

Quantitative online prediction of peptide binding to the major histocompatibility complex

Channa K. Hattotuagama, Pingping Guan, Irimi A. Doytchinova, Christianna Zygouri, Darren R. Flower*

Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, UK

Received 11 November 2002; received in revised form 10 July 2003; accepted 14 July 2003

Abstract

With its implications for vaccine discovery, the accurate prediction of T cell epitopes is one of the key aspirations of computational vaccinology. We have developed a robust multivariate statistical method, based on partial least squares, for the quantitative prediction of peptide binding to major histocompatibility complexes (MHC), the principal checkpoint on the antigen presentation pathway. As a service to the immunobiology community, we have made a Perl implementation of the method available via a World Wide Web server. We call this server MHCpred. Access to the server is freely available from the URL: <http://www.jenner.ac.uk/MHCpred>. We have exemplified our method with a model for peptides binding to the common human MHC molecule HLA-B*3501.

© 2003 Elsevier Inc. All rights reserved.

Keywords: Peptide binding; Partial least squares; Quantitative structure activity relationships; T cell epitope; Major histocompatibility complex

1. Introduction

The advent of the post-genomic era has irrevocably changed the intellectual landscape of the biosciences: its implications suggest that we should be able to gain access to information about biological function at a rate, and on a scale, previously beyond our wildest expectations. The term genome is already passe, and in its wake has come many new definitions. The Biome, and hence biomics, is an all encompassing term covering all such definitions. An oft-neglected part of the Biome is the immunome: the set of antigenic peptides, or immunogenic proteins, within a microorganism, be that virus, bacteria, fungus, or parasite [1,2]. It is also possible to talk of the self-immunome, the set of potentially antigenic self-peptides. This is clearly important within the context of, for example, cancer (the cancer-immunome) and autoimmunity (the auto-immunome), which affect about 30 and 3% of the global population, respectively.

The nature of the immunome is clearly dependent upon the host as much as it is on what we shall, for convenience,

call the pathogen. This is implicit in the terms antigenic and immunogenic. A peptide is not antigenic if the immune system does not respond to it. A good example of this is the major histocompatibility complex (MHC) restriction of T cell responses. A particular MHC allele will have a peptide specificity that may, or may not, overlap, with other expressed alleles, but the total specificity of all alleles, expressed within an individual or population, will not cover the whole of sequence space. Thus, peptides with sequences that do not bind to any of an individual's allelic MHC variants cannot be antigenic within a cellular context. The ability to define the specificity of different MHCs computationally, which we may call *in silico* immunomics, a part of computational vaccinology or computer-aided vaccine design, is an important, but eminently realizable, goal of immunoinformatics, the application of informatics techniques to immunological macromolecules, a newly emergent sub-discipline within bioinformatics. At the heart of computational vaccinology is the challenge of epitope prediction.

Epitopes are chemical moieties recognized by the immune system. While the importance on non-peptide epitopes, such as lipids and carbohydrates, has become well recognized in recent years, peptide T cell and B cell epitopes (as mediated by the cellular and humoral immune systems, respectively) are still the main tools by which the immune system can be examined. The accurate delineation

Abbreviations: MHC, major histocompatibility complex; TCR, T cell receptor; MFI, maximum fluorescence intensity; ER, endoplasmic reticulum; GUI, graphical user interface

* Corresponding author. Tel.: +44-1635-577954; fax: +44-1635-577901/577908.

E-mail address: darren.flower@jenner.ac.uk (D.R. Flower).

of T cell and B cell epitopes, around which polyepitope vaccines are constructed, is the key challenge for informatics with immunobiology.

Techniques for the prediction of B cell epitopes often lack sophistication [3,4], many relying on an elusive knowledge of protein structure [5]. However, much more advanced methods for the prediction of T cell epitopes have been developed [6]. It is now generally accepted that only peptides that bind to MHC at an affinity above a threshold can function as T cell epitopes [23] and that, at least broadly, MHC-peptide affinity is well correlated with the extent of the resulting immune response. It is fair to say that most modern methods for T cell epitope prediction rely, at least conceptually, on predicting the affinity of peptide binding to MHCs.

We have recently developed a 2D-QSAR approach to the prediction of peptide-MHC binding [7], which we have christened the additive method. It is based on the well-known Free-Wilson concept, but utilizes additional terms to account for the effect of side chain correlation. In the present work we introduce a World Wide Web service called MHCpred, which implements this method. MHCpred is composed of a variety of widely distributed human allele-specific quantitative structure activity relationship (QSAR) models built using partial least squares (PLS), a robust multivariate statistical method. This paper deals solely with binding to class I MHCs. In what follows we will describe the server and exemplify our approach by presenting a new allele model of HLA-B*3501 (B3501), a common human class I allele. B3501 is a member of the B5 cross-reactive group (CREG), which includes the closely related alleles HLA-B53, HLA-B51, HLA-B52, HLA-B35, and HLA-B58. HLA-B35 is associated with subacute thyroiditis, where presentation of self-peptides or certain virus peptides by this allele may be involved in mediating pathogenesis, and with accelerated progression from HIV infection to AIDS.

2. Materials and methods

2.1. Server development

2.1.1. Server software

MHCpred is an implementation of the additive method and covers a range of different allele models [7]. MHCpred runs as a CGI server, written in Perl, running under Microsoft windows NT. MHCpred is freely available via the World Wide Web from the URL: <http://www.jenner.ac.uk/MHCpred>. The interface is straightforward and intuitive: the sequence of a protein antigen is entered, an MHC allele and affinity threshold are selected, and the program run. Additionally, an arbitrary motif can be entered to further restrain the search results. The results page produced, subsequently, displays a sorted list of nine amino acid substrings of the entered antigen sequence in order of calculated affinities.

2.1.2. Peptides and binding affinities

Models implemented within the MHCpred server were generated from IC₅₀ values, characterising the affinity of peptides for MHC molecules, collated from the literature and accumulated in the JenPep relational database system [8], our storehouse of quantitative affinity measures for peptide interactions within immunobiology. IC₅₀ values were obtained from radioligand competition assays. The K_D, or equilibrium dissociation constant, of the test peptide can be obtained from these IC₅₀ value using the classic relationship derived by Cheng and Prussoff [9].

$$K_D^i = \frac{IC_{50}}{1 + ([L_{tot}^S]/K_D^S)} \quad (1)$$

where K_Dⁱ is the dissociation constant for the inhibitor or test peptide, K_D^S the dissociation constant for the standard radiolabelled peptide [L_{tot}^S] the total concentration of the radiolabel. This relation holds at the midpoint of the inhibition curve under two principal constraints, the total amount of radiolabel is much greater than the concentration of bound radiolabel and that the concentration of bound test peptide is much less than the IC₅₀. For competition assays, it can be shown that IC₅₀ values may be defined by

$$K_D^i = \frac{[R_{free}](IC_{50} - [RL^i])}{[RL^i]} \quad (2)$$

where [RLⁱ] is the concentration of test peptide bound to MHC and R_{free} the concentration of free MHC. Both [R_{free}] and [RLⁱ] are independent of the test IC₅₀ value. It is clear from this that the dissociation constant varies with IC₅₀, at least within a single experiment. In practice, the variation in IC₅₀ is often sufficiently small that values can be compared between experiments.

2.1.3. Additive method

Extracted IC₅₀ values were first converted to log[1/IC₅₀] values (or -log₁₀[IC₅₀] or pIC₅₀) and used as the dependent variables in a QSAR regression. As eluded to earlier, pIC₅₀ can be related to changes in the free energy of binding ΔG_{bind} ∝ -RT ln IC₅₀. The pIC₅₀ values were predicted from a combination of the contributions (P) of individual amino acids at each position of the peptide and contributions from side chain-side chain interactions

$$pIC_{50} = \text{constant} + \sum_{i=1}^9 P_i + \sum_{j=1}^8 \sum_{i=1}^{9-j} P_i P_{i+j} \quad (3)$$

where the constant 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, and $\sum_{j=1}^8 \sum_{i=1}^{9-j} P_i P_{i+j}$ is a series of summations for pairwise interactions between side chain of increasing sequence separation. In order to simplify this equation, and make it more mathematically tractable, we made use of the observation that class I MHC bound peptides assume extended but twisted conformations,

so that adjacent side chains point in essentially opposite directions, both (1)–(2) and (1)–(3) interactions are possible between side chains. The resulting equation takes the form

$$pIC_{50} = \text{constant} + \sum_{i=1}^9 P_i + \sum_{i=1}^8 P_i P_{i+1} + \sum_{i=1}^7 P_i P_{i+2} \quad (4)$$

The need to handle data matrices with more variables than observations led us to utilize partial least squares, as the predictor of biological activity, as implemented within Sybyl6.7 [10]. Column-filtering and scaling were deactivated and the optimal number of components found via cross-validation using SAMPLS [11]. Summations stop at nine because we are examining only nine amino acid peptides, the commonest length of epitope observed experimentally. Were sufficient data available, we could equally well construct such analyses for other observed epitope lengths: principally 8, 10, or 11 amino acid peptides. Unfortunately, such data is not currently available.

Leave-one-out cross-validation was used to assess the predictive power of the models. These parameters are defined later and q^2 is given by the equation

$$q^2 = 1 - \frac{\text{PRESS}}{\text{SSQ}} \quad (5)$$

where

$$\text{PRESS} = \sum_{i=1}^n (Y_{\text{obs}} - Y_{\text{pred}})^2 \quad (6)$$

$$\text{SSQ} = \sum_{i=1}^n (Y_{\text{obs}} - \bar{Y})^2 \quad (7)$$

Standard error of prediction (SEP) is given by the formula

$$\text{SEP} = \frac{\sqrt{\text{PRESS}}}{n - 1} \quad (8)$$

where n is the number of observations. The non-cross-validated models were assessed using standard multiple linear regression (MLR) parameters: explained variance r^2 , standard error of estimate (SEE), and F ratio and r^2 is given by the formula

$$r^2 = \frac{\sum_{i=1}^n (Y_{\text{pred}} - \bar{Y})^2}{\sum_{i=1}^n (Y_{\text{obs}} - \bar{Y})^2} \quad \text{or} \quad r^2 = \frac{\sum_{i=1}^n (Y_{\text{obs}} - Y_{\text{pred}})^2}{\sum_{i=1}^n (Y_{\text{obs}} - \bar{Y})^2} \quad (9)$$

The standard error of estimate (SEE) is defined as

$$\text{SEE} = \frac{\sqrt{\sum_{i=1}^n (Y_{\text{obs}} - Y_{\text{pred}})^2}}{n - k - 1} \quad (10)$$

where k is the number of variables. The F ratio is given by the formula

$$F_{n,k} = \frac{\sum_{i=1}^n (Y_{\text{pred}} - \bar{Y})^2 / k}{\sum_{i=1}^n (Y_{\text{obs}} - Y_{\text{pred}})^2 / n - k - 1} \quad (11)$$

Table 1

MHCPred server: statistics for the different models currently implemented

	n	q^2	NC	SEP	r^2
Model statistics for additive method					
A*0101	95	0.420	4	0.907	0.997
A*0201	335	0.377	6	0.694	0.731
A*0202	69	0.317	9	0.606	0.943
A*0203	62	0.327	6	0.841	0.963
A*0206	57	0.475	6	0.576	0.989
A*0301	70	0.305	4	0.699	0.972
A*1101	62	0.428	3	0.593	0.977
A*3101	31	0.453	6	0.727	0.990
A*6801	37	0.370	4	0.664	0.974
A*6802	46	0.500	7	0.647	0.983

Multivariate statistics parameters for the different models currently implemented within the MHCPred server. The definition of the different parameters is given in Section 2.

where Y_{pred} is the predicted, Y_{obs} the observed, and \bar{Y} the average dependent variable, in this case IC_{50} s. SEP and SEE are standard errors for prediction, and assess the distribution of errors between observed and predicted values in regression models. The F value is a distance measure between distributions of the predicted values and the error in prediction. As the F value increases, so does confidence that the means and, thus the distributions, are different.

A set of different models was generated. These covered a range of human alleles: A*0101, A*0201, A*0202, A*0203, A*0206, A*0301, A*1101, A*3301, A*6801, A*6802, and B*3501. These exist in high frequency within human populations and significant binding data have been reported for each allele. To further exemplify the method, we also present, more fully, derivation of an additive model for the B*3501 allele (Table 1).

2.1.4. B*3501 model

A set of 52 nonameric peptides was used as a training set (Table 2). Although more than 52 B*3501 nonamer binders could be extracted from the JenPep database, this set was preferred as it was synthesised and analysed by the same experimental group. As before, IC_{50} values were calculated from a quantitative assay based on the inhibition of radiolabeled standard probe peptide FPFKYAAF binding to detergent solubilized MHC molecules. The selected set of peptides was analysed using the approach outlined as described earlier and in Doytchinova et al. [7].

3. Results

3.1. The B*3501 model

Previous, empirical, analysis has delineated an HLA-B*3501-specific motif which exhibits a strong preference for Proline at position 2 and a similar strong preference for hydrophobic and/or aromatic residues (Phe, Met, Leu, Ile, or Tyr) at position 9. Position 4 is mostly occupied

Table 2
List of peptides used in this study of HLA-B*3501

Number	Peptide sequence	Reference	pIC ₅₀ (−log ₁₀ [IC ₅₀])
1	MPLETQLAI	35	7.509
2	LPSDFFPSV	36	6.767
3	HPAAMPHELL	36	6.573
4	MPLETQLAI	37	7.553
5	YPAEITLTW	37	7.268
6	LPSDFFPSV	37	6.526
7	TPYDINQML	37	4.752
8	DPKVKQWPL	37	5.249
9	LPGPKFLQY	37	6.830
10	MPNQAQMRI	38	5.985
11	FAVRPQVPM	38	7.886
12	FPISPIETV	38	5.978
13	LPTNASLSF	38	6.939
14	FPPEGVSIW	38	5.373
15	IPFLTQFKL	38	5.046
16	IPSYKKLIM	38	7.013
17	IPISSWAI	38	7.268
18	FPHCLAFSI	38	7.538
19	WPLLPHVIF	38	6.745
20	LPFRNCRPF	38	6.398
21	SPATLLLVL	38	6.754
22	FPKAGLLII	38	5.885
23	MPFAGLLII	38	9.000
24	LPWHLRFL	38	7.377
25	LPVFTWLAL	38	7.456
26	FPASFFIKL	38	7.824
27	FPVRPQVPL	38	7.658
28	HPQKVTKFM	38	5.558
29	LPSIPVHPI	38	5.532
30	HPEDTGQVF	38	6.686
31	FPFVLAAIL	38	7.585
32	LPQPPICTI	38	4.921
33	LPGCSFSIF	38	6.947
34	YPCTVNFTI	38	6.015
35	VPISHLYIL	38	6.416
36	CPKDGQPSL	38	5.605
37	FPYLVAYQA	38	6.740
38	TPAEVSIVV	38	5.674
39	SPASFFSSW	38	5.373
40	MPREDAHFI	38	6.347
41	LPTTLFQPV	38	5.805
42	GPVTAQVVL	38	4.842
43	IPPSFLQAM	38	6.347
44	VPLSEDQLL	38	5.046
45	CPLERFAEL	38	6.987
46	LPDGQVIWV	38	4.921
47	SPSCPLERF	38	6.421
48	FPVRPQVPT	38	6.585
49	FPVRPQVPM	38	7.886
50	FGVRPQVPL	38	5.432
51	FAVRPQVPL	38	6.830
52	FPVRPQVPL	38	7.432

A list of the 52 peptides used in this study. The peptide sequence, the reference for the experimentally derived data, and the pIC₅₀ are shown [35–38].

by charged residues: Asp, Glu, or Lys. Comparison of the structures of different MHC class I molecules indicates a general mode of peptide binding. While the N and C termini of the peptides are positioned within the peptide binding groove by a network of conserved hydrogen bonds,

Table 3
Amino acid populations with HLA-B*3501 data set

Abundance of amino acids at each peptide position								
1	2	3	4	5	6	7	8	9
M 5	P 49	L 5	E 7	T 5	Q 13	L 10	A 5	I 12
L 12	A 2	S 5	D 5	F 5	F 8	P 4	S 6	V 6
H 3	G 1	A 6	A 3	M 1	P 1	H 3	L 3	L 18
Y 2		Y 2	V 2	I 2	T 2	Q 5	T 4	W 3
T 2		K 3	P 3	K 3	N 2	W 2	M 1	Y 1
D 1		G 2	Q 1	A 3	I 2	M 1	P 10	M 4
F 14		N 1	R 8	P 12	S 3	V 9	Q 2	F 6
I 4		V 9	S 6	G 4	V 3	E 2	R 3	A 1
W 1		I 3	N 1	S 2	K 2	S 3	I 8	T 1
S 3		T 2	L 3	L 4	A 4	F 4	K 2	
V 2		P 2	Y 1	N 1	H 1	I 3	V 4	
C 2		F 4	C 3	R 2	C 1	A 2	F 2	
G 1		H 1	T 4	V 4	L 7	C 1	E 1	
		W 1	H 1	H 1	W 1	Y 2	W 1	
		Q 2	F 1	D 1	G 1	K 1		
		E 1	K 1	E 1	D 1			
		C 1	I 1	Q 1				
		R 1	G 1					
		D 1						

The number of amino acids found in the data set per residue position. For simplicity, one letter code is used.

the anchor residues are held within allele-specific pockets (named A–F) formed by polymorphic side chains, which are specific to a particular allele.

For our set of 52 peptides, we have generated two models using the Additive method: one containing just the amino acid contributions and one with both amino acid and side chain–side chain interactions contributions. The distribution of amino acids at each position is summarized in Table 3. For the amino acids only model, the non-cross-validated parameters are $r^2 = 0.984$, SEE = 0.118 and $F = 410.97$ while the leave-one-out cross-validation (CV-LOO) gives $q^2 = 0.435$ and SEP = 0.710 and the number of components was 6. For the amino acids plus side chain interaction model, the leave-one-out cross-validation gives $q^2 = 0.250$ and SEP = 0.788 with components numbering 5 and the non-cross-validated parameters being $r^2 = 0.985$, SEE = 0.112 and $F = 538.17$. The contributions of the amino acids at different positions are presented in Fig. 1.

As the amino acids only model gives much the better statistics, we will only explore this model here. Using a contribution of 0.2 as a threshold for both positive and negative contributions, and three letter codes for the amino acids, Phe is favoured at position P1, while Ser, Thr, Asp, and Val all make a negative contribution. For position P2, Proline is the preferred amino acid, with Ala being deleterious. For position P3 Trp, Val, and Ile are preferred amino acids, while Gln, Lys, and Pro are deleterious. At position P4, His and Arg are favoured, while Ser and Gly contribute negatively. For position P5, Ser, Arg, Thr, and Phe are preferred, whereas Gly, Val, and Glu are deleterious. For position P6, Phe and Leu are preferred, while Ile, Asp, and Val are deleterious. For position P7, Phe, Leu, and Val are preferred, while Cys,

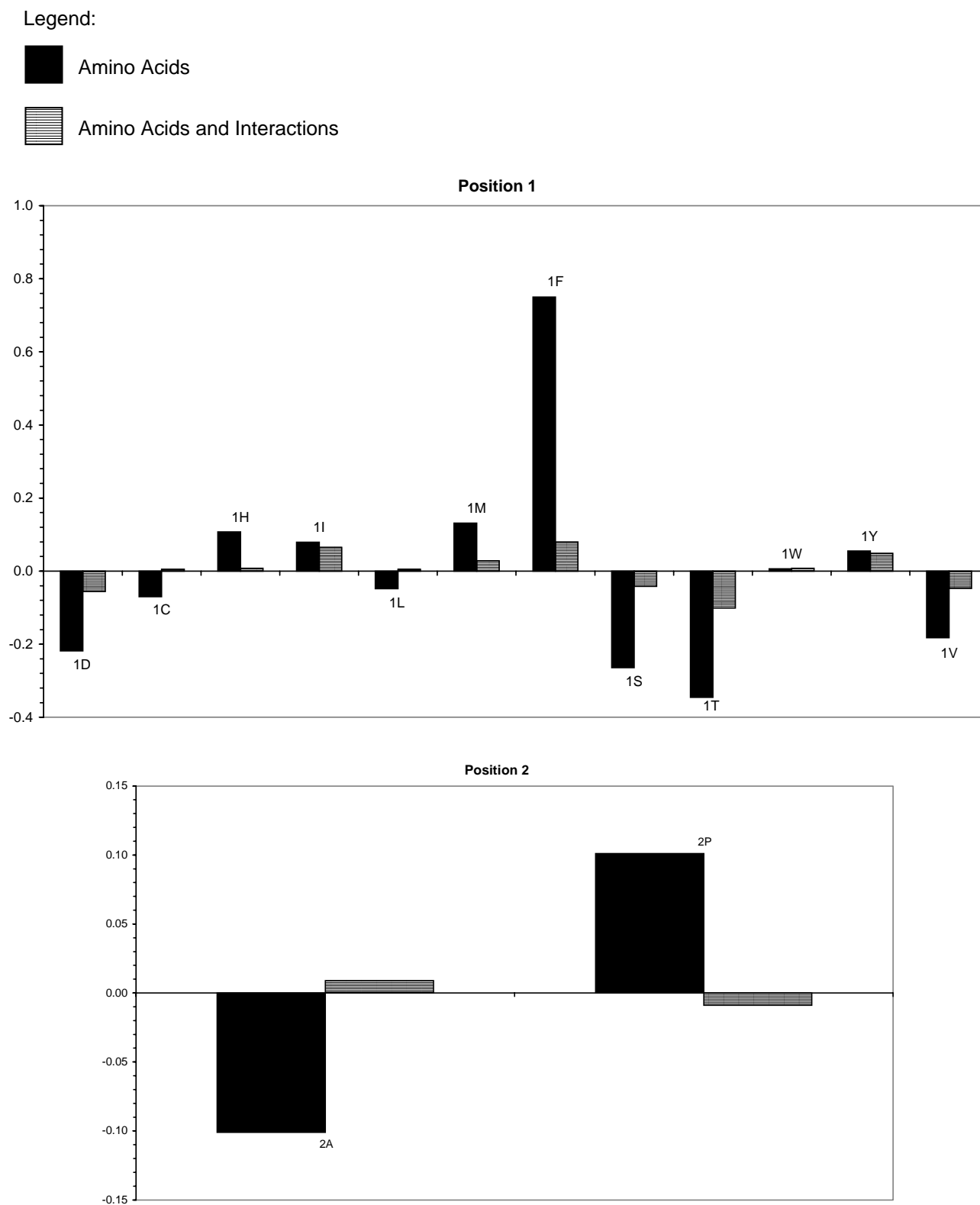


Fig. 1. HLA-B*3501 Model: contribution of position-wise Amino Acids. The contribution made by different individual amino acids at each position of a generalized 9mer HLA-B*3501 binding peptide. A contribution is equivalent to a position-wise Amino Acid regression coefficient obtained by PLS regression, as described in the text. Two models are shown: an amino acids only model and an amino acids plus pairwise side chain–side chain interaction model.

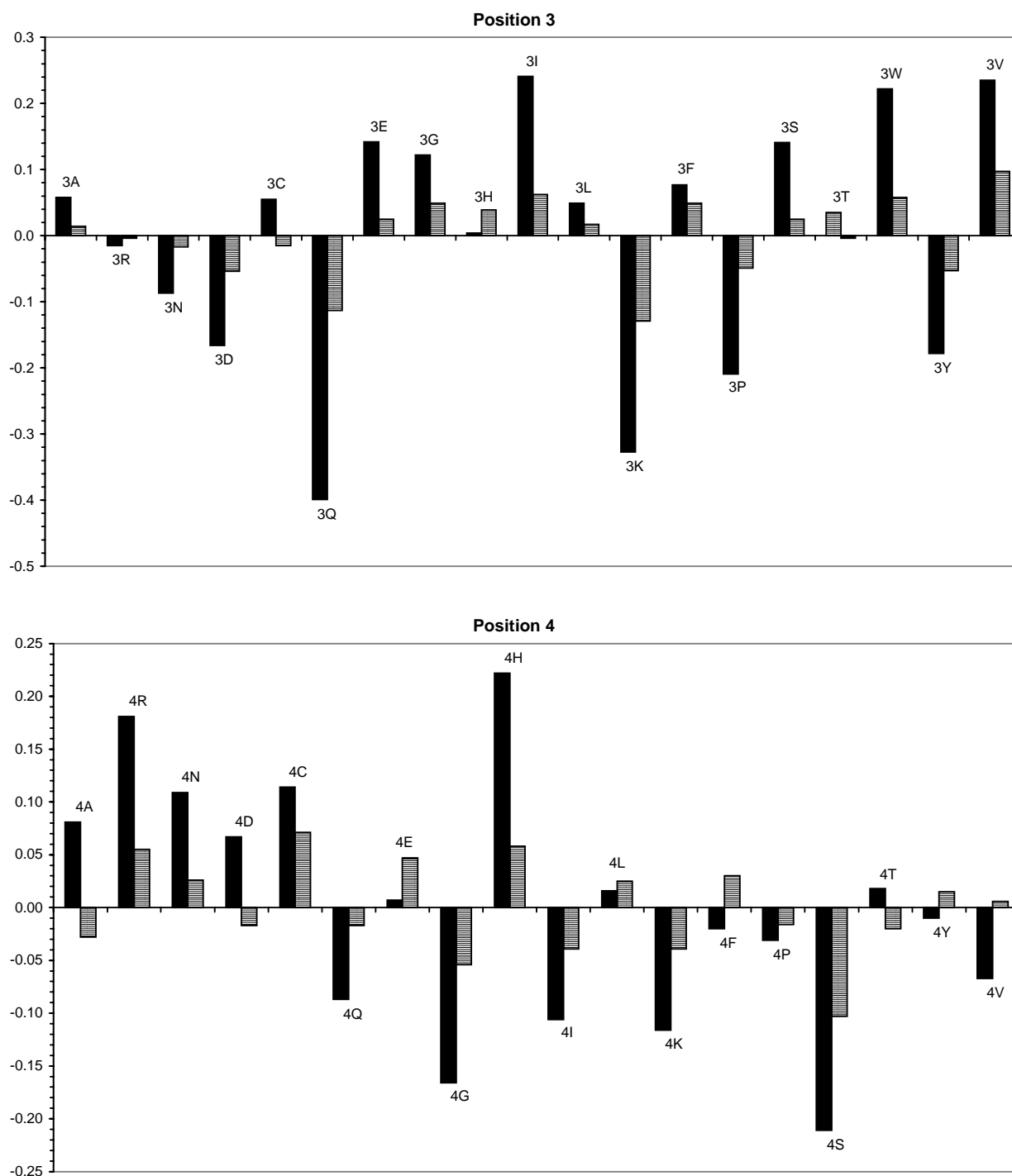


Fig. 1. (Continued)

Ser, and Gln contribute negatively. At position P8, Ala and Lys are preferred, while Thr is deleterious. Finally at P9, Met and Phe contribute positively, whereas amino acids Thr and Trp make negative contributions.

For the with-interactions model, among the (1)–(2) side chain interactions the most favoured is 1F2P; 8P9M, 8P9W and 1F2A are also well tolerated, while 2G3V, 8P9Q, 1F2G and 8P9T contribute negatively. Among the (1)–(3) side

chain interactions 1F3G, 7V9M, and 1F3I are favoured while 1L3G, 2G4R, 7V9Q and 7V9T make significant negative contributions. As might be expected, for the case of HLA-B*3501, the two models (with-interaction and without-interaction) give quite different amino acid contributions in some cases. This is data dependent, as the quality of the interaction data improves, so does the overall statistics.

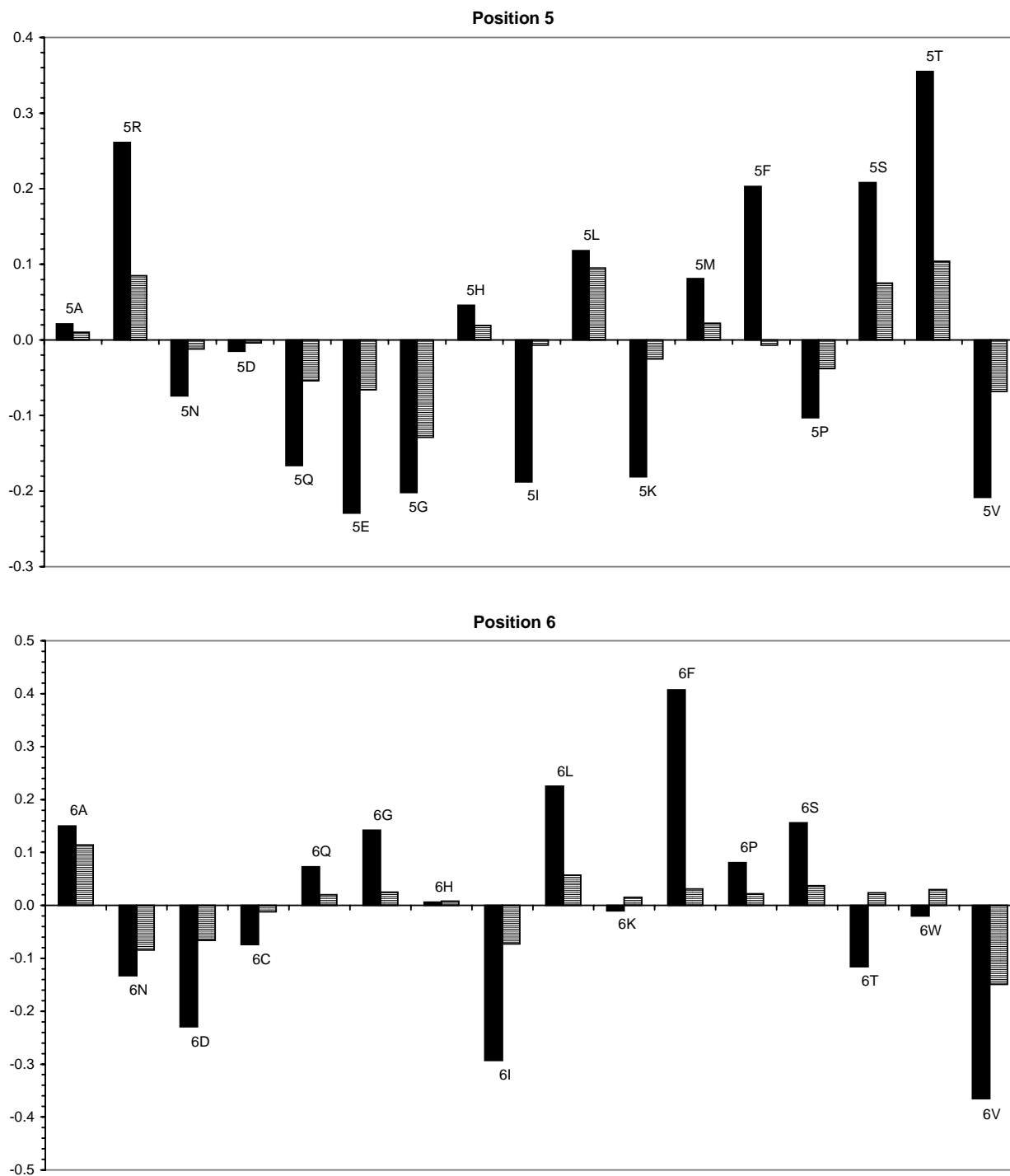


Fig. 1. (Continued)

Some of these observations find explanation in crystallographic data. For example, the complex of HLA-B*3501 with LPPLDITPY suggests that the MHC positions the C terminus in a unique manner solely determined by interactions with polymorphic residues of the heavy chain independent of peptide length. It also reveals an altered conformation resulting in both compression of the peptide

and a shifting of the peptide main-chain. Accordingly, the side chain of P4, P6 and P7 of the *ebna*-peptide are almost entirely exposed to the solvent, whereas, P5 is buried deeply within the groove [12]. The catholic preference for hydrophobic residues at position 9 is a unique feature of HLA-B*3501. The phenomenon may be explained by the presence of residue Ser116, which may hydrogen bond to

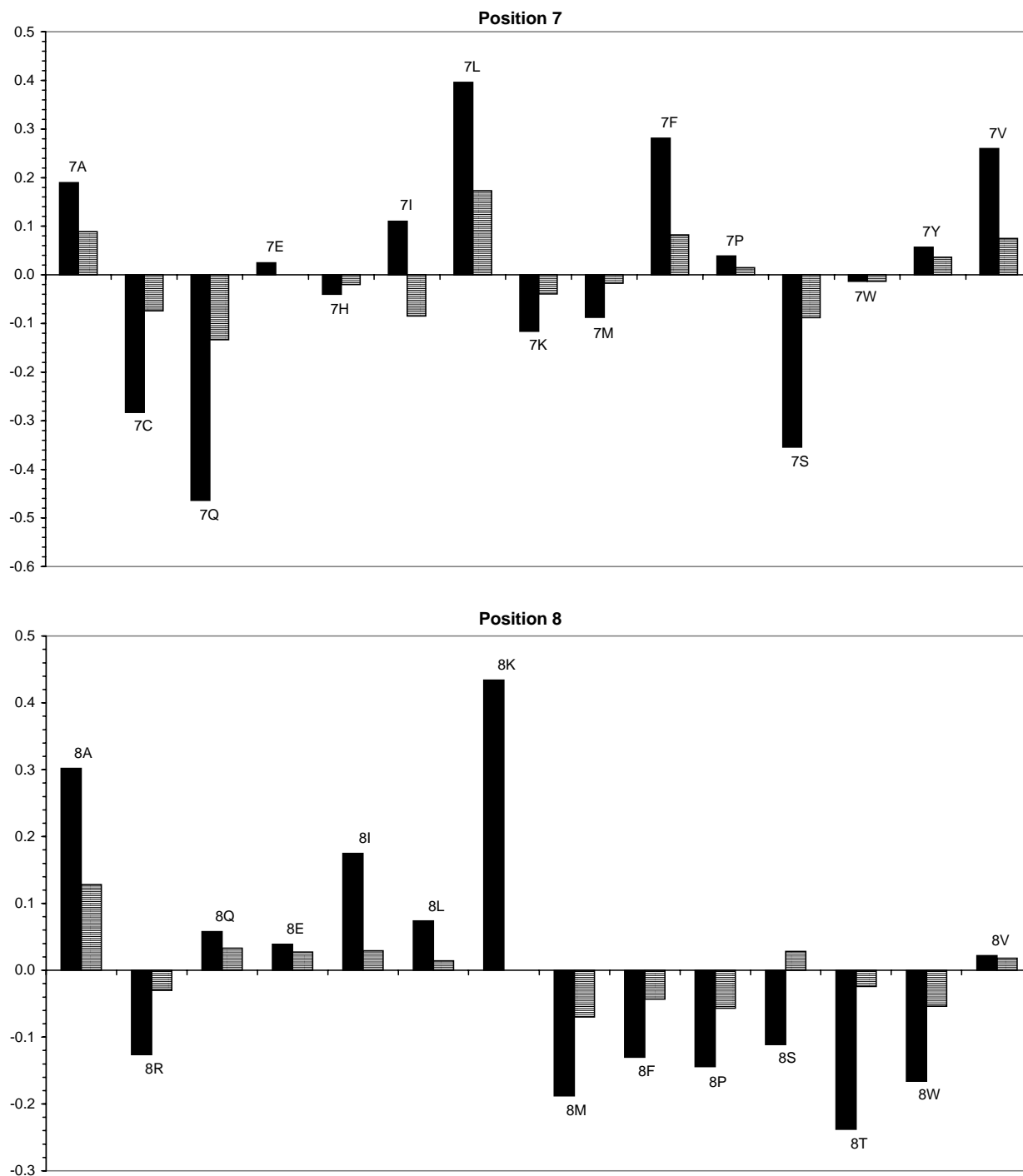


Fig. 1. (Continued)

the Tyr side chain and, through its decreased size, may increase the volume of pocket F allowing larger side chains to be accommodated. On the other hand, Leu81 fills the “end” of the pocket and pushes this residue in the direction of pocket A and closer to the α_1 -helix. This orientation allows the tyrosine-OH group to interact with Ser116 and Tyr74. Tyr99 usually hydrogen bonds with either ASP at P5 or, otherwise, with Tyr9 [12,13]. Another unusual fea-

ture of HLA-B*3501 allele is that the C-terminus of the bound peptide is found closer to the A-pocket than other HLA class I peptides. The altered C-terminal position of the peptide is accompanied by a shift of the N-terminal part of the α_2 helix which seems to be crucial for maintaining the conserved H-bonds between Tyr84, Lys146, Thr143 and the terminal carboxylate of the peptide. These changes are not dependent on peptide length and are characteristic

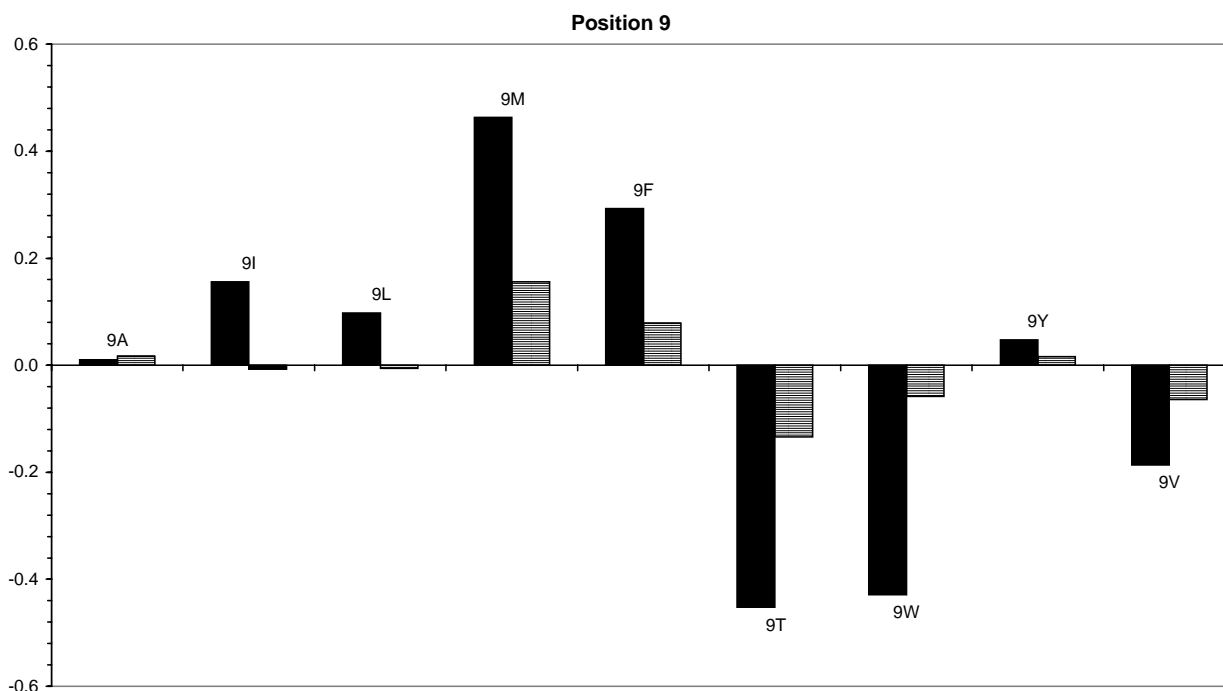


Fig. 1. (Continued).

of HLA-B*3501. Another unique property of HLA-B*3501 is the direct formation of H-bonds between the main chain atoms of P5 and the side chains of Asn70 and Thr73 in contrast to other class I MHC molecules where such interactions are typically mediated via water molecules. However, peptide MHC complexes are relatively few in number relative to the 14000 + binding data we have in our JenPep database [8], and our ability to accurately model or, to use the current terminology, dock a peptide to an MHC is confounded by the extreme flexibility of nine amino acid peptides. Simple sequence changes can give rise to significant, and unexpected, changes in the peptide conformation and thus binding mode. Thus, it is not possible to account for all SAR in this way. Ultimately, sophisticated and powerful methods such as molecular dynamics will become the tool to examine such phenomena.

In developing these methods, we have encountered problems rarely seen in QSAR analyses of small molecules: the size of the peptide molecules being studied, the number of molecules being investigated (maybe 10 times greater) and the great diversity of physicochemical properties associated with each position of the peptide being examined. While the additive method cannot explain the nature of forces involved in such interaction, be that steric, electrostatic, hydrophobic, or hydrogen bonding, or some combination thereof it can quantitatively assess the significance of individual side chains for the affinity, leading to the straightforward design of enhanced binders. Moreover, these results may contain minor, model specific, statistical

peculiarities, however, we would expect that these should decrease as the number and diversity of peptides we study increases.

3.2. The MHCpred server

Currently, MHCpred supports 11 class I HLA allele models: A*0101, A*0201, A*0202, A*0203, A*0206, A*0301, A*1101, A*3301, A*6801, A*6802, and B*3501. A useful result from work on the definition of MHC binding motifs work is that MHC alleles can be clearly grouped together into so-called supertypes, which exhibit broad supermotifs, based on the commonality of their substrate specificity [14]. Most of the models described earlier fall into the A2 and A3 supertypes.

The model for A*0201 has been described in detail before [7], the model for B*3501 is described earlier, and details of the other models will be published elsewhere [15,16] (Table 1). This range of models, which focuses primarily on the HLA-A locus, represents a set of alleles which are widely distributed in the human population and for which considerable binding data is available. As ever, the quality of data is reflected in the quality of predictive models. We anticipate that we will extend this number significantly with an increased number of models for both mouse and human alleles, with an increasing focus on the HLA-B and HLA-C loci.

Although we have implemented the method as a server, it is possible to perform the calculation by hand, and it is illuminating to review this approach. Take, for example, the

peptide FPIRSFVPM

$$\begin{aligned} \text{pIC}_{50} = & \text{constant} + 1\text{F} + 2\text{P} + 3\text{I} + 4\text{R} + 5\text{S} + 6\text{F} + 7\text{V} \\ & + 8\text{P} + 9\text{M} + 1\text{F}2\text{P} + 2\text{P}3\text{I} + 3\text{I}4\text{R} + 4\text{R}5\text{S} \\ & + 5\text{S}6\text{F} + 6\text{F}7\text{V} + 7\text{V}8\text{P} + 8\text{P}9\text{M} + 1\text{F}3\text{I} + 2\text{P}4\text{R} \\ & + 3\text{I}5\text{S} + 4\text{R}6\text{F} + 5\text{S}7\text{V} + 6\text{F}8\text{P} + 7\text{V}9\text{M} \end{aligned}$$

substituting each term by its quantitative value the final calculated pIC_{50} value is 8.723.

$$\begin{aligned} \text{pIC}_{50} = & 6.261 + 0.404 + 0.097 + 0.04 - 0.01 + 0.105 \\ & + 0.083 + 0.093 - 0.031 + 0.230 + 0.501 \\ & + 0.04 \pm \text{absent} \pm \text{absent} + 0.04 \pm \text{absent} \\ & - 0.009 + 0.266 + 0.209 + 0.087 \\ & + 0.066 \pm \text{absent} \pm \text{absent} - 0.015 \\ & + 0.266 = 8.723 \end{aligned}$$

It is important to note that as the number of absent values increases, the chance of incorrect prediction also increases. The coefficients for these missing values can be set to be negative in order to decrease the number of false positives high binders generated. When one wishes to be cautious, then this value is set to be large, when one is keen to be more catholic then it would be set at a considerably lower value. The MHCpred server currently uses a value of 0.0 for both missing amino acid and missing amino acid plus interaction terms.

4. Discussion

Different MHC alleles, both class I and class II, exhibit different peptide specificities: peptides are bound with particular sequence patterns, leading to the development of so-called motifs [17]. Motifs are usually expressed in terms of anchor residues: the presence of certain amino acids at particular positions that are thought to be essential for binding. Taking human class I allele HLA-B*3501 as our example, previous studies have indicated the need for anchor residues at positions 2 (Pro) and 9 (hydrophobic or aromatic residues, such as Phe, Met, Leu, Ile and especially Tyr). Primary anchor residues, although generally deemed to be necessary, are not sufficient for peptide binding, and secondary anchors, residues that are favourable, but not essential, for binding may also be required; other positions show positional preferences for particular amino acids. Moreover, the presence of certain residues at specific positions of a peptide can have a negative effect on binding [13,18,19]. Although motif methods are admirably simple—it is easy to implement either by eye or more systematically scanning protein sequences computationally—there are many problems with the motif approach. Although it is possible to score the relative contributions of primary and secondary anchors to produce a rough and ready measure of binding affinity [20,21],

the most significant problem with the motif approach is that it is, fundamentally, a deterministic method. A peptide is either a binder or is not a binder. A brief reading of the literature shows that motif matches produce many false positives, and are, in all probability, producing an equal number of false negatives. Indeed there are many examples where peptides without both dominant anchors still bind with high affinity.

A more accurate description of this phenomenon is to say that MHCs bind peptides with an equilibrium binding constant dependant on the nature of the bound peptide's sequence. The driving forces behind this binding are precisely the same as those driving drug binding. Within the human population there are an enormous number of different, variant genes coding for MHC proteins, each exhibiting a different peptide-binding sequence selectivity. T cell receptors, in their turn, also exhibit different affinities for pMHC. The combined selectivity of both MHCs and TCRs determines the power of peptide recognition within the immune system and through this phenomenon the recognition of foreign pathogens.

Experimentally, there are many ways to measure binding affinity. IC_{50} values are the most widely quoted binding affinity measures and are calculated from a competitive binding assay [22]. The value given is the concentration required for 50% inhibition of a standard labelled peptide by the test peptide. Therefore, nominal binding affinity is inversely proportional to an IC_{50} value. Once a peptide has bound to a MHC to be recognised by the immune system, the pMHC complex has to be recognised by one of the TCRs of the T cell repertoire. It is generally accepted that a peptide binding to an MHC may be recognized, by a TCR, if it binds with a $\text{pIC}_{50} > 6.3$ [23]. There is some evidence suggesting that as the MHC binding affinity of a peptide rises, the greater the probability that it will be a T cell epitope. The prediction, then, of MHC binding is both the best understood, and, probably, the most discriminating step in the presentation-recognition pathway. A pragmatic solution to the as yet unsolved problem of what will be recognised by the TCR, and thus activate the T cell, is to greatly reduce the number of possible epitopes using MHC binding prediction, and then test the remaining candidates using some measure of T cell activation, such as T cell killing or thymidine incorporation.

In this paper we have exemplified a recently developed method for the prediction of MHC binding [7]: a Free-Wilson based QSAR approach called the additive method. We report here a new example of the method (HLA-B*3501), and an Internet server called MHCpred. There has, in the last 5–10 years, been a sea change in the way in which computer scientists have made their methods available to others: servers have replaced, or are replacing, other methods, such as the distribution of tapes or the use of FTP sites. This change has many corollaries: although the creators of servers can maintain a great degree of secrecy regarding both algorithmic details and the nature of their

implementation, this is compensated for by the easy of use of HTML user interfaces, particularly amongst novices, and questions of interoperability are addressed, to some extent at least, by the use of automated CGI/perl.

A number of methods for the prediction of MHC binding affinity have now been implemented as servers accessible via the World Wide Web. A number of these target class I only, class II only, or target both antigen presentation pathways. Class I methods include PREDICT (<http://sdmc.krdl.org.sg:8080/predict/>), lpprep (<http://reiner.bu.edu/zhiping/lpprep.html>), BIMAS (http://bimas.dcrn.nih.gov/molbio/hla_bind/) [24], PREDEP (<http://bioinfo.md.huji.ac.il/marg/Teppred/mhc-bind/>), NetMHC (<http://www.cbs.dtu.dk/services/NetMHC/>). Class II methods include [MHC-THREAD (<http://www.csd.abdn.ac.uk/~gjl/MHC-Thread/>) [25], Epi-Predict (<http://www.epipredict.de/index.html>) [26], HLA-DR4 binding (<http://www-dcs.nci.nih.gov/branches/surgery/sbprog.html>) [27], ProPred (<http://www.imtech.res.in/raghava/propred/>) [28]. The SYFPEITHI website is the only on-line method that covers both class I and class II (<http://syfpeithi.bmi-heidelberg.com/Scripts/MHCServer.dll/EpiPredict.htm>) [20]. The provenance of certain of these methods is uncertain as many remain unpublished.

In contrast to the methods implemented in these servers, we have taken a quantitative approach to the prediction of MHC binding, which is important for several reasons. First, it allows for a clear and unbiased comparison with measured affinities. The only other quantitative method currently implemented on-line is BIMAS [24]. This server differs from ours in that it predicts kinetic, rather than thermodynamic, properties of the dissociation of the MHC β_2 microglobulin complex, rather than the energetics of peptide-MHC interaction.

Secondly, binding affinity is clearly the product of the interactions of whole peptide with the receptor molecule, and thus methods based on anchor positions [20] are unlikely to generate a low proportion of either false positive or false negative predictions. Thirdly, because our method is quantitative, it has direct application to the important area of heteroclitic peptide generation [29]. Heteroclitic peptides are modified natural, mildly immunogenic peptides with enhanced MHC binding, and, typically, enhanced T cell responses. Use of our approach allows the rapid identification of amino-acid substitutions likely to increase binding and thus T cell activation, with a concomitant benefit in the development of epitope based vaccines.

The additive method relies on the existence of amino acids at a particular position within the training set in order for it to reliably predict the effect of that residue, at that position, on affinity in other peptide sequences. Apart from the addition of other human and mouse allele models, future technical developments will include implementing a descriptor based [30,31], rather than a contribution based, approach to peptide QSAR. This will have a significant impact on the generality of the method and allow us to predict the effect on binding of non-natural amino acids more accurately.

When one begins to write up a paper for publication, experience has taught us to accentuate the positive, obliging us, sometimes at least, to obfuscate the negative. At the prompting of Editor and Referees, we are minded to be more open and candid in discussions. Clearly there is much in our work that might concern referees. As computational chemists practicing QSAR, unfortunately much of it is out of our hands. For example, the peptide sets we use are larger than is typical in the pharmaceutical literature, at least for activity, as opposed to physical property (i.e. ADMET) related, prediction. The peptides themselves are physically large in themselves, and their physical properties are extreme. For example, they can be multiply charged, zwitterionic, and/or exhibit a huge range in hydrophobicity. The sequences, and thus the properties, of the peptides are heavily biased in our peptide sets. This results in part from processes of pre-selection that result in self-reinforcement. Simple motifs are often used to reduce the experimental burden of epitope identification: very sparse sequence patterns are matched and the corresponding peptides tested, with an enormous concomitant reduction in peptide diversity. Moreover, affinity data is of an intrinsically inferior quality: multiple measurements of the same peptide may vary by several orders of magnitude, some values are clearly wrong, a mix of different standard peptides are used in radioligand competition assays, experiments are conducted at different temperatures and over different concentration ranges. We are also performing a “meta-analysis” almost certainly forcing many distinct binding modes into a single QSAR model. In an ideal world we would look at a variety of “internally rich” data from ITC, volumetric analysis, and fluorescence spectroscopy, but to do this on an appropriate scale would be prohibitively time consuming and expensive. Where one might conceive of doing this for a single allele, there are dozen upon dozen of common alleles in the human population and hundreds more at lower frequency in the total population, albeit many have interesting links to disease. To pursue this for all “interesting” alleles is, thus, currently beyond the scope of existing methodology. Compared to all of these caveats concerning data quality, concerns about parameterization pale somewhat. Nonetheless, one may also raise legitimate concerns about the methodology. In CoMFA and CoMSIA, tens of thousands of variables (i.e. GRID positions) are correlated with a single activity measure. We do something similar but derive these variables from a set of peptide sequences. Ours are indicator variables, of course, and we produce a regression equation against those rather than against continuous variables as in CoMFA. Clearly, our method is not perfect, nor do we pretend that it is. Nor, let us be fair, are methods in use by others. The failure of our methods, and those of others, is as much to do with problems relating to the underlying data as it is to do with minor methodological flaws. One principal criticism of our statistical methodology is, however, the chance of over fitting our models. Our X-block is very ‘over-square’ with insufficient degrees of freedom for us to undertake an exhaustive and

completely robust analysis. Moreover, for many terms we have low populations, often only single observations, inflating the associated errors and thus reducing their reliability in prediction. This may give rise to the interesting differences seen in with-interactions and without-interaction models. MHCpred addresses this by offering a choice between with and without models for each allele. We are certainly aware of these dangers and seek to minimise them. For example, we could remove cross-terms with only single observations or try to group amino acids into a reduced number of “types”, i.e. aromatic or charged residues. It is also possible to over emphasise the usefulness of cross-validation and q^2 as measures of performance [32], high values of leave-one-out q^2 are a necessary, but not a sufficient, condition for a model to possess high predictivity. External test sets and randomisation of training data are also important criteria for assessing model quality. For this model, however, it proved difficult to construct meaningful training and training sets, as we have for other alleles. Ultimately, however, we are limited by the data itself. If one visualises our efforts at datamining as a ship, it would be easy for it to founder on the rock of data quality. This is the main issue here, without doubt. With a properly designed training set most of these issues would be resolved. In light of this, we have attempted, and, nonetheless, clearly succeeded in producing useful, if imperfect, models with significant utilitarian value. Eventually, as molecular dynamics methods develop and become common place tools for molecular modelling, these techniques will enable us to develop, in concert with measured data, considerably more accurate and predictive models.

The accurate prediction of epitopes is vital to our goal of developing computer-aided vaccine design or CAVD [33,34]. From the perspective of human disease, a proper understanding of the immune system is vital. Indeed, apart from its key role in oncology and autoimmunity, the immune system has evolved to combat the threat from infectious disease. Disease is the most significant, but also most preventable, causes of death worldwide. In contrast to other causes of death, it can be attacked systematically through the use of biological and chemical entities, such as vaccines and drugs, as well as through other mechanisms such as enhanced public health. The discovery of vaccines-hitherto the province of highly empirical study-is, in the post-genomic era, undergoing a sea change, benefiting from the informational tsunami poised to sweep across much of biology. By making MHCpred freely available, we hope to foster collaboration within the field of computational vaccinology. A deep understanding of the phenomenology of the molecular mechanisms that underlie immunology are being complemented by an increasingly quantitative capacity for prediction. Thus, the ability to reliably predict MHC binding will enable us to analyse microbial immunomes, identifying the most antigenic epitopes and proteins, and thus select sets of favoured putative vaccines.

References

- [1] R. Holtappels, N.K. Grzimek, D. Thomas, M.J. Reddehase, Early gene *m18*, a novel player in the immune response to murine cytomegalovirus, *J. Gen. Virol.* 83 (2002) 311–316.
- [2] R. Holtappels, D. Thomas, J. Podlech, M.J. Reddehase, Two antigenic peptides from genes *m123* and *m164* of murine cytomegalovirus quantitatively dominate CD8 T-cell memory in the H-2d haplotype, *J. Virol.* 76 (2002) 151–164.
- [3] J.L. Pellequer, E. Westhof, PREDITOP, a program for antigenicity prediction, *J. Mol. Graph* 11 (1993) 204–210.
- [4] A.J. Alix, Predictive estimation of protein linear epitopes by using the program PEOPLE, *Vaccine* 18 (1999) 311–314.
- [5] J.M. Thornton, M.S. Edwards, W.R. Taylor, D.J. Barlow, Location of ‘continuous’ antigenic determinants in the protruding regions of proteins, *EMBO J.* 5 (1986) 409–413.
- [6] D.R. Flower, I.A. Doytchinova, K. Paine, P. Taylor, M.J. Blythe, D. Lamponi, C. Zygouri, P. Guan, H. McSparron, H. Kirkbride, Computational vaccine design, in: D.R. Flower (Ed.), *Drug Design, Cutting Edge Approaches*, RSC Publications, London, 2002.
- [7] I.A. Doytchinova, M.J. Blythe, D.R. Flower, An additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201, *J. Proteome Res.* 1 (2002) 263–272.
- [8] M.J. Blythe, I.A. Doytchinova, D.R. Flower, JenPep, a database of quantitative functional peptide data for immunology, *Bioinformatics* 18 (2002) 434–439.
- [9] Y. Cheng, W.H. Prusoff, Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50% inhibition (I_{50}) of an enzymatic reaction, *Biochem. Pharmacol.* 22 (1973) 3099–3108.
- [10] SYBYL 6.7. Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144.
- [11] B.L. Bush, R.B. Nachbar Jr., Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA, *J. Comput. Aided Mol. Des.* 7 (1993) 587–619.
- [12] R. Menssen, P. Orth, A. Ziegler, W. Saenger, Decamer-like conformation of a nona-peptide bound to HLA-B*3501 due to non-standard positioning of the C terminus, *J. Mol. Biol.* 285 (1999) 645–653.
- [13] K.J. Smith, S.W. Reid, D.I. Stuart, A.J. McMichael, E.Y. Jones, J.I. Bell, An altered position of the alpha 2 helix of MHC class I is revealed by the crystal structure of HLA-B*3501, *Immunity* 4 (1996) 203–213.
- [14] F. Sinigaglia, J. Hammer, Motifs and supermotifs for MHC class II binding peptides, *J. Exp. Med.* 181 (1995) 449–451.
- [15] P. Guan, I.A. Doytchinova, D.R. Flower, HLA-A3-supermotif defined by quantitative structure–activity relationship analysis. *Prot Eng.* 16 (2003) 11–18.
- [16] I.A. Doytchinova, D.R. Flower, The HLA-A2-supermotif: A QSAR definition, *J. Comp. Biol.* 1 (2003) 2648–2654.
- [17] A. Sette, S. Buus, E. Appella, J.A. Smith, R. Chesnut, C. Miles, S.M. Colon, H.M. Grey, Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis, *Proc. Natl. Acad. Sci. U.S.A.* 86 (1989) 3296–3300.
- [18] Y. Takamiya, C. Schonbach, K. Nokihara, M. Yamaguchi, S. Ferrone, K. Kano, K. Egawa, M. Takiguchi, HLA-B*3501-peptide interactions: role of anchor residues of peptides in their binding to HLA-B*3501 molecules, *Int. Immunol.* 6 (1994) 255–261.
- [19] J. Sidney, M.F. del Guercio, S. Southwood, V.H. Engelhard, E. Appella, H.G. Rammensee, K. Falk, O. Rotzschke, M. Takiguchi, R.T. Kubo, Several HLA alleles share overlapping peptide specificities, *J. Immunol.* 154 (1995) 247–259.
- [20] H. Rammensee, J. Bachmann, N.P. Emmerich, O.A. Bachor, S. Stevanovic, SYFPEITHI: database for MHC ligands and peptide motifs, *Immunogenetics* 50 (1999) 213–219.
- [21] J. D’Amaro, J.G. Houbiers, J.W. Drijfhout, R.M. Brandt, R. Schipper, J.N. Bavinck, C.J. Melief, W.M. Kast, A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs, *Hum. Immunol.* 43 (1995) 13–18.

- [22] J. Ruppert, J. Sidney, E. Celis, R.T. Kubo, H.M. Grey, A. Sette, Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules, *Cell* 74 (1994) 929–934.
- [23] A. Sette, A. Vitiello, B. Rehman, P. Fowler, R. Nayarsina, W.M. Kast, C.J. Melief, C. Oseroff, L. Yuan, J. Ruppert, The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes, *J. Immunol.* 153 (1994) 5586–5592.
- [24] K.C. Parker, M.A. Bednarek, J.E. Coligan, Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chains, *J. Immunol.* 152 (1994) 163–175.
- [25] M.T. Swain, A.J. Brooks, G.J.L. Kemp, An automated approach to modelling class II MHC alleles and predicting peptide binding, in: Proceedings of the Second IEEE International Symposium on Bio-Informatics and Biomedical Engineering, IEEE Computer Society Press, in press.
- [26] B. Fleckenstein, F. von der Mulbe, J. Wessels, D. Niethammer, K.H. Wiesmuller, From combinatorial libraries to MHC ligand motifs, T-cell superagonists and antagonists, *Jung G. Biologicals* 29 (2001) 179–181.
- [27] C.E. Touloukian, W.W. Leitner, S.L. Topalian, Y.F. Li, P.F. Robbins, S.A. Rosenberg, N.P. Restifo, Identification of a MHC class II-restricted human gp100 epitope using DR4-IE transgenic mice, *J. Immunol.* 164 (2000) 3535–3542.
- [28] H. Singh, G.P. Raghava, ProPred, prediction of HLA-DR binding sites, *Bioinformatics* 17 (2001) 1236–1237.
- [29] S. Tangri, G.Y. Ishioka, X. Huang, J. Sidney, S. Southwood, J. Fikes, A. Sette, Structural features of peptide analogs of human histocompatibility leukocyte antigen class I epitopes that are more potent and immunogenic than wild-type peptide, *J. Exp. Med.* 194 (2001) 833–846.
- [30] S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, Peptide quantitative structure–activity relationships, a multivariate approach, *J. Med. Chem.* 30 (1987) 1126–1135.
- [31] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, *J. Med. Chem.* 41 (1998) 2481–2491.
- [32] A. Golbraikh, A. Tropsha, Beware of q_2 , *J. Mol. Graph Model* 20 (2002) 269–276.
- [33] I.A. Doytchinova, D.R. Flower, Toward the quantitative prediction of T-cell epitopes, CoMFA and CoMSIA studies of peptides with affinity for the Class I MHC molecule HLA-A*0201, *J. Med. Chem.* 44 (2001) 3572–3581.
- [34] I.A. Doytchinova, D.R. Flower, Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex. A three-dimensional quantitative structure–activity relationship study, *Proteins* 48 (2002) 505–518.
- [35] D.L. Doolan, S.L. Hoffman, S. Southwood, P.A. Wentworth, J. Sidney, R.W. Chesnut, E. Keogh, E. Appella, T.B. Nutman, A.A. Lal, D.M. Gordon, A. Oloo, A. Sette, Degenerate cytotoxic T cell epitopes from *P. falciparum* restricted by multiple HLA-A and HLA-B supertype alleles, *Immunity* 7 (1997) 97–112.
- [36] R. Bertonni, J. Sidney, P. Fowler, R.W. Chesnut, F.V. Chisari, A. Sette, Human histocompatibility leukocyte antigen-binding supermotifs predict broadly cross-reactive cytotoxic T lymphocyte responses in patients with acute hepatitis, *J. Clin. Invest.* 100 (1997) 503–513.
- [37] J. Sidney, M.F. del Guercio, S. Southwood, V.H. Engelhard, E. Appella, H.G. Rammensee, K. Falk, O. Rotzschke, M. Takiguchi, R.T. Kubo, et al., Several HLA alleles share overlapping peptide specificities, *J. Immunol.* 154 (1995) 247–259.
- [38] J. Sidney, S. Southwood, M.F. del Guercio, H.M. Grey, R.W. Chesnut, R.T. Kubo, A. Sette, Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules, *J. Immunol.* 157 (1996) 3480–3490.