QSAR and the Prediction of T-Cell Epitopes

Irini A. Doytchinova[1,*] and Darren R. Flower[2]

[1]*Faculty of Pharmacy, Medical University of Sofia, Bulgaria and* [2]*Jenner Institute, Oxford University, UK*

Abstract: Quantitative structure – activity relationships (QSAR) is a well established ligand-based approach to drug design. It correlates changes in the chemical structure of a series of compounds with changes in their biological activities. Peptides of equal length which bind to a certain protein are an excellent target for QSAR. In the present review, we summarize our experience in QSAR studies of peptides acting as T-cell epitopes. T-cell epitopes are protein fragments presented on the cell surface which afford the immune system the opportunity to detect and respond to both intracellular and extracellular pathogens. Epitope-based vaccines are a new generation of vaccines with lower side effects. The process of antigen presentation, which includes proteasome cleavage, TAP and MHC binding, has been modeled and analyzed by QSAR. Derived QSAR models are highly predictive, allowing us to design and test *in vitro* MHC superbinders. All models have been implemented in servers for *in silico* prediction of MHC binders and T-cell epitopes. In practice, better initial *in silico* prediction leads to improved subsequent experimental research on epitope-based vaccines.

Key Words: Proteasome, TAP, MHC class I, MHC class II, additive method.

INTRODUCTION

Vaccination is one of the greatest boons to mankind. Together with other medical discoveries, such as antibiotics, vaccines have greatly reduced mortality and morbidity resulting from infectious diseases. Together with advances in agriculture powered by the Haber process, vaccines have led to an unprecedented burgeoning of the global population and reduction of human misery resulting from uncontrolled epidemics. The history of vaccines began with Edward Jenner's assault on smallpox. On 14th May 1796, he used cowpox, a virus related to smallpox, to build protective immunity in his gardener's son. The culmination of Jenner's work led to the 1980 declaration by the World Health Organization that smallpox had been eradicated. There are now similar global campaigns against polio and tuberculosis. BCG, the main vaccine against tuberculosis, is perhaps, the most widely used vaccine worldwide and is an attenuated or weakened form of the tuberculosis bacterium. Most established vaccines fall into one of two categories: they are either whole viruses or bacteria (albeit chemically treated or attenuated forms thereof) or they are single proteins derived from whole pathogenic bacteria or viruses.

More recently, vaccine discovery has attempted to replace older types of vaccine with rationally designed peptide or DNA vaccines. These vaccines are physically smaller, being based on one or more epitopes. Epitopes are short peptides recognized by the immune system. In the case of vaccines, this recognition should lead to stimulation of specific, protective immunity against disease and mitigation of subsequent infections. As the genomes of pathogenic microbes genomes have been sequenced, epitope-based vaccine discovery has moved firmly into the arena of rationale proteomics.

*Address correspondence to this author at the Faculty of Pharmacy, Medical University of Sofia, Dunav st. 2, Sofia 1000, Bulgaria;
E-mail: idoytchinova@pharmfac.net

ANTIGEN PROCESSING AND PRESENTATION

The main processing pathway for Major Histocompatibility Complex (MHC) class I ligands involves degradation of intracellular viral or self proteins by the proteasome, followed by transport of the products by the transporter associated with antigen processing (TAP) to the endoplasmic reticulum (ER), where peptides are bound to MHC class I molecules, and then presented on the cell surface (Fig. **1**, class I pathway). The proteasome is responsible for generating the C terminus but not the N terminus of the final presented peptide (Crauu *et al*., 1997; Mo *et al.*, 1999; Serwold and Shastri, 1999; Cascio *et al*., 2001). The proteasome is a multimeric proteinase with three active sites: a site with trypsin-like activity (cleavage after basic residues), one with chymotrypsin-like activity (cleavage after hydrophobic residues), and another with peptidylglutamyl-peptide hydrolytic activity (cleavage after acidic residues) (Orlowski and Michaud, 1989; Djaballah, 1992; Orlowski *et al.,* 1993). In addition, in vertebrates there are three γ-interferon-inducible subunits that replace the constitutive subunits (Tanaka and Kasahara, 1998) and assemble the immunoproteasome. The immunoproteasomes have an altered hierarchy of proteosomal cleavage, enhancing cleavage after basic and hydrophobic residues and inhibiting cleavage after acidic residues (van den Eynde and Morel, 2001; Toes *et al.,* 2001). This is in accord with the amino acid preferences for binding to MHC class I molecules at the C terminus (Rammensee *et al.,* 1995).

TAP is an ATP-dependent peptide transport protein that belongs to the ATP-binding cassette (ABC) family of transporters. This family transports across membranes a wide range of molecules, from small sugars to large polypeptides (Monaco *et al.,* 1990). There are two TAP proteins (TAP-1 and TAP-2) which form a transmembrane (TM) heterodimer. Both proteins encode one hydrophobic TM domain and one ATP-binding domain (Meyer *et al.,* 1994). Extant experi-

1570-1646/08 $55.00+.00

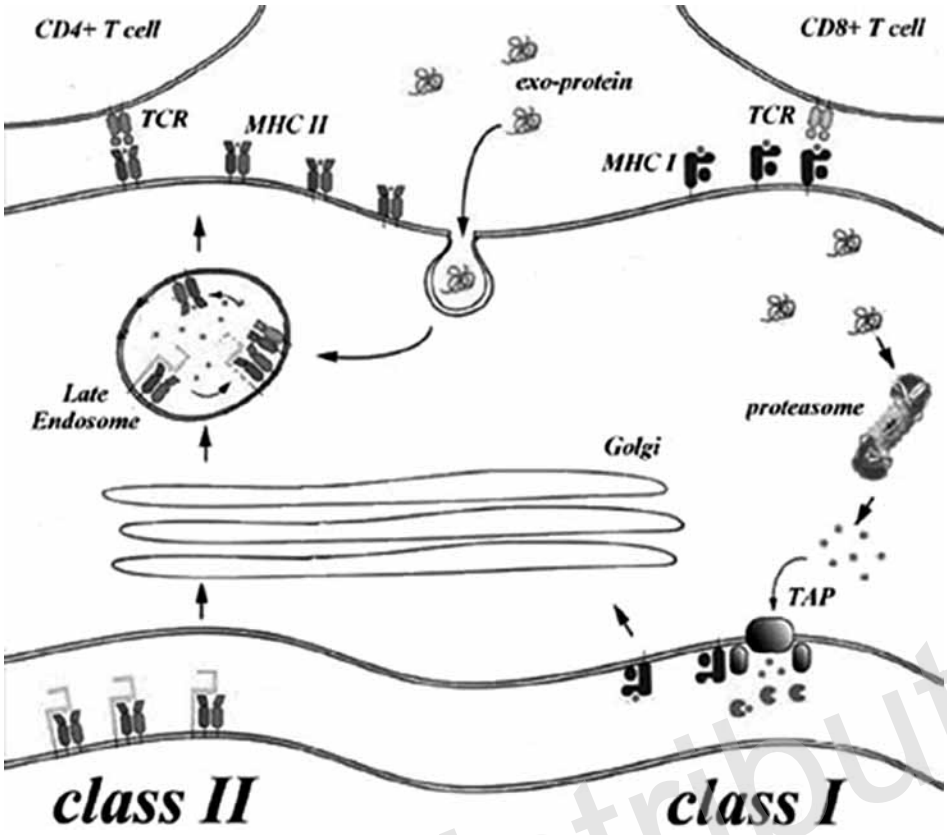


Fig. (1). MHC class I (left) and class II (right) pathways (http://stratikos.googlepages.com/mhc).

mental studies have shown that TAP prefers peptides of eight or more amino acids with hydrophobic or basic residues at the carboxy terminus (Müller *et al.*, 1994; Schumacher *et al.*, 1994). TAP-mediated antigen presentation is important not only for cytosolic antigens but also for most epitopes within membrane or secretory proteins (Lautscham *et al.*, 2003). The TAP-dependent pathway is the principal processing route for peptides binding HLA-A1, HLA-A3, HLA-A11, HLA-A24, HLA-B15 and HLA-B27 (Mormung *et al.*, 1994; de la Salle *et al.*, 1997). Some peptides are able to access the ER *via* other, TAP-independent mechanisms. Examples of alleles exhibiting only partial dependence on TAP include HLA-A2, HLA-A23, HLA-B7 and HLA-B8 (Henderson *et al.*, 1992; Guéguen *et al.*, 1994; Smith and Lutz, 1996; Khanna *et al.*, 1996).

Extracelular bacterial or parasite antigens are taken up into acidified intracellular vesicles by phagocytic cells or endocytosed by other professional antigen-presenting cells and degraded to oligopeptides (Fig. **1**, class II pathway). MHC class II molecules synthesized in the ER pass through such vesicles, bind peptide fragments of the antigen, and then transport them to the cell surface.

MHC proteins are both polygenic (i.e. there are more than one MHC class I and MHC class II genes) and polymorphic (i.e. there are many alleles of each gene) (Janeway, 2001). Each class of MHC has several loci: HLA-A, HLA-B and HLA-C for class I and HLA-DR, HLA-DQ and HLA-DP for class II. MHC alleles may differ by as many as 30 amino acid substitutions, most of them are found within the binding site and are critical for the specificity of peptide binding and therefore for T cell recognition (Saper *et al.*, 1991; Smith *et al.*, 1996). Such an uncommon degree of polymorphism implies a selective pressure to create and maintain it. Different polymorphic MHC alleles have different peptide specificities: each allele binds peptides exhibiting particular sequence patterns.

The complex peptide-MHC class I molecule presented on the cell surface are recognized by CD8 T cells, while the complex peptide-MHC class II molecule – by CD4 T cells. The function of CD8 T cells is to kill infected cells; this is an important means of eliminating sources of new viral particles and obligate cytosolic bacteria, thus freeing the host of infection. CD4 T cells are specialized to activate other cells and fall into two functional classes: Th1 cells, which activate macrophages to kill the intravesicular pathogens they harbor, and Th2 cells or helper T cells, which activate B cells to make antibody. Peptide binding by MHC is, in all likelihood, the bottleneck – that is to say the most discriminating phase - in the recognition of epitopes.

EPITOPE PREDICTION IN PROTEOMICS: STATE OF THE ART

The identification of so-called "binding motifs" began in the 1980s. Motifs seek to characterize peptide specificity of a particular MHC molecule in terms of dominant anchor positions which exhibit strong predilections for a constrained group of amino acids. For example, arguably the most well studied of human MHC proteins - HLA-A*0201 - has anchor

residues at peptide positions p2 (which will accept amino acids Met and Leu) and p9 (which will accept amino acids Leu and Val).

Motifs have proved to be very popular, and very widely exploited; they are both simple to use and very simple to understand. Notwithstanding such lucent simplicity, fundamental technical problems limit their utility: motifs generate many false negatives and many false positives. Moreover, peptides are viewed as either binders or non-binders ignoring the extra dimension of understanding that comes from consideration of affinity. It has been obvious for some time that the whole peptide contributes to the determination of affinity, not just a few anchor residues, and likewise T-cell-mediated immunogenicity. Effective models of binding must employ rather more intricate and complex representations of the biophysical phenomena of binding. There is now a profusion of sequence-based methods for prediction of T-cell epitopes, most relying on the prediction of peptide-MHC binding (Flower *et al.,* 2003). Successfully modelling peptide specificity exhibited by MHCs allows pre-selection of candidate peptides, which, in turn, can help identify immunogenic epitopes. Models which accommodate affinity allow us to modify and modulate affinity, and thus aspects of immunogenicity, in a rationale manner.

Class I MHC alleles have a binding groove which is closed at both ends. Peptides are locked at either end of the groove, which allows us to predict with precision which residues are positioned in the groove. Many methods have been used to predict MHC binding and class I prediction is regarded as being relatively successful, with high reported prediction accuracies (Dönnes and Elofsson, 2002). A succession of ever more sophisticated methods has been applied to the problem: starting with Parker's BIMAS, and progressing through Artificial Neural Networks; Hidden Markov Models; to Support Vector Machines (Flower *et al.,* 2003; Flower, 2003). These have engendered a complete panoply of implementations available *via* the Internet.

Support Vector Machines (SVMs) are an artificial intelligence technique whose inherent accuracy is compelling. Using an appropriate amino acid representation, a single SVM is a binary classifier which identifies a decision boundary between two classes - in this case between epitope and non-epitope - by maximising the margin between them, choosing a linear separation in feature space. As a result of their success, a whole array of SVM-based methods for class I epitope prediction have developed (Dönnes and Elofsson, 2002; Zhao *et al*., 2003; Riedesel *et al.,* 2004; Bhasin and Raghava, 2004a; Bhasin and Raghava, 2004b; Bhasin and Raghava, 2004c; Cui *et al.,* 2007). Most SVM methods undertake discriminant analysis, but there has also been encouraging performance with quantitative prediction using support vector regression (Liu *et al.,* 2006).

The other significant recent trend in epitope prediction has been the use of structures of MHCs and MHC-peptide complexes. These methods use two main techniques: docking and molecular dynamics (MD) simulation. Leveraging earlier work (Altuvia *et al.,* 1995; Schueler-Furman *et al.,* 2000), several approaches apply docking methods (scoring functions derived from computational chemistry or threading methods derived from structural bioinformatics) to identify MHC binders (Tong *et al.,* 2004; Bordner and Abagyan, 2006; Bui *et al.,* 2006; Jojic *et al.,* 2006; Tong *et al.,* 2006a,b). Several workers have used molecular dynamics as a means of realising affinity prediction (Davies *et al.,* 2003; Petrone and Garcia, 2004; Zacharias and Springer, 2004; Fagerberg *et al.,* 2006). However, the availability of computing resources able to sustain the requisite size and duration of simulation necessary for obtaining free energies of binding for medium sized systems such peptide-MHC complexes has hampered development of easily deployable techniques. However, work by Wan and colleagues have begun to address this. They make use of high performance computing deployed *via* the nascent GRID. This approach achieves more realistic simulations of an escalating series of systems of increasing scale (Wan *et al.,* 2004; Wan *et al.,* 2005a; Wan *et al.,* 2005b). Another interesting piece of work (Davies *et al.,* 2006) combines MD with multivariate statistics to produce a hybrid approach to prediction.

In recent times, several attempts to incorporate components of the class I antigen presentation pathway, such as proteasome cleavage (Saxova *et al.,* 2003) and TAP binding (Bhasin and Raghava, 2004b; Doytchinova *et al.,* 2004a), have been made, which have created combined approaches to T-cell epitope prediction (Peters and Sette, 2005; Larsen *et al.,* 2005; Dönnes and Kohlbacher, 2005; Doytchinova *et al.,* 2006a). These approaches show encouraging improvements compared to methods which only rely on MHC-binding. They seek to decrease the number of potential epitopes using subsidiary component stages as additional sequential or concurrent filters.

Prediction of peptide binding to class II MHCs is greatly complicated by their open peptide binding groove. Class II MHCs can, as a result, bind much longer peptides (25+ residues) compared to peptide binding to class I MHCs (at most 15 residues). The grooves of MHC class II alleles will only accommodate 9 to 11 residues of the bound peptide. Class II peptides have the potential complication of being able to bind into the groove in one of several distinct registers (potential alignments between groove and antigenic peptide). Moreover, several studies have indicated that residues while lie outside the binding groove (flanking residues) also affect the magnitude of binding affinity (Carson *et al.,* 1997; Godkin *et al.,* 2001). To complicate the development of effective predictive schemes still further, available data for class II is sparse compared to that available for class I; when coupled to the greater intrinsic complexity of the prediction problem itself, this results in a much reduced level of reported accuracy.

As a consequence of observed problems with the reliability of class II predictions, a wide range of imaginative and innovative approaches have been used in attempts to solve this problem. Pattern recognition techniques used include ANN (Noguchi *et al.,* 2001; Burden and Winkler, 2005) and SVM (Bhasin and Raghava, 2004c; Yang and Johnson, 2005). Typically, a binding core is first estimated or declared, and subsequently the binding affinity is predicted for an unknown peptide from this estimate. This two stage procedure separates the task into a fixed-length problem and an alignment problem. Approaches for solving the dynamic variable-length nature of the class II prediction problem have

however shown promise. Methods include an iterative "meta-search" algorithm (Mallios, 2001), an iterative PLS method (Doytchinova and Flower, 2003a), Hidden Markov Models (Noguchi *et al.*, 2002; Murugan and Dai, 2005), an Ant Colony search (Karpenko *et al.*, 2005), and a Gibbs sampling algorithm (Nielsen *et al.*, 2004).

However, there is another productive, if largely undervalued**,** strand in the prediction of MHC peptide interactions: Quantitative Structure Activity Relationship (QSAR). Until recently, there were few if any published examples that applied QSAR methodology to questions arising from the immune system, nor indeed are there that many papers that apply QSAR techniques to any bioinformatic problem. The difference between QSAR and artificial intelligence methods is primarily a semantic one. In practice they achieve the same goal and work in similar ways, but QSAR techniques tend to be based on different, and possibly more rigorous, types of statistical analyses, including, amongst others, multiple linear and continuum regression, discriminant analysis, and PLS. Both AI and QSAR based methods make use of a representation of molecule structure (either as sequence or in terms of 3D structure) and a measure of binding (either discrete – binders *vs*. non-binders – or in terms of continuous, quantitative affinities).

QSAR IN THE CONTEXT OF PROTEOMICS AND IMMUNOINFORMATICS

The basic idea of the QSAR is that the chemical structure of a molecule determines its biological activity. The chemical structure is represented by a number of calculated and/or experimentally derived descriptors which are correlated with one quantitative or qualitative value representing some measure of activity. Although this relationship was long known intuitively by many scientists, Corwin Hansch is considered officially as the father of QSAR. In his publications in the early 1960s, Hansch for the first time defines quantitatively the relationship between changes in the chemical structure of structurally related compounds and changes in their biological activities (Hansch *et al.*, 1962). Since then, QSAR has become a fundamental ligand-based approach in drug and molecular design, and in environmental risk assessment. Thousands of QSAR models have been published in the literature and the theory and practice of QSAR has entered the medicinal chemistry textbooks.

The deciphering of the human genome poured forth an enormous amount of scientific information and opened up an entirely new era in biology. This discovery is of such important that large areas of science are now divided into pre- and postgenomic eras. In the postgenomic era, high-tech methods are a vital resource for any bioscience. Many new -omics have appeared and developed during the last decade. Proteomics is a research area that includes identification, characterization, and quantification of the proteome. The proteome is the whole protein content expressed by a genome in a cell, tissue, or organism in healthy and disease states. Proteomics can provide information for drug discovery including target identification and validation (Greenbaum *et al.*, 2002; Drummelsmith *et al.*, 2003), lead selection (Bleicher *et al.*, 2003) and optimization (Baker *et al.*, 2002; Kridel *et al.*, 2004), toxicity assessment (Imanishi and Harada, 2004; Keightley *et al.*, 2004), biomarkers discovery (Celis *et al.*, 2000; Park *et al.*, 2002; Zhu *et al.*, 2003; Ding *et al.*, 2004; Petricoin *et al.*, 2004; Kageyama *et al.*, 2004). Only a small part (< 5%) of the proteome, namely the druggable proteome, is readily modulated by a small-molecule drug.

The immunome is that part of the microorganism's proteome that interfaces with the host immune system. It consists of epitopes and antigens able to pass through a series of biochemical processes in the cell, described in the previous section. If there are enough data for each step of the process, like affinity assessments (IC_{50}, BL_{50}, etc.), class affiliation (epitopes *vs*. non-epitopes, binders *vs*. non-binders), then the process can be modeled in order to predict which part of the antigen will act as an epitope. Methods used for compilation and mining of immunological data, process modeling, and epitope prediction comprises immunoinformatics, a subdiscipline within bioinformatics. The appropriate use of immunoinformatics greatly improves the efficiency of immunology research.

One of the principal goals of immunoinformatics is to develop computer aided vaccine design, or computational vaccinology, and apply it to the quest for new vaccines. At the heart of computational vaccinology is epitope prediction. The focus of our recent studies is the development of methods for T-cell epitope prediction. We applied the QSAR philosophy to immunology and vaccinology. In the present review we describe our experience in this field. Antigen processing and presentation were modeled by QSAR methods and the models which result were used subsequently to predict T-cell epitopes. Some of the models were tested experimentally and used in the design of MHC superbinders. All derived models are freely accessible *via* our web-servers MHCPred and EpiJen.

Since the value of QSAR approaches is under appreciated in proteomics and immunoinformatics, a brief description of its basic principles will be given below.

THERMODYNAMIC ASPECTS OF QSAR

The ligand – receptor interaction can be modeled as a reversible bimolecular reaction and represented as

$$L + R \underset{k_2}{\overset{k_1}{\Leftrightarrow}} LR$$

The equilibrium constants of association K_A and dissociation K_D are represented by the ratios of the rate constants and involving reactants and product concentrations.

$$K_A = \frac{k_1}{k_2} = \frac{[LR]}{[L][R]} \qquad K_D = \frac{k_2}{k_1} = \frac{[L][R]}{[LR]}$$

For ligand – receptor interactions the values for K_A vary between 10^2 and 10^{12}, and for K_D – between 10^{-2} and 10^{-12}. At equilibrium the constants are related to the standard free energy of binding ΔG^0:

$$\Delta G^0 = -RT \ln K_A = -RT \ln K_D^{-1} = RT \ln K_D \qquad (1)$$

Considering R = 8.314 $Jmol^{-1}K^{-1}$ and absolute temperature in Kelvin T between 298 (25°C) and 310 (37°C), for

ΔG^0 are obtained values between -10 and -80 kJmol^{-1} (Andrews *et al.*, 1984; Böhm and Klebe, 1996; Babine and Bender, 1997). ΔG^0 is composed of an enthalpic ΔH^0 and entropic $T\Delta S^0$ contribution:

$$\Delta G^0 = \Delta H^0 - T\Delta S^0 \quad (2)$$

Combining equations (1) and (2) gives equation (3) known as the integrated form of the van't Hoff equation when ΔH^0 and ΔS^0 are not temperature dependent.

$$\ln K_A = -\frac{\Delta H^0}{R} \cdot \frac{1}{T} + \frac{\Delta S^0}{R} \quad (3)$$

Equation (3) represents a linear relationship between $\ln K_A$ and 1/T with slope = $\pm\Delta H^0$/R and y-intercept = $\Delta S^0/R$. The sign of the slope depends on the heat effect of the reaction, being positive for exothermic and negative for endothermic ones. It is common practice in the thermodynamic analysis of pharmacological interactions to measure K_A or K_D at several different temperatures, then construct a van't Hoff plot and determine ΔH^0 and ΔS^0. Their sign and value depends on the type of interactions involved in the formation of ligand – receptor complex. Hydrophobic interactions are associated with entropic changes, while electrostatic attractions reflect the enthalpic contribution.

Changes in free energy, enthalpy, entropy and other thermodynamic parameters are related for a particular reaction. When multiple reactions are considered, apparent relationship between some of the thermodynamic parameters may appear. These relationships are termed "extrathermodynamic" because they are not derived from the main principles and, hence, lie outside the domain of traditional thermodynamics (Raffa, 2002). When these extrathermodynamic relationships are applied to a series of structurally related compounds, they can yield insight into their common mechanism of action. Typical extrathermodynamic relationships are enthalpy – entropy compensation and linear free energy relationships.

Enthalpy – entropy compensation is the empirical observation that highly exothermic interactions tend to have large entropy changes, whereas more thermoneutral interactions tend to have less unfavorable entropy changes (Calderone and Williams, 2001). In the ligand – receptor interaction, changes in the enthalpy are considered as a measure of the strength of the interaction, while changes in the entropy account for the disorder in the system. Consider a receptor which consists of several proximal binding sites (Calderone and Williams, 2001). Each site binds a distinct ligand with a certain affinity. If the ligands are linked together and bind the receptor more favorably (positive cooperativity), then the increased free energy of binding will result in a structural tightening of the ligand – receptor complex. The tighter binding (increased enthalpy) leads to a restriction of relative motion of the linked ligands (decreased entropy).

In 1937 Hammett found that the effects of substituents on the reaction rate could be assessed quantitatively by parameters describing the chemical structure of a series of structurally related compounds (Hammett, 1937). He defined this dependence as a linear free energy relationship (LFER): changes in Gibbs energy relates linearly to the logarithm of a reaction rate constant or equilibrium constant. Hansch and colleagues used the Hammett constants σ and *logP* to find a correlation between the structure of a series of structurally related compounds and their biological activity (Hansch *et al.,* 1962). Then, for the first time QSAR was defined as a natural extension of the LFER approach. Subsequently, QSAR has become a widely used ligand – based drug design approach. The extremely wide range of QSAR models, published in the literature, is due to the great variety of molecular descriptors now available. According to a compilation by Todeshini, the number of QSAR descriptors exceeds 1600 (Todeshini and Consonni, 2000).

A UNIVERSAL ADDITIVE METHOD FOR MODELING OF PEPTIDE – PROTEIN INTERACTIONS

The method we used in our studies was called "additive", because it is based on the additivity concept, developed by Free and Wilson, whereby each substituent makes an additive and constant contribution to the biological activity regardless of substituent variation in the rest of the molecule.

$$\text{Biological activity} = \sum_{ij} G_{ij} X_{ij} + \mu$$

In this equation, μ is the overall average of biological activity values and G_{ij} is the activity contributions of the substituent X_i in position j (X_{ij} = 1 if the substituent X_i is in position j; otherwise X_{ij} = 0) (Free and Wilson, 1964). The values of the individual group contributions are calculated by multiple linear regression (MLR). Other models based on the additivity concept are alternative modifications of the Free-Wilson model. The Fujita-Ban modification is a simple linear transformation of the Free-Wilson model, where μ is the activity of the unsubstituted compound predicted by the least-squares method (Fujita and Ban, 1971). In Cammarata's model (Cammarata and Yau, 1970) μ is the experimental activity of the unsubstituted compound (all X_{ij} = H). The models based on the additivity concept are simple to perform and easy to interpret. Because of that they have found wide application in molecular design (Bindal *et al.,* 1982; Gombar, 1986; Nisato *et al.,* 1987; Dalpiaz *et al.,* 1997; Tmej *et al.,* 1998; Tomic *et al.,* 2000; Terada and Nanya, 2000).

We extended the classical Free-Wilson model with cross terms accounting for possible interactions between the amino acids side chains (Doytchinova *et al.,* 2002). Thus, the binding affinity of a nonamer expressed in p-units (negative decimal logarithm of IC_{50} values) could be presented by eqn. 4:

$$pIC_{50} = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_{i,i+1} + \sum_{i=1}^{7} P_{i,i+2} + \sum_{i=1}^{6} P_{i,i+3} + \sum_{i=1}^{5} P_{i,i+4} + \sum_{i=1}^{4} P_{i,i+5} + \sum_{i=1}^{3} P_{i,i+6} + \sum_{i=1}^{2} P_{i,i+7} + P_{i,i+8} \quad (4)$$

where the *const* accounts, at least nominally, for the peptide backbone contribution, $\sum_{i=1}^{9} P_i$ is the sum of amino acids

contributions at each position, $\sum_{i=1}^{8} P_{i,i+1}$ - the sum of cross terms between adjacent amino acids, $\sum_{i=1}^{7} P_{i,i+2}$ - the sum of cross terms between every second amino acids, $\sum_{i=1}^{6} P_{i,i+3}$ - the sum of cross terms between every third amino acids, and so on. As the cross terms account for possible interactions between amino acids, the contributions of the last six terms are negligibly small and the binding affinity of a peptide will depend significantly on the contributions of the amino acids at each position and the cross terms between the adjacent and every second amino acids:

$$pIC_{50} = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_{i,i+1} + \sum_{i=1}^{7} P_{i,i+2} \quad (5)$$

Our experience showed that this model describes well the peptide – protein interaction and gives good predictions when the number of peptides in the training set is above 200. For smaller set, the cross terms could be neglected and eqn. (5) is converted to the simpler eqn. (6):

$$pIC_{50} = const + \sum_{i=1}^{9} P_i \quad (6)$$

Each peptide sequence from the training set of *n* peptides is transformed into a string of 0 and 1. A term is equal to 1 when a certain amino acid at a certain position or a certain interaction between two side-chains exists, otherwise it is 0. Thus, in the case of nonamers, a system of *n* equations with 6180 independent variables is generated. One hundred and eighty variables account for the amino acids contributions (20 aa × 9 positions), 3200 - for the adjacent cross terms, or 1-2 interactions (20 × 20 × 8) and 2800 - for the every second cross terms, or 1-3 interactions (20 × 20 × 7). Variables containing only 0s are omitted. As a dependent variable is used IC_{50}, BL_{50}, class affiliation, etc. The system is solved by the method of partial least squares (PLS).

PLS is a so-called projection method. These methods handle data matrices with more variables than observations very well, and the data can be both noisy and highly collinear. In this situation, conventional statistical methods like multiple regression produce a formula that fits the training data but is unreliable for prediction. PLS forms new *x* variables, named *principal components (PC)*, as linear combinations of the old ones, and then uses them as predictors of the biological activity (Wold, 1995).

In our studies we used PLS methods implemented in SYBYL (Tripos Inc.) and SIMCA (Umetrics Ltd.). The optimal number of components (*PC*) was found by "leave-one-out" cross-validation. The cross-validation (CV) is a practical and reliable method for testing the predictive power of the models. It has become a standard in PLS analysis and is incorporated in all available PLS software (Wold, 1995). In principle, CV is performed by dividing the data into a number of groups, developing a number of parallel models from the reduced data with one of the groups omitted, and then predicting the biological activities of the excluded compounds. When the number of the groups omitted is equal to the number of the compounds in the set, the procedure is named "leave-one-out" (LOO). The predictive power of the models was assessed by the cross-validated coefficient q^2, the standard error of prediction (*SEP*), and the *residuals* between the experimental and predicted binding affinity:

$$q^2 = 1 - \frac{PRESS}{SSQ}$$

$$SEP = \sqrt{\frac{PRESS}{p-1}}$$

residual = pIC_{50}exp – pIC_{50}pred

where *PRESS* is the predictive sum of squares ($\sum_{i=1}^{n} (pIC_{50}\,\text{exp} - pIC_{50}\text{pred})^2$), *SSQ* – the sum of squares of pIC_{50}exp corrected for the mean ($\sum_{i=1}^{n} (pIC_{50}\,\text{exp} - pIC_{50}\text{mean})^2$), *p* is the number of the peptides omitted, pIC_{50}pred is that predicted by the CV-LOO value. The *residuals* between the experimental and predicted pIC_{50} values were classified into 3 categories: below |0.5|, from |0.5| to |1.0| and above |1.0|. A mean |residual| or mean absolute error (MAE) was extracted as well.

The optimal *PC* number was used to derive the final non-cross-validated model. This model was assessed by the explained variance r^2, standard error of estimate (*SEE*), and *F* ratio. When external test sets were used for assessment of the predictive ability of the models, the parameter r^2_{pred} was used:

$$r^2_{pred} = 1 - \frac{PRESS}{SSQ}$$

where *PRESS* and *SSQ* are calculated for the test set, pIC_{50}mean is a mean of experimental pIC_{50} over the training and test sets.

MODELING OF PROTEASOME CLEAVAGE

In order to develop additive models for proteosome cleavage prediction (Doytchinova and Flower, 2006), a training set of 489 naturally processed T-cell epitopes (nonamer peptides) associated with HLA-A and HLA-B molecules was collected from our in-house database AntiJen (http://www.jenner.ac.uk/AntiJen) (Blythe *et al.*, 2002; McSparron *et al.*, 2003; Toseland *et al.*, 2005). A test set of 231 peptides, as used by Saxova *et al.*, (Saxova *et al.*, 2003) to compare the performance of the available methods for proteosome cleavage prediction, was used in our study for external validation. All common T-cell epitopes between the two sets were first excluded from the training set.

The epitopes were presented together with the four flanking amino acids before the N-terminus and the five flanking residues after the C terminus (Fig. **2**). Further, these parent

n4 n3 n2 n1 **p1 p2 p3 p4 p5 p6 p7 p8 p9** | p9' p8' p7' p6' p5'	Cleavage
n4 n3 n2 n1 **p1 p2 p3 p4 p5 p6**	0
n3 n2 n1 **p1 p2 p3 p4 p5 p6 p7**	0
n2 n1 **p1 p2 p3 p4 p5 p6 p7 p8**	0
n1 **p1 p2 p3 p4 p5 p6 p7 p8 p9** |	0
p1 p2 p3 p4 p5 p6 p7 p8 p9 | p9'	0
p2 p3 p4 p5 p6 p7 p8 p9 | p9' p8'	0
p3 p4 p5 p6 p7 p8 p9 | p9' p8' p7'	0
p4 p5 p6 p7 p8 p9 | p9' p8' p7' p6'	0
p5 p6 p7 p8 p9 | p9' p8' p7' p6' p5'	1

Fig. (2). Cleavage site presentation. Peptide positions are given in bold. The positions before the N terminus are denoted as "nn", while the positions after the C terminus – as "pn'". The vertical line shows the cleavage site. When the C terminus of the epitope is located at the middle (position p9 of the decamer), the peptide is considered as positive and takes 1, i. e. cleavage site present. The rest of the overlapped peptides are considered as negative and take 0s (cleavage site not present).

18aa peptides were broken into a set of overlapping decamers. The peptide which contained the C terminus of the epitope at position p9 of the decamer was considered as positive, *i. e.* the cleavage site was present. The rest of the overlapped peptides in each set was considered as negative (cleavage site not present). Thus, the initial training set of 489 epitopes generated 4370 decamers, 489 peptides of them had positive cleavages and 3881 peptides were negative.

As the peptides from the test set had a length of 8-12 residues, some of the parent peptides had lengths different from 18 amino acids. Four parent peptides with no flanking residues after the C terminus were excluded from the test set, since it was not possible to locate the cleavage site. Thereby, the final test set included 227 epitopes. They generated 2100 decamers: 227 peptides were positive and 1873 peptides were negative.

For a set of decamers, the additive method generates a matrix with 200 (20 x 10) columns and a number of rows equal to the number of peptides. A column containing the dependent variable (cleavage *vs*. non-cleavage) is added and the matrix is solved by PLS as implemented in SYBYL 6.9. Models including different positions next to the cleavage site were generated in order to assess the importance of the flanking residues. The prediction rate of T-cell epitopes *vs.* non-T-cell epitopes was measured using Receiver Operating Characteristic (*ROC*) curves (Bradley, 1997). Two variables *sensitivity* (true T-cell epitopes/total T-cell epitopes) and *1-specificity* (false T-cell epitopes/total non-T-cell epitopes) were calculated at different cutoffs. The area under the curve (A_{ROC}) is a quantitative measure of the predictive ability and varies from 0.5 for a random prediction to 1.0 for a perfect prediction. The predictive ability of the models was assessed by LOO-CV on the training set and by external validation on the test set.

Additive models which included different positions before and after the cleavage site were used to assess the importance of flanking amino acids around the C-terminus for accurate proteosome cleavage prediction. Peptide positions were denoted as is shown in Fig. (**2**). Cross terms were omitted as previous studies indicated that the contributions of the positions next to the cleavage site are additive (Altuvia and Margalit, 2000). The ability of the models to discriminate T-cell epitopes from non-T-cell epitopes was assessed by *ROC*-statistics on the training and test sets. The overall performance of the models was very good (all $A_{ROC} > 0.740$, data not shown). Models containing amino acids from both sides of the C-terminus predict better than models which only include flanking positions before the cleavage site. The best performing models for the test set are models p8p9p9'p8' ($A_{ROC} = 0.761$) and p9p9' ($A_{ROC} = 0.759$) (Table **1**).

Our results indicated that positions p9 and p9' are the most significant for the cleavage site. In accordance with these findings, the models derived in the present study show that p8, p9, p9' and p8' are the most influential positions. Cleavage appears after Val, Ile, Tyr, Leu, Lys, Arg, Ala, Phe and Met and/or before Gln, Cys, Glu, Gly, Lys, Arg, Asp, Asn, His and Thr. Arg and Lys make positive contributions at both positions, while Pro, Ser ad Trp contribute negatively at both. The preference for hydrophobic and basic amino acids at the C-termini in our models is compatible with previously reported results based on degradation experiments (Niedermann *et al.,* 1996; Kuttler *et al.,* 2000; Altuvia and Margalit, 2000). These preferences agree with the well established requirements for binding to many MHC class I alleles (Rammensee *et al.,* 1995). Preferences for small (Cys, Gly), polar (Gln, Asn, Thr), positively (Lys, Arg, His) and negatively charged (Glu, Asp) amino acids at p9' are also found in our models. These results, which exclude negatively charged amino acids, are compatible with previously reported preferences at the p9' position (Niedermann *et al.,* 1996; Kuttler *et al.,* 2000; Altuvia and Margalit, 2000). Additionally, the negatively charged aspartic and glutamic acids at p9' position have positive contributions. Among the

amino acids occupying position p8, Glu, His, Cys, Lys and Trp contribute positively and Asp, Tyr, Arg, Phe, Leu and Gln make negative contributions. At the p8' position, Gly, Arg, Glu, Asn, Thr and Ser have positive coefficients, while Tyr, Phe, His, Ile, Met and Trp contribute negatively.

MODELING OF TAP BINDING

We used the additive method to develop a TAP binding prediction model (Doytchinova *et al.,* 2004a). We also evaluated how well this model acts as a pre-selection step in predicting MHC binding peptides. To distinguish between fully and partially TAP-dependent alleles, two data sets were examined. Peptides binding to HLA-A*0201 were selected as representatives of HLA alleles exhibiting partial TAP-dependence and peptides binding to HLA-A*0301 represented fully TAP-dependent HLA alleles.

A set of 163 polyAlanine nonameric peptides was used as a training set. Originally, using the peptide AAASAAAAY as the parent peptide, a set was prepared which included all natural amino acids except cysteine substituted at each position (Daniel *et al.,* 1998). The binding affinities were presented as as $-\log IC_{50}$ values (pIC_{50}). This set was used to develop an additive model for TAP binding (Table **2**).

Two principal components (PC) explain 99.9% of the variance in the set. The most positive contributions to binding belongs to Phe at p9, followed by Phe, Tyr and Trp at p3. The most negative value corresponds to Ser at p9, followed by Pro at p2, Asp and Gly at p9. A TAP binding motif was defined: amino acids that increase TAP binding affinity more than 5 fold (0.699 log unit) were identified as preferred; amino acids that decrease affinity more than 10 fold (1 log unit) were identified as deleterious. No amino acid is strongly preferred at p1, but Glu, Asp and Pro are deleterious. Trp has the highest positive contribution at p2 and Pro and Asp the most negative one. Two groups of amino acids make significant contributions at p3: the first group includes Phe, Tyr and Trp (each making an equal contribution of 1.125 log units) and the second group comprises Ile, Met and Val (coefficients of 0.824). Only Asp and Gly are detrimen-

Table 1. Additive Models for Proteasome Cleavage Prediction

	Модел p9p9'		**Модел p8p9p9'p8'**			
	p9	**p9'**	**p8**	**p9**	**p9'**	**p8'**
A	0.023	-0.012	0.003	0.025	-0.014	0.001
C	-0.031	0.053	0.017	-0.032	0.055	0.005
D	-0.064	0.011	-0.043	-0.065	0.012	-0.009
E	-0.075	0.038	0.078	-0.075	0.037	0.029
F	0.018	-0.029	-0.022	0.019	-0.030	-0.032
G	-0.096	0.035	0.007	-0.097	0.037	0.033
H	-0.049	0.009	0.049	-0.049	0.011	-0.026
I	0.169	-0.046	-0.005	0.169	-0.045	-0.024
K	0.061	0.031	0.010	0.060	0.030	0.006
L	0.100	-0. 043	-0.015	0.096	-0.044	-0.004
M	0.008	-0.020	0.008	0.008	-0.020	-0.013
N	-0.064	0.011	0.006	-0.065	0.012	0.021
P	-0.109	-0.050	0.007	-0.110	-0.049	-0.009
Q	-0.066	0.056	-0.013	-0.066	0.057	0.002
R	0.039	0.022	-0.038	0.041	0.022	0.033
S	-0.092	-0.017	-0.007	-0.092	-0.015	0.013
T	-0.067	0.005	-0.001	-0.066	0.003	0.018
Y	0.134	-0.039	-0.041	0.137	-0.042	-0.044
W	-0.010	-0.011	0.010	-0.009	-0.010	-0.010
V	0.171	-0.006	-0.008	0.172	-0.007	0.009
const	0.101		0.104			

Table 2. Additive Model for TAP Affinity Prediction. The Model Constant is 6.223, r^2 = 0.999, PC = 2

	p1	p2	p3	p4	p5	p6	p7	p8	p9
Ala	0.400	-0.030	-0.240	0.094	-0.173	-0.140	-0.097	-0.178	-0.808
Arg	0.487	0.558	0.347	0.492	0.290	0.527	0.190	-0.016	0.198
Asn	0.444	-0.533	-0.367	0.140	-0.351	0.448	-0.010	0.109	-1.406
Asp	-1.240	-1.074	-1.145	0.191	-0.062	0.022	-0.634	0.109	-1.809
Gln	-0.683	0.558	-0.237	0.316	0.151	0.381	-0.236	-0.174	-0.471
Glu	-1.349	0.035	-0.668	-0.082	-0.351	0.226	-0.053	-0.174	-1.114
Gly	-0.276	-0.791	-1.103	-0.268	-0.430	0.381	-0.685	0.711	-1.687
His	-0.522	-0.706	0.280	0.492	-0.038	0.184	0.093	-0.192	-0.950
Ile	-0.085	0.336	0.824	0.094	0.415	0.749	0.491	-0.091	-0.251
Leu	-0.351	0.637	0.347	0.094	0.591	0.022	0.424	0.146	-0.304
Lys	0.186	0.222	0.125	0.395	-0.086	0.381	-0.394	-0.259	-0.068
Met	-0.027	0.491	0.824	0.316	0.415	-0.189	0.491	0.146	-0.439
Phe	-0.648	0.190	1.125	0.492	0.348	-0.029	0.491	-0.091	1.174
Pro	-1.094	-1.945	0.083	0.395	0.591	0.050	0.792	0.234	-0.633
Ser	-0.073	0.433	0.347	-0.272	-0.086	-0.189	-0.685	0.887	-2.352
Thr	-0.243	0.222	0.011	0.094	0.017	-0.172	-0.146	0.711	-1.292
Trp	-0.419	0.859	1.125	0.492	0.892	-0.136	0.269	0.586	0.308
Tyr	-0.546	0.433	1.125	0.395	0.494	-0.293	0.792	0.074	0.762
Val	-0.012	0.558	0.824	0.249	-0.011	-0.117	0.366	0.146	-0.384
ASC[a]	9.085	10.611	11.147	5.363	5.792	4.636	7.339	5.034	16.410

[a] Absolute sum of contributions.

tal at p3. Low value positive and negative contributions are characteristic of p4. Trp is the preferred amino acid at p5, while Ile is preferred at p6, Pro and Tyr at p7, and Ser, Thr and Gly at p8. There are no strongly disfavoured amino acids at any of these four positions. P9 is very sensitive to changes: Phe and Tyr are favoured, and there are many disfavoured amino acids, including Ser, Asp, Gly, Asn, Thr, Glu, His, and Ala.

Two sets of nonameric peptides were used to test the predictive ability of the additive model for TAP affinity (Fig. **3**). Set A consisted of 47 analogs of the peptide ALAKAAAAV (Daniel *et al.,* 1998). Originally, affinities were presented as IC_{50} values relative to the parent peptide ALAKAAAAV. Set B included 38 nonamers (Daniel *et al.*, 1998) with affinities presented as IC_{50} values relative to the reference peptide RRYNASTEL. The binding affinities of the test peptides were calculated by the additive model and were presented as the logarithm of the relative IC_{50} value ($IC_{50test}/IC_{50reference}$). The correlation between the predicted and experimental $logIC_{50relative}$ (r_{pred}) was used to assess the model predictability. Set A has an r_{pred} of 0.717 and set B has an r_{pred} of 0.832. The high predictive ability of the TAP additive model confirmed the applicability of the additive method for TAP binding affinity prediction.

To assess the TAP contribution to T-cell epitope selection the additive scoring function derived in this study was applied on a set of 317 A*0201 binders, 239 A*0201 non-binders, 76 A*0301 binders and 237 A*0301 non-binders. Binders were extracted from our in-house database AntiJen (Toseland *et al.,* 2005). Non-binders were a gift from Dr. Vladimir Brusic. Receiver Operating Characteristic (ROC) curves (Bradley, 1997) were used to measure the prediction rate of binders *vs.* non-binders. TAP and HLA (Guan *et al.,* 2006) additive scoring functions were applied to predict binders and non-binders. In both cases HLA scoring functions give better predictions than TAP scoring functions (Fig. **4**). However, TAP scores better for fully TAP-dependent A*0301 than for partially TAP-dependent A*0201, A_{ROC} = 0.874 *vs.* A_{ROC} = 0.721. According to the number of false and true negatives at different TAP cutoffs, a lower TAP cutoff ($-logIC_{50}<3.00$) is recommended for A*0201 peptides pre-selection than for A*0301 ($-logIC_{50}<5.00$). Increasing the TAP cutoff drastically increases the number of false negatives for A*0201 but does not affect the number of

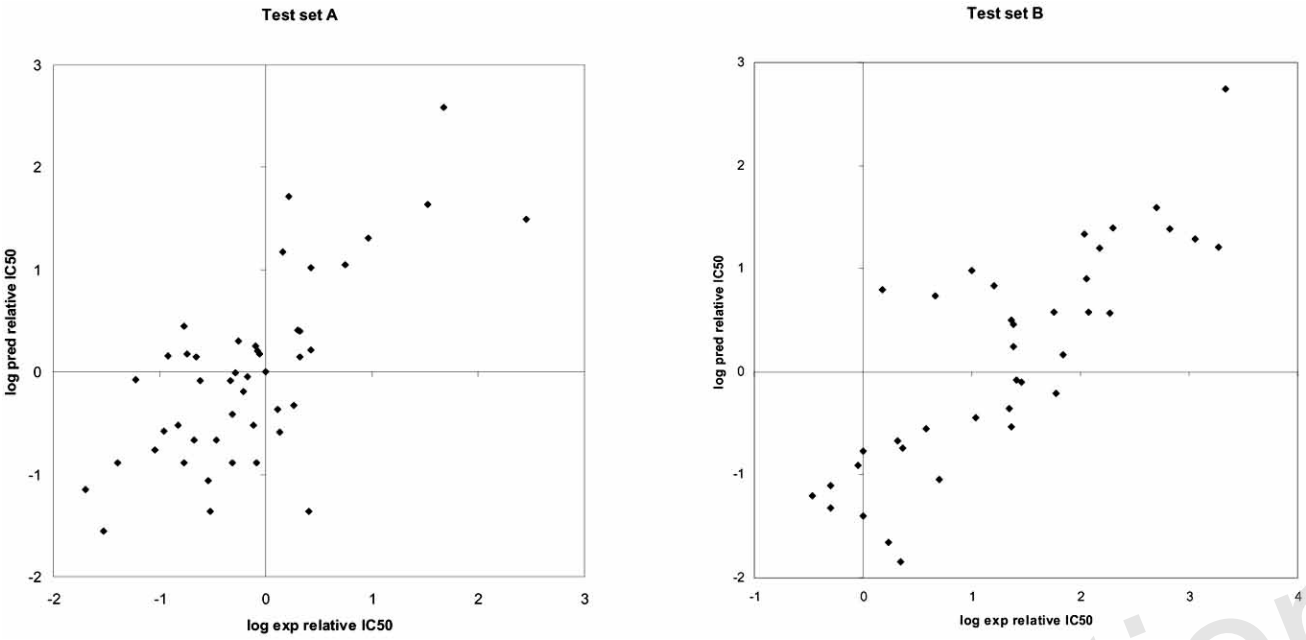
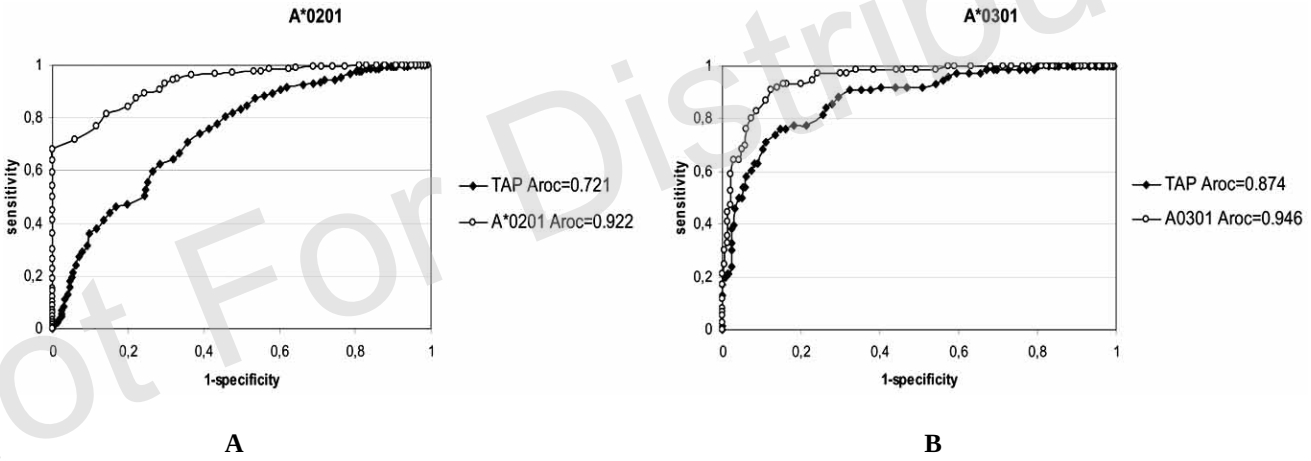


Fig. (3). External validation of the TAP binding model by two test sets. Test set A has r_{pred} = 0.717 (left) and test set B has r_{pred} = 0.832 (right).



A **B**

Fig. (4). *ROC* statistics of the additive models for TAP, HLA-A*0201 (A) и HLA-A*0301 (B) binding affinity prediction.

false negatives for A*0301. Thus, a TAP cutoff of 3.00 eliminates only 24 of the non-binders (10%) for A*0201, whereas, at a TAP cutoff of 5.00, 80 A*0301 non-binders (33%) are eliminated. Unsurprisingly, TAP pre-selection is more efficient for fully TAP-dependent alleles than for partially TAP-dependent alleles.

As TAP transport precedes HLA binding, a conflict will only arise between positions which are deleterious for TAP binding but preferred for HLA binding, but not between TAP preferred and HLA deleterious positions. The absolute sum of contributions indicated that p1, p2, p3 and p9 exhibit the greatest variation in amino acid preference. It is widely assumed that p2 and p9 are the primary anchors and p1 and p3 are secondary anchors for MHC binding. Most HLA alleles prefer peptides with hydrophobic or aromatic amino acids at their C-termini; only the A3 binding motif has positively charged amino acids (Arg or Lys) here. Phe, Tyr and Trp are the preferred amino acids at the C-terminus of TAP binding peptides, whereas Arg makes a small positive contribution and Lys makes a negligible contribution. Ile, Leu and Val exhibit moderate negative values (less than 0.4 log units). Ser, Asp, Gly, Asn, Thr and Glu are all detrimental for TAP binding and this provides a possible explanation as to why few human class I MHC ligands have these amino acids at their C-termini.

There is a great variety of preferred amino acids at anchor p2 in HLA motifs. A2 and A3 supertypes prefer hydrophobic amino acids, A24 prefers aromatic, B7 prefers Pro, B27 prefers positively charged amino acids and B44 prefers negatively charged ones. All these amino acids make positive contributions to TAP binding, except for Pro and Asp. At p2, Pro is a preferred anchor for B7, whereas Asp is pre-

ferred for B44. These are the only points of conflict between TAP and HLA binding preferences. The deleterious effects of Pro and Asp suggest that ligands with Pro and Asp at p2 are unlikely to be transported into the lumen of the ER *via* a TAP-dependent mechanism. Fortunately, these ligands often bear Phe and Tyr at their C-termini, which are strongly preferred by TAP, indicating a potential compensating effect for Pro and Asp.

P1 is the next most sensitive position for TAP binding after p9. It is thought to be a secondary anchor for MHC binding and the side chain occupies pocket A (Ruppert *et al.,* 1993). However, the TAP_19 model (A_{ROC} = 0.563) suggest this position is not overly important for TAP transport. Additionally, the highest negatively contributing amino acids for TAP affinity, Glu and Asp, are common at p1 in many HLA ligands (Rammensee *et al.,* 1995).

Phe, Tyr and Trp at p3 have the highest positive contribution to TAP binding after Phe9, whereas Asp and Gly contribute negatively. The side chain at p3 occupies pocket D in the MHC binding groove and it is thought to be an important secondary anchor (Garboczi *et al.,* 1996). A wide range of amino acids, including Asp and Gly, are available at this position in different MHC ligands, which point to the moderate importance of this position for TAP transport.

Weak amino acid contributions to TAP binding were seen at p4, p5, p6, p7 and p8. Similar results have been found by others (Gubler *et al.,* 1998; Lankat-Buttgereit and Tampé, 1999; Peters *et al.,* 2003). The primary interaction of T-cell receptors (TCR) is with residues 5 to 8 of a class I MHC binding nonapeptide (Garboczi *et al.,* 1996). Thus, antigen recognition by a TCR is in the region of the peptide where TAP exerts minimal selection. Moreover, TAP transport is only one part of the complexity inherent in the emerging picture of class I presentation (Chen and Jondal, 2004; Lautscham *et al.,* 2003).

MODELING OF MHC CLASS I BINDING

The major part of the class I MHC molecule is formed by a transmembrane heavy chain of 44 *kDa* folded into 3 domains α1, α2 and α3 (Krensky and Clayberger, 1996). α1 and α2 form the peptide binding domain, containing the peptide binding groove and the site of interaction with T cell receptors (Jones, 1997). Although not all nine amino acids interact strongly with the binding site, all of them make contact with it (Madden *et al.,* 1993). X-ray data indicate that the MHC peptide-binding site has a 30Å long solvent accessible surface (Madden, 1995), within which six pockets (A to F) have been described (Fig. **5**). Certain pockets are non–polar and make hydrophobic contacts. Others contain polar atoms and could hydrogen bond to bound peptides. Six peptide residues fall into these pockets: they are defined as primary (p2 and p9) and secondary (p1, p3, p6 and p7) anchor positions. The remaining three amino acids (p4, p5 and p8) are solvent accessible and can interact with T cell receptors. They are able to affect MHC binding affinity in several ways: through direct non-bonded interactions with the MHC, by causing conformational changes in anchor residues, and by altering dynamic properties of the whole peptide.

Sequence analysis has shown the peptide domains α1 and α2 to be polymorphic. Twenty residues are the most variable (Parham *et al.,* 1988). Most of these residues contact the peptide, giving MHCs a broad specificity and allowing them to bind a wide variety of peptides (Saper *et al.,* 1991). Sette and Sidney grouped class I alleles into superfamilies based on the overlap between their binding motifs (supermotifs) (Sette and Sidney, 1998). Four superfamilies are known: HLA-A2 (Altfeld *et al.,* 2001), HLA-A3 (Kawashima *et al.,* 1999), HLA-B7 (Coyle and Gutierrez-Ramos, 2001) and HLA-B44 (Sette and Sidney, 1998). Supermotif identification has direct practical implications in epitope-based vaccine development for the prevention of infectious diseases and cancer. Epitope identification is the initial step in the design of epitope based vaccine, and often begins with an *in silico* motif search.

We review here the application of QSAR methods to the definition of A2 and A3 supermotifs. Initially, we applied QSAR methods to peptides binding to the HLA-A*0201 allele (Doytchinova *et al.,* 2002) and then applied them to peptides binding to the HLA-A2 and A3 supertypes. We thus

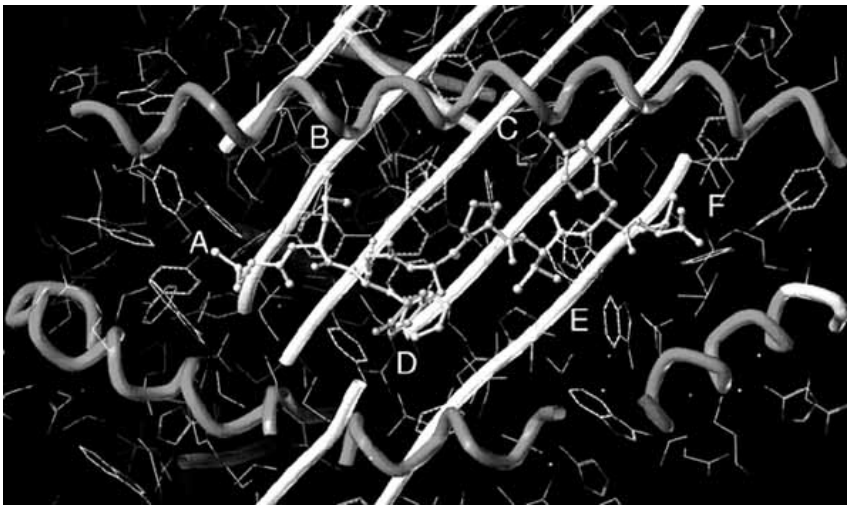


Fig. (5). Peptide binding site on HLA-A*0201 (Madden *et al*., 1993). Alfa-helices are given in dark grey, beta-sheets – in white. The binding site has six pockets denoted from A to F. Peptide LLFGYPVYV (light grey) is shown inside the cleft.

defined extended A2- and A3-supermotifs (Doytchinova and Flower, 2003b; Guan *et al.*, 2003a). The HLA-A2 family is the largest and most diverse allele family at the HLA-A locus, consisting of 55 alleles and is common in all ethnic groups (Sidney *et al.*, 1996a; Ellis *et al.*, 2000; Sette *et al.*, 2001). Within the HLA-A2 family, the most frequent alleles are A*0201, A*0202, A*0203, A*0206 and A*6802. These alleles differ by 1 to 7 amino acids (Schönbach *et al.*, 2000), and these sequence differences alter the peptide binding selectivity of the different A2 alleles. The HLA-A3 supertype covers 44% of the human population and includes 5 main alleles: A*0301, A*1101, A*3101, A*3301 and A*6801 (Sidney *et al.*, 1996a).

Applying the additive method, two types of models were created: one based solely on the amino-acid contributions (amino acids model: AAM) and another based on both amino-acid contributions and amino-acid interactions (amino acids and interactions models: AAIM) (Doytchinova and Flower, 2003b; Guan *et al.*, 2003a). According to the q^2 values the AAMs are more predictive than the AAIMs. This is because certain interactions occur only once. In cross-validation, they appear as missing terms in the equation used for affinity prediction. Prediction error is proportional to the number of missing terms. Missing terms in AAMs are less frequent and so their prediction rate is higher. In contrast, r^2 was slightly lower for the single amino acid models than for the AAIM. The decrease in r^2 shows that the amino acid side-chain interactions are important for the explanation of variance and should be considered in the modelling of the binding process. The statistical parameters for these models are collected in Table **3**.

Amino acids with contributions greater than 0.2 were considered as preferred for a particular allele at the specific position and those with contributions lower than –0.2 were considered as deleterious. Residues identified as preferred for two or more A2/A3-alleles without being deleterious for others were considered as preferred. Residues identified as deleterious for two or more alleles were considered as delete-

Table 3. Statistics of the Additive Models

Model	n	q^2	PC	SEP	r^2	SEE	F	MAE[c]
HLA-A2 Superfamily								
A*0201								
AAM[a]	335	0.377	6	0.694	0.731	0.456	148.66	0.546
AAIM[b]	340	0.337	5	0.726	0.898	0.285	588.88	0.573
A*0202								
AAM	69	0.317	9	0.606	0.943	0.193	109.10	0.546
AAIM	68	0.283	2	0.621	0.748	0.368	96.65	0.511
A*0203								
AAM	62	0.327	6	0.841	0.963	0.197	239.30	0.652
AAIM		<0.300						
A*0206								
AAM	57	0.475	6	0.576	0.989	0.085	728.52	0.443
AAIM		<0.300						
A*6802								
AAM	46	0.500	7	0.647	0.983	0.119	313.30	0.517
AAIM		<0.300						
HLA-A3 Superfamily								
A*0301								
AAM	72	0.436	6	0.680	0.959	0.181	246.90	0.504
AAIM	70	0.305	4	0.699	0.972	0.136	557.37	0.527
A*1101								
AAM	62	0.458	2	0.572	0.829	0.321	143.00	0.507
AAIM	62	0.428	3	0.593	0.977	0.119	821.10	0.467
A*3101								
AAM	30	0.482	3	0.710	0.892	0.325	71.36	0.502
AAIM	31	0.453	6	0.727	0.990	0.098	399.96	0.602
A*6801								
AAM	38	0.531	4	0.594	0.959	0.175	194.85	0.418
AAIM	37	0.370	4	0.664	0.974	0.136	297.48	0.485

[a]AAM: amino acids model; [b]AAIM: amino acids and interactions model; [c]mean absolute error.

rious in the common motif. The supermotifs defined by the additive method are given in Fig. (**6**).

a)

Preferred	F	-	I	G	-	IL	I	F	V
	1	2	3	4	5	6	7	8	9
Deleterious	-	-	T	-	W	S	-	D	A

b)

Preferred	FK	L	IVL	GT	IL	ILY	HI	FKT	VL
	1	2	3	4	5	6	7	8	9
Deleterious	-	VT	CHT	AN	SWY	QS	LT	DER	AT

c)

Preferred	SM	IT	F	FRQ	-	S	FI	RLY	R
	1	2	3	4	5	6	7	8	9
Deleterious	ALQ	N	L	S	GHS	-	-	KSE	Y

Fig. (6). A2-supermotif: a) based on A*0201, A*0202, A*0203, A*0206 and A*6802 alleles; b) based on A*0201, A*0202, A*0203 and A*0206 alleles. c) A3-supermotif.

A2 Supermotif

P2 and C-terminal (p9) are considered as primary anchor positions. Our models indicated significant differences in the amino acid preferences at p2 for A2 alleles. Hydrophobic aliphatic residues such as Leu, Met and Val have well known preferences for this position (Falk *et al.,* 1991; Madden *et al.,* 1993; Ruppert *et al.,* 1993; Parker *et al.,* 1994). However, our models show that Leu and Met are preferred amino acids only for A*0201, A*0202 and A*0203. Leu is deleterious for A*6802 and Met is deleterious for A*0206 and A*6802. Val and Thr are preferred for A*6802. Sidney and colleagues also reveal similar differences in p2 specificities although not so strong as in the present study (Sidney *et al.,* 2001). Comparing the residues forming the pocket B in the different alleles four differences are evident (Schönbach *et al.,* 2000). Three of them (Glu^{63}, Lys^{66} and His^{70}) are disposed at the pocket rim and one (Phe^{9}) at the inner wall. The $Phe^{9}\rightarrow Tyr^{9}$ substitution makes the pocket shallower and long side chains, such as Leu and Met, are no longer accommodated here. Molecular modeling studies hypothesize a possible conformational shift of the aromatic ring of Tyr^{9} into the cavity (Sudo *et al.,* 1995). This conformational change would narrow the size of the B pocket and weaken the entirely hydrophobic state of this pocket. The preferred Val and Thr for A*6802 allele, being deleterious or negative for the other A2-sypertype molecules, denote another point of discrepancy between A*6802 and the remaining A2 alleles. Comparison of the residues forming pocket B shows identity between A*6802 and A*1110, A*2502, A*2613, A*6604, A*6601, A*6602, A*3403, A*3404, A*3402 alleles. None of them except for A*6802 was classified as A2-like allele. At the C-terminal there is a good agreement between the preferences of different alleles. Val is the favored amino acid at this position, Ala is deleterious. Pocket F appears to be the most conserved pocket in the HLA binding cleft (Madden, 1995). The side chain of Tyr^{116} occupies the end of the pocket F and is uncharged, so that the binding site is complementary to small hydrophobic side chains (Saper *et al.,* 1991; Madden *et al.,* 1993).

P1, p3, p5, p6 and p7 are secondary anchor positions (Ruppert *et al.,* 1993; Madden, 1995). Phe is the only one preferred amino acid for p1 in the common motif. Lys is preferred for all alleles except for A*6802. For the last allele Lys is apparently deleterious. The main differences in the amino acids sequences forming this pocket are residues 63 and 66. Glu^{63} and Lys^{66} are substituted for Asn^{63} and Asn^{66} in A*6802 allele (Schönbach *et al.,* 2000). Obviously, the negatively charged Glu^{63} favored the presence of positively charged Lys at p1, while the neutral Asn^{63} is not electrostatically complementary to Lys. Ile is the only preferred amino acid at p3 and Thr is the common deleterious one. Leu and Val are preferred for A*02 alleles but is deleterious for A*6802. P3 side chains of bound peptides fall into pocket D which is a hydrophobic cavity (Bjorkman *et al.,* 1987). There is only one difference in the sequences forming this pocket. Leu^{156} in A*0201 and A*0206 is substituted for Trp in A*0202, A*0203 and A*6802 making a bulky ridge across the center of the cleft. Leu was found to be a preferred residue at p5 for affinity to A2-supertype molecules except for A*6802 where it is negative but not deleterious. Trp is deleterious for three of the five MHC molecules. Ile and Leu are preferred at p6 and Ser is deleterious. The side chain of p6 falls into pocket C. The most dramatic difference between A*6802 and A*02 alleles concerns this pocket. A deep negatively charged pocket at A*6802 is formed by the substitution of Asp for His at position 74 and Gln for His at position 70. This pocket seems suited to bind polar atoms, especially a positively charged side-chains or N-terminus (Lys) (Garrett *et al.,* 1989). Unfortunately, we could not find any published peptide that had Lys at p6 which had been tested for affinity to A*6802. For affinity to A2-supertype molecules Ile is preferred at p7. The side chain at p7 falls into pocket E. Two thirds of the surface area in this pocket is hydrophobic, but Arg^{97} provides a large polar patch on one side of the pocket (Saper *et al.,* 1991). Pocket E can accommodate a variety of complementary peptide side chains, but an incompatible side chain need not prevent complex formation (Madden, 1995).

P4 and p8 are solvent-exposed and may form contacts with the TCR (Madden, 1995). Gly is preferred here. Thr is preferred or positive for the A2-supertype alleles except for A*6802. Phe is a preferred common residue at p8 and Asp is deleterious for four of the five A2 molecules. The A2 supermotif is presented in Fig. (**2a**).

Certain discrepancies between A*6802 and A*02 molecules concerning the amino acids preferences at p1-p9 were

seen in the present study. These discrepancies throw doubt on whether the A*6802 allele belongs to the A2-supertype. The sequence comparison showed that there are only one or two differences in the residues forming the six pockets of A*0201, A*0202, A*0203 and A*0206 molecules. The number of these differences between A*6802 and A*02 molecules is seven residues. Five of them concern pockets A, B and C and are so substantial that they alter the amino acids preferences at the primary anchor p2 and the secondary anchors p1 and p6. The preferred Val and Thr for p2 brings the A*6802 allele closer to the A3-supertype (Sidney *et al.,* 1996b) rather than to the A2-one. But the A3 supermotif requires positively charged residues, such as Arg and Lys, at the C-terminus (Sidney *et al.,* 1996b), which is not true in the case of the A*6802 allele. Obviously, A*6802 is a intermediate allele standing between A2 and A3 supertypes: in anchor p2 it is closer to A3 and in anchor p9 it is nearer to A2. Excluding A*6802 allele, the redefinition of the preferred and deleterious amino acids expands the A2-supermotif (Fig. **2b**). The expansion concerns all positions and especially the anchor p2. One to three new amino acids are added to each position's preferred and deleterious amino acids.

A3 Supermotif

P2 and p9 are generally accepted as primary anchors for the A3 superfamily (Garrett *et al.,* 1989; Matsamura *et al.,* 1992; Falk and Rőtzschke, 1993). The peptide side chain at p2 falls into pocket B and the C-terminal is buried in pocket F (Saper and Bjorkman, 1991; Vasmatzis *et al.,* 1996). Peptides usually have a positively charged residue Arg or Lys at p9 and a variety of hydrophobic residues at p2. A peptide binding motif for the HLA-A3 superfamily has been defined previously by Sidney and colleagues (Sidney *et al.,* 1996b) and Rammensee and colleagues (Rammensee *et al.,* 1995). The supermotif defined in our studies, while in good agreement with previous supermotifs, is more extensive, covering all the nine positions that contact the MHC molecule (Guan *et al.,* 2003a).

P1 is a secondary anchor position. According to the additive method Ser and Met are preferred here. Despite the wide variation in preferences at p2, Ile and Thr were found to be preferred for two of the alleles without being deleterious for the other two. The wide variation for p2 is explained by the polymorphism of residues forming pocket B. Phe^{9} in A*0301 is substituted to Tyr^{9} in A*6801 and A*1101, and to Thr^{9} in A*3101 (Schönbach *et al.,* 2000). The hydroxyl group of Tyr^{9} points towards the inside of the pocket and prevents larger amino acids from reaching the bottom of the pocket (Sudo *et al.,* 1995). Because of this, larger residues like Leu are deleterious for A*6801 and A*1101 but are preferred for A*0301. The change from Glu^{63} to Asn^{63} in A*6801 and A*1101 also changes the conformation of the pocket and stops large amino acids from binding (Vasmatzis *et al.,* 1996). A previous study of pocket B revealed Val^{67} was reoriented in A*6801 and affected amino acid selection (Guo *et al.,* 1993). P3 prefers the hydrophobic residue Phe. Sidney and co-workers (Sidney *et al.,* 1996b) found that peptides with aromatic residue, like Tyr, Phe and Trp, had a 31 fold increase in binding affinity to A*0301. Phe, Arg and Gln are favoured at p4. No amino acid is favoured at p5. Ser, Gly and His are disfavoured here. This position as well as p6 are not particularly important in determining the affinity of peptide binding but may participate in T cell recognition. Ser is well accommodated at p6. P7 is another secondary anchor position (Rammensee *et al.,* 1995). Hydrophobic residues are preferred here. Phe and Ile are strongly preferred by A*0301 and A*1101. Peptide binding studies showed either p3 or p7, together with residues at p2 and p9, induced stable binding of the peptide (Sidney *et al.,* 1996b). Arg, Tyr and Leu were favoured at p8, while Ser, Lys and Glu were deleterious. Positively charged amino acid Arg is the common preferred amino acid at p9. A*6801 and A*3101 preferred Arg, A*1101 favoured the smaller residue Lys, while A*0301 accepted both. Tyr was deleterious at p9, possibly because its aromatic ring was too large for the pocket.

Design of Superbinders

Working solely with peptide data from the literature has a number of drawbacks and weaknesses: reported peptides are highly biased in terms of their position-dependent amino acid composition, often favouring hydrophobic sequences. This arises, in part, from pre-selection processes that result in self-reinforcement. Binding motifs are often used to reduce the experimental burden of epitope identification. Very sparse sequence patterns are matched and the corresponding subset of peptides tested, with an enormous resulting reduction in sequence diversity. This bias is more prominent at the anchor positions, which usually have highly restricted sets of amino acid types. In addition, when working solely with literature data it is not possible to test the predicted binding affinities of newly-designed peptides.

In one of our studies, we determined the binding affinities of a set of 90 nonamer peptides to the MHC class I allele HLA-A*0201 (Doytchinova *et al.,* 2004b) using an in-house FACS-based MHC stabilization assay (Lopes *et al.,* 2003). A good correlation was found between the literature radiolabeled competition assay IC_{50} values and the BL_{50} values from our experiments. Using our BL_{50} values, we derived an additive QSAR model for peptide interaction with HLA-A*0201. The model was applied to design new A2-binding peptides. For this purpose we selected the preferred amino acids at each position and made combinations of them. For some of the positions - 1, 2, 3, 4, 5 and 9 - there were peptides which were obviously strongly preferred, but for other positions - 6, 7 and 8 - a wide range of amino acids were preferred more or less equally. We selected Leu for p2 and Val for p9 as anchors. For p1 Ile and Phe were selected; for p3: Phe, Asp and Trp; for p4: Pro and Asp; for p5: Phe, Leu and Ile; for p6: Pro, Val and Phe; for p7: Pro, Val and Ile; and for p8: Pro, Glu, Thr, Asp and Ser. The combination of all preferred amino acids generated 1620 peptides. Their affinities were predicted by the additive model and the affinities of the top 10 high binders were tested experimentally. The test peptides and their predicted and experimental affinities are given in Table **4**. Notably, these ten peptides all had BL_{50} values higher than those of the best peptides in the training set, with the pre-eminent test peptide having a measured binding affinity more than two orders of magnitude greater than that of the best binder from the training set.

The measured values of the predicted BL_{50} values were highly than our estimates indicating that a possible synergis-

tic effect may operate between amino acids at different positions of the peptide. The newly designed superbinders have much higher affinities than a simple sum of amino acid contributions from different positions might suggest. This phenomenon is an example of positive enthalpic cooperativity (Calderone and Williams, 2001). Generally, where multiple weak noncovalent interactions hold a molecular complex together, the enthalpy of all the individual intermolecular bonding interactions is weakened by extensive intermolecular motion. The noncovalent complex between a peptide and a protein is an excellent example of such a system. As additional interaction sites generate a more strongly bound complex, intermolecular motion is dampened, with all individual interactions becoming more favorable.

Table 4. Superbinders to HLA-A*0201 Designed by the Additive Method

Peptides	pBL_{50} Predicted	pBL_{50} Experimental
ILDPFPPTV	6.786	8.170
ILDPIPPTV	6.534	7.296
ILDPFPVTV	6.755	8.654
ILDDFPPTV	6.631	7.083
ILDDLPPTV	6.367	7.144
ILDDFPVTV	6.600	7.155
ILDPFPPEV	6.836	7.682
ILDPFPPPV	6.685	7.442
ILDPFPITV	6.699	8.139
ILDPLPPTV	6.522	7.145

Experimentally, at least for other systems, the trade-off between intermolecular motion and enthalpic interactions has been shown to account for the way in which entropy and enthalpy compensate for each other. Additionally, according to the fragment-based drug discovery approach, when two fragments binding to different but adjacent sites in an enzyme, if joined together in an ideal fashion then the resulting affinity of the joined molecule will exceed significantly the sum of the fragment affinities (Murray and Verdonk, 2002). The reason is that a fragment loses a significant amount of rigid body rotational and translational entropy when it forms a complex. As the entropy loss has a weak dependence on molecular weight, the sum of the fragment affinities will include two unfavourable rigid body entropy terms while the affinity of the joined molecule will include only one. A peptide binding to a MHC protein is an archetypal example of multiple site binding (Fig. **7**). The six pockets on the binding cleft form a proximal multiple site. The preferred amino acids for each pocket are the fragments. If they are bound together in an ideal fashion, the affinity of the joined molecule should be substantially greater than the sum of the fragment affinities. This approach often is used in the drug discovery (Shuker *et al.,* 1996; Hajduk *et al.,* 1997; Rao and Whitesides, 1997; Rao *et al.*, 1998; Maly *et al.*, 2000; Schaschke *et al.*, 2001) but for the first time we applied it to design MHC superbinders.

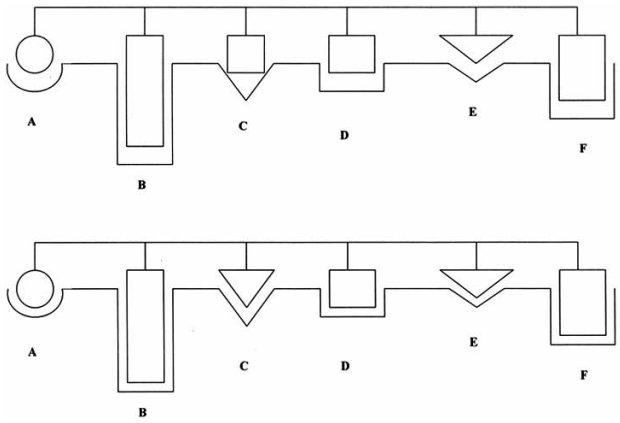


Fig. (7). Peptide binding groove on MHC protein as a proximal multiple binding site. Peptide with non-optimal amino acid binds loosely (up), while peptide with optimal amino acids at each position binds tightly (down).

MODELING OF MHC CLASS II BINDING

Peptides that bind to MHC class II molecules are usually between 10 and 20 residues long, with sizes between 13 and 16 amino acids being the most frequently observed (Rudensky *et al.*, 1991; Hunt *et al.*, 1992; Chicz *et al.*, 1992; Chicz *et al.*, 1993). X-ray data from peptide/MHC class II (Dessen *et al.*, 1997) and TCR/peptide/MHC class II complexes (Hennecke and Wiley, 2002) indicate that nine amino acids are bound in an extended conformation deep in the binding groove of HLA-DR4. A dozen hydrogen bonds between MHC -helices and peptide main chain carbonyl and amide groups are formed. There is one deep pocket that binds the side chain at peptide p1 and there are four shallow pockets that bind side chains at p4, p6, p7 and p9. Side chains at p2, p3, p5 and p8 project prominently toward the T cell. The peptide binding groove of class II molecules is open at both ends and this allows a given peptide to bind in many different ways. This multiple binding ability of peptides results in a lower accuracy for prediction methods compared with those for class I peptides (Brusic *et al.*, 1998).

We have applied the additive method to a set of 82 peptides of 16 amino acids, or less, which bind to the HLA-DRB1*0401 molecule (Doytchinova and Flower, 2003a). In order to address the problem of multiple potential binding, an iterative self-consistent (ISC) PLS-based algorithm was used to select the binding set. Eighty percent of the peptides formed the training set (66 peptides) and 20% a test set (16 peptides). Another set of peptides, all longer than 16 amino acids, was used as a second test set (14 peptides).

The training set of 66 long peptides was presented as a set of nonamers accompanied by the pIC_{50} values of the parent peptide. Only nonamers bearing anchor amino acids (Tyr, Phe, Trp, Leu, Ile, Met, Val) at p1 were selected. The matrix was solved by PLS. LOO-CV is applied to extract the optimum number of components subsequently used to gen-

erate the non-cross-validated model. The previous model is used to predict pIC_{50} values and a new set is extracted. The best predicted nonamers were selected for each peptide, i.e. those with the lowest residual between the experimental and predicted pIC_{50}. The new set is compared to the previous one; if they are the same the final model is obtained. Otherwise, the selection is repeated. The coefficients in the final non-cross-validated model represent the quantitative contributions of each amino acid at each position. The first model had poor predictivity: $q^2 = 0.152$, PC = 1, $r^2 = 0.396$, n=185. Self-consistency was achieved on the seventh iteration. The final model had excellent predictive powers: $q^2 = 0.716$, PC = 4, $r^2 = 0.967$.

All class II prediction methods should be able to overcome the so-called multiplicity problem. This arises both from the indeterminacy of the problem - we do not know a priori which subsequence is the dominant binder - and from the possible degeneracy of the binding process it self. Where a single dominant binding sequence is absent, the measured affinity may be a canonical average of the binding exhibited by several subsequences. These phenomena arise from the binding groove of class II molecules being open at both ends. We may posit that, from a thermodynamic viewpoint, the actual nonameric binding subsequence should have the highest $pIC_{50,}$ or lowest binding energy, among all the nonamers originating from the same long parent peptide. Our analysis of the training set indicates however that the predicted value which is closest to the experimental pIC_{50} is rarely the highest predicted value. We tried three different selection rules to deal with this problem when applied to the test sets: mean, highest value (max) and a combination of both (combi). The last rule selects the mean pIC_{50} when the difference between the highest and lowest predicted pIC_{50} is less than one log unit. Otherwise, it selects the highest predicted value. For both test sets the highest predictivity is given by the combination rule with $r_{pred} = 0.593$ (test set I) and $r_{pred} = 0.655$ (test set II). The performance of the combination rule is not surprising, because when an easily distinguished good binder is not available in the peptide sequence, the binding affinity is a degenerate average of affinities from several binding subsequences.

MHCPred SERVER

To facilitate online T-cell epitope prediction, the models derived by the additive method were implemented as a web server. MHCPred is freely available through the URL http://www.jenner.ac.uk/MHCPred (Guan *et al.,* 2003b; Guan *et al.,* 2006). The server contains models for 11 human MHC class I alleles (HLA-A*0101, HLA-A*0201, HLA-A*0202, HLA-A*0203, HLA-A*0206, HLA-A*0301, HLA-A*1101, HLA-A*3101, HLA-A*6801, HLA-A*6802 and HLA-B*3501), 3 mouse MHC class I alleles (H2-D^b, H2-K^b and H2-K^k), 3 human MHC class II alleles (DRB1*0101, DRB1*0401 and DRB1*0701) and 6 mouse MHC class II alleles (I-A^b, I-A^d, I-A^k, I-A^s, I-E^d and I-E^k). The model for TAP binding affinity prediction also was included in the server. The server accepts protein sequences in plain text format. Two types of models were included in the server: the single amino acid models (this only considers the contributions of the amino acids) and the amino acid with interaction models (which takes into account the contributions of the single amino acids and the 1-2 and 1-3 interactions). Preferred residues at each position can be entered.

The output is arranged in a table and the input sequence is printed at the beginning of the results table. There are two ways to list the output peptides: (i) in ascending order of IC_{50} (nM) values or (ii) according to their position in the input sequences. An IC_{50} cut-off value could be selected. Peptides with predicted binding affinities <500 nM are good binders, whereas those with affinities >5000 nM are considered non-binders. If the user does not enter any value, all the peptides generated from the input sequence will be listed. The binding affinities of those with IC_{50} >5000 nM are not shown, and are replaced by "-". Predicted $-\log IC_{50}$ values are also shown in the table output. There is an option for prediction of binding of mono- and di-amino acid mutations of a peptide. MHCPred takes a single nonamer peptide as the input, substitutes the amino acid at a user-specific position with each of the 20 amino acids and calculates the binding affinities of the new peptides. This option is useful in comparing the binding affinities of heteroclitic analogues of the test peptide. According to the number of missing terms in the model, MHCPred calculates the confidence of prediction for each peptide as a normalized percentage. This feature helps the user to eliminate false-positive predictions and makes the prediction more reliable.

EpiJen SERVER

The next generation of T cell epitope identification methods will focus on integrated multi-step approaches, which subsume proteasome cleavage, TAP transport and MHC binding. The advantages of such integrated methods are higher accuracy and a lower rate of false positive predictions, although they may generate more false negative predictions due to the use of incomplete training sets or high thresholds for individual steps. We have developed a multi-step algorithm for T cell epitope prediction, which we call EpiJen. The method is applied to a set of overlapping peptides generated from a whole protein sequence and acts as a series of filters which successfully reduce the number of potential epitopes. The final set of peptides needed to be tested for epitopes rarely includes more than 5% of the whole sequence. We combine all additive models for binding affinities prediction to human MHC class I alleles and make them publicly available *via* the EpiJen server for T cell epitope prediction (http://www.jenner.ac.uk/EpiJen) (Doytchinova *et al.,* 2006).

The dataflow in EpiJen is shown in Fig. (**4**). Initially, the protein is chopped into overlapping decamers and processed by a proteasome cleavage additive model. A previously derived and tested p1p1' model, as described above (Doytchinova and Flower, 2006) is used. The model takes into account only the contributions of the residues next to the cleavage site: C-terminus and the next aa. Two thresholds, 0.0 and 0.1, can be used here. Threshold 0.0 is recommended for alleles which prefer Phe or Trp at the C-terminus: HLA-A*24, HLA-B*07, HLA-B*27, HLA-B*35, HLA-B*51 and HLA-B*53. The epitopes for other alleles are predicted accurately at a threshold of 0.1. This initial step has a powerful filtering ability: between one half and two thirds of the true negatives were eliminated by this step. The "cleaved" pep-

tides, present as nonamers, are then passed to the next filter: the TAP binding additive model.

The TAP binding additive model has been derived and tested previously (see above) (Doytchinova *et al.*, 2004a). A threshold of 5.00 is recommended for both fully and partially TAP-dependent alleles. Pro and Asp at anchor position 2 has a strong negative effect on TAP binding (Doytchinova *et al.*, 2004a). For that reason, a threshold of 3.0 is recommended for epitopes binding to HLA-B*07, HLA-B*35, HLA-B*40, HLA-B*44, HLA-B*51 and HLA-B*53. The filtering ability of the TAP step is low. Up to 10% of the true negatives are eliminated here. The "transported" peptides move to the next filter: MHC binding.

EpiJen includes 18 additive models which can be used to predict binding to different HLA-A and B alleles. Certain models were developed for single alleles and others developed for allele families. Quantitative data (continuous values like IC_{50}s) were available for certain alleles, for the rest only sequences of binders were known (discontinuous values). The additive models based on continuous values were derived by multiple linear regression (MLR) and those based on discontinuous values by discriminant analysis (DA). The filtering ability of this step is significant: approximately 25-30% of the true negatives are eliminated here. The thresholds for this step are 0.5 for the DA models and 5.3 for MLR models. These thresholds can not be altered by the user. They seek to reduce the number of false positives in long protein sequences.

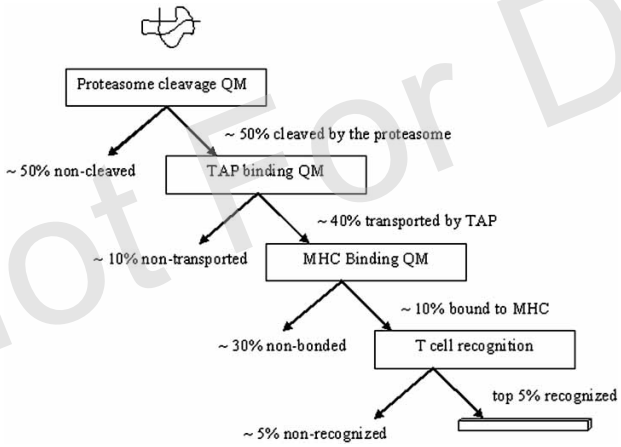


Fig. (8). Dataflow in EpiJen.

All peptides which are presented by MHCs on the cell surface after being cleaved by the proteasome and transported by TAP could potentially be T cell epitopes. However, only a small number of all possible epitopes are actually immunogenic. To reduce the number of false positives we tested different thresholds, which we defined as percentages of available peptides sourced by one protein. The top 5% threshold performed best, giving 85% sensitivity; we recommend it and use it as a default value for this step. Optional are thresholds 2, 3 and 4%.

A set of 160 epitopes and their source proteins were collected from AntiJen (Toseland *et al.*, 2005). They were restricted by the human MHC allele families: HLA-A*01, HLA-A*02, HLA-A*03, HLA-A*11, HLA-A*24, HLA-A*33, HLA-A*68, HLA-B*07, HLA-B*27, HLA-B*35, HLA-B*40, HLA-B*44, HLA-B*51 and HLA-B*53. Six epitopes were promiscuous. Only proteins consisting of less than 1000 amino acids were used in the study. The thresholds were selected as follows: at step 1 (proteasome cleavage) a value of 0.0 was chosen for HLA-A*24, HLA-B*07, HLA-B*27, HLA-B*35, HLA-B*51 and HLA-B*53, and 0.1 for the rest; at step 2 (TAP transport) a value of 3.0 for HLA-B*07, HLA-B*35, HLA-B*51 and HLA-B*53, and 5.0 for the rest; at step 3 (MHC binding) a value of 0.5 was selected for HLA-A*24, HLA-B*27, HLA-B*40 and HLA-B*44, and 5.3 for the rest. For the final step (epitope selection) four thresholds were tested: top 2% to 5%. As the number of non-epitopes generated from each protein was significantly higher than the number of epitopes, only two parameters – *sensitivity* ((true positives/(true positives + false negatives)) and *positive predictive value* (*PPV*) ((true positives/(true positives + false positives)) – were used for comparison. Parameters *accuracy* ((true positives + true negatives)/total) and *specificity* ((true negatives/(true negatives + false positives)) could be misleading. If 98% of the peptides in one source protein are non-epitopes, a model that simply predicts everything as non-epitope will not be very useful, yet it will nonetheless have an overall *accuracy* of 98% and a *specificity* of 100%.

The true positives were 141 (5% cutoff), 132 (4% cutoff), 123 (3% cutoff) and 114 (2% cutoff). False negatives were 25, 34, 43 and 52, while the false positives decreased from 2743 to 2173, 1618 and 1060, respectively. The parameter *sensitivity* varies from 69% (at 2% cutoff) to 85% (at 5% cutoff). The parameter *PPV* diminishes from 10% (at 2% cutoff) to 5% (at 5% cutoff). Thus, our tests indicate that a 5% threshold at the final epitope selection step is sufficient to generate an 85% epitope prediction. This means that by using EpiJen, one need only test 5% of the whole sequence in order to predict 85% of available epitopes.

DISCUSSION

The identification of T-cell epitopes remains a critical step in the development of peptide-based vaccines (Luckey *et al.*, 1998). The first step of such studies is usually *in silico* prediction of potential MHC binders from the sequence of a studied protein, followed by labor-, time- and resource-consuming experiments to verify the natural processing, presentation and T-cell recognition of the predicted peptides. As the veracity of initial *in silico* predictions improves, so subsequent "wet lab" work becomes faster, more efficient, and, ultimately, more successful. A wide range of computer-based algorithms has been developed to help predict T-cell epitopes (for reviews see Schirle *et al.*, 2001; Golgberg *et al.*, 2002; Flower, 2003). In the present review we summarize some of our results, which were derived by applying the additive method to different immunological problems, like proteasome cleavage, TAP binding, MHC class I and class II binding and supermotif definitions.

The additive method is a QSAR technique. QSAR procedures are a powerful, if under-used, tool for *in silico* prediction in bioinformatics. QSAR has found much application, however, in computational drug design, where it can func-

tion as either an engine of interpolation or extrapolation. In interpolation, it can describe the properties of novel or extant molecules, peptides in our case, within a window of measured properties, but it can also be used to explore beyond those boundaries, most often being used to enhance binding affinity. The property of extrapolation into novel property space is the property of QSAR we have exploited in the design of superbinders study.

In most of our papers we have shown that the interpolative powers of our approach work effectively, capturing the essence of prediction (Doytchinova *et al.*, 2002; Guan *et al.*, 2003a; Doytchinova and Flower, 2003a; Doytchinova and Flower. 2003b). Alternatively, the additive method can be used to effectively increase binding affinity in a rational and directed manner, allowing us to design a series of so-called superbinders and, in turn, to use these to explore the effect of systematic substitution of dominant anchor positions. The list of tolerated anchors can be extended to a much larger set than has been commonly envisaged. This has two main implications. First, we need to refine our understanding of the role anchor residues play in peptide binding to MHCs. Second, we need to develop more sophisticated models of binding than those offered by binding motifs as *in silico* prediction devices.

The rational design of high or superbinding peptides is a technique with wide application in a variety of immunological settings. It is a further vindication of the utility of our approach in the prediction of peptide-MHC binding affinity, the principal prerequisite for proteinacious epitopes. Peptides presented by HLA-A2, in particular, would be useful from a vaccination standpoint as they would give rise to immune responses in a high proportion of the HLA diverse population. Perhaps more important, however, is the ability to engineer epitopes with special properties dependent on enhanced or modulated affinity. These might include augmenting the immunogenicity of potential cancer vaccines derived from cancer antigen epitopes, or designing high affinity epitopes, responses to which are reported to be less dependent on CD4 help (Franco *et al.*, 1994). Alternatively, one could design effective and efficacious competitor peptides able to block detrimental responses, as has been done in a murine diabetes model (von Herrath *et al.*, 1998).

We have demonstrated that systematic monosubstitution of high binding peptides produces peptides missing traditional anchors yet retaining high affinity (Doytchinova *et al.*, 2004b). The relative importance of the anchor residues should thus be rethought. One does not require traditional anchors if the rest of the peptide is sufficiently optimised, either artificially, as in this case, or by chance in naturally occurring epitopes. Instead, one should seek more sophisticated and comprehensive models of binding, which are better able to account for all such possibilities. This helps to explain why many observed epitopes are missed when using just anchor motif-based epitope prediction programs. Flexibility as to which amino acids can be tolerated at the anchor positions increases the effective number of peptides that can be presented by a given HLA allele. This augments the chance that a T cell response can be mounted by every individual to each antigen or pathogen. It also has other implications, eg. if multiple amino acids in an epitope can influence peptide – HLA interaction, this may increase opportunities for pathogen escape from CD8 responses *via* alteration of peptide binding to MHC (Borrow and Shaw, 1998).

Greatest advantages of the additive method and related quantitative matrix methods are their easy use and interpretation. All methods derived by us during recent years have been compiled and are freely accessible *via* our servers MHCPred and EpiJen. EpiJen offers many advantages, compared to other integrated methods for T-cell epitope prediction. First, a large quantity of experimental data (more than 2500 peptides) has been used to develop the models. Second, based on the additive method EpiJen combines two well known, widely used methods in drug design (Doytchinova *et al.*, 2002), which have generally proven to be both reliable and predictive: the Free-Wilson method (Free and Wilson, 1964) and PLS (Wold, 1995). Finally, and most importantly, EpiJen uses its models as successive filters: negatives are eliminated at each step rather than their score being summed in order to exceed a global threshold. This is in contrast to alternative methods (Peters and Sette, 2005; Larsen *et al.*, 2005). The combined score, as used by SMM (Peters and Sette, 2005) and NetCTL (Larsen *et al.*, 2005), obscures the final result, because a low (or even negative) TAP and/or proteasome score could be compensated for by a high MHC score. The cellular antigen processing pathway, as modeled in EpiJen, works in a hierarchical or successive manner not in parallel. Peptides that have been eliminated at any of the steps do not continue to the next step. EpiJen can thus be thought of as a more mechanistically meaningful model of overall antigen presentation than other available methods. EpiJen is both a more adaptable and a more flexible approach, which should prove a significant conceptual advantage as combination methods, such as this, evolve in the coming years.

It is well known that "all models are wrong, yet some of them might be useful". Informatic modelling, such as we describe, follows the accumulation of knowledge about a particular mechanism. As knowledge improves, so models will improve. Antigen processing is a very complicated cascade of cellular events. It is clear that cleavage by the proteasome is only one event in antigen presentation. There are many other events, and many of these are proteolytic in nature. Analyses of peptide generation and T-cell epitopes expression in proteasome-inhibited cells suggest that cytoplasmic proteases other than proteasomes may also be involved in antigen processing pathway (Vinitsky *et al.*, 1998; Luckey *et al.*, 1998; Luckey *et al.*, 2001). Tripeptidylpeptidase II (TPPII) was suggested to supply peptides because of its ability to cleave peptides *in vitro* and its upregulation in cells surviving partial proteasome inhibition (Geier *et al.*, 1999). Leucine aminopeptidase was found to generate antigenic peptides from N-terminally extended precursors (Beninga *et al.*, 1998). Puromycin sensitive animopeptidase and bleomycin hydrolase were shown to trim N termini of synthetic peptides (Stoltze *et al.*, 2000). An enzyme located in the lumen in ER and called ERAAP (ER aminopeptidase associated with antigen processing) (Serwold *et al.*, 2002) or ERAP1 (ER aminopeptidase 1) (Saric *et al.*, 2002; York *et al.*, 2002), has been shown to be responsible for the final trimming of the N termini of peptides presented by MHC class I molecules. Recently, it was shown that within the proteasome,

peptides could be formed from noncontiguous parts of the source protein (Hanada *et al.*, 2004; Vigneron *et al.*, 2004). The mechanism of this splicing is not fully understood.

Currently there is insufficient quantitative data about the role of the above mentioned events to allow a precise bioinformatic evaluation of their impact on the antigen processing pathway. Overall, it is clear that, ultimately, many more pathways, involving many more stages, will need to be incorporated into predictive methods if we wish to model the overall process accurately. Given current data, however, EpiJen represents the most accurate and parsimonious approach to antigen prediction.

In conclusion, we have shown that our additive method is of undoubted utility for T-cell epitope prediction and can be used successfully for the design of novel high binding peptides. Indeed, QSAR is a technique able to optimize molecular structure in order to deliver enhanced, reduced, or otherwise modulated, biological properties of any variety that can be measured or classified. We could, for example, use it to optimize the MHC binding affinity of weakly affine peptides, such as putative cancer vaccines. Further, it is equally appropriate for the analysis and manipulation of peptide-MHC complex interaction with T cell receptors as it is peptide affinity for MHC. It is thus a tool of general utility to the immunologist, be they looking to design or enhance epitopes, non-immunogenic competitor peptides, or T cell antagonists.

ACKNOWLEDGEMENTS

The authors thank their colleagues Pingping Guan, Channa Hattotuwagama, Valerie Walshe, Martin Blythe, Debra Clayton (nee Taylor), Shelly Hemsley and Seph Borrow for the fruitful collaborations. The studies were supported by GlaxoSmithKline, Medical Research Council U.K., Biotechnology and Biological Sciences Research Council, U.K., the Department of Health, U.K., the Royal Society, U.K., the Medical University of Sofia, Bulgaria, and the Ministry of Education and Science, Bulgaria.

ABBREVIATIONS

AAM = Amino acids model

AAIM = Amino acids and interactions model

ABC = ATP-binding cassette

AI = Artificial intelligence

ANN = Artificial neuronal networks

APC = Antigen presenting cell

A_{ROC} = Area under curve *sensitivity*/*1-specificity*

CV = Cross validation

ER = Endoplasmic reticulum

HLA = Human leukocyte antigen

LFER = Linear free energy relationship

LOO-CV = Leave one out cross validation

MHC = Major histocompatibility complex

MD = Molecular dynamics

MLR = Multiple linear regression

PC = Principal components

PLS = Partial least squares

QSAR = Quantitative structure - activity relationship

ROC = Receiver operating characteristic

SVM = Support vector machines

TAP = Transporter associated with antigen processing

TB = Tuberculosis

TCR = T-cell receptor

REFERENCES

Altfeld, M. A., Livingston, B., Reshamwala, N., Nguyen, P. T., Addo, M. M., Shea, A., Newman, M., Fikes, J., *et al.* (2001). Identification of novel HLA-A2-restricted human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte epitopes predicted by the HLA-A2 supertype peptide-binding motif. *J. Virol.* **75**: 1301-11.

Altuvia, Y., Schueler, O. and Margalit, H. (1995). Ranking potential binding peptides to MHC molecules by a computational threading approach. *J. Mol. Biol.* **249**: 244-50.

Altuvia, Y. and Margalit, H. (2000). Sequence signals for generation of antigenic peptides by the proteasome: Implications for proteasomal cleavage mechanism. *J. Mol. Biol.* **295**: 879-90.

Andrews, P. R., Craik, D. J. and Martin, J. L. (1984). Functional group contributions to drug-receptor interactions. *J. Med. Chem.* **27**: 1648-57.

Babine, R. E. and Bender, S. L. (1997). Molecular recognition of protein-ligand complexes: applications to drug design. *Chem. Rev.* **97**, 1354-472.

Baker, K., Bleczinski, C., Lin, H., Salazar-Jimenez, G., Sengupta, D., Krane, S. and Cornish, V.W. (2002). Chemical complementation: a reaction-independent genetic assay for enzyme catalysis. *Proc. Natl. Acad. Sci. USA* **99**: 16537-42.

Beninga, J., Rock, K. L. and Goldberg, A. L. (1998). Interferon-gamma can stimulate post-proteasomal trimming of the N terminus of an antigenic peptide by inducing leucine aminopeptidase. *J. Biol. Chem.* **273**: 18734-42.

Bhasin, M. and Raghava, G. P. (2004a). Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* **22**: 3195-204.

Bhasin, M. and Raghava, G. P. (2004b). Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.* **13**: 596-607.

Bhasin, M. and Raghava, G. P. (2004c). SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* **20**: 421-3.

Bindal, M. C., Singh, P. and Gupta, S. P. (1982). Structure - activity studies on hallucinogenic phenylalkylamines using Fujita-Ban approach. *Arzneimittelforschung* **32**: 719-21.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C. (1987). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* **329**: 506-12.

Bleicher, K. H., Bohm, H. J., Muller, K. and Alanine, A. I. (2003). Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2**: 369-78.

Blythe, M. J., Doytchinova, I. A. and Flower, D. R. (2002). JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* **18**: 434-9.

Böhm, H. J. and Klebe, G. (1996). What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs? *Angew. Chem. Int. Ed. Engl.* **35**: 2588-614.

Bordner, A. J. and Abagyan, R. (2006). Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins* **63**: 512-26.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30**: 1145-59.

Brusic, V., Rudy, G., Honeyman, G., Hammer, J. and Harrison, L. (1998). Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* **14**: 121-30.

Bui, H. H., Schiewe, A. J., von Grafenstein, H. and Haworth, I. S. (2006). Structural prediction of peptides binding to MHC class I molecules. *Proteins* **63**: 43-52.

Burden, F. R. and Winkler, D. A. (2005). Predictive Bayesian neural network models of MHC class II peptide binding. *J. Mol. Graph. Model*. **23**: 481-9.

Calderone, C. T. and Williams, D. H. (2001). An enthalpic component in cooperativity: the relationship between enthalpy, entropy, and noncovalent structure in weak associations. *J. Am. Chem. Soc*. **123**: 6262-7.

Cammarata, A. and Yau, S. (1970). Predictability of correlations between in vitro tetracycline potencies and substituent indices. *J. Med. Chem*. **13**: 93-7.

Carson, R. T., Vignali, K. M., Woodland, D. L. and Vignali, D. A. (1997). T-cell receptor recognition of MHC class II-bound peptide flanking residues enhances immunogenicity and results in altered TCR V region usage. *Immunity* **7**: 387-99.

Cascio, P., Hilton, C., Kisselev, A. F., Rock, K. L. and Goldberg, A. L. (2001). 26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide. *EMBO J*. **20**: 2357-66.

Celis, J. E., Wolf, H. and Ostergaard, M. (2000). Bladder squamous cell carcinoma biomarkers derived from proteomics. *Electrophoresis* **21**: 2115-21.

Chen, L. and Jondal, M. (2004). Alternative processing for MHC class I presentation by immature and CpG-activated dendritic cells. *Eur. J. Immunol*. **34**: 952-60.

Chicz, R. M., Urban, R. G., Lane, W. S., Gorga, J. C., Stern, L. J., Vignali, D. A. A. and Strominger, J. L. (1992). Predominant naturally processed peptides bound to HLA DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* **358**: 764-8.

Chicz, R. M., Urban, R. G., Gorga, J. C., Vignali, D. A. A., Lane, W. S. and Strominger, J. L. (1993). Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J. Exp. Med*. **178**: 27-47.

Coyle, A. J. and Gutierrez-Ramos, J. C. (2001). The expanding B7 superfamily: increasing complexity in costimulatory signals regulating T cell function. *Nature* **363**: 203-9.

Craiu, A., Akopian, T., Goldberg, A. and Rock, K. L. (1997). Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proc. Natl. Acad. Sci. USA* **94**: 10850-5.

Cui, J., Han, L. Y., Lin, H. H., Zhang, H. L., Tang, Z. Q., Zheng, C. J., Cao, Z. W. and Chen, Y. Z. (2007). Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol. Immunol*. **44**: 866-77.

Dalpiaz, A., Gessi, S., Varani, K. and Borea, P. A. (1997). De novo analysis of receptor binding affinity data of 8-ethenyl-xantine antagonists to adenosine A1 and A2A receptors. *Arzneimittelforschung* **47**: 591-4.

Daniel, S., Brusic, V., Caillat-Zucman, S., Petrovsky, N., Harrison, L., Riganelli, D., Sinigaglia, F., Gallazzi, F., *et al.* (1998). Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. *J. Immunol.* **161**: 617-24.

Davies, M. N., Sansom, C. E., Beazley, C. and Moss, D. S. (2003). A novel predictive technique for the MHC class II peptide-binding interaction. *Mol. Med*. **9**: 220-5.

Davies, M. N., Hattotuwagama, C. K., Moss, D. S., Drew, M. G. and Flower, D. R. (2006). Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity. *BMC Struct. Biol*. **6**: 5.

de la Salle, H., Houssaint, E., Peyrat, M. A., Arnold, D., Salamero, J., Pinczon, D., Stevanovic, S., Bausinger, H., *et al.* (1997). Human peptide transporter deficiency: importance of HLA-B in the presentation of TAP-independent EBV antigens. *J. Immunol.* **158**: 4555-63.

Dessen, A., Lawrence, C. M., Cupo, S., Zaller, D. M. and Wiley, D.C. (1997). X-ray crystal structure of HLA-DR4 (DRA*0101, DRB1*0401) complexed with a peptide from human collagen II. *Immunity* **7**: 473-81.

Ding, S. J., Li, Y., Tan, Y. X., Jiang, M. R., Tian, B., Liu, Y. K., Shao, X. X., Ye, S. L., *et al.* (2004). From proteomic analysis to clinical significance: overexpression of cytokeratin 19 correlates with hepatocellular carcinoma metastasis. *Mol. Cell Proteomics* **3**: 73-81.

Djaballah, H., Harness, J. A., Savory, P. J. and Rivett, A. J. (1992). Use of serine-protease inhibitors as probes for the different proteolytic activities of the rat liver multicatalitic proteinase complex. *Eur. J. Biochem*. **209**: 629-34.

Dönnes, P. and Elofsson, A. (2002). Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* **3**: 25.

Dönnes, P. and Kohlbacher, O. (2005). Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci.* **14**: 2132-40.

Doytchinova, I. A., Blythe, M. J. and Flower, D. R. (2002). Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. *J. Proteome Res*. **1**: 263-72.

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

Doytchinova, I. and Flower, D. (2003b). The HLA-A2-supermotif: a QSAR definition. *Org. Biomol. Chem*. **1**: 2648-54.

Doytchinova, I., Hemsley, S. and Flower, D. R. (2004a). Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation. *J. Immunol*. **173**: 6813-9.

Doytchinova, I. A., Walshe, V., Jones, N., Gloster, S., Borrow, P. and Flower, D. R. (2004b). Coupling in silico and in vitro analysis of peptide-MHC binding: A bioinformatics approach enabling prediction of superbinding peptides and anchorless epitopes. *J. Immunol*. **172**: 7495-502.

Doytchinova, I. A., Guan, P. and Flower, D. R. (2006). EpiJen: a server for multistep T-cell epitope prediction. *BMC Bioinformatics* **7**: 131.

Doytchinova, I. A. and Flower, D. R. (2006). Class I T cell epitope prediction: improvements using a combination of Proteasome cleavage, TAP affinity, and MHC binding. *Mol. Immunol*. **43**: 2037-44.

Drummelsmith, J., Brochu, V., Girard, I., Messier, N. and Ouellette, M. (2003). Proteome mapping of the protozoan parasite Leishmania and application to the study of drug targets and resistance mechanisms. *Mol. Cell Proteomics* **2**: 146-55.

Ellis, J. M., Henson, V., Slack, R., Ng, J., Hartzman, R. J. and Hurley, C. K. (2000). The frequencies of HLA-A2 alleles in five U.S. population groups: Predominance of A*02011 and identification of HLA-A*0231. *Human Immunology* **61**: 334-40.

Fagerberg, T., Cerottini, J. C. and Michielin, O. (2006). Structural prediction of peptides bound to MHC class I. *J. Mol. Biol*. **356**: 521-46.

Falk, K., Rotzschke, O., Stevanovic, S., Jung, G. and Rammensee, H. G. (1991). Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* **351**: 290-6.

Falk, K. and Rotzschke, O. (1993). Consensus motifs and peptide ligands of MHC class I molecules. *Semin. Immunol*. **5**: 81-94.

Flower, D. R., McSparron, H., Blythe, M. J., Zygouri, C., Taylor, D., Guan, P., Wan, S., Coveney, P. V., *et al.* (2003). Computational vaccinology: quantitative approaches. *Novartis Found Symp*. **254**: 102-20.

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

Franco, A., Southwood, S., Arrhenius, T., Kuchroo, V. K., Grey, H. M., Sette, A. and Ishioka, G. Y. (1994). T-cell receptor antagonist peptides are highly effective inhibitors of experimental allergic encephalomyelitis. *Eur. J. Immunol.* **24**: 940-6.

Free, S. M. Jr. and Wilson, J. W. (1964). A mathematical contribution to structure – activity studies. *J. Med. Chem*. **7**: 395-9.

Fujita, T. and Ban, T. (1971). Structure – activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. *J. Med. Chem*. **14**: 148-52.

Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E. and Wiley, D. C. (1996). Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* **384**: 134-41.

Garrett, T. P., Saper, M. A., Bjorkman, P. J., Strominger, J. L. and Wiley, D. C. (1989). Specificity pockets for the side chains of peptide antigens in HLA-Aw68. *Nature* **342**: 692-6.

Geier, E., Pfeifer, G., Wilm, M., Lucchiari-Hartz, M., Baumeister, W., Eichmann, K. and Niedermann, G. (1999). A giant protease with potential to substitute for some functions of the proteasome. *Science* **283**: 978-81.

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

Godkin, A. J., Smith, K. J., Willis, A., Tejada-Simon, M. V., Zhang, J., Elliott, T. and Hill, A. V. (2001). Naturally processed HLA class II peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-MHC interactions. *J. Immunol*. **166**: 6720-7.

Golgberg, A. L., Cascio, P., Saric, T. and Rock, K. L. (2002). The importance of the proteasome and subsequent proteolitic steps in the generation of antigenic peptides. *Mol. Immunol*. **39**: 147-64.

Gombar, V. (1986). Quantitative Structure - Activity Relationships. Fujita-Ban analysis of beta-adrenergic blocking activity of 1-phenoxy-3-

([(substitutedamido)alkyl] amino)-2-propanols. *Arzneimittelforschung* **36**: 1014-8.

Greenbaum, D. C., Baruch, A., Grainger, M., Bozdech, Z., Medzihradszky, K. F., Engel, J., DeRisi, J., Holder, A. A., *et al*. (2002). A role for the protease falcipain 1 in host cell invasion by the human malaria parasite. *Science* **298**: 2002-6.

Guan, P., Doytchinova, I. A. and Flower, D.R. (2003a). HLA-A3 supermotif defined by quantitative structure-activity relationship analysis. *Protein Eng*. **16**: 11-8.

Guan, P., Doytchinova, I. A., Zygouri, C. and Flower, D. R. (2003b). MHCPred: bringing a quantitative dimension to the online prediction of MHC binding. *Appl. Bioinformatics* **2**: 63-6.

Guan, P., Doytchinova, I., Hattotuwagama, C. and Flower, D. R. (2006). MHCPred 2.0, an updated quantitative T cell epitope prediction server. *Appl. Bioinformatics* **5**: 55-61.

Gubler, B., Daniel, S., Armandola, E. A., Hammer, J., Caillat-Zucman, S. and van Endert, P. M. (1998). Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP. *Mol. Immunol.* **35**: 427-33**.**

Guéguen, M., Biddison, W. and Long, E. O. (1994). T cell recognition of an HLA-A2-restricted epitope derived from a cleaved signal sequence. *J. Exp. Med*. **180**: 1989-94.

Guo, H. C., Madden, D. R., Silver, M. L., Jardetzky, T. S., Gorga, J. C., Strominger, J. L. and Wiley, D. C. (1993). Comparison of the P2 specificity pocket in three human histocompatibility antigens: HLA-A*6801, HLA-A*0201, and HLA-B*2705. *Proc. Natl. Acad. Sci. USA* **90**: 8053-7.

Hajduk, P. J., Sheppard, G., Nettesheim, D. G., Olejniczak, E. T., Shuker, S. B., Meadows, R. P., Steinman, D. H., Carrera, G. M., *et al*. (1997). Discovery of potent nonpeptide inhibitors of Stromelysin using SAR by NMR. *J. Am. Chem. Soc*. **119**: 5818-27.

Hammett, L. P. (1937). The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc*.**59**: 96-106.

Hanada K, Yewdell J. W. and Yang J.C. (2004). Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature* **427**: 252-6.

Hansch, C., Maloney, P. P., Fujita, T. and Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* **194**: 178-80.

Henderson, R. A., Michel, H., Sakaguchi, K., Shabanowitz, J., Appella, E., Hunt, D. F. and Engelhard, V.H. (1992). HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. *Science* **255**: 1264-6.

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

Hunt, D. F., Michel, H., Dickinson, T. A., Shabanowitz, J., Cox, A. L., Sakaguchi, K. and Appella, E. (1992a). Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science* **256**: 1817-20.

Imanishi, S. and Harada, K. (2004). Proteomics approach on microcystin binding proteins in mouse liver for investigation of microcystin toxicity. *Toxicon*. **43**: 651-9.

Janeway, C. A. (2001). Immunobiology: the immune system in health and disease. Churchill Livingstone.

Jojic, N., Reyes-Gomez, M., Heckerman, D., Kadie, C. and Schueler-Furman, O. (2006). Learning MHC I--peptide binding. *Bioinformatics* **22**: e227-35.

Jones, E. Y. (1997). MHC class I and class II structures. *Curr. Opin. Immunol*. **9**: 75-9.

Kageyama, S., Isono, T., Iwaki, H., Wakabayashi, Y., Okada, Y., Kontani, K., Yoshimura, K., Terai, A., *et al.* (2004). Identification by proteomic analysis of calreticulin as a marker for bladder cancer and evaluation of the diagnostic accuracy of its detection in urine. *Clin. Chem.* **50**: 857-66.

Karpenko, O., Shi, J. and Dai, Y. (2005). Prediction of MHC class II binders using the ant colony search strategy. *Artif. Intell. Med*. **35**: 147-56.

Kawashima, I., Tsai, V., Southwood, S., Takesako, K., Sette, A. and Celis, E. (1999). Identification of HLA-A3-restricted cytotoxic T lymphocyte epitopes from carcinoembryonic antigen and HER-2/neu by primary in vitro immunization with peptide-pulsed dendritic cells. *Cancer Res*. **59**: 431-5.

Keightley, J. A., Shang, L. and Kinter, M. (2004). Proteomic analysis of oxidative stress-resistant cells: a specific role for aldose reductase overexpression in cytoprotection. *Mol. Cell Proteomics* **3**: 167-75.

Khanna, R., Burrows, S. R., Moss, D. J. and Silins, S. L. (1996). Peptide transporter (TAP-1 and TAP-2)-independent endogenous processing of Epstein-Barr virus (EBV) latent membrane protein 2A: implications for cytotoxic T-lymphocyte control of EBV-associated malignancies. *J. Virol*. **70**: 5357-62.

Krensky, A. M. and Clayberger, C. (1996). Structure of HLA molecules and immunosuppressive effects of HLA derived peptides. *Int. Rev. Immunol*. **13**: 173-85.

Kridel, S. J., Axelrod, F., Rozenkrantz, N. and Smith, J. W. (2004). Orlistat is a novel inhibitor of fatty acid synthase with antitumor activity. *Cancer Res.* **64**: 2070-5.

Kuttler, C., Nussbaum, A. K., Dick, T. P., Rammensee, H.-G., Schild, H. and Hadeler, K.-P. (2000). An algorithm for the prediction of proteasome cleavages. *J. Mol. Biol*. **298**: 417-29.

Lankat-Buttgereit, B. and Tampé, R. (1999). The transporter associated with antigen processing TAP: structure and function. *FEBS Lett*. **464**: 108-12.

Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Brunak, S., Lund, O. and Nielsen, M. (2005). An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* **35**: 2295-303.

Lautscham, G., Rickinson, A. and Blake, N. (2003). TAP-independent antigen presentation on MHC class I molecules: lessons from Epstein-Barr virus. *Microbes Infect.* **5**: 291-9.

Liu, W., Meng, X., Xu, Q., Flower, D. R. and Li, T. (2006). Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics* **7**: 182.

Lopes, A. R., Jaye, A., Dorrell, L., Sabally, S., Alabi, A., Jones, N. A., Flower, D. R., De Groot, A., *et al.* (2003). Greater CD8+ TCR heterogeneity and functional flexibility in HIV-2 compared to HIV-1 infection. *J. Immunol.* **171**: 307-16.

Luckey, C. J., King, G. M., Marto, J. A., Venketeswaran, S., Maier, B. F., Crotzer, V. L., Colella, T. A., Shabanowitz, J., *et al*. (1998). Proteasomes can either generate or destroy MHC class I epitopes: evidence for nonproteosomal epitope generation in the cytosol. *J. Immunol.* **161**: 112-21.

Luckey, C. J., Marto, J. A., Partridge, M., Hall, E., White, F. M., Lippolis, J. D., Shabanowitz, J., Hunt, D. F., *et al.* (2001). Differences in the expression of human class I MHC alleles and their associated peptides in the presence of proteasome inhibitors. *J. Immunol.* **167**: 1212-21.

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

Madden, D. R., Garboczi, D. N. and Wiley, D. C. (1993). The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell* **75**: 693-708.

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

Mallios, R. R. (2001). Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics* **17**: 942-8.

Maly, D. C., Choong, I. C. and Ellman, J. A. (2000). Combinatorial target-guided ligand assembly: Identification of potent subtype-selective c-Src inhibitors. *Proc. Natl. Acad. Sci. USA* **97**: 9367-72.

Matsumura, M., Fremont, D. H., Peterson, P. A. and Wilson, I. A. (1992). Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* **257**: 927-34.

McSparron, H., Blythe, M. J., Zygouri, C., Doytchinova, I. A. and Flower, D. R. (2003). JenPep: A novel computational information resource for immunobiology and vaccinology. *J. Chem. Inf. Comput. Sci.* **43**: 1276-87.

Meyer, T. H., van Endert, P. M., Uebel, S., Ehring, B. and Tampé, R. (1994). Functional expression and purification of the ABC transporter complex associated with antigen processing (TAP) in insect cells. *FEBS Lett.* **351**: 443-7.

Mo, X. Y., Cascio, P., Lemerise, K., Goldberg, A. L. and Rock, K. (1999). Distinct proteolytic processes generate the C and N termini of MHC class I-binding peptides. *J. Immunol*. **163**: 5851-9.

Monaco, J., Cho, S. and Attaya, M. (1990). Transport protein genes in the murine MHC – possible implications for antigen processing. *Science* **250**: 1723-6.

Mormung, F., Neefjes, J. J. and Hämmerling, G. J. (1994). Peptide selection by MHC-encoded TAP transporters. *Curr. Opin. Immunol.* **6**: 32-7.

Müller, K. M., Ebensperger, C. and Tampé, R. (1994). Nucleotide binding to the hydrophilic C-terminal domain of the transporter associated with antigen processing (TAP). *J. Biol. Chem*. **269**: 14032-7.

Murray, C. W. and Verdonk, M. L. (2002). The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comp. Aided Mol. Dis*. **16**: 741-53.

Murugan, N. and Dai, Y. (2005). Prediction of MHC class II binding peptides based on an iterative learning model. *Immunome Res*. **1**: 6.

Niedermann, G., King, G., Butz, S., Birsner, U., Grimm, R., Shabanowitz, J., Hunt, D. F. and Eichmann, K. (1996). The proteolytic fragments generated by vertebrate proteasomes: structural relationships to major histocompatibility complex class I binding peptides. *Proc. Natl. Acad. Sci. USA* **93**: 8572-7.

Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S. and Lund, O. (2004). Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* **20**: 1388-97.

Nisato, D., Wadnon, J., Callet, G., Mettefeu, D., Assens, J. L., Plouzane, C., Tonnerre, B., Pliska, V., *et al*. (1987). Renin inhibitors. Free-Wilson and correlation analysis of the inhibitory potency of a series of pepstatin analogues on plasma renin. *J. Med. Chem*. **30**: 2287-91.

Noguchi, H., Hanai, T., Honda, H., Harrison, L. C. and Kobayashi, T. (2001). Fuzzy neural network-based prediction of the motif for MHC class II binding peptides. *J. Biosci. Bioeng*. **92**: 227-31.

Noguchi, H., Kato, R., Hanai, T., Matsubara, Y., Honda, H., Brusic, V. and Kobayashi, T. (2002). Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J. Biosci. Bioeng*. **94**: 264-70.

Orlowski, M. and Michaud, C. (1989). Pituitary multicatalytic proteinase complex. Specificity of components and aspects of proteolitic activity. *Biochemistry* **28**: 9270-8.

Orlowski, M., Cardozo, C. and Michaud, C. (1993). Evidence for the presence of five distinct proteolytic components in the pituitary multicatalytic proteinase complex. Properties of two components cleaving bonds on the carboxyl side of branched chain and small neutral amino acids. *Biochemistry* **16**: 1563-72.

Parham, P., Lomen, C. E., Lawlor, D. A., Ways, J. P., Holmes, N., Coppin, H. L., Salter, R. D., Wan, A. M., *et al*. (1988). *Proc. Natl. Acad. Sci. USA* **85**: 4005-9.

Park, K. S., Kim, H., Kim, N. G., Cho, S. Y., Choi, K. H., Seong, J. K. and Paik, Y. K. (2002). Proteomic analysis and molecular characterization of tissue ferritin light chain in hepatocellular carcinoma. *Hepatology* **35**: 1459-66.

Parker, K. C., Bednarek, M. A. and Coligan, J. E. (1994). Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol*. **152**: 163-75.

Parker, K. C., Shields, M., DiBrino, M., Brooks, A. and Coligan, J. E. (1995). Peptide binding to MHC class I molecules: implications for antigenic peptide prediction. *Immunol. Res*. **14**: 34-57.

Peters, B., Bulik, S., Tampé, R., van Endert, P. M. and Holzhütter, H. G. (2003). Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol*. **171**: 1741-9.

Peters, B. and Sette, A. (2005). Generating quantitative models describing the sequence specificity of biological process with the stabilized matrix method. *BMC Bioinformatics* **6**: 132.

Petricoin, E. F., Ornstein, D. K. and Liotta, L. A. (2004). Clinical proteomics: Applications for prostate cancer biomarker discovery and detection. *Urol. Oncol*. **22**: 322-8.

Petrone, P. M. and Garcia, A.E. (2004). MHC-peptide binding is assisted by bound water molecules. *J. Mol. Biol*. **338**: 419-35.

Raffa, R. B. (2001). Introduction: on the relevance of thermodynamics to pharmacology. In *Drug – Receptor Thermodynamics: Introduction and Applications*, (Raffa, R. B., ed.) Wiley pp. 3-12.

Rammensee, H. G., Friede, T. and Stevanoviic, S. (1995). MHC ligands and peptide motifs: first listing. *Immunogenetics* **41**: 178-228.

Rao, J. and Whitesides, G. M. (1997). Tight binding of a dimeric derivative of vancomycin with dimeric L-Lys-D-Ala_D-Ala. *J. Am. Chem. Soc*. **119**: 10286-90.

Rao, J., Lahiri, J., Isaacs, L., Weis, R. M. and Whitesides, G. M. (1998). Trivalent system from Vancomycin: D-Ala-D-Ala with higher affinity than Biotin: Avidin. *Science* **280**: 708-11.

Riedesel, H., Kolbeck, B., Schmetzer, O. and Knapp, E. W. (2004). Peptide binding at class I major histocompatibility complex scored with linear functions and support vector machines. *Genome Inform*. **15**: 198-212.

Rudensky, A. Y., Preston-Hulbert, P., Soon-Cheol, H., Barlow, A. and Janeway Jr., C. A. (1991). Sequence analysis of peptides bound to MHC class II molecules. *Nature (London)* **353**: 622-7.

Ruppert, J., Sidney, J., Celis, E., Kubo, R. T., Grey, H. M. and Sette, A. (1993). Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. *Cell* **74**: 929-37.

Saper, M. A., Bjorkman, P. J. and Wiley, D.C. (1991). Refined structure of the human class histocompatibility antigen HLA-A2 at 2.6 Å. *J. Mol. Biol*. **219**: 277-319.

Saric, T., Chang, S.-C., Hattori, A., York, I. A., Markant, S., Rock, K. L., Tsujimoto, M. and Goldberg, A. L. (2002). An IFN-γ-induced aminopeptidase in the ER, ERAPI, trims precursors to MHC class I-presented peptides. *Nat. Immunol*. **3**: 1169-76.

Saxova, P., Buus, S., Brunak, S. and Keşmir, C. (2003). Predicting proteosomal cleavage sites: a comparison of available methods. *Int. Immunol*. **15**: 781-7.

Schaschke, N., Matschiner, G., Zettll, F., Marquardt, U., Bergner, A., Bode, W., Sommerhoff, C. P. and Moroder, L. (2001). Bivalent inhibition of human beta-tryptase. *Chem. Biol*. **8**: 313-27.

Schirle, M., Weinschenk, T. and 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.

Schölkopf, B. and Smola, A. J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, Mass. MIT Press.

Schönbach, C., Koh, J. L., Sheng, X., Wong, L. and Brusic, V. (2000). FIMM, a database of functional molecular immunology. *Nucleic Acids Res*. **28**: 222-4.

Schueler-Furman, O., Altuvia, Y., Sette, A. and Margalit, H. (2000). Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci*. **9**: 1838-46.

Schumacher, T. N. M., Kantesaria, D. V., Heemels, M. T., Ashton-Rickardt, P. G., Shepherd, J. C., Früh, K., Yang, Y., Peterson, P. A., *et al.* (1994). Peptide length and sequence specificity of the mouse TAP1/TAP2 translocator. *J. Exp. Med*. **179**: 533-40.

Serwold, T. and Shastri, N. (1999). Specific proteolytic cleavages limit the diversity of the pool of peptides available to MHC class I molecules in living cells. *J. Immunol*. **162**: 4712-19.

Serwold, T., Gonzalez, F., Kim, J., Jacob, R. and Shastri, N. (2002). ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature* **419**: 480-3.

Sette, A. and Sidney, J. (1998). HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. *Curr. Opin. Immunol*. **10**: 478-82.

Sette, A., Livingstone, B., McKinney, D., Appella, E., Fikes, J., Sidney, J., Newman, M. and Chesnut, R. (2001). The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation. *Biologicals* **29**: 271-6.

Shuker, S. B., Hajduk, P. J., Meadows, R. P. and Fesik, S. W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**: 1531-4.

Sidney, J., Grey, H. M., Kubo, R. T. and Sette, A. (1996a). Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. *Immunol. Today* **17**: 261-6.

Sidney, J., Grey, H. M., Southwood, S., Celis, E., Wentworth, P. A., del Guercio, M. F., Kubo, R. T., Chesnut, R. W., *et al*. (1996b). Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide binding repertoires of common HLA molecules. *Hum. Immunol*. **45**: 79-93.

Sidney, J., Southwood, S., Mann, D. L., Fernandez-Vina, M. A., Newman, M. J. and Sette, A. (2001). Majority of peptides binding HLA-A*0201 with high affinity crossreact with other A2-supertype molecules. *Hum. Immunol*. **62**: 1200-16.

Smith, K. D. and Lutz, C. T. (1996). Peptide-dependent expression of HLA-B7 on antigen processing-deficient T2 cells. *J Immunol*. **156**: 3755-64.

Stoltze, L., Schirle, M., Schwarz, G., Schroeter, C., Thompson, M. W., Hersh, L. B., Kalbacher, H., Stevanović, S., *et al.* (2000). Two new proteases in the MHC class I processing pathway. *Nat. Immunol.* **1**: 413-8.

Sudo, T., Kamikawaji, N., Kimura, A., Date, Y., Savoie, C. J., Nakashima, H., Furuichi, E., Kuhara, S., *et al*. (1995). Differences in MHC class I self peptide repertoires among HLA-A2 subtypes. *J. Immunol*. **155**: 4749-56.

Tanaka, K. and Kasahara, M. (1998). The MHC class I ligand-generating system: roles of immunoproteasomes and the interferon-g-inducible proteasome activator PA28. *Immunol. Rev*. **163**: 161-76.

Terada, Y. and Nanya, K. (2000). Free-Wilson analysis of the antibacterial activity of fluoronaphthyridines against various microbes. A new application of indicator variables. *Pharmazie* **55**: 133-5.

Tmej, C., Chiba, P., Huber, M., Richter, E., Hitzler, M., Schaper, K. J. and Ecker, G. (1998). A combined Hansch/Free-Wilson approach as predictive tool in QSAR studies on propafenone-type modulators of multidrug resistance. *Arch. Pharm. (Weinheim)* **331**: 233-40.

Todeschini, R. and Consonni, V. (2000). *Handbook of molecular descriptors. Methods and Principles in Medicinal Chemistry* (Mannhold, R., Kubinyi, H., Timmerman, H., eds.) vol.11 Wiley - VCH, Weinheim, 2000.

Toes, R. E., Nussbaum, A. K., Degermann, S., Schirle, M., Emmerich, N. P., Kraft, M., Laplace, C., Zwinderman, A., *et al.* (2001). Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J. Exp. Med*. **194**: 1-12.

Tomic, S., Nilsson, L. and Wade, R. C. (2000). Nuclear receptor-DNA binding specificity: A COMBINE and Free-Wilson QSAR analysis. *J. Med. Chem*. **43**: 1780-92.

Tong, J. C., Tan, T. W. and Ranganathan, S. (2004). Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci*. **13**: 2523-32.

Tong, J. C., Bramson, J., Kanduc, D., Chow, S., Sinha, A. A. and Ranganathan, S. (2006a). Modeling the bound conformation of Pemphigus Vulgaris-associated peptides to MHC Class II DR and DQ Alleles. *Immunome Res*. **2**: 1.

Tong, J. C., Zhang, G. L., Tan, T. W., August, J. T., Brusic, V. and Ranganathan, S. (2006b). Prediction of HLA-DQ3.2beta ligands: evidence of multiple registers in class II binding peptides. *Bioinformatics* **22**: 1232-8.

Toseland, C. P., Taylor, D. J., McSparron, H., Hemsley, S. L., Blythe, M. J., Paine, K., Doytchinova, I. A., Guan, P., *et al.* (2005). AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res*. **1**: 4.

van den Eynde, B. J. and Morel, S. (2001). Differential processing of class-I-restricted epitopes by the standard proteasome and the immunoproteasome. *Curr. Opin. Immunol*. **13**: 147-53.

Vasmatzis, G., Zhang, C., Cornette, J. L. and DeLisi, C. (1996). Computational determination of side chain specificity for pockets in class I MHC molecules. *Mol. Immunol*. **33**: 1231-9.

Vigneron, N., Stroobant, V., Chapiro, J., Ooms, A., Degiovanni, G., Morel, S., van der Bruggen, P., Boon, T., *et al*. (2004). An antigenic peptide produced by peptide splicing in the proteasome. *Science* **304**: 587-90.

Vinitsky, A., Anton, L. C., Snyder, H. L., Orlowski, M., Bennink, J. R. and Yewdell, J. W. (1997). The generation of MHC class I-associated peptides is only partially inhibited by proteasome inhibitors: involvement of nonproteasomal cytosolic proteases in antigen processing. *J. Immunol.* **159**: 554-64.

von Herrath, M. G., Coon, B., Lewicki, H., Mazarguil, H., Gairin, J. E. and Oldstone, M. B. (1998). *In vivo* treatment with a MHC class I-restricted blocking peptide can prevent virus-induced autoimmune diabetes. *J. Immunol*. **161**: 5087-96.

Wan, S., Coveney, P. V. and Flower, D. R. (2005a). Molecular basis of peptide recognition by the TCR: affinity differences calculated using large scale computing. *J. Immunol*. **175**: 1715-23.

Wan, S., Coveney, P. V. and Flower, D. R. (2005b). Peptide recognition by the T-cell receptor: comparison of binding free energies from thermodynamic integration, Poisson-Boltzmann and linear interaction energy approximations. *Philos Transact. A Math. Phys. Eng. Sci*. **363**: 2037-53.

Wan, S., Coveney, P. and Flower, D.R. (2004). Large-scale molecular dynamics simulations of HLA-A*0201 complexed with a tumor-specific antigenic peptide: can the alpha3 and beta2m domains be neglected? *J. Comput. Chem*. **25**: 1803-13.

Wold, S. (1995). PLS for Multivariate Linear Modeling. In *Chemometric Methods in Molecular Design* (van de Waterbeemd, H., ed.) VCH, Weinheim pp. 195-218.

Yang, Z. R. and Johnson, F. C. (2005). Prediction of T-cell epitopes using biosupport vector machines. *J. Chem. Inf. Model*. **45**: 1424-8.

York, I. A., Chang, S.-C., Saric, T., Keys, J. A., Favreau, J. M., Goldberg, A. L. and Rock, K. L. (2002). The ER aminopeptidase ERAPI enhances or limits antigen presentation by trimming epitopes to 8-9 residues. *Nat. Immunol.* **3**: 1177-84.

Zacharias, M. and Springer, S. (2004). Conformational flexibility of the MHC class I alpha1-alpha2 domain in peptide bound and free states: a molecular dynamics simulation study. *Biophys. J*. **87**: 2203-14.

Zhao, Y., Pinilla, C., Valmori, D., Martin, R. and Simon, R. (2003). Application of support vector machines for T-cell epitopes prediction. *Bioinformatics* **19**: 1978-84.

Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J. and Kovach, J. S. (2003). Detection of cancer-specific markers amid massive mass spectral data. *Proc. Natl. Acad. Sci. USA* **100**: 14666-71.

Received: **October 18, 2007** **Revised**: **February 8, 2008** **Accepted**: **February 12, 2008**