Bioinformatic Study of Antigen Presentation by HLA class II

by
Pau Marc Muñoz Torres

Thesis submitted to
Universitat Autònoma of Barcelona
in partial fulifillment of the requeriments
for the degree of
Doctor of Philosophy

Director Dr Xavier Daura Director Dr Juan A Cedano

Tesis Doctoral UAB/ANY 2013
PhD Program - Protein Structure and Function
Institut de Biotecnologia i de Biomedicina

Als meus pares i germanes

Voler l'impossible ens cal,
i no que mori el desig.

Marià Villangomez.

Eivissa 1913-2002
L´any del seu centenari.

Agraïments

Ithaki (Ιθάκη) - Κωνσταντῖνος Καβάφης

Σὰ βγεῖς στὸν πηγαιμὸ γιὰ τὴν Ἰθάκη,
νὰ εὔχεσαι νά ᾽ναι μακρὺς ὁ δρόμος,
γεμάτος περιπέτειες, γεμάτος γνώσεις.

Τοὺς Λαιστρυγόνας καὶ τοὺς Κύκλωπας,
τὸν θυμωμένο Ποσειδῶνα μὴ φοβᾶσαι,
τέτοια στὸν δρόμο σου ποτέ σου δὲν θὰ βρεῖς,
ἂν μέν᾽ ἡ σκέψις σου ὑψηλή, ἂν ἐκλεκτὴ
συγκίνησις τὸ πνεῦμα καὶ τὸ σῶμα σου ἀγγίζει.

Τοὺς Λαιστρυγόνας καὶ τοὺς Κύκλωπας,
τὸν ἄγριο Ποσειδῶνα δὲν θὰ συναντήσεις,
ἂν δὲν τοὺς κουβανεῖς μὲς στὴν ψυχή σου,
ἂν ἡ ψυχή σου δὲν τοὺς στήνει ἐμπρός σου.

Νὰ εὔχεσαι νά ᾽ναι μακρὺς ὁ δρόμος.
Πολλὰ τὰ καλοκαιρινὰ πρωινὰ νὰ εἶναι
ποῦ μὲ τί εὐχαρίστηση, μὲ τί χαρὰ
θὰ μπαίνεις σὲ λιμένας πρωτοειδωμένους.

Νὰ σταματήσεις σ᾽ ἐμπορεῖα Φοινικικά,
καὶ τὲς καλὲς πραγμάτειες ν᾽ ἀποκτήσεις,
σεντέφια καὶ κοράλλια, κεχριμπάρια κ᾽ ἔβενους,
καὶ ἡδονικὰ μυρωδικὰ κάθε λογῆς,
ὅσο μπορεῖς πιὸ ἄφθονα ἡδονικὰ μυρωδικά.

Σὲ πόλεις Αἰγυπτιακὲς πολλὲς νὰ πᾷς,
νὰ μάθεις καὶ νὰ μάθεις ἀπ᾽ τοὺς σπουδασμένους.
Πάντα στὸ νοῦ σου νά ᾽χεις τὴν Ἰθάκη.
Τὸ φθάσιμον ἐκεῖ εἶν᾽ ὁ προορισμός σου.

Ἀλλὰ μὴ βιάζεις τὸ ταξίδι διόλου.
Καλλίτερα χρόνια πολλὰ νὰ διαρκέσει.
Καὶ γέρος πιὰ ν᾽ ἀράξεις στὸ νησί,
πλούσιος μὲ ὅσα κέρδισες στὸν δρόμο,
μὴ προσδοκώντας πλούτη νὰ σὲ δώσει ἡ Ἰθάκη.

Ἡ Ἰθάκη σ᾽ ἔδωσε τ᾽ ὡραῖο ταξίδι.
Χωρὶς αὐτὴν δὲν θά ᾽βγαινες στὸν δρόμο.
Ἄλλα δὲν ἔχει νὰ σὲ δώσει πιά.

Κι ἂν πτωχικὴ τὴν βρεῖς, ἡ Ἰθάκη δὲν σὲ γέλασε.
Ἔτσι σοφὸς ποὺ ἔγινες, μὲ τόση πεῖρα,
ἤδη θὰ τὸ κατάλαβες οἱ Ἰθάκες τὶ σημαίνουν.

Viatge a Ítaca - Constantino Kavafis (Pau Riba)

Quan surts per fer el viatge cap a Ítaca,
has de pregar que el camí sigui llarg,
ple d'aventures, ple de coneixences.
Els Lestrígons i els Cíclops,
l'aïrat Posidó, no te n'esfereeixis:
són coses que en el teu camí no trobaràs,
no, mai, si el pensament se't manté alt, si una
emoció escollida
et toca l'esperit i el cos alhora.

Els Lestrígons i els Cíclops,
el feroç Posidó, mai no serà que els topis
si no els portes amb tu dins la teva ànima,
si no és la teva ànima que els dreça davant teu.

Has de pregar que el camí sigui llarg.
Que siguin moltes les matinades d'estiu
que, amb quina delectança, amb quina joia!
entraràs en un port que els teus ulls ignoraven;
que et puguis aturar en mercats fenicis
i comprar-hi les bones coses que s'hi exhibeixen,
corals i nacres, mabres i banussos
i delicats perfums de tota mena:
tanta abundor com puguis de perfums delicats;
que vagis a ciutats d'Egipte, a moltes,
per aprendre i aprendre dels que saben.

Sempre tingues al cor la idea d'Ítaca.
Has d'arribar-hi, és el teu destí.
Però no forcis gens la travessia.
És preferible que duri molts anys
i que ja siguis vell quan fondegis a l'illa,
ric de tot el que hauràs guanyat fent el camí,
sense esperar que t'hagi de dar riqueses Ítaca.

Ítaca t'ha donat el bell viatge.
Sense ella no hauries pas sortit cap a fer-lo.
Res més no té que et pugui ja donar.
I si la trobes pobra, no és que Ítaca t'hagi enganyat.
Savi com bé t'has fet, amb tanta experiència,
ja hauràs pogut comprendre què volen dir les Ítaques.

El viatge

Ara fa aproximadament uns vuit anys que un dia assolellat vaig trepitjar el IBB per primer cop: vaig entrar-hi i vaig preguntar pel Doctor Daura, anava vestit més o menys de forma elegant perquè volia causar una bona impressió. A la secretaria em varen indicar el camí, havia de baixar les escales i tombar a mà dreta, dos cops, era la primera porta del passadís.

Vaig baixar les escales i em vaig trobar un home tot alt davant d'un laboratori fosc en el que semblava que no hi havia ningú. Si miraves cap a dalt, a les finestres que tocaven al sostre, només hi veies ampolles buides. Com qui no vol la cosa li vaig tornar a fer la pregunta, "Perdoni, on és el despatx del Doctor Daura?" i la resposta va ser "Qui el busca?". Un cop em vaig haver presentat varem passar cap al seu despatx. Aquell dia sols varen sortir de la seva boca tres frases, "Què vols?", a la qual li vaig respondre amb un rotllo impressionant, "Aquí no fem bioinformàtica, fem biologia computacional", però sincerament, aquest "petit detall", a mi tampoc m'importava gaire, el que volia era no tornar a trepitjar un laboratori "humit", per al final deixar anar un "tu m'interesses". Poc temps després, vaig descobrir que al laboratori fosc hi havia vida.

Recordo el primer dia de feina, o almenys el propòsit que em vaig fer: el meus companys m'havien de veure com una persona seriosa, no s'havia de notar que m'agradava la festa i les dones com a tot bon eivissenc. Però els gens em trairen. Al cap de poc d'haver començat, quan arribava al laboratori, en sortia una rossa i se'm va escapar un "guau!", crec que no va passar ni un segon perquè de fons se sentís "GERUPPA!": ja tenia mal nom! I es va "oficialitzar" quan es va convertir en el meu usuari en el sistema.

Aquest va ser el començament de la meva aventura en aquestes terres. Una aventura que m'ha portat a ports que desconeixia per a conèixer-vos a tots vosaltres. Hem anat a fer tapes, celebrat fondues, treballat, rigut i plorat, anat de

festa (i quines festes), après coses que a vegades tenien a veure amb les ciències de la vida i altres amb la ciència de viure, coses que han fet que avui jo sigui qui sóc.

A tots vosaltres, Gràcies.

Xavi, Juan, Dolores i Mercè; gràcies pel temps i la paciència que m'heu dedicat (reconec que n'heu tingut molta amb mi) i, sobre tot, gràcies per la oportunitat que m'heu donat d'aprendre de vosaltres, és impagable.

I ara toca anomenar a sa meva família, gràcies per ser sempre a es meu costat, sense valtros no hagués fet res del que he fet.

Ara, si em perdoneu, haig de continuar el meu viatge, Ítaca m'espera, ens veiem en arribar-hi.

Pau, octubre del 2013

Contents

Summary **15**

Objectives **17**

Chapter 1. Introduction **19**

Major histocompatibility complex and the immune response 21

Immune response 21

Activation of B and T cells 21

CD8+ and CD4+ T cells 22

T-cell receptors 23

T cell recognition of antigen 23

The major histocompatibility complex 24

The MHC class I and class II functional protein complexes 25

Antigen presentation pathways 26

Cytosolic processing and MHC class I presentation 26

Endosomal processing and MHC class II presentation 27

HLA class II polymorphism 29

Peptide binding to MHC class II molecules 31

HLA class II peptide databases 32

Prediction of HLA class II epitopes 33

References 40

Chapter 2. HLA2db: a suite for the prediction of HLA class II epitopes based on non-assisted self-learning procedures **49**

Abstract 51

Introduction 52

Materials and methods 55

Data collection 55

Identification of epitope core sequences from multiple-match 9-residue segments 59
Definition of an initial binding profile 60
PSSM focusing and refinement 62
Pocket inheritance 65
Epitope prediction 66
Maintenance, growth and stability of the database 67
Results and discussion 68
Query sequences 68
Available HLA class II molecules 68
Scanning of proteomes 72
Integrating user data to generate a new profile or improve an existing one 72
Immunome calculation 73
HLA2db in the context of other predictors 74
References 74

Chapter 3. A Support Vector Machine for the prediction of MHC class II epitopes based on amino-acid distances 81

Abstract 83
Introduction 83
Implementation 84
Data collection 84
SVM generation 85
Conclusion 89
References 89

Chapter 4. Discussion 93

Coherence of the original epitope dataset 95
Protein-data storage and retrieval 98
Data codification and organisation 99
Codification of bacterial and human proteomes 99
Database schema 100
Binding-motif calculation procedure 102
Peptide binding core 102
Computation phases 103
Dataset enrichment with Blosum tables 103

Predictor training....104
References....111

Conclusions....113

Summary

Understanding how peptides are selectively bound and presented by major histocompatibility complex class II molecules (MHC class II or HLA class II in humans) is of outmost importance for its broad implications in human health, from infection to autoimmunity or cancer. The aim of this thesis was to develop a computational strategy to identify HLA class II binding patterns for a variety of alleles and use this knowledge to predict their capacity to bind specific peptide sequences. To make an effective use of the prediction algorithm, a web-based platform for the analysis of large peptide or protein sets, including various functionalities, was also devised.

In order to accomplish these objectives, the work was divided into three different stages. The first stage consisted in the construction of a *postgresql* relational database to store all the information required for and generated by the algorithms developed. The required, uploaded information (subject to updates) consisted of known HLA class II epitopes and the translated genomes of a list of pathogenic bacterial species and human. In addition, the database was designed to include a private section for the upload of user-owned epitope information, which the owner may use in combination with the public data. In a second stage two predictors were developed, one using position-specific scoring matrices (PSSMs) and the other one using a support vector machine (SVM). PSSM development was performed using an iterative optimisation protocol, starting from the alignment of known epitopes to identify HLA class II binding cores (9-residue segments) and incorporating additional information such as allele conservation and non-binders at different phases of the refinement. For SVM construction, the epitope core was defined using the corresponding PSSM and the parameters for the SVM with a radial-basis-function (RBF) kernel were set up individually for each molecule to

get the best performance. In the third stage, two web pages were constructed, one for each predictor. The servers share a common part in which the user can introduce peptide or protein sequences in Fasta format to perform an analysis that delivers both putative epitopes and their localization in a selected proteome. In addition, the PSSM-based server allows the user to upload his/her own sequences to elucidate new HLA class II binding patterns and perform predictions with them.

Objectives

The primary objective of this thesis has been to create a method and software to predict peptide binding to HLA class II molecules. The main characteristics of the resulting web-based platform should be:

- Ease of use.
- Capacity to perform unsupervised analyses.
- Capacity to auto-update the background data.
- Capacity to automatically discard ambiguous data.
- Enable users to:
  - Analyse their own data sets, in a private manner if wanted.
  - Derive patterns using a combination of private and public data, keeping resulting patterns private if wanted.
  - Use private and/or public patterns to perform predictions.
- Identify potential epitopes for selected HLA class II molecules in user-provided lists of peptides or proteins and localise the same binding motifs in a selected proteome (translated genome of pathogenic bacterial species or human).
- Identify potential epitopes for selected HLA class II molecules in full proteomes.

Chapter 1. Introduction

Major histocompatibility complex and the immune response

This section introduces the subject of this thesis, the MHC class II proteins and their role in immunity. It is general knowledge immunology and should serve the non-expert reader to situate these proteins in their functional context and visualise the domain of application of the computational methods described in Chapters 2 and 3, core of this thesis. The text book Janeway's Immunobiology [1] has been used as main reference.

Immune response

An immune response is a body's integrated response against antigens (foreign material). The fight against an infection by a pathogen may involve a body's response at two complementary levels: the innate immune response, which acts at the first stages of the infectious process by restraining the penetration of the pathogenic agent into the organism acting against conserved molecular patterns in a non-specific and immediate manner, and the adaptive immune response, developed as an adaptation to infection with the pathogen and providing a specific response against it. A fundamental characteristic of the adaptive immune system is that it can generate immunological memory, so that a fast and stronger response is produced in subsequent infections by the same agent, i.e. providing protective immunity against it. The cells in charge for the adaptive immune response are the antigen-specific lymphocytes, i.e. B lymphocytes (B cells) and T lymphocytes (T cells).

Activation of B and T cells

Lymphocytes require activation to perform their protective function. Lymphocytes that have not yet been activated by antigen are known as naive lymphocytes, while

those that have met the antigen and differentiated into fully functional lymphocytes are known as effector lymphocytes. In addition to the binding of antigen to lymphocyte receptors, lymphocyte responses require a second signal that comes from cells carrying co-stimulatory molecules on their surface. In the case of B cells, T cells perform this co-stimulation. After binding of an antigen to a B-cell receptor (BCR) on the cell surface and appropriate co-stimulation by T cells, the lymphocyte proliferates and differenciates into plasma cells, the effector form of B lymphocytes. They produce antibodies, a secreted form of the BCR specific for the same antigen. The BCR and antibody protein complexes are generically known as immunoglobulins. On the other hand, activation of naive T cells is dependent on specialised antigen-presenting cells (APCs) such as dendritic cells (most important in this respect), macrophages and B cells. Stimulation by one of these cell types together with binding of antigen to a T-cell receptor (TCR) may result in proliferation and differentiation into one of several different types of effector T lymphocytes: cytotoxic T cells, which kill cells that are infected with viruses or other intracellular pathogens, helper T cells, which provide the essential additional signals that activate antigen-stimulated B cells to differentiate and produce antibody, and regulatory T cells, which supress the activity of other lymphocytes and help control immune responses. During an immune response, some of the activated B and T cells differenciate into memory cells.

CD8+ and CD4+ T cells

T lymphocytes are responsible for the so-called cell-mediated immune response of adaptive immunity, which is needed to control infection by intracellular pathogens. T lymphocytes are composed of two main classes, distinguished by carrying either the CD8 or CD4 protein on their surface. These proteins help determining the way T cells interact with other cells. Cytotoxic T cells carry CD8 (CD8+ T cells), while T cells dedicated to the activation of other cells, as opposed to killing them, carry CD4 (CD4+ T cells). The two major types of effector CD4+ T cells are called T_H1 and T_H2. They are both involved in combating bacterial infections, albeit in different ways. T_H1 cells activate infected macrophages so that they perform their

bactericidal activity, but may also adopt a helper-T-cell role activating B cells to produce antibody. T_H2 cells, on the other hand, are entirely dedicated to the latter function.

T-cell receptors

T cells display thousands of identical antigen receptors on their surface, each one consisting of two different polypeptide chains, α and β, linked by a disulfide bond. The α:β heterodimers are very similar in structure to the Fab fragment of an immunoglobulin. Both chains of the TCR have an amino-terminal variable region (V) with homology to an immunoglobulin V domain, a constant region (C) with homology to an immunoglobulin C domain, a short segment connecting the C domain to the membrane, similar to an immunoglobulin hinge region and containing the cysteine residue that forms the inter-chain disulfide bond, a transmembrane domain, and a short cytoplasmic tail. Both chains have carbohydrate side chains attached to the C and V domains.

T cell recognition of antigen

In contrast to antigen recognition by B cells, which involves direct binding of BCRs and antibodies to free, intact antigen, T cells recognise short protein fragments resulting from antigen processing. The recognition by TCRs of peptides derived from antigen requires that they are presented to the receptor by membrane glycoproteins of the major histocompatibility complex (MHC) on the surface of cells. The TCR interacts with the complex by making contacts with both the MHC molecule and the antigen peptide. This introduces and extra dimension to antigen recognition by T cells, known as MHC restriction, where recognition depends on the right combination of peptide and MHC molecule. MHC restriction is one of the basis for the phenomenon of alloreactivity, by which T cells respond to non-self or allogenic MHC molecules, e.g. from transplants. There are two main types of MHC molecules, called MHC class I and MHC class II, distinguished by their genome localisation, structure, and target peptides. During antigen recognition, CD8 or CD4

molecules (depending on the type of T cell) associate with the TCR on the T-cell surface and bind to invariant sites on the immunoglobulin-like domains of the MHC molecule, away from the peptide-binding site. They act as co-receptors ensuring that the MHC:ligand complex is recognised by the right T-cell, i.e. a CD8+ T cell if the peptide is presented by an MHC class I molecule and a CD4+ T cell if it is presented by a MHC class II molecule.

The major histocompatibility complex

The first MHC gene products were discovered on the surface of white blood cells, becoming known as leukocyte antigens. This is why the human MHC is also referred to as the human leukocyte antigen (HLA) complex. They were originally studied for their ability to confer tolerance (histocompatibility) following tissue grafts or, later, organ transplants, but their primary function is to provide protection against pathogens. The MHC is a complex of polymorphic and co-dominant genes. This means that most people are heterozygous for the MHC and express two different forms of each molecule. In its extended definition, the MHC is encoded along 7.6 Mb of the short arm of chromosome 6 in humans and contains 421 loci [2,3]. The genes are organized in clusters, facilitating the co-expression of those proteins that are physically or functionally associated. In particular, the classical MHC class I and MHC class II genes are found in clusters localised toward the telomeric and centromeric regions of the extended MHC, respectively [3]:

- **HLA class I supercluster:** comprises the classical class I genes (*HLA-A*, *-B* and *-C*). They are expressed in almost all cells and provide a mechanism to display fragments of foreign proteins synthesized in the cytosol to CD8+ T cells, enabling detection and clearing, for example, of cells expressing viral or tumour proteins. This cluster also contains so-called non-classical class I genes (*HLA-E*, *-F*, *-G*, *HFE* and 12 pseudogenes) and class I-like genes (*MICA*, *MICB*, and 5 pseudogenes).

- **HLA class II cluster:** comprises the classical class II genes (*HLA-DP*, *-DQ*, *-DR* and pseudogenes). They are expressed by APCs and present fragments of

proteins originating in intracellular vesicles to CD4+ T cells, enabling, for example, the activation of macrophages infected by bacteria or the activation of B cells after internalization of free antigen. This cluster also contains so-called non-classical class II genes (*HLA-DM* and *-DO*).

The combination of MHC genes presented by a single chromosome is known as the MHC haplotype, and each haplotype confers different characteristics of peptide recognition and presentation.

The MHC class I and class II functional protein complexes

The MHC class I molecule consists of an α chain encoded in the MHC and non-covalently associated with a smaller chain, β_2-microglobulin, which is not polymorphic and is encoded in a different chromosome. The α chain has 3 domains and is the only one to span the membrane. The membrane-bound α_3 domain and the β_2-microglobulin closely resemble immunoglobulin domains. The α_1 and α_2 domains form the peptide-binding groove, which concentrates much of the polymorphism of these molecules.

The MHC class II molecule consists of a non-covalent complex of two chains, α and β, each organised in two domains, α_1 and α_2 and β_1 and β_2. Both chains are encoded within the MHC, span the membrane by C-terminal sequences consecutive to the immunoglobulin-like α_2 and β_2 domains and end with a short cytoplasmic tail. The peptide-binding cleft is formed by two domains from different chains, α1 and β1. As in the MHC class I molecule, the binding site is delimited by a β-sheet (floor) and two α-helices (walls), but in this case the cleft is open at the ends, therefore allowing the termini of the peptide to extend beyond the binding groove. Thus, while MHC class I molecules bind peptides of 8-13 amino-acid residues, peptides that bind to MHC class II molecules are not constrained in length and may be as long as 30 residues.

Antigen presentation pathways

Inside a host cell, an infectious organism can replicate either in the cytosol, which communicates with the nucleus, or in the vesicular system, which communicates with the extracellular system. The immune system has developed different paths for the detection and elimination of pathogens from the cytosol and the vesicular system, mediated by MHC class I and CD8+ T cells in the first case and by MHC class II and CD4+ T cells in the second [4,5], although there is significant cross talk between these two paths. Virus and certain bacteria replicate in the cytosol or in the nuclear compartment, whereas many other bacteria and some parasites replicate in the endosomes and lysosomes constituting the vesicular system. Exogenous antigens from extracellular pathogens or other pathogen-infected cells can also enter the vesicular system (by phagocytosis, receptor-mediated endocytosis or pinocytosis) or the cytosol (by active translocation after phagocytosis or receptor-mediated endocytosis) of specialised APCs.

Cytosolic processing and MHC class I presentation

Protein degradation in the cytosol is performed mainly by a large multi-catalytic protease complex called proteasome. The proteasome is part of the ubiquitin-dependent degradation pathway for cytosolic proteins. The implication of the proteasome in the production of peptide ligands for MHC class I molecules has been demonstrated by various means [6], although it is not clear whether the proteasome is the only cytosolic protease capable of generating peptides for transport into the endoplasmic reticulum (ER), where they will meet MHC class I molecules. Chaperones protect these peptides from complete degradation in the cytoplasm before translocation to the ER. However, many of these peptides are too long to readily bind to MHC class I molecules. Thus, while the C-terminal ends are generally produced by cleavage in the proteasome, the N-terminal ends may be shortened in the ER by an aminopeptidase called ERAAP (endoplasmic reticulum aminopeptidase associated with antigen processing).

The folding and assembly of MHC class I molecules takes place in the lumen of the endoplasmic reticulum (ER), which they do not leave unless they bind their antigen peptide. Newly synthesized MHC class I α chains that enter the ER bind to the membrane-bound chaperon calnexin, which retains the polypeptide in a partly folded state. When the β chain binds this complex, the partially folded dimer dissociates from calnexin and binds to a complex of proteins known as the MHC class I loading complex. One component of this complex, calreticulin is also a chaperon. A second component is the TAP associated protein tapasin (encoded by a gen within the MHC), which forms a bridge between the MHC molecule and TAP (transporter associated with antigen processing), an ATP-dependent peptide transporter of the ABC family that transports peptides (with some level of specificity) from the cytosol to the ER and which is also encoded in the MHC. A third component is Erp57, a thiol oxidoreductase that may have a role in breaking and reforming the disulfide bond in the α_2 domain during peptide loading. Calnexin, calreticulin and Erp57 are part of the cell's general protein-quality control system. The loading complex seems to be essential both to maintain the MHC class I molecule in a state that can bind a peptide and to carry out a so-called peptide-editing function, which consists in the exchange of low-affinity peptides for higher affinity ones. The binding of a peptide to the heterodimer finally releases it from the loading complex so that the MHC class I molecule and its bound peptide can be exported to the cell surface by vesicular transport. Most of the peptides transported by TAP will not bind the MHC molecules in that cell and will then be transported back to the cytosol by a different ATP-dependent transport mechanism.

Endosomal processing and MHC class II presentation

Bacteria and parasites that replicate inside intracellular vesicles in macrophages and extracellular pathogens and proteins that are internalised into endocytic vesicles (e.g. BCR-mediated endocytosis of antigens by B cells), are reduced and degraded by proteases within the vesicles. The material that enters the cells through endocytosis is contained in endosomes, which become increasingly acidic as they move to the interior of the cell, eventually fusing with lysosomes. The

endosomes and lysosomes contain proteases that are activated at low pH, such as the cathepsins (cysteine proteases) B, D, S and L, the last two of which play a predominant role in antigen processing. Reduction of disulfide bonds to enable digestion is carried out primarily by IFN-γ-induced lysosomal thiol reductase (GILT). Autophagy, by which cytosolic proteins and organelles are delivered to lysosomes for degradation within the normal process of protein turnover, provides one of the sources of cross-talk with the MHC class I pathway.

The biosynthetic pathway of MHC class II molecules, like that of MHC class I, starts with translocation of the nascent chains to the ER, and they must therefore be prevented from binding to peptides transported into the ER lumen from the cytosol or to polypeptides being synthesized by the cell. This is accomplished by binding to a protein known as the MHC class II-associated invariant chain (Ii), which forms trimers. Each Ii subunit binds to an MHC class II α:β heterodimer blocking its groove. Assembly of this nine-chain complex requires calnexin. When the assembly is completed, the complex is released from calnexin and transported out of the ER. Ii has a second function, which is the delivery of the MHC class II molecules to a specialised low-pH endosomal compartment called MIIC (MHC class II compartment), where peptide loading can occur. In this compartment, Ii is cleaved by acid proteases such as cathepsin S in several steps. The initial cleavages generate a truncated form of Ii that remains bound to the MHC class II molecule and to the membrane. Subsequent cleavage releases the MHC class II molecule from the membrane-associated fragment of Ii, leaving a short fragment called CLIP (class II-associated invariant-chain peptide) still bound to the MHC molecule and blocking the binding groove. By fusion of the MIIC with incoming endosomes the MHC class II molecule eventually enters the cell's endosomal pathway and encounters and binds peptides. The MIIC contains a special type of MHC class II molecule, called HLA-DM in humans, which binds to empty MHC class II molecules stabilising them. It catalyses the release of CLIP, the subsequent loading of peptides and the already mentioned peptide-editing function. HLA-DM does not bind itself peptides, as the region of the groove is closed in this molecule. Stable complexes are finally transported to the cell surface. As with MHC class I

molecules, MHC class II molecules in uninfected cells bind peptides derived from self-proteins.

HLA class II polymorphism

Two properties of the MHC make it difficult for pathogens to evade the immune system. First, the MHC is polygenic: it contains several different MHC class I and class II genes, so that every person possesses a set of MHC molecules with different peptide-binding specificities. Second, the MHC is highly polymorphic, i.e. there are multiple variants (alleles) of each gene in the population as a whole (see Figure 1.1 for HLA allele nomenclature). The genes encoding the α and β chains of MHC class II molecules are contiguous within the MHC [3]. There are three pairs of classical MHC class II α- and β-chain genes, called *HLA-DR*, *HLA-DP* and *HLA-DQ*. Although both chains contribute to the formation of the peptide-binding groove and both can be polymorphic, the β chains are much more polymorphic than the α chains. The DR α-chain is encoded by the *HLA-DRA* locus and shows basically no polymorphism (Table 1.1). The DR β-chain is encoded by four loci, *HLA-DRB1* (most variable), *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5*. However no more than two functional loci are present on a single chromosome, i.e. *DRB1* plus any one of the three other *DRB* genes. In HLA-DQ both the α and β chains vary greatly and are encoded by loci *HLA-DQA1* and *HLA-DQB1*, respectively. In HLA-DP the α and β chains are encoded by loci *HLA-DPA1* and *HLA-DPB1*, respectively. Thus, the three sets of genes of one chromosome may give rise to four types of MHC class II molecules (2 DR, 1 DQ and 1 DP), each with different peptide specificities. The high polymorphism, with most individuals being heterozygous, and the co-dominant expression of the MHC products, means that the number of MHC class II molecules expressed in an individual may actually double. The number of different MHC molecules may be increased still further by the combination of α and β chains encoded by different chromosomes, although not all combinations may form a stable dimer. The number of MHC class II molecules an individual may express is nevertheless small compared to the vast number of antigens it may have to react to. Therefore, MHC molecules need to have limited specificity.

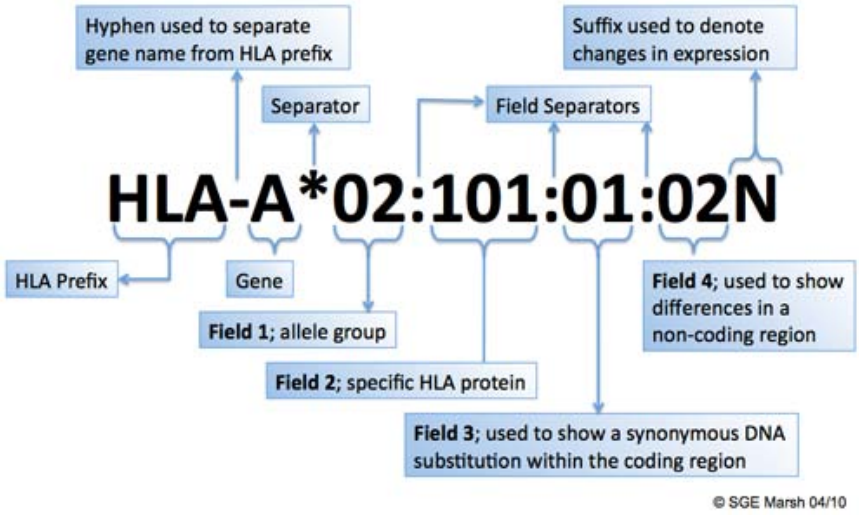


Figure 1.1. Nomenclature of HLA alleles [7,8] (http://hla.alleles.org/nomenclature/naming.html). Image courtesy of Prof. SGE Marsh, HLA Informatics Group, Anthony Nolan Research Institute, London, UK.

Table 1.1. Classical HLA class II genes and allele and protein numbers at the IMGT/HLA database (July 2013) [7].

Gene	Alleles	Proteins	Description
HLA-DRA	7	2	DR α-chain
HLA-DRB1	1355	1005	DR β1-chain determining specificities DR1, DR2, DR3, DR4, DR5 etc.
HLA-DRB3	58	46	DR β3-chain determining DR52 and Dw24, Dw25, Dw26 specificities
HLA-DRB4	15	8	DR β4-chain determining DR53 specificity
HLA-DRB5	20	17	DR β5-chain determining DR51 specificity
HLA-DQA1	51	32	DQ α-chain
HLA-DQB1	415	277	DQ β-chain
HLA-DPA1	37	19	DP α-chain
HLA-DPB1	190	147	DP β-chain

The products of individual MHC alleles can differ by up to 20 amino-acid residues [7]. Most of the differences are found in exposed surfaces of the N-terminal domains and in the peptide-binding groove in particular, thus conferring distinct binding specificities to each resulting MHC class II molecule. Some of the polymorphic residues in MHC molecules are located in the α-helices that flank the peptide-binding groove, in the contact region with the TCR. This contributes, together with induced differences in the conformation of the peptide, to the different recognition of the same antigen peptide by different MHC class II molecules (MHC restriction).

Peptide binding to MHC class II molecules

Since the first MHC class II crystal structure was reported in 1993 [9,10], more than 60 crystallographic structures of the antigen recognition domains of MHC class II molecules have been resolved and deposited in the PDB (Protein Data Bank [11]), mostly with bound peptide. In particular, at the time of this writing (July 2013) the IEDB (Immune Epitope Database [12,13]) reports 37 HLA class II:peptide structures, 1 HLA-DP, 5 HLA-DQ and 31 HLA-DR. The different MHC class II molecules are highly conserved at the structural level, even across species. First observations from the initial HLA-DR1 structures were that peptides bind in an extended conformation (Figure 1.2) that projects from both ends of an open-ended antigen-binding groove [9] and that pockets in the peptide-binding site accommodate five of the peptide side chains, explaining specificity, while a considerable number of hydrogen bonds between conserved DR1 residues and the backbone of the peptide provide a universal mode of peptide binding distinct from that used by MHC class I [10,14]. MHC class II usually binds peptides 9 to 30 residues long. The extended conformation adopted by the 9-residue peptide core results in the side chains of peptide residues at positions P1, P4, P6 and P9 being directed into the MHC class II peptide-binding groove. The residues at these positions in the peptide are termed anchor residues because the interactions of their side chains with distinctive pockets in the binding groove further stabilize the MHC class II:peptide complex. Other peptide side chains may establish contacts

with the binding groove, most notably at position P7, which is directed sideways within the groove and can therefore be described as binding to a P7 pocket [15]. The binding pockets contain mostly polymorphic residues, such that different alleles bind specific groups of residues in each pocket [16,17]. Despite this selectivity component, however, the anchor positions tend to be highly degenerate, making the prediction of class II epitopes difficult.

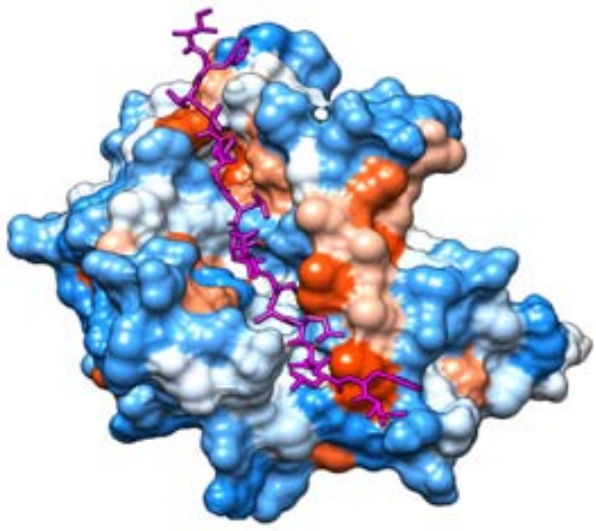


Figure 1.2. Example of peptide bound to an MHC class II molecule [18]

The role of interactions between the peptide flanking regions (positions preceding P1 and succeeding P9) and the MHC class II molecule is still unclear, but there exist evidences that these segments, far from completely superfluous, can have an affect on CD4+ T cell antigen recognition [19].

HLA class II peptide databases

Advances in high-throughput immunoproteomics methods [20,21] have enabled the generation of vast libraries of HLA ligands. Table 1.2 lists databases including experimentally determined HLA class II ligands. Importantly, some of these databases report both binding and non-binding peptides, an essential feature for

the training of prediction algorithms. The Immune Epitope Database (IEDB) [22], with over 67.000 peptides assayed for 186 HLA class II alleles (July 2013), has quickly become the major reference and was the database chosen for the training of the algorithms developed in this thesis.

Table 1.2. List of public databases containing HLA class II ligands.

Database name	URL	Ref.
SYFPEITHI	http://www.syfpeithi.de/	[23]
AntiJen	http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm	[24]
MHCBN	http://www.imtech.res.in/raghava/mhcbn	[25]
EPIMHC	http://imed.med.ucm.es/epimhc/	[26]
IEDB	http://www.immuneepitope.org/	[22]

Prediction of HLA class II epitopes

The ability to predict HLA class II epitopes in a protein or a proteome has a number of fundamental medical applications [27]:

- Design of vaccines for cancer and infectious diseases by means of antigen discovery. This can be achieved by using in-silico tools to perform a rapid screening of whole genomes, specific proteomes (e.g. cancer proteomes) or protein families. The aim of the screening is to search for proteins with antigenic properties [28], reducing the time and cost of cell-based or *in-vivo* screens.

- Design of new protein therapeutics, free of HLA class II epitopes, which is a common problem of recombinant proteins. To this end, in-silico tools may be used to search for MHC class II motives in recombinant proteins with potential use as therapeutics, enabling the design of mutant, non-reactive sequences, thus preventing immune neutralisation of the therapeutic agent.

- Antigen discovery in whole genomes or in specific proteomes (e.g. cancer proteomes) for purposes other than vaccine design. For example, for the understanding of immune processes.

- Discovery of cross-reactive HLA class II epitopes in the context of autoimmunity. Cross-reactivity may be caused by allergenic proteins with high homology to human proteins, i.e. in celiac disease, or by pathogen mimicry as a result of an evolutive process to avoid host defences. These cases can be detected by host-pathogen proteome comparisons to find homologue regions in both organisms that could be recognized by MHC class II molecules.

- Discovery of immunogenic determinants of self in the context of grafts and transplants. This could be addressed computationally by the study of MHC restriction, using the relevant MHC allele populations.

- Determination of HLA class II epitopes associated with IgE responses in the context of allergy. The combination of methods for the identification of HLA class II and antibody epitopes may be used in this context, as well as for the development of vaccines inducing both CD4+ and antibody responses.

The identification of peptides that bind to MHC class II molecules is significantly more complex than equivalent predictions for MHC class I. These difficulties are due to the fact that the 9-residue peptide core is inserted in a longer sequence, complicating its determination, and the range of amino acids that may occupy the anchor positions is larger than in the MHC class I case. Many approaches have been proposed in order to overcome this setback. In the iterative self-consistent approach [29] a first alignment of 9-residue segments from the problem peptide set is performed by fixing the P1 position (e.g. using a simple motif predictor, see below). The resulting set of 9-residue sequences is then reduced until convergence by using an iterative process in which, at each iteration, motives with the lower scores are removed from the training set. Other approaches align peptides as found in crystallographic structures or create an initial basis set of sequences using a prediction method created previously.

The uncertainty introduced by the unknown peptide length may be partly alleviated in the future by current efforts toward the prediction of antigen pre-processing for presentation to CD4+ T cells [30]. An efficient prediction of antigen pre-processing would enable an early discrimination of peptides, as currently all possible 9-residue sequences from a protein need to be considered. Knowing the right sequence length would also allow the introduction of the potential effect of flanking residues in the predictions of binding.

Many bioinformatics tools are available for the prediction of HLA class II epitopes. They are based on existing database knowledge on sequence and/or structure of experimentally probed peptide binders and non-binders. The prediction algorithms learn from this data by using different approaches, from straightforward statistics (i.e. generating sequence motifs or position-specific scoring matrices), to machine-learning techniques such as artificial neural networks and support vector machines [31]. In the following paragraphs I introduce the main concepts and available public servers.

There are different ways to classify prediction methods on the basis their output, the kind of the data used or the mathematical approach. The output given by a predictor may be used to divide the methods into two different categories. Qualitative methods classify the epitopes as binders/non-binders or strong/weak binders, the prediction being usually based on the presence/absence of certain amino-acid residues at the anchoring positions. Quantitative methods, on the other hand, return a numeric binding score meant to predict the peptide's binding affinity. If the methods are classified on the type of data used, they may be divided in sequence-based methods, where binding patterns are elucidated using only the amino-acid sequence, and structure-based methods, where both the peptide's amino-acid sequence and the structural features of the specific HLA class II molecule are considered. Finally, methods may be classified on the basis of the underlying mathematical approach. Within this classification, a first group of methods make use of position-specific scoring matrices (PSSMs) [32], which are two-dimensional matrices commonly used to represent motifs. In the current context, PSSMs contain one column for each position of interest, i.e. the anchoring

residues (P1, P4, P6, P7 and P9) or residues 1 to 9 if the entire epitope core is considered, and one row for each symbol of the alphabet, i.e. the 20 amino acids. The final score for the peptide results from adding the individual scores of each amino acid in the query sequence for the corresponding position in the HLA class II binding groove. This representation assumes that each residue of a peptide independently contributes to HLA class II binding. Although the peptide length is fixed, peptides of various sizes can be analysed by generating all possible 9-residue sequences contained in them. A second group of prediction methods make use of Hidden Markov Models (HMMs) [33], a technique commonly used for pattern recognition. HMMs are statistical models in which the system is assumed to follow a Markov process. The objective is to determine the unknown (hidden) parameters of the model from the observable parameters. In this case, the observables are the training peptide sequences, the number of states is optimised for the training data set and transitions between states and their emission of symbols (amino acids) are governed by probabilities reflecting the observed data. After the training phase, the model is used to assign binding-likelihood values to query peptide sequences. Advantages of HMMs are their capacity to treat peptides of varying sizes at once (and discover multiple hidden binding patterns in them) and consider correlations between adjacent residues (through the Markov transition probabilities). A third group of methods use Artificial Neural Networks (ANNs) [34], which are also of common use for classification and pattern recognition. An ANN consists of nodes (computational elements) that receive signals via interconnecting arcs. An ANN can be trained to recognize a pattern by strengthening signals (adjusting arc weights) and by adjusting activation thresholds for individual nodes. The advantages of ANNs are that they are adaptive, are effective with nonlinear data, and are tolerant to a certain level of erroneous data. However, they need to determine a large number of parameters and therefore require larger amounts of binding data than simpler prediction methods. Unlike HMMs, they require, as with PSSMs, the pre-alignment of the peptides. A fourth group of methods make use of support vector machines (SVMs) [35]. A SVM is a classification technique that tries to differentiate members of different populations in a sample by building hyper-planes between them so as to discriminate between the characteristics of the

different members in the sample. Their predictions are of similar accuracy to ANN and HMM predictions, with the added advantage that they can be trained on relatively small peptide datasets. A fifth group of methods make use of quantitative structure-activity relationships (QSAR) [36]. The objective of the QSAR approach is to establish a relationship between the chemical structure and the biological response. To this end, the chemical structure is translated into a quantitative description, followed by a statistical modelling. Thus, peptides may be described by global and local (amino-acid level) descriptors of various sorts (of physical or chemical nature) and the resulting models fitted to quantitative binding data (e.g. IC50s). Excess of number of terms and descriptors may quickly lead to over-fitting if large training sets are not available. Finally, structure-based approaches may use a variety of techniques to assess peptide binding quantitatively, from molecular modelling and simulation [37] to 3D-QSAR [38]. These latter type of methods are, however, not amenable to fast, automated predictions.

The most popular web-servers oriented to epitope prediction are:

- SYFPEITHI [23] is an early database for MHC ligands and peptide motifs with an associated MHC class I and II predictor. The predictor, probably the simplest one still in function, was built with motif-based 20x9 PSSMs.

- Proped [39] is a TEPITOPE [40] web service. TEPITOPE was the first program to use pocket profiles to extrapolate PSSM data from one HLA class II molecule to another, generating so-called virtual matrices. It can make predictions for 51 HLA-DR molecules. A recent extension, TEPITOPEpan [41], promises unlimited coverage of HLA-DR molecules.

- Rankped [42] is another PSSM-based method. To derive the PSSMs, the peptide core of each epitope was deduced from structural or sequence-similarity alignments of peptides presented by the different MHC class II molecules. Once peptides were aligned, all repeated cores were collapsed to one and a profile was created using PROFILEWEIGHT [43] or BLK2PSSM [44].

- HLA-DR4Pred [45] is a predictor dedicated to the HLA-DR1*04:01 molecule. It was trained using 1154 peptides (587 presented by HLA-DR1*04:01) spread between 5 groups to perform cross-validation (5 training rounds using 4 groups as a training set and 1 as a test set during different rounds). Two predictors were developed, one using a SVM approach and another one using an ANN approach. In both cases the peptide is codified using a binary vector of 20 positions for each of the 9 core amino acids. The SVM was trained using different kernels (linear, polynomial, RBF and sigmoid) and the RBF kernel was finally selected for the server. The ANN was trained using standard feed-forward backpropagation with 1 hidden layer. The number of cycles and loops performed were fixed internally by the method.

- ARB matrix [46], which stands for Average Relative Binding matrix, is a PSSM-based predictor based on affinities rather than frequencies. Thus, the scoring-matrix elements are in this case a function of the binding affinity (IC50) of peptides having the specific amino acid at the specific position relative to the affinity of all other peptides. After matrix generation, the scoring for the 9-residue segments were fitted to IC50 values by means of different regression techniques.

- SVMHC [47] is a SVM predictor. The singularity here is that performance was measured using the Matthews correlation (MC), which is often used in binary classification to determine if the obtained results have a good correlation with the expected ones.

- SVRMHC [48] is a SVM regression method in which the produced model depends only on a subset of the training data (marginal data points with scores near the threshold are dismissed). Models were constructed for MHC class II molecules with IC50 values available for at least 50 peptides. Peptide cores were first defined using the iterative self-consistent strategy. For each MHC molecule, six different configurations were attempted resulting from three different kernel functions (linear, polynomial and RBF) in combination with two sequence encoding schemes (sparse encoding and 11-factor encoding).

- SMM-align [49] uses a Gibbs sampler method [50] to generate a 20x9 weight matrix. SMM-align seeks to identify the weight matrix that optimally reproduces the measured IC50 values for each peptide by using a Metropolis Monte Carlo procedure. The peptide sequences are presented to SMM-align using two sequence-encoding schemes, sparse encoding and Blosum50. The final prediction score for a 9-residue peptide is then calculated as the average of the sparse and Blosum encoded predictions. The threshold used for the definition of peptide binders is an IC50 of 500 nM. This method uses the peptide length as a training parameter input, i.e. it incorporates the flanking residues to the prediction.

- NN-align [51] is an ANN-based method that allows for simultaneous identification of the MHC class II binding core and binding affinity. NN-align was trained using an algorithm that allows for correction of bias in the training data due to redundant binding core representation. Information on the residues flanking the peptide-binding core is also incorporated.

- NetMHCIIpan [52] is another ANN-based method. The input sequences were presented to the neural network in three distinct manners: sparse encoding, Blosum50 encoding and a mixture of the two. Peptide flanking residues were also incorporated (to a maximum of three per side). The binding core and flanking residues in the peptide training set were identified with SMM-align. Finally, the HLA sequence (contact residues) was also incorporated as input to the ANN. Thus, starting from data on 14 HLA-DR molecules and taking both peptide and HLA sequence information into account, NetMHCIIpan is meant to generalize and predict peptide binding for any other HLA-DR molecule of known sequence, similarly as TEPITOPEpan.

- IEDB-AR [53] predicts peptide binding by generating a consensus score between NN-align, SMM-align, TEPITOPE and NetMHCIIpan scores.

- MHC2pred [54] uses a standard SVM-based approach.

- Epitop [55] is a QSAR-based method that uses principal components analysis (PCA) to construct a PSSM. It builds on the proteochemometrics approach [56], which uses both protein and ligand descriptors (as opposed to standard QSAR which makes only use of ligand descriptors). A single proteochemometric model could potentially predict peptide binding to many MHC proteins. Amino acids in the 9 core positions are encoded with 3 descriptors (volume, polarity and hydrophobicity). HLA polymorphic residues as well as cross-terms between adjacent peptide residues and between peptide-HLA residues are incorporated. The model was derived using the iterative self-consistent approach, were IC50s were used as target.

- EpiDOCK [57] is the first structure-based prediction server. Three initial crystallographic structures (of DR, DQ and DP molecules) were used as templates for the modelling of a total of 23 HLA class II molecules. Single amino-acid substitutions were sued to construct a virtual combinatorial library of peptides. AutoDock [58] and GOLD [59] were used to dock the peptides to the HLA molecules. The resulting scores were use to generate a 20x9 PSSM for each molecule.

References

1. Murphy K, Travers P, Walport M (2008) Janeway's Immunobiology. New York: Garland Science, Taylor & Francis Group, LLC.

2. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, et al. (2003) The DNA sequence and analysis of human chromosome 6. Nature 425: 805–811. doi:10.1038/nature02055.

3. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, et al. (2004) Gene Map of the Extended Human Mhc. Nat Rev Genet 5: 889–899. doi:http://dx.doi.org/10.1038/nrg1489.

4. Lipscomb MF, Masten BJ (2002) Dendritic Cells: Immune Regulators in Health and Disease. Physiol Rev 82: 97–130. doi:10.1152/physrev.00023.2001.

5. Jensen PE (2007) Recent advances in antigen processing and presentation. Nat Immunol 8: 1041–1048. doi:10.1038/ni1516.

6. Goldberg AL, Cascio P, Saric T, Rock KL (2002) The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. Mol Immunol 39: 147–164. doi:10.1016/S0161-5890(02)00098-6.

7. Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, et al. (2011) The IMGT/HLA database. Nucleic Acids Res 39: D1171–D1176. doi:10.1093/nar/gkq998.

8. Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, et al. (2010) Nomenclature for factors of the HLA system, 2010. Tissue Antigens 75: 291–455. doi:10.1111/j.1399-0039.2010.01466.x.

9. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, et al. (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364: 33–39. doi:10.1038/364033a0.

10. Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, et al. (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature 368: 215–221. doi:10.1038/368215a0.

11. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, et al. (2012) The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res 41: D475–D482. doi:10.1093/nar/gks1200.

12. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, et al. (2010) The Immune Epitope Database 2.0. Nucleic Acids Res 38: D854–D862. doi:10.1093/nar/gkp1004.

13. Ponomarenko J, Papangelopoulos N, Zajonc DM, Peters B, Sette A, et al. (2011) IEDB-3D: structural data within the immune epitope database. Nucleic Acids Res 39: D1164–D1170. doi:10.1093/nar/gkq888.

14. Schueler-Furman O, Altuvia Y, Margalit H (2001) Examination of possible structural constraints of MHC-binding peptides by assessment of their native structure within their source proteins. Proteins Struct Funct Bioinforma 45: 47–54. doi:10.1002/prot.1122.

15. Jones Ey, Fugger L, Strominger JL, Siebold C (2006) MHC class II proteins and disease: a structural perspective. Nat Rev Immunol 6: 271–282. doi:http://dx.doi.org/10.1038/nri1805.

16. Rammensee HG, Friede T, Stevanoviíc S (1995) MHC ligands and peptide motifs: first listing. Immunogenetics 41: 178–228. doi:10.1007/BF00172063.

17. Bondinas GP, Moustakas AK, Papadopoulos GK (2007) The spectrum of HLA-DQ and HLA-DR alleles, 2006: a listing correlating sequence and structure with function. Immunogenetics 59: 539–553. doi:10.1007/s00251-007-0224-8.

18. Muixí L, Gay M, Muñoz-Torres PM, Guitart C, Cedano J, et al. (2011) The peptide-binding motif of HLA-DR8 shares important structural features with other type 1 diabetes-associated alleles. Genes Immun 12: 504–512. doi:10.1038/gene.2011.26.

19. Cole DK, Godkin A (2013) Re-directing CD4+ T cell responses with the flanking residues of MHC class II-bound peptides: the core is not enough. Front T Cell Biol 4: 172. doi:10.3389/fimmu.2013.00172.

20. Schirle M, Weinschenk T, Stevanović 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. doi:10.1016/S0022-1759(01)00459-8.

21. Purcell AW, Gorman JJ (2004) Immunoproteomics Mass Spectrometry-based Methods to Study the Targets of the Immune Response. Mol Cell Proteomics 3: 193–208. doi:10.1074/mcp.R300013-MCP200.

22. Kim Y, Ponomarenko J, Zhu Z, Tamang D, Wang P, et al. (2012) Immune epitope database analysis resource. Nucleic Acids Res 40: W525–W530. doi:10.1093/nar/gks438.

23. Rammensee H-G, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50: 213–219. doi:10.1007/s002510050595.

24. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, et al. (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res 1: 4. doi:10.1186/1745-7580-1-4.

25. Lata S, Bhasin M, Raghava GP (2009) MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. BMC Res Notes 2: 61. doi:10.1186/1756-0500-2-61.

26. Reche PA, Zhang H, Glutting J-P, Reinherz EL (2005) EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics 21: 2140–2141. doi:10.1093/bioinformatics/bti269.

27. De Groot AS (2006) Immunomics: discovering new targets for vaccines and therapeutics. Drug Discov Today 11: 203–209. doi:10.1016/S1359-6446(05)03720-7.

28. Davies MN, Flower DR (2007) Harnessing bioinformatics to discover new vaccines. Drug Discov Today 12: 389–395. doi:10.1016/j.drudis.2007.03.010.

29. Doytchinova IA, Flower DR (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: 2263–2270. doi:10.1093/bioinformatics/btg312.

30. Hoze E, Tsaban L, Maman Y, Louzoun Y (2013) Predictor for the effect of amino acid composition on CD4+ T cell epitopes preprocessing. J Immunol Methods 391: 163–173. doi:10.1016/j.jim.2013.02.006.

31. Lafuente E, Reche P (2009) Prediction of MHC-Peptide Binding: A Systematic and Comprehensive Overview. Curr Pharm Des 15: 3209–3220. doi:10.2174/138161209789105162.

32. Reche PA, Glutting J-P, Zhang H, Reinherz EL (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics 56: 405–419. doi:10.1007/s00251-004-0709-7.

33. Noguchi H, Kato R, Hanai T, Matsubara Y, Honda H, et al. (2002) Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. J Biosci Bioeng 94: 264-270. doi: 10.1016/S1389-1723(02)80160-8.

34. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics 14: 121-130. doi:10.1093/bioinformatics/14.2.121.

35. Bhasin M, Raghava GP (2004) SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics 20: 421-423. doi:10.1093/bioinformatics/btg424.

36. Doytchinova IA, Walshe V, Borrow P, Flower DR (2005) Towards the chemometric dissection of peptide--HLA-A*0201 binding affinity: comparison of local and global QSAR models. J Comput Aided Mol Des 19: 203-212. doi:10.1007/s10822-005-3993-x.

37. Muixí L, Carrascal M, Alvarez I, Daura X, Martí M, et al. (2008) Thyroglobulin Peptides Associate In Vivo to HLA-DR in Autoimmune Thyroid Glands. J Immunol 181: 795–807.

38. Hattotuwagama CK, Doytchinova IA, Flower DR (2007) Toward the prediction of class I and II mouse major histocompatibility complex-peptide-binding affinity: in silico bioinformatic step-by-step guide using quantitative structure-activity relationships. Methods Mol Biol 409: 227-245. doi:10.1007/978-1-60327-118-9_16.

39. Singh H, Raghava GPS (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17: 1236–1237. doi:10.1093/bioinformatics/17.12.1236.

40. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, et al. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotech 17: 555–561. doi:10.1038/9858.

41. Zhang L, Chen Y, Wong HS, Zhou S, Mamitsuka H, et al. (2012) TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. PLoS One 7: e30483. doi:10.1371/journal.pone.0030483.

42. Reche PA, Glutting J-P, Reinherz EL (2002) Prediction of MHC class I binding peptides using profile motifs. Hum Immunol 63: 701–709. doi:10.1016/S0198-8859(02)00432-9.

43. Thompson JD, Higgins DG, Gibson TJ (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. Comput Appl Biosci CABIOS 10: 19–29.

44. Henikoff JG, Henikoff S (1996) Using substitution probabilities to improve position-specific scoring matrices. Comput Appl Biosci CABIOS 12: 135–143. doi:10.1093/bioinformatics/12.2.135.

45. Bhasin M, Raghava GPS (2003) Prediction of Promiscuous and High-Affinity Mutated MHC Binders. Hybrid Hybridomics 22: 229–234. doi:10.1089/153685903322328956.

46. Bui H-H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, et al. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB

matrix applications. Immunogenetics 57: 304–314. doi:10.1007/s00251-005-0798-y.

47. Dönnes P, Kohlbacher O (2006) SVMHC: a server for prediction of MHC-binding peptides. Nucleic Acids Res 34: W194–197. doi:10.1093/nar/gkl284.

48. Wan J, Liu W, Xu Q, Ren Y, Flower DR, et al. (2006) SVRMHC prediction server for MHC-binding peptides. BMC Bioinformatics 7: 463. doi:10.1186/1471-2105-7-463.

49. Nielsen M, Lundegaard C, Lund O (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics 8: 238. doi:10.1186/1471-2105-8-238.

50. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, et al. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics 20: 1388-1397. doi:10.1093/bioinformatics/bth100.

51. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics 10: 296. doi:10.1186/1471-2105-10-296.

52. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, et al. (2008) Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS Comput Biol 4: e1000107. doi:10.1371/journal.pcbi.1000107.

53. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, et al. (2008) Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Res 36: W513–518. doi:10.1093/nar/gkn254.

54. Lata S, Bhasin M, Raghava GPS (2007) Application of machine learning techniques in predicting MHC binders. Methods Mol Biol Clifton NJ 409: 201–215. doi:10.1007/978-1-60327-118-9_14.

55. Dimitrov I, Garnev P, Flower DR, Doytchinova I (2010) EpiTOP—a proteochemometric tool for MHC class II binding prediction. Bioinformatics 26: 2066–2068. doi:10.1093/bioinformatics/btq324.

56. Dimitrov I, Garnev P, Flower DR, Doytchinova I (2010) Peptide binding to the HLA-DRB1 supertype: A proteochemometrics analysis. Eur J Med Chem 45: 236–243. doi:10.1016/j.ejmech.2009.09.049.

57. Atanasova M, Patronov A, Dimitrov I, Flower DR, Doytchinova I (2013) EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. Protein Eng Des Sel PEDS. doi:10.1093/protein/gzt018.

58. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, et al. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30: 2785-2791. doi:10.1002/jcc.21256.

59. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267: 727–748. doi:10.1006/jmbi.1996.0897.

Chapter 2. HLA2db: a suite for the prediction of HLA class II epitopes based on non-assisted self-learning procedures

Abstract

Recent advances in proteomics have substantially increased the amount of available data on peptides that bind to HLA class II. This has led to a significant improvement in accuracy and coverage of bioinformatic predictions of HLA class II binding peptides. The suite presented here represents another step in this direction. We have created a web-based set of tools, following a scoring-matrix-type approach, for the identification in query sequences of peptide motifs that bind to specific HLA class II molecules and for the search of these motifs in entire proteomes of microorganisms and human. This set of tools has been especially devised to meet the needs of the immunologist, assisting the analysis of motifs in antigen presentation by HLA class II in the contexts of infection, autoimmunity or vaccine design. With this objective, the applications have been designed to: a) maximise the versatility of the queries, with no restrictions on the amount or length of input sequences to be evaluated for epitope prediction; b) enable the scanning of over 240 microbial proteomes and the human proteome for predicted motifs; c) enable a private use of the suite's profile-development tools so that the user may incorporate his/her own peptide libraries to derive new profiles or improve existing ones, shortening the time between the acquisition of the experimental data, the refinement of the resulting library and the determination of binding motifs; d) enable the combined analysis of query sequences against all the HLA class II molecules available in the suite, making it possible to eventually analyse a full proteome for epitopes in relation to all HLA class II molecules in a single shot; e) periodically expand the number and quality of HLA class II binding profiles available for prediction by automatic updates incorporating new data from reference databases. This service is freely accessible at http://bioinf.uab.cat/hla2db.

Introduction

Major histocompatibility complex (MHC) class II molecules –human leukocyte antigen or HLA class II in human– are cell-surface glycoproteins that present (mostly) exogenous peptides to CD4+ T lymphocytes. MHC class II loci are extremely polymorphic. Allelic variation between MHC class II molecules of different individuals accounts for the differential ability to bind and display antigenic peptides and plays also a major role in autoimmunity [1]. MHC class II molecules are constituted by two chains, α and β, each having two domains, α1 and α2 and β1 and β2. The α2 and β2 immunoglobulin-like domains are bound to a transmembrane region anchoring the MHC-II molecule to the cell membrane, while the heterodimer of α1 and β1 forms the peptide-binding groove. Although both chains contribute to the peptide-binding groove and both can be polymorphic, the β chains are for unknown reasons much more polymorphic than the α chains. There are three classical MHC class II molecules in human, HLA-DP, HLA-DQ and HLA-DR, encoded by the HLA complex on chromosome 6. In HLA-DP the α and β chains are encoded by loci *HLA-DPA1* and *HLA-DPB1*, respectively. In HLA-DQ both the α and β chains are variable and are encoded by polymorphic loci *HLA-DQA1* and *HLA-DQB1*, respectively. A person often expresses two α-chain and two β-chain variants that may form 4 DQ isoforms. The basically invariant DR α-chain is encoded by the non-polymorphic *HLA-DRA* locus. Different DR β-chains are encoded by four loci, *HLA-DRB1* (most variable), *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5*. However no more than two functional loci are present on a single chromosome, i.e. *DRB1* plus any one of the three other *DRB* genes.

The class II peptide-binding site is formed by one α-helix and four β-strands from each of the membrane-distal domains, α1 and β1. The sheet floors and helical walls define a groove suitable for binding 9 peptide residues within a cleft that is open at both ends, thus enabling the binding of a broad range of peptide lengths, typically up to 30 residues. One fundamental feature of HLA molecules is their ability to form stable complexes with a large number of different peptide sequences. This capacity arises from the interaction between conserved HLA residues distributed

along the binding groove and the peptide main chain, thus providing a certain level of sequence-independent affinity for peptide ligands. The affinity and specificity of binding is increased by the interaction of residue side chains at specific positions in the peptide sequence (anchor residues) with various pockets within the binding groove of the molecule. Thus, of the 9 core peptide residues binding the HLA surface, residues at position 1, 4, 6, 7 and 9 typically occupy the corresponding pockets –P1, P4, P6, P7 and P9– in the HLA groove. These pockets contain mostly polymorphic residues, such that different alleles bind specific groups of residues in each pocket [2, 3]. Despite this selectivity component, however, the anchor positions tend to be highly degenerate, making the prediction of class II epitopes difficult.

Advances in high-throughput immunoproteomics methods [4, 5] have enabled the generation of vast libraries of MHC ligands [6-13]. These peptide libraries have allowed the training of computational algorithms to recognize and predict MHC-binding peptides, based solely on sequence or on sequence and structure [14, 15]. Sequence methods make use of pattern-recognition approaches. A pattern is a representation of the over-populated amino acids in specific positions of the peptide sequence. In the simplest implementation, predictions based on patterns do not generate a score but two possible states, presence or absence of the pattern. This simple approach, however, does neither discriminate positive patterns by their affinity nor allow for compensating effects across pockets. This can be overcome by use of a position-specific scoring matrix (PSSM), which assigns a probability to each amino acid at each position enabling the calculation of a combined score for the peptide. In general, this probability is derived directly from normalised observed frequencies [7, 16], but may also incorporate experimental affinity data (IC_{50} values) [17, 18]. In some cases, the matrices have been built by assembly of pre-classified pocket profiles, generated from multiple alignments of HLA sequences and structural identification of pocket polymorphisms [19, 20]. Limitations of PSSMs include the inability to capture positional dependences between amino acids (i.e. correlations) or to jointly analyse sequences of variable length. Correlations may be accounted for, for example, by use of additional 20x20

matrices describing the co-existence of atoms at given positions, in a QSAR-type parameterisation procedure [21, 22].

Other HLA class II binding predictors are based on Hidden Markov Models (HMMs) [23] and machine-learning techniques, including artificial neural networks (ANNs) [23, 24] and support vector machines (SVMs) [25]. Although in principle more powerful, machine-learning techniques are particularly vulnerable to poorly representative training sets and to noise (arising from the experimental uncertainty of binding measurements and potential errors in peptide-database annotations), two relevant issues in the context of many HLA class II molecules. To compensate for technique-specific limitations, some applications provide a consensus from a combination of different methodologies [26, 27].

At a different level, structure-based approaches overcome the limitations of the above-mentioned high-throughput methods (i.e. accounting for residue correlations, variable sequence length, interactions outside the core region, etc.) by evaluating on physical grounds the interaction between the peptide and the HLA molecule, providing an estimate for the free energy of binding. These procedures require, however, intensive computation and are therefore amenable only to the study of small peptide sets. They also require previous knowledge of the structure of the HLA molecule, although three-dimensional-structure modelling based on homology templates is often a valid alternative for this family of proteins. The structure of the complex with the peptide can be then derived with algorithms such as pDOCK [28]. Binding affinity can be finally assessed using physically based scoring functions trained on experimental data [29, 30] or empirical potentials and statistical-mechanics expressions in combination with sampling algorithms such as molecular-dynamics simulation [31].

Here we present HLA2db, a PSSM-based, unsupervised, online system for the identification of HLA class II binding motifs in polypeptide sequences. Its functionality includes the identification and scoring of potential epitopes in query sequences entered by the user (with no limitation in number of sequences or their length) in relation to the selected HLA class II molecule (18 HLA-DR, 8 HLA-DQ and 5 HLA-DP molecules are currently available). The selection may include a single

HLA molecule, all HLA2db-available molecules individually, or all molecules combined, the latter case providing the number of HLA molecules recognising each predicted epitope. It can also scan microbial proteomes (currently 243) and the human proteome for matches of the binding motifs identified within the input sequences. By registering, the user may also upload private, experimentally validated peptide sequences for the automatic generation / improvement of the PSSM of a specific HLA molecule. The system performs its own checks to ensure that the new sequences make a valid and significant contribution to the PSSM, and the new PSSM is kept for the user's private use in successive queries. A maintenance system has been also set up, such that each time the internal database is updated with the incorporation of new information from public peptide libraries all the public PSSMs are automatically recalculated. By doing this, the quality of the predictions shall progressively improve.

The service has been implemented to support vaccine and autoimmunity / tolerance studies, allowing the screening of both microbial and human proteomes for the presence of predicted HLA class II epitopes and providing information on the level of promiscuity of these epitopes by performing the analysis against the complete set of HLA molecules available. The robustness of the system is illustrated by the possibility of entering a FASTA file [32] with the entire human proteome and execute it against all 31 available HLA class II molecules. This query, which outputs its results in about 8 hours, provides a theoretical human auto-immunome (restricted to the 31 HLA molecules).

Materials and methods

Data collection

Bacterial proteomes were obtained from PATRIC (Pathosystems Resource Integration Center) [33] and corresponding sequences were downloaded from Uniprot [34]. The human proteome was downloaded from HPRD (Human

Proteome Reference Database) [35]. Epitope-presentation data was extracted from Immune Epitope Database (IEDB). All this data was imported to a PostgreSQL database-management system.

Each human and bacterial protein sequence was decomposed into 9-residue segments by running a 9-residue window over the sequence with a 1-residue step. For each 9-residue segment, the protein code (Uniprot) and starting position in the protein sequence were stored.

IEDB entries that contain information on HLA class II binding and refer to peptides with no chemical modifications were selected. The corresponding information on peptide residue sequence, qualitative binding measurement (positive or negative) and binding HLA molecule was retrieved and stored. Repeated entries (from different studies relating the same peptide to the same HLA class II protein) were collapsed to a single one using the following rule: if the number of entries with positive binding was larger than the number of entries with negative binding, the peptide was annotated as positive; else, if the number of entries with negative binding was larger, the peptide was annotated as negative; otherwise (i.e. the number of entries with positive and negative binding were equal) the peptide was discarded. The total number of positives and negatives incorporated per HLA molecule for the construction of the current HLA2db release is given in Table 2.1.

Table 2.1. Initial set of binding and non-binding peptides, from IEDB, for each molecule in the database (molecules ordered by number of binders).

Molecule	No. of binders	No. of non-binders
HLA-DRB1*01:01	6597	1800
HLA-DRB1*04:01	1997	1140
HLA-DRB1*07:01	1602	769
HLA-DRB1*15:01	1572	719
HLA-DRB5*01:01	1443	532
HLA-DRB1*11:01	1422	835
HLA-DRB1*03:01	1375	1272
HLA-DRB1*08:02	997	525
HLA-DRB4*01:01	994	403
HLA-DRB1*04:05	992	316

Molecule	No. of binders	No. of non-binders
HLA-DRB1*09:01	974	299
HLA-DRB1*13:02	906	406
HLA-DRB1*04:04	804	389
HLA-DRB3*01:01	750	677
HLA-DQA1*05:01/DQB1*03:01	612	200
HLA-DQA1*05:01/DQB1*02:01	577	262
HLA-DPA1*02:01/DPB1*01:01	423	190
HLA-DQA1*03:01/DQB1*03:02	415	247
HLA-DRB1*12:01	364	254
HLA-DQA1*01:02/DQB1*06:02	345	91
HLA-DQA1*04:01/DQB1*04:02	344	100
HLA-DQA1*01:01/DQB1*05:01	317	349
HLA-DRB3*02:02	234	177
HLA-DPA1*03:01/DPB1*04:02	217	171
HLA-DPA1*01:03/DPB1*02:01	209	144
HLA-DPA1*02:01/DPB1*05:01	188	281
HLA-DQA1*05:01/DQB1*03:02	180	187
HLA-DPA1*01/DPB1*04:01	176	162
HLA-DRB1*04:07	155	4
HLA-DRB1*13:01	137	180
HLA-DRB1*04:02	124	202
HLA-DQA1*05:01/DQB1*04:01	119	37
HLA-DRB1*16:02	118	3
HLA-DPA1*02:01/DPB1*02:01	115	22
HLA-DQA1*05:01/DQB1*06:02	99	76
HLA-DQA1*01:04/DQB1*05:03	94	27
HLA-DPA1*02:01/DPB1*04:01	90	155
HLA-DPA1*02:01/DPB1*04:02	88	109
HLA-DRB1*01:02	78	37
HLA-DRB1*08:01	74	38
HLA-DQA1*03:01:02	68	24
HLA-DQA1*05:01/DQB1*06:04	67	39
HLA-DRB1*04:03	59	120
HLA-DPA1*01:03/DPB1*04:01	49	30
HLA-DPA1*02:01/DPB1*09:01	46	18
HLA-DRB1*11:04	44	41
HLA-DRB1*01:03	42	11
HLA-DQA1*01:02/DQB1*06:04	42	22
HLA-DPA1*02:01/DPB1*20:01	42	9
HLA-DRB1*04:06	38	55

Molecule	No. of binders	No. of non-binders
HLA-DQA1*01:02/DQB1*05:02	37	14
HLA-DRB1*14:01	36	12
HLA-DQA1*01:02:04	36	18
HLA-DQA1*02:01/DQB1*02:02	35	17
HLA-DQA1*01:02:02	35	24
HLA-DQA1*05:01:01	34	15
HLA-DQA1*03:01/DQB1*03:01	34	28
HLA-DRB1*11:02	32	36
HLA-DRB1*11:03	28	38
HLA-DRB1*03:02	27	35
HLA-DPA1*01:03/DPB1*03:01	22	1
HLA-DRB1*03:03	21	6
HLA-DRB1*16:01	16	39
HLA-DRB1*13:03	15	3
HLA-DPA1*01:03/DPB1*04:02	15	5
HLA-DQA1*02:01/DQB1*02:01	14	1
HLA-DRB3*03:01	13	3
HLA-DRB1*15:02	13	26
HLA-DRB1*10:01	13	0
HLA-DRB1*14:02	12	28
HLA-DRB1*13:04	11	1
HLA-DRB1*03:05	11	6
HLA-DRB1*53:01	10	8
HLA-DRB5*02:02	8	6
HLA-DRB5*02:01	8	7
HLA-DQA1*05:01/DQB1*05:01	7	9
HLA-DQA1*03:01/DQB1*02:01	7	1
HLA-DQA1*05:01/DQB1*02:02	6	7
HLA-DPA1*01:03:01	5	6
HLA-DRB1*52:01	3	9
HLA-DRB1*04:11	3	0
HLA-DQB1*05:03	3	2
HLA-DQA1*03:01:01	3	11
HLA-DPA1*02:02/DQB1*03:19	3	0
HLA-DRB5*01:02	2	0
HLA-DRB4*01:03	2	0
HLA-DRB1*15:03	2	0
HLA-DRB1*08:04	2	0
HLA-DQA1*03:02/DQB1*04:01	2	0
HLA-DQA1*03:02/DQB1*03:03	2	6

Molecule	No. of binders	No. of non-binders
HLA-DQA1*02:01:02	2	3
HLA-DQA1*01:04:03	2	0
HLA-DQA1*01:03/DQB1*06:03	2	0
HLA-DQA1*01:01:01	2	3
HLA-DPA1*02:01/DPB1*11:01	2	0
HLA-DRB1*13:05	1	0
HLA-DRB1*03:04	1	0
HLA-DQB1*05:02	1	4
HLA-DQA1*05:05/DQB1*03:01	1	8
HLA-DQA1*05:05:01	1	6
HLA-DQA1*05:01/DQB1*06:03	1	2
HLA-DQA1*05:01/DQB1*06:01	1	3
HLA-DQA1*05:01/DQB1*04:02	1	3
HLA-DQA1*04:01:02	1	2
HLA-DQA1*03:02:01	1	0
HLA-DQA1*03:01/DQB1*04:01	1	3
HLA-DQA1*02:01/DQB1*03:03	1	0
HLA-DQA1*01:03/DQB1*06:01	1	3
HLA-DQA1*01:03:01	1	0
HLA-DRB1*08:03	0	6
HLA-DQA1*05:01/DQB1*03:03	0	4
HLA-DQA1*03:02:03	0	1
HLA-DQA1*01:01/DQB1*05:03	0	1
HLA-DPA1*02:01/DPB1*03:01	0	1
HLA-DPA1*02:01:01	0	3

Identification of epitope core sequences from multiple-match 9-residue segments

For each HLA class II molecule in our database (i.e. with binding data retrieved from IEDB) a preliminary analysis was performed of all peptides with annotated positive binding. The aim of this analysis was to identify all sequences of 9 or more residues shared by two or more peptides. In general, this will correspond to nested sets, since the probability of finding a 9-residue sequence match in different proteins is relatively small. To this end, a pairwise comparison between all peptides was performed using the following procedure: peptide A is compared to

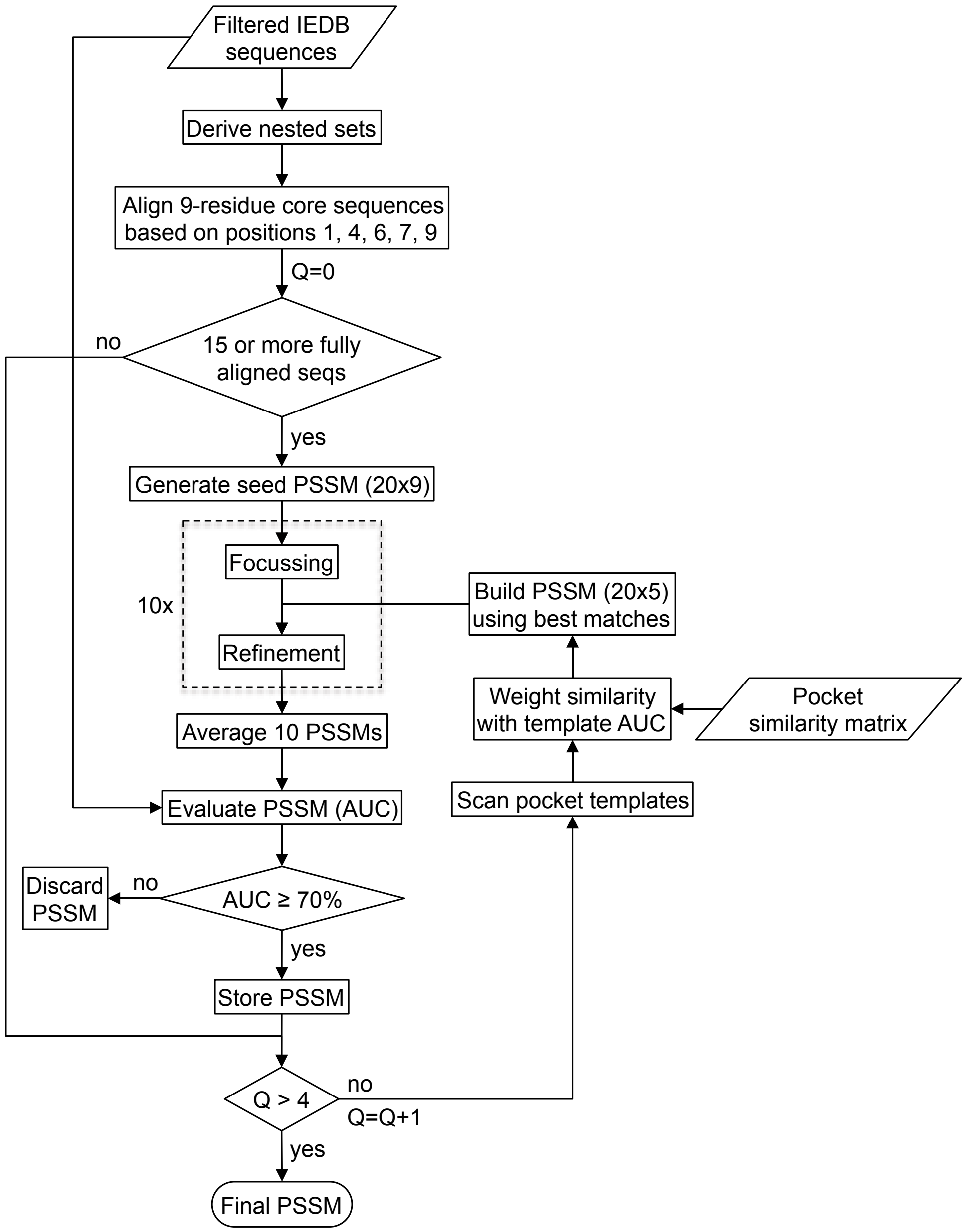
peptide B; if they share a segment of 9 or more residues this segment is isolated and compared to peptide C (for a match of 9 or more residues); this process is continued until the comparison has run over all peptides and the segment has reached a minimum common size of 9 residues. Once the common segment of A (if any) is identified, the algorithm runs the comparisons starting from peptide B. The process is iterated until the last peptide is used as reference for pairwise comparisons. All different 9-residue segments resulting from this analysis are labelled as putative epitope (core) sequences in relation to the specific HLA molecule, thus constituting the initial dataset for profile development. Longer segments found to be common to more than one peptide are discarded, since their HLA binding positions cannot be unequivocally assigned. However, they will be recovered at later stages of the global algorithm (Figure 2.1).

Definition of an initial binding profile

Amino-acid residues at positions 1, 4, 6, 7 and 9 of an epitope core sequence are expected to bind the corresponding pockets in the binding groove of the HLA class II molecule. Starting from the previous dataset of 9-residue core sequences assigned to a specific HLA molecule, an initial binding profile was derived as follows. First, the 9-residue core sequences were aligned on the basis of positions 1, 4, 6, 7, and 9 using ClustalW2 [36] with a percent identity for delay (MAXDIV) of 30% and a prohibitively high penalty for gap opening. After multiple alignment, the largest group of fully aligned sequences were selected as initial set of core sequences representing the binding profile. This strict selection of sequences was chosen to avoid the inclusion of potentially miss-annotated peptides at the initial stages of the process, by distinguishing the sequences that have similar characteristics at the amino-acid-residue level from those that diverge from the most common pattern. The global algorithm is such that these initially discarded sequences can be recovered at later stages.

If at the end of this process the set contained 15 or more core sequences, each of them coming from at least 2 original peptides, it was used as initial training set for the construction of a seed position-specific scoring matrix (PSSM) from the

[illegible]


Filtered IEDB sequences
Derive nested sets
Align 9-residue core sequences based on positions 1, 4, 6, 7, 9
Q=0
15 or more fully aligned seqs
no
yes
Generate seed PSSM (20x9)
10x
Focussing
Refinement
Build PSSM (20x5) using best matches
Weight similarity with template AUC
Pocket similarity matrix
Average 10 PSSMs
Evaluate PSSM (AUC)
Scan pocket templates
Discard PSSM
no
AUC ≥ 70%
yes
Store PSSM
Q > 4
no
Q=Q+1
yes
Final PSSM

[illegible]

PSSM focusing and refinement

After the construction of the initial seed PSSM, three new sequence datasets were generated for each HLA class II molecule. First, all IEDB peptides with annotated positive binding were recovered. 90% of these peptides (randomly chosen) were used to generate a positive-binding training set. The remaining 10% were kept as a positive-binding test set. All IEDB peptides with annotated negative binding were also taken and decomposed into 9-residue segments by running a 9-residue window over the sequence with a 1-residue step. A number of these sequences (randomly chosen) equal to the number of sequences in the positive-binding test set were incorporated into a corresponding negative-binding test set. All remaining segments were then incorporated into a negative-binding training set. The positive and negative-binding test sets were combined into a single test set.

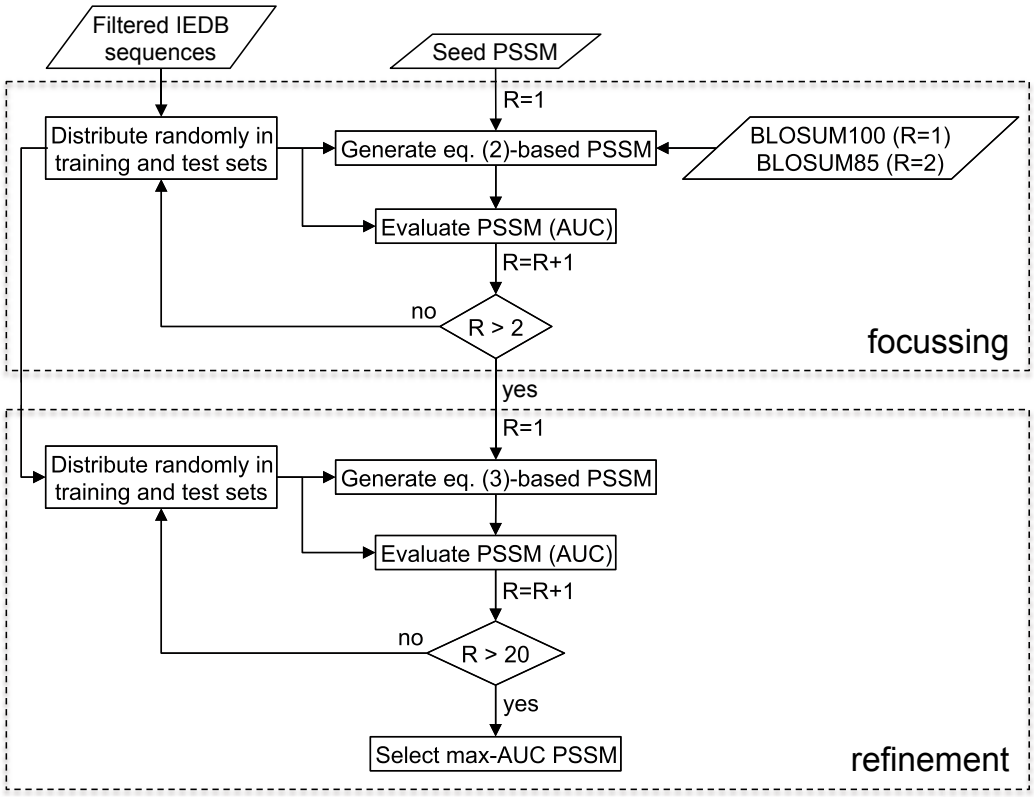
Using these datasets, the seed PSSM was evolved using the following two-phase iterative procedure (Figure 2.2). At each round the three datasets are rebuilt by a new random resampling of peptides.

1. *PSSM focusing*: In this phase, the bias introduced by the use of a small number of core sequences in the construction of the seed PSSM is alleviated by considering potential amino-acid substitutions as given by BLOSUM tables.

First, the seed PSSM is used to evaluate the sequences of the positive-binding training set. All sequences from this set are decomposed into 9-residue segments by running a 9-residue window over the sequence with a 1-residue step. The evaluation of the resulting 9-residue sequences is performed by adding the scores of all nine positions:

$$S_S = \sum_{p=1}^{9} S_{ip} \tag{2.1}$$

where S_s is the segment score and S_{ip} the score of amino acid i in position p as given by the PSSM table. For segments coming from the same original peptide, only the segment with highest score is taken. These segments conform a new working list of putative core sequences.


Filtered IEDB sequences
Seed PSSM
R=1
Distribute randomly in training and test sets
Generate eq. (2)-based PSSM
BLOSUM100 (R=1)
BLOSUM85 (R=2)
Evaluate PSSM (AUC)
R=R+1
no
R > 2
focussing
yes
R=1
Distribute randomly in training and test sets
Generate eq. (3)-based PSSM
Evaluate PSSM (AUC)
R=R+1
no
R > 20
yes
Select max-AUC PSSM
refinement

[illegible]

[illegible]

[illegible]

$$[illegible] = \sum_{[illegible]} [illegible] \qquad (2)$$

[illegible]

To introduce information on non-binding peptides, a negative-binding PSSM is then generated from the frequencies observed in the negative-binding training set and subtracted from the PSSM obtained with eq. (2.2), leading to a new master PSSM.

As indicated, the training and test sets are then rebuilt by a new random distribution of sequences and the procedure is iterated using BLOSUM85. The performance of the PSSM at each round is evaluated by running a prediction over the epitopes of the test set (see *Epitope prediction* below) and calculating the area under the curve (AUC) of the receiver operating characteristic (ROC) curve [37].

2. *PSSM refinement*: This phase consists of 20 rounds of PSSM re-generation. The following equation is used to generate the PSSM at each round, following otherwise the same scheme described for the focusing phase:

$$S_{ip} = \log \frac{f_{ip}^{obs}}{f_i^{bg}} \tag{2.3}$$

Where f_{ip}^{obs} is the observed frequency of amino acid *i* at position *p* and f_i^{bg} is the background frequency of the amino acid in proteins.

Note that the main difference between the focusing and refinement phases is the use of BLOSUM tables in the former to correct for low statistics. Of the 20 rounds of refinement, the PSSM with largest AUC is selected.

The entire procedure (focusing plus refinement) is performed 10 times independently, each of them starting with a different (random) distribution of sequences in the training and test sets. This replicates are performed to limit the dependence of the final results on the initial random distribution. The resulting 10 PSSMs are then averaged to generate a new PSSM. If the AUC of this average PSSM is below 70% it is discarded (i.e. the corresponding HLA class II molecule remains with no associated PSSM), otherwise it is stored (i.e. the PSSM for the given HLA class II molecule is established and will only be modified in later steps if its performance can be improved).

Pocket inheritance

Poorly performing PSSMs may be improved by making use of homology relations at the level of binding pockets, a strategy used with success by other authors to construct so-called virtual matrices [19, 20]. The implemented procedure consists of the following steps.

Data pre-processing: This step involves a one-time analysis of binding pockets for the generation of a similarity matrix. First, the binding pockets P1, P4, P6, P7 and P9 of the three classical HLA class II types, DP, DQ and DR, were analysed for the identification of the residues potentially involved in the interaction with the epitope. The identification was performed manually by inspection of representative three-dimensional structures (PDB entry codes 3LQZ, 1JK8 and 1A6A, for peptide-bound representatives of DP, DQ and DR, respectively; http://www.rcsb.org) using pymol [38]. For DP and DQ both chains (alpha and beta) where analysed, while for DR only the beta chain was considered. The residues with side chain oriented to the pocket cavity and a distance to the peptide's anchoring residue not larger than 7 Å were considered to be directly or indirectly related to binding. To assign the corresponding residues in other molecules, all DP, DQ and DR sequences available in IMGT/HLA [8] were downloaded and multiple alignments were performed with ClustalW2 [36]. After the binding residues had been identified for every pocket of every DP, DQ and DR molecule, all binding pockets of the same type (P1, P4, P6, P7 or P9) were pairwise compared. The comparison was performed using a measure of the physicochemical distance between two pockets. This distance is calculated in a four-dimensional space using the four principal components of a PCA (Principal Component Analysis) of 237 physicochemical properties describing each of the 20 amino acids, as described by Venkatarajan and Braun [39]. A smaller distance between pockets involves a higher physicochemical similarity. The resulting matrix of similarities was stored for use within the PSSM-derivation algorithm.

Implementation, i.e. creating new profiles from existing ones: After the initial refinement phase described above, an attempt to improve the PSSMs is made by

inheritance of pocket profiles (Figure 2.1). To this end, each HLA class II molecule is evaluated for best matches of its pockets with (template) pockets of molecules having an established PSSM. The match is evaluated as the similarity between the query and template pockets (physicochemical distance) weighted by the performance of the PSSM of the molecule that contains the template pocket. The smaller the quotient between distance and template AUC the better the match. In other words, at equal similarity, inheritance will proceed from the best-performing PSSM. Best-matches for the five pockets of a given HLA class II molecule can be then used to construct a five-column PSSM for this molecule by inheritance of the corresponding columns from the PSSMs of the template HLAs. This PSSM is then re-submitted to the refinement phase, where it develops again into a 9-column PSSM. This pocket inheritance procedure is iterated three additional times (Figure 2.1) two enable the use of rescued molecules (with a PSSM overcoming the 70% threshold) to rescue further molecules.

Epitope prediction

Epitope prediction is performed after each round of PSSM focusing and refinement and is also available for online queries. To this end, the peptides in the test set, when deriving a PSSM, or the user input sequences, when dealing with an online query, are decomposed into 9-residue segments by running a 9-residue window over each sequence with a 1-residue step. The evaluation of the 9-residue segments is then performed by adding the scores of all nine positions as given by the PSSM (eq. (2.1)). In the case of an online query, the service presents as output all segments with a score S_S above a threshold value whose default corresponds to the point of the ROC curve where the difference between specificity and sensitivity is minimal. This tries to reflect the experimental situation, where a sequence can in principle present different binding motifs with different affinities. When deriving a PSSM, the segment scores are combined into a peptide score (for each peptide in the test set) for further evaluation of the AUC of the PSSM. The peptide score S_P is calculated using an empirical equation chosen to optimize the ROC (see Chapter 4):

$$S_P = \frac{9}{L}\left(1 + \sum_{n=1}^{L-8} \lambda^{S_S}\right) \quad (2.4)$$

where L is the peptide length, 9 refers to the segment length, the sum runs over the number of 9-residue segments generated from the peptide (L-8) and λ (=3) is a fitted parameter.

Maintenance, growth and stability of the database

All steps of the algorithm described (Figure 2.1) have been automatised. This enables both straightforward updates of HLA2db following relevant updates of the parent databases (i.e. Uniprot [34] and HPRD [35] for proteomes and IEDB [13] for experimentally determined epitopes) and the unassisted generation or improvement of PSSMs using epitope sequences uploaded by the user. The derivation of a pocket-similarity matrix, which is the only manual step, should in principle need no regular updates. Potential changes of format of the information captured from the parent databases will be identified by implemented read-time checks, and shall require limited re-programing of the corresponding routines.

The stability of the database and the prediction tool has been particularly surveilled. At the data-integrity level, the potential corruption of PSSMs by incorporation of poor quality data is avoided by *i)* the use of stringent criteria in the initial filtering of uploaded epitope sequences (see *Definition of an initial binding profile* above), and *ii)* the condition that only those data that improve the ROC curve for an established PSSM are finally taken into account. At the execution level, both stability and speed are favoured by the pre-processing of all proteome sequences available within HLA2db.

Results and discussion

HLA2db is a PSSM-based, unsupervised, online system for the identification of HLA class II binding motifs in polypeptide sequences. Its functionality and distinguishing elements are discussed here.

Query sequences

A main functionality of HLA2db is the identification and scoring of potential epitopes in query sequences entered by the user in relation to a selected HLA class II molecule. The sequences may be entered online or uploaded as a file in FASTA format [32]. A sequence may contain up to 3 asterisks (*) as a wildcard indicating that any amino acid can be present at the specified positions. A distinguishing characteristic of this tool, relative to other available servers, is the possibility to enter any number of sequences of any length. For example, a full proteome may be uploaded for a one-shot analysis against a specific HLA class II molecule or, alternatively, against all available molecules (see *Immunome calculation* below).

Available HLA class II molecules

In the current release, HLA2db is able to evaluate sequences for binding to 31 different HLA class II molecules (18 HLA-DR, 8 HLA-DQ and 5 HLA-DP) with a reasonably high predictive power (Table 2.2). The criterion used to flag an HLA molecule as available is that the prediction performance of its associated PSSM must be characterised by a ROC curve with an AUC value equal or higher than 70%. Lower values generally originate from an insufficient number of known epitope sequences for the derivation of a profile describing the molecule's binding preferences or to the presence of noise in the data, due to experimental uncertainties or miss-annotations in the reference epitope database. The inherent capacity of HLA class II molecules to associate to large peptide repertoires particularly stresses the requirement of sufficiently representative peptide sets for the training of any type of predictor. The (rapid) growth of the amount of epitope

data at IEDB should translate in both the improvement of HLA2db's predictive power for molecules already available and the incorporation of new molecules flagged as available.

A very important aspect of the implemented algorithm is the possibility to use the profiles derived for HLA molecules for which there is sufficient peptide-binding data available to derive the profiles for additional molecules by exploiting binding-pocket homologies. In the current release, this feature has enabled the inclusion of 17 of the total 31 molecules available.

When performing a query, the selection may include a single HLA molecule, all available molecules individually, or all molecules combined. The first option outputs the 9-residue segments from the query sequences that are predicted to be epitopes for the selected HLA molecule. Although a molecule-dependent default value for the score cut-off is suggested (corresponding to the point of the ROC curve where the difference between specificity and sensitivity is minimal) the user may choose a different threshold. To facilitate the choice, the minimum and maximum scores are also indicated. Higher values correspond to a higher specificity (fewer false positives) and a lower sensitivity (more false negatives). The second option provides an equivalent output including the predictions for all HLA molecules available in the database. In this case, the choice of the score cut-off is limited to using default values, values corresponding to 90% specificity or values corresponding to 90% sensitivity. The third option provides an alternative output, consisting on the number of HLA molecules that recognise each predicted epitope, enabling a rapid assessment of epitope promiscuity. The choice of the score cut-off is in this case also limited to default, 90% specificity or 90% sensitivity.

Table 2.2. Statistics for the set of HLA class II molecules available in HLA2db.

Molecule	AUC	Score threshold	Sensitivity	Specificity	Accuracy	PPV	NPV
HLA-DPA1*01:03/DPB1*02:01	0.87	2.99	0.80	0.80	0.80	0.85	0.73
HLA-DPA1*02:01/DPB1*01:01	0.87	3.10	0.83	0.82	0.83	0.91	0.69
HLA-DPA1*02:01/DPB1*02:01	0.84	4.10	0.88	0.88	0.88	0.97	0.60
HLA-DPA1*02:01/DPB1*05:01	0.85	3.29	0.79	0.79	0.79	0.72	0.85
HLA-DPA1*03:01/DPB1*04:02	0.88	3.35	0.84	0.83	0.83	0.86	0.80
HLA-DQA1*01:01/DQB1*05:01	0.85	3.30	0.77	0.78	0.78	0.76	0.79
HLA-DQA1*01:02/DQB1*06:02	0.83	3.33	0.78	0.80	0.78	0.93	0.49
HLA-DQA1*03:01/DQB1*03:02	0.81	3.21	0.75	0.74	0.74	0.83	0.64
HLA-DQA1*04:01/DQB1*04:02	0.82	3.26	0.77	0.77	0.77	0.92	0.50
HLA-DQA1*05:01/DQB1*02:01	0.76	2.96	0.70	0.71	0.71	0.84	0.52
HLA-DQA1*05:01/DQB1*03:01	0.80	3.17	0.74	0.73	0.74	0.89	0.48
HLA-DQA1*05:01/DQB1*03:02	0.74	3.37	0.70	0.69	0.70	0.69	0.71
HLA-DQA1*05:01/DQB1*04:01	0.80	3.60	0.82	0.82	0.82	0.93	0.59
HLA-DRB1*01:01	0.71	3.72	0.66	0.66	0.66	0.88	0.35
HLA-DRB1*03:01	0.74	3.29	0.68	0.68	0.68	0.70	0.66
HLA-DRB1*04:01	0.75	3.51	0.69	0.69	0.69	0.79	0.56
HLA-DRB1*04:02	0.80	3.40	0.74	0.74	0.74	0.64	0.82
HLA-DRB1*04:04	0.82	3.31	0.75	0.75	0.75	0.86	0.60
HLA-DRB1*04:05	0.74	3.42	0.68	0.68	0.68	0.87	0.40
HLA-DRB1*07:01	0.75	3.43	0.69	0.69	0.69	0.82	0.51

Molecule	AUC	Score threshold	Sensitivity	Specificity	Accuracy	PPV	NPV
HLA-DRB1*08:02	0.73	3.26	0.67	0.67	0.67	0.79	0.52
HLA-DRB1*09:01	0.74	3.15	0.69	0.68	0.69	0.88	0.40
HLA-DRB1*11:01	0.77	3.57	0.71	0.71	0.71	0.80	0.59
HLA-DRB1*12:01	0.77	3.31	0.70	0.70	0.70	0.77	0.62
HLA-DRB1*13:01	0.82	3.61	0.75	0.75	0.75	0.69	0.80
HLA-DRB1*13:02	0.75	3.33	0.69	0.70	0.69	0.84	0.50
HLA-DRB1*15:01	0.76	3.25	0.68	0.69	0.68	0.83	0.50
HLA-DRB3*01:01	0.74	3.36	0.68	0.68	0.68	0.71	0.66
HLA-DRB3*02:02	0.85	3.22	0.80	0.80	0.80	0.84	0.75
HLA-DRB4*01:01	0.70	3.58	0.65	0.65	0.65	0.82	0.43
HLA-DRB5*01:01	0.75	3.22	0.69	0.68	0.68	0.85	0.44

The statistics have been evaluated using the entire set of epitope sequences (see Table 2.1).

AUC: area under the receiver-operating-characteristic (ROC) curve; gives the probability that the PSSM will rank a randomly chosen positive binder higher than a randomly chosen negative one.

Score threshold: score separating binders from non binders; corresponds to the point of the ROC curve where the difference between sensitivity and specificity is minimal; all following quantities refer to this point.

Sensitivity: number of true positives relative to the sum of true positives and false negatives; high sensitivity indicates few false negatives.

Specificity: number of true negatives relative to the sum of false positives and true negatives; high specificity indicates few false positives.

Accuracy: sum of true positives and true negatives relative to the total.

Positive predictive value (PPV) or precision: number of true positives relative to the sum of true and false positives.

Negative predictive value (NPV): number of true negatives relative to the sum of true and false negatives.

Scanning of proteomes

A distinguishing aspect of HLA2db is the possibility to scan a database of bacterial proteomes (currently from a list of 243) and the human proteome for matches of the binding motifs identified within the query sequences. This is simply achieved by selecting the proteome of interest. In addition to the list of predicted epitopes and their scores, the output will then indicate the proteins of the selected proteome that contain each of the binding motifs identified (positions 1, 4, 6, 7 and 9 of an epitope). A link to the corresponding Uniprot entry is given for the motif-containing protein.

Integrating user data to generate a new profile or improve an existing one

By logging into the server, the user may also upload private, experimentally validated peptide sequences for the automatic generation / improvement of the PSSM for a specific HLA molecule. The system performs its own checks to ensure that the new sequences make a valid and significant contribution to the PSSM (see Materials and methods) and, if so, the new PSSM is kept for the user's private use in successive queries.

Three potential situations are envisaged. In the first one, the user would upload additional sequences for a molecule already available in HLA2db. These sequences would then be combined with the IEDB sequences used to generate the public PSSM to create a new data set for further PSSM development. Nevertheless, a new PSSM would only be established if it showed a performance superior to that of the public one. In the second situation, the user would upload additional sequences for a molecule present at HLA2db but not available, i.e. with AUC < 70%. As in the previous case, the user sequences would be combined with the IEDB ones for further PSSM development. A new PSSM would in this case be established only if it presented an AUC ≥ 70% at the end of the process. In the third situation, the user would upload a significant number of sequences for a molecule not present in

HLA2db (i.e. neither at IEDB). In such case PSSM development would proceed from scratch with the user's data set alone, subject to the same procedures and criteria used for the public ones.

The user may still decide that a PSSM improved with his/her private data can be made publicly available. This can be done by contacting the server manager at the address provided.

Additional services available to the registered user are the possibility to consult various data relative to the profile, including the PSSM table, a WebLogo representation [40] and performance statistics (incl. the ROC curve).

Immunome calculation

The immunome is described as the set of epitopes derived from a proteome (human or pathogen) that are presented to the host immune system in the context of MHC class I and class II molecules or that engage antibodies, eliciting an immune response [41]. Following this idea, HLA2db enables the evaluation of all possible 9-residue segments obtained from a full fragmentation of a proteome, scoring them against the complete list of HLA class II molecules available. This is achieved by uploading a FASTA file containing the full proteome of interest and selecting the "All molecules individually" or "All molecules combined" options. The output will then provide the list of 9-residue segments from the proteome that are predicted to bind at least one of the available HLA molecules. The output format will depend on the HLA-selection option, as explained above, and the number of epitopes and confidence of the prediction will depend on the chosen score cut-off, as also explained. If the user chooses the same species for proteome scanning, the output will, in addition, indicate the protein (with link to Uniprot) to which each of the predicted epitopes belongs.

HLA2db in the context of other predictors

The statistics provided in Table 2.2 may only be taken as an internal test. Proper comparison to the performance of other available predictors requires an independent test set with a significant number of sequences per HLA class II molecule (independent meaning that it has not been used in the training of the predictors). Evaluations of predictors using fresh data sets (before deposition in databases) have been recently performed, for example, by Wang and collaborators [26, 27]. As they point out, the limited number of available epitope sequences for most HLA class II molecules makes it unwise to discard part of the data when training the algorithms with the purpose of having and independent test set. During the development of the PSSMs, we have used a strategy based on random resampling of the epitope-sequence dataset for the generation of the training and test sets at each training/evaluation step in the algorithm. The evaluation of the final PSSMs, which statistics are reported in Table 2.2, is then performed using the full set o epitope sequences. Therefore, Table 2.2 provides an indication of the capacity of our predictor to explain the data on which it has been trained.

IEDB has become a standard reference repository of epitope sequences for HLA class II molecules. One should therefore expect the predictor described here to have a performance similar to other PSSM-based predictors that use this repository as reference and somewhat lower than predictors based on artificial intelligence algorithms [27]. Nevertheless, the HLA2db server offers a number of functionalities, highlighted above, that are not standardly found in other servers and may be relevant to the immunologist.

References

1. Hammer J, Sturniolo T, Sinigagua F (1997) HLA Class II Peptide Binding Specificity and Autoimmunity. In: Frank J. Dixon, editor. Advances in Immunology. Academic Press, Vol. 66. pp. 67–100.

2. Rammensee HG, Friede T, Stevanoviíc S (1995) MHC ligands and peptide motifs: first listing. Immunogenetics 41: 178–228. doi:10.1007/BF00172063.

3. Bondinas GP, Moustakas AK, Papadopoulos GK (2007) The spectrum of HLA-DQ and HLA-DR alleles, 2006: a listing correlating sequence and structure with function. Immunogenetics 59: 539–553. doi:10.1007/s00251-007-0224-8.

4. Schirle M, Weinschenk T, Stevanović S (2001) Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. Journal of Immunological Methods 257: 1–16. doi:10.1016/S0022-1759(01)00459-8.

5. Purcell AW, Gorman JJ (2004) Immunoproteomics Mass Spectrometry-based Methods to Study the Targets of the Immune Response. Mol Cell Proteomics 3: 193–208. doi:10.1074/mcp.R300013-MCP200.

6. Brusic V, Rudy G, Harrison LC (1998) MHCPEP, a database of MHC-binding peptides: Update 1997. Nucl Acids Res 26: 368–371. doi:10.1093/nar/26.1.368.

7. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50: 213–219.

8. Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SGE (2001) IMGT/HLA Database — a sequence database for the human major histocompatibility complex. Nucleic Acids Res 29: 210–213. doi:10.1093/nar/29.1.210.

9. Schönbach C, Koh JLY, Flower DR, Wong L, Brusic V (2002) FIMM, a database of functional molecular immunology: update 2002. Nucl Acids Res 30: 226–229. doi:10.1093/nar/30.1.226.

10. Sathiamurthy M, Hickman H d., Cavett J w., Zahoor A, Prilliman K, et al. (2003) Population of the HLA Ligand Database. Tissue Antigens 61: 12–19. doi:10.1034/j.1399-0039.2003.610102.x.

11. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, Flower DR (2003) JenPep: A Novel Computational Information Resource for Immunobiology and Vaccinology. J Chem Inf Comput Sci 43: 1276–1287. doi:10.1021/ci030461e.

12. Lata S, Bhasin M, Raghava GP (2009) MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. BMC Research Notes 2: 61. doi:10.1186/1756-0500-2-61.

13. Salimi N, Fleri W, Peters B, Sette A (2012) The immune epitope database: a historical retrospective of the first decade. Immunology 137: 117–123. doi:10.1111/j.1365-2567.2012.03611.x.

14. Lafuente E, Reche P (2009) Prediction of MHC-Peptide Binding: A Systematic and Comprehensive Overview. Current Pharmaceutical Design 15: 3209–3220. doi:10.2174/138161209789105162.

15. Tong JC, Tan TW, Ranganathan S (2007) Methods and protocols for prediction of immunogenic epitopes. Brief Bioinform 8: 96–108. doi:10.1093/bib/bbl038.

16. Reche PA, Glutting J-P, Zhang H, Reinherz EL (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics 56: 405–419. doi:10.1007/s00251-004-0709-7.

17. Bui H-H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, et al. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics 57: 304–314. doi:10.1007/s00251-005-0798-y.

18. Nielsen M, Lundegaard C, Lund O (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics 8: 238. doi:10.1186/1471-2105-8-238.

19. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, et al. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA

microarrays and virtual HLA class II matrices. Nat Biotech 17: 555–561. doi:10.1038/9858.

20. Singh H, Raghava GPS (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17: 1236–1237. doi:10.1093/bioinformatics/17.12.1236.

21. Doytchinova IA, Blythe MJ, Flower DR (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: 263–272. doi:10.1021/pr015513z.

22. Guan P, Doytchinova IA, Zygouri C, Flower DR (2003) MHCPred: a server for quantitative prediction of peptide–MHC binding. Nucl Acids Res 31: 3621–3624. doi:10.1093/nar/gkg510.

23. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucl Acids Res 33: W172–W179. doi:10.1093/nar/gki452.

24. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics 10: 296. doi:10.1186/1471-2105-10-296.

25. Dönnes P, Kohlbacher O (2006) SVMHC: a server for prediction of MHC-binding peptides. Nucl Acids Res 34: W194–W197. doi:10.1093/nar/gkl284.

26. Wang P, Sidney J, Dow C, Mothé B, Sette A, et al. (2008) A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach. PLoS Comput Biol 4: e1000048. doi:10.1371/journal.pcbi.1000048.

27. Wang P, Sidney J, Kim Y, Sette A, Lund O, et al. (2010) Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics 11: 568. doi:10.1186/1471-2105-11-568.

28. Khan JM, Ranganathan S (2010) pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. Immunome Res 6: S2. doi:10.1186/1745-7580-6-S1-S2.

29. Tong JC, Zhang GL, Tan TW, August JT, Brusic V, et al. (2006) Prediction of HLA-DQ3.2β Ligands: evidence of multiple registers in class II binding peptides. Bioinformatics 22: 1232–1238. doi:10.1093/bioinformatics/btl071.

30. Bordner AJ (2010) Towards Universal Structure-Based Prediction of Class II MHC Epitopes for Diverse Allotypes. PLoS ONE 5: e14383. doi:10.1371/journal.pone.0014383.

31. Muixí L, Carrascal M, Alvarez I, Daura X, Martí M, et al. (2008) Thyroglobulin Peptides Associate In Vivo to HLA-DR in Autoimmune Thyroid Glands. J Immunol 181: 795–807.

32. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. PNAS 85: 2444–2448.

33. Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, et al. (2011) PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. Infect Immun 79: 4286–4298. doi:10.1128/IAI.00207-11.

34. The UniProt Consortium (2011) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research 40: D71–D75. doi:10.1093/nar/gkr981.

35. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database—2009 update. Nucl Acids Res 37: D767–D772. doi:10.1093/nar/gkn892.

36. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680. doi:10.1093/nar/22.22.4673.

37. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861–874. doi:10.1016/j.patrec.2005.10.010.

38. The PyMOL Molecular Graphics System, Version 1.2r3pre (n.d.). Schrödinger, LLC.

39. Venkatarajan MS, Braun W (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. J Mol Model 7: 445–453. doi:10.1007/s00894-001-0058-5.

40. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: A Sequence Logo Generator. Genome Res 14: 1188–1190. doi:10.1101/gr.849004.

41. De Groot AS (2006) Immunomics: discovering new targets for vaccines and therapeutics. Drug Discovery Today 11: 203–209. doi:10.1016/S1359-6446(05)03720-7.

Chapter 3. A Support Vector Machine for the prediction of MHC class II epitopes based on amino-acid distances

Abstract

The identification of MHC class II epitopes is a fundamental step in many studies of immunological processes, including immune responses to infection and undesired processes such as transplant rejection and the development of autoimmunity. Identifying MHC class II epitopes experimentally is, however, time and resource consuming and computational approaches have emerged as a powerful prediction aid. Many methods have been used to this end, some of the most successful being based on support vectors machines (SVMs). Building on our previous work on HLA class II epitope prediction, we have developed sdHLa2, and algorithm and web server that combines that makes use of position-specific scoring matrices (PSSMs) to develop SVMs. The sdHLA2 web server has been designed to: a) maximise the versatility of the queries, with no restrictions on the amount or length of input sequences to be evaluated for epitope prediction, and b) enable the scanning of over 240 microbial and the human proteome for predicted motifs. sdHLA2 is freely accessible at http://bioinf.uab.cat/sdhla2

Introduction

The proteins coded by the major histocompatibility complex (MHC) class II –human leukocyte antigen or HLA class II in human– play an important role in the immune response to infection by presenting exogenous epitopes to CD4+ T lymphocytes, but have also a major participation in undesired processes such as those leading to autoimmunity [1]. Peptide presentation by HLA class II molecules has therefore implications in both health and disease, and understanding and being able to predict the affinity of HLA class II proteins for specific peptides has been one of the priorities of the field. Recent advances in high-throughput immunoproteomics methods [2,3] have allowed the identification of large sets of epitopes now available in databases [4–10] and which can be used to train different prediction services [11–20].

Here we present sdHLA, an unsupervised, online system for the identification of HLA class II binding motifs in polypeptide sequences using a support vector machine. Its functionality includes the identification and scoring of potential epitopes in query sequences entered by the user (with no limitation in number of sequences or their length) in relation to the selected HLA class II molecule (18 HLA-DR, 8 HLA-DQ and 5 HLA-DP molecules are currently available). The selection may include a single HLA molecule, all sdHLA2 available molecules individually, or all molecules combined, the latter case providing the number of HLA molecules recognising each predicted epitope. It can also scan microbial proteomes (currently 243) and the human proteome for matches of the binding motifs identified within the input sequences. A maintenance system has been also set up, such that each time the internal database is updated with the incorporation of new information from public peptide libraries the SVMs for all molecules are automatically recalculated. By doing this, the quality of the predictions shall progressively improve.

The service has been implemented to support vaccine and autoimmunity / tolerance studies, allowing the screening of both microbial and human proteomes for the presence of predicted HLA class II epitopes and providing information on the level of promiscuity of these epitopes by performing the analysis against the complete set of HLA molecules available.

Implementation

Data collection

Epitopes were downloaded from the Immune Epitope Database (IEDB) [10]. They were then filtered and classified as positive and negative binders according to IEDB annotations and following the criteria described in Chapter 2. For positive binders, the 9-residue peptide core (binding segment) was determined using the PSSMs developed previously (see Chapter 2). Negative binders were fragmented

into 9-residue segments by running a 9-residue window over the sequence with a 1-residue step. For each HLA class II molecule, positive and negative binders were distributed randomly between a test set (10% of positive binders and the same number of negative binders) and a training set (all positive and negative binders not contained in the test set).

SVM generation

A Support Vector Machine [21] is a supervised-learning algorithm that separates different populations in a sample by constructing a hyperplane or set of hyperplanes in a high-dimensional space. Is a type of linear classifier, i.e. makes the classification decision based on the value of a linear combination of the characteristics (features) of the object to be classified, which are presented to the machine in a vector called a feature vector. A fundamental property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin (minimum distance between the objects and the hyperplane). Objects on the margin are called the support vectors. As the sample is often not linearly separable in the original space, SVMs make use of the so-called kernel trick, by which the original space is (generally non-linearly) transformed into a new (much higher-dimensional) space in which the data can be classified linearly. The trick consists on using an algorithm that only requires dot products between the vectors in the new space and choosing the mapping such that these high-dimensional dot products can be computed within the original space by means of a kernel function. Common non-linear functions include the polynomial, sigmoidal and radial-basis-function (RBF) kernels, the latter being the most popular. The effectiveness of an SVM depends on the selection of kernel, the kernel's parameters, and the soft-margin parameter C, which is determines the trade-off between misclassification and simplicity of the decision surface.

A SVM was generated for each HLA class II molecule using the SVM light PERL module [22]. The amino acids at positions 1 to 9 of the peptide core were codified using the first four vectors of a principal-components analysis (PCA) by Venkatarajand and Braun [23], which describes the 20 amino acids using 237

different physico-chemical features. The SVM server was trained in two different steps. First, linear, sigmoidal, polynomial and RBF kernels were tested by using a bench of parameters with the HLA-DRB1*04:01 molecule as test case, as explained in [24]. Performance was measured using the receiver operating characteristic (ROC) curve [25]. The best performance was reached with the RBF kernel, which was the used to train the SVMs for the remaining molecules. The RBF kernel has a single parameter called gamma (γ), which weights the influence of each object in the final model. Although the value of γ is ideally −1/n, it is often selected together with C by a grid search with exponentially growing sequences. All parameter combinations were evaluated using a ROC curve and those with the best performance were chosen for implementation in the web server (see statistics in Table 3.1). As discussed in Chapter 2, the statistics provided in Table 3.1 may only be taken as an internal test. Proper comparison to the performance of other available predictors requires the availability of an independent test set with a significant number of sequences per HLA class II molecule.

Table 3.1. Statistics for the set of HLA class II molecules available in sdHLA2.

Molecule	AUC	Score threshold	Sensitivity	Specificity	Accuracy	PPV	NPV	C	γ
HLA-DPA1*01:03/DPB1*02:01	0.97	-0.06	0.97	0.97	0.97	0.98	0.95	8	3
HLA-DPA1*02:01/DPB1*01:01	0.97	0.40	0.96	0.96	0.96	0.98	0.91	5	2
HLA-DPA1*02:01/DPB1*02:01	0.85	1.00	0.97	0.92	0.96	0.98	0.85	8	1
HLA-DPA1*02:01/DPB1*05:01	0.98	0.00	0.97	0.92	0.94	0.89	0.98	9	6
HLA-DPA1*03:01/DPB1*04:02	0.97	-3.01	0.96	0.97	0.96	0.97	0.95	6	7
HLA-DQA1*01:01/DQB1*05:01	0.98	-0.42	0.94	0.94	0.94	0.94	0.95	6	2
HLA-DQA1*01:02/DQB1*06:02	0.96	1.00	0.98	0.96	0.98	0.99	0.94	6	2
HLA-DQA1*03:01/DQB1*03:02	0.98	0.13	0.96	0.96	0.96	0.97	0.93	7	2
HLA-DQA1*04:01/DQB1*04:02	0.96	0.66	0.96	0.96	0.96	0.99	0.88	5	1
HLA-DQA1*05:01/DQB1*02:01	0.97	0.07	0.96	0.96	0.96	0.98	0.91	5	6
HLA-DQA1*05:01/DQB1*03:01	0.97	0.22	0.97	0.97	0.97	0.99	0.92	10	2
HLA-DQA1*05:01/DQB1*03:02	0.97	-0.49	0.95	0.95	0.95	0.95	0.95	5	1
HLA-DQA1*05:01/DQB1*04:01	0.89	-0.73	0.95	0.95	0.95	0.98	0.86	10	6
HLA-DRB1*01:01	0.97	1.00	0.96	0.96	0.96	0.99	0.88	9	10
HLA-DRB1*03:01	0.98	0.00	0.97	0.95	0.96	0.96	0.97	7	8
HLA-DRB1*04:01	0.98	0.00	0.95	0.97	0.96	0.98	0.92	9	9
HLA-DRB1*04:02	0.97	-0.31	0.96	0.96	0.96	0.94	0.98	4	2
HLA-DRB1*04:04	0.98	0.02	0.98	0.98	0.98	0.99	0.96	4	8
HLA-DRB1*04:05	0.98	0.45	0.97	0.97	0.97	0.99	0.91	2	3
HLA-DRB1*07:01	0.98	0.00	0.96	0.96	0.96	0.98	0.92	8	7

Molecule	AUC	Score threshold	Sensitivity	Specificity	Accuracy	PPV	NPV	C	γ
HLA-DRB1*08:02	0.99	-0.01	0.95	0.96	0.96	0.98	0.92	9	4
HLA-DRB1*09:01	0.98	0.82	0.97	0.97	0.97	0.99	0.91	5	3
HLA-DRB1*11:01	0.98	0.00	0.96	0.96	0.96	0.98	0.94	3	5
HLA-DRB1*12:01	0.98	-0.09	0.96	0.96	0.96	0.97	0.94	3	3
HLA-DRB1*13:01	0.96	0.00	0.94	0.93	0.94	0.92	0.95	9	4
HLA-DRB1*13:02	0.98	0.00	0.97	0.97	0.97	0.98	0.93	7	8
HLA-DRB1*15:01	0.98	0.00	0.96	0.96	0.96	0.98	0.91	9	7
HLA-DRB3*01:01	0.98	0.00	0.95	0.95	0.95	0.96	0.95	5	6
HLA-DRB3*02:02	0.98	0.05	0.97	0.99	0.98	0.99	0.96	6	7
HLA-DRB4*01:01	0.97	1.00	0.97	0.97	0.97	0.99	0.92	3	10
HLA-DRB5*01:01	0.98	0.09	0.96	0.96	0.96	0.99	0.91	8	6

The statistics have been evaluated using the entire set of epitope sequences (see Table 2.1).

AUC: area under the receiver-operating-characteristic (ROC) curve; gives the probability that the SVM will rank a randomly chosen positive binder higher than a randomly chosen negative one.

Score threshold: score separating binders from non binders; corresponds to the point of the ROC curve where the difference between sensitivity and specificity is minimal; all following quantities refer to this point.

Sensitivity: number of true positives relative to the sum of true positives and false negatives; high sensitivity indicates few false negatives.

Specificity: number of true negatives relative to the sum of false positives and true negatives; high specificity indicates few false positives.

Accuracy: sum of true positives and true negatives relative to the total.

Positive predictive value (PPV) or precision: number of true positives relative to the sum of true and false positives.

Negative predictive value (NPV): number of true negatives relative to the sum of true and false negatives.

C: soft-margin parameter of the SVM; γ: single parameter of the RBF kernel.

Conclusion

As explained above, our method uses a codification based on amino-acid descriptors, corresponding to a series of physical and chemical characteristics of each amino acid, which prevents the system from over-learning (as compared to binary descriptions of amino acids). Moreover, the performance of the method will rapidly improve as the number and quality of the peptides found at repositories increases, as SVM performance is known to depend heavily on the size of the training set. As shown in Table 3.1, the SVMs developed here have a very high capacity to recognise patterns already seen during training, relative to the PSSM-based predictor described in Chapter 2, which uses the same pool of training and test peptides. As already discussed in Chapter 2, proper statistical validation will however require the availability of new independent peptide sets. Besides the SVM being intrinsically a more powerful method than that based on PSSMs, in this case SVM development profited from the knowledge already captured by the PSSMs, as the output from the first method was used to feed the training set of 9-residue peptides for the second one.

References

1. Hammer J, Sturniolo T, Sinigaglia F (1997) HLA class II peptide binding specificity and autoimmunity. Adv Immunol 66: 67–100.

2. Schirle M, Weinschenk T, Stevanović 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. doi:10.1016/S0022-1759(01)00459-8.

3. Purcell AW, Gorman JJ (2004) Immunoproteomics: Mass spectrometry-based methods to study the targets of the immune response. Mol Cell Proteomics 3: 193–208. doi:10.1074/mcp.R300013-MCP200.

4. Brusic V, Rudy G, Harrison LC (1998) MHCPEP, a database of MHC-binding peptides: Update 1997. Nucleic Acids Res 26: 368–371. doi:10.1093/nar/26.1.368.

5. Rammensee H-G, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50: 213–219. doi:10.1007/s002510050595.

6. Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SGE (2001) IMGT/HLA Database—a sequence database for the human major histocompatibility complex. Nucleic Acids Res 29: 210–213. doi:10.1093/nar/29.1.210.

7. Schönbach C, Koh JLY, Flower DR, Wong L, Brusic V (2002) FIMM, a database of functional molecular immunology: update 2002. Nucleic Acids Res 30: 226–229. doi:10.1093/nar/30.1.226.

8. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, Flower DR (2003) JenPep: a novel computational information resource for immunobiology and vaccinology. J Chem Inf Comput Sci 43: 1276–1287. doi:10.1021/ci030461e.

9. Lata S, Bhasin M, Raghava GPS (2009) MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. BMC Res Notes 2: 61. doi:10.1186/1756-0500-2-61.

10. Salimi N, Fleri W, Peters B, Sette A (2012) The immune epitope database: a historical retrospective of the first decade. Immunology 137: 117–123. doi:10.1111/j.1365-2567.2012.03611.x.

11. Reche PA, Glutting J-P, Zhang H, Reinherz EL (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics 56: 405–419. doi:10.1007/s00251-004-0709-7.

12. Bui H-H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, et al. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics 57: 304–314. doi:10.1007/s00251-005-0798-y.

13. Nielsen M, Lundegaard C, Lund O (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics 8: 238. doi:10.1186/1471-2105-8-238.

14. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, et al. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol 17: 555–561. doi:10.1038/9858.

15. Singh H, Raghava GPS (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17: 1236–1237. doi:10.1093/bioinformatics/17.12.1236.

16. Guan P, Doytchinova IA, Zygouri C, Flower DR (2003) MHCPred: a server for quantitative prediction of peptide–MHC binding. Nucleic Acids Res 31: 3621–3624. doi:10.1093/nar/gkg510.

17. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic Acids Res 33: W172–W179. doi:10.1093/nar/gki452.

18. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics 10: 296. doi:10.1186/1471-2105-10-296.

19. Dönnes P, Kohlbacher O (2006) SVMHC: a server for prediction of MHC-binding peptides. Nucleic Acids Res 34: W194–197. doi:10.1093/nar/gkl284.

20. Khan JM, Ranganathan S (2010) pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. Immunome Res 6 Suppl 1: S2. doi:10.1186/1745-7580-6-S1-S2.

21. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20: 273-297.

22. Joachims T (1999) Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. Schölkopf B, Burges C and Smola A (Eds.), MIT-Press.

23. Venkatarajan MS, Braun W (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. Mol Model Annu 7: 445–453. doi:10.1007/s00894-001-0058-5.

24. Hsu CL, Chang c (2003) A Practical Guide to Support Vector Classication. Available: http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf.

25. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27: 861–874. doi:10.1016/j.patrec.2005.10.010.

Chapter 4. Discussion

Coherence of the original epitope dataset

Data extraction, manipulation and analysis are key points to any project attempting to make predictions by training algorithms on existing data. Given the high amount of data used in the work presented here, it would be clearly impossible to check this data manually before using it to train an algorithm for the prediction of HLA class II presentation. Yet, scientists working with databases are generally aware of the existence of miss-annotated or incoherent data in these repositories. Indeed at the meeting of the European Federation of Immunological Societies in Berlin (September 2009), Dr A. Sette, from Immune Epitope Database (IEDB), requested the collaboration of researchers to correct possible errors in this database.

Contradictory information is often found in databases as a result of difficulties in comparing the different types of measurements used to generate the data. In the case of HLA class II peptide presentation this mostly arises from the use of different techniques to measure binding (Table 4.1), potentially leading to contradictory conclusions for the same HLA class II – epitope pair. Mistakes can also result from typos introduced at the time of incorporating new data to the database. We have estimated the percentage of miss-annotated peptides at IEDB to be around 10% by revising all HLA-DRB1*04:01 peptides contained in this database. The most common mistake detected is an unclear relation between the given quantitative cut-off and the qualitative assignment of peptide presentation. Some examples are given in Table 4.2.

Another source of mistakes may result from sample miss-definition. HLA nomenclature was thought to name genes instead of proteins, emphasizing silent mutations. For this reason, some times a single peptide may seem to be presented by different HLA molecules that in fact are the same, resulting in the decrease of sample size and limiting any statistical test performance. In the present study this problem has been addressed by considering only information related to the protein (gene locus, serologic family and codifying allele) and discarding the rest.

Table 4.1. Different test and measurement units used by HLA class II studies [1].

Assay Type	Assay Response	Assay Units
Cell bound MHC - Fluorescence	Association (or direct binding)	EC50 nM
Cell bound MHC - Fluorescence	Association (or direct binding)	Ka M^{-1}
Cell bound MHC - Fluorescence	Association (or direct binding)	Kon (nM^{-1} s^{-1})
Cell bound MHC - Fluorescence	Association (or direct binding)	t1/2 (min)
Cell bound MHC - Fluorescence	Competition (or equilibrium binding)	IC50 nM
Cell bound MHC - Fluorescence	Competition (or equilibrium binding)	Kd nM
Cell bound MHC - Fluorescence	Dissociation	Koff (s^{-1})
Cell bound MHC - Fluorescence	Dissociation	t1/2 (min)
Cell bound MHC - Radioactivity	Competition (or equilibrium binding)	IC50 nM
Cell bound MHC - Radioactivity	Competition (or equilibrium binding)	Kd nM
Cell bound MHC - Radioactivity	Dissociation	t1/2 (min)
Cell bound MHC - T cell response	Competition (or equilibrium binding)	IC50 nM
Lysate - Radioactivity	Association (or direct binding)	EC50 nM
Lysate - Radioactivity	Competition (or equilibrium binding)	Kd nM
Lysate - Radioactivity	Dissociation	Koff (s^{-1})
Lysate - Radioactivity	Dissociation	t1/2 (min)
Purified MHC - Fluorescence	Association (or direct binding)	EC50 nM
Purified MHC - Fluorescence	Association (or direct binding)	Kon (nM^{-1} s^{-1})
Purified MHC - Fluorescence	Association (or direct binding) approximating Kd	EC50 nM
Purified MHC - Fluorescence	Competition (or equilibrium binding)	IC50 nM
Purified MHC - Fluorescence	Competition (or equilibrium binding)	Kd nM
Purified MHC - Fluorescence	Competition (or equilibrium binding) approximating Kd	IC50 nM
Purified MHC - Fluorescence	Dissociation	Koff (s^{-1})
Purified MHC - Fluorescence	Dissociation	t1/2 (min)
Purified MHC - Fluorescence	Dissociation	Tm (^{o}C)
Purified MHC - Radioactivity	Competition (or equilibrium binding)	IC50 nM
Purified MHC - Radioactivity	Competition (or equilibrium binding)	Kd nM
Purified MHC - Radioactivity	Competition (or equilibrium binding) approximating Kd	IC50 nM
Purified MHC - Radioactivity	Dissociation	t1/2 (min)
Purified MHC - X-ray Crystallography	Structure (crystal, NMR, etc.)	Angstroms

Finally, mistakes can be introduced by incorrect donor genotyping. Heterozygous individuals for HLA class II genes, displaying to different haplotypes, represent a

particularly difficult case when trying to discriminate which molecule is responsible for presenting a certain peptide.

Table 4.2. Example of miss annotated peptides at the Databases.

Sequence	Qualitative value	Problem found	Correction	Reference
SPFGQAAAGDKPS	Negative	Ic50 value missed	Annotated as negative	[2]
TDVNRYRYSNNYEAIPHIS	Positive	A percentage was taken as an IC50 value	Qualitative value was changed	[3]
SKPKVYQWFDLRK	?	Not found	Removed from the list	[4]
KSKKHMNHDGEKKKVKKLKD	Positive	Incoherent relation between QV and the Ic50	Corrected using IC50	[5]

The presence of all these potential sources of errors in data retrieved from databases implies that some type of control system needs to be applied before calculations, in order to eliminate incoherences and ensure the best possible performance. This is especially the case when one attempts to implement a non-supervised server, capable of periodically auto-downloading new data from the databases and rebuilding the predictor. We have dealt with miss-annotations by eliminating or re-annotating all those entrances suspicious to be in incorrect. After doing this (see Materials and Methods in Chapter 2) prediction performance improved around a 10%. Clearly, when applying strict criteria to perform the initial filtering there is a risk to discard valid data. This can be a problem when the amount of data is already very limited.

The argumentation given above suggests a need for protocol standardization.

Protein-data storage and retrieval

The development of new techniques for genome sequencing and analysis has increased the amount of information available in a relative short time (Figure 4.1), calling for the development of new techniques to store and retrieve data in an acceptable run time. This goal can be achieved either by increasing computational resources or improving data organization.

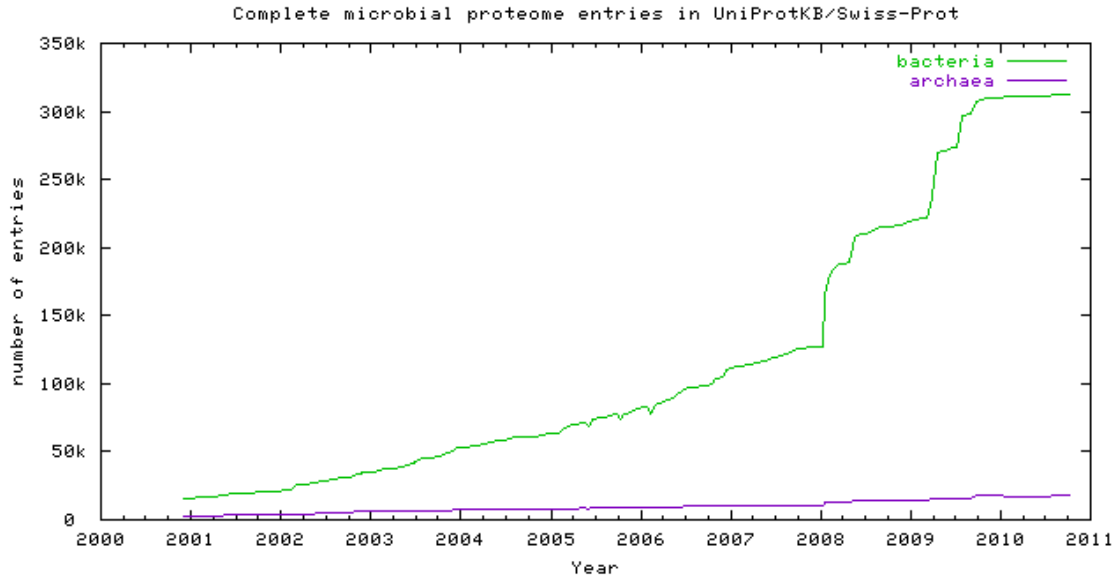


Figure 4.1. Increase of UniprotKB/Swiss-prot database in the last 11 years.

As we expected a continuous growth of our internal database as more sequences become available in public databases, optimizing data storage has been a cornerstone of our work. We addressed this issue by testing different techniques to reduce the required storage support, emphasizing data codification and optimizing the indexation of epitope motifs.

Different initiatives were taken into account. On the one hand, in the PSSM approach the scores for each binding motif are not stored, but calculated in real time. When asterisks are introduced in the sequence only the scores for resulting peptides in which an asterisk occupies an anchoring position are calculated. On the

other hand, all possible epitope motifs in a protein are coded as a number using a function.

Data codification and organisation

Data storage is crucial to the performance of the application. The storage must satisfy both a quick data retrieval and a minimum disk space. Those objectives are fulfilled by data storage optimization.

A motif can be codified in two different ways. The simplest one is storing the character corresponding to each amino acid in the motive. Alternatively, one can also store it as a numeric code. Storing motifs with characters costs 1 byte per character, i.e. 5 bytes per motive. Storing the motif as an integer represents 32 bits of memory (4 bytes). Despite this difference may seem small, when using large datasets it becomes significant. In addition, this type of codification optimizes table indexation, and reduces the virtual memory needed for preforming a query, increasing search velocity.

Codification of bacterial and human proteomes

Each human and bacterial protein sequence was decomposed into 9-residue segments by running a 9-residue window over the sequence with a 1-residue step. The putative-motif code (PMC) for the 9-residue segment was coded using the following equation

$$PMC = 1 + A9 + 20^1 A7 + 20^2 A6 + 20^3 A4 + 20^4 A1 \quad (4.1)$$

where An is a code (assigned arbitrarily from 1 to 20) to an amino acid at the anchoring position n.

For each 9-residue segment, the protein code (Uniprot), starting position in the protein sequence and PMC were stored.

Database schema

A database is an organized collection of data. Nowadays, these collections are stored in a digital format using complex software for their management, known as database management systems (DBMS). The data collection together with the DBMS is called a database system. In choosing a DBMS we looked for a good balance between data-retrieval velocity and storage optimisation. Although MySQL was a clear initial option, we finally moved to PostgreSQL because of its capacity to use functions to create indexes.

Database design is critical for effective data storage and retrieval. To achieve this goal, data was grouped and codified in different tables. As seen in Figure 4.2 for the PSSM case, tables were divided according to data content and functionality. The Processed-data repository contains all information resulting from data calculations. The Raw-data repository contains all data collected from on-line databases. Tables under Users control are exclusively dedicated to the management and control of the data of registered users. Finally, Database-management tables control database updates.

Tables contained in the Processed-data repository are: Fs, including all the statistical data for each HLA molecule processed and the user who submitted the calculation. PSSMs table is where PSSMs resulting from calculations are stored. Each field in this table is assigned to a particular user (ALL for public data), so that public and private data do not mix. This section also contains all the tables oriented to provide proteomic support to the service. These are Sp, Names and Local. The Sp table contains the names of species in the database and their Uniprot taxon identification (taxid). This information is related to the Names table by the taxid, which also contains the protein Swiss-prot code and the name for each analysed protein. Local table contains all putative motifs and their position in the proteins. The Pockets table contains the amino acids conforming the pockets of all examined HLA class II sequences. Finally, the Pocomp table stores the distances resulting from comparing pockets of different molecules.

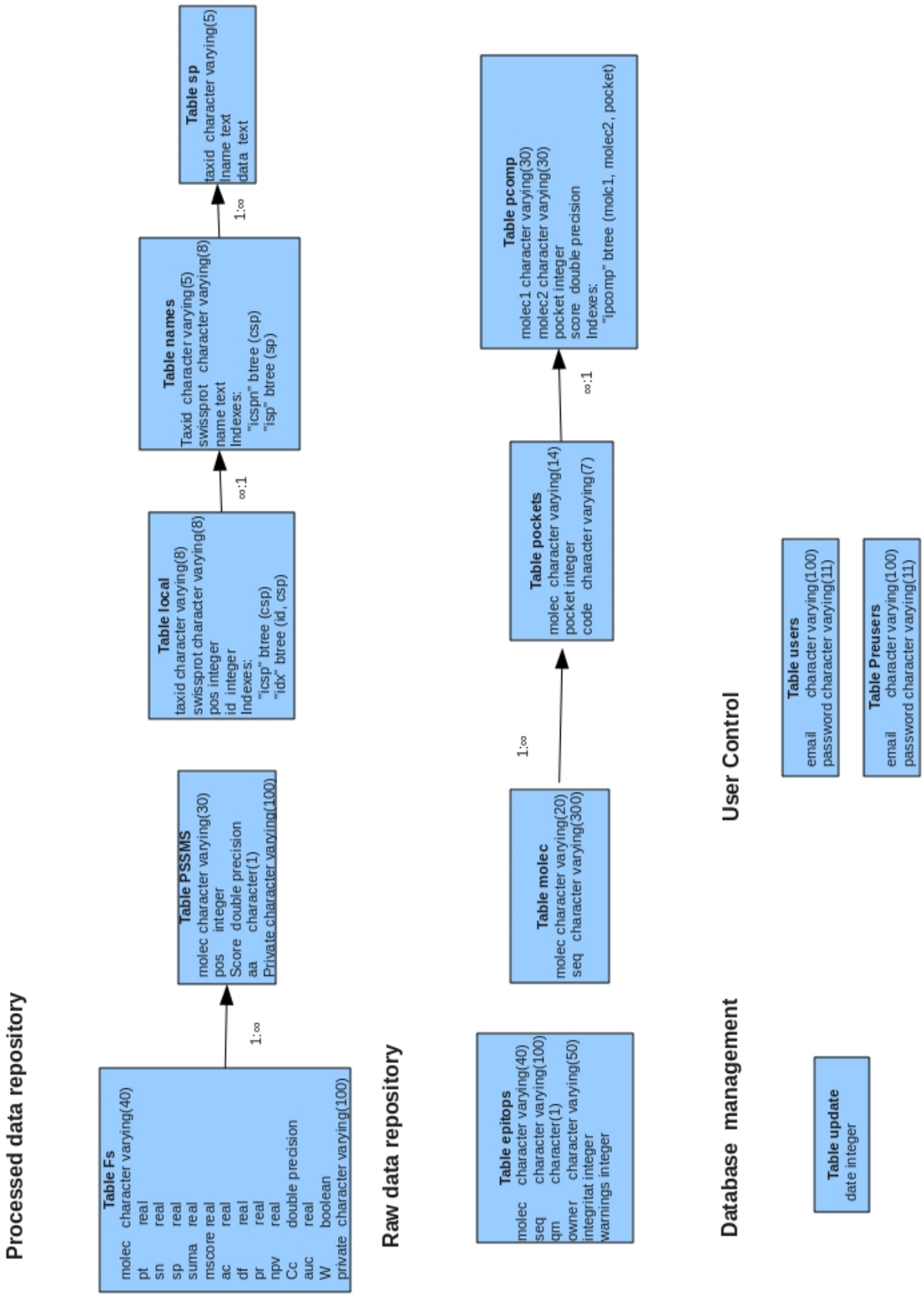


Figure 4.2. Schema of HLA2db (PSSM-based) database organisation.

The Raw-data repository contains all sequences of proteins and epitopes as downloaded from databases. This information is contained in the Molec and Epitopes tables, respectively.

A group of tables related to user management and database updates was also set up. Under the User control section two tables have been implemented, Preusers and Users. The first one stores information on users once a registration request has been received. If a confirmation is received, the information is finally stored in the Users table, otherwise it is deleted. Finally, the Update table contains the date of the last database update.

Binding-motif calculation procedure

Peptide binding core

As mentioned in the introduction, the length of peptides binding to HLA class II varies form 8 to 30 residues, from which only 9 residues occupy the protein's binding grove. The definition of this peptide core is therefore central to the work presented here, as the statistical analysis on which both the PSSM and SVM approaches rely depend directly on it.

The actual peptide length depends on the protein processing taking place before presentation [6]. An interesting characteristic of this process is the possibility of producing peptides with the same core but different ends, as a result of the action of proteases with cleavage preferences. These peptides are known as nested sets. This property allows us to define the peptide core by identifying the common part between two or more peptides (see Materials and Methods in Chapter 2). Note that this can be extended to other potentially existing peptides binding to the same HLA molecule and having the same 9-residue core but originating from a different protein (generally an ortholog). Although this would not be strictly a nested set, our protocol does not distinguish between these two cases.

In contrast with the iterative self-consistent method (see Introduction), which strongly depends on the knowledge on the amino acid at P1, our technique is less prone to the introduction of noise and less information dependent. In the one hand, using nested sets to find the epitope core is independent of any previous knowledge and allows finding new amino acids able to bind to P1. On the other hand, it reduces the inclusion of non-binding core sequences at early stages of the training, which is crucial to the correct development of the refinement and the final result.

In comparison to the alignment of sequences resulting from extracting information from crystallographic structures, our method is less specific but the amount of data available is much more extense, facilitating a wider coverage of HLA molecules.

Computation phases

Dataset enrichment with Blosum tables

When using the nested-set-like approach mentioned in the previous section for the identification of peptide cores, the number of available peptides per HLA class II molecule becomes relatively small. This is especially so because we discard any peptides not entering one of these nested-set-like groups (see Materials and Methods in Chapter 2). To be able to deal with HLA molecules ending up with an insufficient number of representative peptides for a statistically significant analysis, a methodology based on BLOSUM tables was developed.

Blosum tables are a group of two-dimensional matrices resulting from computing the amino-acid exchange probability in aligned proteins with a certain degree of divergence (the percent identity is given as an appendix number to the Blosum name). Each time an amino-acid substitution occurs in a conserved protein region, the new amino acid should have similar properties to ensure the conservation of protein structure and function. Therefore, the more often an exchange between two amino acids exist, the more similar they are. Thus, the elements of a Blosum

matrix can be interpreted as a degree of similarity between amino-acid pairs, being the higher scores for the more similar and exchangeable amino acids.

Here, Blosum tables are used to find new putative core sequences with an exchangeable amino acid at the anchoring positions with those resulting from peptide comparison. Once this correlation is done, a PSSM can be computed, which is used to perform the first prediction of cores from the training set of peptide binders. This process is repeated in 3 rounds (focusing rounds), and the resulting tables are not yet considered for the final predictor.

In contrast with iterative self-consistent method, which scans the proteome to find 9-residue sequences with a certain amino acid at P1, to then perform a statistical analysis to evaluate each nonamer and exclude from the training set those with lower score until convergence, our method starts from few specific peptide cores to move to larger sets, allowing to find new amino acids that may eventually fit into a pocket without previous experimental evidence. This capacity to find new potential relations between pockets and amino acids becomes even more evident when comparing to methods taking crystallographic information to perform the statistics.

Predictor training

Why should the sequences be unique?

A PSSM reflects the frequency of a certain amino acid at each of the 9 positions in the core, and is the nexus between rounds. In this sense, the inclusion of repeated core sequences in the sample used to perform the statistics could lead to overweighting those amino acids in the repeated sequence, resulting in matrix corruption and a decrease of the prediction performance.

What is the meaning of the formulae used for PSSM construction?

The specific characteristics of the method developed here are partly dictated by the formulae used to construct the PSSM and make the predictions. The calculation was performed in two different ways depending on the phase of the process (see

Materials and Methods in Chapter 2). In the focusing phase, amino-acid frequencies at each position were corrected using BLOSUM tables. In the refinement phase, the PSSM was refined using a statistical formula. To this end, 22 different formulae were tested. The performance obtained with each formula was measured using the area under the curve (AUC) in a Receiver Operating Characteristic (ROC) curve during various rounds of prediction – calculus.

The equation chosen to calculate the statistics derives from a binomial distribution, and measures the relation between the frequency of each amino acid in the sample and its frequency in nature (see eq. (2.3) in Chapter 2). That is, it considers existence or non-existence of the amino acid at the position and, in the former case, whether it is above or below the expected frequency (in nature). A logarithmic function is used to increase the differences between values.

This equation works with small samples provided the frequencies are limited to single positions, i.e. synergies or correlations between amino acids at different positions in the sequence are ignored. It is for this reason that a second approach based on SVMs was later implemented.

Formulae derived from a normal distribution or from G or Chi-square were also evaluated but, as expected, were found inappropriate. In addition, BLOSUM-based calculations were seen to fail when used as unique approach. Although there is a correlation between amino-acid properties and epitope presentation, it does not take into account pocket characteristics as conformation and flexibility, which are also determinant for binding.

Data corruption, over-learning and convergence

A predictor needs to be trained to gain the ability to discriminate different cases. In an ideal situation, the initial set provides all possible cases that could occur in nature and the training would be a one-step calculation. Unfortunately, the most common scenario is to have a set of data more or less representative of the real population. Here, training deals with the dangers of over-learning and matrix corruption.

Data corruption occurs when the data from which statistics are calculated is miss-annotated. In those cases, sequences considered positive are classified as negative and vice-versa. This can easily occurs with low affinity epitopes that are on the boundary between presented and non-presented. As a result, the method's discrimination capability is affected, being reflexed as a decrease of the AUC value in a ROC curve. This is caused because the miss-annotated sequences introduce a bias into the PSSM that is fixed at following rounds, being the error amplified. Normally, data fixation occurs from the third to the fifth rounds (see Figure 4.3A), at which point the learning process should be stopped [7, 8].

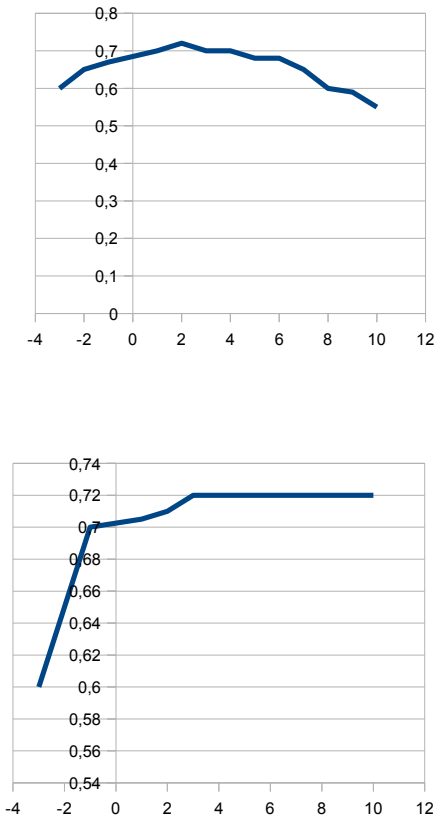


Figure 4.3. Examples of the two possible scenarios in the evolution of the AUC in a ROC curve during a learning process: corrupted sample (upper panel) and non-corrupted sample (lower panel).

A second scenario is data convergence (Figure 4.3B). It happens when data discrimination cannot be improved any further. This may happen if the initial dataset is sufficiently large to fully describe the underlying population or, on the

contrary, if the dataset is too small to show the native diversity, leading also to a situation in which positions are over-fixed and the method looses the possibility of finding cases not included in the training set. Over-fixing positions can be avoided simply by stopping the process before right when convergence occurs, which is normally at the fifth round [7, 8].

Over-learning is another problem and is usually connected to early convergence. To avoid it, the statistics must be performed on a large sample displaying the native population diversity, allowing the construction of both a training and a test set [7]. Another possible initiative to minimize over-learning is to allow the algorithm to have enough freedom to report not only the patterns seen in the training set but also patterns similar to those. When starting from a sufficiently large dataset, over-learning can be avoided by setting a threshold that equilibrates the sensitivity and specificity of the method.

An approach taken to partially avoid these problems was to re-build all sets at each step of the processing. This way, almost all peptides in the repository were used to feed both the train and test sets. Once this procedure was set up, it became apparent that the result was highly dependent on the quality of the sets at the very initial rounds. For this reason, the process was repeated 20 times, with initial random distribution of peptides in the training and test sets (see Materials and Methods in Chapter 2). This way the process gained stability.

Are some pockets more important than the others?

The idea that some pockets are more important than others is wide spread among immunologists. To tests this hypothesis a formula that reflects the possible differences between pockets was tested:

$$S_c = \sum_{i=1}^{9} f_i S_i \qquad (4.2)$$

where S_c is the score for the 9-residue core, S_i is the score for the amino acid at position i and f_i is obtained from a G-test (see below) if i=1,4,6,7,9 (anchoring positions) and is set equal to 1 otherwise. This equation weights the score of each

amino acid bind at each position according to pocket importance. It was calculated by measuring the deviation of the distribution of amino acids in each pocket from the distribution in nature by using a G-test. We made the assumption that the higher the difference between the expected (from distribution in nature) and observed frequencies, i.e. the larger the G-test value, the more important the pocket is and, accordingly, the higher its weight in the final score.

Contrary to our expectation, the inclusions of these weight factors in the formula produced a decrease of the performance of the method, reflected by a decrease of the AUC of the ROC. This result may have two different interpretations. A first one is that the PSSM may already incorporate implicitly this information. In this case, the inclusion of the weight factors in eq. (4.2) would result in over-weighting. A second one is that epitope presentation may not depend as strongly as often assumed on specific positions but, rather, on the global energetics of the interaction between peptide and binding groove, including entropic effects. Indeed, everything we know today about molecular interactions goes in this direction. After this result, we decided not to introduce weight factors to distinguish the relevance of the different pockets in the binding groove.

Why do we use a positive- and a negative-binding PSSM?

PSSMs for positive and negative binders were separately constructed and subtracted (negative from positive). The objective was double, to reduce the background noise from amino acids with similar frequencies in both tables and to make the differences between good and bad binders more evident in the learning process.

Why does the prediction improve by considering the different possible binding motifs in a peptide and their lengths?

One of the initial objectives of this work was to adjust the score predicted for binding to existing IC50 measurements. This objective has not been fulfilled. This can be due to many factors, among them the already discussed intrinsic error of experimental measurements of binding, illustrated by different values reported for

the same peptide. As a matter of fact, the actual binding properties of a peptide will result from the combination of all its potential binding modes, with appropriate thermodynamic weights. This is because a peptide longer than 9 residues can potentially present more than one binding motif. Our approach follows this precise reasoning, such that in analysing a given peptide the scores of all possible binding motifs are added and normalised by the length of the peptide. The normalisation factor is introduced to take implicitly into account the larger loss of entropy of the longer sequences, which is negative to binding. When using this approach, the prediction performance increases (see Table 4.3), suggesting that this is indeed an effect present in nature.

Table 4.3. Example comparison of prediction performance considering only the 9-residue segment with the larger score or the normalised sum for all segments.

HLA type	Largest score	Normalised sum of scores
DRB1*0101	0.68	0.73
DRB1*0301	0.64	0.68
DRB1*0401	0.66	0.67
DRB1*0404	0.73	0.76
DRB1*0405	0.69	0.72
DRB1*0701	0.75	0.77
DRB1*0802	0.66	0.68
DRB1*1101	0.75	0.79
DRB1*1302	0.67	0.68
DRB1*1501	0.64	0.69
DRB4*0101	0.65	0.75
DRB5*0101	0.75	0.78
Mean	0.69	0.72
Min	0.64	0.67
Max	0.75	0.79

Why does the molecule rescue work?

HLA class II molecules are highly structurally conserved. In line with previous approaches [8] we considered the possibility of finding news patterns by using combinations of those previously described. To implement this idea it was

necessary to decide on a measure of pocket similarity. With this objective, the study by Venkarajam and Braun [9] was used. They performed a principal components analysis (PCA) to describe the 20 amino acids using 237 different physico-chemical features. We decided to use the four principal eigenvectors from this PCA to represent describe the amino acids. The distance in this four-dimensional space between the amino acids conforming two equivalent pockets in two different HLA molecules can then be measured (see Materials and Methods in Chapter 2). The smaller the distance between the two pockets, the higher their similarity. If one of these pockets is already represented in a PSSM, the corresponding column can be inherited by the second molecule.

5-residue vs. 9-residue profiles

It is widely accepted that only the five anchoring positions are relevant for peptide binding to HLA class II. The corresponding pockets are named P1, P4, P6, P7 and P9 (see Introduction). Building on this hypothesis, when constructing the PSSMs we initially based our calculations on only these positions in the peptide. However, comparison with PSSMs built using the 9 residues of the peptide core showed that the performance of the resulting 9-column matrices was superior. This suggests that all 9 residues occupying the binding groove are relevant to binding, as discussed previously. This was further supported by results from the support vector machines. Generation of SVMs with the five anchoring positions and with all nine positions showed that the 9-residue-based SVMs required lower space complexity than 5-residue-based ones. In the example given in Table 4.4 we can see that both cases had the same performance with an AUC around 74% for HLA-DRB1*04:01, but the parameter values for the 9-residue model were smaller than the parameters for the 5-residue model. This is important because the larger the values of gamma and C in a SVM model, the more complex are the hyperplanes separating the populations and the more difficult to interpret the results.

Table 4.4. Comparison between the performances of the SVM fed with sequences of nine residues or only with the five anchoring residues

Positions	g	c	AUC
5	2	49	0.74
9	2	25	0.74

References

1. https://dst.liai.org/Assay_response_units_MHC.html

2. Abel LCJ, Iwai LK, Viviani W, Bilate AM, Faé KC, et al. (2005) T cell epitope characterization in tandemly repetitive Trypanosoma cruzi B13 protein. Microbes Infect 7: 1184–1195. doi:10.1016/j.micinf.2005.03.033.

3. Patarroyo ME, Bermúdez A, Salazar LM, Espejo F (2006) High non-protective, long-lasting antibody levels in malaria are associated with haplotype shifting in MHC–peptide–TCR complex formation: a new mechanism for immune evasion. Biochimie 88: 775–784. doi:10.1016/j.biochi.2006.01.005.

4. Texier C, Pouvelle S, Busson M, Hervé M, Charron D, et al. (2000) HLA-DR Restricted Peptide Candidates for Bee Venom Immunotherapy. J Immunol 164: 3177–3184.

5. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, et al. (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364: 33–39. doi:10.1038/364033a0.

6. Nielsen M, Lund O, Buus S, Lundegaard C (2010) MHC Class II epitope predictive algorithms. Immunology 130: 319–328. doi:10.1111/j.1365-2567.2010.03268.x.

7. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402. doi:10.1093/nar/25.17.3389.

8. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, et al. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotech 17: 555–561. doi:10.1038/9858.

9. Venkatarajan MS, Braun W (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. Mol Model Annu 7: 445–453. doi:10.1007/s00894-001-0058-5.

Conclusions

1. An internal database of MHC class II epitopes has been built. The system incorporating peptides into this database is able to deal, to a certain extent, with miss-annotations and ambiguities.

2. An algorithm to automatically generate HLA class II epitope profiles starting from the experimental data gathered in this database has been established.

3. Two predictors of binding of peptides to 31 HLA class II molecules have been developed, one based on position-specific scoring matrices and one based on support vector machines. Both predictors can be trained without any expert supervision.

4. A database that helps researchers to localize the list of predicted epitopes in a variety of proteomes (243 pathogenic bacteria and human) has been set up.

5. A new tool that helps researchers train PSSMs for new HLA class II molecules or improve existing ones based on the user's own peptide datasets has been set up.

6. The training of both the PSSM- and SVM-based methods shows that the binding of peptides to HLA class II molecules depends not only on the amino acids present at the anchoring positions of the peptide core (1, 4, 6, 7 and 9), but on the entire 9-residue sequence.

7. Binding of a peptide to HLA class II cannot be reduced to the best-binding core but it will be a function of all potential binding arrangements, i.e. peptides

presented *in vivo* may contain more than one 9-residue segment with the capacity to bind to the HLA molecule.