European Journal of Medicinal Chemistry 45 (2010) 236–243

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Peptide binding to the HLA-DRB1 supertype: A proteochemometrics analysis

Ivan Dimitrov [a], Panayot Garnev [a], Darren R. Flower [b], Irini Doytchinova [a,*]

[a] Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st, 1000 Sofia, Bulgaria
[b] The Jenner Institute, Oxford University, Compton, RG20 7 NN, Berkshire, UK

ARTICLE INFO

Article history:
Received 20 June 2009
Received in revised form
4 September 2009
Accepted 29 September 2009
Available online 13 October 2009

Keywords:
Proteochemometrics
QSAR
Epitope prediction
MHC class II

ABSTRACT

A proteochemometrics approach was applied to a set of 2666 peptides binding to 12 HLA-DRB1 proteins. Sequences of both peptide and protein were described using three *z*-descriptors. Cross terms accounting for adjacent positions and for every second position in the peptides were included in the models, as well as cross terms for peptide/protein interactions. Models were derived based on combinations of different blocks of variables. These models had moderate goodness of fit, as expressed by r^2, which ranged from 0.685 to 0.732; and good cross-validated predictive ability, as expressed by q^2, which varied from 0.678 to 0.719. The external predictive ability was tested using a set of 356 HLA-DRB1 binders, which showed an r^2_{pred} in the range 0.364–0.530. Peptide and protein positions involved in the interactions were analyzed in terms of hydrophobicity, steric bulk and polarity.

1. Introduction

Major histocompatibility complex (MHC) proteins, also known as human leukocyte antigens (HLA), are glycoproteins which bind within the cell short peptides, also called epitopes, derived from host and/or pathogen proteins, and present them at the cell surface for inspection by T-cells. T-cell recognition is a fundamental mechanism underlying the adaptive immune system through the action of which the host identifies and responds to foreign antigens [1].

There are two classes of MHC molecules: class I and class II. MHC class I molecules typically present peptides from proteins synthesized within the cell (endogenous processing pathway). MHC class II proteins primarily present peptides derived from endocytosed extracellular proteins (exogenous processing pathway). Both classes of MHC proteins are extremely polymorphic. More than 3500 molecules are listed in IMGT/HLA database [2]. MHC class I proteins are encoded by three loci: HLA-A, HLA-B and HLA-C. MHC class II proteins also are encoded by three loci: HLA-DR, HLA-DQ and HLA-DP. The peptide binding site of class I proteins has a closed cleft, formed by a single protein chain (α-chain) [3]. Usually, only short peptides of 8–11 amino acids bind in extended conformation. In contrast, the cleft of class II proteins is open-ended, allowing much longer peptides to bind, although only 9 amino acids actually occupy the site. The class II cleft is formed by two separate protein chains: α and β [3]. Both clefts have binding pockets, corresponding to primary and secondary anchor positions on the binding peptide. The combination of two or more anchors is called a motif. The experimental determination of motifs for every allele is prohibitively expensive in terms of labor, time and resources. The only practical alternative is to make use of a bioinformatics approach.

Many bioinformatics methods exist to predict peptide-MHC class II binding (for a recent review, see Ref. [1]). These approaches can be classified into three groups, according to the underlying methodology employed: quantitative matrices (QM), artificial neural networks (ANN), or support vector machines (SVM). Although MHC class II binding predictions are more complex than MHC class I predictions, most of the available servers have good predictive ability. They are used to preselect suitable targets for subsequent experimental validation. The alternative systematic MHC binding mapping is costly and time-consuming because it requires synthesis and testing of large numbers of overlapping peptides corresponding to the whole target protein sequence. High affinity MHC binders can be potential vaccine candidates, as T-cells only recognize and respond to peptides bound to MHC molecules.

All available methods for MHC binding prediction treat each set of peptides binding to a particular MHC protein separately, developing models for peptide binding prediction for only one particular protein target. In contrast, proteochemometrics, which is a recent QSAR approach developed by Wikberg et al. [4], deals with ligands that bind to a set of similar proteins. Proteochemometrics is specifically designed to solve QSAR tasks where a set of ligands binds to a set of related proteins. In a conventional QSAR analysis, the *X* matrix of descriptors only includes chemical information from ligands; in a proteochemometrics analysis the *X* matrix

* Corresponding author.
E-mail address: idoytchinova@pharmfac.net (I. Doytchinova).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.09.049

```
                10         20         30         40         50         60         70         80         90
DRB1*0101   GDTRPRFLWQ LKFECHFFNG TERVRLLERC IYNQEESVRF DSDVGEYRAV TELGRPDAEY WNSQKDLLEQ RRAAVDTYCR HNYGVGESFT
DRB1*0301   --------EY STS------- -----Y-D-Y FH----N--- ------F--- ---------- ---------- K-GR--N--- -----V----
DRB1*0401   --------E- V-H------- -----F-D-Y F-H---Y--- ---------- ---------- ---------- K--------- ----------
DRB1*0404   --------E- V-H------- -----F-D-Y F-H---Y--- ---------- ---------- ---------- ---------- -----V----
DRB1*0405   --------E- V-H------- -----F-D-Y F-H---Y--- ---------- ------S--- ---------- ---------- ----------
DRB1*0701   ---Q------ G-YK------ ----QF---L F-----F--- ---------- ------V--S ------I--D --GQ---V-- ----------
DRB1*0802   --------EY STG--Y---- -----F-D-Y F-----Y--- ---------- ---------- ------F--D ---L------ ----------
DRB1*0901   ---Q----K- D--------- -----Y-H-G ------N--- ---------- ------V--S ------F--R ---E---V-- ----------
DRB1*1101   --------EY STS------- -----F-D-Y F-----Y--- ------F--- -------E-- ------F--D ---------- ----------
DRB1*1201   --------EY STG--Y---- ---------H FH----LL-- ------F--- ------V--S ------I--D ---------- ----AV----
DRB1*1302   --------EY STS------- -----F-D-Y FH----N--- ------F--- ---------- ------I--D E--------- ----------
DRB1*1501   ---------- P-R------- -----F-D-Y F--------- ------F--- ---------- ------I--- A--------- -----V----
```

Fig. 1. Sequence alignment of HLA-DRB1 proteins (first 90 residues), used in the study.

contains information from both proteins and ligands. One single proteochemometrics model could potentially predict peptide binding to a whole group of MHC proteins. Proteochemometrics has been successfully applied to various classes of G-protein coupled receptors [5–7], antibodies [8], and viral proteases [9,10].

In the present study, a proteochemometrics approach was applied to a set of 2666 peptides binding to 12 HLA-DRB1 proteins. The aim of this study was to develop a QSAR model for binding prediction to a set of HLA-DRB1 proteins and to reveal key ligand–receptor interactions within this set that help determine ligand specificity.

2. Computational methods

2.1. Data sets

Two ligand sets were used in the study: training and test. The training set was used to develop proteochemometrics models whose predicting ability was then assessed using the test set. The training set consisted of 2666 peptides of different lengths which were bound to 12 HLA-DRB1 proteins. Data was extracted from the Immune Epitope Database (http://www.immuneepitope.org) in September 2008, according to the following criteria: Alleles: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0404, DRB1*0405, DRB1*0701, DRB1*0802, DRB1*0901, DRB1*1101, DRB1*1201, DRB1*1301 and DRB1*1501; Assay: Purified MHC – Radioactivity Competition; Quantitative measurement; Units: IC_{50} nM. From a set of overlapping peptides with different lengths, only the longest peptide was included in the training set. Peptide binding affinities were originally assessed using a quantitative assay based on the inhibition of binding of a radiolabeled standard peptide to detergent-solubilized MHC molecules and presented as $pIC_{50} = \log(1/IC_{50})$ [11,12].

The test set included peptides binding to the same DRB1 proteins as the training set. The data was extracted from the AntiJen database [13] and their affinities were assessed using the same radiolabeled assay. All binders common to both sets were deleted from the test set. The final test set consisted of 356 binders.

The HLA class II proteins included in the study belonged to the HLA-DR1 serotype: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0404, DRB1*0405, DRB1*0701, DRB1*0802, DRB1*0901, DRB1*1101, DRB1*1201, DRB1*1301 and DRB1*1501. The protein sequences were collected from the IMGT/HLA database (http://www.ebi.ac.uk/imgt/hla). The HLA class II binding site is formed by ~35 amino acids from the first 80 residues of the α-chain and the first 90 residues of the β-chain [14–18]. As the HLA-DR α-chain (DRA chain) exhibits no binding site polymorphism, only HLA-DR1 β-chains (DRB1 chains) were used in our analysis. DRB1 chains contain 18 polymorphic amino acids in the binding site. They occupy positions 9, 11, 13, 26, 28, 30, 38, 47, 57, 60, 67, 70, 71, 74, 77, 78, 85 and 86 (Fig. 1). Some of the amino acids interact with peptide backbone, others with peptide side chains.

2.2. Description of ligands and proteins

The peptides used in the study were of different length. The MHC binding site can only accommodate 9 amino acids, thus each binder was presented as a set of overlapping nonamers, each with the same IC_{50} values as the parent peptide. Each nonamer was encoded as a string comprising three z-descriptors (z_1, z_2 and z_3) per amino acid. z-Descriptors relate to hydrophobicity, steric effects and polarizability [19]. The set of 27 (9×3) descriptors formed the L block. Cross terms for adjacent positions (L12) and for every second position (L13) as well as a combination of them (L123) were included in the models to deal with the non-linearity. Cross term L123 accounts for every three adjacent positions in the peptide.

The polymorphic amino acids of the HLA-DRB1 proteins were also encoded using three z-scales. The set of 54 (18×3) descriptors formed the P block. The binding site has five pockets, corresponding to primary and secondary peptide anchor positions (Fig. 2). Positions 1 is a primary anchor while positions 4, 6, 7 and 9 are secondary. The binding pockets are named after the anchor position. Only polymorphic amino acids which interacted with peptide side chains were considered. They were as follows: pocket 1 – Val/Ala$^{85\beta}$ and Gly/Val$^{86\beta}$, pocket 4 – Phe/Ser/His/Tyr/Gly/Arg$^{13\beta}$, Leu/Tyr/Phe$^{26\beta}$, Glu/Asp/His$^{28\beta}$, Gln/AspArg$^{70\beta}$, Arg/Lys/Glu/Ala$^{71\beta}$, Ala/Arg/Gln/Leu/Glu$^{74\beta}$ and Tyr/Val$^{78\beta}$, pocket 6 – Leu/Ser/Val/GlyAsp/Pro$^{11\beta}$, Phe/Ser/His/Tyr/Gly/Arg$^{13\beta}$ and Glu/Asp/His$^{28\beta}$, pocket 7 – Glu/Asp/His$^{28\beta}$ and Cys/Tyr/Leu/Gly/His$^{30\beta}$, pocket 9 – Trp/Glu/Lys$^{9\beta}$, Cys/Tyr/Leu/Gly/His$^{30\beta}$, Val/Leu$^{38\beta}$ and Asp/Ser/Val$^{57\beta}$.

Cross terms for peptide/protein amino acid interactions in each pocket were included in the X matrix and formed the LP block of variables. The whole X matrix consisted of five blocks of descriptors: L, L12, L13, L123, P and LP. The proteochemometrics QSAR models derived here can be summarized as follows:

$$pIC_{50} = b + \sum(a_1 * \mathrm{L}) + \sum(a_2 * \mathrm{L12}) + \sum(a_3 * \mathrm{L13}) + \sum(a_4 * \mathrm{L123}) + \sum(a_5 * \mathrm{P}) + \sum(a_6 * \mathrm{LP}),$$

where a_n are PLS coefficients showing the contribution of each term to the binding affinity. Positive a_n mean favorite contribution of a term, while negative a_n point to non-favorite or even deleterious contributions. The models in this study are based on different combinations between the six blocks, as blocks L and P are present in all models. The models were derived by iterative self-consistent PLS based (ISC-PLS) algorithm.

2.3. QSAR by ISC-PLS algorithm

The training set of 2666 DRB1 binders was presented as a set of overlapping nonamers accompanied by the pIC_{50} values of the corresponding parent peptide. Only nonamers bearing anchor residue at position 1 (Tyr, Phe, Trp, Leu, Ile, Met, Val and Ala) were selected. The iterative self-consistent (ISC) algorithm [20,21] based

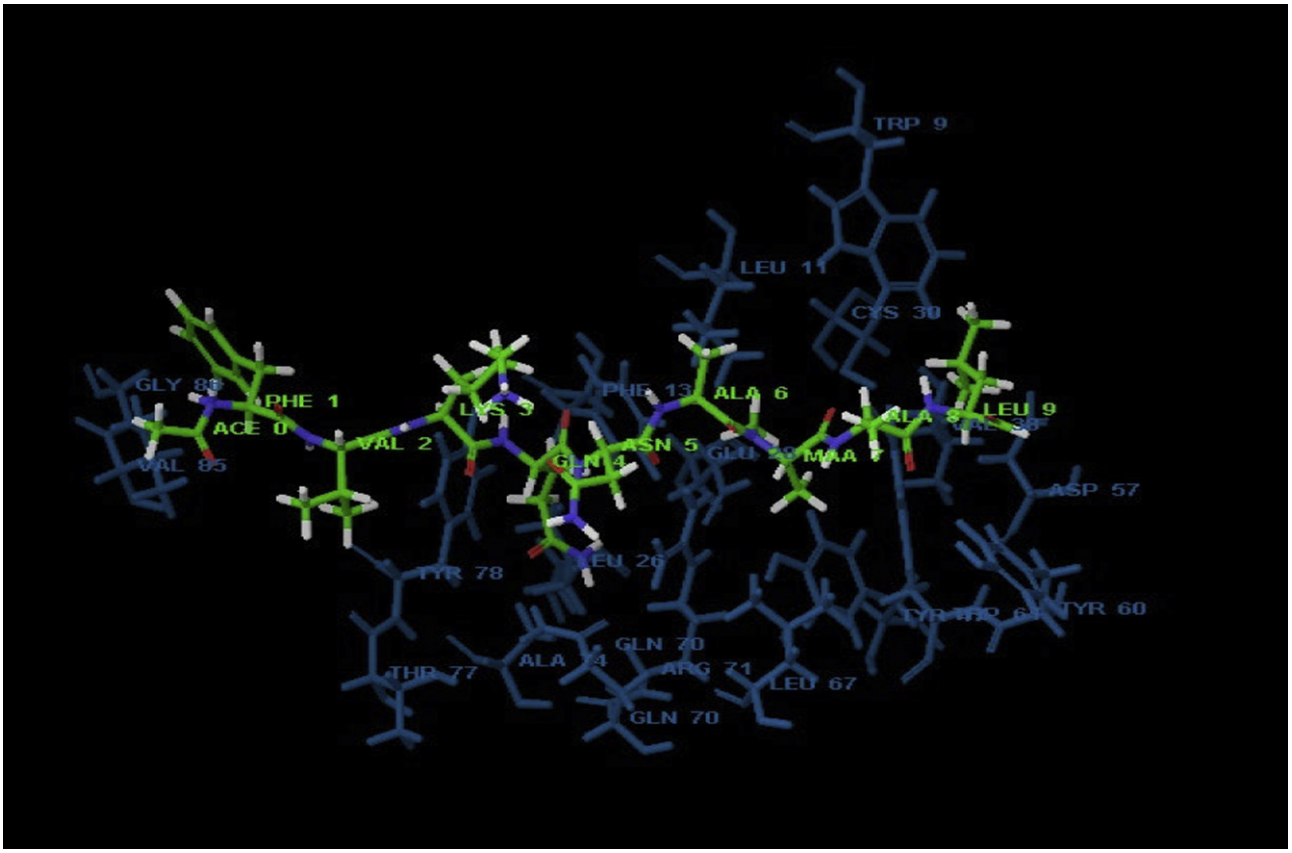

Fig. 2. Binding of peptide FVKQNA(MAA)AL to HLA-DR1 allele (pdb code: 1pyw). Peptide is given in green, protein is given in blue. Only polymorphic DRB1*0101 protein positions are labeled. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

on the partial least squares (PLS) method [22] was used to develop the proteochemometrics QSAR model. Briefly, the initial training set included all nonamers with anchors at position 1 ($n = 10670$). This was used to extract the first model. The optimum number of principal components (PCs) was derived by cross-validation in 7 groups. The first model was used to predict pIC_{50}s of the initial set and the best predicted nonamers from each parent peptide formed a second training set. This second set was used to extract the second QSAR model, which predicts pIC_{50}s of the initial training set. The best predicted nonamers from each parent peptide were selected and placed in a third training set. The selection procedure was repeated until the peptides in consecutive derived training sets were the same at the 99% level. The PLS method handles data matrices with more variables than observations very well, and this data can be both noisy and highly collinear. PLS forms new X variables (PC) as linear combinations of old variables, and then uses them to predict biological activity.

The models were assessed using r^2 (goodness of fit), q^2 (cross-validation in 7, 5 and 2 groups), and r^2_{pred} (external validation by test set). Q^2 values are mean of 200 runs. Simca-P 8.0 [23] was used to undertake PLS calculations.

2.4. Variable importance in projection (VIP)

VIP is the sum of the variable influence over all model dimensions and is a measure of variable importance [24]. High VIP values (VIP > 1.0) indicate good correlation between the variable and biological activity. Only the top 10 VIPs from each model were considered in the study.

3. Results

The proteochemometrics models derived in the present study are shown in Table 1. The models were developed by including different blocks of variables in the X matrix. These were assessed in terms of r^2, q^2 and r^2_{pred}. The top 10 important variables from each model are given in Table 2. The correlations between pIC_{50}(pred) and pIC_{50}(exp) for the test set are given in Fig. 3.

3.1. L + P model

The L + P model includes the L and P blocks of variables and explains 70% of the variance in the training set ($r^2 = 0.697$). The cross-validation in 7 groups gave $q^2 = 0.688$. Cross-validations in 5 and 2 groups gave q^2 value close to q^2_{CV7}. However, the correlation between the predicted and experimental pIC_{50} values of the external test set was poor with $r^2_{pred} = 0.364$ (Fig. 3A). Among the

Table 1
Proteochemometrics models assessed by r^2 (goodness of fit), q^2 (cross-validation in 7, 5 and 2 groups) and r^2_{pred} (external validation by test set).

Model	r^2	q^2_{CV7}	q^2_{CV5}	q^2_{CV2}	PC	r^2_{pred}
L + P	0.697	0.688	0.689	0.689	3	0.364
L + L12 + P	0.726	0.716	0.717	0.717	3	0.530
L + L12 + L13 + P	0.732	0.719	0.719	0.717	3	0.471
L + L123 + P	0.701	0.689	0.689	0.690	3	0.404
L + P + LP	0.691	0.686	0.686	0.686	2	0.369
AnchorL + P	0.685	0.678	0.678	0.667	3	0.431

Table 2
The top 10 most important variables for each model. VIP – variable importance in projection.

L + P model			L + P + L12 model			L+ P + L12 + L13 model		
Variable	VIP	Coefficient	Variable	VIP	Coefficient	Variable	VIP	Coefficient
z_3(L1)	3.4778	0.5326	z_1(L2)	3.6867	−0.1052	z_1(L3)	3.4068	−0.0933
z_2(L1)	2.7863	0.1663	z_1(L1)z_1(L2)	3.6340	0.0305	z_1(L1)z_1(L3)	3.2101	0.0251
z_1(P11)	1.4744	−0.0084	z_1(P11)	1.7192	−0.0081	z_3(L6)	2.4285	0.1331
z_2(P26)	1.4571	−0.0207	z_2(P26)	1.6931	−0.0199	z_1(P11)	1.8687	−0.0073
z_3(P30)	1.4233	0.0098	z_3(P30)	1.6482	0.0098	z_2(P26)	1.8332	−0.0194
z_1(L1)	1.3637	0.0803	z_3(P26)	1.4815	−0.0267	z_3(P30)	1.7885	0.0092
z_1(L2)	1.3375	−0.0455	z_1(P13)	1.4412	−0.0077	z_3(P26)	1.6140	−0.0228
z_3(L6)	1.3345	0.0953	z_2(P28)	1.3738	−0.0465	z_1(P13)	1.5677	−0.0075
z_3(L2)	1.3268	−0.0867	z_3(P11)	1.3644	−0.0014	z_2(P28)	1.4988	−0.0392
z_3(P26)	1.2934	−0.0244	z_1(P9)	1.3531	−0.0080	z_3(P11)	1.4981	0.0017
L + P + L123 model			L + P + LP model			AnchorL + P model		
Variable	VIP	Coefficient	Variable	VIP	Coefficient	Variable	VIP	Coefficient
z_3(L1)	2.5788	0.3717	z_1(L4)	3.0344	−0.0394	z_3(L6)	2.6363	0.2335
z_3(L6)	2.3680	0.1641	z_1(L4)z_1(P28)	3.0103	−0.0115	z_2(L4)	1.8335	−0.1208
z_2(L1)	2.2907	0.1267	z_1(L4)z_1(P26)	2.9849	0.0088	z_1(L4)	1.8074	−0.0753
z_1(L3)	1.7714	−0.0629	z_1(L4)z_1(P78)	2.8801	0.0247	z_2(L1)	1.6511	0.1151
z_1(P11)	1.6205	−0.0068	z_1(L4)z_1(P70)	2.8738	−0.0145	z_3(L1)	1.6059	0.2887
z_2(P26)	1.6053	−0.0172	z_1(L4)z_1(P71)	2.8108	−0.0133	z_2(L6)	1.4394	−0.0997
z_2(L4)	1.6030	−0.0827	z_2(P26)	1.8617	−0.0206	z_1(P11)	1.3175	−0.0062
z_3(P30)	1.5586	0.0098	z_1(P11)	1.8374	−0.0108	z_2(P26)	1.3115	−0.0151
z_3(P71)	1.4761	0.0449	z_3(P30)	1.7593	0.0108	z_3(P71)	1.2893	0.0442
z_3(P26)	1.4251	−0.0245	z_1(P13)	1.6009	−0.0050	z_3(P30)	1.2730	0.0094

top 10 most important variables are descriptors of ligand positions L1, L2 and L6 and protein positions P11, P26 and P30.

3.2. L + P + L12 model

The L12 block includes cross terms between adjacent ligand positions. It was added to the previous model improving r^2 and q^2 slightly and r^2_{pred} significantly (see Fig. 3B). The corresponding values are 0.726, 0.716 and 0.530. The most important variables in this model originate mainly from the HLA-DBR1 proteins: positions P9, P11, P13, P26, P28 and P30. The ligand is represented by position L2. The most important cross term is z_1(L1)z_1(L2).

3.3. L + P + L12 + L13 model

The L13 block includes cross terms between every second ligand positions. The addition of L13 block to the L + P + L12 model slightly improves r^2 and q^2_{CV7} (0.732 and 0.719, respectively) but reduces r^2_{pred} (0.471) (Fig. 3C). Thus, the L13 block brings more noise than signal into the model. Protein descriptors dominate this model: positions P11, P13, P26, P28 and P30. The most important ligand positions are L3 and L6 and the most important cross term is z_1(L1)z_1(L3).

3.4. L + P + L123 model

L123 block accounts for every three neighbor positions in the ligand. Adding it to L + P model slightly improves r^2, q^2_{CV7} and r^2_{pred} (0.701, 0.689 and 0.404, respectively) (Fig. 3D). The derived model is worse compared to the L + P + L12 and L + P + L12 + L13 models. The most important variables are ligand positions L1, L3, L4 and L6 and protein positions P11, P26, P30 and P71. There were no cross term among the top 10 VIPs.

3.5. L + P + LP model

The LP block contains ligand–protein cross terms accounting for interactions between side chains in the peptide and binding site pockets. Our initial expectation was that this model would outperform others in terms of goodness of fit and predictive ability. Surprisingly, the presence of the LP block in the L + P model does not improve r^2, q^2_{CV7} and r^2_{pred} (0.691, 0.686 and 0.369, respectively) (Fig. 3E). The interactions between ligand position L4 and the residues forming pocket 4 in the binding site (protein positions P26, P28, P70, P71 and P78) are among the ten most important variables. The ligand is represented by position L4 and the protein by positions P11, P13, P26 and P30.

3.6. AnchorL + P model

The L block in AnchorL + P model contains *z*-descriptors only for the anchor positions of the binding peptide. These positions are L1, L4, L6, L7 and L9. The model has slightly reduced r^2 and q^2_{CV7} (0.685 and 0.678, respectively) compared to the L + P model but significantly higher r^2_{pred} (0.431) (Fig. 3F). The top ten VIPs include ligand positions L1, L4 and L6 and protein positions P11, P26, P30 and P71.

4. Discussion

The MHC class II binding site is composed of an open-ended cleft formed by two antiparallel α-helices, each one belonging to either α or β protein chains. Peptides bind in an extended conformation deep in the binding cleft with the termini extending out of cleft at either end (Fig. 2). Only 9 residues are bound in the cleft, although typically class II binders are longer. About a dozen hydrogen bonds between conserved class II protein residues and peptide main chain carbonyl and amide groups are formed. The DR1 binding groove contains one deep pocket which accepts the side chain of the ligand primary anchor position L1 (pocket 1) and four shallow pockets corresponding to the side chains of the ligand secondary anchor positions L4, L6, L7 and L9 (pockets 4, 6, 7 and 9, respectively). In all DR alleles pocket 1 prefers hydrophobic aromatic or aliphatic amino acids, such as Tyr, Phe, Trp, Leu, Ile, Met and Val. The preferences of secondary pockets are more diverse and depend on the local protein structure.

To our knowledge, the present study is the first to apply a proteochemometrics approach to MHC binding prediction. The proteochemometrics models developed here are reliable and robust

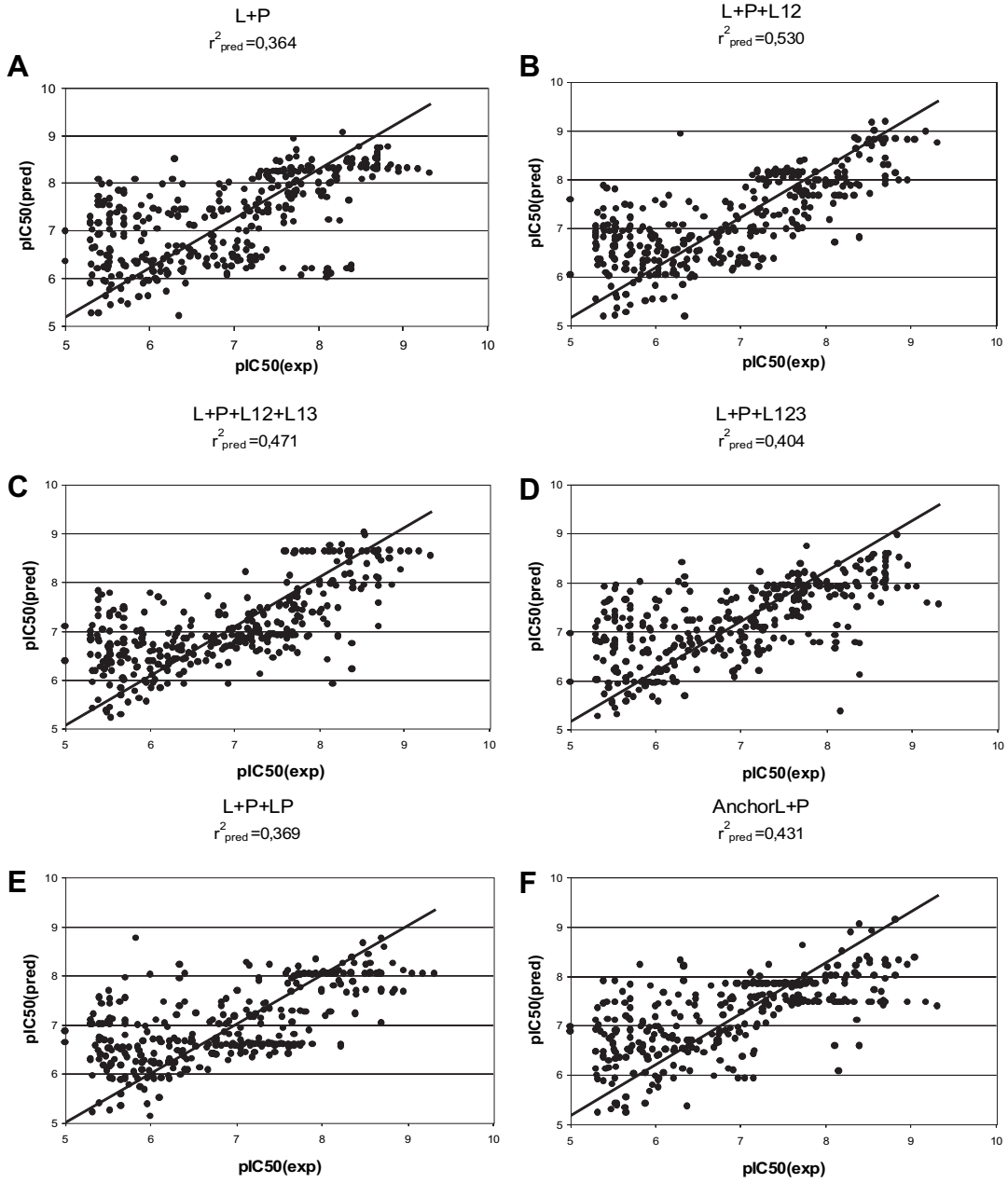

Fig. 3. Correlation between predicted and observed pIC_{50} values (r^2_{pred}) for the external test set ($n = 356$).

predictive tools for the accurate identification of peptides with a high affinity for MHC molecules. The *in silico* prediction of potential MHC binders from the sequence of a studied protein is the most critical step in the identification of immunogenic epitopes and the development of epitope-based vaccines [25]. The efficiency and success of subsequent experimental work is dependent on the accuracy of initial prediction. The epitope is the immunological quantum that constitutes the smallest entity recognized by the immune system. T-cell epitopes are peptides of varying lengths recognized as complexes with class I and class II MHC molecules. Recognition of these complexes is undertaken by the T-cell receptor expressed on the surface of T-cells. Epitope recognition leads to the activation of T-cells and the downstream immune response. This includes the formation of memory cells upon which a successful recall response upon subsequent infection depends. Thus the identification of peptides affine for MHC molecules and thus immunogenic T-cell epitopes is a necessary if not sufficient pre-requisite for the discovery and development of safe and efficacious vaccines.

In the present study the proteochemometrics approach was applied to 2666 peptides binding to 12 HLA-DRB1 alleles. Several models were derived based on different combinations of variable blocks describing both ligands and proteins (Table 1). Models have moderate goodness of fit, as expressed by r^2, ranging from 0.685 to 0.732. Their internal predictive ability, as expressed by q^2_{CV7}, was good, varying from 0.678 to 0.719. The cross-validations in 5 and 2 groups gave q^2 values close to those from cross-validation in 7 groups. The most important feature of a QSAR model is its ability to extrapolate, predicting activities or affinities of compounds not included in the training set. This external predictive ability usually is assessed by r^2_{pred}. Models had r^2_{pred} in the range 0.364–0.530. The highest r^2_{pred} value belongs to L + P + L12 model. This model was considered the most predictive.

The large difference between q^2 values and r^2_{pred} for some of the models suggests that either the model is overfitting the training set or the test set is substantially different from the training data. Using the same test set in all models shows that r^2_{pred} depends on the model and not the test set. Robust models work well on sets where poor models fail. As a rule of thumb, models with higher q^2 also have higher r^2_{pred}. It should also be mentioned that both the binding peptide and the binding site can exhibit extraordinary flexibility. Although the *z*-descriptors consider amino acid flexibility in a subtle manner, the overall peptide conformational changes and movements are not considered explicitly in the models. However, our previous studies showed conclusively that variations in peptide binding conformation do not affect significantly the predictability of the models [26].

Analysis of these models identifies the most important protein residues and peptide positions for accurate binding prediction (Table 2). Ligand position L1 is in the top ten most important variables in three of the six models (L + P, L + P + L123 and AnchorL + P). As the training set contains only nonamers starting with preferred hydrophobic anchor residues (Tyr, Phe, Trp, Leu, Ile, Met, Val and Ala), the models distinguish between them in terms of two other properties: steric bulk (z_2) and electronic properties (z_3). The positive coefficients for z_2 and z_3 mean that bulky and aromatic residues, like Tyr, Phe and Trp, are preferred to small aliphatic ones.

Although ligand position L2 is a non-anchor position, it is present among the VIPs in two models (L + P and L + P + L12). The negative coefficients for z_1 and z_3 in the L + P model suggests that hydrophobic and aliphatic amino acids, like Ile, Leu, Met and Val, are preferred at this position. The negative z_3 coefficient, solely presented in the L + P + L12 model, adds polar amino acids, like Lys, Gln, Arg and Thr, to the preferences. It was found that replacement of Lys at L2 with Ala greatly decreased the affinity of this peptide for DR1 molecule [18]. At the same time, the L2 side chain points out of the binding site and interacts with the T-cell receptor [18]. Together, these data indicate that L2 may play a dual role in the presentation of peptide by DRB1 proteins.

Similarly, ligand position L3 is generally considered not to be an anchor for MHC binding but rather interacts with the T-cell receptor [14]. The negative coefficients for z_1 in L + P + L12 + L13 and L + P + L123 models indicate a preference for hydrophobic residues at this position. X-ray data shows that although L3 Phe is a prominent, solvent exposed residue, it nestles in a hydrophobic shelf of the α-helix of HLA-DRB1*1501 [14].

Position L4 is a secondary anchor position for MHC class II binding. It is present in the top 10 VIPs in three models (L + P + L123, L + P + LP and AnchorL + P). The negative coefficients for z_1 and z_2 mean that small hydrophobic residues here should increase binding affinity. In fact, a great variety of amino acids are found at this position, and it is a key position for classifying DR molecules into supertypes [27]. Depending on the structural features of binding pocket 4, DR alleles accept either aromatic or aliphatic residues as well as negatively or positively charged ones [27].

Ligand position L6 is also a secondary anchor for peptide binding to DR molecules, and is a key position for DR classification. L6 is among the most important variables in four models (L + P, L + P + L12 + L13, L + P + L123 and AnchorL + P) represented mainly by positive coefficient for z_3 and once by a negative coefficient for z_2. Pocket 6 is a polar, negatively charged pocket dominated by protein position P28. A great variety of amino acids can bind here, apart from negatively charged Asp and Glu [14].

Positions L5, L7, L8 and L9 are not represented in the top 10 VIPs in any of the models, although L7 and L9 are secondary anchors for MHC class II binding.

The most important protein positions for the binding prediction are P9, P11, P13, P26, P28, P30 and P71 (Table 2). Position P9 is part of pocket 9. Among the DR alleles considered in this study polymorphism at this position includes Trp, Glu and Lys (Fig. 1). P9 is presented in L + P + L12 model by z_1. z_1 accounts for amino acid hydrophobicity and distinguishes between hydrophobic Trp and hydrophilic Glu and Lys. Position P11 appears in all models mainly through the contribution of its z_1 component. P11 determines the depth of polar pocket 6 [14]. Some of the proteins contain hydrophobic residues at this position, like Leu, Val and Pro, while others possess hydrophilic Ser and Asp. P13 is an extremely polymorphic protein position. It forms the wall between pockets 4 and 6. In L + P + L12, L + P + L12 + L13 and L + P + LP models it is represented by its z_1 component. Thus, the polymorphism is reduced to hydrophobic Phe and Tyr and hydrophilic Ser, His and Arg. P26 is part of pocket 4 and exists in all models through negative values of z_2 and z_3. z_2 and z_3 considered together distinguish aromatic and aliphatic residues. Three amino acids exist at this position: the aliphatic Leu and the aromatic Phe and Tyr. P28 appears only in two models, being represented by z_2 (L + P + L12 and L + P + L12 + L13). It is consistent with the negative charge of pocket 6, except in DRB1*0901 ($His^{28\beta}$). P30 is among the top 10 VIPs in all models, presented by its z_3 contribution. It lies between pockets 7 and 9. In terms of z_3, the high polymorphism is reduced to electron-rich Cys, His and Tyr and aliphatic Leu. Finally, P71, as represented by z_3, has a high VIP in two models (L + P + L123 and AnchorL + P). The polymorphism at P71 has important consequences in determining the available space for the side chain at peptide position L4 [14]. When positively charged Arg and Lys are here, negatively charged Asp and Glu are preferred at L4. *Vice versa*, if negatively charged Glu exists at P71, L4 preferences are for Arg and Lys [27].

The remaining polymorphic protein positions have only a low impact on binding prediction for DR alleles.

Among the most important variables are several cross terms (Table 2). The terms $z_1(L1)z_1(L2)$ (L + P + L12 model) and $z_1(L1)z_1(L3)$ (L + P + L12 + L13 model) have positive PLS coefficients. These results imply that hydrophobic residues (negative z_1 values) at positions L2 and L3 will increase the binding affinity, as the nonamers considered in the study start with hydrophobic amino acids (negative z_1 values at L1). This is in good agreement with the negative PLS coefficients for $z_1(L2)$ and $z_1(L3)$ discussed above.

A set of peptide–protein cross terms reflecting the interactions between ligand position L4 and protein positions P26, P28, P70, P71 and P78, forming pocket 4, has a great impact on prediction in the L + P + LP model (Table 2). As both ligand and protein residues are present as cross terms including z_1 values, the interactions are likely dominated by hydrophobicity. However, not all amino acids forming pocket 4 are hydrophobic. Residues at P28, P70 and P71 are polar or even charged (negatively or positively) and have positive z_1 values. This variety explains the different preferences for polar or non-polar amino acids observed at position L4.

Compared to five years ago [28], significant logistic advances have recently been made [29,30], yet problems still abound for immunoinformatic prediction. Recently, both structure- [31] and data-driven [32] prediction of antibody-mediated epitopes have been shown to be inadequate. Only for class I MHC peptide binding prediction are results seen to be both satisfactory and relatively accurate. However, several comparative studies indicate that class II T-cell epitope prediction in particular is typically poor and unreliable [33–35].

For many alleles, the creation of meaningful and useful test and training sets remains distinctly problematic. Properly designed training sets will address most such issues. Data diversity and data quality, as well as data quantity, are key issues. As diversity in

peptide sequence and affinity increases, so does the generality of generated models. Highly degenerate data or data with a very narrow affinity range often prove difficult. Predictive models can be tested using a complex array of techniques involving cross-validation, test sets, randomisation, and the rest. The optimal strategy for testing is the use of experimental validation involving the blind prediction and testing of new peptides.

For T-cell epitope prediction, the major issue remains the availability of data. Similarly, over 3500 different MHC alleles are known to exist in the global human population, indicating the extreme potential for distinct peptide specificities. The situation is exacerbated by the logistic constraints on sampling the specificity of even a single allele. For class II prediction, the inherently catholic nature of peptide specificity, as well as issues such as the effect of flanking residues and the possibility of alternative binding registers, combine to make the problem much more complex and intractable relative to class I prediction. In addressing these issues, two main approaches have been taken. One is the development of so-called supertypes [27,36] which seek to reduce the overall continuum of peptide binding into more discrete regions which exhibit clear commonalities of binding. The other is the development of pan-MHC methods [37,38], which seek to extrapolate beyond known data to generate more extensive binding predictions.

The proteochemometrics approach explored here utilizes both these ideas; it enlarges the available peptide space of binders and also builds the nature of the supertype directly into a synoptic pseudo-meta-analysis. It allows us to explore in some detail the interactions between peptide and protein residues, which has only previously been explored in those limited cases where structural data is available [39], but which can now be extended to any allele where binding data are available. This should allow us far greater insight into the quantitative contribution made by individual residues within separate alleles in determining peptide-binding specificity. More exciting still, proteochemometrics should lead to an appreciable increase in the robustness of predictions made across the group, as well as hopefully increasing the accuracy for particular alleles. The exploitation of this powerful technique is just beginning, and we expect to develop these ideas further in subsequent publications.

Acknowledgements

The research was supported by The National Science Fund of Ministry of Education and Science, Bulgaria (Grant 02-115/2008). DRF received salary support from a Senior Jenner Fellowship and the Wellcome Trust Grant WT079287MA; he is a Jenner Institute Investigator.

References

[1] D.R. Flower, Epitopes: the immunological quantum. in: D.R. Flower (Ed.), Bioinformatics for Vaccinology. Wiley–Blackwell, Chichester, UK, 2008, pp. 94–95.
[2] J. Robinson, M.J. Waller, P. Parham, N. de Groot, R. Bontrop, L.J. Kennedy, P. Stoehr, S.G.E. Marsh, IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. 31 (2003) 311–314.
[3] C.A. Janeway, P. Travers, M. Walport, J.D. Capra, Antigen recognition by T lymphocytes, in: Immunobiology: the immune system in health and disease. Current Biology Publications, London, 1999, pp. 115–162.
[4] M. Lapinsh, P. Prusis, A. Gutcaits, T. Lundstedt, J.E.S. Wikberg, Development of proteo-chemometrics: a novel technology for the analysis of drug–receptor interactions. Biochim. Biophys. Acta 1525 (2001) 180–190.
[5] M. Lapinsh, P. Prusis, S. Uhlén, J.E.S. Wikberg, Improved approach for proteochemometrics modeling: application to organic compound – amine G protein-coupled receptor interactions. Bioinformatics 21 (2005) 4289–4296.
[6] M. Lapinsh, S. Veiksina, S. Uhlén, R. Petrovska, I. Mutule, F. Mutulis, S. Yahorava, P. Prusis, J.E.S. Wikberg, Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes. Mol. Pharmacol. 67 (2005) 50–59.
[7] P. Prusis, S. Uhlén, R. Petrovska, M. Lapinsh, J.E.S. Wikberg, Prediction of indirect interactions in proteins. BMC Bioinformatics 7 (2006) 167.
[8] I. Mandrika, P. Prusis, S. Yahorava, M. Shikhagie, J.E.S. Wikberg, Proteochemometric modeling of antibody–antigen interactions using SPOT synthesized peptide arrays. Protein Eng. Des. Sel. 20 (2007) 301–307.
[9] A. Kontijevskis, P. Prusis, R. Petrovska, S. Yahorava, F. Mutulis, I. Mutule, J. Komorowski, J.E.S. Wikberg, A look inside HIV resistance through retroviral protease interaction maps. PLoS Comput. Biol. 3 (2007) 424–435.
[10] P. Prusis, M. Lapins, S. Yahorava, R. Petrovska, P. Niyomrattanakit, G. Katzenmeier, J.E.S. Wikberg, Proteochemometrics analysis of substrate interactions with denque virus NS3 proteases. Bioorg. Med. Chem. 16 (2008) 9369–9377.
[11] 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-A*0201 molecules. Cell 74 (1993) 929–937.
[12] A. Sette, J. Sidney, M.-F. del Guercio, S. Southwood, J. Ruppert, C. Dalberg, H.M. Grey, R.T. Kubo, Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. Mol. Immunol. 31 (1994) 813–822.
[13] C.P. Toseland, D.J. Taylor, H. McSparron, S.L. Hemsley, M.J. Blythe, K. Paine, I.A. Doytchinova, P. Guan, C.K. Hattotuwagama, D.R. Flower, AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical and cellular data. Immunome Res. 1 (2005) 4.
[14] K.J. Smith, J. Pyrdol, L. Gauthier, D.C. Wiley, K.W. Wucherpfennig, Crystal structure of HLA-DR2 (DRA*0101, DRB1*1501) complexed with a peptide from human myelin basic protein. J. Exp. Med. 188 (1998) 1511–1520.
[15] J. Hennecke, A. Carfi, D.C. Wiley, Structure of a covalently stabilized complex of a human αβ T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1. EMBO J. 19 (2000) 5611–5624.
[16] J. Hennecke, D.C. Wiley, Structure of a complex of the human α/β 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 (2002) 571–581.
[17] Z. Zavala-Ruiz, E.J. Sundberg, J.D. Stone, D.B. DeOliveira, I.C. Chan, J. Svendsen, R.A. Mariuzza, L.J. Stern, Human class II MHC protein HLA-DR1 bound to a designed peptide related to influenza virus hemagglutinin, FVKQNA(-MAA)AL, in complex with staphylococcal enterotoxin C3 variant 3B2 (SEC3-3B2). J. Biol. Chem. 278 (2003) 44904–44912.
[18] E.F. Resloniec, R.A. Ivey III, K.B. Whittington, A.H. Kang, H.-W. Park, Crystallographic structure of a rheumatoid arthritis MHC susceptibility allele, HLA-DR1 (DRB1*0101), complexed with the immunodominant determinant of human type II collagen. J. Immunol. 177 (2006) 3884–3892.
[19] S. Hellberg, M. Sjostrom, B. Skagerberg, S. Wold, Peptide quantitative structure–activity relationships, a multivariate approach. J. Med. Chem. 30 (1987) 1126–1135.
[20] I.A. Doytchinova, D.R. Flower, 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 (2003) 2263–2270.
[21] R.R. Mallios, An iterative approach to class II predictions. in: D.R. Flower (Ed.), Immunoinformatics: Predicting Immunogenicity in Silico, Methods in Molecular Biology, 409. Humana Press, 2007, pp. 341–353.
[22] S. Wold, PLS for multivariate linear modeling. in: H. van de Waterbeemd (Ed.), Chemometric methods in molecular design. VCH, Weinheim, Germany, 1995, pp. 195–218.
[23] Simca-P 8.0. Umetrics UK Ltd., Wokingham Road, RG42 1PL, Bracknell, UK.
[24] L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, S. Wold, Multi- and Megavariate Data Analysis. Umetrics AB, Umeå, Sweden, 2006, pp. 85–87.
[25] A. Sette, M. Newman, B. Livingston, D. McKinney, J. Sidney, G. Ishioka, S. Tangri, J. Alexander, J. Fikes, R. Chestnut, Optimizing vaccine design for cellular processing, MHC binding and TCR recognition. Tissue Antigens 59 (2002) 443–451.
[26] 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.
[27] I.A. Doytchinova, D.R. Flower, *In silico* identification of supertypes for class II MHCs. J. Immunol. 174 (2005) 7085–7095.
[28] D.R. Flower, Towards *in silico* prediction of immunogenic epitopes. Trends Immunol. 24 (2003) 667–674.
[29] B. Peters, H.H. Bui, S. Frankild, M. Nielson, C. Lundegaard, E. Kostem, D. Basch, K. Lamberth, M. Harndahl, W. Fleri, S.S. Wilson, J. Sidney, O. Lund, S. Buus, A. Sette, A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput. Biol. 2 (2006) e65.
[30] H.H. Lin, S. Ray, S. Tongchusak, E.L. Reinherz, V. Brusic, Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. BMC Immunol. 9 (2008) 8.
[31] J.V. Ponomarenko, P.E. Bourne, Antibody–protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct. Biol. 7 (2007) 64.
[32] M.J. Blythe, D.R. Flower, Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci. 14 (2005) 246–248.

[33] Y. El-Manzalawy, D. Dobbs, V. Honavar, On evaluating MHC-II binding peptide prediction methods. PLoS ONE 3 (2008) e3268.
[34] H.H. Lin, G.L. Zhang, S. Tongchusak, E.L. Reinherz, V. Brusic, Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. BMC Bioinformatics 9 (Suppl. 12) (2008) S22.
[35] U. Gowthaman, J.N. Agrewala, In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion. J. Proteome Res. 7 (2008) 154–163.
[36] I.A. Doytchinova, P. Guan, D.R. Flower, Identifying human MHC supertypes using bioinformatic methods. J. Immunol. 172 (2004) 4314–4323.
[37] H. Zhang, C. Lundegaard, M. Nielsen, Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. Bioinformatics 25 (2009) 83–89.
[38] M. Nielsen, C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, S. Buus, O. Lund, Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS Comput. Biol. 4 (2008) e1000107.
[39] M.N. Davies, C.K. Hattotuwagama, D.S. Moss, M.G. Drew, D.R. Flower, Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity. BMC Struct. Biol. 6 (2006) 5.