Scholarly Research Exchange

Volume 2008 • Article ID 360572 • doi:10.3814/2008/360572

Research Article

HomoSAR: An Integrated Approach Using Homology Modeling and Quantitative Structure-Activity Relationship for Activity Prediction of Peptides

Raghuvir R. S. Pissurlenkar and Evans C. Coutinho

Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz (E), Mumbai 400098, India

Correspondence should be addressed to Evans C. Coutinho, evans@bcpindia.org

Received 24 March 2008; Revised 23 May 2008; Accepted 12 August 2008

3D-QSAR of peptides is a daunting task. The difficulty in peptide QSAR arises due to the sheer number of conformational degrees of freedom for peptides that makes alignment in a 3D grid an overwhelming task. In this paper, we propose a method of QSAR where the alignment of peptides is shifted from 3D space to 1D space, making the alignment of peptides a very simple proposition. The method called *HomoSAR*, is based on an integrated approach that uses the principles of homology modeling in conjunction with the QSAR formalism to predict and design new peptide sequences. The peptides to be studied are subjected to a multiple sequence alignment which is followed by scoring every position in the peptide sequence against a reference peptide in the alignment, through calculation of similarity indices. The *similarity indices* obtained for each position (amino acid residue) in the peptide form the "descriptor" values (independent variables) which are then correlated to the biological activity of the peptide by G/PLS techniques. As an application, the methodology has been illustrated for the dataset of nonamer peptides that bind to the Class I major histocompatibility complex (MHC) molecule HLA-A*0201 as this dataset has been extensively studied. The models generated have statistically significant correlation coefficients and predictive r^2. The cross validated coefficients (q^2) are in an acceptable range. The *HomoSAR* approach identifies amino acids and properties that are preferred or detrimental at every position in the peptide sequence. The approach is simple to use and is able to extract all information contained in the dataset to explain the underlying structure activity relationships. The approach is applicable to peptide sequences which are not all of uniform length.

1. Introduction

Peptides and proteins form one of the important components of all biological systems. Peptides are nature's choice to maintain homeostasis and combat disease conditions and endowed them with high potency (low dose), specificity, and selectivity (reduced side effects). The peptide backbone is highly flexible, and the side chains of amino acids have the ability to adopt a conformation complementary to the active site of the receptor so as to match the bulk, hydrophobic, and electrostatic forces. For example, in the case of alpha-conotoxins, which are a class of nicotinic acetylcholine receptor (nAChR) antagonists, a single amino acid substitution in alpha-conotoxin PnIA shows a shift in the selectivity for the mammalian neuronal nicotinic acetylcholine receptor subtypes [1]. The flexibility of the backbone and the side chains also setup difficulties in the rational design of peptide drugs. The process of experimental lead optimization of peptides becomes an exponentially cumbersome procedure as the peptide length increases. For example, the design of a dipeptide would require an experimental scan through 20×20 combinations of amino acids to unravel the entire SAR of the dipeptide. Thus, the rational design of peptides is still a daunting task.

The QSAR techniques have the power and ability to quickly optimize a peptide sequence given a dataset of peptides with known biological activity. However, this is mostly restricted to the 2D-QSAR methods. There are very few examples in the literature that deal with the design and SAR of peptides by 3D-QSAR approaches, for example, the 3D-QSAR studies of DPP-IV dipeptides [2], MHC binding peptides [3–7], and recently, studies

on the δ opioid peptides—rubiscolins [8]. The 3D-QSAR techniques are not without their inherent problems. The difficulty of obtaining a unique alignment of peptides for 3D-QSAR analysis makes the application of CoMFA or CoMSIA techniques a daunting proposition. However, CoMFA and CoMSIA are the preferred methods for small molecule 3D-QSAR. The alignment procedure becomes more complex and uncertain as the degrees of freedom increase with increasing number of rotatable bonds. However, receptor-based alignment methods may provide a way out albeit at a high computational expense. It is for these reasons that 2D-QSAR techniques being much simpler and quicker than the 3D techniques are often used for studying peptides [9–24]. The power of 2D-QSAR results is dictated by the property spaced spanned.

In view of these short comings, we propose a method for QSAR of peptides where the central element of alignment is translated from 3D space to 1D space, which can be executed easily and accurately, while preserving the information content. The approach is called *HomoSAR* and is distantly related to the kernel-based approach of Salomon and Flower [23]. There are three basic steps in *HomoSAR*. The first step in *HomoSAR* is the *sequence alignment* of the peptides in the training and test sets separately as followed in the homology modeling or comparative protein modeling procedures. The multiple sequence alignment is the method of choice since it takes into account all the peptides in the dataset. The second step of the approach scores all the peptides of the dataset against a reference peptide in the alignment using *similarity indices*. The similarity indices calculated for each and every position in the peptide sequence is related to the binding activity in the third step by a suitable statistical algorithm. The three individual steps in HomoSAR are discussed in some details below.

The central step in this approach is the so-called *sequence alignment*. Sequence alignment is used for the detection of correspondences between amino acids of a reference peptide/protein and those of the query peptide/protein and can be related to the structure and activity of the peptides. The alignment of amino acid sequences is a crucial step in homology modeling due to which many different methods and programs have been published and are still being developed. The earliest attempt to clarify the structural similarity between protein sequences was by Needleman and Wunsch [25]. Variants of this algorithm have been developed independently by others and applied in many fields. ALIGN, BESTFIT, and GAP [26] are some of the computer-based programs which are being widely used for sequence alignment. The original Needleman and Wunsch algorithm was written to handle only a pair of sequences, whereas several other programs have been developed to handle multiple sequence alignment. Recent ones in this category are CLUSTALW [27], MAXHOM [28], and so forth. *HomoSAR* uses the multiple sequence alignment over the pairwise alignment due to the ability of the algorithm to handle multiple sequences and thus reduce the bias of a single reference.

The second step following sequence alignment is scoring or weighting the aligned sequences. In homology modeling, this is provided by the so-called *homology matrices* which makes use of the most probable amino acid substitutions according to the physical, chemical, or statistical properties. From the various available matrices [29–35], the following ones are frequently applied: *identity matrix, codon substitution matrix*, *mutation matrix* (Dayhoff or PAM 250 matrix), and *physical property matrices*. *HomoSAR* uses one of the above-mentioned scoring matrices for the multiple sequence alignment; however, for the QSAR analysis, *similarity indices* calculated from specific amino acid properties are used. These indices are calculated for every amino acid in the peptide sequence in relation to the amino acid in that position in a reference peptide, as identified by the alignment procedure.

The third step involves relating the similarity indices for the amino acids in the sequences with the biological activity through the use of a robust statistical method which is efficient enough to identify relationships with statistical significance. The G/PLS algorithm is the statistical method used in the third step of *HomoSAR*, which through its evolutionary nature is able to pick out descriptor variables that have the closest relation to the biological activity.

Homology modeling helps in identifying the similarity between different peptide/protein sequences on the basis of mutation, identity, or hydrophobic pattern of the sequences. This means that similar/related sequences will have similar structures and in turn similar function. It is well accepted that activity is related to structure, therefore variation in peptide sequences can be related to the variation in their activity distribution. Thus, the procedure of sequence alignment of peptides/proteins does establish a relationship between the activities and the sequences/structures, but is unable to quantify this relationship. On the other hand, QSAR which deals with the relationship of structure with activity establishes this relationship in a mathematical formulation. *HomoSAR* attempts to draw the strengths of homology modeling also called homology modeling, to overcome some of the limitations inherent in peptide QSAR approaches. Thus, a union of the principles of homology modeling and the QSAR formalism can establish a novel means of understanding in a quantitative fashion the variation in peptide sequence with activity. *HomoSAR* could also be used to address the difficulties of correlating both sequence diversity and variation in length with activity. The relationship between the length of the peptide chain and activity is not very obvious. An increase or decrease in peptide length often has a variable effect on the activity. For every biological effect, there is an optimum length of the peptide for which the activity is the highest, and deviation from this optimal length reflects directly on the activity. Thus identifying the optimum length for peptides is not always easy, though recognition of residues for affinity and activity may be somewhat simple.

If all the peptides binding to a given receptor are of uniform length, then the overlay of the peptides is straight forward if the active site permits a snug fit of the peptides. However, if the active site encloses a large space, then peptides with varying length could be translated in relation to each other so as to attain a tighter binding in the active

site. In such cases, a simple overlay of peptides cannot be used to impose the condition that the peptides share the same binding mode, but a good understanding of the binding mode can be gained through the sequence alignment technique in *HomoSAR*.

2. Computational Details

We demonstrate the *HomoSAR* methodology on a dataset of 128 nonameric peptides belonging to the HLA-0201*A series. This dataset was chosen simply because it is one of the established peptide dataset in terms of structural diversity, wide distribution of activity and has been well characterized both by theoretical and experimental studies [3–7, 16–23]. It is the best test bed for the validation of the *HomoSAR* methodology. The dataset was divided into a training set (87 molecules) and a test set (41 molecules) randomly on the basis of the activity values as shown in Table 1. For the present QSAR studies, the binding affinities of the peptides in the dataset were compiled from the literature [36–48] and transformed as $\mathrm{pIC}_{50}(-\log \mathrm{IC}_{50})$ values in terms of the molar concentration.

2.1. Multiple Sequence Alignment of the Peptides. The first step in *HomoSAR* involves an alignment of all the peptides in the dataset, shown in Figure 1. The alignment was executed using the *DNASIS Max* [49] sequence alignment software running on a Windows platform. The peptides sequences in both the training and test sets were aligned separately, aligned by the multiple sequence alignment strategy. The peptide *102* in the dataset was chosen as the reference peptide for scoring (*vide infra*) following the alignment step. The eight nonapeptides cocrystallized with the MHC protein and whose structures have been solved by X-ray crystallography (PDB codes are 1AKJ, 1DUZ, 1HHG, 1HHJ, 1OGA, 1QEW, 1QSE, and 1QSF) were also included in the peptide alignment, as a check against alignment results obtained by the multiple sequence alignment protocol.

2.2. Similarity Indices. Following alignment of the peptides in the dataset, the second step in *HomoSAR* involves calculating a similarity index for every amino acid position in the peptide sequence against the amino acid in the same position in the reference peptide (see Figure 1), as established by the alignment rule.

The similarity index (S) between peptide A (the reference peptide) and peptide B for "ith" position in the sequences, for a given physicochemical property, is given by

$$S^{\mathrm{AB}}_{[P]_i} = \frac{2 \times P^{\mathrm{A}}_i \times P^{\mathrm{B}}_i}{\left(P^{\mathrm{A}}_i\right)^2 + \left(P^{\mathrm{B}}_i\right)^2}, \tag{1}$$

where $S^{\mathrm{AB}}_{[P]_i}$ is the similarity between peptides A and B at the "ith" position in the peptide sequences for the physicochemical property P; P^{A}_i and P^{B}_i are the physicochemical property of the amino acid in the respective peptide sequences A and B at the "ith" position. The denominator is a normalizing factor.

2.3. Physicochemical Properties [P] for Computing Similarity Indices. The properties [P] used to calculate the similarity indices (S) (1) are the properties of amino acids such as *isotropic surface area (ISA), electronic charge index (ECI), hydrophobicity (HS), molar refractivity (MR), total dipole moment (TDM), and total lipole moment (TLM)*. The similarity values for the peptides (1) are used as the X-variables (descriptors) for derivation of the QSAR models. These properties were selected as they describe the steric, electronic, and hydrophobic nature of the amino acids that are key descriptors of the binding process. The significance of these properties used to calculate the similarity indices are discussed below.

2.3.1. Isotropic Surface Area (ISA). *Isotropic surface area* (ISA) is the portion of the solute molecule which is accessible for nonspecific interactions with water. The nonspecific interactions are those between water and solute molecules other than hydrogen bond interactions. The ISA is calculated as the sum of the surfaces over the side chain atoms accessible to nonspecific solvent interactions. Surfaces which interface the waters of hydration and the solute are excluded from the ISA [13]. Thus ISA provides a means to quantify hydrophobic nature of the solute molecules.

2.3.2. Electronic Charge Index (ECI). *Electronic charge index (ECI)* is the sum of the absolute value of the CNDO/2 charges of the side-chain atoms [13]. It is a measure of the local polarity at the amino acid side chain. A significant contribution of ECI to activity may indicate the presence of dipolar interactions of the side chain with the receptor site. It is calculated by the following formula:

$$\mathrm{ECI} = \sum |q_i|, \tag{2}$$

where q_i is the atomic charge of the ith atom in the amino acid side chain.

2.3.3. Valence Relative Chirality Index (^{v}RCI). *Valence relative chirality index* allows distinction between the R and S chiral isomers which the regular physicochemical properties cannot distinguish as reflected in the activity of the molecule. In the *relative chirality indices* [50] calculation, the three groups in descending priority attached to the chiral center are viewed from a reference point to calculate the new chirality metric. The groups/atoms a, b, c and d are then assigned valence delta value (δ^v) according to the method of Hall and Kier. The group delta value for any group (δ^v_i) attached to a chiral carbon is calculated as

$$\delta^v_i = \delta^v_{n1} + \left(\frac{\delta^v_{n2}}{2}\right) + \left(\frac{\delta^v_{n3}}{4}\right) + \left(\frac{\delta^v_{n4}}{8}\right) + \cdots + \left(\frac{\delta^v_{nN}}{2^{(N-1)}}\right), \tag{3}$$

where $n1$ is the atom attached directly to the chiral center (nearest neighbor), $n2$ is attached to $n1$, $n3$, to $n2$, and so on.

Table 1: HLA-A*0201 dataset (used for studying QSAR by the *HomoSAR* approach) with the experimental [pIC_{50}] and predicted affinity [pIC_{50}].

(a)

Training set							
Low affinity							
Sr. no.	Peptide	Expt pIC_{50}	Pred pIC_{50}	Sr. no.	Peptide	Expt pIC_{50}	Pred pIC_{50}
1	VALVGLFVL	5.15	5.23	9	SLHVGTQCA	5.84	6.10
2	VCMTVDSLV	5.15	5.31	11	SLNFMGYVI	5.88	6.26
3	HLESLFTAV	5.30	5.92	12	NLQSLTNLL	6.00	5.76
4	GTLVALVGL	5.34	5.45	13	FVTWHRYHL	6.03	6.71
5	LLSCLGCKI	5.45	5.53	15	QVMSLHNLV	6.17	6.54
6	LQTTIHDII	5.50	5.30	18	ALAKAAAAI	6.21	6.51
7	TLLVVMGTL	5.58	6.06				
Intermediate affinity							
19	GLGQVPLIV	6.30	5.99	39	TLGIVCPIC	6.82	6.59
20	MLDLQPETT	6.34	6.52	40	CLTSTVQLV	6.83	6.72
21	LLSSNLSWL	6.34	6.46	42	FLCKQYLNL	6.88	7.23
22	GLACHQLCA	6.38	6.20	43	FAFRDLCIV	6.89	6.55
23	LIGNESFAL	6.42	6.12	44	FLEPGPVTA	6.90	7.30
24	ALAKAAAAV	6.42	6.24	45	ALAKAAAAA	6.95	6.48
25	LLAVGATKV	6.48	6.79	46	LMAVVLASL	6.95	7.01
26	KLPQLCTEL	6.48	6.65	47	YVITTQHWL	6.98	7.52
27	ALAKAAAAL	6.51	6.54	48	LLCLIFLLV	7.00	7.03
28	WILRGTSFV	6.56	6.72	49	HLAVIGALL	7.00	6.28
29	IISCTCPTV	6.58	6.73	50	ITAQVPFSV	7.02	7.52
30	FLGGTPVCL	6.62	6.80	51	YLEPGPVTL	7.06	7.54
31	ALIHHNTHL	6.62	7.02	52	YTDQVPFSV	7.07	7.20
32	NLSWLSLDV	6.64	6.62	53	NLYVSLLLL	7.11	7.40
33	YMIMVKCWM	6.66	6.72	54	ILHNGAYSL	7.13	7.43
34	VLQAGFFLL	6.68	6.81	55	SIISAVVGI	7.16	7.39
35	GTLGIVCPI	6.71	6.28	56	VVMGTLVAL	7.17	7.75
36	VILGVLLLI	6.79	7.16	57	YLEPGPVTI	7.19	6.95
37	VTWHRYHLL	6.79	6.78	58	GLSRYVARL	7.25	6.96
38	PLLPIFFCL	6.80	6.74				
High affinity							
60	VLLDYQGML	7.33	7.10	82	MLGTHTMEV	7.85	7.27
61	YLEPGPVTV	7.34	7.25	83	LLFGYPVYV	7.89	8.02
63	YLSPGPVTA	7.38	7.22	84	ILKEPVHGV	7.92	7.51
65	IIDQVPFSV	7.40	7.45	85	YLMPGPVTV	7.93	8.04
66	SVYDFFVWL	7.44	7.53	86	WLDQVPFSV	7.94	7.49
67	ITWQVPFSV	7.46	8.01	87	KTWGQYWQV	7.96	7.91
68	ITYQVPFSV	7.48	8.19	89	YLAPGPVTA	8.03	7.82
69	GLYSSTVPV	7.48	7.34	90	YLYPGPVTV	8.05	7.61
71	LLLCLIFLL	7.59	7.24	91	LLMGTLGIV	8.10	7.25
73	VLIQRNPQL	7.64	7.49	92	YLWPGPVTV	8.13	7.50
74	SLYADSPSV	7.66	7.66	93	FLLTRILTI	8.15	8.24
75	RLLQETELV	7.68	7.24	94	GLLGWSPQA	8.24	8.21
76	ILSQVPFSV	7.70	7.57	95	ILYQVPFSV	8.31	8.22
77	IMDQVPFSV	7.72	7.60	96	GILTVILGV	8.35	8.06
78	QLFEDNYAL	7.76	7.31	99	YLFPGPVTA	8.50	8.05
79	ALMDKSLHV	7.77	7.19	101	ILFQVPFSV	8.70	8.35
80	YAIDLPVSV	7.80	7.85	102	ILWQVPFSV	8.77	8.62
81	FVWLHYYSV	7.82	8.00				

(b)

Test set							
Low affinity							
Sr. no.	Peptide	Expt pIC_{50}	Pred pIC_{50}	Sr. no.	Peptide	Expt pIC_{50}	Pred pIC_{50}
106	LTVILGVLL	5.58	5.91	110	TVILGVLLL	6.07	6.33
107	HLLVGSSGL	5.79	5.77	111	WTDQVPFSV	6.15	6.66
108	LLVVMGTLV	5.87	5.74	112	AIAKAAAAV	6.18	6.45
109	GIGILTVIL	6.00	5.61				
Intermediate affinity							
113	VLHSFTDAI	6.38	6.18	122	TLHEYMLDL	6.73	7.56
114	AAAKAAAAV	6.40	6.18	123	TLDSQVMSL	6.79	6.99
115	ILTVILGVL	6.42	6.20	124	HLYQGCQVV	6.83	6.53
116	MLLAVLYCL	6.48	6.83	125	QLFHLCLII	6.89	6.32
117	AVAKAAAAV	6.50	6.66	126	ITDQVPFSV	6.95	7.31
119	ILDEAYVMA	6.62	6.12	127	ALCRWGLLL	7.00	6.83
120	YLEPGPVTA	6.67	7.40	129	HLYSHPIIL	7.13	6.53
121	LLWFHISCL	6.68	6.56	131	FTDQVPFSV	7.21	7.57
High affinity							
132	YMLDLQPET	7.31	7.49	142	YLAPGPVTV	7.82	7.47
133	RLMKQDFSV	7.34	8.09	143	VVLGVVFGI	7.85	7.5
134	KLHLYSHPI	7.35	8.22	144	MMWYWGPSL	7.92	7.25
135	ITMQVPFSV	7.40	7.71	145	ILAQVPFSV	7.94	7.49
137	KIFGSLAFL	7.48	7.18	147	FLLSLGIHL	8.05	7.41
138	ALVGLFVLL	7.59	6.69	148	ILMQVPFSV	8.13	7.80
139	YLSPGPVTV	7.64	6.67	149	YLFPGPVTV	8.24	8.09
140	GLYSSTVPV	7.70	6.97	150	YLMPGPVTA	8.37	7.96
141	YLYPGPVTA	7.77	7.77	151	YLWPGPVTA	8.50	7.97

Peptide	P_{-4}	P_{-3}	P_{-2}	P_{-1}	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{+1}	P_{+2}
a102 [A]	–	–	–	–	I	L	W	Q	V	P	F	S	V	–	–
a001 [B]	–	–	–	–	–	–	V	A	L	V	G	L	F	V	L
$S^{AB}_{ISA_i}$	0	0	0	0	0	0	0.93	0.57	0.97	1	0.21	0.25	0.91	−0.68	−0.75
a102 [A]	–	–	–	–	I	L	W	Q	V	P	F	S	V	–	–
a021 [B]	L	L	S	S	N	L	S	W	L	–	–	–	–	–	–
$S^{AB}_{ISA_i}$	−0.75	−0.75	0	0	0.24	1	0.22	0.22	0.97	0	0	0	0	0	0
a102 [A]	–	–	–	–	I	L	W	Q	V	P	F	S	V	–	–
a152 [B]	–	–	–	–	F	L	D	Q	V	P	F	S	V	–	–
$S^{AB}_{ISA_i}$	0	0	0	0	0.97	1	0.2	1	1	1	1	1	1	0	0

Figure 1: A picture of the alignment of HLA-A*0201 peptides along with positionwise similarity indices calculated by (1). $S^{AB}_{ISA_i}$ similarity indices for the query peptide [B] aligned against the reference peptide [A] for positions P_{-4} to P_{+2}, the similarity indices have been calculated by (1) using the property—isotropic surface area (ISA).

The relative chirality indices ($^{\nu}$RCIs) for a pair of enantiomers are calculated as

$$
\begin{aligned}
{}^{\nu}\mathrm{RCI}_R &= \delta_a^{\nu} + \left(\delta_a^{\nu} + \delta_a^{\nu}\delta_b^{\nu}\right) + \left(\delta_a^{\nu} + \delta_a^{\nu}\delta_b^{\nu} + \delta_a^{\nu}\delta_b^{\nu}\delta_c^{\nu}\right) + \delta_a^{\nu}\delta_b^{\nu}\delta_c^{\nu}\delta_d^{\nu},\\
{}^{\nu}\mathrm{RCI}_S &= \delta_a^{\nu} + \left(\delta_a^{\nu} + \delta_a^{\nu}\delta_c^{\nu}\right) + \left(\delta_a^{\nu} + \delta_a^{\nu}\delta_c^{\nu} + \delta_a^{\nu}\delta_b^{\nu}\delta_c^{\nu}\right) + \delta_a^{\nu}\delta_b^{\nu}\delta_c^{\nu}\delta_d^{\nu}.
\end{aligned}
\tag{4}
$$

2.3.4. Hydrophobicity Scale (HS). The estimated hydrophobic effects [51] (kcal/mol) are values based on the contribution of the hydrophobic effect to the burial of each type of amino acid residue and side chain, obtained by analyzing the multitude of hydrophobicity scales. The scale estimates the free energy for transferring a residue from water to a

nonaqueous solvent, that is, the affinity of a residue for the solvent. The hydrophobic scale for the amino acid side chains is calculated as the difference between the estimated hydrophobic effect for the individual amino acid burial and that for glycine residue. It describes the thermodynamics of the partitioning of nonpolar compounds between water and a nonaqueous phase. This scale has been calculated to overcome the flaws of a set of previous hydrophobic scales which account for the partitioning of the amino acid residue between aqueous and organic solvents.

2.3.5. Total Dipole Moment (TDM). It is a partial charge-dependent parameter calculated on the basis of the center of charge over the substitution as the origin [52–55]. *Tsar3.3* [56] uses an empirical procedure called *Charge-2* for the rapid evaluation of partial atomic charges, which utilizes two fundamental chemical concepts; the inductive effect in saturated molecules and Hückel molecular orbital calculations for π systems. The total dipole moment along the amino acid side chain describes the electrostatic interaction at the receptor site. It is calculated as follows:

$$\vec{\mu} = e\sum_i \vec{r_i} q_i, \tag{5}$$

where $\vec{r_i}$ is the distance of the "ith" atom from the origin and "q_i" is the atomic charge of the "ith" atom.

2.3.6. Total Lipole Moment (TLM). The lipole of a molecule is a measure of the lipophilic distribution [57]. It is calculated from the sum of atomic log P values. This property has been calculated for the amino acid side chains using *Tsar3.3* [55]. It is calculated using

$$\vec{L} = \sum_i \vec{r_i} l_i, \tag{6}$$

where $\vec{r_i}$ is the distance of an "ith" atom from the origin and "l_i" is the atomic log P of the "ith" atom.

2.3.7. Molar Refractivity (MR). The molar refractive index of a molecule is a combined measure of its size and polarizability [57]. This fragment constant thermodynamic descriptor relates the effect of substituents on a reaction center from one type of process to another. The basic idea behind the use of such a descriptor is that similar changes in structure are likely to produce similar changes in reactivity, ionization, and binding. It can be experimentally determined or theoretically calculated using empirical rules. This property has been calculated using the method described by Vishwanadhan et al. as implemented in *Tsar3.3* [56]. It is calculated as

$$\mathrm{MR} = \left(\frac{(n^2-1)}{(n^2+2)}\right)\frac{(\mathrm{MW})}{d}, \tag{7}$$

where, "n" is the refractive index, "MW" molecular weight, and "d" is the density of the substituent group.

A few other similarity indices were derived from the above described properties. These new similarity indices were derived for "*dipeptide*" pairs, that is, neighboring amino acids (i, $i+1$) and denoted as $S^{\mathrm{AB}}_{[P]_{ij}}$; for "*tripeptide*" segments, that is, amino acids in a 1–3 relationship (i, $i+1$, $i+2$) and denoted as $S^{\mathrm{AB}}_{[P]_{ijk}}$, using one of the above mentioned properties $[P]$.

The similarity indices for peptides "A" and "B" for "*dipeptide*" pairs, *that is*, neighboring residues "i" and "$i+1$" are given by

$$S^{\mathrm{AB}}_{[P]_{ij}} = S^{\mathrm{AB}}_{[P]_i} \times S^{\mathrm{AB}}_{[P]_j}, \tag{8}$$

where $j = i+1$ and $S^{\mathrm{AB}}_{[P]_i}$, $S^{\mathrm{AB}}_{[P]_j}$, are similarity indices calculated using (1) for positions "i" and "j," using property $[P]$.

Likewise, the similarity between peptides "A" and "B" computed for "*tripeptide*" segments, that is, three successive amino acids "i," "j," and "k" is

$$S^{\mathrm{AB}}_{[P]_{ijk}} = S^{\mathrm{AB}}_{[P]_i} \times S^{\mathrm{AB}}_{[P]_j} \times S^{\mathrm{AB}}_{[P]_k}, \tag{9}$$

where $j = i+1$ and $k = i+2$ and $S^{\mathrm{AB}}_{[P]_i}$, $S^{\mathrm{AB}}_{[P]_j}$, $S^{\mathrm{AB}}_{[P]_k}$ are the similarity indices calculated using (1) for positions "i," "j," and "k" using property $[P]$.

Likewise, three other variables were calculated.

The total similarity between peptides A and B is given by

$$SS^{\mathrm{AB}}_{[P]} = \sum_i S^{\mathrm{AB}}_{[P]_i}, \tag{10}$$

where $S^{\mathrm{AB}}_{[P]_i}$ is the similarity index for position "i" in the two sequences according to (1). Likewise,

$$DS^{\mathrm{AB}}_{[P]} = \sum_{ij} S^{\mathrm{AB}}_{[P]_{ij}} \tag{11}$$

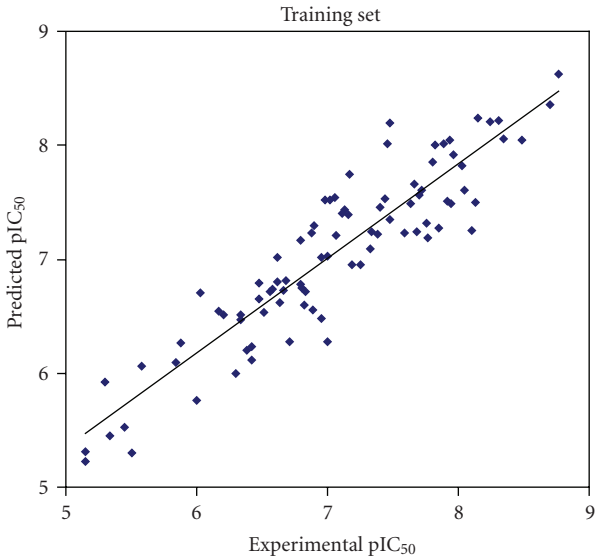
is the sum of the similarity indices for all dipeptides in the sequences A and B, as defined by (8).

Moreover,

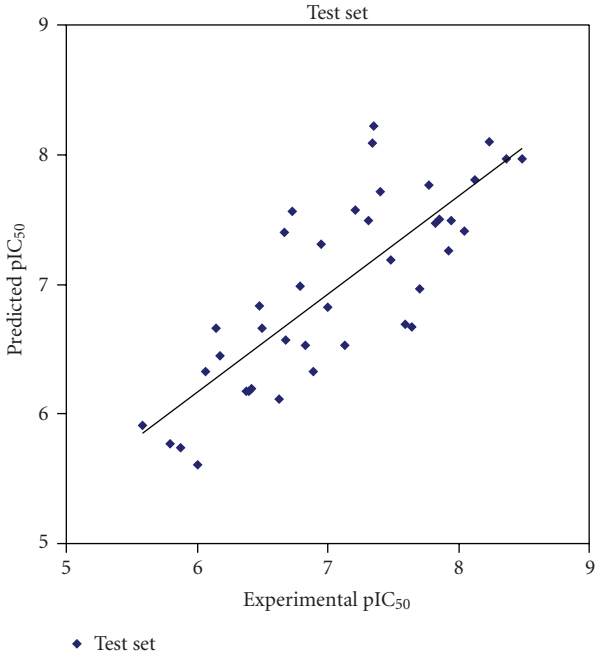
$$TS^{\mathrm{AB}}_{[P]} = \sum_{ijk} S^{\mathrm{AB}}_{[P]_{ijk}} \tag{12}$$

is the sum of the similarity indices for all tripeptides motifs in the sequences A and B, as defined by (9).

Every amino acid in the query sequence is assigned a similarity index (1) on the basis of a particular amino acid property $[P]$ against the amino acid at that particular position in the reference peptide (see Figure 1), as defined by the sequence alignment rule. When there is a gap in the alignment, that is, no amino acid can be matched in the query sequence, the position in the query is assigned a zero (0) value for the similarity index (see Figure 1), while in the situation where a gap occurs in the alignment, because no amino acid match occurs in the reference sequence but an amino acid is found in the query sequence, then this position in the query sequence is penalized with a negative value of its similarity index calculated against glycine. The matrix containing the similarity indices calculated for a particular property $[P]$ for all sequences in the training set



(a)



(b)

FIGURE 2: (a) Plot of experimental versus predicted activity for the training set. (b) Plot of experimental versus predicted activity for the test set.

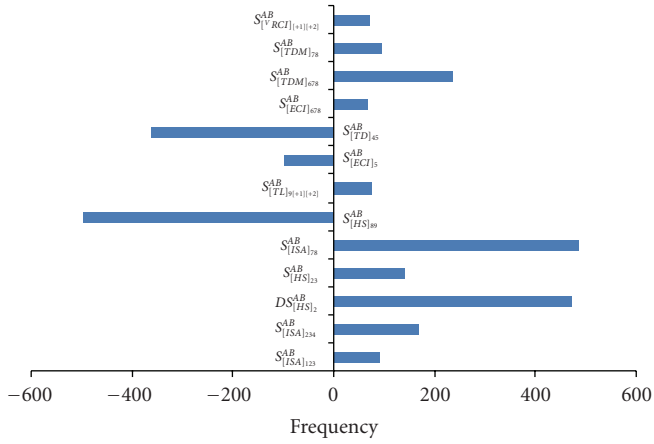
forms the X-variables in the QSAR table which is correlated with the biological activity (Y-variable). During the multiple sequence alignment, there are peptide sequences which translate to the right or left of the reference peptide. The amino acids in the query peptides which are aligned to the right of the first amino acid in the reference peptide are marked by additional position numbers with a negative sign while the amino acids in the query peptide which are aligned to the left of the last amino acid in the reference peptide are marked with positive position numbers, as seen in Figure 1.

2.4. QSAR Models and Statistics. The regression procedure, the third step in the *HomoSAR*, was carried out with the program—*Cerius2* (v4.11 Accelrys Inc., San Diego, Calif, USA) [58] running on a RedHat Linux Enterprise WS 4.0 workstation and on an SGI Fuel workstation (Silicon Graphics Inc., Calif, USA). Other modeling and computations were carried out using *InsightII* (v2005L Accelrys Inc., USA) [58] running on a RedHat Linux Enterprise WS4.0 workstation. All QSAR equations were generated with the genetic function approximation/partial least squares (G/PLS) method [59, 60] as implemented in *Cerius2*, with 10 000 generations, a population size of 500, a smoothness value (d) of 1.0, 6 PLS components, and no scaling of descriptors. The models were generated with equation lengths varying from 7 to 11. The rest of the parameters were set at their default values. The QSAR models were generated for similarity indices calculated for all properties described in Sections 2.3.1 to 2.3.7 collectively. The total X-variables (the similarity indices) numbered 315.

3. Results and Discussion

In a previous paper, we had reported the Hansch approach using specific properties of the amino acids as descriptors, and the Free-Wilson method to understand the SAR of HLA-A*0201 nonamer peptides [24]. The approaches were able to throw light on how the variation in amino acids at the nine positions influence the activity; but could not shed light on why minor similarities or dissimilarities in the peptide sequences cause large variation in the activity. The method also falls short in explaining whether all the peptides have the same binding pose within the MHC protein. It is not always true that peptides of the same length have the same binding mode. There is a possibility that one sequence may glide or translate in the binding pocket relative to the other amino acid sequence of the same length, thus affecting the binding affinity. Thus simply overlaying peptides in the active site (the atom-based alignment in CoMFA) may be insufficient in understanding peptide QSAR. *HomoSAR* is a QSAR technique that is based on homology modeling which is an efficient tool in identifying peptide/protein sequences that have a strong underlying relationship in terms of structure and function (activity). The method also uses similarity indices that are based on amino acid properties that reflect important binding attributes (electrostatic, steric, and hydrophobic) to score the peptide sequences aligned against the reference sequence.

The training set and the test sets were separately aligned against the 8 peptides whose crystal structures have been solved. The 8 peptides show perfect sequence alignment without any gaps; which is in harmony with their identical binding modes as seen in the X-ray structures. This places confidence in the alignment results for both the training and test set peptides.



(a)

$S^{AB}_{[^{V}RCI]_{[+1][+2]}}$	Is the relative valence chiral index over 1st and 2nd position dipeptide pair after the C-terminal
$S^{AB}_{[TDM]_{78}}$	Is the total dipole moment over 7th and 8th position dipeptide pair
$S^{AB}_{[TDM]_{678}}$	Is the total dipole moment over 6th, 7th and 8th position tripeptide segment
$S^{AB}_{[ECI]_{678}}$	Is the electronic charge index over 6th, 7th and 8th position tripeptide segment
$S^{AB}_{[TDM]_{45}}$	Is the total dipole moment over 4th and 5th position dipeptide pair
$S^{AB}_{[ECI]_{5}}$	Is the electronic charge index over 5th position
$S^{AB}_{[TLM]_{9[+1][+2]}}$	Is the total dipole moment over 9th, 1st and 2nd position after C-terminal
$S^{AB}_{[HS]_{89}}$	Is the hydrophobic scale over 8th and 9th dipeptide pair
$S^{AB}_{[ISA]_{78}}$	Is the isotropic surface area over 7th and 8th dipeptide pair
$S^{AB}_{[HS]_{23}}$	Is the hydrophobic scale over 2nd and 3rd dipeptide pair
$DS^{AB}_{[HS]}$	Is the hydrophobic scale over the entire peptide
$S^{AB}_{[ISA]_{234}}$	Is the isotropic surface area over 2nd, 3rd and 4th tripeptide segment
$S^{AB}_{[ISA]_{123}}$	Is the isotropic surface area over 1st, 2nd and 3rd tripeptide segment

(b)

Figure 3: Frequency of appearance of the physicochemical property associated at different positions in the sequence in the *HomoSAR* models.

The models derived by *HomoSAR* along with their statistical data are presented in Table 2. All models constructed are statistically significant. The models were internally validated using cross-validation by the leave-one-out (LOO) and leave-group-out (LGO) protocols and by boot strapping. The models were also tested for their predictive power on a test set. The predictive $r^2(r^2_{\text{pred}})$ for the models is given in Table 2. The plot of the experimental versus predicted binding affinities for the best model is given in Figure 2. The affinities predicted by the best *HomoSAR* model are given in Table 1. All the 500 equations were analyzed to identify the properties associated with each position in the peptide sequence that best accounts for the biological activity. The frequency of appearance of each property at the different positions in the peptide sequence in the QSAR equation is shown by the bar graph in Figure 3. The results of the QSAR models for the HLA-A*0201 dataset indicating the preferred nature and type of the amino acid at each position in the sequence are discussed below.

The term $DS^{\text{AB}}_{[\text{HS}]}$ appears with high frequency in the QSAR equations; it is the sum of the similarity indices for hydrophobicity of "*dipeptide*" pairs in the sequence, thus indicating the prevalence of hydrophobic character over the entire length of the peptide as a significant attribute for activity. This is perfectly in line with the nature of the binding cavity of the MHC protein [61]. The models also emphasize hydrophobic character for residues at the 2nd and the 3rd positions of the nonamer peptide. This is in complete harmony with all QSAR studies reported on this dataset [3, 19, 24]. Further, at position 4, a small increase in the hydrophobic nature is predicted to improve affinity of the peptide.

The models speak of the need to strike a balance for amino acids at positions 7 and 8; these should be residues with sufficient hydrophobic character as well as a capacity for dipolar interaction with the receptor. This is supported by the X-ray crystal structures of the HLA-A*0201 complexes, which show residues like tyrosine, tryptophan,

Table 2: *HomoSAR* models with the statistical data.

Sr. no.	*HomoSAR* models	n	No. of terms	PRESS	r^2	r^2_{CV} (LOO)	r^2_{CV} (LGO)	BSr^2	r^2_{random}	r^2_{pred}
(1)	$pIC_{50} = 6.83 + 1.88 * S^{AB}_{[ISA]_{78}} - 2.18 * S^{AB}_{[HS]_{89}} + 0.90 * S^{AB}_{[TDM]_{678}} - 0.615683 * S^{AB}_{[ECI]_5} + 1.85 * S^{AB}_{[HS]_{23}} - 0.42 * S^{AB}_{[TLM]_{12}}$	87	7	15.37	0.80	0.75	0.79	0.78	0.31	0.42
(2)	$pIC_{50} = 4.41 + 0.24 * DS^{AB}_{[HS]} + 1.91 * S^{AB}_{[TDM]_{9[+1][+2]}} + 1.20 * S^{AB}_{[HS]_{23}} - 2.36 * S^{AB}_{[HS]_{89}} + 1.44 * S^{AB}_{[ISA]_{78}} + 0.72 * S^{AB}_{[TDM]_{678}} - 0.63 * S^{AB}_{[ECI]_5}$	87	8	14.49	0.83	0.70	0.83	0.82	0.34	0.62
(3)	$4.94 - 2.55 * S^{AB}_{[HS]_{89}} + 1.49 * S^{AB}_{[HS]_{23}} + 0.59 * S^{AB}_{[TDM]_{678}} + 1.61 * S^{AB}_{[ISA]_{789}} + 0.17 * DS^{AB}_{[HS]} + 1.83 * S^{AB}_{[TLM]_{9[+1][+2]}} - 0.73 * S^{AB}_{[ECI]_5} + 0.47 * S^{AB}_{[TDM]_7}$	87	9	13.62	0.85	0.77	0.79	0.81	0.36	0.67
(4)	$4.93 - 2.43 * S^{AB}_{[HS]_{89}} - 0.67 * S^{AB}_{[ECI]_5} + 0.77 * S^{AB}_{[TDM]_{678}} + 1.49 * S^{AB}_{[HS]_{23}} + 0.20 * DS^{AB}_{[HS]} + 1.93 * S^{AB}_{[TLM]_{9[+1][+2]}} + 0.31 * S^{AB}_{[MR]_{[-1]12}} + 1.59 * S^{AB}_{[ISA]_{78}} - 0.23 * S^{AB}_{[TLM]_1}$	87	10	13.44	0.86	0.77	0.73	0.84	0.07	0.67
(5)	$4.63 + 1.04 * S^{AB}_{[TLM]_{9[+1][+2]}} + 1.75 * S^{AB}_{[ISA]_{789}} - 0.45 * S^{AB}_{[ECI]_{456}} - 0.35 * S^{AB}_{[TDM]_5} - 2.39 * S^{AB}_{[HS]_{89}} - 0.31 * S^{AB}_{[^{\nu}RCI]_{123}} + 0.22 * DS^{AB}_{[HS]} + 0.73 * S^{AB}_{[TDM]_{678}} + 1.51 * S^{AB}_{[HS]_{234}} + 0.80 * S^{AB}_{[MR]_{9[+1][+2]}}$	87	11	12.78	0.85	0.79	0.83	0.84	0.31	0.67

Table 3: Some of the newly designed peptides with their affinities for the HLA-A*0201 molecule as predicted by the best *HomoSAR* model.

Sr. no.	Peptide	Predicted pIC_{50}
(1)	RLWDRPPTV	9.26
(2)	RLYRRASTV	7.95
(3)	YLWDFPPEV	8.46
(4)	RLMTFFPSV	7.85
(5)	WIWQVPFRV	9.19

and phenylalanine at these positions making dipolar contacts in the binding pocket.

The term $S^{AB}_{[ECI]_{456}}$—the electronic similarity index for the "*tripeptide*" segment spanning positions 4, 5, and 6—emerges with a negative frequency. This means that the electronic character of the amino acids at the three positions 4, 5, and 6 needs to be lowered to an optimal level to enhance binding; this is more so for the 5th position in the sequence. This insight into the requirements for positions 4, 5, and 6 was not revealed in the "*descriptor-based QSAR*" study [24], but the observations are in line with earlier papers [3, 19].

It is appealing to note from the terms appearing in the *HomoSAR* models that there needs to be a considerable increase in the electronic property of the amino acid occupying positions 6 and 7, while maintaining sufficient hydrophobic character at these positions. This requirement is in agreement with the "*binary QSAR*" approach [24].

There is a titular appearance of the similarity terms for the extended positions 10 and 11 (see Figure 1) at the C-terminal end of the peptide. These terms show that an increase in the chain length at the C-terminal end is possible; however there can be no extension at the N-terminal end. This is in accordance with the fact that decapeptides do show decent levels of biological activity [61]. The standard QSAR methods are unable to extract this information about the peptide length and activity.

The analysis of the *HomoSAR* models has led to the design of some new peptides with affinity higher than the peptides listed in Table 1. The peptide sequences with their predicted affinities are given in Table 3.

4. Conclusions

The complexity in peptide design by 3D-QSAR methods arises because of several variables: first, the large number of degrees of freedom that makes secondary structure determination difficult. Second, as the peptide length increases from two to ten, the probability of arriving at the optimal alignment is very remote. The problem aggravates when peptides of varying length have the same level of activity. For this reason, while 2D/3D QSAR has been very successful in the design and discovery of small molecules, the successful applications in peptides are far and few in between. The *HomoSAR* approach is an attempt to solve the problem of peptide QSAR by primarily moving the crucial step of alignment in 3D-QSAR from 3D space to the less complex 1D space. This has been achieved by adopting the principles of homology modeling into the QSAR formalism. As an application to the MHC class of peptides, the technique was able to extract all known SARs reported for their

class as well as reveal a few that were hitherto unknown. The *HomoSAR* approach is also able to give an idea of the relative binding mode the query peptides can have in relation to the reference peptide. Thus, this technique can be gainfully employed to understand and optimize the relationship between activity and the position and nature of amino acids in any peptide sequence, without resorting to the cumbersome 3D spatial analysis. In conclusion, *HomoSAR* as a union of homology modeling and QSAR principles is a useful tool in the medicinal chemists' armamentarium to design peptide ligands.

Acknowledgments

The All India Council for Technical Education (AICTE, New Delhi) is acknowledged for support in developing the *HomoSAR* approach through Grant F. no. 8022/RID/NPRDJ/RPS-5/2003-04, and Department of Sciences and Technology (DST, New Delhi) is also thanked for providing some of the computational facilities under its FIST Program (SR/FST/LSI-163/2003).

References

[1] M. Loughnan, T. Bond, A. Atkins, et al., "α-conotoxin EpI, a novel sulfated peptide from *Conus episcopatus* that selectively targets neuronal nicotinic acetylcholine receptors," *The Journal of Biological Chemistry*, vol. 273, no. 25, pp. 15667–15674, 1998.

[2] W. Brandt, T. Lehmann, A. Barth, and S. Fittkaui, "Molecular modeling and CoMFA investigations of the serine proteases thermitase and dipeptidyl peptidase IV and their inhibitors," *Journal of Molecular Graphics*, vol. 11, pp. 277–278, 1993.

[3] I. A. Doytchinova and 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," *Journal of Medicinal Chemistry*, vol. 44, no. 22, pp. 3572–3581, 2001.

[4] I. A. Doytchinova and D. R. Flower, "Quantitative approaches to computational vaccinology," *Immunology and Cell Biology*, vol. 80, no. 3, pp. 270–279, 2002.

[5] I. A. Doytchinova and D. R. Flower, "Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three-dimensional quantitative structure-activity relationship study," *Proteins: Structure, Function, and Bioinformatics*, vol. 48, no. 3, pp. 505–518, 2002.

[6] I. A. Doytchinova and D. R. Flower, "A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif," *Journal of Computer-Aided Molecular Design*, vol. 16, no. 8-9, pp. 535–544, 2002.

[7] M. N. Davies, C. K. Hattotuwagama, D. S. Moss, M. G. B. Drew, and D. R. Flower, "Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity," *BMC Structural Biology*, vol. 6, article 5, pp. 1–13, 2006.

[8] J. Caballero, M. Saavedra, M. Fernández, and F. D. González-Nilo, "Quantitative structure-activity relationship of rubiscolin analogues as δ opioid peptides using comparative-molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA)," *Journal of Agricultural and Food Chemistry*, vol. 55, no. 20, pp. 8101–8104, 2007.

[9] P. H. A. Sneath, "Relations between chemical structure and biological activity in peptides," *Journal of Theoretical Biology*, vol. 12, no. 2, pp. 157–195, 1966.

[10] A. Kidera, Y. Konishi, M. Oka, T. Ooi, and H. A. Scheraga, "Statistical analysis of the physical properties of the 20 naturally occurring amino acids," *Journal of Protein Chemistry*, vol. 4, no. 1, pp. 23–55, 1985.

[11] S. Hellberg, M. Sj"ostr"om, B. Skagerberg, and S. Wold, "Peptide quantitative structure-activity relationships, a multivariate approach," *Journal of Medicinal Chemistry*, vol. 30, no. 7, pp. 1126–1135, 1987.

[12] M. Cocchi and E. Johansson, "Amino acids characterization by GRID and multivariate data analysis," *Quantitative Structure-Activity Relationships*, vol. 12, no. 1, pp. 1–8, 1993.

[13] E. R. Collantes and W. J. Dunn III, "Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues," *Journal of Medicinal Chemistry*, vol. 38, no. 14, pp. 2705–2713, 1995.

[14] A. Zaliani and E. Gancia, "MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 3, pp. 525–533, 1999.

[15] N. El Tayar, R.-S. Tsai, P.-A. Carrupt, and B. Testa, "Octan-1-ol-water partition coefficients of zwitterionic α-amino acids. Determination by centrifugal partition chromatography and factorization into steric/hydrophobic and polar components," *Journal of the Chemical Society, Perkin Transactions 2*, no. 1, pp. 79–84, 1992.

[16] I. A. Doytchinova and 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*, vol. 19, no. 17, pp. 2263–2270, 2003.

[17] I. A. Doytchinova, M. J. Blythe, and D. R. Flower, "Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201," *Journal of Proteome Research*, vol. 1, no. 3, pp. 263–272, 2002.

[18] P. Guan, I. A. Doytchinova, and D. R. Flower, "HLA-A3 supermotif defined by quantitative structure-activity relationship analysis," *Protein Engineering*, vol. 16, no. 1, pp. 11–18, 2003.

[19] I. A. Doytchinova, P. Guan, and D. R. Flower, "Quantitative structure-activity relationships and the prediction of MHC supermotifs," *Methods*, vol. 34, no. 4, pp. 444–453, 2004.

[20] D. R. Flower, H. McSparron, M. J. Blythe, et al., "Computational vaccinology: quantitative approaches," *Novartis Foundation Symposium*, vol. 254, pp. 102–120, 2003.

[21] P. Guan, I. A. Doytchinova, V. A. Walshe, P. Borrow, and D. R. Flower, "Analysis of peptide-protein binding using amino acid descriptors: prediction and experimental verification for human histocompatibility complex HLA-A*0201," *Journal of Medicinal Chemistry*, vol. 48, no. 23, pp. 7418–7425, 2005.

[22] I. A. Doytchinova, V. A. Walshe, P. Borrow, and D. R. Flower, "Towards the chemometric dissection of peptide-HLA-A*0201 binding affinity: comparison of local and global QSAR models," *Journal of Computer-Aided Molecular Design*, vol. 19, no. 3, pp. 203–212, 2005.

[23] J. Salomon and D. R. Flower, "Predicting class II MHC-peptide binding: a kernel based approach using similarity scores," *BMC Bioinformatics*, vol. 7, article 501, pp. 1–11, 2006.

[24] R. R. S. Pissurlenkar, A. K. Malde, S. A. Khedkar, and E. C. Coutinho, "Encoding type and position in peptide QSAR: application to peptides binding to class I MHC molecule HLA-A*0201," *QSAR and Combinatorial Science*, vol. 26, no. 2, pp. 189–203, 2007.

[25] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[26] J. Devereux, P. Haeberli, and O. Smithies, "A comprehensive set of sequence analysis programs for the VAX," *Nucleic Acids Research*, vol. 12, no. 1, part 1, pp. 387–395, 1984.

[27] C. Sander and R. Schneider, "Database of homology-derived protein structures and the structural meaning of sequence alignment," *Proteins: Structure, Function, and Bioinformatics*, vol. 9, no. 1, pp. 56–68, 1991.

[28] G. D. Schuler, S. F. Altschul, and D. J. Lipman, "A workbench for multiple alignment construction and analysis," *Proteins: Structure, Function, and Bioinformatics*, vol. 9, no. 3, pp. 180–190, 1991.

[29] M. Vingron and P. Argos, "A fast and sensitive multiple sequence alignment algorithm," *Computer Applications in the Biosciences*, vol. 5, no. 2, pp. 115–121, 1989.

[30] D. R. Boswell and A. D. McLachlan, "Sequence comparison by exponentially-damped alignment," *Nucleic Acids Research*, vol. 12, no. 1, part 2, pp. 457–464, 1984.

[31] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed., vol. 5, supplement 3, pp. 345–352, National Biomedical Research Foundation, Washington, DC, USA, 1978.

[32] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: detection of distantly related proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 13, pp. 4355–4358, 1987.

[33] J. L. Risler, M. O. Delorme, H. Delacroix, and A. Henaut, "Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix," *Journal of Molecular Biology*, vol. 204, no. 4, pp. 1019–1029, 1988.

[34] D. M. Engelman, T. A. Steitz, and A. Goldman, "Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins," *Annual Review of Biophysics and Biophysical Chemistry*, vol. 15, pp. 321–353, 1986.

[35] G. H. Gonnet, M. A. Cohen, and S. A. Benner, "Exhaustive matching of the entire protein sequence database," *Science*, vol. 256, no. 5062, pp. 1443–1445, 1992.

[36] Y. Rongcun, F. Salazar-Onfray, J. Charo, et al., "Identification of new HER2/neu-derived peptide epitopes that can elicit specific CTL against autologous and allogeneic carcinomas and melanomas," *The Journal of Immunology*, vol. 163, no. 2, pp. 1037–1044, 1999.

[37] L. Rivoltini, Y. Kawakami, K. Sakaguchi, et al., "Induction of tumor-reactive CTL from peripheral blood and tumor-infiltrating lymphocytes of melanoma patients by in vitro stimulation with an immunodominant peptide of the human melanoma antigen MART-1," *The Journal of Immunology*, vol. 154, no. 5, pp. 2257–2265, 1995.

[38] M. R. Parkhurst, E. B. Fitzgerald, S. Southwood, A. Sette, S. A. Rosenberg, and Y. Kawakami, "Identification of a shared HLA-A*0201-restricted T-cell epitope from the melanoma antigen tyrosinase-related protein 2 (TRP2)," *Cancer Research*, vol. 58, no. 21, pp. 4895–4901, 1998.

[39] W. M. Kast, R. M. P. Brandt, J. Sidney, et al., "Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins," *The Journal of Immunology*, vol. 152, no. 8, pp. 3904–3912, 1994.

[40] A. Sette, A. Vitiello, B. Reherman, et al., "The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes," *The Journal of Immunology*, vol. 153, no. 12, pp. 5586–5592, 1994.

[41] M. R. Parkhurst, M. L. Salgaller, S. Southwood, et al., "Improved induction of melanoma-reactive CTL with peptides from the melanoma antigen gp100 modified at HLA-A*0201-binding residues," *The Journal of Immunology*, vol. 157, no. 6, pp. 2539–2548, 1996.

[42] A. Vitiello, A. Sette, L. Yuan, et al., "Comparison of cytotoxic T lymphocyte responses induced by peptide or DNA immunization: implications on immunogenicity and immunodominance," *European Journal of Immunology*, vol. 27, no. 3, pp. 671–678, 1997.

[43] M.-F. del Guercio, J. Sidney, G. Hermanson, et al., "Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype," *The Journal of Immunology*, vol. 154, no. 2, pp. 685–693, 1995.

[44] V. Tsai, S. Southwood, J. Sidney, et al., "Identification of subdominant CTL epitopes of the gp100 melanoma-associated tumor antigen by primary in vitro immunization with peptide-pulsed dendritic cells," *The Journal of Immunology*, vol. 158, no. 4, pp. 1796–1802, 1997.

[45] Y. Kawakami, S. Eliyahu, C. Jennings, et al., "Recognition of multiple epitopes in the human melanoma antigen gp100 by tumor-infiltrating T lymphocytes associated with in vivo tumor regression," *The Journal of Immunology*, vol. 154, no. 8, pp. 3961–3968, 1995.

[46] T. S. Jardetzky, W. S. Lane, R. A. Robinson, D. R. Madden, and D. C. Wiley, "Identification of self peptides bound to purified HLA-B27," *Nature*, vol. 353, no. 6342, pp. 326–329, 1991.

[47] A. Y. Rudensky, P. Preston-Hurlburt, S.-C. Hong, A. Barlow, and C. A. Janeway Jr., "Sequence analysis of peptides bound to MHC class II molecules," *Nature*, vol. 353, no. 6345, pp. 622–627, 1991.

[48] K. C. Parker, M. A. Bednarek, and J. E. Coligan, "Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains," *The Journal of Immunology*, vol. 152, no. 1, pp. 163–175, 1994.

[49] *DNasisMax2.7*, Hitachi Software Engineering America, Ltd. MiraiBio Group, South San Francisco, Calif, USA, 2007.

[50] R. Natarajan, S. C. Basak, and T. S. Neumann, "Novel approach for the numerical characterization of molecular chirality," *Journal of Chemical Information and Modeling*, vol. 47, no. 3, pp. 771–775, 2007.

[51] P. A. Karplus, "Hydrophobicity regained," *Protein Science*, vol. 6, no. 6, pp. 1302–1307, 1997.

[52] R. J. Abraham and G. H. Grant, "Charge calculations in molecular mechanics. V. Silicon compounds and π bonding," *Journal of Computational Chemistry*, vol. 9, no. 3, pp. 244–256, 1988.

[53] R. J. Abraham and P. E. Smith, "Charge calculations in molecular mechanics IV: a general method for conjugated systems," *Journal of Computational Chemistry*, vol. 9, no. 4, pp. 288–297, 1988.

[54] R. J. Abraham and P. E. Smith, "Charge calculations in molecular mechanics 7: application to polar π systems incorporating nitro, cyano, amino, C=S and thio substituents," *Journal of Computer-Aided Molecular Design*, vol. 3, no. 2, pp. 175–187, 1989.

[55] R. J. Abraham, G. H. Grant, I. S. Haworth, and P. E. Smith, "Charge calculations in molecular mechanics. Part 8

partial atomic charges from classical calculations," *Journal of Computer-Aided Molecular Design*, vol. 5, no. 1, pp. 21–39, 1991.

[56] *Tsar3.3*, Oxford Molecular Ltd., Oxford Science Park, Oxford, UK, 2002.

[57] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, and R. K. Robins, "Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics," *Journal of Chemical Information and Computer Sciences*, vol. 29, pp. 163–172, 1989.

[58] *Cerius2 4.11 and InsightII, v2005L*, Accelrys Inc., San Diego, Calif, USA, 2005.

[59] W. J. Dunn III, S. Wold, U. Edlund, S. Hellberg, and J. Gasteiger, "Multivariate structure-activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: the PLS method," *Quantitative Structure-Activity Relationships*, vol. 3, no. 4, pp. 131–137, 1984.

[60] R. D. Cramer III, J. D. Bunce, D. E. Patterson, and I. E. Frank, "Cross-validation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies," *Quantitative Structure-Activity Relationships*, vol. 7, pp. 18–25, 1988.

[61] H.-D. Holtje, W. Sippl, D. Rognan, and G. Folkers, "Example for the modeling of protein-ligand complexes: Antigen presentation by MHC class I," in *MolecularModeling: Basic Principles and Applications*, H.-D. Holtje and G. Folkers, Eds., pp. 179–215, WILEY-VCH GmbH & Co. KGaA, Weinheim, Germany, 2nd edition, 2003.