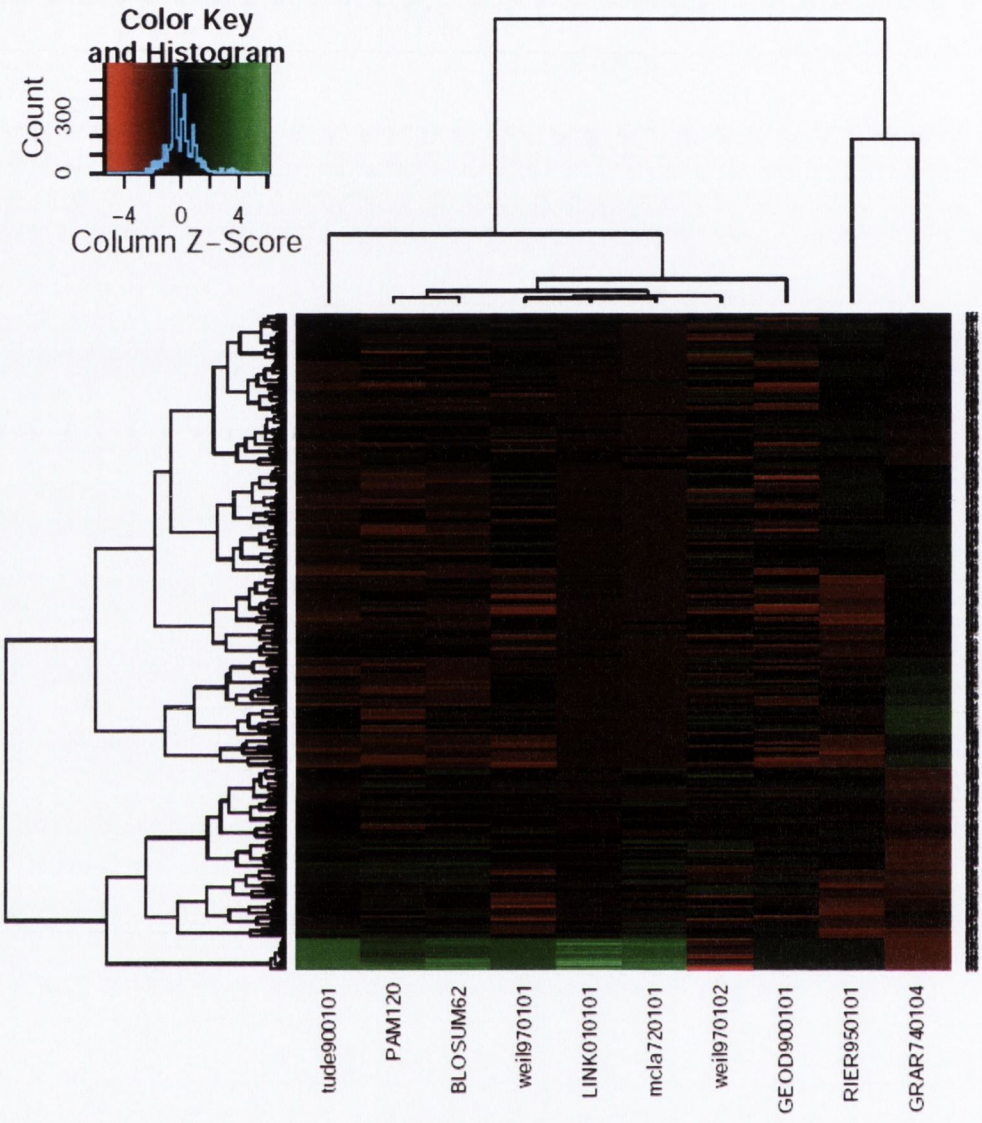


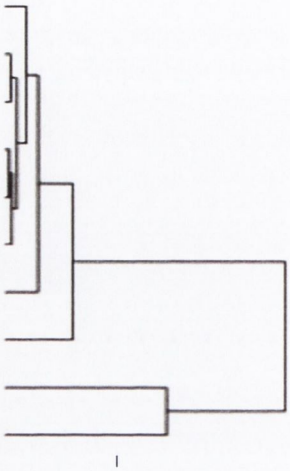
Figure 2-7 Heatmap and dendrogram of the ten selected amino acid similarity matrices

The clustering of each matrix (top) shows the grouping of each matrix based on similar information content with regard to amino acid similarities/substitution probabilities. The clustering on the vertical axis shows clustering of similar scores in each matrix e.g. the lowermost cluster shows the scores for substitution of identical amino acids for each matrix.

2.3.8 Influence of amino acid substitution matrix on Gribskov profile analysis–based MHC affinity predictions

The LOOCV prediction values for each of the different matrices using the Gribskov-based method (Table 2-6 (A)) show the apparent differences between each matrix in terms of epitope prediction. These results suggest the MCLA720201 matrix as the best matrix for MHC binding prediction. The second set of analyses compared the ability of the best scoring matrix based on cross validation (MCLA720101) to predict the affinity of the peptides in the validation set when implemented as part of the GPA algorithm. For completion, correlation coefficients for data from all substitution matrices were examined, but not used for selection of the affinity prediction matrix.

A	Cut-off 3 (R^2)	Cut-off 4 (R^2)	Raw (R^2)
TUDE900101	0.270	0.236	0.301
PAM120	0.088	0.146	0.227
BLOSUM62	0.215	0.222	0.293
WEIL970101	0.371	0.202	0.269
LINK010101	0.346	0.210	0.285
MCLA720101	0.395	0.320	0.364
WEIL970102	0.236	0.129	0.046
RIER950101	0.030	-0.051	0.052
GEOD900101	0.157	0.075	0.120
GRAR740104	-0.178	-0.283	-0.352
n=	85	208	341



B	Cut-off 3 (r)	Cut-off 4 (r)	Raw (r)
TUDE900101	0.694	0.666	0.659
PAM120	0.559	0.614	0.626
BLOSUM62	0.523	0.609	0.635
WEIL970101	0.207	0.260	0.276
LINK010101	0.734	0.618	0.597
MCLA720101	0.680	0.663	0.656
WEIL970102	-0.582	-0.611	-0.548
RIER950101	-0.239	0.041	0.027
GEOD900101	0.354	0.367	0.386
GRAR740104	-0.541	-0.574	-0.598
n=	30	30	30

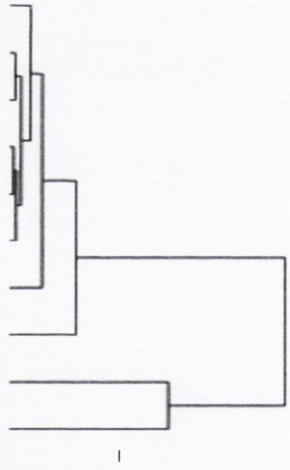


Table 2-6 Evaluation of the optimal substation matrix for epitope prediction

The upper table shows the correlation coefficients attained for each matrix using LOOCV (A). The lower table (B) shows the Pearson correlation coefficient for each set of predictions using the validation set. For each comparison the best score is highlighted in green and the worst in red. The validation set correlation for the substitution matrix selected using the LOOCV data is highlighted in blue. The dendrogram derived using R (see also figure 2.5) and illustrated above is located alongside the table to show the similarities and differences between the predictive abilities of the different matrices.

2.3.9 Effects of dataset composition on the model building and predictivity characteristics of the QSAR algorithm

The differing abilities of the additive method to produce a well fitting model in addition to its ability to predict novel sequence affinities were assessed using the modelled binding data. The results (Table 2-7, below) show that sole use of high affinity data limits the model fit -in addition to predictive power- of the method, as evidenced by the lowest R^2 values for any training data subset. Overall, the mixed affinity data provided the best combination of both model building and validation. The single, substitution dataset provided a poor model R^2 value (0.335), yet generated the highest validation R^2 value (0.962). The medium and low affinity datasets provided lower R^2 values than either the mixed affinity or single substitution datasets yet outperformed the high-affinity dataset.

	R^2 Value	
Dataset	**Model (Q^2)**	**Validation**
Mixed Affinity	0.944	0.824
Single Substitution	0.335	0.962
High Affinity	0.057	0.205
Medium Affinity	0.244	0.464
Low Affinity	0.438	0.662
n	180	1000

Table 2-7 Model building and validation R^2 values for a variety of training set compositions for a theoretical allele

Highest values are highlighted in green while lowest values are highlighted in red.

2.3.10 Optimisation of latent variables

The effects of varying the number of latent variables on the predictive abilities of the additive method show a clear preference for a lower rather than higher number of latent variables. As these results are generated using a theoretical model the indicated number of latent variables can not be thought of as wholly applicable to actual experimental data. However, the general trend suggests

that a lower %CV allows the use of more latent variables, thus generating better model predictions. This would suggest that, when generating models from published literature, the use of a small number of latent variables would be appropriate. We therefore elected to fix the number of latent variables at 3 for the remaining analyses.

	CV %						
nLV	**0%**	**5%**	**10%**	**15%**	**20%**	**25%**	**30%**
1	0.634	0.619	0.504	0.439	0.328	0.251	0.218
2	0.801	0.716	0.422	0.327	0.194	0.115	0.092
3	0.857	0.712	0.318	0.239	0.139	0.074	0.057
4	0.895	0.688	0.349	0.187	0.107	0.051	0.039
5	0.924	0.662	0.219	0.152	0.084	0.036	0.030
6	0.937	0.669	0.189	0.140	0.075	0.028	0.027
7	0.947	0.596	0.163	0.128	0.062	0.021	0.021
8	0.955	0.557	0.140	0.118	0.052	0.017	0.017
9	0.962	0.521	0.124	0.111	0.047	0.015	0.015
10	0.968	0.487	0.114	0.104	0.042	0.013	0.014

Table 2-8 R^2 values for a series of data with modelled experimental error versus an increasing number of latent variables

The % CV (Coefficient of variation) increases from left to right along the X-axis and the number of latent variables (nLV) increase from top to bottom along the Y axis. The best observed values for each %CV are indicated in green while the worst are indicated in red.

2.3.11 Re-evaluation of the re-optimised binding prediction algorithms

The results of the post-optimisation validation experiments show that for all dataset compositions, the Gribskov algorithm particularly benefited (Table 2-9, below). The same was not true for the PLS based algorithm which showed moderate decreases in predicted affinity for the cleaned datasets and moderate increases for the results generated using the raw dataset. As the HMM based method was not subject to any optimisations, the results are unchanged.

		Algorithm			
Dataset	**n**	**Gribskov**	**HMM**	**PLS**	**Pruning**
Cleaned - Cut-off: 3	85	0.637 (+115)	0.516 (0)	0.207 (-35)	N
Cleaned - Cut-off: 4	208	0.655 (+55)	0.524 (0)	0.510 (-46)	N
Raw Dataset	341	0.651 (+31)	0.632 (0)	0.647 (+26)	N
Raw Dataset	341	-	-	0.618 (+8)	Y

Table 2-9 Correlation coefficients of predicted vs. actual affinities for the HLA-A2 validation dataset

The figures in brackets contain the difference between the optimised correlation coefficients and the coefficients reported prior to optimisation (see also Table 2-4, above). Increases in correlation are indicated in green, decreases in red.

2.3.12 Evaluation of nonameric core alignment methods

Unlike the MHC class I data, the MHC class II data need to be aligned to highlight their core binding region. Two methods were compared for identifying/aligning these core regions for use with the profile based algorithms.

2.3.12.1 Alignment using Clustalw

Alignment using the Clustalw program is summarised using the sequence LOGOs shown below (Figure 2-8). It can easily be seen in the first of these images, that the alignment required the removal of particularly redundant data as these data have the capacity to skew the output. However, having removed the highly redundant data using our wordmatching routine (Section 2.2.13) the alignments still show little signal when compared to background. Even so, with a minor signal and effective alignment of core regions we might expect to be able to visually identify the core region peptides. This was not the case and the Logo seems to show a more distributed alignment of sequences, which would suggest that Clustalw alignment is unsuitable for alignment of nonameric cores from MHC class II binding peptides.

A

B

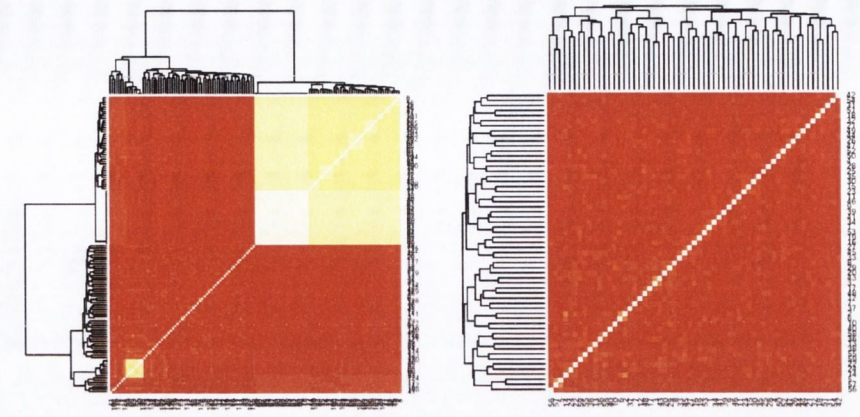
Figure 2-8 Sequence Logo Showing the information content of a Clustalw alignment before (A) and after removal of repetitive data (B)

The cleaned dataset presents much less information as can be seen from the number of bits for each residue indicated on the vertical axis.

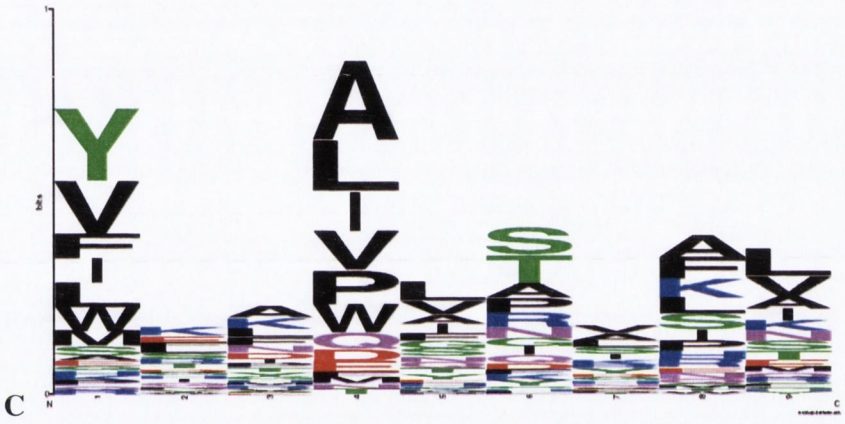
2.3.12.2 Alignment using motifs

Using the binding motif from the SYFPEITHI website (Rammensee et al., 1999) implemented as a regular expression, only 18 out of 134 potential core binding regions were identified. This would result in elimination of a large percentage of the available training data; therefore, it was decided to implement the motif as a binary position specific scoring matrix using scores of 1 for preferred residues and scores of 0 for all other residues. This matrix based approach allowed the selection of the most appropriate core region from all sequences.

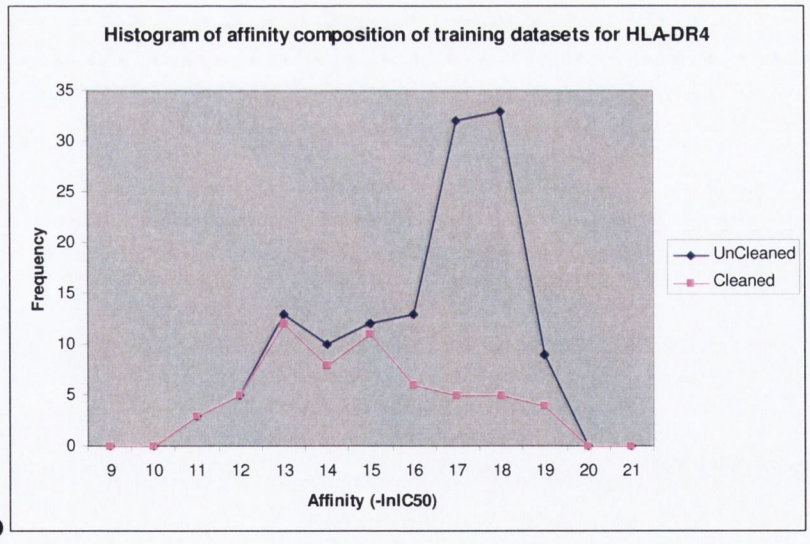
Visual inspection of the heatmap and dendrogram for the core region data (Figure 2-9, below) clearly shows the effect of repetitive sequence removal on the overall composition of the data. After processing using the wordmatching routine, the data comprised 59 unique core sequences. The Logo plotted from the resultant core region data is also shown below (Figure 2-9, below) and shows a clear information content for regions known to be important in antigen-MHC interaction. Additionally, it shows the effect of information based alignment and effective data cleaning when compared with the Clustalw based alignments.



A B



C



D

Figure 2-9 Characteristics of cleaned and raw HLA-DR4 training datasets

Heatmap and dendrogram of the HLA-DR4 core binding data pre and post processing (A and B). Logo Showing the information content of DR4Dw4 core regions aligned using a binary PSSM based on the defined binding motif after cleaning of over-represented data (C). The affinity histogram of the cleaned and uncleaned data is also shown (D)

2.3.13 Predictivity of algorithms for MHC class II binding using motif aligned training data

The correlation coefficients for predictions using each algorithm are shown in Table 2-10, below. The uncleaned dataset shows a clear advantage in terms of predictive abilities when compared to its processed counterpart. PLS in contrast to the other algorithms still retains a reasonable ability to render accurate affinity predictions when trained on the processed data. For reference, the Q^2 for QSAR model building are also included in the graph, however, they show little concordance with the abilities of the QSAR algorithm as measured by prediction of values in the test set. The pruning algorithm was also implemented in this set of analyses using the unprocessed data as a starting point. The pruning algorithm actually reduced the predictivity of the QSAR algorithm, although it did generate a greater model Q^2 value.

	Algorithm				
Dataset	**Gribskov**	**HMM**	**QSAR**	**Pruning**	**QSAR Q^2**
Raw	0.334	0.192	0.337	N	0.326
Cleaned	-0.040	0.060	0.283	N	0.011
Raw	-	-	0.206	Y	0.614

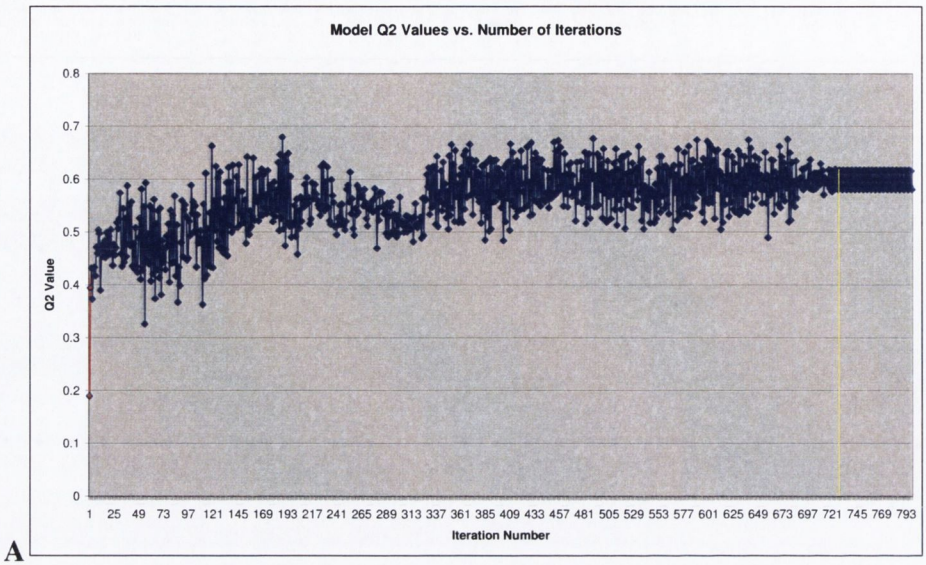
Table 2-10 Correlation coefficients of biological measurements and predictions as rendered using a combination of dataset variations and algorithms

Pearson correlation coefficients are shown for validation set predictions (n=42) of each algorithm trained using either cleaned or raw data. The Q^2 value from LOOCV is also included for reference and subsequent discussion. .

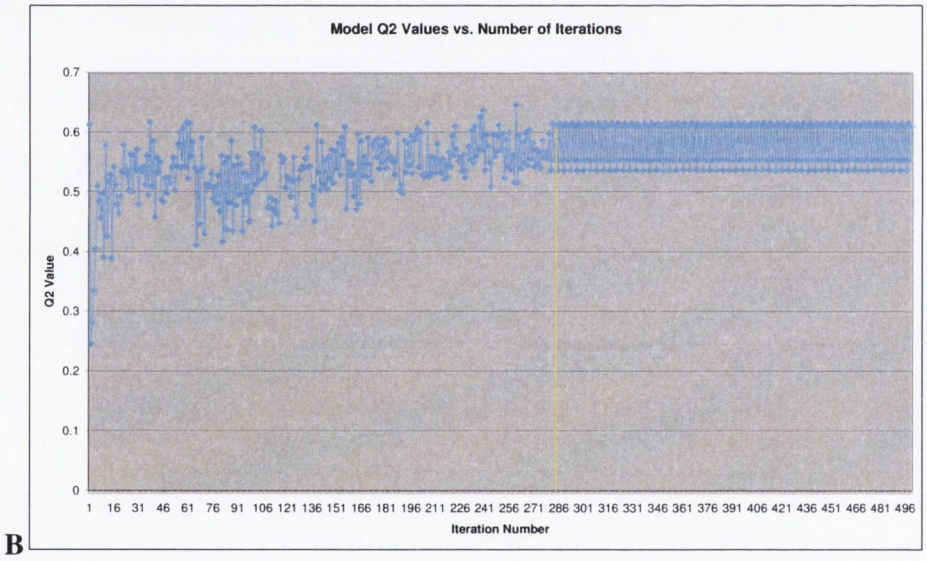
2.3.14 Evaluation of epitope predictions using a modified Iterative Self Consistent Algorithm

Both versions of the modified Iterative self consistent algorithm (i.e. selecting either the closest-to-*in-vitro* or highest numeric values as the likely nonameric core) were implemented and the fitness of each iteration as measured by model Q^2 plotted on a graph to illustrate the relationship (Figure 2-10, below). The first approach to be evaluated selected the highest scoring nonameric sub-sequence as the best match and failed to converge but instead fell into a repeating pattern after the 727^{th} iteration. The model building was therefore allowed to run for 727 iterations and the model generated at this iteration used to make predictions of the test set data. The correlation coefficient for these predictions (r=0.0018) can be thought to show no correlation given the number of validation data (n=42).

The second variation of the algorithm also fell into a repeating pattern at the 286^{th} iteration. At this iteration, the model was used to generate the test set predictions. As above the predictions rendered using the modified ISC algorithm failed to show any correlation (r=0.0009, n=42)



A



B

Figure 2-10 Graphs showing the Q^2 values of models built at successive iterations of a modified ISC algorithm

Graphs illustrate the Q^2 values of models built at successive iterations of a modified ISC algorithm. Two alternative approaches were used to select the most appropriate nonameric core at each iteration; the highest scoring sub-sequence (A), or the closest to the *in vitro* sequence (B). The correlation coefficients for the test set data were 0.0018 and 0.0009 respectively. The largest observable increase in model Q^2 value is between the first and second iteration for each approach.

2.4 Discussion

A crucial resource for any researcher working on epitope prediction is an expertly curated database of relevant sequence information as is provided in the AntiJen database (Toseland et al., 2005). This database contains binding data for a host of human and non-human MHC alleles and provided a suitable starting point for our evaluation and optimisation of MHC binding prediction algorithms. The first step in our analysis involved the formation of a set of data for both of our test alleles. The information content easily allows the selection of peptide affinities measured under similar laboratory conditions which considerably reduces bias in the training and validation sets. Other reports for which the data lacked this level of annotation forced researchers to adopt a more 'fuzzy' classification of binding abilities (Honeyman et al., 1998) which undoubtedly reduced the information content of the training data.

The second step in our dataset preparation involved removing varying levels of overlap between the elements of the dataset; this is a normal approach in motif based studies (Reche et al., 2002). The cut-off values represented the amount of common residues between each element and were an empirically determined trade-off between dataset size and information bias. The eventual approach chosen was to use two cleaned datasets, generated using two separate cut off values and a raw/uncleaned dataset. This approach would allow us to assess both dataset size and information overlap as determinants of binding model predictive quality. In order to account for the error and unmodelled variation in our training dataset we were also able to derive a series of modelled MHC class I binding data and use these data for optimisation of the QSAR algorithm. Modelled data was based on the independent binding of side-chains (IBS) hypothesis which has been shown to be a realistic approximation in a number of studies (Peters et al., 2003, Sturniolo et al., 1999); in this, interactions between adjacent residues were not modelled in the data.

The three MHC class I datasets (one raw and two cleaned) were used as a base from which to assess the quantitative predictive abilities of three different

algorithms. The first algorithm was based on Gribskov's profile analysis (Gribskov et al., 1987) and represented traditional profile based methods. The second algorithm utilised HMMs, an approach more rooted in machine learning methods (Brusic et al., 2002, Mamitsuka, 1998). The final approach utilised a PLS-QSAR method initially defined by Doytchinova *et al* (Doytchinova et al., 2002). This range of methods combined with the aforementioned datasets allowed the reliable examination of the strengths and weaknesses of the various approaches. An optimisation procedure was also carried out where possible to investigate whether the methods could benefit from tweaking to suit the chosen application.

The initial examination of the prediction routines showed the PLS algorithm with included pruning routine to be the superior method when using the uncleaned training set. However, it was subsequently noted that a number of elements in the training sets were also represented in the test set. This was remedied by removing the common elements from the training set and re-running the validation. This resulted in a significant alteration of the 'leader board' and showed the HMM based algorithm coupled with the largest dataset to be the best routine for HLA-A2 prediction. With smaller datasets the Gribskov algorithm proved the most successful. This 'success' is a likely consequence of its ability to extrapolate from simple amino acid frequencies by coupling them with biochemical and evolutionary information from the PAM120 matrix. A somewhat surprising finding was the lack of negative effect when using uncleaned data. The approach of removing bias – in the form of overrepresented residues - from the training dataset has been used in many studies (Bui et al., 2005, Reche et al., 2002). In the case of the Gribskov algorithm it is likely that normalisation per column i.e. dividing by the number of epitopes being tested, likely diluted the impact of overrepresented residues as the sample number increased. It is also possible that the over-represented data is similarly over-represented in the training set. However, the inclusion of common data was controlled for, making this scenario less likely. It is also probable that molecules of high affinity for a given allele are more commonly over-represented and thus the bias is actually beneficial in terms of *in vitro* prediction. The QSAR algorithm is not subject to such bias as its base in the

field of QSAR means it is designed to equate changes in molecular composition with corresponding changes in IC_{50} values. The downside to this method is the inability to incorporate biochemical information from substitution matrices which renders it less adept a generating strong predictions when training datasets are small and lack data overlap.

The lack of data overlap immediately stands out as being detrimental for QSAR model building. This is due to the ability of overlapping data to reduce the likelihood of random error being modelled as biological effect. Similarly, this occurs with artificial neural networks which in some cases contain up to 9000 interconnections and not enough data to cover each interconnection even once (Peters et al., 2003). This is also likely to have been to be the case for Doytchinova *et al* (Doytchinova et al., 2002) when modelling interactions between the three adjacent amino acids of MHC binding peptides - as evidenced by the low model Q^2 value despite good predictivity. The removal of such interaction data generated more accurate predictions for Peters *et al* (Peters et al., 2003). However, in terms of extending current statistical methods beyond the current motif based methods, the additive method would represent the next most likely approach, as it allows one extend the number of interconnections (i.e. to incorporate interactions between adjacent residues) in tandem with increases in dataset size. Additionally, the adjacent interactions of amino acid side chains are thought to affect the conformation of the peptide while outside of the binding groove (Doytchinova et al., 2002) and are therefore not allele specific. Thus, a generalised model of side chain interactions may be generated which could be incorporated into binding models for all alleles, thus, reducing the requirement for a large volume of data for each separate model. This selective information would not be possible with ANN based approaches used to-date, which keep such data in hidden layers.

The next step in the analysis involved optimising algorithms where possible, to attempt to improve the predictivity of *in vitro* MHC affinity. This was thought to be particularly relevant for the profile methods which are optimised for detection of evolutionary divergent sequences as opposed to prediction of binding affinity. The first step in the profile algorithm optimisation was the

assessment of the optimal normalisation method. Normalising by column (number of residues) would effectively represent the enrichment for certain residues for each relative position. The alternative method of normalising by row (number of occurrences of a particular amino acid) would serve to highlight at which position certain residues are preferred. For all datasets the column normalisation was shown to be the best approach by a large margin. The next factor to be optimised was the choice of substitution matrix. The default matrices used in profile analyses are the BLOSUM and PAM matrices which are well suited to detection of homologous proteins (Gribskov et al., 1987, Altschul et al., 1990, Henikoff and Henikoff, 1996). However, to our knowledge no attempt has been made to determine if the information content of these matrices is the best suited to the task of epitope prediction. To this end we obtained a selection of different matrices from the AAIndex website (Kawashima and Kanehisa, 2000) which were determined to represent a sufficient span of amino acid similarity measures (Section 2.3.7). The initial assessment of each matrix was tested using LOOCV and the optimal matrix selected based on this value. Of the ten matrices tested the MCLA720101 chemical similarity matrix proved the most successful at this stage. As chemical composition is arguably the sole determinant of peptide-MHC affinity, the suitability of MCLA720101 for the task of epitope prediction would be almost expected. The relatively poor performance of the PAM and BLOSUM matrices is not surprising as they are derived from the substitution probabilities of evolutionarily conserved sequences and chemical structure is unlikely to be the sole determinant of evolutionary amino acid substitution rates. Unsurprisingly, the worst performing matrix was the GRAR740104 chemical distance matrix i.e. the opposite of our best performing matrix, which serves to reinforce the relevance of chemical similarity to peptide-MHC binding. The validation stage of the assessment was to determine the ability of our chosen substitution matrix to predict the binding affinities of our validation dataset. In this task, the chosen matrix (MCLA720101) again, outperformed the majority of matrices, reinforcing the appropriateness of the selection.

Optimisation of the HMM algorithm was not feasible as it involved the use of compiled software (Eddy, 1998) which would prove difficult to reengineer. As the remaining algorithm (QSAR) was an in-house implementation, it could be

optimised. This optimisation was first performed using the modelled MHC class I binding data. This optimisation served to inform our choice of model building parameters based on the abilities of the statistical method, before application to biological data. The first optimisation step assessed the impact of dataset composition on the resultant models and their ability to generate quantitative predictions of validation set binding affinities. Surprisingly, the high affinity dataset generated the worst predictions a likely consequence of high affinity peptides requiring a rather limited repertoire of residues at key binding positions. This is also an unusual finding in that preselection of high affinity peptides is a normal factor in some algorithms. The medium and low affinity datasets did not perform as poorly, but similarly failed to generate very high R^2 values. The optimally composed datasets were the mixed-affinity and single-substitution which differed remarkably in their Q^2 values. The single substitution based dataset was generated from a series of single amino substituted analogues of a modelled MHC class I binding peptide which would cover every residue at each of the nine pockets only once. The low Q^2 values observed for this peptide are unsurprising as these figures are calculated using LOOCV and if the majority of residues at each pocket are present only once, each LOOCV training set has absent data for the test peptide at each iteration, leading to less accurate predictions. Thus, we can see that the impact of residues missing from the training set can have serious ramifications for prediction accuracy. This also brings into question the utility of the Q^2 statistic as a globally reliable indicator of model robustness especially given the prevalence of substituted peptide data in the epitope databases. On datasets such as the mixed affinity dataset the Q^2 value may lend a lot to the comprehension of model fit as there is a high degree of overlap between peptides and thus predictions may be more reliable. A real world example of this phenomenon is evidenced by the predicted affinities of the class II data (Table 2-9). The processed data contains both redundant and non-redundant training datasets. The non-redundant training set generates a poor model Q^2 value (0.011) yet generates the best validation score (r=0.283) for any of the algorithms trained using the cleaned training data. However, the uncleaned dataset exhibits both high model Q^2 values and reasonable validation correlation figures (r=0.337). The implementation of training set pruning is

thus similarly flawed as the tendency to increase the model Q^2 by removing outliers from the dataset may result in removing data which may make the external predictions less robust. This factor is confirmed by the work of Doytchinova *et al* who report that a number of the peptides removed from the training set because of high residual error, had missing values (e.g. 20% - 38%) and were thus poorly predicted (Doytchinova et al., 2002). This was also the case in our own analyses where pruning of the training set consistently reduced the applicability of the subsequent model to the validation set (Table 2-5, Table 2-9, Table 2-10). This was especially true for the HLA-DR4 model which gave a very high Q^2 value (0.614), yet generated a relatively low correlation coefficient (r=0.208).

The next assessment of the QSAR algorithm evaluated the impact of the number of latent variables and the %CV in the predictive abilities of the model. The use of modelled data was also necessary here as no biological dataset exists which would allow such a comparison. The results of this analysis showed that large numbers of components would only serve to increase the R^2 values generated using the validation dataset if the measurements were free of random variation – a factor that is impossible in practice. However, the results did illustrate that a low number of components would generate the best predictions of our validation data when random variation was modelled into the training data. This is most likely due to a large number of components modelling noise as signal, and thus overfitting the model to the training data. From this analysis it was decided to reduce the number of components to three to allow for experimental error, yet still allow the algorithm to account for factors not present in the modelled data.

The re-evaluation of our test algorithms demonstrated the validity of the optimisations especially for the Gribskov algorithm which, as established earlier, clearly benefits from the use of chemical as opposed to evolutionary information (Section 2.3.7). The most notable example of this increased predictive ability can be found in the model built using the smallest training set in which the correlation between predicted and experimental affinities increased from 0.522 to 0.637. Conversely, the correlations for the smaller

datasets decreased marginally when using the QSAR algorithm optimisations but increased marginally for the larger uncleaned dataset. Irrespective of dataset size the Gribskov method seems to give consistent predictions, which further reinforces the appropriate nature of the matrix information content for epitope prediction. Additionally, as the substitution matrix is not specific to a single allele it is likely to be equally useful for other alleles. In spite of the variable differences in the QSAR algorithm as a result of optimisations, it was elected to restrict the number of components to three for the remainder of the analysis as the difference in predicted affinity correlation coefficients was slight and it was considered more appropriate to model as little biological error as possible.

The focus of the work was then turned to class II epitope prediction which also presents its own experimental issues. Firstly, the IC_{50} measurement is an indication of the propensity of the entire test peptide to affect the binding of a labelled indicator to an MHC molecule of interest. Thus, the only method by which one can locate the core binding regions is via 'walking' of short peptide sequence affinity measurements which span the length of the parent sequence. This approach is not always performed as it is unnecessary for a number of studies. Thus, one must often work with peptides for which core binding regions have yet to be determined.

Many approaches have been utilised to locate the binding regions of MHC class II binding peptides from motif discovery programs such as MEME (Reche et al., 2002) to more custom designed approaches (Doytchinova and Flower, 2003). In this study we assessed three methods; Clustalw (Chenna et al., 2003) a program used in traditional multiple sequence alignment, the ISC algorithm defined by Doytchinova *et al* (Doytchinova and Flower, 2003), and motif based alignments. The initial assessment was that of the two traditional methods Clustalw and motif alignments. Implemented as a motif and scanned using a regular expression the motif method failed to highlight core regions in many of our known binders – only 18 of our 134 known binders contained a motif match. To circumvent this obvious lack of sensitivity the motif was implemented as a binary position specific scoring matrix which would identify

the most likely core region based on the established motif. The information content of each approach was evaluated using sequence logos. From examination of the sequence logos it was evident that the information content of the motif based selection routine was far superior. This was possibly due in most part to the motif based method being specific for the task of core region localisation which resulted in block alignments, in contrast to the Clustal results which gave a more distributed alignment. Additionally, the Clustalw method assumes that the sequences which it is aligning are divergent; this assumption places constraints of the information return from an alignment which appear to be incompatible with the task of epitope prediction. Moreover, a similar lack of transferability to MHC binding was also reported by Chang *et al* (Chang et al., 2007). To circumvent such limitations Reche *et al* used the motif discovery program MEME to create block alignments for MHC class II binding peptides (Reche et al., 2004).

Data from the motif alignment were cleaned to remove overrepresented data in the form of single substituted peptides and synthetic analogues. From this, both the cleaned and uncleaned data were used to build binding models using all three algorithms. For the cleaned dataset the only algorithm to generate a reasonable set of predictions was the PLS algorithm which can generate robust models with low affinity data. This observation is particularly important as the majority of peptides in the training dataset were of low affinity. This would lead the profile based algorithm to create a model in which the major representatives of peptide-MHC interaction were of low affinity. In contrast, the PLS algorithm explicitly models the effects of sequence on affinity meaning that even low affinity data can provide suitable training material. The HMM algorithm performed rather poorly for both cleaned and uncleaned data - a likely result of HMMs requiring larger amounts of data for epitope prediction (Brusic et al., 2004). The clear-cut benefits of the QSAR algorithm were less pronounced for the uncleaned dataset as both it and the Gribskov based algorithm performed comparably. This can be explained by examination of the effects of dataset cleaning (Figure 2-9). The cleaned dataset can be seen to contain very little high affinity data, however, the uncleaned dataset has a greater number of high affinity sequences which exhibit a high degree of

sequence similarity. This does not seem to adversely bias the profile method as the normalisation method emphasises the most abundant sequences which are high affinity. This emphasis of high affinity sequences leads to less reliance on the low affinity data. In fact one could argue that presented with a minimal of high affinity sequences and an appropriate substitution matrix, a reasonably robust model could be generated for practically any MHC molecule. Yet one cannot be guaranteed such assumptions apply to all datasets. Accordingly, a more affinity oriented approach such as the PLS algorithm would generate the more reliable model should sufficient data be available.

The final stage of the study assessed the Iterative Self Consistent (ICS) algorithm originally defined by Doytchinova *et al* (Doytchinova and Flower, 2003). The inability of this method to generate a reasonably robust model of peptide-MHC affinity is most likely due to the omission of a requirement for a bulky-hydrophobic residue at position 1, as required in the initial article. This step when initiated by Doytchinova *et al* was akin to a loose motif based alignment which allowed a second level of selection to choose the most appropriate motif. The in-house implementation of this algorithm also ran for a greater number of iterations than the original implementation as a likely result of the alterations made to the algorithm. Based on this lack of self consistency without preselection the ICS algorithm was not thought to be as suitable for our chosen task as the motif alignment coupled with the QSAR or GPA algorithms.

As stated by Flower (Flower, 2003b) sequence-based methods are limited by available data. The limitations of publicly available data range from simple inter-laboratory bias to a more insidious bias towards over-representation of motif influenced sequences. However, until such time as sufficient computer time and power are easily available to run routine atomistic-based analyses, the use of statistical methods is likely to remain the norm.

2.5 Summary

In order to best characterise putatively antigenic regions from prolamin proteins it was decided to generate a quantitative peptide-HLA-DQ2 binding model. Thus, it was necessary to first establish the applicability of a variety of algorithms to the task of quantitative affinity prediction. Our analyses demonstrated that each algorithm has individual strengths and weaknesses. However, it was recognised that the re-optimised Gribskov's profile analysis-based algorithm was particularly advantageous when the number of data available was low. Alternatively, the QSAR based algorithm performed better with larger datasets, especially if the training dataset was composed of substituted analogues. Accordingly, the choice of algorithm in following chapters would be a function of the available data.

3 Examination of the relationship between peptide length and MHC class II affinity

3.1 Introduction

The comparatively poor peptide-MHC class II binding affinity predictions found previously (Section 2.3.13) has necessitated a more thorough examination of factors that differ between class I and class II peptide-MHC interactions. Because MHC class II molecules have an open ended binding groove they do not restrict the length of peptides accommodated within the groove. A number of studies have found some influence on binding and T cell recognition from residues outside of the main nonameric core.

Residues which fall outside of the core nonameric binding region but within an extended binding groove have been referred to as peptide-flanking residues (PFRs) (Moudgil et al., 1998). The importance of PFRs for T cell recognition has been well investigated. It is a widely accepted fact that these flanking regions can contribute to T cell recognition of a peptide-MHC complex (Arnold et al., 2002). Arnold and colleagues reported their findings with respect to T cell recognition of PFRs showing that certain T cell responses were completely dependent on the residues at P-1 and P11. It has also been shown that the chemical properties of the residues located at these crucial recognition sites could influence T cell recognition with residues capable of salt-bridge or hydrogen bonding with the TCR most favoured (Arnold et al., 2002). Similar findings were reported by Stepniak *et al* for immunostimulatory HLA-DQ2 binding peptides derived from gliadin and glutenin proteins (Stepniak et al., 2005). The findings of both authors can be partially explained by the results of X-ray crystallographic studies of the peptide-MHC class II-TCR trimolecular complex which demonstrated that the CDR3 regions of the TCR α and β chains tend to locate over the p5 residue in the binding groove(Hennecke et al., 2000, Hennecke and Wiley, 2002).

It has also been demonstrated that PFRs contribute strongly to peptide-MHC stability in addition to providing an increased measurable affinity for the MHC binding groove, in particular the residue at position P-1 (Nelson et al., 1993). Nelson *et al* also reported that the effect of increased stability was greater for an increased peptide length of six amino acids compared to four. They also

found the effect to differ depending on which terminus was elongated. Furthermore, a previous study by Srinivasan *et al* showed an increased affinity as a result of peptide elongation on a single study peptide (Srinivasan et al., 1993). Yet, the exact mechanism by which such peptide elongations would serve to increase affinity remains unclear. Although it is theoretically possible that the ragged termini may interact with the MHC molecules outside the groove, such a schema has not been demonstrated, nor has it been observed in X-ray crystallographic structures.

In addition to the PFRs and sequence length being able to affect MHC class II affinity the number of binding registers within a peptide has also been demonstrated to play a role in the immune response. Nanda *et al* were able to uncover evidence which suggested T cells could recognise multiple registers in a singular peptide (Nanda et al., 1995). Prior to the study by Nanda *et al*, biphasic dissociation kinetics for a selection of MHC class II binding peptides had been reported (Witt and McConnell, 1994). Following that, a number of authors have studied the properties of multiple register binding (Bankovich et al., 2004). One such study by McFarland *et al* demonstrated that two distinct binding registers within a single 16 residue peptide were capable of stimulating a T cell response *in vitro* (McFarland et al., 1999). Both Seamons *et al* (Seamons et al., 2003) and He *et al* (He et al., 2002) have also studied the capacity of register shifting to influence the immune response. Both studies focussed on the reactivity to myelin basic protein in a murine model of experimental allergic encephalomyelitis (EAE) and found compelling evidence for the presence of multiple binding registers as crucial determinants of T cell reactivity. Shan *et al* also found evidence for multiple register binding in a single peptide when examining the proteolytic fragments of gliadin digests. Their studies located a 33 residue peptide which contained six partially overlapping copies of known T cell stimulatory peptides. When examining the potency of the interaction between this 33-mer peptide and the HLA-DQ2 molecule Xia *et al* found no evidence for the observed high affinity being caused by formation of complexes between peptide and more than one MHC class II molecule (Xia et al., 2005). This finding would suggest the formation

of a dimeric peptide-MHC class II complex to be the normal means of interaction.

To-date, the majority or peptide MHC class II predictions do not specifically account for the length of the peptide in either the test or training set (Reche et al., 2004, Sturniolo et al., 1999). Doytchinova *et al* stratified test data into residues of greater than or less than 16 residues in addition to selecting only peptides of less than 16 residues in length for the training set. However, the study by Doytchinova failed to return a correlation coefficient as large as that obtained using class I binding data. We hypothesised that the length of a peptide binding to an MHC class II molecule has an impact on the reported binding affinity. To further this examination we elected to test two alternative hypotheses for peptide binding to HLA-DQ2; the single binding register (SBR) model which assumes the single highest affinity nonameric core within a peptide determines its capacity to be naturally presented, and the multiple binding register (MBR) model which considers multiple nonameric binders within a peptide to be a better indicator of its ability to be naturally presented. In order to assess this model and test its applicability to the problem of peptide-MHC class II binding prediction we implemented a qualitative sliding window algorithm as an application of the MBR model. The results of this predictive algorithm were compared with those of two comparable algorithms based on a SBR model.

3.2 Methods

3.2.1 Dataset formation

Data for the study were obtained by querying the AntiJen database for all sequences know to bind human MHC class II molecules. These data were imported into a spreadsheet and then further filtered to exclude those data for which no radiometrically determined IC_{50} value was present. IC_{50} values were then transformed to their $-\ln IC_{50}$. This database was then output to a text file containing epitope sequence, -lnIC50 value, pubmed ID of the source article, experimental conditions and the MHC binding allele. This database formed the source data for interrogation using an in-house algorithm.

3.2.2 Data query algorithm

To analyse the data it was decided to search for peptides for which elongation events were recorded. To minimise the possibility of experimental error and different experimental conditions, elongation events were only searched for within single publications. In order to perform this search a python algorithm was implemented to search with a single publication for any instances of affinity determination for a sequence and a longer counterpart.

To maximise the information return from the available data a 'greedy' algorithm was implemented in tandem with a 'non-greedy' algorithm to search for elongation events. The greedy algorithm (Figure 3-1, below) serves to increase the information return by iteratively considering each peptide in order of increasing length as the query peptide and then searching within the publication for longer counterparts. This serves to consider elongations of the shortest peptide to be as valid as those of interim elongations. The non-greedy algorithm omits from the search space any peptide identified as part of a previous elongation event.

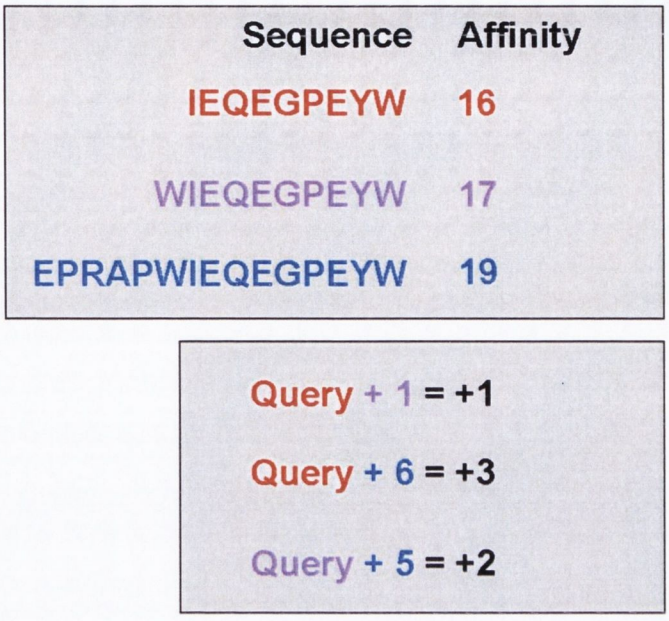


Figure 3-1: Illustration of the elongation event search algorithm.

The upper section contains theoretical affinity data for a set of peptides within a single publication. The lower section shows the differences in peptide length and associated changes in affinity (colour coded by peptide). The algorithm searches within a publication for any instance of a sequence exhibiting a longer counterpart. The algorithm then returns the length difference and affinity difference for each recorded elongation event (i.e. Query+1=+1, and Query +6=+3). In the 'greedy' algorithm interim elongation peptides are also used as query peptides which can increase the return of information for a set of peptide elongations (Query+5=+2).

3.2.3 Statistical analysis of elongation data

All statistical analyses were performed using the Minitab® software package (Minitab Inc) for statistical analysis. Discrete datasets were isolated for particular analyses such as the chi-square analysis using Microsoft Excel®. The usage of different foundation datasets is highlighted in the results section.

3.2.4 Evaluation of the single binding register (SBR) model of MHC class II presentation

3.2.4.1 Curation of a set of naturally processed HLA-DQ2 binding peptides

A set of naturally processed peptide sequences and their experimentally determined core binding regions were provided by Prof F. Koning (Leiden University Medical Centre, Leiden, Holland). The sequences of these peptides were obtained by eluting bound peptides from purified MHC as described by van de Wal *et al* (van de Wal et al., 1996). The sequences of the eluted peptides were determined using mass spectrometry. The principal binding determinant for each eluted sequence was determined by truncation studies as previously performed by van de Wal *et al* (van de Wal et al., 1996).

In order to determine the parent sequence of each of the eluted peptides BLAST searches were performed against all human proteins in the Refseq database using the NCBI BLAST server (http://www.ncbi.nlm.nih.gov/BLAST). To account for the usage of short sequences the E-value threshold was increased to >10000. Parent proteins were selected as the top match to contain 100% sequence identity with the query peptide.

3.2.5 Implementation of a novel sliding window based epitope prediction algorithm

The sliding window algorithm was developed as an extension of the single binding register (SBR) based additive method. Predictions were performed using the PSSM derived in section 4.2.6 and the method defined in section 2.2. This approach generates a series of affinities for each potential nonameric sequence in the parent protein. The predicted nonameric affinities were then used as the basis for the sliding window analysis (Figure 3-2, below).

For the sliding window approach the algorithm was designed to iteratively calculate the number of potential nonameric binders in each window based on a predefined cut-off threshold and window size. For example, in a window

containing 5 overlapping nonamers, if two of these nonamers had predicted binding affinities above the threshold the window score is returned as 2. Mathematically, this calculation can be represented as:

$$S(l,a) = \sum_{b=1}^{l} A(a+b) \text{; where } A(a+b) > c,$$

Where:

$S(l,a)$ is the score for a window for length l overlapping nonamers beginning at position a in a protein sequence,

$A(x)$ is the affinity for a nonameric sequence beginning at position x in our protein of interest,

c is the affinity cut-off value.

In order to identify those windows which contained an overrepresentation of nonamers, the mean and standard deviation was calculated for the returned window scores of each full protein sequence. Each window score was then assigned a classification based on whether the window score was greater than either 1 or 2 standard deviations above the mean. This approach allowed the use two levels of stringency when assessing the algorithm.

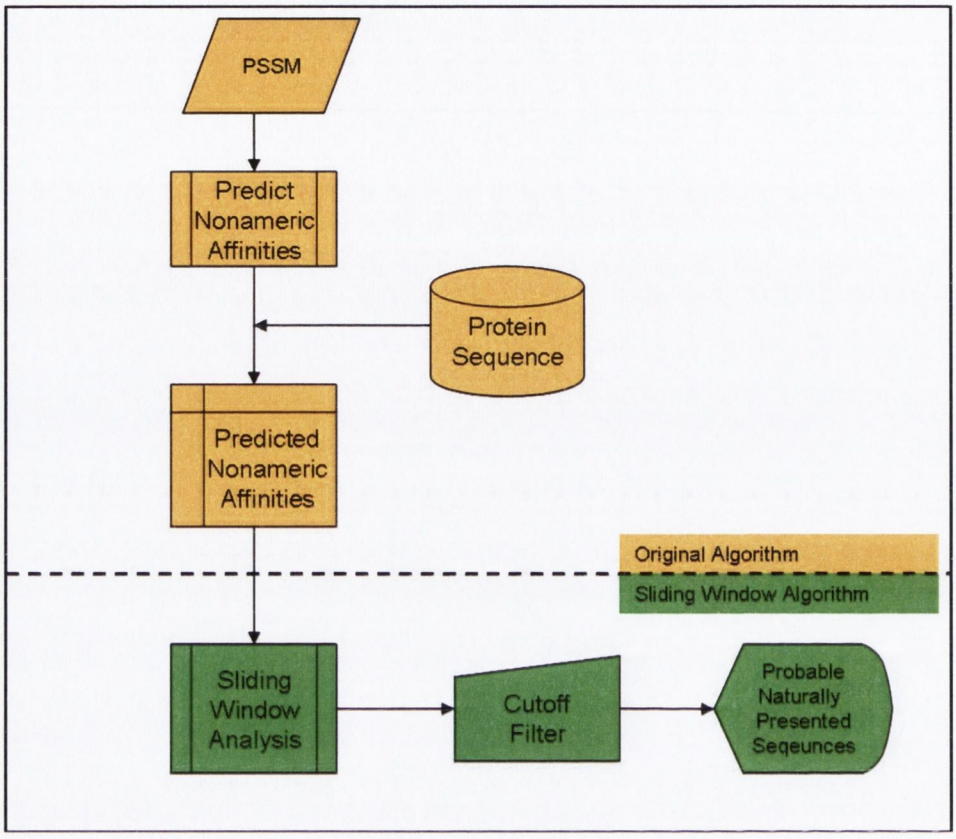


Figure 3-2: Algorithmic layout of the sliding window method for MHC class II epitope prediction.

The original method is outlined above the line (orange) while the sliding window algorithm extension is shown below the line (green).

3.2.6 Epitope prediction for characterised HLA-DQ2 binders

Prediction of putative antigenic regions in the characterised HLA-DQ2 binders was performed using the PLS derived matrix in both SBR and sliding window formats. As a comparison, these methods were also compared to the predictions rendered by the RANKPEP server (Reche et al., 2004). The RANKPEP server was chosen as a comparative method because it is one of the few services to offer HLA-DQ2 epitope prediction. Additionally, when benchmarking their iterative self consistent algorithm Doytchinova *et al* found that, of the services analysed, it correctly identified more regions as being antigenic than either SYFPEITHI, MHC-Thread or ProPred (Doytchinova and Flower, 2003).

3.2.6.1 Detection of binding sequences using the sliding window algorithm

Predicted binding regions were defined using the sliding window algorithm as outlined above (Section 3.2.5). The percentage of the nonameric sub-sequences from the parent peptide covered at 2 standard deviations above the mean was determined by obtaining the average coverage across all analysed. It was necessary to obtain these figures to obtain comparable nonameric sub-sequence coverage to the comparison methods.

3.2.6.2 Detecting nonameric binders using the PLS derived model

Nonamer affinities were predicted using the method described in section 2.2.12 and the HLA-DQ2 binding model created in section 4.2.6. In order to achieve comparable coverage to the sliding window algorithm the top 5% of predicted affinities were chosen for comparison. The returned results were deemed to have correctly predicted a binder if the predicted core nonamer was contained within the decameric core of the experimentally determined peptide.

3.2.6.3 Rankpep

For each peptide in the dataset of naturally presented peptides the parent protein was submitted to the Rankpep server (http://bio.dfci.harvard.edu/Tools/rankpep.html) for analysis using the HLA-DQ2 prediction matrix. A cutoff value of 5% was used, which selected the top 5% of predicted binders. The returned results were deemed to have correctly predicted a binder if the predicted core nonamer was contained within the decameric core of the experimentally determined peptide.

3.2.7 Evaluation of the single binding register (SBR) model of MHC binding by algorithm comparison

In order to evaluate the single binding register (SBR) model in peptide-MHC class II interactions two approaches were utilised to predict naturally presented molecules known to bind HLA-DQ2. Predictions based on the SBR model

were made using both the RANKPEP server and the QSAR model generated below (section 4.3.5.2). Predictions based on the multiple binding register (MBR) model were made using a novel sliding window based algorithm (Section 3.2.5). Two factors necessitated controlling for in the evaluation. Firstly, each method would have a different likelihood of predicting the correct core region by chance if the percentage of nonameric sub-sequences varied between prediction methods. Thus, each method was to be allowed an equal percentage of nonamers in the returned predictions. However, the characterised core regions from the naturally presented peptides were decamers and thus contained two nonameric sub-sequences (i.e. a sequence contains n-8 nonameric sub-sequences). This would render the SBR methods twice as likely to correctly identify the core region by chance. To counter this it was decided to equalise the methods by allowing the MBR based method a two-fold greater percentage nonamer coverage. Thus, the MBR based method was allowed approximately 10% nonamer coverage while the SBR based methods were allowed 5% coverage. This ensured an equal likelihood for each method of predicting the core region by chance.

When examining results it was found that where equal percentages of residues were covered, the sliding window method outperformed both SBR-based methods. For this reason a more formal evaluation was carried out and each method was also evaluated to return the top predictions covering both 18 and 45 percent of residues in the query protein. These predictions were then evaluated to determine the proportion of correctly identified core decameric peptides at each level of sequence coverage.

3.3 Results

3.3.1 Dataset Characteristics

The initial list of epitopes known to bind MHC class II molecules consisted of 2193 peptides binding 36 distinct alleles. The majority of alleles represented were of the HLA-DR class (Table 3-1, below). Within the dataset DRB1*0401 was the most prominent allele and 25.76 percent of the characterised epitopes were known to bind to it. Data were present from the three major MHC class II subclasses i.e. HLA-DR, -DP, and –DQ.

For comparison, the greedy and non-greedy search algorithms were both used to form separate datasets. Analysis of the MHC class II binding data using both the greedy and non-greedy algorithms generated datasets of 1279 and 275 elongation events, respectively. For both datasets the data covered 19 distinct HLA alleles, with alleles from the DR, DP and DQ classes (Figure 3-3, Figure 3-4, Table 3-2, Table 3-3). From these findings the advantages of using the greedy algorithm are evident. Histograms of both datasets' affinity compositions are provided as an initial indication as to the general effects of peptide elongation. Visual examination of the affinity histogram from the greedy dataset (Figure 3-3) shows a pattern similar to a bell curve with an elongated right tail, showing a definite trend towards increased affinity across the observed binding events. Hereafter all calculations were performed using data generated from the greedy algorithm to optimise the amount of available information.

Allele	Count	Percent
DPB1*0401	25	1.14
DPB1*0402	25	1.14
DQA1*0101/DQB1*0505	1	0.05
DQA1*0102/DQB1*0602	1	0.05
DQA1*0103/DQB1*0603	1	0.05
DQA1*0201/DQB1*0201	2	0.09
DQA1*0203	1	0.05
DQA1*0301	1	0.05
DQA1*0301/DQB1*0301	43	1.96
DQA1*0301/DQB1*0302	23	1.05
DQA1*0301/DQB2*0302	1	0.05
DQA1*0302	1	0.05
DQA1*0302/DQB1*0401	23	1.05
DQA1*0501/DQB1*0201	124	5.65
DQA1*0501/DQB1*0301	1	0.05
DQB1*0302	20	0.91
DRB1*0101	268	12.22
DRB1*0201	2	0.09
DRB1*0301	180	8.21
DRB1*0401	565	25.76
DRB1*0402	1	0.05
DRB1*0404	59	2.69
DRB1*0405	78	3.56
DRB1*0406	10	0.46
DRB1*0801	22	1
DRB1*0802	52	2.37
DRB1*0803	3	0.14
DRB1*0901	49	2.23
DRB1*1101	84	3.83
DRB1*1201	28	1.28
DRB1*1302	50	2.28
DRB1*1501	222	10.12
DRB3*0101	26	1.19
DRB4*0101	63	2.87
DRB5*0101	137	6.25
	N=2193	

Table 3-1: Composition of the peptide HLA binding dataset

Count represents the number of peptide binding data for each allele, percent is the percentage of total peptide data

Allele	Count	Percent
DPB1*0401	38	2.97
DPB1*0402	38	2.97
DQA1*0301/DQB1*0301	64	5
DQA1*0501/DQB1*0201	114	8.91
DRB1*0101	124	9.7
DRB1*0301	1	0.08
DRB1*0401	77	6.02
DRB1*0404	2	0.16
DRB1*0405	3	0.23
DRB1*0802	3	0.23
DRB1*0803	2	0.16
DRB1*0901	3	0.23
DRB1*1101	3	0.23
DRB1*1201	3	0.23
DRB1*1302	3	0.23
DRB1*1501	682	53.32
DRB3*0101	24	1.88
DRB4*0101	1	0.08
DRB5*0101	94	7.35
	N=1279	

Table 3-2: Composition of the dataset of elongation events compiled using the greedy algorithm.

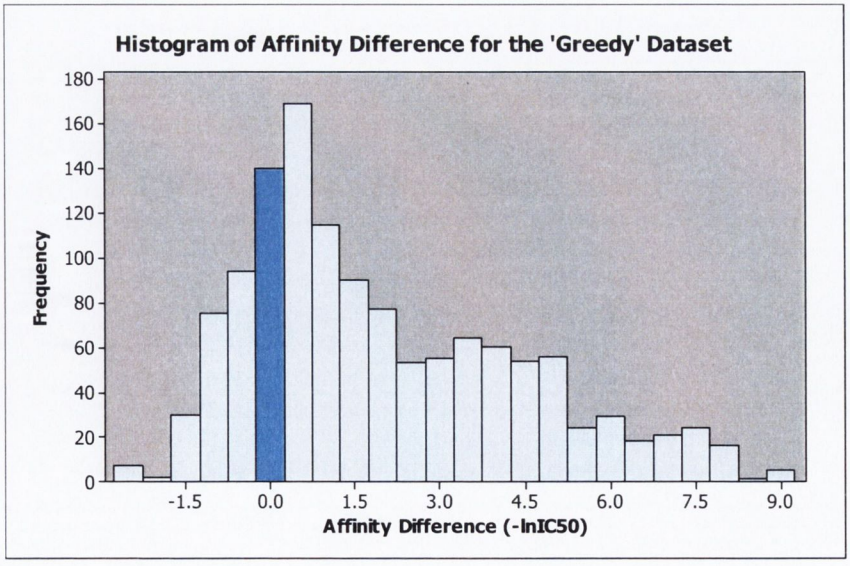


Figure 3-3: Histogram of the affinity differences observed in the elongation dataset compiled using the 'greedy' algorithm.

Allele	Count	Percent
DPB1*0401	8	3.14
DPB1*0402	8	3.14
DQA1*0301/DQB1*0301	18	7.06
DQA1*0501/DQB1*0201	40	15.69
DRB1*0101	34	13.33
DRB1*0401	30	11.76
DRB1*0404	2	0.78
DRB1*0405	3	1.18
DRB1*0802	3	1.18
DRB1*0803	2	0.78
DRB1*0901	3	1.18
DRB1*1101	3	1.18
DRB1*1201	3	1.18
DRB1*1302	3	1.18
DRB1*1501	61	23.92
DRB3*0101	7	2.75
DRB4*0101	1	0.39
DRB5*0101	25	9.8
	N=255	

Table 3-3: Composition of the dataset of elongation events compiled using the non-'greedy' algorithm

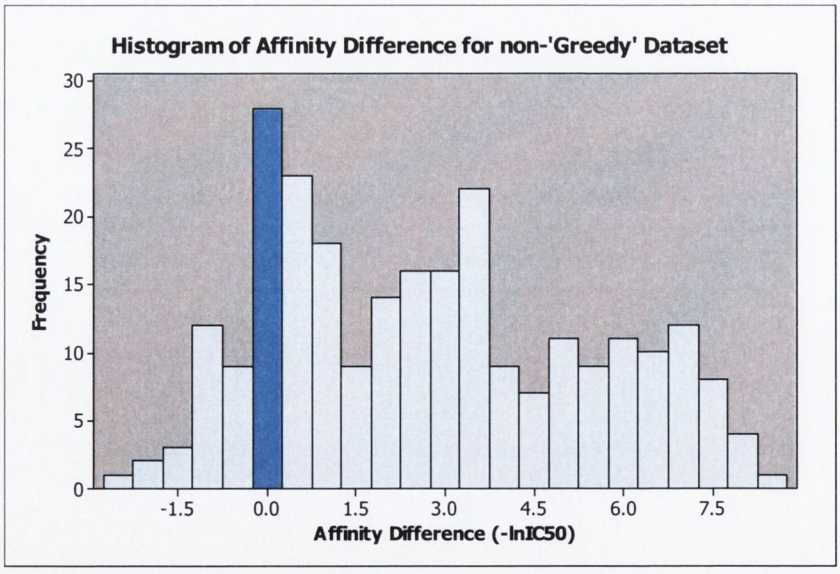


Figure 3-4: Histogram of the affinity differences observed in the elongation dataset compiled using the non-'greedy' algorithm.

3.3.2 Characterising the relationship between increases in sequence length and MHC class II affinity

In order to best represent the relationship between increases in peptide length and reported binding affinity, simple summary statistics would be insufficient as they would fail to account for the differences in principal binding determinants and experimental conditions. Therefore, when examining the relationship between peptide length increases and affinity, it was necessary to utilise a different measure of variance. For this purpose the mean affinity increase for each class of peptide elongations (e.g. 1 or two residue additions) was iteratively calculated. For each iteration, the data from one publication was omitted from the calculation and the mean value for each class returned. The relationship was then plotted as the mean +/- 2 standard deviations of the returned mean value of each iteration (Figure 3-5). This serves to reduce the impact of multiple sources of potential bias on the data while still allowing visual examination of the reliance of values on select publications.

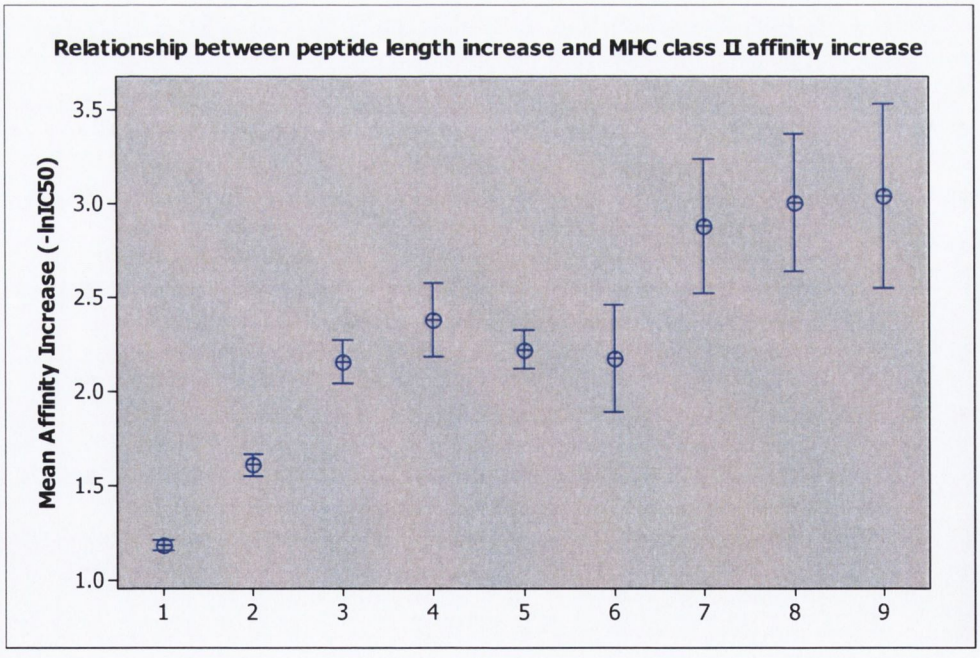


Figure 3-5: Relationship between sequence elongation length and alterations in affinity. The number of amino acids by which the peptide length was increased is compared with the associated difference in affinity. On average it can be seen that a larger increase in length will return a larger increase in affinity. Fold changes are represented using their natural logarithm on the vertical axis.

To further examine this association of peptide length and affinity the relationship between final peptide length and affinity was stratified by length of increase and a regression line for each class of peptide length increase plotted. As insufficient data were available for additions of greater than 5 residues, only increases of between one and five residues were analysed. A clearly observable relationship is evident across the different classes of peptide length increase, showing a decreasing effect as a result of peptide elongation as a peptide increases in length. From the plotted graph (Figure 3-6, below), one can also see that an approximate point exists (at 18-20 residues in length), beyond which the regression lines suggest no advantage in peptide elongation.

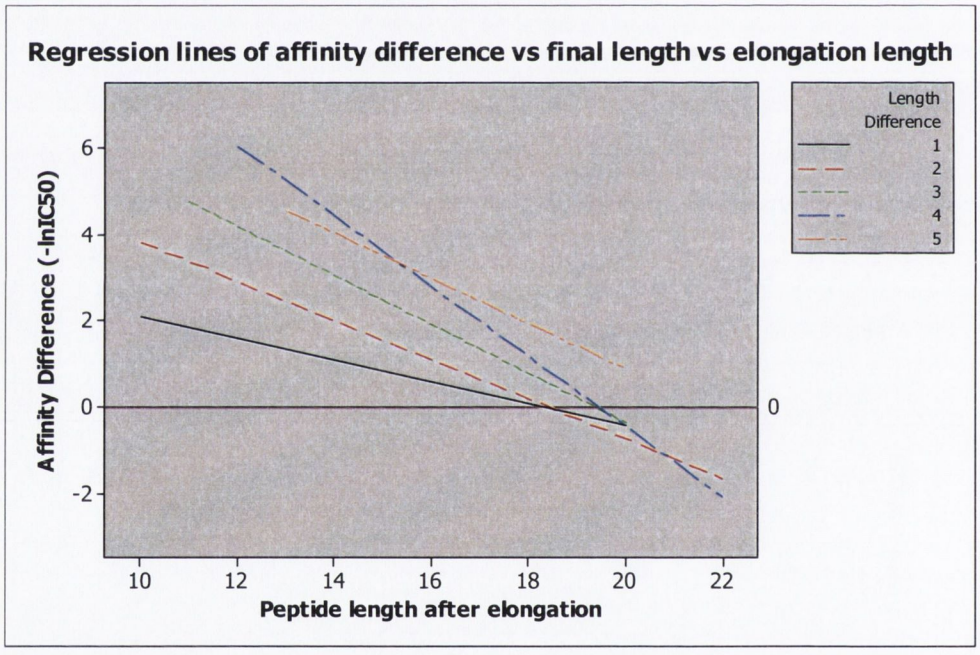


Figure 3-6: Regression plot of classes of peptide length additions.

A separate regression line relating observed difference in affinity to final peptide length for each class of peptide elongation (i.e. different length increases) is illustrated.

3.3.3 Chi Square Analysis

To examine whether the observed effects of increased peptide length on the reported affinity were specific to either terminus in the elongation set a chi-square analysis was performed. For the purposes of this analysis all elongation events were classified by whether the event resulted in either a positive or negative on the reported binding affinity. Additionally, events were classified based on whether they occurred at the amino or carboxy terminus or both. A summary of the observed incidence of terminus additions and positive and negative effects on affinity are shown in Table 3-4 (below).

	Amino	Both	Carboxy	All
Decreased (Observed)	120	80	92	292
Decreased (Expected)	120.3	81.5	90.2	292
Increased (Observed)	407	277	303	987
Increased (Expected)	406.7	275.5	304.8	987
All	527	357	395	1279

Table 3-4: Chi Square Table for the observed elongation events.

Rows show effect (Affinity Increase or Decrease), whereas, the columns show the terminus at which the elongation was observed.

The chi-square analysis showed no significant difference to exist between the increases or decreases in peptide affinity based on the terminus at which the addition was recorded ($p=0.959$).

Dataset of peptides eluted from HLA-DQ2

Decameric Core	AccessionNo.	Description
FMANIPLLLY	NM_001042466	* prosaposin isoform c preproprotein
YEPVLIEILV	NM_001042466	* prosaposin isoform c preproprotein
FRAVTELGRP	XM_001124749	PREDICTED: hypothetical protein
YVALDFEQEM	NM_001083538	protein expressed in prostate, ovary, testis, and placenta 2
YSADLPLPLP	NM_024423.1	desmocollin 3 isoform Dsc3b preproprotein
VTAEELESYM	NM_016176.2	calcium binding protein Cab45 precursor
VVAGDEWLFE	NM_017458.2	major vault protein
EVAAEEPNAA	NM_007355.2	heat shock 90kDa protein 1, beta
IIPIQEEEEE	NM_021950.3	membrane-spanning 4-domains, subfamily A, member 1
ATAEEEEDFG	NM_178014.2	tubulin, beta polypeptide
VEEEAEEPYE	NM_201414.1	amyloid beta A4 protein precursor, isoform c
EISLVGDDLD	NM_014718.3	calsyntenin 3
VISDDEPGYD	NM_000194.1	hypoxanthine phosphoribosyltransferase 1
IADASEDQVF	NM_003105.3	sortilin-related receptor containing LDLR class A repeats preproprotein
EAAVEEEEEE	NM_003299.1	tumor rejection antigen (gp96) 1
EEAEEEEEEE	XM_037523.12	PREDICTED: similar to SET domain c ontaining 1A
IAPVAEEEAT	NM_002300.4	lactate dehydrogenase B
IAAENEEDEH	NM_004309.3	Rho GDP dissociation inhibitor (GDI) alpha
IAAEIEHFIH	NM_001001586	Na+/K+ -ATPase alpha 1 subunit isoform b proprotein
AEPELEEELF	NM_001243.3	tumor necrosis factor receptor superfamily, member 8 isoform 1 precursor

Table 3-5: Decameric core regions of peptides experimentally determined to be naturally presented by HLA-DQ2.

The Refseq accession number and description of the parent peptides are also shown. *Two of the naturally processed peptides originate from the same parent protein

3.3.4 Comparison of SBR and MBR based methods

Evaluation of the two models by comparative analysis of the SBR and MBR based predictive algorithms revealed a lower sensitivity for the MBR based algorithm (Table 3-6). This would tend to suggest the SBR model to be the more appropriate for prediction of our test data. Instances of true and false positives by algorithm and sequence are also outlined in Table 3-8.

	Sensitivity
Sliding Window	0.5
QSAR-SBR	0.6
Rankpep	0.6

Table 3-6 Comparative sensitivity of the SBR and MBR based methods

Sensitiviies of the three different algorithms are presented as the proportion of decameric core peptides detected. While the two SBR-based algorithms detect 60% of peptides in the test set, the MBR-based sliding window algorithm detects only 50%.

Interestingly, the sliding window algorithm had the unaccounted for effect of increasing method sensitivity when predictions were normalised to the percentage of residues covered by an algorithm. By examining the sensitivity of the algorithms at 18 and 45 percent residue coverage the proportion the sliding window algorithm outperformed SBR based algorithms (Table 3-7, Figure 3-7). Of the two SBR based methods the QSAR based method outperformed Rankpep at 18% coverage.

Protein Coverage (%)	18	45
Sliding Window	0.5	0.9
QSAR-SBR	0.45	0.6
Rankpep	0.3	0.6
	Sensitivity	

Table 3-7: Algorithm sensitivities at 18% and 45% residue coverage.

The proportion of true positives detected by each algorithm at 18 and 45 percentage sequence residue coverage is outlined in the above table. The sensitivities are denoted by the proportion of true positives detected. (n=20). The optimal approach at each level of coverage is highlighted in green.

Core Decapeptide	QSAR 5%	Rankpep 5%	SW-2SD
FMANIPLLLY	Y	N	N
YEPVLIEILV	Y	N	N
FRAVTELGRP	N	N	N
YVALDFEQEM	N	N	Y
YSADLPLPLP	Y	Y	Y
VTAEELESYM	Y	Y	Y
VVAGDEWLFE	N	N	N
EVAAEEPNAA	Y	Y	Y
IIPIQEEEEE	N	Y	N
ATAEEEEDFG	Y	Y	Y
VEEEAEEPYE	Y	Y	Y
EISLVGDDLD	N	N	N
VISDDEPGYD	N	Y	N
IADASEDQVF	N	N	N
EAAVEEEEEE	Y	Y	Y
EEAEEEEEEE	N	N	N
IAPVAEEEAT	Y	Y	Y
IAAENEEDEH	Y	Y	Y
IAAEIEHFIH	Y	Y	N
AEPELEEELF	Y	Y	Y

Table 3-8: Breakdown of naturally processed peptide detection by core decapeptide, algorithm and percentage sequence coverage.

True positives are highlighted in green, false negatives are highlighted in red.

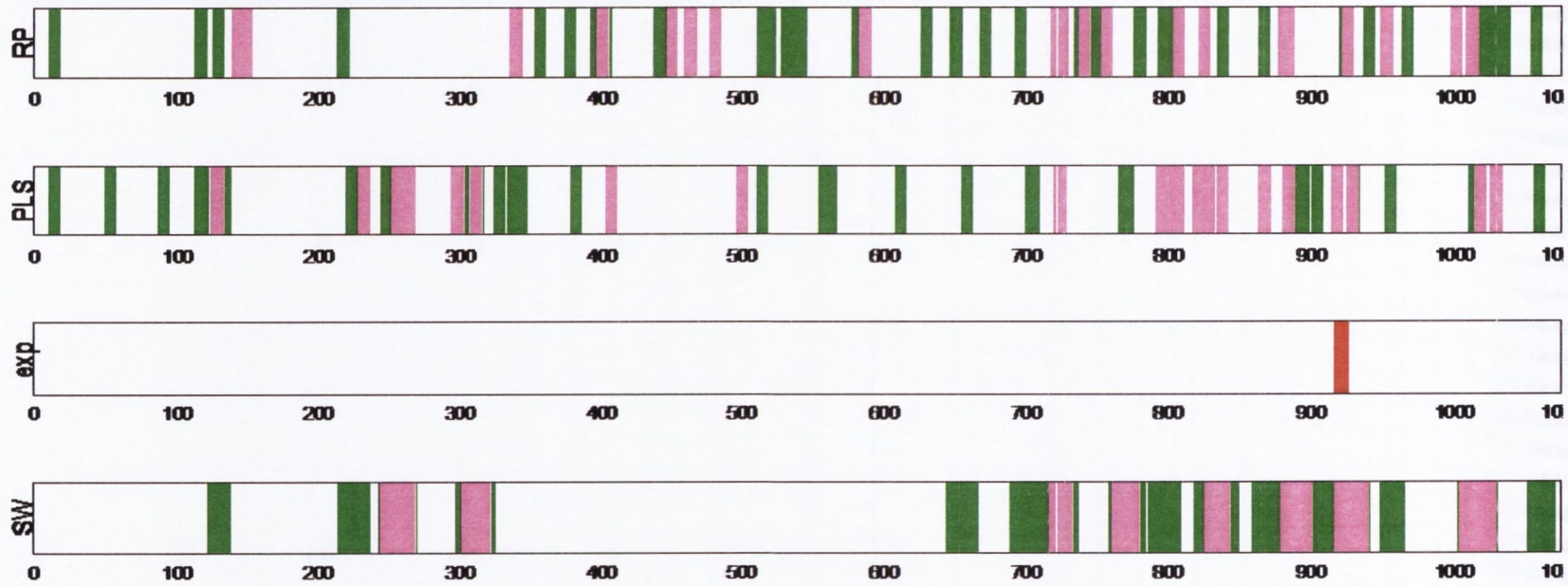


Figure 3-8: Spread of epitope coverage by algorithm compared to the experimentally determined core decameric sequence.

The above figure shows the locations of the regions highlighted as putatively antigenic by the RANKPEP server (RP), the PLS-QSAR algorithm (PLS) and the sliding window algorithm (SW). Areas highlighted at 18% coverage are coloured purple, whereas those highlighted at 45% coverage are coloured green. Relative location within the protein is demarcated by the numbers below each sequence representation. The experimentally determined decameric core region is highlighted in red

3.4 Discussion

When attempting to derive a relationship between sequence length and observed MHC affinity, one is limited by a number of factors. The most obvious and insurmountable of these factors is the presence of different principal binding determinants in the majority of sequences which would significantly bias any simple length-affinity correlation. That is, if one examines the affinity of two sequences of different length and low sequence similarity, it is impossible to ascertain whether any observed affinity difference is due to differences in either the principal binding determinant or the sequence length. Thus, in order to examine such a relationship one must normalise all factors except for sequence length before analysing its affect on affinity. The approach taken for this study was to search within publications for affinity measurements for a single sequence and an elongated equivalent. The benefits of this approach were twofold; firstly, inter-laboratory variation in affinity measurements would be excluded. Secondly, bias from variations in the principal binding determinant would also be reduced, meaning the bias would be removed from intra- but not inter- sequence comparisons. Bias from inter-sequence comparisons was unavoidable, but analyses were optimised for the required task so as to avoid representaion of bias as effect.

To begin our analysis, it was first necessary to extract any data fitting the selection criteria for a compiled dataset of MHC class II binding peptides. This data filtering was performed using an in-house algorithm implemented using the python language. In order to ensure that the greedy algorithm did not create any undesirable skew, data were generated using both a greedy and a non-greedy version of the algorithm. Results were compared by visual examination of the resultant affinity difference histograms. For both algorithms the datasets generated a predominately positive affinity difference with similar magnitude of spread. From this it was possible to conclude that neither algorithm generated any detrimental bias. Thus, the data generated from the greedy algorithm were chosen as the optimal source for further analysis; this was especially true given the size difference between the two datasets (n=1279 using the greedy algorithm vs. n=255 for the non-greedy algorithm).

Having compiled and chosen the optimal dataset, the available data were analysed to assess the impact of sequence length on measured affinity. The immediately obvious effect of peptide elongation was evident from the histogram (Figure 3-3), which showed the majority of affinity differences to be positive. However, further analysis was required to ascertain the magnitude of the effect when controlling for length of elongation and peptide length after elongation.

The first such analysis required a visual measure of the relationship between the magnitude of the length increase and any increase in affinity. A direct comparison would be uninformative, as many of the sequences would be subject to bias from differing principal binding determinants and different final peptide lengths. For this reason, it was elected to generate a simple average for each length of increase and use a simple measure of how reliant the reported average was on particular publications. From this analysis (Figure 3-5) it can clearly be seen that for a larger increase in length the difference in affinity is - on average - a larger increase. As data were quite sparse for larger increases in length, the resultant dependence on particular publications can be easily noted (from the larger 'error' bars). As one might expect the average response is also not linear and seems to level out at larger increases in length. That is, the effect shows a ceiling which would be in agreement with the law of diminishing marginal returns. This would seem to be quite a logical outcome for such an effect as, if no ceiling existed for the effect, sequence length would be the most important determinant of peptide-MHC affinity.

The second analysis performed on the elongation data was designed to further examine the ceiling effect of peptide elongation. For this analysis it was decided to correlate the reported affinity difference with the peptide length after peptide elongation. For this purpose it was necessary to stratify the data based on the length of peptide elongation as our previous analysis showed this to be an important determinant of the observed affinity difference. The results of this examination are –unsurprisingly- in agreement with our previous finding that longer peptide additions result in a greater increase in affinity.

However, the correlation graph shows what would appear to be a clearly defined ceiling to the effect, at approximately 19 residues in length. This effect could not be examined any further as adequate data were not available. It is possible, however, that for the majority of MHC class II binding peptides attaining a certain length may lead to formation of secondary and tertiary structures which are less compatible with MHC class II binding, leading to the observed lack of effect for peptide length increases over a set length. An exception to this rule may be the 33-amino acid peptide initially describe by Shan *et al* (Shan et al., 2002) which is known to bind HLA-DQ2 and stimulate disease associated T cells. Despite the unusual length of this peptide it is likely it fails to impede its MHC class II binding abilities because of its proline rich nature which would help it form a type II polyproline helix in solution. Such a secondary structure is the normal conformation in the MHC class II binding groove (Stern et al., 1994). To further our analysis, a chi-square analysis was performed to establish whether the observed effect of increased peptide-MHC class II affinity was terminus specific. A preponderance of favourable interactions at any particular terminus might suggest that the peptide elongations at that particular terminus facilitated particularly favourable interactions either inside or outside of the binding groove. The results of the analysis showed no statistically significant bias to exist ($p=0.959$) pointing to a more general, terminus independent effect on affinity.

In this chapter we have also examined what we have termed the single binding register (SBR) model of peptide MHC class II interaction. The foundation of such a model is based heavily on common characteristics of class I and class II molecules, namely, the ability to bind peptides through favourable interactions between peptide amino acid side chains and 'pockets' in the MHC binding groove. One may consider this to be the default model for describing peptide-MHC class II binding as other possibilities are comparatively poorly described. Examination of the sequences and affinities of well described MHC class II binding peptides led us to question the SBR model. While the SBR model could explain the occurrence of motifs in many well characterised MHC class II binders (van de Wal et al., 1996, Stern et al., 1994, Kim et al., 2004), it falls short when we consider our observation that peptide length and MHC affinity

bear a demonstrable association. It also falls short of explaining the reported antigenicity of multiple registers within the same peptide (Bankovich et al., 2004, Seamons et al., 2003, He et al., 2002). We hypothesised that multiple binding registers in a single peptide can exhibit cooperativity in MHC binding. Tong *et al* found that peptides with multiple registers are quite common among high affinity peptides (Tong et al., 2006), which would lend further validity to this assertion. Two alternative hypotheses describing peptide binding to HLA-DQ2 were tested: the single binding register (SBR) model which assumes the single highest affinity nonameric core within a peptide determines its capacity to be naturally presented; and the multiple binding register (MBR) model which would imply that multiple nonameric binders within a peptide are a better indicator of its ability to be naturally presented.

The approach used in this assessment was similar to that of Peters *et al* who examined the independent binding of side-chains (IBS) hypothesis using a selection of algorithms which either agreed or disagreed with the IBS hypothesis (Peters et al., 2003). This evaluation was performed using two prediction algorithms based on the SBR model and a single novel algorithm based on the MBR model. The use of two SBR based algorithms served a dual purpose as it also allowed the comparison of the PLS-based methods and the motif-based algorithms used for the RANKPEP server. Comparing method sensitivities and specificities is a complicated process as one cannot be guaranteed that those areas predicted as non-binders are correctly identified as such. This is due to the lack of availability of data for non-binders within a particular protein. Therefore, in order to make our comparison more equitable it was elected to ensure that each algorithm made predictions that covered an approximately equal percentage of the nonameric sub-sequences present in the parent. Thus, each algorithm would have an equal probability of predicting the core decameric peptide by chance alone.

For our assessment both SBR based algorithms outperformed the MBR based sliding window algorithm. The finding that the MBR based model showed a lesser degree of sensitivity would tend to suggest the SBR model to be the better model of peptide-MHC class II interaction. This is especially true since

the same PSSM was used for both the SBR-based QSAR and sliding window predictions, making it less likely that the observed difference is a function of two different model building algorithms. In *A Brief History of Time*, Stephen Hawking states that "a theory is a good theory if it satisfies two requirements: It must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements, and it must make definite predictions about the results of future observations" (Hawking, 1988). Additionally, Hawking states that theories can only be thought of as being provisional, in that, if they fail to agree with recorded observations they must be reformed. It would appear from these results that the reformation of the SBR model in favour of the MBR model would be inappropriate. The MBR theory satisfies only one of the two criteria outlined by Hawking. That is, it agrees with current observations such as the reported incidence of multiple registers in high affinity peptides (Tong et al., 2006), and with studies which support the biological relevance of multiple registers in binding interactions (Bankovich et al., 2004, McFarland et al., 1999, Rabinowitz et al., 1997, Seamons et al., 2003, Viner et al., 1996, He et al., 2002, Nanda et al., 1995, Shan et al., 2002). However, its capacity for making predictions about the results of future observations, as evidenced by our comparison of methods, is not as effective as the current SBR model. It must also be stressed, however, that the algorithm based on the MBR model assumes that the aggregation of nonameric sub-sequences above a set threshold is a sufficient representation of the model. If the MBR model was indeed the *in vivo* method of peptide-MHC class II interaction other factors not incorporated into the sliding window algorithm such as the affinity of individual nonameric sub-sequences for the MHC molecule may be quite important. Thus, while our study has failed to provide evidence that would favour the MBR model additional studies such as affinity measurements with register-limited peptides would provide a more clear-cut picture of the significance of multiple registers within MHC class II binding peptides.

An unforeseen benefit of the sliding window algorithm was its ability to return a greater sensitivity of core peptide detection per percentage of residues covered. At both levels of protein coverage the SBR-based algorithms were

outperformed by the sliding window approach. In fact, the sliding window algorithm correctly predicted 90% of binders at 45% sequence coverage compared to only 60% for the QSAR and RANKPEP algorithms (Table 3-7). This effect is a likely consequence of the ability of the sliding window approach to condense the testable regions to areas of high nonamer content. Thus, for an equal amount of residues one gains more nonamer coverage and a higher probability of identifying the true antigenic regions. The spread of predictions in relation to the actual experimentally determined decameric core of one of the naturally processed peptides is presented in Figure 3-7. It can clearly be seen from this schema that the sliding window predictions condense areas of potential antigenicity to definite locations within the parent protein. This is in stark contrast to the comparison methods which were based upon SBR model and would require a greater amount of experimentation to confirm antigenicity. Implemented as part of a combined *in vitro* and *in silico* immunogenicity screening regime, the sliding window algorithm approach may have the capacity to increase workflow efficiency.

The implications of our findings spread into not only epitope prediction but also general immunology and vaccinology. To begin, the fact that increasing the length of a peptide can - within limits - increase its MHC class II affinity carries implications for those developing vaccines as it may provide a route to increase efficient peptide MHC interaction at little additional cost. For example, if addition of three residues to either terminus of a putative immunogenic peptide leads to a two-fold increase in peptide-MHC class II complex formation it would effectively reduce the molar peptide concentration necessary to observe a biological effect. Secondly, ability for a longer peptide to bind MHC class II in the course of natural antigen presentation would bias the peptide-MHC class II repertoire towards more recently degraded peptides which would presumably be longer. A mechanism such as this may prevent the body from recognising harmless ubiquitous peptides in the course of infection.

3.5 Summary

In this chapter the relationship between peptide length and MHC class II affinity was examined. Additionally, a new model of peptide MHC class II interaction was also proposed and tested. The results of our *in silico* analysis clearly demonstrated that, within limits, the elongation of a peptide can markedly increase affinity for MHC class II molecules. It was proposed that the mechanism for this effect may be an increased number of binding registers as a result of peptide elongation. This hypothesis was tested using available data but was not supported any more than the current model which assumes the interaction of a single binding register with the MHC class II binding groove. The results of these analyses mean that the length of peptides in training and test sets should be controlled for and recognised. Additionally, the use of epitope prediction algorithms from the first chapter, which assume the interaction of a single binding register with MHC class II, are appropriate.

4 Creation and evaluation of a resource for prediction of prolamin immunogenicity in coeliac disease

4.1 Introduction

As mentioned previously, the HLA-DQ2 mediated T-cell response to coeliac associated antigens is thought to represent a crucial prerequisite for development of the coeliac inflammatory process. Indeed, many studies have focussed on mapping these T cell stimulatory regions from prolamin proteins in order to better understand the pathophysiology and possibly develop approaches to a safe alternative to the standard gluten-free diet.

The initial characterisation of a HLA-DQ2 restricted T cell line from the gut of coeliac disease patients was performed by Sjostrom *et al* (Sjostrom et al., 1998). The authors of this article also concluded that gliadin deamidation was important for creation of active T cell epitopes. This deamidation-enhanced immunoreactivity was further explored by Vader *et al* who characterised additional T cell epitopes based on a tTG deamidation algorithm (Vader et al., 2003a). The same year as the article by Sjostrom *et al,* an article by van de Wal *et al* was published which characterised the first HLA-DQ8 restricted T cell epitope from small intestinal T cells of coeliac disease patients (van de Wal et al., 1998). Subsequently, van de Wal *et al* also described the involvement of glutenin in the coeliac immune response (van de Wal et al., 1999). These studies coupled with the characterised link between coeliac disease and HLA-DQ2 form the basis of our current understanding of coeliac disease pathogenesis. Numerous additional studies of T cell epitopes have also been conducted in coeliac disease research which have further extended the information available on the T cell mediated immune response. For example, Vader *et al* recognised and characterised the contribution of sequence similarity to the observed immune response to gliadin related proteins (Vader et al., 2003a). Qiao *et al* added further to the role of proline in the T cell mediated immune response by demonstrating the importance of proline spacing in characterised epitopes (Qiao et al., 2005). Arentz-Hansen *et al* successfully characterised the first HLA-DQ2 T cell epitope from avenin adding further support to the tentative classification of oats as unsafe for consumption by patients with coeliac disease (Arentz-Hansen et al., 2004). The exclusivity of the immune response to well characterised canonical epitopes was further ruled

out by Vader *et al* who characterised a diverse T cell response in childhood coeliac disease (Vader et al., 2002b). This last finding has lead some to consider epitope focussing to be a feature of the developed CD inflammatory process (Martucci and Corazza, 2002). All of the studies of T cell epitopes have added to the general level of understanding of coeliac disease. However, a further contribution to the physiological understanding of the pathophysiology was provided by Shan *et al* (Shan et al., 2002, Shan et al., 2005). The authors of both studies characterised peptides from the gliadin proteins resistant to proteolytic digestion by normal brush border enzymes. The resistance of these peptides to digestion was related to their proline content. This relationship shed more light on the findings of Arentz-Hansen *et al,* who demonstrated that T cell epitopes of gliadins were predominantly located within proline rich regions (Arentz-Hansen et al., 2002, Arentz-Hansen et al., 2000a). Thus reactivity to coeliac disease prolamins was characterised as being the product of prolamin primary structure i.e. through the content and spacing of proline residues, peptide-HLA-DQ2 affinity and propensity for deamidation by tTG.

The characterisation of the link between antigenicity and prolamin structure has provided the potential for alternative treatments to the strict gluten free diet. Among the potential therapeutic options suggested are HLA-DQ2 blockers, dietary prolyl endopeptidase supplementation and tissue transglutaminase inhibitors (Sollid and Khosla, 2005). Any of these options could potentially result in a reliable pharmacologic treatment. However, an alternative to the current strict gluten free diet may be possible through reducing the immunogenicity of current coeliac-'unfriendly' grains. Such a diet would provide the best of both worlds for coeliac disease patients as it would reduce the difficulty involved in maintaining a strict gluten free diet without incurring the almost unavoidable side effects of medication. Initial investigations to determine the viability of a low immunogenicity strain of wheat have been promising. For example, Molberg *et al* have been able to exploit the characterised evolution of common hexaploid bread wheat to localise immunodominant epitopes to the wheat chromosome D by use of T cell stimulation assays (Figure 4-1) (Molberg et al., 2005). Similarly, a separate exclusively bioinformatics study by van Herpen *et al,* found sets of

characterised stimulatory epitopes across all ancestral genomes (van Herpen et al., 2006). These two studies provide evidence that a well structured breeding strategy coupled with selection of the least immunogenic strains of parent grains could provide a suitable starting point for the creation of a 'coeliac-safe' bread wheat.

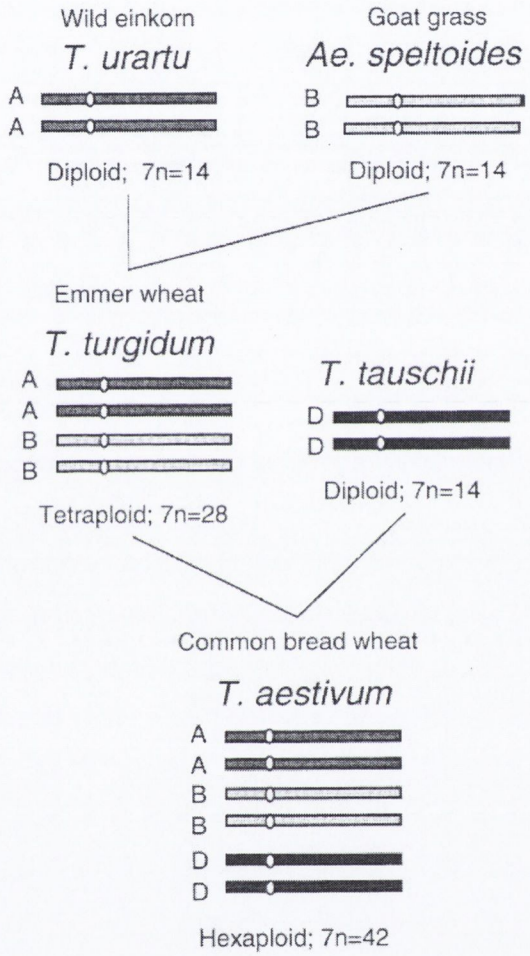


Figure 4-1 Simplified overview of the evolution of hexaploid wheat

Adapted from Molberg *et al* (Molberg et al., 2005)

To further the goal of selectively breeding and creating 'coeliac-safe' wheat it was envisaged that a computational resource may reduce the experimental burden by allowing researchers pre-select less immunogenic strains based on *in silico* screening. The goal of this chapter was to create and assess such a resource in the context of current knowledge of coeliac disease and prolamin immunogenicity.

4.2 Materials and Methods

4.2.1 Prolamin screening resource design

The front-end of the protein query resource was designed using Adobe Dreamweaver® (Adobe) with minor manual editing. The html-based front-end was linked to a query system using cgi as implemented in the Python language. The dynamic elements of the resource were coded using the Python scripting language. In order to facilitate updating of the resource or addition of novel characterised epitopes to the epitope database, the PSSMs and database of characterised epitopes were stored as text files instead of being coded into the final resource.

Filters and pre-processing steps were implemented as optional steps in the analysis and were selected to best represent the current processes of known relevance in coeliac disease pathogenesis. The predicted deamidation by tTG of selected glutamine residues was implemented using the algorithm defined by Vader *et al* (Vader et al., 2002a). A proline mask was implemented to eliminate the occurrence of proline residues at the relative positions 2, 4, 7 and 9 in predicted residues and was intended to replicate the observations of Qiao *et al* (Qiao et al., 2005). An additional filter was implemented to select only proline rich peptides - the exact number of proline residues to be determined by the user. The PSSM used for the HLA-DQ2 epitope predictions was generated in section 4.2.6.

The analysis on a query sequence was designed to focus on either characterised epitopes or novel predictions based on a combination of predicted affinity and selected filters as outlined below (Figure 4-2).

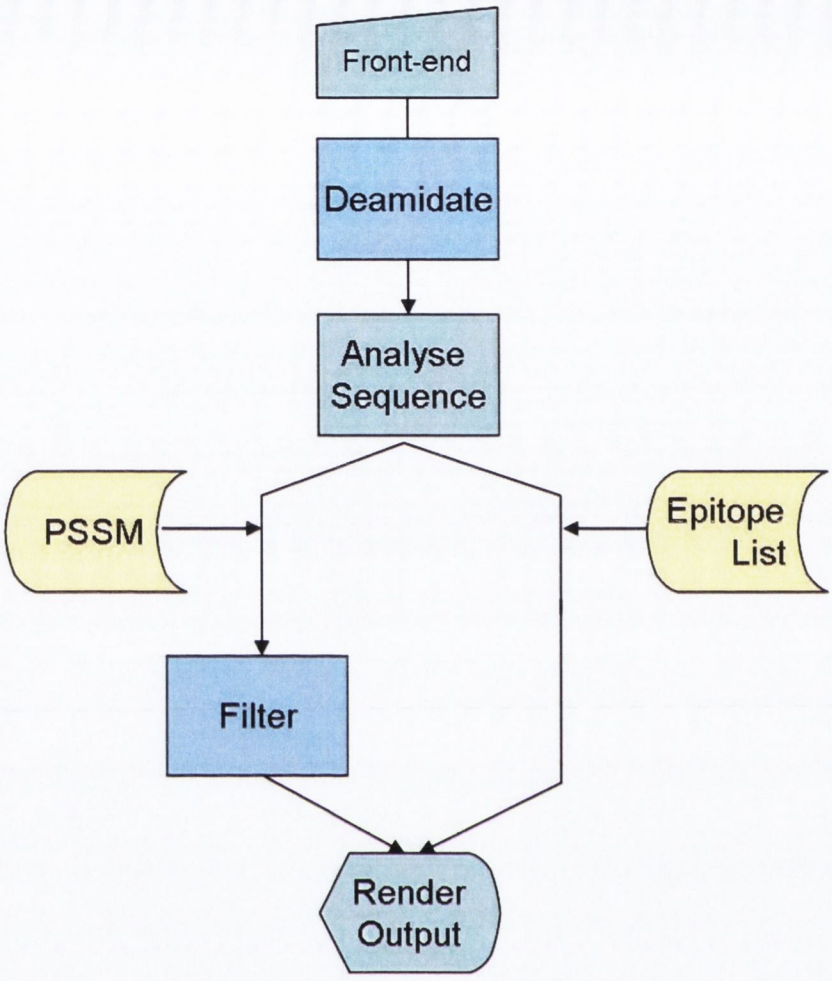


Figure 4-2: Flow chart of the prolamin screening resource.

Optional processes are highlighted in blue. The PSSMs and epitope list can be regularly updated and are highlighted in yellow. The 'Front-end' refers to the HTML based input page.

4.2.2 Formation of a database of known T-Cell stimulatory prolamin epitopes

An initial pubmed search was carried out for any articles containing the words 'epitope' and either 'celiac' or 'coeliac'. The resultant articles were downloaded as fulltext and screened for any sequences of known T-cell epitopes. Entries were added to the epitope list and included only peptide sequence and the pubmed ID of the source article. Where deamidation sites were indicated, the deamidated equivalent was also added to the database. Entries with ambiguous residues (e.g. X or B) at any site were not included in the database.

4.2.3 Creation of a database of prolamin proteins for prediction evaluation

The Uniprot resource was queried using the publicly available SRS service on the EBI web server (http://srs.ebi.ac.uk). The database was queried for sequences bearing the words 'gliadin', 'hordein', 'secalin' or 'avenin' in the sequence description. Sequences bearing the words 'precursor' or being less than 100 amino acids in length were omitted from the final search results as were results from species outside the Pooideae subfamily. The final sequences were output to a text file in FASTA format.

4.2.4 In-house implementation of a HLA-DQ2 binding assay

Attempts to implement an in-house method for HLA-DQ2 binding affinity prediction were complicated by a lack of a positive control for any assay. This therefore led to many variations in the methods described below. For the sake of brevity and because the development of an in-house assay was not the primary focus of this study, only the most frequently used variant of each method is described.

4.2.4.1 Cell culture

DUCAF and VAVY cell lines were obtained form the European Collection of Cell Cultures (http://www.ecacc.org.uk). These cells were chosen for their homozygous expression of HLA-DQ2 (DQA*0501, DQB*0201). Cells were cultured as per ECACC recommendations. Briefly, cells were cultured in 10% Heat Inactivated FBS in RPMI with addition of a penicillin/streptomycin solution (Gibco) is sterile vented tissue culture flasks (Falcon) in a 37°C incubator with 5% CO_2. Cells were maintained at a concentration of $1x10^5$ to $1x10^6$ cells per ml by addition of extra media and splitting of culture to separate sterile flasks. Cell viability counts were performed using an EB/AO method as described below (Section 4.2.4.2). HLA-DQ2 expression was validated using a flow cytometric method (Section 4.2.4.8, below).

4.2.4.2 Determination of viable cell concentration

The concentration of leukocytes in a solution was determined using a haemocytometer, and EB/AO stain. Diluted sample was applied under the coverslip of a Neubauer haemocytometer. Cells were visualised under the microscope using UV light, while a small degree of white light was used to visualise the slide markings. Viable cells (fluorescing green) were counted and their concentration per ml calculated.

4.2.4.3 Lysate purification

Cell lysates were extracted from both DUCAF and VAVY cell lines. Cells were first washed once by centrifugation in original media and resuspension in cold PBS, followed by a second centrifugation step and final resuspension in cold PBS. Throughout the remainder of the procedure all solutions were kept on ice and all centrifugation steps were preformed in a refrigerated centrifuge at 4°C. Cells were counted and resuspended at $4x10^6$cells/ml in lysis buffer (20mM Tris/HCl pH 7.5, 5mM MgCl2, 1% NP-40). Immediately prior to use 45µL of complete inhibitor mix (2ml dH2O + 1 complete protease inhibitor tablet (Roche)) was made and added for every ml of lysis buffer. The resuspended cells were placed on an agitator on ice for one hour. The cell solution was then spun down for twenty minutes at 4000rpm. The resultant supernatant was aliquoted and immediately frozen at -80°C.

4.2.4.4 Sandwich ELISA

100µL of a diluted antibody clone (SPV-L3) was added to appropriate wells of a certified Maxisorp™ plate (Nunc) using carbonate binding buffer at a concentration of 2µg/100µL. Plates were covered and incubated for 2 hours at 37°C. Wells were washed four times with 300µL of 0.05% Tween 20 in PBS per well using an automated washer (Thermo Scientific). To block the wells 200µL of 1%BSA in PBS was added to each well and incubated for 2 hours at 37°C. Wells were again washed four times with 300µL of 0.05% Tween20 in PBS. To each appropriate well, either 100µL of cell lysate, 50µL of cell lysate and 50uL of 1%BSA in PBS, or 100µL of 1% BSA in PBS were added. The plate was then covered and incubated overnight at 4°C. Wells were washed four times with 0.05% Tween20 in PBS. 100µL of a 1:100 dilution of the

biotinylated antibody clone 1a3 (Biodesign, now Meridian Life Science, Inc) in PBS was added to each well. The plate was covered and incubated for 2 hours at 37°C. After incubation the plate was washed four times with PBS-Tween and 100µL of TMB liquid substrate added to each well. The TMB solution was allowed develop for 15 minutes at room temperature before being stopped by addition of 50µL of 1M H_2SO_4. Absorbance for each well was measured at 620nm on a 96-well plate reader (Wallac).

4.2.4.5 Affinity Chromatography

Affinity chromatography was carried out using a combination of an in-house method (Byrne et al., 2007) and a method used by Sidney *et al* (Sidney et al., 2002). Briefly, Protein A agarose was equilibrated to room temperature before adding 2mls 50% gel slurry to an 8ml capacity reusable glass column (Pierce). Columns were equilibrated with 5 column volumes of PBS before application of 1ml of antibody diluted 1:100 in buffer (25mM Tris, 150mM NaCl, pH 7.2). Prior to addition of the cell lysate the column was equilibrated with 10ml of wash buffer (10mM Tris-Base pH8.0, 1%NP-40). Cell lysate was passed over the columns twice. Columns were then washed with 50ml of wash buffer and subsequently washed with 10 ml of detergent (PBS, pH 7.4, 0.4% n-octyl-D-glucopyransoide (Sigma)). MHC class II molecules were eluted into neutralising buffer (50mM diethlyamine, 150mM NaCl, 0.4% w/v n-octylglucoside, pH 11.5) by addition of 1ml of elution buffer (2M Glycine, pH 2.5). The neutralised solution was then concentrated using Centriprep 30 centrifugal concentrators (Centriprep). To ensure antibody had successfully bound the Protein A beads the neutralising buffer was also passed over the beads to elute any bound antibody. The eluted solution was also concentrated using Centriprep 30 centrifugal concentrators (Centriprep). All concentrated solutions were then run on standard SDS page gels which were stained using both EZblue (Sigma) or silver stain using a standard protocol.

4.2.4.6 Immunoprecipitation

Prior to conducting the assay, cell lysate was preconcentrated to increase the possibility of successful binding. 100µL of anti-HLA-DQ Ab (SPV-L3) was added to the concentrated cell lysate (0.5ml) and the solution incubated

overnight on an agitator at 4°C. 0.5ml of protein A slurry (Pierce) was added to the Ab-lysate solution and incubated for 2 hours at 4°C. Protein A agarose beads were then washed ten times in TBS by centrifugation at 2500xg and resuspended in an equal volume of SDS page loading buffer. In order to analyse the bound protein, the bead containing solution was boiled for 5 minutes to denature the protein and the solution loaded directly onto the SDS gel. Protein visualisation was performed using both EZblue (Sigma) and silver stain protocols.

4.2.4.7 Western Blotting

Western blotting was carried out as previously described by our group (Byrne et al., 2007). Briefly, an SDS-PAGE gel of the cell lysate was run and protein was blotted onto PVDF membranes (0.8mA/cm^2) using a semi-dry apparatus (Apollo). The membrane was blocked at 4°C overnight with 5% non-fat dried milk (Marvel) including 0.5% Tween (Sigma). The SPV-L3 and 1a3 antibodies (Biodesign) were diluted 1:1000 and 1:500 in 5% non-fat dried milk including 0.5% Tween and applied to the PVDF membrane. After 2 hours of incubation at room temperature and thorough washing with PBS containing 0.1% Tween, the membranes were incubated with a 1:1000 dilution of rabbit anti-mouse conjugated to horseradish peroxidase (DAKO). Blots were visualised using chromagenic substrate DAB (3,3'-Diaminobenzidine) and hydrogen peroxide.

4.2.4.8 Flow Cytometry

Flow cytometric analysis was performed on cultured cells to confirm expression of HLA-DQ2. Cells to be examined were removed from incubation and centrifuged at 1400rpm for 10 minutes and subsequently resuspended in FACS buffer (Becton Dickinson) at 10^7 - 10^6 cells/ml. 100µL of the cell suspension were added to a series of prelabelled 5ml tubes (Falcon). Cells were blocked with a 1:10 dilution of rabbit serum (Becton Dickinson) by addition of 5µL to each 100µL cell aliquot and incubation a room temperature for 30 minutes. Cells were spun at 1400rpm and resuspended in 100µL FACS buffer. 5µL of each appropriate antibody of Media/FACS buffer control were added to the resuspended cells and incubated for 30 minutes at 4°C. Cells were washed with 1ml of FACS buffer. Cells were resuspended in 100µL of a 1:50 dilution

of secondary antibody – FITC conjugated rabbit anti-mouse (Becton Dickinson). Cell solutions were incubated for 30 minutes at 4°C and washed twice in 1ml of FACS buffer before being resuspended in 500µL of fixing solution (Becton Dickinson). Samples were analysed using a Becton Dickinson FACScan and the mean fluorescence channel examined.

4.2.5 HLA-DQ2 binding assay

100µL of a diluted antibody clone SPV-L3 was added to appropriate wells of a certified Maxisorp™ plate (Nunc) using carbonate binding buffer at a concentration of 2µg/100µL. Plates were covered and incubated for 2 hours at 37°C. Wells were washed four times with 300µL of wash buffer (0.05% Tween20 in PBS) per well using an automated washer (Thermo Scientific). To block the wells 200µL of 1%BSA in PBS was added to each well and incubated for 2 hours at 37°C. Wells were again washed four times with 300µL of wash buffer per well. To each well 50µL of cell lysate (freshly thawed on ice) was added, together with 50uL of 1%BSA in PBS. The plate was then covered and incubated overnight at 4°C. Peptide dilutions were made in 10% DMSO. 50µL of peptide solution and 50µL binding buffer (Appendix J) were added to appropriate wells. Plates were covered and incubated for 48 hours at 37°C. Wells were washed four times with wash buffer. 100µL of a 1:1000 dilution of Europium-Streptavidin (PerkinElmer) in 1% BSA-TBS was added to each well and incubated for 45 minutes at room temperature with agitation. The plate was then washed four times with wash buffer and 150µL of enhancement solution (PerkinElmer) added to each well. Plates were then incubated on a shaker at room temperature for 15 minutes prior to being read on the manual input of an AutoDELFIA® plate reader (PerkinElmer).

4.2.6 Creation and Validation of a HLA-DQ2 binding model

As demonstrated previously (Chapter 3), the length of a sequence can have an effect on its reported affinity for a given MHC molecule. This would undoubtedly have repercussions for the chosen predictive routine. The PLS algorithms was chosen to interpret the available training data as it can readily

deal with a range of affinities, and can incorporate additional single residue analogues in order to increase predictivity. In contrast to the previous approaches, however, a much greater level of human intervention was utilised to create the DQ2 binding model, in order to account for the paucity of available data.

4.2.6.1 Data Selection and Processing

The HLA-DQ2 training set was initially compiled as for the HLA-DR4 dataset (Section 2.2.4). Briefly, data were selected to provide the largest dataset of peptides assayed under homogeneous experimental conditions. The largest collections of peptides available from the AntiJen database were tested against the radiolabelled peptide sequence: KPLLIIAEDVEGEY. The majority of peptides in this set were from analogue substitution studies and therefore best suited to analysis with the PLS algorithm. Peptides of greater than 11 or less than 9 amino acids in length were removed from the dataset to reduce any undesirable bias as a result of increased peptide length. Additionally, any peptides for which a truncated version was also present in the training set were also removed.

A set of data absent from the AntiJen database were those from a publication by Vader *et al* (Vader et al., 2003b). These peptides were tested under different experimental conditions than those of the main training set yet shared two common data points. The two common data points were used to construct a regression line in Microsoft Excel® and the resultant regression equation used to transform the data from the Vader *et al* report, to render it compatible with the remaining data. As this approach had not been previously evaluated two datasets were compiled for comparative purposes: one containing the data from the Vader *et al* paper and one without.

A validation peptide set was also generated. This set of validation data was formulated from data of peptides tested against the second most abundant indicator peptide (YPFIEQEGPEFFDQE). This dataset was also reduced in size to remove peptides of greater than 15 amino acids in length. A more

stringent cut-off value would have been implemented except it would have resulted in a smaller validation set.

4.2.6.2 HLA-DQ2 Model building

As reported above, the majority of peptides in the training set resulted from single residue substitution studies. For this reason all peptides in the training set were manually aligned. The model was generated using the QSAR algorithm with three latent variables and without pruning. To account for missing values in the training set a system of amino acid grouping was applied similar to that of Bui *et al* (Table 4-1, below) (Bui et al., 2005). In this approach, any data missing as a result of absence from the training set had the singular compositions created as an average of the contributions of those group members present in the training set.

Residues	Common Property
STC	Small Polar
MILV	Hydrophobic
DE	Negatively Charged
RHK	Positively Charged
QN	Polar
FYW	Aromatic

Table 4-1: Groupings of amino acids used to account for missing data in the training dataset

Amino acid grouping are the same as those used by Bui *et al* (Bui *et al* , 2005)

4.2.6.3 HLA-DQ2 Model Validation

In order to validate the binding model the shortest peptides (15 residues or less) tested against the second most abundant indicator peptide were selected as outlined above (Section 4.2.6.1). For each peptide each overlapping nonameric core was assessed for its potential binding affinity for HLA-DQ2. From these predicted nonameric affinities the highest affinity was chosen as the representative affinity for the entire sequence. The correlation of these

predicted affinities with *in vitro* measures was then evaluated by using the Pearson correlation coefficient.

4.2.7 Validation of the affinity prediction and filtering algorithms for detection of characterised T cell stimulatory epitopes.

To assess the sensitivity of the affinity based prediction, coupled with the applied filters and predicted deamidation algorithm, the database of known HLA-DQ2 restricted T cell stimulatory peptides was used (see also section 4.2.2, above). To define an appropriate affinity cut-off for the affinity based predictions, a subset of known T cell epitopes was selected for optimisations (n=20) For each characterised epitope in the optimisation set, an affinity based prediction was made for each nonameric sub-sequence and the highest affinity prediction selected as the representative affinity. The mean and standard deviation of the predicted affinity measurements were then calculated and two cut-off levels (8.4 and 10) were determined based on the returned results i.e. 8.4 and 10 represent the mean minus two and one standard deviations, respectively. Those epitope sequences not used for optimisation were used to form the validation dataset (n=47).

Method sensitivity was assessed for a variety of selection criteria. Sensitivity was represented as the proportion of true positives identified from the validation epitope dataset. Evaluations of affinity and filter based selection criteria were performed using the prolamin screening resource. To determine the proportion of nonamers to pass the selection criteria, a concatenated sequence was formed using all of the sequences obtained in section 4.2.3. This sequence was analysed using the chosen selection criteria as implemented in the prolamin screening resource.

4.2.8 Screening the gliadin family of proteins for putative antigenic regions

The prolamin protein sequences obtained in section 4.2.3 were screened using the prolamin screening resource. Screening was optimised for specificity and

an affinity cut-off or 10 was implemented in tandem with a proline cutoff of 3. Results of screening were tabulated and visually related to protein of interest by use of a phylogenetic tree (section 4.2.9).

4.2.9 Phylogenetic tree generation

Protein sequences obtained in section 4.2.3, above, were firstly aligned using a locally implemented version of Clustalw (Chenna et al., 2003, Higgins and Sharp, 1988). The resultant alignment in clustalw format was left unedited and converted to Mega format using the supplied format conversion tool with the Mega software (Kumar et al., 2004). The MEGA3 software was also used for phylogenetic tree reconstruction. A bootstrap test of phylogeny was implemented using a maximum parsimony method; branches with a bootstrap score of less that 70 were collapsed to generate the final tree.

In order to visualise different characteristics of proteins as a result of analysis using the prolamin analysis resource, branches were condensed/expanded or coloured to suit the task and saved as jpeg files.

4.3 Results

4.3.1 Implementation of the prolamin screening resource

Implementation of the prolamin screening resource was overall successful. The resource was hosted on an internal server using the Apache http server. All queries were processed in less than a single second due to the optimised coding and design. Graphical representations of characterised epitopes were to an acceptable standard and hyperlinking to source articles executed correctly. The resource was tested using both Mozilla Firefox and Microsoft Internet Explorer and results displayed correctly in both browsers.

Location and sequence of characterised binders within the query protein

Graphical Overview

Select sequence entries to link to source article through pubmed

Table of characterised epitopes

A	Sequence	C
143	QQPFPQQPQQPFPQ	156
182	QQPFPQQPQQPFPQ	195
246	QQPFPQQPQQPFPQ	259
145	PFPQQPQQPF	154
184	PFPQQPQQPF	193
248	PFPQQPQQPF	257
295	PFPQQPQQPF	304

Select sequence entries to link to source article through pubmed

Figure 4-3 Example output from the characterised epitope search algorithm.

4.3.2 Formation of a database of known T-Cell stimulatory sequences and prolamin epitopes

The literature search known T-cell stimulatory epitopes yielded a total of 101 epitopes from six articles (5.4Appendix H). 67 of the epitopes were in native form with 34 being deamidated equivalents. The prolamin sequence search

returned 40 gliadin-like sequences from a variety of grain species, the accession numbers of which are represented in Figure 4-9.

4.3.3 Set-up and optimisation of a HLA-DQ2 binding assay

Setup and optimisation of an in-house binding assay was unsuccessful. A number of approaches were utilised to examine peptide-MHC affinity and/or purify HLA-DQ2 for use in such an assay. While, flow cytometric assays demonstrated the presence of HLA-DQ2 on the cell surface, additional assays such as the immunoprecipitation and affinity chromatography reactions failed to show its presence in purified cell lysate (Figure 4-4). The quantitative assays employed to detect HLA-DQ2 in cell lysates failed to demonstrate any detectable amount of protein. The sandwich ELISA failed to generate any signal that exceeded background and the western blot did not yield any detectable bands (results not shown). These results were true for cell lysate extracted from both the VAVY and DUCAF cell lines.

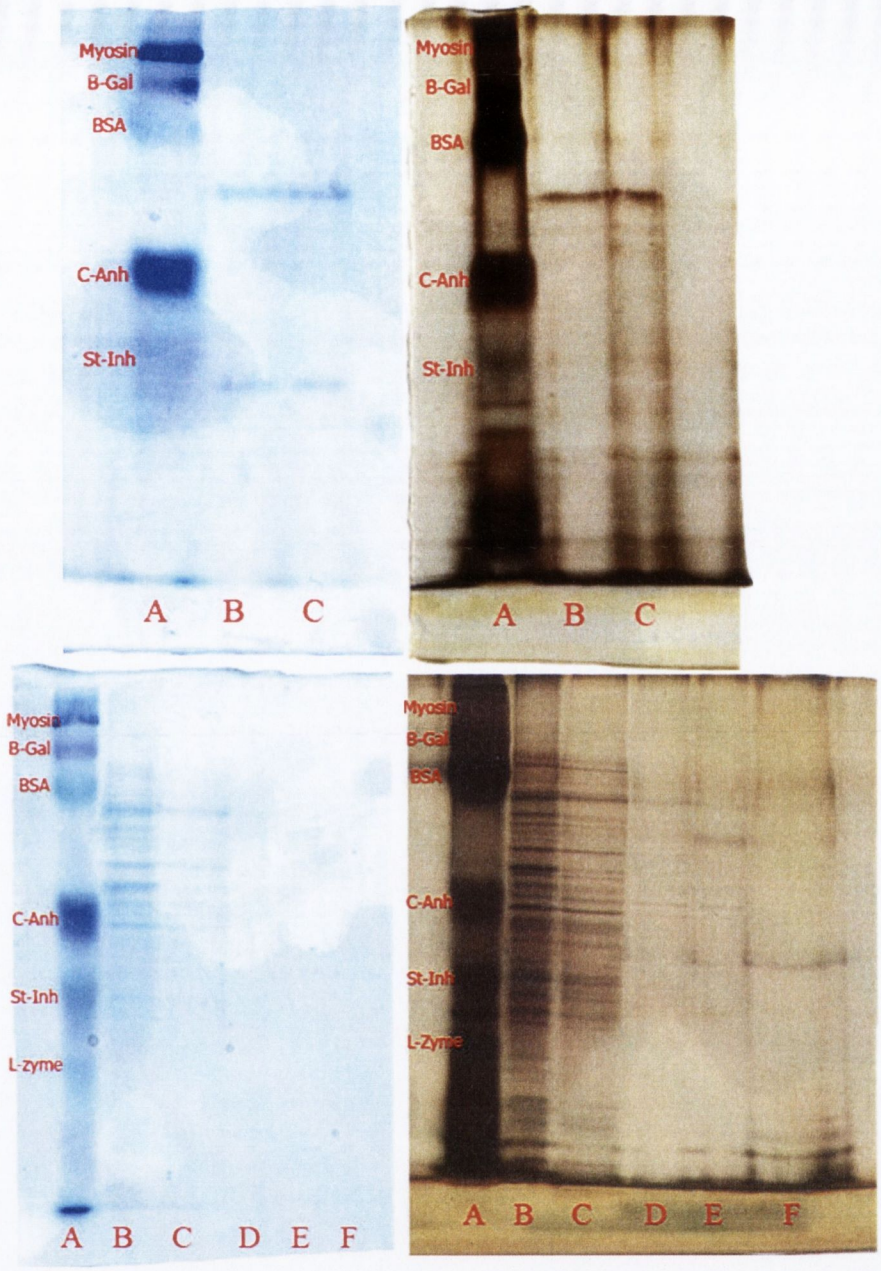


Figure 4-4 Representative SDS -Page gels of preparations using affinity chromatography and immunoprecipitation

Representative gels of proteins purified using immunoprecipitation (top) and affinity chromatography (bottom). Gels are SDS-Page stained using EZblue reagent (left) and silver stain (right). For the immunoprecipitation method the sample was run in duplicate (lanes B and C). For the affinity chromatography method eluate from the initial wash steps were added to wells B-D, the detergent eluate was added to lane E and the antigen eluate added to well F. For all gels, the molecular weight ladder was added to lane A. None of the protein bands in the SDS-page gels showed enrichment for proteins corresponding to HLA molecule fragments.

4.3.4 HLA-DQ2 Binding Assay Control Run

Using the VAVY cell lysate and a labelled indicator peptide a dose response curve was generated for both Ab coated and uncoated wells. Although a dose response relationship was observed, a similar (although slightly higher) response was observed in wells which contained no capture Ab for HLA-DQ2, which would suggest a non-specific binding interaction or a low HLA-DQ2 specific signal.

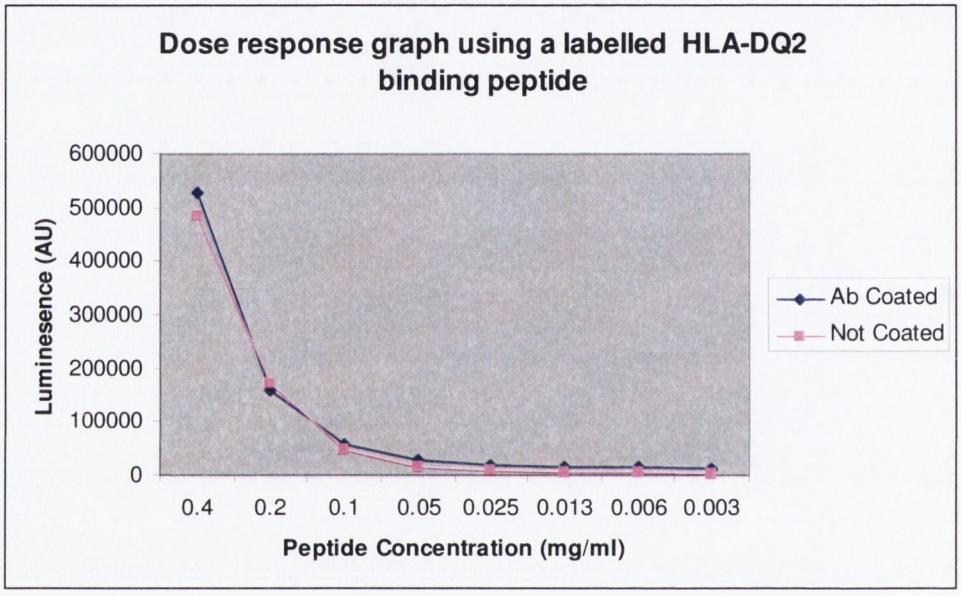


Figure 4-5: Dose response graph using a HLA-DQ2 binding peptide.

Signal generated using the method defined by Vader *et al* using wells that were either coated or uncoated with HLA-DQ capture antibody.

4.3.5 HLA-DQ2 Model Building and Validation

4.3.5.1 Dataset Composition

The DQ2 training dataset comprised 74 unique peptides, the majority of which were derived from substituted analogues of high affinity peptides (Appendix E). The dataset available from the article by Vader *et al* contained binding affinities for 8 sequences known to bind HLA-DQ2, but tested using a different assay system. Of these eight peptides, two were common to the 74 peptides from the AntiJen database. These two peptides were used to generate a regression equation as outlined below (Figure 4-6). This regression equation was then used to transform the remaining data from the article by Vader *et al* to

a comparable level with that from AntiJen. Two datasets were composed from these data; the first contained only sequences from the AntiJen database (n=74), the second contained data from both the AntiJen database and the paper by Vader *et al* (n=80). Inclusion of the data from the paper by Vader *et al* was considered the best approach as this increased the information content in the low affinity region (Figure 4-7).

The validation dataset contained 12 unique sequences with associated IC_{50} values. Each sequence was less than, or equal to 15 residues in length. None of the validation peptides were common to the training data.

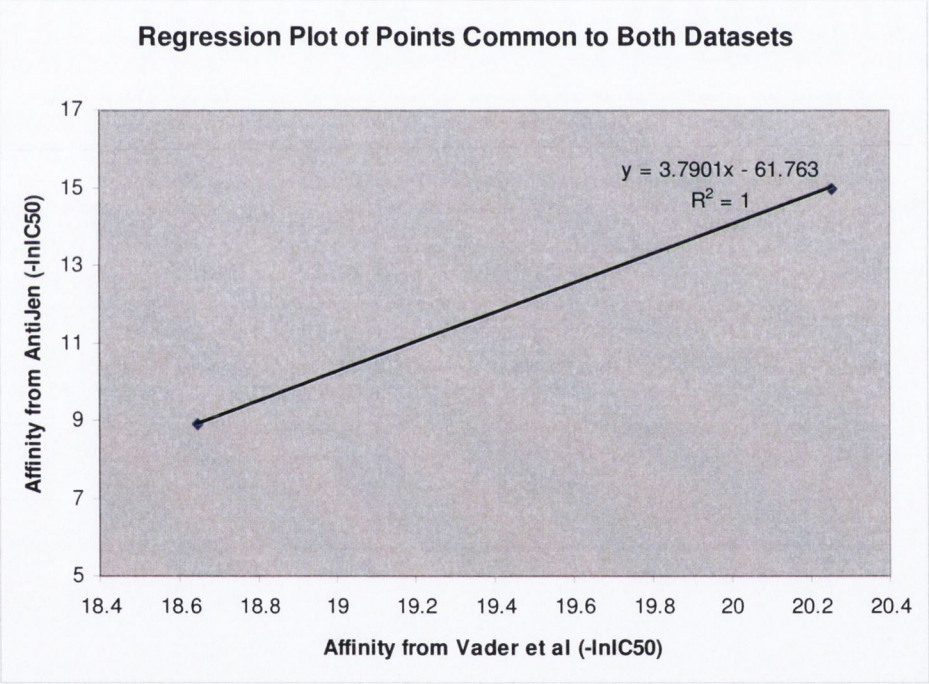


Figure 4-6: Regression graph and equation for data common to the AntiJen dataset and the dataset from Vader et al.

A standard regression graph is used to illustrate the relationship between readouts for the same peptide using two different assay systems. The R^2 value of 1 is not unexpected as there are only two points from which to draw the regression line. The regression equation and R^2 value are visible in the upper right of the chart.

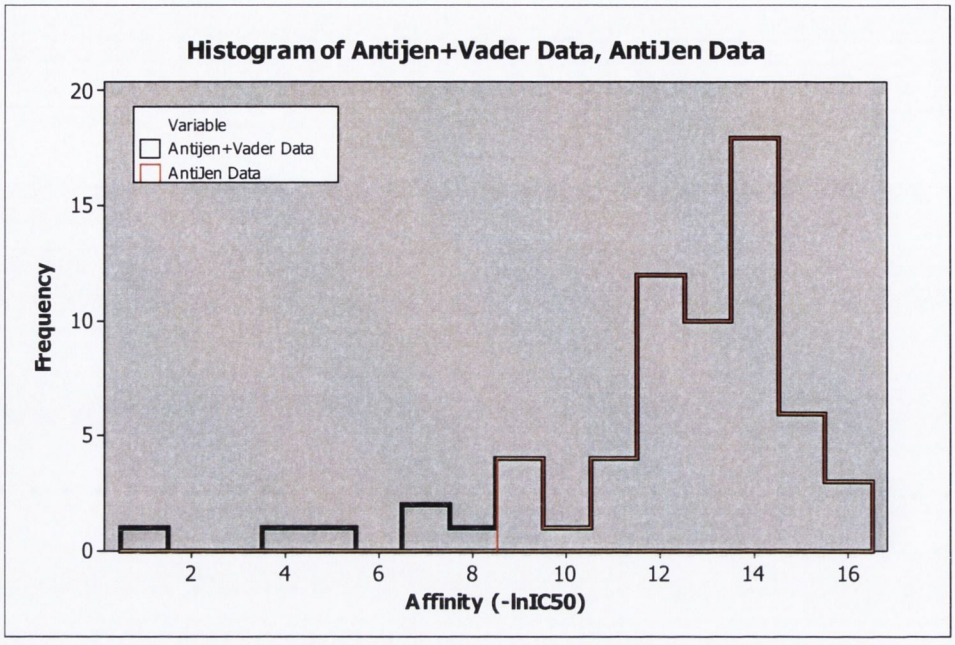


Figure 4-7: Histogram of data affinity composition

Data from both the AntiJen database (AntiJen) and the combined data from the AntiJen database and paper by Vader et al (AntiJen+Vader) are shown. Inclusion of the Vader data increases the information content at lower affinities.

4.3.5.2 Model Generation using the PLS algorithm

Model Q^2 values for both potential training datasets were compared to ensure the inclusion of data from the paper by Vader *et al* did not adversely affect the model building process. The inclusion of this data increased the model Q^2 value; therefore, the model based on this dataset was chosen for use in further studies. Missing data (MD) were updated in the working PSSM; full listings for both versions of the PSSM are presented in the appendices (0).

	Model Q^2
AntiJen	0.310
AntiJen + Vader	0.526

Table 4-2: Model Q^2 values for both variants of the HLA-DQ2 training dataset.

Model Q^2 values for the data from the AntiJen database (AntiJen) and the combined data from the AntiJen database and the article by Vader et al (AntiJen+Vader)

4.3.5.3 Model Validation

In order to validate the HLA-DQ2 binding model, predicted affinities for data in the validation dataset were compared to the *in vitro* measures using the Pearson correlation coefficient. The preselected binding model (Antijen + Vader + MD Correction) gave a correlation coefficient of 0.575. For comparison the predictions for both training sets, with and without MD correction, were also analysed and presented. Results are reasonably comparable with the preselected model except for the Raw AntiJen trained dataset which was relatively low, but, seemed to benefit from the MD correction and use of the AntiJen+Vader dataset.

	Raw PSSM	MD Corrected PSSM
AntiJen	0.440	0.542
AntiJen + Vader	0.559	0.575

Table 4-3: Correlation coefficients for validation data using both training datasets and MD correction.

Correlation coefficient for validation data as predicted using the PSSM trained on the full dataset (AntiJen+Vader) and incorporating missing data corrections (highlighted in green). Additional correlation coefficients are provided for information purposes.

4.3.6 Detection of characterised binders: benchmarking selection criteria

The sensitivity of the prolamin screening resource for a variety of selection criteria was assessed. Initially, however, appropriate affinity cut-off levels were determined by obtaining the mean and standard deviation of the predicted affinities for a subset of known HLA-DQ2 restricted epitopes. For these 18 epitopes, the mean predicted affinity was 11.8 with a standard deviation of 1.7. Using these figures, two cut-off values (10 and 8.4) were determined by subtracting either one or two standard deviations from the mean.

The sensitivity and the proportion of nonameric sub-sequences returned using the chosen selection criteria are outlined in Table 4-4, Table 4-5 and Figure 4-8. From these results one can easily note that as sensitivity increases, the proportion of nonamers to pass the filters also increases. The lack of

specificity, necessitated by the relatively low affinity cut-off value necessary to detect prolamin peptides can be overcome to a large extent using the proline filter and mask. By selecting these masks one can easily reduce the number of nonamers passing the selection criteria, thus, increasing the specificity at the expense of sensitivity.

	Sensitivity				
Affinity Cut-off	Raw	Deamidated	Proline Filter	Proline Mask	Proline count
10	0.89	0.96	0.67	0.52	3
10	0.89	0.96	0.94	0.69	2
8.4	0.96	0.98	0.67	0.58	3
8.4	0.96	0.98	0.95	0.78	2

Table 4-4 Detection sensitivity of characterised T cell stimulatory epitopes

Sensitivity is represented as the proportion of actual T cell stimulatory epitopes identified using the specified selection criteria

	Proportion of nonamers to pass selection criteria				
Affinity Cut-off	Raw	Deamidated	Proline Filter	Proline Mask	Proline Count
10	0.59	0.75	0.18	0.04	3
10	0.59	0.75	0.33	0.09	2
8.4	0.87	0.95	0.25	0.06	3
8.4	0.87	0.95	0.45	0.13	2

Table 4-5 Proportions of nonameric sub-sequences to pass affinity cutoff and filters

The proportion of nonamers to pass the specified selection criteria from all of the prolamin proteins in the test set

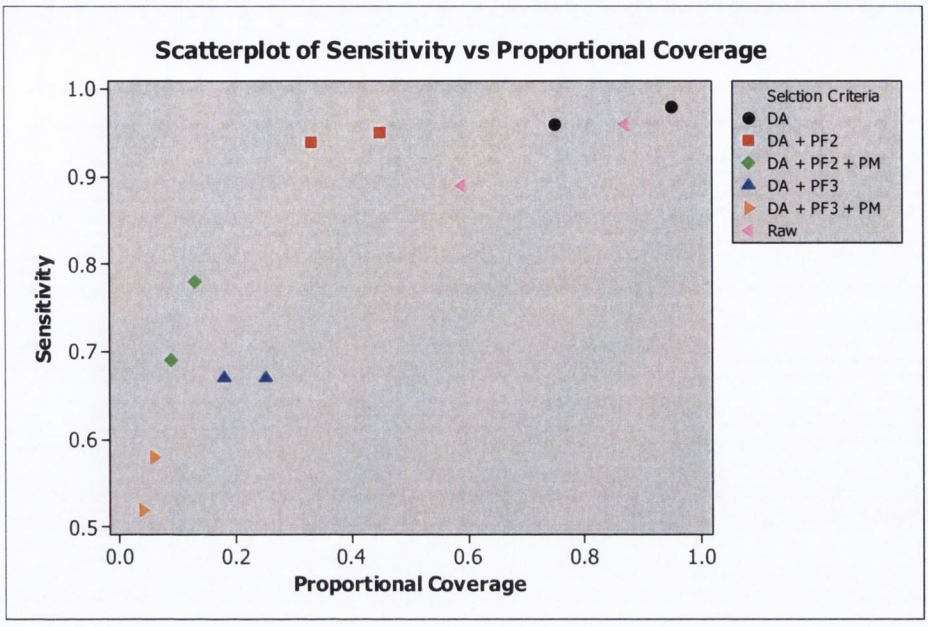


Figure 4-8 Scatterplot of sensitivity vs proportional coverage

The above scatterplot compares the proportion of nonameric sub-sequences to pass the selection criteria for a representative gliadin like sequence and the sensitivity of those criteria for locating characterised epitopes (Table 4-4 and Table 4-5). In all cases the observed lesser sensitivity or proportional coverage for points within subsets is due to the application of a higher affinity cut-off (i.e. 10 vs. 8.4). The filters applied to the sequences were the deamidation algorithm (DA), proline filter at a proline count of two (PF2) or three (PF3) and the proline mask (PM).

4.3.7 Creation of a phylogenetic tree of the gliadin family of proteins

The phylogenetic tree of the gliadin family proteins (Figure 4-9) accurately identified those proteins that came from the same grain families e.g. secalin from rye and avenin from oats. Additionally, proteins such as the Alpha/Beta and Omega gliadins were also localised to separate branches.

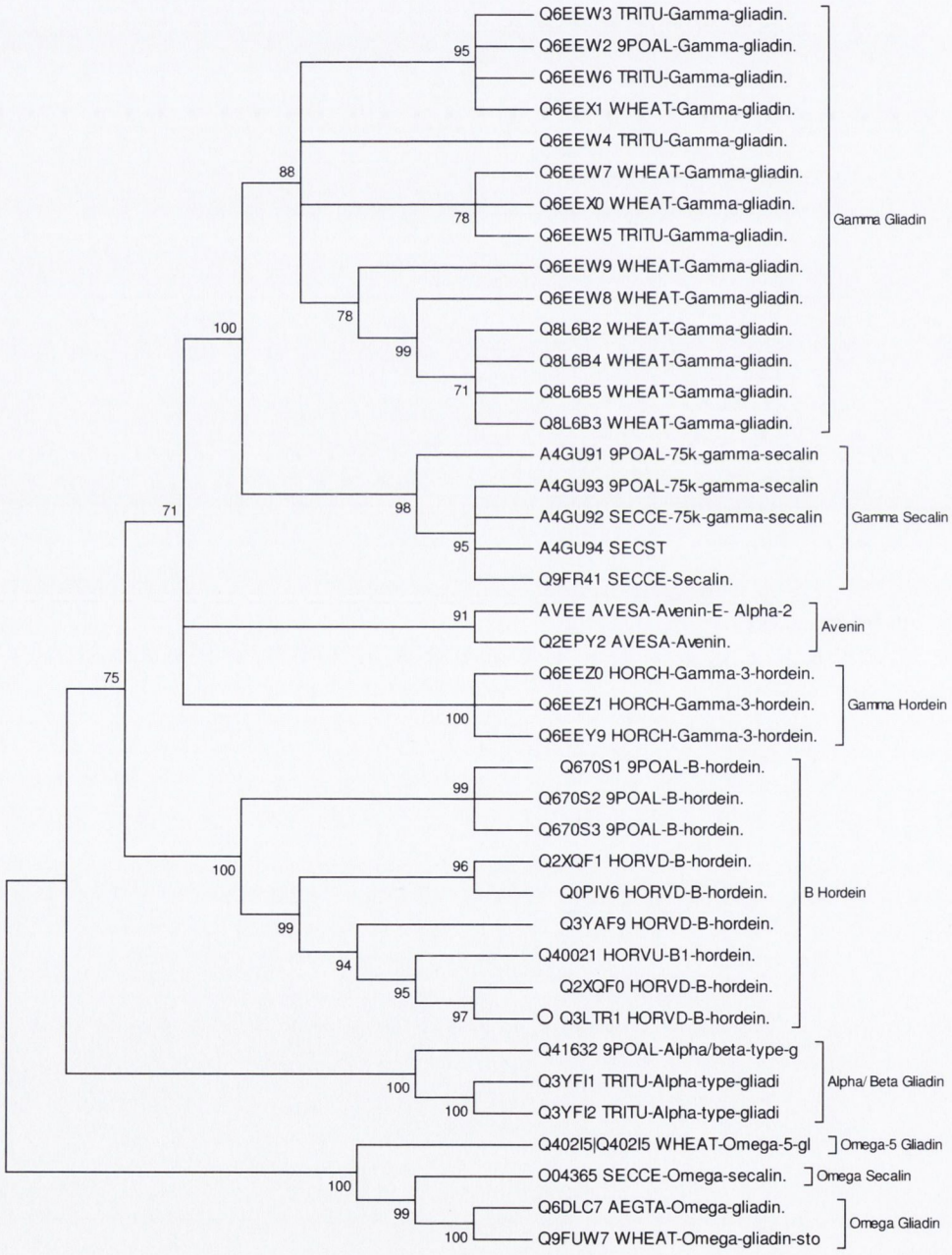


Figure 4-9 Condensed phylogenetic tree of gliadin-like prolamins

4.3.8 Screening the gliadin family of proteins for putative antigenic regions

Screening of the gliadin family proteins provided some interesting insights into the general properties of the proteins in addition to highlighting sequences of potential benefit for development of a less immunogenic wheat. Firstly, when focussing on the gamma gliadins alone one can quite easily see that the number of putative immunogenic sub-sequences varies considerably. For example, 5% of the nonameric sub-sequences in the gliadin protein Q6EEW4 pass our selection criteria, which is roughly twice as great as the 2.6% from the Q6EEW3 sequence (Figure 4-10). This finding is also mirrored by the relatively large amount of characterised T cell epitopes in the Q6EEW4 sequence when compared to the Q6EEW3 sequence (Table 4-6).

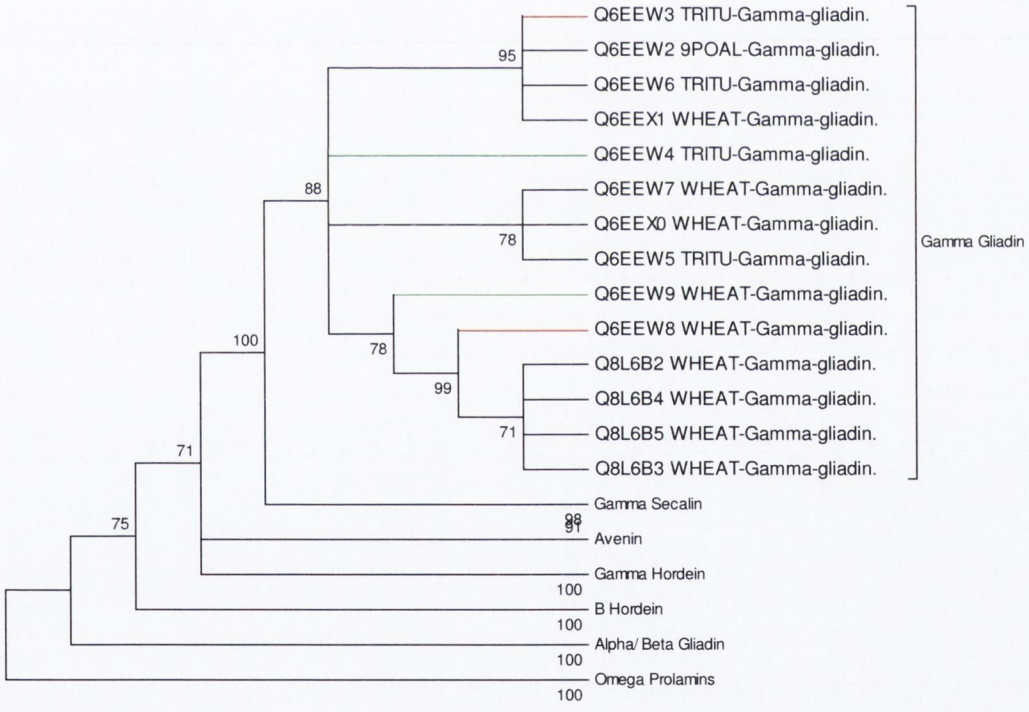


Figure 4-10 Condensed tree showing those gliadin sequences with high and low putative antigenic segments.

Those sequences for which the putative antigenic nonamers are below 3% are shown in red, those with greater than 4% putative antigenic nonamers are highlighted in green.

Protein	Residues Deamidated	Proline Filter and Mask	Characterised Epitopes
>Q6EEW3_TRITU-Gamma-gliadin.	11.98	2.56	0.00
>Q6EEW2_9POAL-Gamma-gliadin.	13.45	3.75	0.00
>Q6EEW6_TRITU-Gamma-gliadin.	12.73	3.75	0.00
>Q6EEX1_WHEAT-Gamma-gliadin.	13.09	3.75	0.00
>Q6EEW4_TRITU-Gamma-gliadin.	13.82	4.12	0.00
>Q6EEW7_WHEAT-Gamma-gliadin.	13.48	3.65	0.00
>Q6EEX0_WHEAT-Gamma-gliadin.	13.14	3.38	0.38
>Q6EEW5_TRITU-Gamma-gliadin.	11.72	3.03	0.00
>Q6EEW9_WHEAT-Gamma-gliadin.	15.76	4.95	1.65
>Q6EEW8_WHEAT-Gamma-gliadin.	12.06	2.81	0.80
>Q8L6B2_WHEAT-Gamma-gliadin.	12.90	3.69	0.74
>Q8L6B4_WHEAT-Gamma-gliadin.	12.90	3.69	0.74
>Q8L6B5_WHEAT-Gamma-gliadin.	12.90	3.69	0.74
>Q8L6B3_WHEAT-Gamma-gliadin.	12.19	3.69	0.74
>A4GU91_9POAL-75k-gamma-secalin.	17.00	7.64	0.45
>A4GU93_9POAL-75k-gamma-secalin.	17.58	8.05	0.22
>A4GU92_SECCE-75k-gamma-secalin.	17.80	7.16	0.22
>A4GU94_SECST-75k-gamma-secalin.	18.26	7.34	0.22
>Q9FR41_SECCE-Secalin.	18.02	7.16	0.22
>Q09114_AVEE_AVESA-Avenin	11.54	0.57	1.15
>Q2EPY2_AVESA-Avenin.	9.43	0.00	0.00
>Q6EEZ0_HORCH-Gamma-3-hordein.	10.60	0.36	0.00
>Q6EEZ1_HORCH-Gamma-3-hordein.	10.60	0.36	0.00
>Q6EEY9_HORCH-Gamma-3-hordein.	10.21	0.36	0.00
>Q670S1_9POAL-B-hordein.	11.04	3.09	0.00
>Q670S2_9POAL-B-hordein.	10.13	3.02	0.00
>Q670S3_9POAL-B-hordein.	11.11	3.36	0.00
>Q2XQF1_HORVD-B-hordein.	10.57	2.72	0.00
>Q0PIV6_HORVD-B-hordein.	12.76	3.90	0.00
>Q3YAF9_HORVD-B-hordein.	13.33	3.77	0.00
>Q40021_HORVU-B1-hordein.	10.70	3.42	0.00
>Q2XQF0_HORVD-B-hordein.	12.46	3.81	0.00
>Q3LTR1_HORVD-B-hordein.	13.10	3.55	0.00
>Q41632_9POAL-Alpha/beta-type-gliadin.	12.16	1.74	0.00
>Q3YFI1_TRITU-Alpha-type-gliadin.	13.64	1.44	0.36
>Q3YFI2_TRITU-Alpha-type-gliadin.	13.29	1.80	0.36
>Q402I5|Q402I5_WHEAT-Omega-5-gliadin.	26.88	0.93	0.00
>O04365_SECCE-Omega-secalin.	23.25	8.02	0.29
>Q6DLC7_AEGTA-Omega-gliadin.	23.98	9.11	1.04
>Q9FUW7_WHEAT-Omega-gliadin	23.21	9.19	1.10

Table 4-6 Distribution of predictions across the prolamin proteins

All results are percentages. For predicted deamidations the percentage refers to the predicted percentage of residues to be deamidated by tissue transglutaminase. Those predictions to pass the affinity cutoff in addition to the proline filter and proline mask after predicted deamidation (Proline Mask and Filter) and the percentage characterised epitopes refer to the percentage of nonameric sub-sequences to return a positive hit.

Secondly, when one examines the family of gliadin-like proteins some remarkable differences exist. For example, the percentage of residues predicted to be deamidated using the deamidation algorithm was considerably higher for the omega type gliadins (Figure 4-11, Table 4-6). Additionally, the percentage of nonamers to pass the affinity and proline filter and mask varied across the different groups of sequences with the greatest percentage being present in omega gliadin, omega secalin and gamma secalin with strong under-representations present in Avenin, Hordein and Omega-5 gliadin.

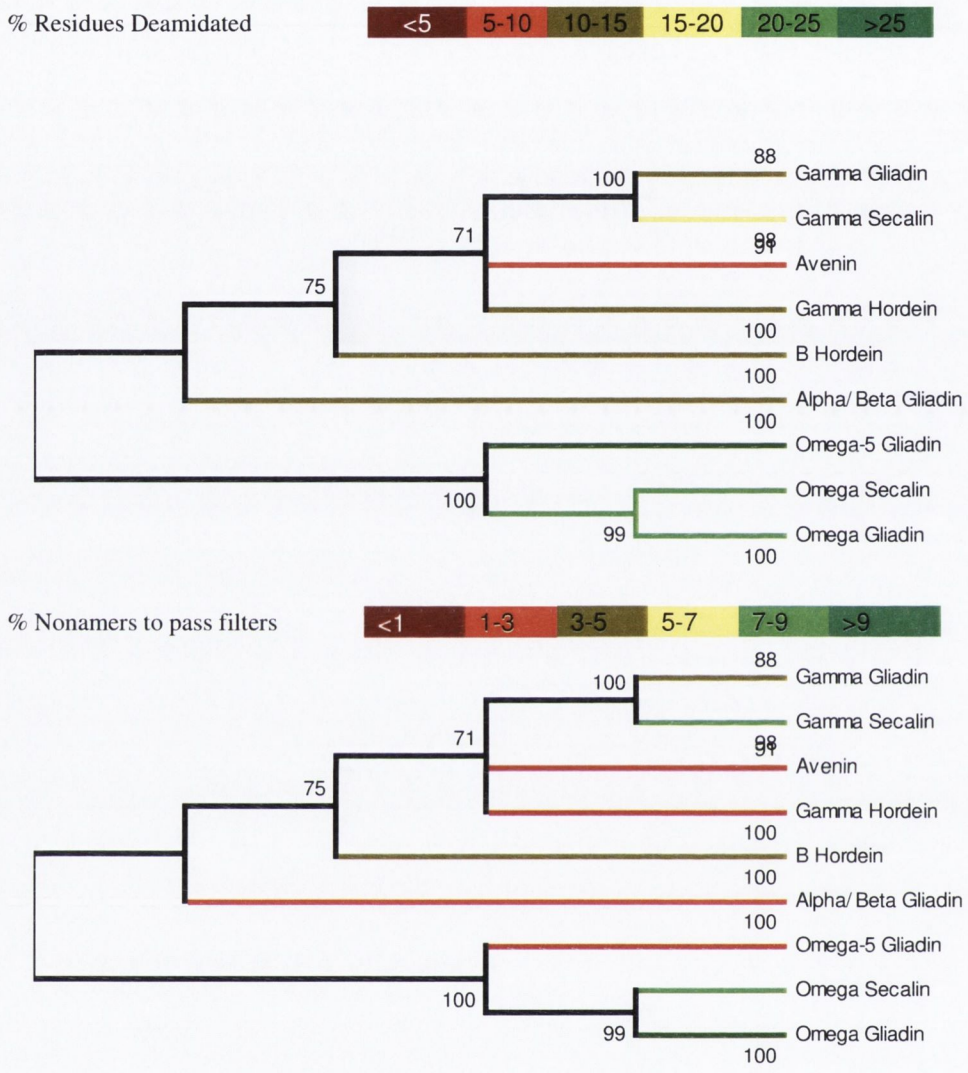


Figure 4-11: Breakdown of predicted deamidated residues and putative antigenic regions by inferred phylogeny.

Grouping of predicted characteristics by protein sequence families. Colour keys are indicated above each of the two phylogenetic trees.

4.4 Discussion

For this study it was necessary to create a queryable resource for *in silico* screening of prolamin sequences. The requirements for the resource were simple; it had to be lightweight in terms of computational burden, expandable to accommodate additional HLA molecules if necessary, easy to update and relevant to coeliac disease. For this purpose we chose to implement an html based resource to be used on the intranet which could potentially be made available over the World Wide Web. Design and implementation of the project were implemented using cgi-scripting which allows the results page to reflect the input sequence i.e. each results page should be sequence specific. Additionally, it was also decided to implement filters based on studies in the field of coeliac disease which were shown to increase the selectivity of the method. The filters and pre-processing steps included a deamidation filter based on the algorithm defined by Vader *et al* (Vader et al., 2002a) which demonstrated a strong association between propensity for tTG mediated deamidation and T cell reactivity, a proline filter to select proline rich epitopes based on the association between proline rich regions and antigenicity (Shan et al., 2002, Arentz-Hansen et al., 2002, Shan et al., 2005), and lastly a proline mask to replicate the findings that proline residues are restricted to certain locations within the canonical binding register (Qiao et al., 2005).

Having designed and coded our prolamin screening resource, the next step in our study was the construction of a robust model of peptide-HLA-DQ2 binding. Optimisation of the binding prediction algorithms suggested the optimal approach to be a combination of the QSAR algorithm and single-substitution based training data. To supplement the publicly available data it was initially intended to generate supplementary data in-house. Using our modelling data and information about the observed effect of peptide length of peptide affinity (Section 2.2.3 and section 3.3.2), we elected to generate a set of binding data using single substitutions of a known high affinity nonameric peptide. However, using a combination of approaches we were unsuccessful in setting up a competitive assay for HLA-DQ2 binding. The drawback of all of the implemented approaches was the lack of a positive control to optimise the

assays. Thus, it cannot be stated with certainty that sufficient HLA-DQ2 failed to populate the cell lysate. Nevertheless, the variety of approaches attempted strongly suggests the absence of our test protein from all lysate preparations. This may have been due to any one of a number of factors, from cell stress down-regulating HLA-DQ2 expression to an ineffective level of protease inhibitors. Additionally it has been observed that HLA-DQ2 molecules can be difficult to purify for use in competitive binding assays (John Sidney, personal communication). This step of our analysis was omitted for time concerns and the model was constructed with publicly available data instead.

Construction of the HLA-DQ2 binding model using publicly available data involved a greater level of supervision than is traditional for epitope prediction algorithms. This manual curation of the binding model occurred at two stages; initially, when selecting training data and subsequently, when compensating for missing data. Curation of the training dataset involved selecting those data that were thought to provide the best model building information, based on our findings in chapters 2 and 3. To this end, we enriched our dataset for short peptides, many of which were formed from substitution studies. Additionally, data from an article by Vader *et al* (Vader et al., 2003b) were incorporated into the dataset. Although these binding affinities were determined under different experimental conditions, transformation of the peptide affinities was performed by linear regression. This assumed linear relationship between peptide affinities was deemed to be a reasonable assumption and was validated in part by the increased model Q^2 value generated with the AntiJen+Vader dataset (Table 4-2). In combination with the manual curation of the training dataset, a correction for missing data was implemented based on the amino acid grouping used by Bui *et al* (Bui et al., 2005). This correction was deemed necessary in order to compensate for the lack of binding data for a number of amino acids at positions in the binding groove. The validation correlation coefficient for the final model is comparable with that found by Doytchinova *et al* when using an iterative self consistent algorithm (Doytchinova and Flower, 2003). It is quite possible that a superior correlation value may have been obtained for a validation set of nonameric or decameric sequences, which would not incur influences from sequence length. However, given the data availability, such a

dataset would be unlikely to contain sufficient data to return a reliable measure of correlation.

Having generated and validated a robust model of peptide-HLA-DQ2 interaction, a database of epitopes known to stimulate T cells from coeliac disease patients was generated. In total 101 characterised epitopes were found by searching the literature. The screening resource allows one perform one of two separate queries; either a search for characterised epitopes of which there are 101, or, prediction of novel immunogenic regions using the HLA-DQ2 binding model and selection filters.

As *in silico* screening for characterised epitopes requires only minimal evaluation (i.e. a sub-sequence either is or is not present in a protein), therefore validation was only required for predicted immunogenic sequences. The first validation step involved determining the sensitivity of the method, for this purpose two affinity cut-offs were determined using a subset of known epitopes. These two affinity cut-offs were combined with a variety of filtering options to determine the proportion of true positives, based on the selection criteria. The sensitivity when all filters were applied was between 0.52 and 0.78. This sensitivity could climb as high as 0.98 without applying filters but the number of background sequences to pass the filter would also increase (Table 4-5). This second evaluation of the method cannot be strictly thought of as a measure of sensitivity as it cannot account for the occurrence of false positives. However, it does show that a remarkable number of sequences pass the selection criteria if using affinity alone. This is likely due to the fact that gliadin peptides do not exhibit an extraordinarily high affinity for HLA-DQ2 when compared with some well characterised binders. The relatively low affinity of gliadin peptides is evident in Figure 4-7 which shows that the gliadin peptides enrich the low affinity data in the training set. However by implementing all filters and the deamidation algorithm it was possible to reduce the proportion of sub-sequence passing the selection criteria to a proportion between 0.04 and 0.13. Plotting of the sensitivity versus the 'specificity' for the range of parameters allowed easy visualisation of the trade-off involved in increasing the method sensitivity (Figure 4-8). However, the

lack of a good negative control sequence meant that a ROC curve could not be generated, as the number of false positives could not be determined.

The next step in the analysis involved an assessment of the predictions from the screening resource. A subset of the returned values was presented in tabular form (Table 4-6) and represented diagrammatically using condensed variants of a purpose generated phylogenetic tree. The first such tree (Figure 4-10) contained a representation of the percentage of predicted immunogenic regions within the gamma gliadin proteins. Interestingly, it became evident that some of the sequences contained up to twice as many predicted immunogenic regions as others. Reassuringly, for the putatively highly immunogenic (Q6EEW4) sequence, a suitably high level of characterised epitopes was also found when compared with the remaining sequences (Table 4-6).

To follow on from this analysis the distribution of putative immunogenic sequences and putative deamidation sites was analysed across the family of gliadin-like proteins (Figure 4-11). Surprisingly, the well characterised immunogenic gamma gliadins contained only average amounts of deamidation sites and putative antigenic regions when compared to the other sequences. Omega-5 gliadin contained the largest percentage of putative deamidation sites yet contained the lowest percentage of nonamers to pass the selection criteria. This may partly explain the finding that omega-5 gliadin is found to be involved in the less common type-I hypersensitivity to ingestion of gliadin (Palosuo et al., 2001). It has also been shown that the transglutaminase mediated crosslinking of omega-5 gliadin can increase IgE reactivity (Palosuo et al., 2003). Omega secalin, omega gliadin and gamma secalin all showed an over-representation of putative antigenic regions when compared with other family members. Interestingly, sequences like the gamma, alpha and omega gliadins, show a putative level of reactivity that seems to be species specific i.e. all of the sequences show a level of putative reactivity that is greater than that of avenin, for example. This may partially explain the localisation of prolamin immunogenicity to wheat, barley and rye. Again, the much debated immunogenicity of avenin may benefit from evaluation using this resource. Of the two avenin sequences tested, one showed no putatively antigenic regions,

while the other sequence contained a minimal amount. Interestingly, the mildly antigenic avenin protein, was suggested to be related to coeliac disease pathogenesis in previous studies (Rocher et al., 1992), although this antigenicity was not confirmed outside of immunoblotting. Findings such as this create the possibility that certain strains of oats may harbour immunogenic regions whereas others may not. Thus, the initial development of an irrefutably coeliac-safe strain of oats may generate an interesting proof-of-concept for reduction of immunogenicity in coeliac disease associated grains. These findings demonstrate that the implemented screening resource can render predictions that may be of use in the pre-selection of minimally immunogenic prolamins for the development of coeliac-friendly grains.

4.5 Summary

In this chapter a computational resource for screening prolamins for potential HLA-DQ2 dependent immunogenicity was created. The resource merged information available from a HLA-DQ2 binding motif with current knowledge of tTG mediated deamidation and proline dependence. Validation of the predictive abilities of the resource showed it to be quite capable of identifying characterised epitopes. The resource was then used to screen a dataset of gliadin-like prolamins. The results of the identified confirmed the antigenicity of characterised immunogenic proteins in addition to highlighting the potential immunogenicity of lesser studied prolamins such as the omega gliadins.

5 General Discussion

Through the course of this thesis it has been demonstrated, that using computational methods it is possible to obtain reasonable estimates of the affinity one might observe if one were to conduct *in vitro* binding assays. It was subsequently demonstrated that the length of a peptide can be a determining factor in peptide-MHC class II affinity, as peptide elongations tend to increase MHC class II affinity. However, a mechanism underlying this effect could not be conclusively determined. As these methods were part of a larger study, the results from the first two experimental chapters formed the basis for the last body of work which involved the creation of and evaluation of an *in silico* prolamin screening resource.

5.1 Review of Results

In Chapter 2, the focus of our work was the implementation, optimisation and evaluation of algorithms for the quantitative prediction of peptide binding to both MHC class I and MHC class II molecules. The three algorithms are established methods in the field of immunoinformatics yet have quite different origins. The Gribskov's profile analysis (GPA) algorithm originated from protein domain identification; hidden Markov models are similarly used in protein domain identification and homology searching and generate more complex statistical models than profiles; the QSAR based method originated in the field of computational chemistry. Of the three methods, only the QSAR algorithm accounted for the observed affinity of a peptide when constructing the binding model. It was surprising then, that the two methods which did not account for observed affinity, generated predictions which correlated with those observed *in vitro* equally as well as those of the QSAR-algorithm in some instances. Additionally, the methods were re-optimised to suit the task of epitope prediction. Principally, the GPA algorithm was evaluated using a series of amino acid substitution matrices, each of which accounted for different common properties between amino acids. Using an in-house program, LOOCV was implemented in order to choose the most optimal substitution matrix. The best matrix for the task of epitope prediction was unsurprisingly a chemical similarity matrix. The QSAR algorithm was also optimised for our epitope prediction studies. Using data modelling it was possible to determine the

optimal number of latent variables and the optimal dataset composition for the QSAR algorithm. Additionally, the effect of cleaning the dataset to remove over-represented sequences was also assessed. The results of these analyses showed a demonstrable benefit attributable to method optimisations. This effect was especially pronounced for models built using the GPA algorithm with small numbers of peptides. The effect of dataset cleaning was shown to be detrimental to the model building process as larger training datasets consistently delivered better binding models compared with their cleaned counterparts.

Having reoptimised and reassessed the algorithms using the MHC class I datasets the resultant optimised methods were assessed for their ability to predict peptide binding to MHC class II. For this purpose, it was necessary to align the core regions of the known MHC class II binding peptides of the training set. Two approaches were taken; alignment using a standard protein alignment program, Clustalw, and alignment using a known HLA-DR4 binding motif. Alignment using the binding motif proved much more successful than the Clustalw alignment and generated a greater information content per position in the binding groove. This set of aligned nonameric core peptides was cleaned and the cleaned and raw datasets used to build binding models for HLA-DR4. As with the MHC class I epitope prediction algorithms, the dataset cleaning actually reduced the predictivity of the methods. Examination of the sequence affinity histogram revealed that, prior to dataset cleaning, the dataset contained many high affinity (albeit heavily overrepresented) sequences. Hence, the presence of a cohort of high affinity peptides was necessary for proper model building using the HMM and GPA algorithms. Interestingly, the QSAR based method managed to generate a relatively robust model using the cleaned dataset of only medium affinity peptides, as its design takes peptide affinity into consideration. A modified version of the iterative self consistent algorithm originally described by Doytchinova *et al* was also tested but found to be ineffective, generating the lowest recorded correlations, most likely as a result of the modified parameters. As a result of our method evaluations, we were able to create a binding model using the most appropriate algorithm for the dataset size and composition for our prolamin screening resource.

Following our implementation and evaluation of the affinity prediction algorithms, the impact of sequence length on reported affinity was assessed. This was done to further refine our understanding of peptide-MHC class II affinity measurements for affinity prediction and to better understand the impact of peptide length in selection of peptides for MHC class II binding. From this study we were able to characterise a positive relationship between peptide length and MHC class II binding affinity. Additionally, it was assessed whether this relationship was the result of a multivalency due to the creation of multiple overlapping nonamers each capable of binding MHC class II. Our results failed to show this to be the case although it would be necessary to perform separate structured *in vitro* tests using register limited analogues to definitively rule out such a mechanism.

Having developed and optimised the methods it was possible to create a HLA-DQ2 binding model. Based on our results in the previous chapters it was decided to construct the model using publicly available data and supplement this with data generated in-house. The in-house data would be generated using single substituted analogues of a known high affinity binder, and interpreted using the QSAR algorithm as our data modelling suggested this to be an optimal approach (Chapter 2). The generated data would also be only nine residues in length to remove any bias as a result of peptide length. Unfortunately, in the time allotted it was impossible to optimise the binding assay to run in-house. The fault of the assay to operate correctly was due to a very low signal to noise ratio which may have been due to any one of a number of factors. The most likely explanation is that a poor level of HLA-DQ2 expression or a suboptimal solution at any stage of the assay optimisation, may have resulted in a low concentration of HLA-DQ2 in the cell lysate rendering it unsuitable for use in a binding assay. In spite of this however, it was still possible to generate a binding model which generated correlation values comparable with those generated by Doytchinova *et al* using their ISC algorithm (Doytchinova and Flower, 2003). This binding model was then used to form the basis for prediction of immunogenic regions from prolamin proteins as implemented in our prolamin screening resource. In addition to

affinity prediction, the prolamin screening resource also contained a number of filters to increase the specificity of the predictions. For example, a proline mask was implemented to eliminate those nonamers above the affinity cutoff with proline residues at p2, p4, p7 or p9. Additionally, it was also possible to filter only those residues above cutoff which contained a pre-ordained number of proline residues. Also, the resource contains a database of known T cell stimulatory sequences from gliadin which can allow one to screen a prolamin sequence for characterised epitopes. Using all of the currently available gliadin-like prolamin sequences from swissprot, the screening resource was tested to ascertain its usability. The results of this analysis provided some interesting insights into prolamin protein toxicity and the relative abundance of glutamine residues suitable for tTG mediated deamidation, as was discussed previously (Section 4.4).

5.2 Implications for peptide-MHC affinity prediction

To the best of our knowledge this is the first study to compare the efficacies of the QSAR, HMM and profile based methods for quantitative peptide-MHC class II binding affinity prediction. It is also one of the first studies to compare quantitative methods of peptide-MHC class II affinity prediction. Reassuringly, a number of our findings agreed with the literature; for example, profile based methods seemed to show the better predictivity with smaller training datasets whereas the more complex methods (HMMs and QSAR) seemed to benefit from the extra data. This would agree with the assertions of Brusic, who proposed the selective use of predictive algorithms based on dataset size (Brusic et al., 2004). The optimisation of the GPA algorithm by alteration of the amino acid substitution matrix particularly boosted its performance. This finding may be of particular importance for those generating binding models using a limited number of peptide sequences, as are used in the RANKPEP resource (Reche et al., 2002). Furthermore, the sliding window algorithm used in Chapter 3 unexpectedly provided a potential increase in the current efficiency of *in vitro* epitope screening based on predictive algorithms. For example, at 45% residue coverage the sliding window algorithm correctly

identified 90% of the characterised binders whereas the other approaches only identified 60%.

A set of software was also written to perform all of the analyses which can be re-used and made available to researchers studying epitope prediction. This also makes the additive method freely available for the first time as prior to now it had only been implemented using commercially available software (Doytchinova et al., 2002). This lack of availability on an open-source platform likely limited the use of the additive method in comparative studies in the past. Within the confines of this thesis the enhanced clarity and familiarity with epitope prediction algorithms was of critical benefit for chapters 3 and 4. As the computer programs were readily available and optimised, usage of the correct algorithm for the dataset composition was greatly simplified and more informed choices could be made.

5.3 Implications for the study of immunology

In addition to being of benefit to those engaged in epitope prediction, the results from Chapter 3 may be of great importance for the study of immunology. Sercarz posited that longer peptides have a greater binding affinity for MHC class II (Sercarz and Maverakis, 2003); this assertion was based on observations in a number of articles such as that by Srinivasan *et al* (Srinivasan et al., 1993). However, to the best of our knowledge this is the first study to directly test this hypothesis across a range of proteins and MHC class II alleles. The results of our study were compelling and across 1279 identified peptide elongation events for peptides binding to 19 distinctive MHC class II alleles, the affinity predominately increased. We were also able to demonstrate that the length of the affinity increase was also a deciding factor in the observed increase in length, with larger length increases leading to larger affinity increases. It was also possible to identify an approximate 'ceiling length' after which the addition of further residues will have a negative or null effect.

From these results, it can be posited that the effect of increased peptide affinity as a result of increased length is in fact a general one and not peculiar to a restricted subset of peptides. The implications of such an effect are many; to

begin, the field of subunit vaccine design may gain increased vaccine efficacy by increasing peptide length or eliminating potential proteolytic cleavage sites. In fact, the reduction of proteolytic enzyme cleavage sites would also be likely to increase the molar concentration of peptide available for binding, thus increasing both affinity and molar concentration. An additional *in vivo* role of this increased affinity may be to increase reactivity to recently degraded peptides. Presumable those peptides most recently degraded may be less completely digested and - as a result of increased length - have an increased affinity for MHC class II relative to shorter peptides. Such an effect would increase the likelihood of T cells recognising more recently endocytosed peptides and form a type of 'triage' for T cell reactivity.

5.4 Implications for coeliac disease and gluten sensitivity

The implemented prolamin screening resource was successfully able to return good levels of prediction of characterised binders without excessively high background when using all of the implemented filters. Additionally, the database of characterised epitopes also allows for screening for sub-sequences previously determined to be immunogenic. All in all, this renders the prolamin screening resource as a useful tool for screening prolamin proteins in order to identify those with the lowest potential immunogenicity. As illustrated in the flowchart below (Figure 5-1), one could use the resource to initially identify and exclude those sequences which lack characterised immunogenic sequences. The remaining sequences could then be screened to identify those with the lowest number of putative antigenic sub-sequences for further testing using *in vitro* assays.

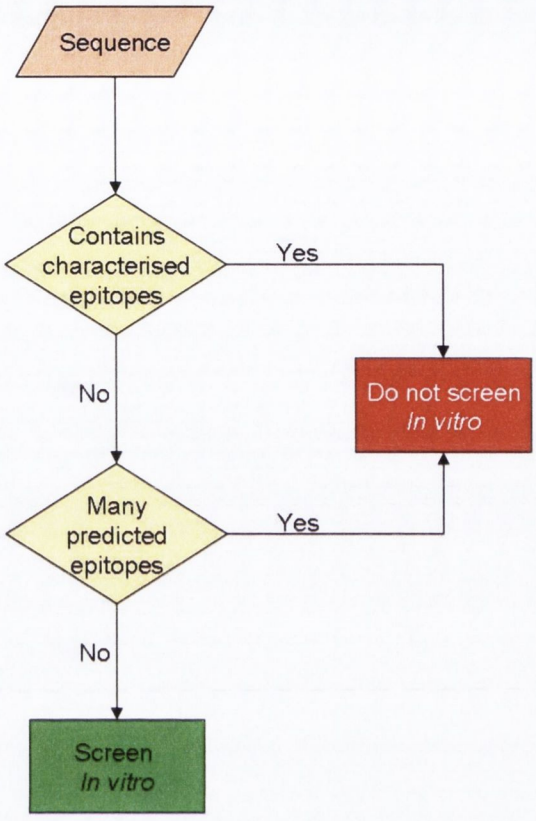


Figure 5-1 Decision tree for screening prolamin sequences using *in vitro* assays

Initial investigations using the prolamin screening resource allowed the identification of putative, moderately immunogenic sub-sequences among the gamma gliadins. Additionally, the omega gliadin and omega secalin proteins, in addition to the gamma secalins, were also predicted to harbour greater immunogenicity than the gamma gliadin gliadins. The fact that research has focussed on gamma fractions does not eliminate the possibility that omega gliadins are potentially as immunogenic if not more so. In fact, previous studies have shown that the omega gliadin fraction can trigger the characteristic inflammatory process *in vivo* (Ciclitira et al., 1984). More recent studies have also highlighted the greater diagnostic specificity of the omega fraction when used for serological screening (Chirdo et al., 2000). Interestingly, omega-5 gliadin which seems to have diverged significantly from the other omega proteins analysed in the study, was predicted to be the least immunogenic, and yet was also predicted to contain the greatest proportion of potential deamidation sites. These factors may add further to the discovery of omega-5 gliadin as the allergen in wheat-dependent, exercise induced anaphylaxis (WDEIA) and the increased IgE reactivity observed after tTG mediated cross linking (Palosuo et al., 2001, Palosuo et al., 2003). Moreover, any inability for the omega-5 gliadin fractions to elicit a strong cellular immune response may result in the more Th2-like humoral response observed by Palosuo *et al.* Furthermore, the link between tTG mediated cross-linking and IgE reactivity may indicate that proposed tTG inhibitors may be of use in the treatment of WDEIA.

It must be pointed out however that the applicability of the prolamin screening resource is peculiar to HLA-DQ2 and omits resources to characterise immunogenicity in the context of HLA-DQ8. The reasons for this are two-fold: firstly, the majority of coeliac patients express HLA-DQ2, which means a larger population with which to test and apply reduced immunogenicity wheat; secondly, the larger body of data available for HLA-DQ2 binding affinity measurements and characterised epitopes allow construction of more accurate binding models and better validation. However, the modular nature of the prolamin screening resource allows the inclusion of extra alleles as needs dictate and the quantity of available data increases.

The results of our investigations in Chapter 3 also hold implications for coeliac disease as the best characterised binding sequence is 33 amino acids in length, and a number of highly similar peptides exist throughout the prolamins (Shan et al., 2002). The 33mer peptide described by Shan *et al* also contains multiple overlapping epitopes. Our findings that longer peptides tend to exhibit a higher MHC class II affinity may partially explain the immunodominance of this peptide. The observed effect of a markedly increased length-dependent MHC class II affinity may apply to this excessively long peptide and explain its immunodominance. It would stand to reason that the identification of potential enzyme cleavage sites in an analogue of this peptide may indicate a natural mechanism for peptide shortening, and thus, reduction of immunogenicity.

References

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ARENTZ-HANSEN, E. H., MCADAM, S. N., MOLBERG, O., KRISTIANSEN, C. & SOLLID, L. M. (2000a) Production of a panel of recombinant gliadins for the characterisation of T cell reactivity in coeliac disease. *Gut,* 46**,** 46-51.

ARENTZ-HANSEN, H., FLECKENSTEIN, B., MOLBERG, O., SCOTT, H., KONING, F., JUNG, G., ROEPSTORFF, P., LUNDIN, K. E. & SOLLID, L. M. (2004) The molecular basis for oat intolerance in patients with celiac disease. *PLoS Med,* 1**,** e1.

ARENTZ-HANSEN, H., KORNER, R., MOLBERG, O., QUARSTEN, H., VADER, W., KOOY, Y. M., LUNDIN, K. E., KONING, F., ROEPSTORFF, P., SOLLID, L. M. & MCADAM, S. N. (2000b) The intestinal T cell response to alpha-gliadin in adult celiac disease is focused on a single deamidated glutamine targeted by tissue transglutaminase. *J Exp Med,* 191**,** 603-12.

ARENTZ-HANSEN, H., MCADAM, S. N., MOLBERG, O., FLECKENSTEIN, B., LUNDIN, K. E., JORGENSEN, T. J., JUNG, G., ROEPSTORFF, P. & SOLLID, L. M. (2002) Celiac lesion T cells recognize epitopes that cluster in regions of gliadins rich in proline residues. *Gastroenterology,* 123**,** 803-9.

ARNOLD, P. Y., LA GRUTA, N. L., MILLER, T., VIGNALI, K. M., ADAMS, P. S., WOODLAND, D. L. & VIGNALI, D. A. (2002) The majority of immunogenic epitopes generate CD4+ T cells that are dependent on MHC class II-bound peptide-flanking residues. *J Immunol,* 169**,** 739-49.

BANKOVICH, A. J., GIRVIN, A. T., MOESTA, A. K. & GARCIA, K. C. (2004) Peptide register shifting within the MHC groove: theory becomes reality. *Mol Immunol,* 40**,** 1033-9.

BASTA, S. & ALATERY, A. (2007) The cross-priming pathway: a portrait of an intricate immune system. *Scand J Immunol,* 65**,** 311-9.

BHASIN, M. & RAGHAVA, G. P. (2004) Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine,* 22**,** 3195-204.

BLYTHE, M. J. & FLOWER, D. R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci,* 14**,** 246-8.

BOOTS, A. M., VERHEIJDEN, G. F., SCHONINGH, R., VAN STAVEREN, C. J., BOS, E., ELEWAUT, D., DE KEYSER, F., VEYS, E., JOOSTEN, I. & RIJNDERS, A. W. (1997) Selection of self-reactive peptides within human aggrecan by use of a HLA-DRB1*0401 peptide binding motif. *J Autoimmun,* 10**,** 569-78.

BRUSIC, V. (2003) From immunoinformatics to immunomics. *J Bioinform Comput Biol,* 1**,** 179-81.

BRUSIC, V., BAJIC, V. B. & PETROVSKY, N. (2004) Computational methods for prediction of T-cell epitopes--a framework for modelling, testing, and applications. *Methods,* 34**,** 436-43.

BRUSIC, V., PETROVSKY, N., ZHANG, G. & BAJIC, V. B. (2002) Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol Cell Biol,* 80, 280-5.

BUI, H. H., SIDNEY, J., PETERS, B., SATHIAMURTHY, M., SINICHI, A., PURTON, K. A., MOTHE, B. R., CHISARI, F. V., WATKINS, D. I. & SETTE, A. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics,* 57, 304-14.

BUUS, S., LAUEMOLLER, S. L., WORNING, P., KESMIR, C., FRIMURER, T., CORBET, S., FOMSGAARD, A., HILDEN, J., HOLM, A. & BRUNAK, S. (2003) Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens,* 62, 378-84.

BYRNE, G., RYAN, F., JACKSON, J., FEIGHERY, C. & KELLY, J. (2007) Mutagenesis of the catalytic triad of tissue transglutaminase abrogates coeliac disease serum IgA autoantibody binding. *Gut,* 56, 336-41.

CHAKRABORTY, A. K., DUSTIN, M. L. & SHAW, A. S. (2003) In silico models for cellular and molecular immunology: successes, promises and challenges. *Nat Immunol,* 4, 933-6.

CHANG, K. Y., SURI, A. & UNANUE, E. R. (2007) Predicting peptides bound to I-Ag7 class II histocompatibility molecules using a novel expectation-maximization alignment algorithm. *Proteomics,* 7, 367-77.

CHENNA, R., SUGAWARA, H., KOIKE, T., LOPEZ, R., GIBSON, T. J., HIGGINS, D. G. & THOMPSON, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res,* 31, 3497-500.

CHIRDO, F. G., RUMBO, M., CARABAJAL, P., MAVROMATOPULOS, E., CASTAGNINO, N., ANON, M. C. & FOSSATI, C. A. (2000) Determination of anti-omega-gliadin antibodies in serologic tests for coeliac disease. *Scand J Gastroenterol,* 35, 508-16.

CHORZELSKI, T. P., BEUTNER, E. H., SULEJ, J., TCHORZEWSKA, H., JABLONSKA, S., KUMAR, V. & KAPUSCINSKA, A. (1984) IgA anti-endomysium antibody. A new immunological marker of dermatitis herpetiformis and coeliac disease. *Br J Dermatol,* 111, 395-402.

CICLITIRA, P. J., EVANS, D. J., FAGG, N. L., LENNOX, E. S. & DOWLING, R. H. (1984) Clinical testing of gliadin fractions in coeliac patients. *Clin Sci (Lond),* 66, 357-64.

CIECHANOVER, A. (1994) The ubiquitin-proteasome proteolytic pathway. *Cell,* 79, 13-21.

CRESSWELL, P., ACKERMAN, A. L., GIODINI, A., PEAPER, D. R. & WEARSCH, P. A. (2005) Mechanisms of MHC class I-restricted antigen processing and cross-presentation. *Immunol Rev,* 207, 145-57.

DANCHIN, E., VITIELLO, V., VIENNE, A., RICHARD, O., GOURET, P., MCDERMOTT, M. F. & PONTAROTTI, P. (2004) The major histocompatibility complex origin. *Immunol Rev,* 198, 216-32.

DE GROOT, A. S. & BERZOFSKY, J. A. (2004) From genome to vaccine--new immunoinformatics tools for vaccine design. *Methods,* 34, 425-8.

DE GROOT, A. S. & RAPPUOLI, R. (2004) Genome-derived vaccines. *Expert Rev Vaccines,* 3, 59-76.

DE OLIVEIRA, D. B., HARFOUCH-HAMMOUD, E., OTTO, H., PAPANDREOU, N. A., STERN, L. J., COHEN, H., BOEHM, B. O., BACH, J., CAILLAT-ZUCMAN, S., WALK, T., JUNG, G., ELIOPOULOS, E., PAPADOPOULOS, G. K. & VAN ENDERT, P. M. (2000) Structural analysis of two HLA-DR-presented autoantigenic epitopes: crucial role of peripheral but not central peptide residues for T-cell receptor recognition. *Mol Immunol,* 37, 813-25.

DEAVIN, A. J., AUTON, T. R. & GREANEY, P. J. (1996) Statistical comparison of established T-cell epitope predictors against a large database of human and murine antigens. *Mol Immunol,* 33, 145-55.

DEFRANCO, A. L., LOCKSLEY, R. M. & ROBERTSON, M. (2007) The MHC and Polymorphism of MHC Molecules. *Immunity: The Immune Response in Infections and Inflammatory Disease.* New Science Press.

DEL CARPIO, C. A., HENNIG, T., FICKEL, S. & YOSHIMORI, A. (2002) A combined bioinformatic approach oriented to the analysis and design of peptides with high affinity to MHC class I molecules. *Immunol Cell Biol,* 80, 286-99.

DESSEN, A., LAWRENCE, C. M., CUPO, S., ZALLER, D. M. & WILEY, D. C. (1997) X-ray crystal structure of HLA-DR4 (DRA*0101, DRB1*0401) complexed with a peptide from human collagen II. *Immunity,* 7, 473-81.

DIETERICH, W., EHNIS, T., BAUER, M., DONNER, P., VOLTA, U., RIECKEN, E. O. & SCHUPPAN, D. (1997) Identification of tissue transglutaminase as the autoantigen of celiac disease. *Nat Med,* 3, 797-801.

DIETERICH, W., LAAG, E., SCHOPPER, H., VOLTA, U., FERGUSON, A., GILLETT, H., RIECKEN, E. O. & SCHUPPAN, D. (1998) Autoantibodies to tissue transglutaminase as predictors of celiac disease. *Gastroenterology,* 115, 1317-21.

DOYTCHINOVA, I. A., BLYTHE, M. J. & FLOWER, D. R. (2002) Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. *J Proteome Res,* 1, 263-72.

DOYTCHINOVA, I. A. & FLOWER, D. R. (2002) Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three-dimensional quantitative structure-activity relationship study. *Proteins,* 48, 505-18.

DOYTCHINOVA, I. A. & FLOWER, D. R. (2003) Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics,* 19, 2263-70.

DOYTCHINOVA, I. A. & FLOWER, D. R. (2005) In silico identification of supertypes for class II MHCs. *J Immunol,* 174, 7085-95.

DOYTCHINOVA, I. A., WALSHE, V. A., JONES, N. A., GLOSTER, S. E., BORROW, P. & FLOWER, D. R. (2004) Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J Immunol,* 172, 7495-502.

DUFFY, M. J., LYNN, D. J., LLOYD, A. T. & O'SHEA, C. M. (2003) The ADAMs family of proteins: from basic studies to potential clinical applications. *Thromb Haemost,* 89, 622-31.

EDDY, S. R. (1998) Profile hidden Markov models. *Bioinformatics,* 14, 755-63.

EKINS, S., NIKOLSKY, Y., BUGRIM, A., KIRILLOV, E. & NIKOLSKAYA, T. (2007) Pathway mapping tools for analysis of high content data. *Methods Mol Biol,* 356, 319-50.

FALK, K., ROTZSCHKE, O., STEVANOVIC, S., JUNG, G. & RAMMENSEE, H. G. (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature,* 351, 290-6.

FARRELL, R. J. & KELLY, C. P. (2002) Celiac sprue. *N Engl J Med,* 346, 180-8.

FEIGHERY, C. (1999) Fortnightly review: coeliac disease. *Bmj,* 319, 236-9.

FELSENSTEIN, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol,* 266, 418-27.

FLOWER, D. R. (2003a) Databases and data mining for computational vaccinology. *Curr Opin Drug Discov Devel,* 6, 396-400.

FLOWER, D. R. (2003b) Towards in silico prediction of immunogenic epitopes. *Trends Immunol,* 24, 667-74.

GIANFRANI, C., TRONCONE, R., MUGIONE, P., COSENTINI, E., DE PASCALE, M., FARUOLO, C., SENGER, S., TERRAZZANO, G., SOUTHWOOD, S., AURICCHIO, S. & SETTE, A. (2003) Celiac disease association with CD8+ T cell responses: identification of a novel gliadin-derived HLA-A2-restricted epitope. *J Immunol,* 170, 2719-26.

GIUDICELLI, V. & LEFRANC, M. P. (1999) Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics,* 15, 1047-54.

GRIBSKOV, M., MCLACHLAN, A. D. & EISENBERG, D. (1987) Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A,* 84, 4355-8.

GRIFFIN, M., CASADIO, R. & BERGAMINI, C. M. (2002) Transglutaminases: nature's biological glues. *Biochem J,* 368, 377-96.

GROSS, D. M., FORSTHUBER, T., TARY-LEHMANN, M., ETLING, C., ITO, K., NAGY, Z. A., FIELD, J. A., STEERE, A. C. & HUBER, B. T. (1998) Identification of LFA-1 as a candidate autoantigen in treatment-resistant Lyme arthritis. *Science,* 281, 703-6.

GULUKOTA, K., SIDNEY, J., SETTE, A. & DELISI, C. (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol,* 267, 1258-67.

HAMMER, G. E., GONZALEZ, F., CHAMPSAUR, M., CADO, D. & SHASTRI, N. (2006) The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules. *Nat Immunol,* 7, 103-12.

HAWKING, S. (1988) *A Brief History of Time*, Bantam.

HE, X. L., RADU, C., SIDNEY, J., SETTE, A., WARD, E. S. & GARCIA, K. C. (2002) Structural snapshot of aberrant antigen presentation linked to

autoimmunity: the immunodominant epitope of MBP complexed with I-Au. *Immunity,* 17, 83-94.

HENIKOFF, J. G. & HENIKOFF, S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci,* 12, 135-43.

HENNECKE, J., CARFI, A. & WILEY, D. C. (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, 5611-24.

HENNECKE, J. & WILEY, D. C. (2002) Structure of a complex of the human alpha/beta T cell receptor (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity. *J Exp Med,* 195, 571-81.

HIGGINS, D. G. & SHARP, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene,* 73, 237-44.

HILL, A. V. (2006) Aspects of genetic susceptibility to human infectious diseases. *Annu Rev Genet,* 40, 469-86.

HOLZHUTTER, H. G., FROMMEL, C. & KLOETZEL, P. M. (1999) A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. *J Mol Biol,* 286, 1251-65.

HONEYMAN, M. C., BRUSIC, V., STONE, N. L. & HARRISON, L. C. (1998) Neural network-based prediction of candidate T-cell epitopes. *Nat Biotechnol,* 16, 966-9.

HYATT, G., MELAMED, R., PARK, R., SEGURITAN, R., LAPLACE, C., POIROT, L., ZUCCHELLI, S., OBST, R., MATOS, M., VENANZI, E., GOLDRATH, A., NGUYEN, L., LUCKEY, J., YAMAGATA, T., HERMAN, A., JACOBS, J., MATHIS, D. & BENOIST, C. (2006) Gene expression microarrays: glimpses of the immunological genome. *Nat Immunol,* 7, 686-91.

JARDETZKY, T. S., BROWN, J. H., GORGA, J. C., STERN, L. J., URBAN, R. G., CHI, Y. I., STAUFFACHER, C., STROMINGER, J. L. & WILEY, D. C. (1994) Three-dimensional structure of a human class II histocompatibility molecule complexed with superantigen. *Nature,* 368, 711-8.

JOJIC, N., REYES-GOMEZ, M., HECKERMAN, D., KADIE, C. & SCHUELER-FURMAN, O. (2006) Learning MHC I--peptide binding. *Bioinformatics,* 22, e227-35.

KAWASHIMA, S. & KANEHISA, M. (2000) AAindex: amino acid index database. *Nucleic Acids Res,* 28, 374.

KELLEY, J., WALTER, L. & TROWSDALE, J. (2005) Comparative genomics of major histocompatibility complexes. *Immunogenetics,* 56, 683-95.

KESMIR, C., NUSSBAUM, A. K., SCHILD, H., DETOURS, V. & BRUNAK, S. (2002) Prediction of proteasome cleavage motifs by neural networks. *Protein Eng,* 15, 287-96.

KHANNA, R. (2004) Predictive algorithms and T cell epitope mapping. *J Immunol,* 173, 2895; author reply 2895-6.

KIM, C. Y., QUARSTEN, H., BERGSENG, E., KHOSLA, C. & SOLLID, L. M. (2004) Structural basis for HLA-DQ2-mediated presentation of

gluten epitopes in celiac disease. *Proc Natl Acad Sci U S A,* 101, 4175-9.

KLEIN, J. & SATO, A. (2000) The HLA system. Second of two parts. *N Engl J Med,* 343, 782-6.

KORBER, B., LABUTE, M. & YUSIM, K. (2006) Immunoinformatics comes of age. *PLoS Comput Biol,* 2, e71.

KUBY, J. (1994) *Immunology,* New York, W.H. Freeman.

KUMANOVICS, A., TAKADA, T. & LINDAHL, K. F. (2003) Genomic organization of the mammalian MHC. *Annu Rev Immunol,* 21, 629-57.

KUMAR, S., TAMURA, K. & NEI, M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform,* 5, 150-63.

LEE, S., MILLER, S. A., WRIGHT, D. W., ROCK, M. T. & CROWE, J. E., JR. (2007) Tissue-specific regulation of CD8+ T-lymphocyte immunodominance in respiratory syncytial virus infection. *J Virol,* 81, 2349-58.

LEFRANC, M. P. (2005) IMGT, the international ImMunoGeneTics information system(R): a standardized approach for immunogenetics and immunoinformatics. *Immunome Res,* 1, 3.

LEFRANC, M. P., GIUDICELLI, V., BUSIN, C., MALIK, A., MOUGENOT, I., DEHAIS, P. & CHAUME, D. (1995) LIGM-DB/IMGT: an integrated database of Ig and TcR, part of the immunogenetics database. *Ann N Y Acad Sci,* 764, 47-9.

LIVINGSTONE, D. J., MANALLACK, D. T. & TETKO, I. V. (1997) Data modelling with neural networks: advantages and limitations. *J Comput Aided Mol Des,* 11, 135-42.

LOGEAN, A., SETTE, A. & ROGNAN, D. (2001) Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorg Med Chem Lett,* 11, 675-9.

LUNDIN, K. E., SCOTT, H., FAUSA, O., THORSBY, E. & SOLLID, L. M. (1994) T cells from the small intestinal mucosa of a DR4, DQ7/DR4, DQ8 celiac disease patient preferentially recognize gliadin when presented by DQ8. *Hum Immunol,* 41, 285-91.

LUSCOMBE, N. M., GREENBAUM, D. & GERSTEIN, M. (2001) What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med,* 40, 346-58.

LYNN, D. J., LLOYD, A. T., FARES, M. A. & O'FARRELLY, C. (2004) Evidence of positively selected sites in mammalian alpha-defensins. *Mol Biol Evol,* 21, 819-27.

MADDEN, D. R. (1995) The three-dimensional structure of peptide-MHC complexes. *Annu Rev Immunol,* 13, 587-622.

MAIURI, L., CIACCI, C., RICCIARDELLI, I., VACCA, L., RAIA, V., AURICCHIO, S., PICARD, J., OSMAN, M., QUARATINO, S. & LONDEI, M. (2003) Association between innate response to gliadin and activation of pathogenic T cells in coeliac disease. *Lancet,* 362, 30-7.

MAMITSUKA, H. (1998) Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins,* 33, 460-74.

MARTIN, M. P., GAO, X., LEE, J. H., NELSON, G. W., DETELS, R., GOEDERT, J. J., BUCHBINDER, S., HOOTS, K., VLAHOV, D., TROWSDALE, J., WILSON, M., O'BRIEN, S. J. & CARRINGTON, M. (2002) Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat Genet,* 31, 429-34.

MARTUCCI, S. & CORAZZA, G. R. (2002) Spreading and focusing of gluten epitopes in celiac disease. *Gastroenterology,* 122, 2072-5.

MCFARLAND, B. J., SANT, A. J., LYBRAND, T. P. & BEESON, C. (1999) Ovalbumin(323-339) peptide binds to the major histocompatibility complex class II I-A(d) protein using two functionally distinct registers. *Biochemistry,* 38, 16663-70.

MCLYSAGHT, A., BALDI, P. F. & GAUT, B. S. (2003) Extensive gene gain associated with adaptive evolution of poxviruses. *Proc Natl Acad Sci U S A,* 100, 15655-60.

MCLYSAGHT, A., HOKAMP, K. & WOLFE, K. H. (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet,* 31, 200-4.

MENG, W. S., VON GRAFENSTEIN, H. & HAWORTH, I. S. (2000) Water dynamics at the binding interface of four different HLA-A2-peptide complexes. *Int Immunol,* 12, 949-57.

MHC SEQUENCING CONSORTIUM (1999) Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature,* 401, 921-3.

MOLBERG, O., UHLEN, A. K., JENSEN, T., FLAETE, N. S., FLECKENSTEIN, B., ARENTZ-HANSEN, H., RAKI, M., LUNDIN, K. E. & SOLLID, L. M. (2005) Mapping of gluten T-cell epitopes in the bread wheat ancestors: implications for celiac disease. *Gastroenterology,* 128, 393-401.

MOUDGIL, K. D., SERCARZ, E. E. & GREWAL, I. S. (1998) Modulation of the immunogenicity of antigenic determinants by their flanking residues. *Immunol Today,* 19, 217-20.

MOUNT, D. W. (2004) *Bioinformatics : sequence and genome analysis,* Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

NANDA, N. K., ARZOO, K. K., GEYSEN, H. M., SETTE, A. & SERCARZ, E. E. (1995) Recognition of multiple peptide cores by a single T cell receptor. *J Exp Med,* 182, 531-9.

NELSON, C. A., PETZOLD, S. J. & UNANUE, E. R. (1993) Identification of two distinct properties of class II major histocompatibility complex-associated peptides. *Proc Natl Acad Sci U S A,* 90, 1227-31.

NUSSBAUM, A. K., KUTTLER, C., HADELER, K. P., RAMMENSEE, H. G. & SCHILD, H. (2001) PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics,* 53, 87-94.

OLAUSSEN, R. W., JOHANSEN, F. E., LUNDIN, K. E., JAHNSEN, J., BRANDTZAEG, P. & FARSTAD, I. N. (2002) Interferon-gamma-secreting T cells localize to the epithelium in coeliac disease. *Scand J Immunol,* 56, 652-64.

PALOSUO, K., VARJONEN, E., KEKKI, O. M., KLEMOLA, T., KALKKINEN, N., ALENIUS, H. & REUNALA, T. (2001) Wheat

omega-5 gliadin is a major allergen in children with immediate allergy to ingested wheat. *J Allergy Clin Immunol,* 108, 634-8.

PALOSUO, K., VARJONEN, E., NURKKALA, J., KALKKINEN, N., HARVIMA, R., REUNALA, T. & ALENIUS, H. (2003) Transglutaminase-mediated cross-linking of a peptic fraction of omega-5 gliadin enhances IgE reactivity in wheat-dependent, exercise-induced anaphylaxis. *J Allergy Clin Immunol,* 111, 1386-92.

PELTE, C., CHEREPNEV, G., WANG, Y., SCHOENEMANN, C., VOLK, H. D. & KERN, F. (2004) Random screening of proteins for HLA-A*0201-binding nine-amino acid peptides is not sufficient for identifying CD8 T cell epitopes recognized in the context of HLA-A*0201. *J Immunol,* 172, 6783-9.

PERELSON, A. S. (2002) Modelling viral and immune system dynamics. *Nat Rev Immunol,* 2, 28-36.

PETERS, B., TONG, W., SIDNEY, J., SETTE, A. & WENG, Z. (2003) Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics,* 19, 1765-72.

PIERTNEY, S. B. & OLIVER, M. K. (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity,* 96, 7-21.

QIAO, S. W., BERGSENG, E., MOLBERG, O., JUNG, G., FLECKENSTEIN, B. & SOLLID, L. M. (2005) Refining the rules of gliadin T cell epitope binding to the disease-associated DQ2 molecule in celiac disease: importance of proline spacing and glutamine deamidation. *J Immunol,* 175, 254-61.

RABINOWITZ, J. D., TATE, K., LEE, C., BEESON, C. & MCCONNELL, H. M. (1997) Specific T cell recognition of kinetic isomers in the binding of peptide to class II major histocompatibility complex. *Proc Natl Acad Sci U S A,* 94, 8702-7.

RAMMENSEE, H., BACHMANN, J., EMMERICH, N. P., BACHOR, O. A. & STEVANOVIC, S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics,* 50, 213-9.

RECHE, P. A., GLUTTING, J. P. & REINHERZ, E. L. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol,* 63, 701-9.

RECHE, P. A., GLUTTING, J. P., ZHANG, H. & REINHERZ, E. L. (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics,* 56, 405-19.

REIMERS, M. & CAREY, V. J. (2006) Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol,* 411, 119-34.

REINHERZ, E. L., TAN, K., TANG, L., KERN, P., LIU, J., XIONG, Y., HUSSEY, R. E., SMOLYAR, A., HARE, B., ZHANG, R., JOACHIMIAK, A., CHANG, H. C., WAGNER, G. & WANG, J. (1999) The crystal structure of a T cell receptor in complex with peptide and MHC class II. *Science,* 286, 1913-21.

ROBINSON, J. H. & DELVIG, A. A. (2002) Diversity in MHC class II antigen presentation. *Immunology,* 105, 252-62.

ROCHER, A., COLILLA, F., ORTIZ, M. L. & MENDEZ, E. (1992) Identification of the three major coeliac immunoreactive proteins and

one alpha-amylase inhibitor from oat endosperm. *FEBS Lett,* 310, 37-40.

RODGERS, J. R. & COOK, R. G. (2005) MHC class Ib molecules bridge innate and acquired immunity. *Nat Rev Immunol,* 5, 459-71.

ROGNAN, D., LAUEMOLLER, S. L., HOLM, A., BUUS, S. & TSCHINKE, V. (1999) Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem,* 42, 4650-8.

ROITT, I. M. & DELVES, P. J. (2001) *Roitt's essential immunology,* Oxford, Blackwell Scientific Publications.

SCHIRLE, M., WEINSCHENK, T. & STEVANOVIC, S. (2001) Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. *J Immunol Methods,* 257, 1-16.

SCHNEIDER, T. D. & STEPHENS, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res,* 18, 6097-100.

SEAMONS, A., SUTTON, J., BAI, D., BAIRD, E., BONN, N., KAFSACK, B. F., SHABANOWITZ, J., HUNT, D. F., BEESON, C. & GOVERMAN, J. (2003) Competition between two MHC binding registers in a single peptide processed from myelin basic protein influences tolerance and susceptibility to autoimmunity. *J Exp Med,* 197, 1391-7.

SEGEL, L. A. (2001) Controlling the immune system: diffuse feedback via a diffuse informational network. *Novartis Found Symp,* 239, 31-40; discussion 40-51.

SERCARZ, E. E. & MAVERAKIS, E. (2003) Mhc-guided processing: binding of large antigen fragments. *Nat Rev Immunol,* 3, 621-9.

SETTE, A., BUUS, S., APPELLA, E., SMITH, J. A., CHESNUT, R., MILES, C., COLON, S. M. & GREY, H. M. (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc Natl Acad Sci U S A,* 86, 3296-300.

SHAN, L., MOLBERG, O., PARROT, I., HAUSCH, F., FILIZ, F., GRAY, G. M., SOLLID, L. M. & KHOSLA, C. (2002) Structural basis for gluten intolerance in celiac sprue. *Science,* 297, 2275-9.

SHAN, L., QIAO, S. W., ARENTZ-HANSEN, H., MOLBERG, O., GRAY, G. M., SOLLID, L. M. & KHOSLA, C. (2005) Identification and analysis of multivalent proteolytically resistant peptides from gluten: implications for celiac sprue. *J Proteome Res,* 4, 1732-41.

SIDNEY, J., DEL GUERCIO, M. F., SOUTHWOOD, S. & SETTE, A. (2002) The HLA molecules DQA1*0501/B1*0201 and DQA1*0301/B1*0302 share an extensive overlap in peptide binding specificity. *J Immunol,* 169, 5098-108.

SIJTS, A. J., VILLANUEVA, M. S. & PAMER, E. G. (1996) CTL epitope generation is tightly linked to cellular proteolysis of a Listeria monocytogenes antigen. *J Immunol,* 156, 1497-503.

SJOSTROM, H., LUNDIN, K. E., MOLBERG, O., KORNER, R., MCADAM, S. N., ANTHONSEN, D., QUARSTEN, H., NOREN, O., ROEPSTORFF, P., THORSBY, E. & SOLLID, L. M. (1998) Identification of a gliadin T-cell epitope in coeliac disease: general

importance of gliadin deamidation for intestinal T-cell recognition. *Scand J Immunol,* 48, 111-5.

SOLLID, L. M. (2002) Coeliac disease: dissecting a complex inflammatory disorder. *Nat Rev Immunol,* 2, 647-55.

SOLLID, L. M. & KHOSLA, C. (2005) Future therapeutic options for celiac disease. *Nat Clin Pract Gastroenterol Hepatol,* 2, 140-7.

SOLLID, L. M., MARKUSSEN, G., EK, J., GJERDE, H., VARTDAL, F. & THORSBY, E. (1989) Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *J Exp Med,* 169, 345-50.

SOLLID, L. M., MOLBERG, O., MCADAM, S. & LUNDIN, K. E. (1997) Autoantibodies in coeliac disease: tissue transglutaminase--guilt by association? *Gut,* 41, 851-2.

SRINIVASAN, M., DOMANICO, S. Z., KAUMAYA, P. T. & PIERCE, S. K. (1993) Peptides of 23 residues or greater are required to stimulate a high affinity class II-restricted T cell response. *Eur J Immunol,* 23, 1011-6.

STEPNIAK, D. & KONING, F. (2006) Celiac disease--sandwiched between innate and adaptive immunity. *Hum Immunol,* 67, 460-8.

STEPNIAK, D., VADER, L. W., KOOY, Y., VAN VEELEN, P. A., MOUSTAKAS, A., PAPANDREOU, N. A., ELIOPOULOS, E., DRIJFHOUT, J. W., PAPADOPOULOS, G. K. & KONING, F. (2005) T-cell recognition of HLA-DQ2-bound gluten peptides can be influenced by an N-terminal proline at p-1. *Immunogenetics,* 57, 8-15.

STERN, L. J., BROWN, J. H., JARDETZKY, T. S., GORGA, J. C., URBAN, R. G., STROMINGER, J. L. & WILEY, D. C. (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature,* 368, 215-21.

STURNIOLO, T., BONO, E., DING, J., RADDRIZZANI, L., TUERECI, O., SAHIN, U., BRAXENTHALER, M., GALLAZZI, F., PROTTI, M. P., SINIGAGLIA, F. & HAMMER, J. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol,* 17, 555-61.

TENG, M. K., SMOLYAR, A., TSE, A. G., LIU, J. H., LIU, J., HUSSEY, R. E., NATHENSON, S. G., CHANG, H. C., REINHERZ, E. L. & WANG, J. H. (1998) Identification of a common docking topology with substantial variation among different TCR-peptide-MHC complexes. *Curr Biol,* 8, 409-12.

TONG, J. C., ZHANG, G. L., TAN, T. W., AUGUST, J. T., BRUSIC, V. & RANGANATHAN, S. (2006) Prediction of HLA-DQ3.2beta ligands: evidence of multiple registers in class II binding peptides. *Bioinformatics,* 22, 1232-8.

TOSELAND, C. P., CLAYTON, D. J., MCSPARRON, H., HEMSLEY, S. L., BLYTHE, M. J., PAINE, K., DOYTCHINOVA, I. A., GUAN, P., HATTOTUWAGAMA, C. K. & FLOWER, D. R. (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res,* 1, 4.

TROWSDALE, J. (1995) "Both man & bird & beast": comparative organization of MHC genes. *Immunogenetics,* 41, 1-17.

TROWSDALE, J. & PARHAM, P. (2004) Mini-review: defense strategies and immunity-related genes. *Eur J Immunol,* 34, 7-17.

VADER, L. W., DE RU, A., VAN DER WAL, Y., KOOY, Y. M., BENCKHUIJSEN, W., MEARIN, M. L., DRIJFHOUT, J. W., VAN VEELEN, P. & KONING, F. (2002a) Specificity of tissue transglutaminase explains cereal toxicity in celiac disease. *J Exp Med,* 195, 643-9.

VADER, L. W., STEPNIAK, D. T., BUNNIK, E. M., KOOY, Y. M., DE HAAN, W., DRIJFHOUT, J. W., VAN VEELEN, P. A. & KONING, F. (2003a) Characterization of cereal toxicity for celiac disease patients based on protein homology in grains. *Gastroenterology,* 125, 1105-13.

VADER, W., KOOY, Y., VAN VEELEN, P., DE RU, A., HARRIS, D., BENCKHUIJSEN, W., PENA, S., MEARIN, L., DRIJFHOUT, J. W. & KONING, F. (2002b) The gluten response in children with celiac disease is directed toward multiple gliadin and glutenin peptides. *Gastroenterology,* 122, 1729-37.

VADER, W., STEPNIAK, D., KOOY, Y., MEARIN, L., THOMPSON, A., VAN ROOD, J. J., SPAENIJ, L. & KONING, F. (2003b) The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten-specific T cell responses. *Proc Natl Acad Sci U S A,* 100, 12390-5.

VAN DE WAL, Y., KOOY, Y. M., DRIJFHOUT, J. W., AMONS, R. & KONING, F. (1996) Peptide binding characteristics of the coeliac disease-associated DQ(alpha1*0501, beta1*0201) molecule. *Immunogenetics,* 44, 246-53.

VAN DE WAL, Y., KOOY, Y. M., VAN VEELEN, P., VADER, W., AUGUST, S. A., DRIJFHOUT, J. W., PENA, S. A. & KONING, F. (1999) Glutenin is involved in the gluten-driven mucosal T cell response. *Eur J Immunol,* 29, 3133-9.

VAN DE WAL, Y., KOOY, Y. M., VAN VEELEN, P. A., PENA, S. A., MEARIN, L. M., MOLBERG, O., LUNDIN, K. E., SOLLID, L. M., MUTIS, T., BENCKHUIJSEN, W. E., DRIJFHOUT, J. W. & KONING, F. (1998) Small intestinal T cells of celiac disease patients recognize a natural pepsin fragment of gliadin. *Proc Natl Acad Sci U S A,* 95, 10050-4.

VAN HEEL, D. A., FRANKE, L., HUNT, K. A., GWILLIAM, R., ZHERNAKOVA, A., INOUYE, M., WAPENAAR, M. C., BARNARDO, M. C., BETHEL, G., HOLMES, G. K., FEIGHERY, C., JEWELL, D., KELLEHER, D., KUMAR, P., TRAVIS, S., WALTERS, J. R., SANDERS, D. S., HOWDLE, P., SWIFT, J., PLAYFORD, R. J., MCLAREN, W. M., MEARIN, M. L., MULDER, C. J., MCMANUS, R., MCGINNIS, R., CARDON, L. R., DELOUKAS, P. & WIJMENGA, C. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet,* 39, 827-9.

VAN HERPEN, T. W., GORYUNOVA, S. V., VAN DER SCHOOT, J., MITREVA, M., SALENTIJN, E., VORST, O., SCHENK, M. F., VAN VEELEN, P. A., KONING, F., VAN SOEST, L. J., VOSMAN, B., BOSCH, D., HAMER, R. J., GILISSEN, L. J. & SMULDERS, M. J.

(2006) Alpha-gliadin genes from the A, B, and D genomes of wheat contain different sets of celiac disease epitopes. *BMC Genomics,* 7**,** 1.

VINER, N. J., NELSON, C. A., DECK, B. & UNANUE, E. R. (1996) Complexes generated by the binding of free peptides to class II MHC molecules are antigenically diverse compared with those generated by intracellular processing. *J Immunol,* 156**,** 2365-8.

WITT, S. N. & MCCONNELL, H. M. (1994) Formation and dissociation of short-lived class II MHC-peptide complexes. *Biochemistry,* 33**,** 1861-8.

XIA, J., SOLLID, L. M. & KHOSLA, C. (2005) Equilibrium and kinetic analysis of the unusual binding behavior of a highly immunogenic gluten peptide to HLA-DQ2. *Biochemistry,* 44**,** 4442-9.

ZANONI, G., NAVONE, R., LUNARDI, C., TRIDENTE, G., BASON, C., SIVORI, S., BERI, R., DOLCINO, M., VALLETTA, E., CORROCHER, R. & PUCCETTI, A. (2006) In celiac disease, a subset of autoantibodies against transglutaminase binds toll-like receptor 4 and induces activation of monocytes. *PLoS Med,* 3**,** e358.

Appendices

*Appendix A Training set data for HLA-A*0201*

Sequence	-lnIC50	IC50(nM)	REFERENCE
AIIDPLIYA	15.16	238	CANCER RESEARCH 1997 57 4348-4355
ALIICNAII	13.93	893	CANCER RESEARCH 1997 57 4348-4355
ALVARAAVL	11.41	11111	CANCER RESEARCH 1997 57 4348-4355
AVMDPFIYA	14.09	758	CANCER RESEARCH 1997 57 4348-4355
CLALSDLLV	14.85	357	CANCER RESEARCH 1997 57 4348-4355
FLALIICNA	16.60	39	CANCER RESEARCH 1997 57 4348-4355
FLAMLVLMA	13.08	2083	CANCER RESEARCH 1997 57 4348-4355
FLCWGPFFL	17.09	37	CANCER RESEARCH 1997 57 4348-4355
FLHLTLIVL	12.21	5000	CANCER RESEARCH 1997 57 4348-4355
FLSLGLVSL	15.40	206	CANCER RESEARCH 1997 57 4348-4355
GLAANQTGA	12.95	2381	CANCER RESEARCH 1997 57 4348-4355
GLFLSLGLV	14.44	535	CANCER RESEARCH 1997 57 4348-4355
GLVSLVENA	14.45	532	CANCER RESEARCH 1997 57 4348-4355
ILLGIFFLC	12.10	5556	CANCER RESEARCH 1997 57 4348-4355
ILLLEAGAL	12.25	4762	CANCER RESEARCH 1997 57 4348-4355
LLCLVVFFL	14.21	676	CANCER RESEARCH 1997 57 4348-4355
LLVSGSNVL	13.47	1408	CANCER RESEARCH 1997 57 4348-4355
LVLMAVLYV	13.62	1210	CANCER RESEARCH 1997 57 4348-4355
MLARACQHA	13.32	1639	CANCER RESEARCH 1997 57 4348-4355
NGMLIMCNA	11.00	16667	CANCER RESEARCH 1997 57 4348-4355
RLHKRQRPV	12.74	2941	CANCER RESEARCH 1997 57 4348-4355
SLCFLGAIA	11.51	10000	CANCER RESEARCH 1997 57 4348-4355
SLNSTPTAI	14.27	633	CANCER RESEARCH 1997 57 4348-4355
SLVENALVV	14.66	429	CANCER RESEARCH 1997 57 4348-4355
SVIDPLIYA	13.64	1190	CANCER RESEARCH 1997 57 4348-4355
SVMDPLIYA	16.30	83	CANCER RESEARCH 1997 57 4348-4355
TILIGVFVV	12.30	4545	CANCER RESEARCH 1997 57 4348-4355
TILLGIFFL	15.38	96	CANCER RESEARCH 1997 57 4348-4355
TILLGVFIF	12.30	4545	CANCER RESEARCH 1997 57 4348-4355
TLPRARRRV	10.60	25000	CANCER RESEARCH 1997 57 4348-4355
VLETAVGLL	15.25	238	CANCER RESEARCH 1997 57 4348-4355
YLILIMCNS	11.98	6250	CANCER RESEARCH 1997 57 4348-4355
YLVLINCNS	13.04	2174	CANCER RESEARCH 1997 57 4348-4355
ALPYWNFAT	13.40	1515	CANCER RESEARCH 1998 58 4895-4901
ALVGLFVLL	17.39	28	CANCER RESEARCH 1998 58 4895-4901
FVTWHRYHL	13.51	1351	CANCER RESEARCH 1998 58 4895-4901
LIGNESFAL	14.69	417	CANCER RESEARCH 1998 58 4895-4901
LLSCLGCKI	12.30	4545	CANCER RESEARCH 1998 58 4895-4901
LLVVMFTLV	13.51	1351	CANCER RESEARCH 1998 58 4895-4901
QVMSLHNLV	13.87	943	CANCER RESEARCH 1998 58 4895-4901
SAANDPIFV	12.30	4545	CANCER RESEARCH 1998 58 4895-4901
SLDDYNHLV	17.47	26	CANCER RESEARCH 1998 58 4895-4901

TLDSQVMSL	15.15	263	CANCER RESEARCH 1998 58 4895-4901
TLLVVMGTL	12.85	2632	CANCER RESEARCH 1998 58 4895-4901
VALVGLFVL	11.70	8333	CANCER RESEARCH 1998 58 4895-4901
VCMTVDSLV	11.85	7143	CANCER RESEARCH 1998 58 4895-4901
VLHSFTDAI	14.21	676	CANCER RESEARCH 1998 58 4895-4901
VMGTLVALV	17.36	29	CANCER RESEARCH 1998 58 4895-4901
VTWHRYHLL	15.64	161	CANCER RESEARCH 1998 58 4895-4901
GLRDLAVAV	13.43	1470	CELL 1993 74 929-937
HLEGKVILV	11.70	8333	CELL 1993 74 929-937
HLSLRGLPV	11.29	12500	CELL 1993 74 929-937
HLYQGCQVV	15.73	147	CELL 1993 74 929-937
LLWKGEGAV	15.34	217	CELL 1993 74 929-937
PLTSIISAV	12.61	3333	CELL 1993 74 929-937
RLCVQSTHV	11.00	16666	CELL 1993 74 929-937
SLYAVSPSV	16.56	64	CELL 1993 74 929-937
WLSLLVPFV	18.63	6.9	CELL 1993 74 929-937
DIEITCVYC	14.25	649	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
ELTEVFEFA	14.51	500	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
FKDLFVVYR	14.51	500	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
FLNTLSFVC	13.87	943	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
FQQLFLNTL	14.69	417	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
GLYNLLIRC	15.82	135	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
KCIDFYSRI	14.69	417	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
KLPDLCTEL	14.32	602	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
KTVLELTEV	16.06	106	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
LFLNTLSFV	14.01	820	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
LQDIEITCV	16.27	86	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
LYNLLIRCL	14.21	676	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
SLQDIEITC	14.45	532	CLINICAL CANCER RESEARCH 2001 7 SUPP. 788-795
FAFRDLCIV	16.05	87.7	EURO.JOURNAL OF IMMUNOLOGY 1996 26 97-101
FLLSLGIHL	18.55	7.7	EURO.JOURNAL OF IMMUNOLOGY 1996 26 97-101
ILAGYGAGV	15.97	115.5	EURO.JOURNAL OF IMMUNOLOGY 1996 26 97-101
KLPQLCTEL	12.10	5555.6	EURO.JOURNAL OF IMMUNOLOGY 1996 26 97-101
PLLPIFFCL	16.01	76.9	EURO.JOURNAL OF IMMUNOLOGY 1996 26 97-101
TLGIVCPIC	16.11	66.7	EURO.JOURNAL OF IMMUNOLOGY 1996 26 97-101
VLQAGFFLL	16.30	33.3	EURO.JOURNAL OF IMMUNOLOGY 1996 26 97-101
YMLDLQPET	16.99	35.7	EURO.JOURNAL OF IMMUNOLOGY 1996 26 97-101
FLCKQYLNL	15.25	238	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
FLCKQYLNL	15.25	238	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
GLYSSTVPV	17.48	20	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
HLESLFTAV	12.21	5000	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
HLLVGSSGL	13.34	1613	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
LLGCAANWI	12.21	5000	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
LLSSNLSWL	14.60	455	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
NLGNLNVSI	16.39	76	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
NLQSLTNLL	13.82	1000	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
NLSWLSLDV	15.58	77	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
NLYVSLLLL	16.38	77	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678
SLNFMGYVI	13.54	1316	EURO.JOURNAL OF IMMUNOLOGY 1997 27 671-678

YLVAYQATV	17.09	20.4	HUMAN IMMUNOLOGY 2001 62 1200-1216
GLIMVLSFL	17.63	22	IMMUNITY 1997 7 97-112
GLLGNVSTV	17.55	24	IMMUNITY 1997 7 97-112
KILSVFFLA	19.11	5	IMMUNITY 1997 7 97-112
VLAGLLGNV	17.78	19	IMMUNITY 1997 7 97-112
FMDGTMSQV	16.45	72	IMMUNOLOGICAL REVIEWS 2002 188 136-146
KTWGKYWQV	17.50	25	IMMUNOLOGICAL REVIEWS 2002 188 136-146
YLESGSVTA	16.43	73	IMMUNOLOGICAL REVIEWS 2002 188 136-146
YMDGTMSQV	16.42	74	IMMUNOLOGICAL REVIEWS 2002 188 136-146
ALCRWGLLL	16.12	100	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
ALIHHNTHL	15.25	238.1	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
ILDEAYVMA	15.25	238.1	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
ILHNGAYSL	16.41	74.6	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
KIFGSLAFL	17.22	33.3	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
QLFEDNYAL	17.87	17.3	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
QLMPYGCLL	15.42	201	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
RLLQETELV	17.69	20.8	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
SIISAVVGI	16.48	69.4	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
VLIQRNPQL	17.60	22.7	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
VVLGVVFGI	18.06	14.3	INTERNATIONAL JOURNAL OF CANCER 1998 78 202-208
ALMPLYACI	17.96	10	JOURNAL OF CLINICAL INVESTIGATION 1997 100 503-513
SLFNTIATL	13.80	1014	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLFNTIAVL	16.18	94	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLFNTVATL	13.52	1339	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLFNTVAVL	16.73	54	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYITVATL	13.46	1428	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYITVAVL	11.00	16667	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNAIATL	17.14	36	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNAVATL	17.36	29	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNLVAVL	17.43	27	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNTIATL	14.29	620	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNTIAVL	17.06	39	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNTISVL	15.12	270	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNTITVL	15.70	152	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNTVATL	16.81	50	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566

SLYNTVAVL	15.87	128	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
SLYNTVSTL	15.65	159	JOURNAL OF CLINICAL INVESTIGATION 1998 101 2559-2566
FLYNRPLSV	15.74	146	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
FMGAGSKAV	14.28	631	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
FVDYNFTIV	15.24	240	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
FVNHDFTVV	15.02	300	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
GIRPYEILA	17.23	33	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
IAGGVMAVV	15.45	196	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
IVMGNGTLV	13.82	998	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
KLFPEVIDL	15.41	203	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
LLLLGLWGL	17.63	22	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
LLPSLFLLL	15.89	125	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
RLTEELNTI	13.95	871	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
SVYVDAKLV	16.10	102	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
VLAKDGTEV	16.52	67	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
VLLPSLFLL	17.14	36	JOURNAL OF CLINICAL INVESTIGATION 1998 102 1062-1071
HLYSHPIIL	16.75	38	JOURNAL OF EXPERIMENTAL MEDICINE 1995 181 1047-1058
LLAQFTSAI	16.81	50	JOURNAL OF EXPERIMENTAL MEDICINE 1995 181 1047-1058
YMDDVVLGA	16.36	31	JOURNAL OF EXPERIMENTAL MEDICINE 1995 181 1047-1058
ALAKAAAAV	15.34	125	JOURNAL OF IMMUNOLOGY 1994 152 2874-2881
LLDVPTAAV	17.89	17	JOURNAL OF IMMUNOLOGY 1994 152 2874-2881
SLLPAIVEL	17.55	24	JOURNAL OF IMMUNOLOGY 1994 152 2874-2881
YLLPAIVHI	17.83	18	JOURNAL OF IMMUNOLOGY 1994 152 2874-2881
FLCKQYLNL	17.36	29	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
FLGGTPVCL	15.25	238	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
FLLTRILTI	18.60	7.1	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
GLLGWSPQA	18.56	5.8	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
GLSRYVARL	16.60	42	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
IISCTCPTV	15.15	263	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
ILLLCLIFL	15.76	143	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
ILSPFMPLL	16.92	45	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592

KLHLYSHPI	17.16	17	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
LLCLIFLLV	16.11	101	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
LLCLIFLLV	16.11	101	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
LLLCLIFLL	17.47	26	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
LLWFHISCL	15.39	208	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
MMWYWGPSL	18.24	12	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
QLFHLCLII	15.86	130	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
SLYADSPSV	17.70	14	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
VLLDYQGML	16.43	47	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
WILRGTSFV	15.10	278	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
WILRGTSFV	15.10	278	JOURNAL OF IMMUNOLOGY 1994 153 5586-5592
GIGILTVIL	13.82	1000	JOURNAL OF IMMUNOLOGY 1995 154 2257-2265
ILTVILGVL	14.78	381	JOURNAL OF IMMUNOLOGY 1995 154 2257-2265
LTVILGVLL	12.85	2632	JOURNAL OF IMMUNOLOGY 1995 154 2257-2265
SLHVGTQCA	13.45	1439	JOURNAL OF IMMUNOLOGY 1995 154 2257-2265
TTAEEAAGI	12.39	4167	JOURNAL OF IMMUNOLOGY 1995 154 2257-2265
TVILGVLLL	13.98	847	JOURNAL OF IMMUNOLOGY 1995 154 2257-2265
VILGVLLLI	15.62	164	JOURNAL OF IMMUNOLOGY 1995 154 2257-2265
AAGIGILTV	14.12	395	JOURNAL OF IMMUNOLOGY 1995 154 3961-3968
MLLAVLYCL	14.92	333	JOURNAL OF IMMUNOLOGY 1995 154 3961-3968
YMNGTMSQV	17.03	40	JOURNAL OF IMMUNOLOGY 1995 154 3961-3968
AMFQDPQER	13.22	1818	JOURNAL OF IMMUNOLOGY 1995 154 5934-5943
GTLGIVCPI	15.36	193	JOURNAL OF IMMUNOLOGY 1995 154 5934-5943
LQTTIHDII	12.67	3157	JOURNAL OF IMMUNOLOGY 1995 154 5934-5943
MLDLQPETT	13.91	462	JOURNAL OF IMMUNOLOGY 1995 154 5934-5943
TLHEYMLDL	14.82	188	JOURNAL OF IMMUNOLOGY 1995 154 5934-5943
AAAKAAAAV	14.73	400	JOURNAL OF IMMUNOLOGY 1995 154 685-693
AIAKAAAAV	14.22	667	JOURNAL OF IMMUNOLOGY 1995 154 685-693
ALAKAAAAA	16.00	113	JOURNAL OF IMMUNOLOGY 1995 154 685-693
ALAKAAAAI	14.30	615	JOURNAL OF IMMUNOLOGY 1995 154 685-693
ALAKAAAAL	14.99	308	JOURNAL OF IMMUNOLOGY 1995 154 685-693
ALAKAAAAM	17.03	40	JOURNAL OF IMMUNOLOGY 1995 154 685-693
AVAKAAAAV	14.95	320	JOURNAL OF IMMUNOLOGY 1995 154 685-693
DPKVKQWPL	14.22	667	JOURNAL OF IMMUNOLOGY 1995 154 685-693
LLFGYPVYV	18.16	13	JOURNAL OF IMMUNOLOGY 1995 154 685-693
FLEPGPVTA	15.88	126.4	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
FTDQVPFSV	16.61	61.4	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
IIDQVPFSV	17.03	40	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ILAQVPFSV	18.28	11.5	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ILDQVPFSV	16.77	52	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ILMQVPFSV	18.71	7.5	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ILSQVPFSV	17.73	20	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
IMDQVPFSV	17.77	19.1	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ITAQVPFSV	16.16	95.6	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ITFQVPFSV	16.53	66.2	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ITMQVPFSV	17.03	40	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ITSQVPFSV	14.27	637	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ITWQVPFSV	17.17	34.9	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
ITYQVPFSV	17.22	33.1	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548

WLDQVPFSV	18.28	11.5	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
WLEPGPVTA	14.01	827.3	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
WTDQVPFSV	14.15	716.7	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLAPGPVTA	18.49	9.3	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLAPGPVTV	18.00	15.2	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLEPGPVTI	16.55	65	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLEPGPVTL	16.25	87.5	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLEPGPVTV	16.91	45.5	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLFPGPVTV	18.97	5.8	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLMPGPVTV	18.26	11.7	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLSPGPVTA	17.00	41.4	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLSPGPVTV	17.60	22.8	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLWPGPVTV	18.71	7.5	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLYPGPVTA	17.90	16.9	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YLYPGPVTV	18.54	8.9	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
YTDQVPFSV	16.27	86	JOURNAL OF IMMUNOLOGY 1996 157 2539-2548
GLGQVPLIV	14.51	500	JOURNAL OF IMMUNOLOGY 1997 158 1796-1802
HLAVIGALL	16.09	100	JOURNAL OF IMMUNOLOGY 1997 158 1796-1802
KTWGQYWQV	18.02	11.1	JOURNAL OF IMMUNOLOGY 1997 158 1796-1802
LLAVGATKV	14.91	333.3	JOURNAL OF IMMUNOLOGY 1997 158 1796-1802
RLMKQDFSV	16.91	45.5	JOURNAL OF IMMUNOLOGY 1997 158 1796-1802
YLEPGPVTA	15.27	94.3	JOURNAL OF IMMUNOLOGY 1997 158 1796-1802
ALTVVWLLV	15.87	128	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
ALYGALLLA	18.75	7.2	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
FLAGALLLA	14.33	599	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
FLYAALLLA	10.97	17177	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
FLYGALLAA	18.88	6.3	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
FLYGALRLA	18.76	7.1	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
FLYGALVLA	17.06	39	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
GLCFFGVAL	12.39	4167	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
IAATYNFAV	16.19	93	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
LLVFACSAV	14.60	455	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
LVSLLTFMI	13.16	1923	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
QMTFHLFIA	13.30	1667	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
RMYGVLPWI	17.36	29	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
SLLTFMIAA	18.48	9.4	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
SLSRFSWGA	13.91	909	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
VIHAFQYVI	13.62	1220	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
VVHFFKNIV	9.90	50000	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
YALTVVWLL	15.94	119	JOURNAL OF IMMUNOLOGY 1997 159 4943-4951
SAICSVVRR	13.46	1429	JOURNAL OF IMMUNOLOGY 1998 161 4447-4455
ALLAGLVSL	16.39	76.3	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
ALMDKSLHV	17.88	17.2	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
AMVGAVLTA	16.40	75.5	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
AVIGALLAV	17.84	17.9	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
FLLRWEQEI	17.48	25.6	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
FLPWHRLFL	16.00	112.3	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
FLWGPRALV	16.61	61	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
FVWLHYYSV	18.00	15.2	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976

IMPGQEAGL	16.55	64.9	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
LLAGLVSLL	16.17	95.2	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
LLAVLYCLL	17.22	33.3	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
LLLEAGALV	18.18	6.7	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
LLWSFQTSA	18.00	15.2	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
LMAVVLASL	16.01	111.1	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
MLGTHTMEV	18.06	14.3	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
RIWSWLLGA	16.12	100	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
RLLGSLNST	15.61	166.7	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
SIIDPLIYA	14.60	454.6	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
SLADTNSLA	14.60	454.6	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
SVYDFFVWL	16.83	35.7	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
VLTALLAGL	16.32	82	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
VVMGTLVAL	16.30	66.7	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
YAIDLPVSV	17.96	15.6	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
YVITTQHWL	15.86	104.2	JOURNAL OF IMMUNOLOGY 1998 161 6970-6976
HMWNFISGI	18.00	15.2	JOURNAL OF IMMUNOLOGY 1999 162 6681-6689
LLFLLLADA	15.34	217.4	JOURNAL OF IMMUNOLOGY 1999 162 6681-6689
VLVGGVLAA	15.50	185.2	JOURNAL OF IMMUNOLOGY 1999 162 6681-6689
WMNRLIAFA	15.92	122	JOURNAL OF IMMUNOLOGY 1999 162 6681-6689
YLVTRHADV	14.60	454.5	JOURNAL OF IMMUNOLOGY 1999 162 6681-6689
ITDQVPFSV	16.12	70	JOURNAL OF IMMUNOLOGY 2000 164 2354-2361
YIGEVLVSM	16.53	66	JOURNAL OF IMMUNOLOGY 2001 167 3223-3230
YIGSVLISV	18.81	6.8	JOURNAL OF IMMUNOLOGY 2001 167 3223-3230
AIHNVVHAI	15.61	166	JOURNAL OF IMMUNOLOGY 2003 170 2719-2726
QLIPCMDVV	15.73	147	JOURNAL OF IMMUNOLOGY 2003 170 2719-2726
VLQQSTYQL	16.66	58	JOURNAL OF IMMUNOLOGY 2003 170 2719-2726
ALSTGLIHL	14.98	312.5	JOURNAL OF VIROLOGY 1995 69 2462-2470
DLMGYIPLV	16.34	80	JOURNAL OF VIROLOGY 1995 69 2462-2470
FLLLADARV	17.83	17.9	JOURNAL OF VIROLOGY 1995 69 2462-2470
GLQDCTMLV	17.59	23	JOURNAL OF VIROLOGY 1995 69 2462-2470
ALGLVCVQA	14.91	333.3	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
ALGLVCVQM	12.21	5000	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
ALREEEEGV	15.25	238.1	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
FLWGPRALA	15.94	119	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
GLLGDNQIM	13.84	980.4	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
GLVCVQAAT	13.04	2173.9	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
IMPKTGFLI	14.25	649.4	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
KVADLVGFL	14.37	574.7	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
KVLEYVIKV	16.71	55.6	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
LVLGTLEEV	15.67	156.3	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
RALAETSYV	14.51	500	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
SLHCKPEEA	14.77	384.6	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
VADLVGFLL	12.90	2500	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
YVIKVSARV	14.85	357.1	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
YVLVTCLGL	13.04	2173.9	MOLECULAR IMMUNOLOGY 1994 31 1423-1430
FLPSDFFPS	15.96	117	MOLECULAR IMMUNOLOGY 1994 31 813-822
GILGFVFTL	18.93	6	MOLECULAR IMMUNOLOGY 1994 31 813-822
LLGRNSFEV	15.97	38	MOLECULAR IMMUNOLOGY 1994 31 813-822

AIYHPQQFV	14.98	313	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
ALLSDWLPA	16.18	94.3	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
ALVLLMLPV	17.28	31.2	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
AMKADIQHV	15.61	167	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
AMLQDMAIL	16.14	98	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
DMWEHAFYL	15.84	132	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
FLLPDAQSI	14.77	384.6	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
FVVALIPLV	18.70	7.6	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
GLFLTTEAV	17.29	31	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
GLMTAVYLV	18.54	8.9	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
GLVDFVKHI	15.34	217.3	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
GLYGAQYDV	15.20	250	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
GLYLSQIAV	16.16	96.1	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
GLYRQWALA	15.50	185.1	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
GLYYLTTEV	17.69	20.8	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
ILFTFLHLA	19.04	5.4	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
ILLSIARVV	14.60	454.5	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
ILSSLGLPV	16.81	50	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
KLAGGVAVI	14.85	357	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
LLACAVIHA	15.20	250	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
LLFLGVVFL	16.81	50	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
LLFRFMRPL	17.15	35.7	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
LLPLGYPFV	14.91	333.3	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
LLWQDPVPA	16.91	45.4	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
LMIGTAAAV	16.35	79	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
LMLPGMNGI	15.25	238	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
MALLRLPLV	16.76	52.6	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
MLASTLTDA	15.20	250	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
MLGNAPSVV	15.30	227.2	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
MLQDMAILT	15.61	167	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
MTYAAPLFV	18.10	13.8	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
RLLDDTPEV	16.16	96.1	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
RLMIGTAAA	15.30	227	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
RLPLVLPAV	19.09	5.1	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
RLVSGLVGA	15.70	152	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
RMFAANLGV	17.15	35.7	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
RMPAVTDLV	15.89	125	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
SLAGFVRML	16.01	111.1	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
SLLEIGEGV	16.14	98	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
SLYFGGICV	18.36	10.6	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
TVLRFVPPL	16.38	76.9	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
VLLLDVTPL	16.81	50	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
WLLIDTSNA	14.85	357.1	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
YLDLALMSV	19.02	5.5	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
YLLALRYLA	18.42	10	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
YLSEGDMAA	15.04	294.1	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
YLSQIAVLL	18.23	12.1	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
YLYVHSPAL	19.04	5.4	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215
YTYKWETFL	17.36	29	PROC.NATL.ACAD.SCI. USA 2000 97 12210-12215

Appendix B HLA-A2 Validation Set

EPITOPE	-lnIC50	IC50(nM)	REFERENCE
ALBRWGLLV	17.83	18	JOURNAL OF IMMUNOLOGY 2001 167 787-796
ALCRWGLLL	16.12	100	JOURNAL OF IMMUNOLOGY 2001 167 787-796
ALNKMFBQV	15.47	192	JOURNAL OF IMMUNOLOGY 2001 167 787-796
ALNKMFCQL	14.12	735	JOURNAL OF IMMUNOLOGY 2001 167 787-796
ATVGIMIGV	17.15	35.7	JOURNAL OF IMMUNOLOGY 2001 167 787-796
CLTSTVQLV	15.39	208	JOURNAL OF IMMUNOLOGY 2001 167 787-796
CQLAKTCPV	15.38	208.3	JOURNAL OF IMMUNOLOGY 2001 167 787-796
FLWGPRALV	17.28	31.3	JOURNAL OF IMMUNOLOGY 2001 167 787-796
HLYQGCQVV	15.79	139	JOURNAL OF IMMUNOLOGY 2001 167 787-796
ILHNGAYSL	16.41	75	JOURNAL OF IMMUNOLOGY 2001 167 787-796
IMIGVLVGV	16.50	68.5	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KASEYLQLV	15.70	151.5	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KIFGSLAFL	17.15	35.7	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KIWEELSVL	14.85	357.1	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KLBPVQLWV	16.89	46	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KLCPVQLWV	15.92	122	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KLFCQLAKV	16.14	98	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KLFGSLAFV	18.97	5.8	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KMFBQLAKV	15.89	125	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KMFCQLAKT	14.92	333	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KTCPVQLWV	14.14	725	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KVAELVHFL	16.50	68.5	JOURNAL OF IMMUNOLOGY 2001 167 787-796
KVFGSLAFV	17.73	20	JOURNAL OF IMMUNOLOGY 2001 167 787-796
QIIGYVIGT	14.98	312.5	JOURNAL OF IMMUNOLOGY 2001 167 787-796
SIISAVVGI	16.48	69.4	JOURNAL OF IMMUNOLOGY 2001 167 787-796
SLPPPGTRV	15.92	122	JOURNAL OF IMMUNOLOGY 2001 167 787-796
SMPPPGTRV	15.58	172	JOURNAL OF IMMUNOLOGY 2001 167 787-796
STPPPGTRV	13.91	909	JOURNAL OF IMMUNOLOGY 2001 167 787-796
VLLGVVFGV	19.85	2.4	JOURNAL OF IMMUNOLOGY 2001 167 787-796
VLYGPDTPI	15.42	200	JOURNAL OF IMMUNOLOGY 2001 167 787-796
VLYGPDTPV	17.73	20	JOURNAL OF IMMUNOLOGY 2001 167 787-796
VVLGVVFGI	18.08	14	JOURNAL OF IMMUNOLOGY 2001 167 787-796
YLSGANLNL	17.40	27.8	JOURNAL OF IMMUNOLOGY 2001 167 787-796
YLSGANLNV	16.43	73	JOURNAL OF IMMUNOLOGY 2001 167 787-796

Appendix C HLA-DR4 Training Set

EPITOPE	-lnIC50'	IC50(nM)	REFERENCE
AGLLGNVSTVLLGGV	11.8807	6923	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
AKYVKQNTLKLAT	17.6098	22.5	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
APYHFDLSGHAFGSMAKKGE	17.5044	25	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
DIEKKIAKMEKASSVFNVVNS	15.1978	251	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
DNVLDHLTGRSC	10.702	22500	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
DTPYLDITYHFVMQRLPL	13.9411	882	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
EFWEFDLPGIKA	14.9211	331	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
EKVYLAWVPAHKGIG	18.4933	9.3	JOURNAL OF VIROLOGY 2001 75 4195-4207
EKYVKQNTLKLAT	17.2575	32	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
EPQGSTYAASSATSVD	14.0691	776	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
ESEFQAALSRKVAKL	14.9487	322	CLINICAL CANCER RESEARCH 2003 9 947-954
EVWREEAYHAADIKD	12.4067	4091	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
FATCFLIPLTSQFFLP	12.2496	4787	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
FNNFTVSFWLRVPK…	14.4637	523	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
FRKYTAFTIPSINNE	15.0397	294	JOURNAL OF VIROLOGY 2001 75 4195-4207
GACPKYVKQNTLKLATGMR	17.1679	35	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
GEIYKRWIILGLNKI	13.3045	1667	JOURNAL OF VIROLOGY 2001 75 4195-4207
GLAYKFVVPGAATPY	16.9166	45	JOURNAL OF IMMUNOLOGY 2000 165 1123-1137
HKYVKQNTLKLAT	17.6787	21	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
HNWVNHAVPLAMKLI	18.2384	12	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
HSNWRAMASDFNLPP	18.3254	11	JOURNAL OF VIROLOGY 2001 75 4195-4207
IFSKASDSLQLVFGIE	12.7206	2989	CLINICAL CANCER RESEARCH 2003 9 947-954
IGCWYCRRRNGYRAL	11.338	11912	PROC.NATL.ACAD.SCI. USA 2000 97 400-405
IKLPIILAFATCFLIP	14.144	720	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
IPQEWKPAITVKVLPA	16.7914	51	JOURNAL OF IMMUNOLOGY 1998 160 3363-

			3373
KKYVKQNTLKLAT	17.7788	19	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
KRWIILGLNKIVRMY	14.2728	633	JOURNAL OF VIROLOGY 2001 75 4195-4207
KSKYKLATSVLAGLL	18.7632	7.1	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
KTAVQMAVFIHNFKR	13.9109	909	JOURNAL OF VIROLOGY 2001 75 4195-4207
KVYLAWVPAHKGIGG	15.6734	156	JOURNAL OF VIROLOGY 2001 75 4195-4207
KYKIAGGIAGGLALL	12.4938	3750	JOURNAL OF IMMUNOLOGY 2000 165 1123-1137
KYVKQNTLKLATGMR	17.5044	25	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
LAAIIFLFGPPTALRS	13.9792	849	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
LGNWQYFFPVIFSKASDSL	13.912	908	CLINICAL CANCER RESEARCH 2003 9 947-954
LLKYRARAPVTKA...	14.2871	624	CLINICAL CANCER RESEARCH 2003 9 947-954
LTQYFVQENYLEYRQVPG	12.2517	4777	CLINICAL CANCER RESEARCH 2003 9 947-954
LTSQFFLPALPVFTWL	15.3573	214	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
LVNLLIFHINGKIIK	12.3934	4146	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
LWWSTMYLTHHYFVDL	13.8306	985	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
MNYYGKQENWYSLKK	17.6787	21	JOURNAL OF IMMUNOLOGY 2000 165 1123-1137
MRKLAILSVSSFLFV	12.063	5769	JOURNAL OF IMMUNOLOGY 2000 165 1123-1137
NLSNVLATITTGVLDI	12.0501	5844	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
NVKYLVIVFLIFFDL	14.0979	754	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
PAYEKLSAEQSPPPY	12.7027	3043	PROC.NATL.ACAD.SCI. USA 2000 97 400-405
PAYVKQNTLKLAT	17.5452	24	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PEYVKQNTLKLAT	18.1583	13	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PHHTALRQAILCWGELMTLA	16.2806	85	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
PHYVKQNTLKLAT	17.3911	28	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKFVKQNTLKLAT	18.4207	10	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKIVKQNTLKLAT	16.2235	90	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKVVKQNTLKLAT	15.57	173	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170

PKYAKQNTLKLAT	16.5038	68	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYEKQNTLKLAT	15.7126	150	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYHKQNTLKLAT	16.8112	50	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYKKQNTLKLAT	15.9959	113	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYLKQNTLKLAT	16.6124	61	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYQKQNTLKLAT	17.1679	35	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVAQNTLKLAT	18.0842	14	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVEQNTLKLAT	16.8112	50	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVHQNTLKLAT	17.8329	18	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKANTLKLAT	16.8731	47	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKENTLKLAT	17.5044	25	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKHNTLKLAT	16.6124	61	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKKNTLKLAT	12.7169	3000	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKLNTLKLAT	16.4748	70	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKNNTLKLAT	15.895	125	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQATLKLAT	17.3911	28	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQETLKLAT	17.6787	21	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQHTLKLAT	17.0857	38	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQKTLKLAT	17.8901	17	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQLTLKLAT	16.0983	102	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNALKLAT	13.9209	900	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNSLKLAT	16.2235	90	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTAKLAT	16.5644	64	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTEKLAT	15.8405	132	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTHKLAT	15.57	173	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170

PKYVKQNTIKLAT	15.4003	205	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTKKLAT	13.0046	2250	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLALAT	17.0857	38	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLELAT	16.4606	71	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLHLAT	17.0097	41	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKAAT	18.9315	6	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKFAT	16.2235	90	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKHAT	17.1679	35	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKIAT	17.3221	30	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKKAT	14.3131	608	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLAA	18.9315	6	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLAH	17.0097	41	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLAK	16.3044	83	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLAS	17.2575	32	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
PKYVKQNTLKLAT	16.9166	45	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLATGMR	16.6628	58	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLAY	16.2574	87	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLFT	17.0097	41	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLHT	16.5961	62	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLKT	17.1969	34	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKLST	16.8521	48	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLKQAT	17.5452	24	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLLLAT	16.6979	56	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLRLAT	16.5801	63	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNTLSLAT	16.0504	107	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170

PKYVKQNTQKLAT	16.8112	50	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQNVLKLAT	16.2235	90	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQQTLKLAT	16.7159	55	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKQTTLKLAT	17.2575	32	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVKTNTLKLAT	16.0693	105	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVRQNTLKLAT	17.5044	25	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVSQNTLKLAT	17.6787	21	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PKYVVQNTLKLAT	16.18	94	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PLKAEIAQRLEDV	11.1075	15000	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
PRYVKQNTLKLAT	17.3911	28	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
PSYVKQNTLKLAT	17.3911	28	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170
QEIDPLSYNYIPVNSN	15.796	138	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
QGQMVHQAISPRTLN	16.6289	60	JOURNAL OF VIROLOGY 2001 75 4195-4207
QKQITKIQNFRVYYR	14.8455	357	JOURNAL OF VIROLOGY 2001 75 4195-4207
QYIKANSKFIGITE	11.1075	15000	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
RHNWVNHAVPLAMKLI	18.0842	14	JOURNAL OF IMMUNOLOGY 2000 165 1123-1137
RNGYRALMDKS…	12.6313	3268	PROC.NATL.ACAD.SCI. USA 2000 97 400-405
RVYQEPQVSPPQRAET	17.0597	39	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
SGNWQYFFPVIFSKASDSL	12.7443	2919	CLINICAL CANCER RESEARCH 2003 9 947-954
SPAIFQSSMTKILEP	12.5427	3571	JOURNAL OF VIROLOGY 2001 75 4195-4207
SSIIFGAFPSLHSGCC	15.9192	122	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
SSVFNVVNSSIGLIM	14.614	450	JOURNAL OF IMMUNOLOGY 2000 165 1123-1137
THHYFVDLIGGAMLSL	14.7145	407	JOURNAL OF IMMUNOLOGY 1998 160 3363-3373
TNNPPIPVGEIYKRWIILGL	15.2018	250	JOURNAL OF CLINICAL INVESTIGATION 2001
VDAQGTLSKIFKLGGRDSRS	11.6183	9000	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
WEFVNTPPLVKLWYQ	17.3911	28	JOURNAL OF VIROLOGY 2001 75 4195-4207
YFFPVIFSKASDSLQL	14.1594	709	CLINICAL CANCER RESEARCH 2003 9 947-954

YGSDTITLPCRIKQFINMWQE	13.3411	1607	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
YKLNFYFDLLRAKL	10.702	22500	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
YKTIAYDEEARR	10.702	22500	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
YLDNIKDNVGKMED	12.2061	5000	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
YTLLQAAPALDKLK...	13.3411	1607	JOURNAL OF IMMUNOLOGY 1992 149 2634-2640
YVKQNTLKLATGMR	16.2235	90	JOURNAL OF IMMUNOLOGY 1993 151 3163-3170

Appendix D HLA-DR4 Validation Set

EPITOPE	pIC50	IC50(nM)(f)	REFERENCE
CLRYTVDKSKPKA	16.118	100	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
CYKLEHPVTGCGER	13.479	1400	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
DKSKPKAYQWFDL	12.612	3333	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
DTKCYKLEHPVTGC	16.629	60	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
EGRCLRYTVDKSK	16.772	52	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
ELGRFKHTDACCR	15.272	233	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
ERTEGRCLRYTVD	11.806	7463	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
FKHTDACCRTHDM	12.255	4762	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
FNLIDTKCYKLEHP	11.251	13000	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
FVGKMYFNLIDTKC	15.088	280	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
GKMYFNLIDTKCYK	15.376	210	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
GPNELGRFKHTDA	12.971	2326	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
GRCLRYTVDKSKP	17.391	28	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
GRFKHTDACCRTH	14.915	333	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
IDTKCYKLEHPVTG	15.936	120	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
ISSYFVGKMYFNLI	12.206	5000	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
KCYKLEHPVTGCGE	14.378	570	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
KPKAYQWFDLRKY	12.842	2646	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
KSKPKAYQWFDLR	14.392	562	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
LGRFKHTDACCRT	15.519	182	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
LIDTKCYKLEHPVT	14.783	380	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
LRYTVDKSKPKAY	15.936	120	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
MYFNLIDTKCYKLE	10.724	22000	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
NELGRFKHTDACC	15.686	154	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
NLIDTKCYKLEHPV	13.633	1200	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
PKYVKQNTLKLAT	16.939	44	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
PNELGRFKHTDAC	15.125	270	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
RCLRYTVDKSKPK	16.06	106	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
RFKHTDACCRTHD	13.665	1163	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
RTEGRCLRYTVDK	13.218	1818	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
RYTVDKSKPKAYQ	13.122	2000	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
SGPNELGRFKHTD	12.794	2778	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
SKPKAYQWFDLRK	12.231	4878	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
SSGPNELGRFKHT	12.101	5556	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
SYFVGKMYFNLIDT	12.899	2500	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
TEGRCLRYTVDKS	16.986	42	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
TISSYFVGKMYFNL	11.513	10000	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
TKCYKLEHPVTGCG	15.088	280	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
VGKMYFNLIDTKCY	13.027	2200	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
YFNLIDTKCYKLEH	11.176	14000	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
YFVGKMYFNLIDTK	14.45	530	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184
YTVDKSKPKAYQW	11.626	8929	JOURNAL OF IMMUNOLOGY 2000 164 3177-3184

Appendix E HLA-DQ2 Antijen Training Set

EPITOPE	-lnIC50	IC50(nM)	REFERENCE
AIEQEGPEYW	14.894	340	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
EDIEIIPIGEE	13.228	1800	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
IEQEGPEYW	13.634	1199	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
PWIEQEGPEYW	15.636	162	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WAEQEGPEYW	12.800	2760	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIAQEGPEYW	14.757	390	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEAEGPEYW	15.376	210	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEKEGPEYW	13.576	1270	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQAGPEYW	15.202	250	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEAPEYW	15.782	140	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGAEYW	14.014	820	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGKEYW	8.839	145000	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPAYW	14.378	570	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPEAW	15.202	250	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPEKW	14.158	710	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPEYA	13.966	860	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPEYK	11.400	11200	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPEYW	15.202	250	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPEYW	15.125	270	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPEYW	15.050	291	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEGPKYW	11.651	8710	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQEKPEYW	14.592	460	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIEQKGPEYW	11.996	6170	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WIKQEGPEYW	14.014	820	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
WKEQEGPEYW	11.739	7980	EURO.JOURNAL OF IMMUNOLOGY 1996 26 2764-2772
AAAAAVAAEAY	16.475	70	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
ATSSNVMEERY	13.465	1420	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
EWTSSNVMEER	11.560	9540	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
TSSNVMEERY	12.592	3400	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WASSNVMEERY	13.642	1190	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTASNVMEERY	13.910	910	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSANVMEERY	14.489	510	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSAVMEERY	15.648	160	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNAMEERY	13.339	1610	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNAMEERY	11.455	10600	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNDMEERY	13.739	1080	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNEMEERY	13.739	1080	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNQMEERY	12.228	4890	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVAEERY	13.569	1280	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMAERY	14.064	780	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEARY	12.586	3420	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEARY	11.356	11700	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEDRY	13.458	1430	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEEAY	13.371	1560	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEER	10.134	39700	INTERNATIONAL IMMUNOLOGY 1996 8 177-182

WTSSNVMEERA	11.907	6740	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERA	11.534	9790	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERF	13.659	1170	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERK	8.963	128000	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERL	13.659	1170	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERM	13.545	1310	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERN	12.589	3410	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERR	13.315	1650	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERS	13.458	1430	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERY	15.202	250	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERY	14.077	770	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERY	13.932	890	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERY	13.776	1040	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERY	13.600	1240	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMEERY	13.576	1270	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNVMELRY	11.659	8640	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WTSSNWMEERY	9.460	77900	INTERNATIONAL IMMUNOLOGY 1996 8 177-182
WIEQEGPEYA	12.710	3020	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYD	12.373	4230	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYE	12.460	3880	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYF	14.231	660	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYK	10.561	25900	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYL	13.515	1350	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYN	12.251	4780	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYQ	12.272	4680	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYS	12.320	4460	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYV	13.291	1690	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
WIEQEGPEYW	11.801	7500	INTERNATIONAL IMMUNOLOGY 1998 10 1229-1236
PFPQPELPY	8.940	131000	JOURNAL OF EXPERIMENTAL MEDICINE 2000 191

Appendix F HLA-DQ2 Training Data from Vader et al

Epitope	Published IC50	-lnIC50	Transformed -lnIC50
PQPELPYPQ	15	18.015	6.543
PFPQPELPY*	8	18.644	8.926
FRPEQPYPQ	30	17.322	3.916
PYPEQPEQP	65	16.549	0.986**
PQQSFPEQE	14	18.084	6.805
IIQPEQPAQ	10	18.421	8.080
PFSEQEQPV	25	17.504	4.607
IEQEGPEYW*	1.6	20.253	15.026

*Denotes peptides used for the regression analysis

**Peptide score too low and may skew analysis

Appendix G HLA-DQ2 Validation Data

EPITOPE	-lnIC50	IC50(nM)	REFERENCE
GVAGLLVALAV	14.547	481	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
MSSGSFINISV	12.625	3289	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
YKTIAFDEEARR	13.137	1971	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
KPLLIAEDVEGEY	16.939	44	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
EEDIEIIPIQEEEY	16.089	103	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
LTEWTSSNVMEERY	15.163	260	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
IDVWLGGLAENFLPY	17.632	22	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
YPFIEQEGPEFFDQE	17.504	25	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
EPRAPWIEQEGPEYW	16.612	61	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
LCQVFADATPTGWGL	16.549	65	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
YQSYGPSGQYTHEFD	13.501	1369	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108
FVNQHLCGSHLVEAL	11.799	7511	JOURNAL OF IMMUNOLOGY 2002 169 5098-5108

Appendix H Experimentally determined T cell epitopes

Sequence	Reference
IIQPQQPAQ	Gastroenterology 2002 122 1729-37
PFRPQQPYPQPQPQ	Gastroenterology 2002 122 1729-37
PFSQQQQSPF	Gastroenterology 2002 122 1729-37
PFSQQQQQ	Gastroenterology 2002 122 1729-37
QSEQSQQPFQPQ	Gastroenterology 2002 122 1729-37
QIPQQPQQF	Gastroenterology 2002 122 1729-37
QLPQQPQQF	Gastroenterology 2002 122 1729-37
IIQPEQPAQ	Gastroenterology 2002 122 1729-37
PFRPEQPYPQPQPQ	Gastroenterology 2002 122 1729-37
PFSEQQESPF	Gastroenterology 2002 122 1729-37
PFSEEQEQ	Gastroenterology 2002 122 1729-37
QSEESQQPFQPQ	Gastroenterology 2002 122 1729-37
QIPEQPQQF	Gastroenterology 2002 122 1729-37
QLPEQPQQF	Gastroenterology 2002 122 1729-37
PFPQPQLPY	Gastroenterology 2002 123 803-809
PQPQLPYPQ	Gastroenterology 2002 123 803-809
PYPQPQLPY	Gastroenterology 2002 123 803-809
PQQSFPQQQ	Gastroenterology 2002 123 803-809
FPQQPQQPYPQQP	Gastroenterology 2002 123 803-809
FSQPQQQFPQPQ	Gastroenterology 2002 123 803-809
LQPQQPFPQQPQQPYPQQPQ	Gastroenterology 2002 123 803-809
WPQQQPFPQPQQPFCQQPQR	Gastroenterology 2002 123 803-809
QFPQTQQPQQPFPQPQQTFP	Gastroenterology 2002 123 803-809
QFPQTQQPQQPFPQPQQTFP	Gastroenterology 2002 123 803-809
PFPQPQQTFPQQPQLPFPQQ	Gastroenterology 2002 123 803-809
PQQPFPQPQQPQQPFPQSQQ	Gastroenterology 2002 123 803-809
PQQPFPQPQQQFPQPQQPQQ	Gastroenterology 2002 123 803-809
QFPQPQQPQQSFPQQQQPAI	Gastroenterology 2002 123 803-809
LRPLFQLAQGLGIIQPQQPA	Gastroenterology 2002 123 803-809
LGIIQPQQPAQLEGIRSLVL	Gastroenterology 2002 123 803-809
PFPQPELPY	Gastroenterology 2002 123 803-809
PQPELPYPQ	Gastroenterology 2002 123 803-809
PYPQPELPY	Gastroenterology 2002 123 803-809
PQQSPFEQE	Gastroenterology 2002 123 803-809
FPEQPEQPYPQQP	Gastroenterology 2002 123 803-809
FSQPEQEFPQPQ	Gastroenterology 2002 123 803-809
QPFPQPQLPYPQPQLPY	Gastroenterology 2003 125 1105-1103
QQFPQPQQPFPQQP	Gastroenterology 2003 125 1105-1103
QPFPQPQQPFPQSQ	Gastroenterology 2003 125 1105-1103
QLQPFPQPQLPYPQ	Gastroenterology 2003 125 1105-1103
PQQPFPQPQQPFRQ	Gastroenterology 2003 125 1105-1103
QYQPYPEQQEPFVQ	Gastroenterology 2003 125 1105-1103
QYQPYPEQQQPFVQ	Gastroenterology 2003 125 1105-1103
PQPFRPQQPYPQPQPQ	Gastroenterology 2003 125 1105-1103
PQQPQQSFPQQQRPF	Gastroenterology 2003 125 1105-1103
PQQPQQSFPQQPQR	Gastroenterology 2003 125 1105-1103

QQPFPQQPQQPFPQ	Gastroenterology 2003 125 1105-1103
QQPFVQQQQPFVQQ	Gastroenterology 2003 125 1105-1103
QPFPQPELPYPQPELPY	Gastroenterology 2003 125 1105-1103
QEFPQPEQPFPQQP	Gastroenterology 2003 125 1105-1103
QPFPQPEQPFPQSQ	Gastroenterology 2003 125 1105-1103
QLQPFPQPELPYPQ	Gastroenterology 2003 125 1105-1103
PQQPFPQPEQPFRQ	Gastroenterology 2003 125 1105-1103
PEQPFPQPEQPFPQ	Gastroenterology 2003 125 1105-1103
QYQPYPEQEEPFVQ	Gastroenterology 2003 125 1105-1103
QYQPYPEQEQPFVQ	Gastroenterology 2003 125 1105-1103
PQPFRPEQPYPQPQPQ	Gastroenterology 2003 125 1105-1103
PEQPQQSFPEQERPF	Gastroenterology 2003 125 1105-1103
PEQPEQSFPEQPQR	Gastroenterology 2003 125 1105-1103
QQPFPEQPEQPFPQ	Gastroenterology 2003 125 1105-1103
EQPFVPPPQPFVQQ	Gastroenterology 2003 125 1105-1103
SEQYQPYEQQEPFVQQQQ	PLoS Med 2004 1
YQPYPEQQEPFV	PLoS Med 2004 1
SEQYQPYEQEEPFVQQQQ	PLoS Med 2004 1
YQPYPEQEEPFV	PLoS Med 2004 1
IIQPQQPAQ	Immunogenetics 2005 57 8-15
PQQSFPQQQ	Immunogenetics 2005 57 8-15
PQPQLPYPQ	Immunogenetics 2005 57 8-15
PYPQPQLPY	Immunogenetics 2005 57 8-15
PFSQQQQSPF	Immunogenetics 2005 57 8-15
PFPQQPQQPF	Immunogenetics 2005 57 8-15
IIQPEQPAQ	Immunogenetics 2005 57 8-15
PQQSFPEQQ	Immunogenetics 2005 57 8-15
PQPELPYPQ	Immunogenetics 2005 57 8-15
PYPQPELPY	Immunogenetics 2005 57 8-15
PFSEEQESPF	Immunogenetics 2005 57 8-15
PFPEQPEQPF	Immunogenetics 2005 57 8-15
YLQLQPFPQPQLPYP	J Clin Invest 2006 116 2226-36
PFPQPQLPYPQPQLP	J Clin Invest 2006 116 2226-36
PFPQPQLPYPQPQLP	J Clin Invest 2006 116 2226-36
PFPQPQLPYPQPQLP	J Clin Invest 2006 116 2226-36
PFPQPQLPYPQPQLP	J Clin Invest 2006 116 2226-36
YPSGQGSFQPSQQNP	J Clin Invest 2006 116 2226-36
GSFQPSQQNPQAQGS	J Clin Invest 2006 116 2226-36
QAQGSVQPQQLPQFE	J Clin Invest 2006 116 2226-36
WPQQQPFPQPQQPFCQQPQR	J Clin Invest 2006 116 2226-36
WPQQQPFPQPQQPFCQQPQR	J Clin Invest 2006 116 2226-36
QQPFCQQPQRTIPQPHQTFH	J Clin Invest 2006 116 2226-36
HQPQQTFPQPQQTYPHQPQQ	J Clin Invest 2006 116 2226-36
QQTYPHQPQQQFPQTQQPQQ	J Clin Invest 2006 116 2226-36
QFPQTQQPQQPFPQPQQTFP	J Clin Invest 2006 116 2226-36
PFPQPQQTFPQQPQLPFPQQ	J Clin Invest 2006 116 2226-36
QQPQLPFPQQPQQPFPQPQQ	J Clin Invest 2006 116 2226-36
QQPQLPFPQQPQQPFPQPQQ	J Clin Invest 2006 116 2226-36
PQQPFPQSQQPQQPFPQPQQ	J Clin Invest 2006 116 2226-36
PQQPFPQSQQPQQPFPQPQQ	J Clin Invest 2006 116 2226-36
QFPQPQQPQQSFPQQQQPAI	J Clin Invest 2006 116 2226-36
QFPQPQQPQQSFPQQQQPAI	J Clin Invest 2006 116 2226-36

SFPQQQQPAIQSFLQQQMNP	J Clin Invest 2006 116 2226-36
IHSVAHSIIMQQEQQQGVPI	J Clin Invest 2006 116 2226-36
IHSVAHSIIMQQEQQQGVPI	J Clin Invest 2006 116 2226-36

HLA-DQ2 binding model PSSMs

	1	2	3	4	5	6	7	8	9
A	0.18	1.08	1.53	0.35	1.09	0.42	0.28	0.57	0.06
R	0.00	-0.89	0.00	0.00	0.00	0.00	0.00	0.57	0.45
N	0.00	0.00	-0.26	0.00	0.00	0.00	0.00	0.00	-0.10
D	0.00	0.00	0.00	0.53	0.00	0.00	0.50	0.00	-1.06
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	0.00	-0.48	-0.39	0.28	-1.50	-0.44	-0.61	-0.79	-0.92
E	0.00	0.41	0.00	0.76	-0.44	0.28	1.23	0.00	-0.81
G	0.00	0.00	0.00	0.00	-0.21	0.00	0.00	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	1.65	-0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00	0.31	0.00	0.12	0.00	0.84
K	-0.61	0.31	-0.12	-0.83	0.50	-2.09	-0.51	0.34	-2.49
M	0.00	0.00	0.00	0.00	0.60	0.00	0.00	0.00	0.55
F	-0.89	-0.18	0.00	0.00	-0.79	0.00	0.00	0.00	1.12
P	-0.65	0.00	-0.14	-0.44	0.43	1.84	-0.44	-0.75	0.00
S	0.34	0.19	-0.61	-0.79	0.00	0.00	0.00	0.00	0.27
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
W	0.00	0.00	0.00	-1.37	0.00	0.00	0.00	0.00	1.02
Y	0.00	0.00	0.00	0.00	0.00	0.00	-0.58	0.07	1.32
V	0.00	0.00	0.00	1.51	0.00	0.00	0.00	0.00	-0.24

PSSM derived using AntiJen+Vader dataset without missing data correction.

	1	2	3	4	5	6	7	8	9
A	0.18	1.08	1.53	0.35	1.09	0.42	0.28	0.57	0.06
R	-0.61	-0.89	-0.12	-0.83	0.50	-2.09	-0.51	0.57	0.45
N	0.00	-0.48	-0.26	0.28	-1.50	-0.44	-0.61	-0.79	-0.10
D	0.00	0.41	0.00	0.53	-0.44	0.28	0.50	0.00	-1.06
C	0.34	0.19	-0.61	-0.79	0.00	0.00	0.00	0.00	0.27
Q	0.00	-0.48	-0.39	0.28	-1.50	-0.44	-0.61	-0.79	-0.92
E	0.00	0.41	0.00	0.76	-0.44	0.28	1.23	0.00	-0.81
G	0.00	0.00	0.00	0.00	-0.21	0.00	0.00	0.00	0.00
H	-0.61	-0.29	-0.12	-0.83	0.50	-2.09	-0.51	0.46	-1.02
I	1.65	-0.44	0.00	0.00	0.46	0.00	-0.23	0.07	0.90
L	1.65	-0.44	0.00	0.00	0.31	0.00	0.12	0.07	0.84
K	-0.61	0.31	-0.12	-0.83	0.50	-2.09	-0.51	0.34	-2.49
M	1.65	-0.44	0.00	0.00	0.60	0.00	-0.23	0.07	0.55
F	-0.89	-0.18	0.00	0.07	-0.79	0.00	0.00	0.00	1.12
P	-0.65	0.00	-0.14	-0.44	0.43	1.84	-0.44	-0.75	0.00
S	0.34	0.19	-0.61	-0.79	0.00	0.00	0.00	0.00	0.27
T	0.34	0.19	-0.61	-0.79	0.00	0.00	0.00	0.00	0.27
W	-0.89	-0.18	0.00	-1.37	-0.79	0.00	0.00	0.00	1.02
Y	1.65	-0.44	0.00	0.00	0.46	0.00	-0.58	0.07	1.32
V	-0.89	-0.18	0.00	1.51	-0.79	0.00	0.00	0.00	-0.24

PSSM derived using AntiJen+Vader dataset with missing data correction.

Appendix I Materials and Suppliers

Description	Item Number	Supplier
DUCAF HLA-Defined Cell line	88052019	ECACC
VAVY HLA-Defined Cell line	88052023	ECACC
Complete protease inhibitor mix	187 35 80	Roche
NP-40 Substitute	74385-1L	Sigma-Aldrich
Penicillin Streptomycin	15240-062	Gibco
Phosphate Buffered Saline Tablets	BR0014G	Oxoid
40% Acrylamide	EC-852	National Diagnostics, Hull, England
TEMED	T-1835	Sigma-Aldrich
UltraPure 10X TRIS/Glycine	EC-830	National Diagnostics, Hull, England
Dried Skimmed Milk		Marvel (UK)
Tween 20	536980 246	Merck-Schuchdart, Germany
APS (Ammonium Peroxodisulphate)	10032	BDH Laboratory Supplies, Poole, England
Trizma Base	T1503-1KG	Sigma-Aldrich
TMB Liquid substrate system	T8665-1L	Sigma-Aldrich
Bovine Albumin Fraction V	A-7906	Sigma-Aldrich
Immobilised Protein A	20333	Pierce
Octyl-B-D-glucopyranoside	O8001	Sigma-Aldrich
Diethylamine	D0806	Sigma

Appendix J Buffers and Solutions

EB/AO reagent

The Ethidium Bromide/ Aciduric Orange (EB/AO) reagent was made up by adding 10 ml of stock Aciduric Orange (0.1%) to 4ml of stock Ethidium Bromide (0.4%). This mixture was then made up to 1 litre using 0.9% sodium chloride (NaCl) solution (Baxter). The container in which it was stored was wrapped in aluminium foil and stored at 4°C.

10% Heat inactivated foetal calf serum

Foetal Calf Serum (FCS) was inactivated through heating to 56°C for a total of 30 minutes. The heat inactivated FCS was then placed in sterile aliquots and frozen until use at -20°C. Prior to use the required amount of frozen aliquots were allowed thaw and were then made up to ten times their original volume with PBS to yield a 10% foetal calf serum solution.

Phosphate buffered saline (PBS)

Phosphate buffered saline (PBS) was prepared by addition of a single PBS tablet (Oxoid) per 100ml of distilled water. Tablets were allowed to dissolve in the solution at room temperature.

Coating buffer.

Coating buffer was prepared by making 8.4g $NaHCO_3$ and 3.56g Na_2CO_3 up to 1 litre in distilled water. The pH was adjusted to 9.5 using HCL and/or NaOH where necessary. The coating buffer was stored at 4°C for a maximum of 10 days.

Wash buffer

Wash buffer was made by formulating a 0.05% Tween-20 solution in PBS, this solution was used within the same day.

Binding Buffer

For each well the binding buffer was made to contain 25µL of 4X citrate phosphate solution, 14µL dH_2O, 9µL of complete inhibitor mix (2ml dH_2O plus one complete inhibitor tablet (Roche)), 1µL of 5% Tween20 and 1µL of 5% NP-40.

4X Citrate Phosphate Solution

150mM Na_2HPO_4 adjusted to pH5.4 with citrate