Side-chain conformational space analysis (SCSA): A multi conformation-based QSAR approach for modeling and prediction of protein–peptide binding affinities

Peng Zhou · Xiang Chen · Zhicai Shang

Received: 19 March 2008 / Accepted: 11 September 2008 / Published online: 8 October 2008

Abstract In this article, the concept of multi conformation-based quantitative structure–activity relationship (MCB-QSAR) is proposed, and based upon that, we describe a new approach called the side-chain conformational space analysis (SCSA) to model and predict protein–peptide binding affinities. In SCSA, multi-conformations (rather than traditional single-conformation) have received much attention, and the statistical average information on multi-conformations of side chains is determined using self-consistent mean field theory based upon side chain rotamer library. Thereby, enthalpy contributions (including electrostatic, steric, hydrophobic interaction and hydrogen bond) and conformational entropy effects to the binding are investigated in terms of occurrence probability of residue rotamers. Then, SCSA was applied into the dataset of 419 HLA-A*0201 binding peptides, and nonbonding contributions of each position in peptide ligands are well determined. For the peptides, the hydrogen bond and electrostatic interactions of the two ends are essential to the binding specificity, van der Waals and hydrophobic interactions of all the positions ensure strong binding affinity, and the loss of conformational entropy at anchor positions partially counteracts other favorable nonbonding effects.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9245-0) contains supplementary material, which is available to authorized users.

P. Zhou · X. Chen · Z. Shang (✉)
Institute of Molecular Design and Molecular Thermodynamics, Department of Chemistry, Zhejiang University, Hangzhou, China
e-mail: shangzc@zju.edu.cn

Keywords Multi-conformation-based quantitative structure–activity relationship · Side-chain conformational space analysis · Rotamer library · Self-consistent mean field theory · Protein–peptide complex · HLA-A*0201-restricted CTL epitope

Introduction

Quantitative structure–activity relationships (QSAR) represent an attractive approach for predicting compound activities when compared to more elaborate computational approaches, meriting in high speed and low cost. [1, 2]. QSAR researches in history could be divided into four stages: (i) 2D-QSAR, typically as linear free energy relationship proposed by Hansch [3], indicator variable analysis of Free and Wilson [4], and a series of topological indices developed by Winer, Randic, Balaban, et al. [5–7]; (ii) 3D-QSAR, mainly including alignment-dependent and alignment-independent. For the former, typical achievements are comparative molecular field analysis (CoMFA) proposed by Cramer et al. [8] and similar comparative molecular property analysis in following (e.g., CoMSIA [9] and CoMMA [10] etc.); while the latter is mostly contributed by grid-independent descriptors (GRIND) [11]; (iii) Multi dimension-QSAR (MD-QSAR). 4D-QSAR [12] proposed by Hopfinger et al. is the pioneering work, with a series of later works contributed by Vedani and co-workers [13]; (iv) Structure based-QSAR (SB-QSAR). Earlier, comparative binding energy analysis (COMBINE) [14] was proposed as the representation of SB-QSAR, and following, relevant works were gradually developed by many research groups including ours [15–17]. Pseudo receptor-based QSAR (PRB-QSAR) can be viewed as a transitional stage from 3D-QSAR to SB-QSAR, with typical methods of genetic evolved receptor model (GERM), pseudo atomic receptor model (PAEM), and receptor surface model (RSM) developed by Walters, Hahn and Chen et al. [18–20]. Therefore, QSAR essentially experiences from simple, intuitive to complex, abstractive transformations, and the future QSAR would be focused on how to really reflect active states of drug molecules, how to comprehensively analyze receptor–ligand binding information, and how to derive interpretable statistical model.

As it has been known that conformational characteristics of ligands in complex systems are a key factor to QSAR study, and in early 3D-QSAR approaches, often molecular "low-energy conformation" was directly served for this purpose. However in investigations of numerous complex crystal structures later, it was revealed that drug molecules exerting physiological effects were not always in its low-energy conformation, so the concept of "active conformation" was introduced. But, receptor–ligand binding is dynamic in physiological environment, with molecular property determined upon the ensemble of all possible conformations, i.e., "multi-conformation". On the other hand, nonbonding effects are the driving force for biomolecular recognition and association. In early CoMFA, only electrostatic and steric interactions were given consideration, and in later developed molecular field descriptors, hydrophobic force and hydrogen bond were added to improve the statistical quality and interpretability of the generated models. All these nonbonding interactions are regarded as enthalpy effects. However, another nonbonding effect, i.e., loss of conformational entropy [21], conventionally neglected in traditional QSAR studies.

In view of that, the multi conformation-based quantitative structure–activity relationship (MCB-QSAR) is proposed in this study. On the basis of side chain rotamer library defined by Lovell et al. [22], self-consistent mean field theory [23, 24] was used to determine conformation distributions of peptide side chains in the complex, then the nonbonding interactions including hydrogen bond, electrostatic, steric, hydrophobic interaction and entropy loss were calculated and analyzed based on the statistical average of conformation distributions. Here, this method is tentatively called the SCSA. SCSA inherits in many features of conformational analysis, receptor informational analysis and nonbonding potential analysis, etc., therefore being a comprehensive MCB-QSAR method. However, instead of the single fixed conformation in traditional QSARs, SCSA introduces multi conformational information, thereby effectively boosting modeling statistical quality and interpretability. Employing SCSA into QSAR modeling of 419 HLA-A*0201 binding peptides, the resulted statistical models provide insights into the mechanisms of HLA-A*0201 protein presenting antigen peptides, which may provide promising candidates for vaccine design.

Theoretical background and methodology

Research objects of SCSA

Protein–peptide interaction (PPI) plays an important role in most biological processes, such as antibody–antigen recognition, enzyme–substrate binding and receptor–hormone association. Thus, that PPI has driven considerable interests is due to its much significance in many subjects as biology, immunology and medicine [25]. SCSA aims to variant peptide ligands binding with a single protein receptor. Often the case is that the protein structure has been already solved by X-ray diffraction or constructed by homology modeling, and moreover, one or more protein–peptide complexes are also structurally known (which takes the co-crystallization structures or derives from molecular docking).

Overview of SCSA

For the protein–peptide complex, the backbone is usually viewed as a rigid body that is little changed, but the side chains are transforming among different discrete rotamers. In investigation of numerous protein crystal structures, conformation number of residue side chains is very limited in packing state [26], and possible rotamers of different residue side chains were derived from cluster analysis on structurally known proteins in protein data bank (PDB) [27], then collected together to define the rotamer library [28]. For the reason that we are only concerned about different peptides binding with a single receptor (protein), SCSA takes into account conformations of residue side chains of both the ligands (peptides) and active sites of the receptor. Thereafter, self-consistent mean field theory [23, 24] is employed to analyze side chain conformations of different peptides in sample set, resulting in probabilities of different rotamers in peptides. Subsequently, protein–peptide interactions are evaluated for enthalpy effect (indicated by statistical average over nonbonding interactions of all rotamers) and entropy effect (resulted from Boltzmann's formulation). The QSAR model is then constructed to relate these nonbonding effects (independent variables) with peptide bioactivities (dependent variables) by statistical modeling techniques.

Details in SCSA

(i) Protein–peptide complex. Protein–peptide complex can directly take the X-ray crystal structure. However, when crystal structure is unknown or unclear enough, theoretical pathway is considered to model and refine the complex structures, with relevant methods including homology modeling [29], molecular docking [30], molecular mechanics [31] and molecular dynamics [32], etc. Usually, hydrogen atoms, especially important to receptor–ligand interactions (i.e., hydrogen bonds and steric clashes are both directly determined by hydrogen position), are not explicitly presented in protein structural files downloaded from the PDB database, so prior to SCSA calculations, the protein is required to add all hydrogen atoms, which was fulfilled by Reduce program [33].

(ii) Self-consistent mean field theory (SCMFT). SCMFT proposed by Koehl and Delarue is widely used in protein conformation analysis and entropic calculations [23, 24], and here used to calculate probability distribution of side chain rotamers of peptides in the complex. Each residue's side-chain conformation is modeled as a rotamer with limited number of discrete states. The backbone-independent penultimate rotamer library (Fig. 1) used here was developed by Lovell et al. [22], recommended by Dunbrack Jr [28] for the various applications.

SCMFT calculation is a self-consistent process by iteration solving conformational matrix (CM), with matrix element CM(a, b) indicating the probability of residue a in rotamer b. Calculating procedure of SCMFT is briefly demonstrated as following:

In SCMFT, effective potential (E) of rotamer k of side-chain i is as:

$$E(i,k) = U(x_{ik}) + U(x_{ik},x_0) + \sum_{\substack{j=1\\ j\neq i}}^{N} \sum_{l=1}^{K_j} \mathrm{CM}(j,l)U\left(x_{ik},x_{jl}\right) \quad (1)$$

where x_{ik} is the atomic coordinates of side chain i in rotamer k; first term $U(x_{ik})$ is the potential for this rotamer

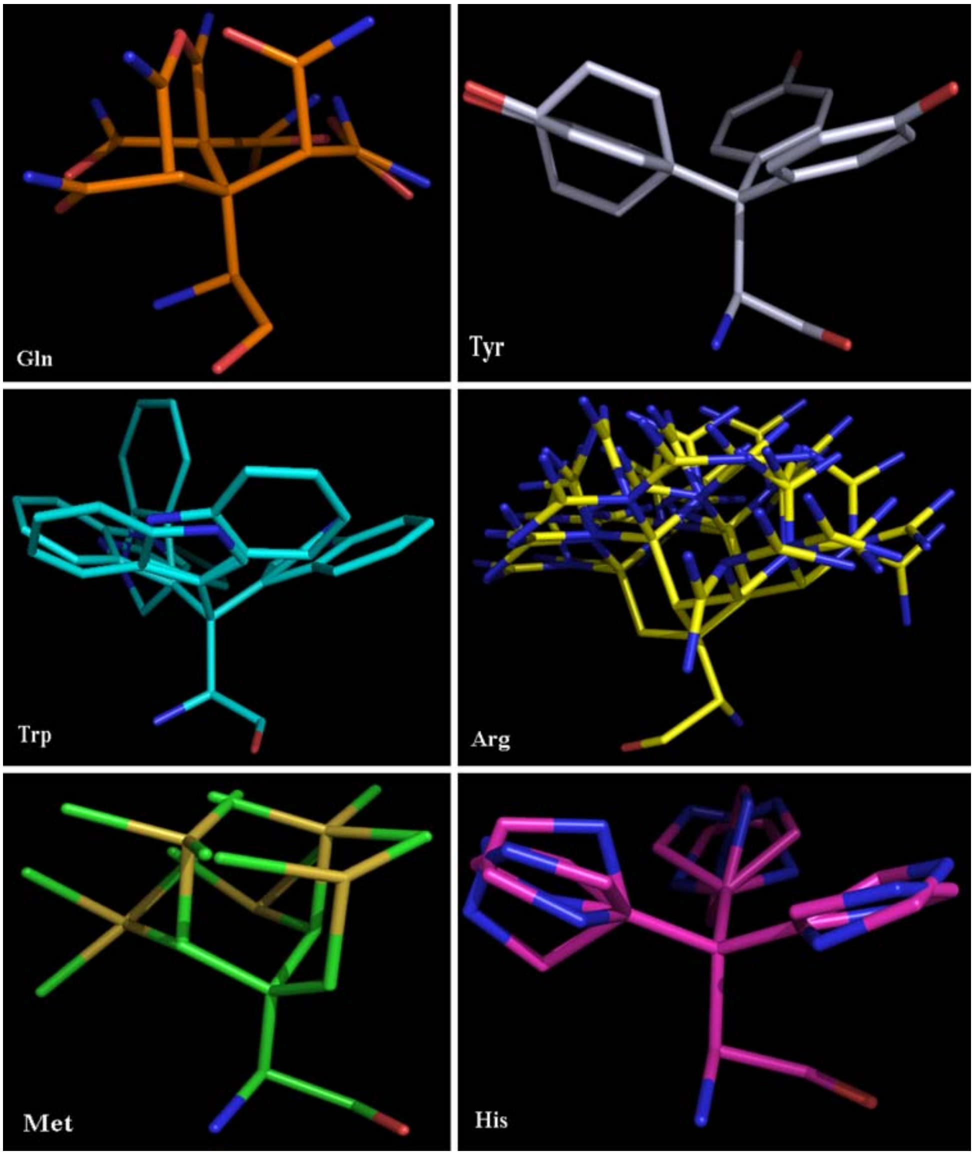

Fig. 1 Illustration of the side chain rotamers for amino acids Gln, Tyr, Trp, Arg, Met, and His contained in the penultimate rotamer library [22]. This figure was produced using PyMOL [72]

alone; the second term $U(x_{ik}, x_0)$ is the potential for this rotamer in the context of fixed coordinates (main chain, C_β atoms, residues in active site of protein, etc.); third term $\sum_{j=1, j\neq i}^{N} \sum_{l=1}^{K_j} \mathrm{CM}(j,l)U(x_{ik}, x_{jl})$ is the interaction potential between the side chain i in rotamer k and all other side chains in all possible rotamers. CM(j, l) is a conformational matrix element, giving the probability that the conformation of side chain j is described by rotamer l; $U(x_{ik}, x_{jl})$ is the potential between the k rotamer of side-chain i and the l rotamer of side-chain j. The double sum is over all side chains j (1 to N, not equal to i), and over all rotamers (1 to K_j) for each side chain j [34]. In SCSA, side-chain multi conformations of ligand peptide and of active site residues of the protein are given considerations. Thus in Eq. 1, coordinates of non-active site residues of protein are fixed, with the interaction potential with rotamer x_{ik} incorporated into $U(x_{ik}, x_0)$ term. In this study, potential function uses the nonbonding term of OPLS-AA force field [35, 36], expressed as the following:

$$U(A,B) = U_{\mathrm{vdW}} + U_{\mathrm{coulombic}} = \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} f_{ij}\left[4\omega_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right) + k_c\frac{q_i q_j}{\varepsilon_p r_{ij}}\right] \quad (2)$$

where $f_{ij} = 1.0$ except for intramolecular 1,4-interactions for which $f_{ij} = 0.5$; r_{ij} is the distance between atoms i and j; ω_{ij} is van der Waals (vdW) well depth, σ_{ij} is vdW equilibrium distance, taken from reference [36]; k_c is Coulomb's constant, ε_p is the protein dielectric constant; q_i is the charge of atom i. Parameters included in Eq. 2 are suggested by Chowdry et al. [37] that interactions between atoms greater than 10 Å apart are neglected. To prevent favorable Coulombic interactions from overwhelming the repulsive contribution of the vdW potential, attractive Coulombic interactions are capped by an adjustable parameter with a default value of −12.45 kcal/mol (−0.3e^2/A). Conversely, the vdW repulsive energies are capped with a default value of 1,000 kcal/mol. A dielectric constant of 8.0 is used.

Iterative solution of CM: at step 0, initial conformational matrix CM_0 is assigned to the residues of ligand peptide and the active site residues of protein, with probability distribution of residue rotamers in CM_0 directly taking the statistical probability in rotamer library; At step $t + 1$, effective potential of rotamers for each residue is calculated by Eq. 1 coupled with CM_t, and subsequently, matrix elements of CM_{t+1} are calculated using Boltzmann distribution law; such an iteration is repeated until convergence, i.e., $\mathrm{CM}_{t+1} = \mathrm{CM}_t$ (or $\mathrm{CM}_{t+1} \approx \mathrm{CM}_t$).

(iii) Enthalpy effect in protein–peptide interactions. Enthalpy effects include electrostatic, steric, hydrophobic interaction and hydrogen bond. To analyze the contributions of each residue in ligand peptides to the binding, all the nonbonding interactions are separately calculated. Moreover, nonbonding interaction between rotamers of the peptide residue and protein are evaluated by the statistical average:

$$\Delta H_i^{\mathrm{nonbond}} = \sum_{k=1}^{K_i} \mathrm{CM}(i,k)\Delta H_{ik}^{\mathrm{nonbond}} \quad (3)$$

$\Delta H_{ik}^{\mathrm{nonbond}}$denotes nonbonding interactions between the rotamer k of residue i in peptide and the protein, and *nonbond* can be any one of electrostatic, steric, hydrophobic or hydrogen bond interactions; CM(i, k) is the probability of residue i in rotamer k. When calculating peptide rotamers in interactions with the protein, the protein is viewed as single fixed conformation (taking the crystal structure).

In SCSA, electrostatic and steric interactions of rotamer ik are calculated by Coulombic and vdW terms ($\Delta H_{ik}^{\mathrm{coulombic}}$ and $\Delta H_{ik}^{\mathrm{vdW}}$) in OPLS-AA potential function; while hydrophobic interactions and hydrogen bonds are calculated by Eqs. 4 and 5:

$$\Delta H_{ik}^{\mathrm{hydrophobic}} = -\sum_{\substack{l\in \mathrm{rotamer},ik \\ j\in \mathrm{receptor}}} (S_l\rho_l + S_j\rho_j)e^{-r_{lj}} \quad (4)$$

$$\Delta H_{ik}^{\mathrm{hydrogenbond}} = \sum_{\substack{l\in \mathrm{rotamer},ik \\ j\in \mathrm{receptor}}} \cos^4\theta\left[\frac{C_{lj}}{r_{lj}^{12}} - \frac{D_{lj}}{r_{lj}^{10}}\right] \quad (5)$$

Equation 4 is an empirical hydrophobic interaction function [17], in which S_l denotes solvent-accessible surface area [38] of the atom (or group) l of rotamer ik; ρ_l is the hydrophobicity of atom (or group) l, usually derived from experimental data fitting [39], and here atomic hydrophobicity takes Eisenberg scale [40]. Eq. 5 is the hydrogen bond term in the scoring function of AutoDock [41], and the parameters C and D are assigned to give a maximal well depth of 5 kcal/mol at 1.9 Å for O…H and N…H, and a depth of 1 kcal/mol at 2.5 Å for S…H; $\cos^4\theta$ is a directional weight based on the angle θ away from ideal bonding geometry and is fully described by Goodford et al. [42, 43], the hydrogen bond energy is set to zero when $\theta \leq 90°$.

(iv) Entropy effect in protein–peptide interactions. Based upon Boltzmann's law, the conformational matrix CM is used to estimate the conformational entropy of every side chains of peptide in protein–peptide complex [34]:

$$S_i^{\mathrm{complex}} = -R\sum_{k=1}^{K_i} \mathrm{CM}(i,k)\ln \mathrm{CM}(i,k) \quad (6)$$

where R is the universal gas constant. When peptide is in free state (no binding with protein), the conformational entropy

S_i^{free} of residues is evaluated by Monte Carlo simulation by Creamer [44]. Therefore, loss of conformational entropy of residue i in peptide are expressed as conformational free energy:

$$-T\Delta S_i = -T\left(S_i^{\text{complex}} - S_i^{\text{free}}\right) \tag{7}$$

T is thermodynamic temperature, taking 300 K.

(v) Genetic algorithm-partial least square regression (GA-PLS) modeling. By SCSA calculations, 5 nonbonding terms are generated for each peptide residue (i.e., electrostatic, steric, hydrophobic, hydrogen bond and conformational free energy), indicating the statistical average over all its multi conformations (rotamers). Thus for a peptide with length of N, there are totally $5 \times N$ nonbonding terms (i.e., SCSA descriptors) which are used as independent variables for QSAR modeling. For a QSAR data set, not all the structural descriptors are related significantly to binding affinity. So prior to modeling, redundant descriptors should be deleted in order to improve model stability and predictive power especially when variables number is large [45, 46]. In QSAR, genetic algorithm (GA) is widely used for variable selection [47], while partial least square (PLS) regression is a classical multi-dimension statistical modeling technique, potent to compressing and filtering complicated information involved in independent variable matrix X [48, 49]. In view of that, GA-PLS is recommended to perform QSAR modeling between SCSA descriptors and binding affinities.

Dataset

Numerous peptides and analogues were found to have potential bioactivities [50], e.g., bradykinins, oxytocin analogues, bitter tasting peptides, antipeptides and antimicrobial peptides, etc. Here, peptide-major histocompatibility complex (MHC) interactions are ideally selected for SCSA analysis, ascribed to the following: (i) Many crystal structures of MHC–peptide complexes have been already solved [51]; (ii) Abundant experimental data of binding affinities ensures the statistical reliability of QSAR modeling (e.g., Peters et al. [52] recently publish a benchmarking database comprising 48,828 quantitative peptide-binding affinity measurements of different MHC I binding peptides; Moreover, Toseland et al. [53] have constructed the AntiJen database which also records numerous quantitative affinity entries of diverse peptides binding with immune receptors); (iii) Independent binding of side-chains (IBS) hypothesis proposed by Parker et al. [54, 55] pointed out that in MHC–peptide complex, each residue side chain binds independently of the rest of the peptide, that complies with that each peptide residue is independently calculated in SCSA; (iv) Immune system is a very important but extremely complex human tissue, and recently arisen computer-aided vaccine design (CAVD) [56] powerfully improves the developments of antigen-presenting mechanism, vaccine design and immunotherapy. Thus that performing CAVD studies by QSAR strategies becomes important, driving considerable interest [57].

HLA-A*0201, a kind of human MHC molecules, is coded by 0201-type allele at the HLA-A locus, serving as one of the most frequent class I alleles in many different populations. For example, it is expressed in approximately 50% of Caucasians [58] and has been demonstrated to play a critical role in antigen presentation of both viral antigens [59] and tumor antigens from a variety of cancers [60]. The two anchor residues fall into the hydrophobic pocket of the peptide-binding cleft of HLA-A*0201, and they are normally referred as positions P_2 and P_9, so the HLA-A*0201 binding nonapeptides are presented as $P_1P_2P_3P_4P_5P_6P_7P_8P_9$. 473 HLA-A*0201-restricted CTL epitopes (nonapeptides) were collected from the AntiJen database [53], and the binding affinity IC_{50} we used here was assessed by the required doses of replacing 50% of radiolabeled standard peptide from the HLA-A*0201-standard peptide complex in given time. To ensure different IC_{50} has comparability, we only selected the 419 samples of which the standard peptide is HBVc18227 (FLPSDYEPSV). Amongst, 138 are high-affinity peptides ($IC_{50} \leq 50$ nM, $pIC_{50} \geq 7.301$), 185 peptides have intermediate affinity (50 nM $< IC_{50} \leq 500$ nM, $7.301 > pIC_{50} \geq 6.301$), and 96 are low-affinity peptides ($IC_{50} > 500$ nM, $pIC_{50} < 6.301$).

An excellent QSAR model should be robust, stable and predictable, and usually cross-validated correlative coefficient q^2 is used to evaluate modeling quality. However, Tropsha et al. [61–63] asserted that a more reliable result should be tested by external validation. Thus, the 419 peptides were divided into two portions: a training set to establish model and a test set to evaluate model. D-optimal [64, 65] was used to serve this purpose, and as a result, 69 out of 419 peptides were selected to enter into the test set. In D-optimal calculation, peptide structures were characterized by orthogonal coding [66] coupled with principal component analysis (PCA), i.e., each position of nonapeptide was characterized by 20 binary variables, and different variable (binary) indicates occurrence or not of a certain amino acid (1 denotes presence and 0 for absence). Then PCA was employed to compress variable space for all the 419 samples in order to speeding up calculations of D-optimal.

Sequences, experimental and calculated binding affinities (expressed as pIC_{50}) of 419 nonapeptides are provided as Table S1 in Supplementary Material.

Results and discussion

SCSA calculation and QSAR modeling

The three-dimensional crystal structure of HLA-A*0201-peptide (LLFGYPVYV) complex at 1.8 Å resolution by X-ray diffraction was revealed in Fig. 2a, wherein two α-helixes and one β-sheet make up the peptide-binding cleft of the HLA-A*0201 molecule, with the embedded peptide unfolding. Figure 2b shows the peptide backbone (red marked) extracted from the complex and side chain rotamers. About 26 peptides extracted from crystal complexes were aligned using ProFit program [67] (see Table S2 in Supplementary Material), indicating that in most cases, the backbone Cα atoms were well overlapped with each other (RMS < 0.5 Å). Therefore, backbone conformations of diverse HLA-A*0201-binding peptides are supposed to be approximate, and this assumption is similar to Doytchinova's report [68]. In view of that, backbone of LLFGYPVYV in the crystal structure of HLA-A*0201-LLFGYPVYV complex (PDB ID: 1DUZ) was selected as the template. The crystal structure of 1DUZ was refined by Khan et al. [69] on the basis of earlier works by Madden et al. [70], with its resolution highly achieving 1.8 Å. Prior to calculation, the crystal structure was preprocessed: (i) Adding hydrogen atoms; (ii) Removing crystal water; (iii) Completing the missing atoms, fragments and groups; (iv) Assigning correctly bond order, atomic type in aromatic ring and protonated/ionized state.

Taking the preprocessed HLA-A*0201-peptide backbone as the template, we employed SCSA to calculate nonbonding interactions between the HLA-A*0201 protein and multi rotamers of the 419 samples. During this process, active site residues of the HLA-A*0201 protein adopted our previous definitions [17]. Due to HLA-A*0201-restricted CTL epitopes are all nonapeptides, 5 × 9 = 45 nonbonding interaction terms (SCSA descriptors) were generated for each peptide, referring as variables V_1–V_{45}. Amongst, V_1–V_5 are 5 SCSA descriptors for position 1, V_6–V_{10} are orderly 5 SCSA descriptors for P_2, and so forth. After that, the model was built using 350 training samples, and the generated SCSA descriptors of the independent variable matrix $\mathbf{X}_{350\times45}$ were related to dependent variable vector $\mathbf{Y}_{350\times1}$ (affinities). Modeling methods are PLS and GA-PLS. GA-PLS parameters settings: Population size, 100; Genmax, 200; Mutation rate, 1.0%; Hybridization and crossover, 2 points; Fitness function, cross-validation; Data pretreatment, autoscaling. According to the suggestion by Shao [71], internal stability validation is performed by leave-one-out cross-validation (LOO-CV) and leave-group (1/5)-out cross-validation (LGO-CV). External validation on the GA-PLS model was implemented using 69 test samples based on Tropsha's statistics [61].

In this study, SCSA was implemented using in-house program written in C++; SIMCA-P 10.0 (Umetrics AB, Umea, Sweden) and Matlab 7.4 (MathWorks, Natick, MA, USA) were used for the PLS and GA-PLS calculations.

Analysis of the statistical models

PLS model. Without GA-variable selection, the PLS model was directly constructed using all SCSA descriptors. Number of principal components (NPCs), fitting correlation coefficient r^2, LOO-CV q^2, LGO-CV q^2 and root mean square error of estimation (RMSEE) are 5, 0.704, 0.378, 0.214, and 0.492, respectively. We can see the modeling stability is low ($q^2 < 0.4$) in spite of a high fitting ability ($r^2 > 0.7$). Thus the PLS model may be overfitted due to the strong noise involved in SCSA descriptors.

GA-PLS model. Subject to variable selection, the GA-PLS model was built. Ultimately 32 significant variables were selected from 45 SCSA descriptors, and the model's statistics NPCs, r^2, LOO-CV q^2, LGO-CV q^2 and RMSEE are 4, 0.692, 0.579, 0.511, and 0.517, respectively. Compared with the PLS model, the GA-PLS model is greatly improved, especially with the cross-validation q^2 greatly increased, and this model meets the recommended criteria by Golbraikh et al. ($r^2 > 0.6$, $q^2 > 0.5$) [62]. In

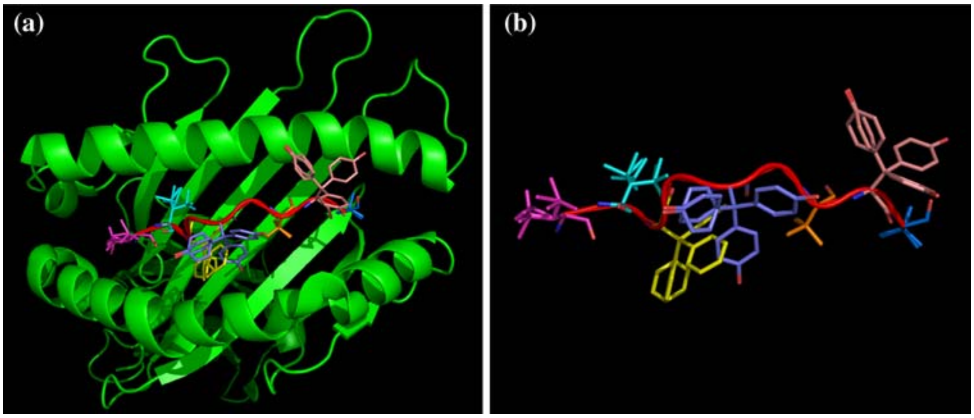

Fig. 2 **a** Crystal structure of HLA-A*0201-LLFGYPVYV complex (PDB ID: 1DUZ); **b** spatial conformation of peptide backbone (colored in red) and side chain rotamers (based on penultimate rotamer Library [22]). This figure was produced using PyMOL [72]

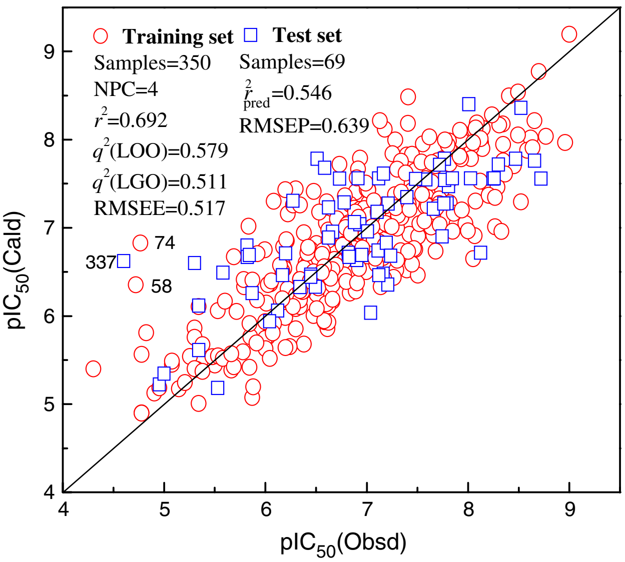

Fig. 3 The calculated versus observed activities for 419 peptide using GA-PLS

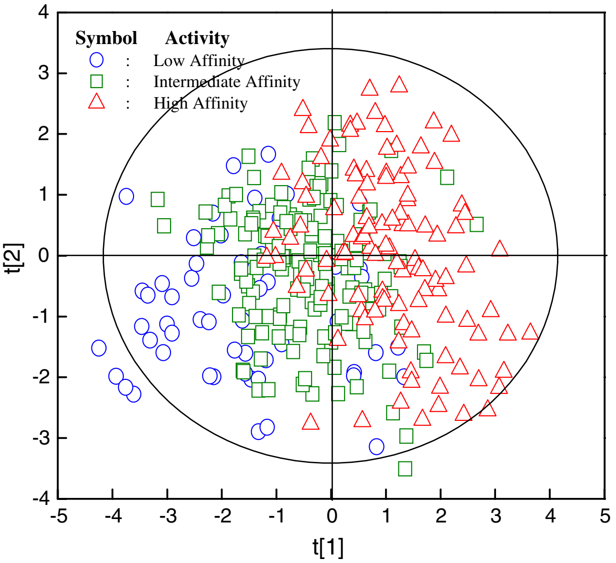

Fig. 4 Scoring scatters for 350 training samples at the top two principal component spaces (low affinity: $pIC_{50} < 6.301$ ($IC_{50} > 500$ nM); intermediate affinity: $7.301 > pIC_{50} > 6.301$ (50 nM $< IC_{50} <$ 500 nM); high affinity: $pIC_{50} > 7.301$ ($IC_{50} <$ 50 nM))

Fig. 3, training set samples are marked by circle to reveal the correlation between the observed and calculated values in the GA-PLS model, indicating except for two outliers, most samples are dispersed around an origin-passed diagonal. By analysis, the probable basis of these exceptions is rationalized as follows: for peptides number 58 (FLFGSLAFL) and number 74 (FLYAALLLA), their observed affinities are only 4.721 and 4.765, respectively. However, their residues at P_2 and P_9 are each classical anchor residues Leu and Ala. Usually, such peptides are supposed to have high binding affinities. Thus, that the overestimation of these two samples by the GA-PLS model may be due to their low experimental affinities ($pIC_{50} < 5.0$). It is believed that special structures and binding affinities, of course including experiment errors, may all contribute to large calculation errors. In traditional QSAR studies, outliers are conventionally removed from the model, however this may risk losing valuable information. In contrast, the GA-PLS model has been already qualified, and these outliers were cautiously kept. Figure 4 shows the scoring scatters for 350 training set samples in the top two principal component spaces, wherein most samples fall within the Hotelling T2 ellipse of 95% confidence level except for 4 low-affinity samples. In investigations, these 4 outliers are greatly different from other samples in high-dimensional structural characteristics. Besides, peptides are increasingly distributed from left-low to the right-upper in terms of binding affinities in this figure, showing obvious regularity (peptides in different affinity ranks are marked by different symbols). Therefore, the PLS principal component spaces are believed to favorably reflect the structural and active features of peptides. Subsequently, a rigorous statistical diagnosis was further implemented for the GA-PLS model by Y random permutations test and normal probability of standardized residual analysis [49], indicating there are no morbidities such as abnormality, overfitting and random correlation, etc. in this model.

External validation. The GA-PLS model was validated by performing predictions on 69 test set samples, and in Fig. 3, test samples are marked by square to indicate the correlation between the predicted and the observed affinities, with correlation coefficient r^2_{pred} and RMSEP (root mean square error of prediction) are 0.546 and 0.639, respectively. Amongst, sample number 337 has a large positive error, similar to the case of samples number 58 and 74 in training set, the anchor residues P_2 and P_9 of sample number 337 are hydrophobic Leu and Val, but the observed affinity is very low (only 4.602), so this sample (number 337) was overestimated by the GA-PLS model. For test set, Tropsha et al. [61] proposed diagnostic statistics as external correlation coefficient q^2_{pred}, $r^2_{0,\mathrm{ext}}$, $r'^2_{0,\mathrm{ext}}$ and predicting slopes k, k' to evaluate models. Thereby, a predictable QSAR model is identified by the following criteria:

$$\frac{r^2_{\mathrm{pred}} - r^2_{0,\mathrm{ext}}}{r^2_{\mathrm{pred}}} < 0.1 \text{ or } \frac{r^2_{\mathrm{pred}} - r'^2_{0,\mathrm{ext}}}{r^2_{\mathrm{pred}}} < 0.1 \qquad (8)$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \qquad (9)$$

where $r^2_{0,\mathrm{ext}}$ and $r'^2_{0,\mathrm{ext}}$are the coefficients of determination for the regression through origin (predicted versus observed activities $r^2_{0,\mathrm{ext}}$, and observed versus predicted activities $r'^2_{0,\mathrm{ext}}$), and k together with k' are the slopes of the origin-passed regression line. Based on such criteria, the Tropsha's statistics of GA-PLS-predicted results was calculated and listed in Table 1. The predictive power of the GA-PLS model satisfies Eqs. 8 and 9. Note that the external correlation coefficients r^2_{pred}, $r^2_{0,\mathrm{ext}}$ and $r'^2_{0,\mathrm{ext}}$ on test set are extremely approximate each other, suggesting the model possesses a favorable unbiasedness. By such a rigorous statistical diagnosis, the GA-PLS model is shown to have excellent stability and predictability.

Analysis of different positions in peptides

SCSA descriptors, coefficients, and variable importance in the projection (VIP) of the variables of the GA-PLS model are summarized in Table 2. VIP is the sum of the variable influence over all modeling dimensions and is a measure of variable importance. Higher VIP values indicate good correlation between the variable and the data [49]. It can be seen that 32 variables, including hydrogen bond of P_1,

Table 1 Statistics on test set predicted by the GA-PLS model

Common statistics		Tropsha's statistics						
r^2_{pred}	RMSEP	q^2_{ext}	$r^2_{0,\mathrm{ext}}$	$r'^2_{0,\mathrm{ext}}$	$\frac{r^2_{\mathrm{pred}} - r^2_{0,\mathrm{ext}}}{r^2_{\mathrm{pred}}}$	$\frac{r^2_{\mathrm{pred}} - r'^2_{0,\mathrm{ext}}}{r^2_{\mathrm{pred}}}$	k	k'
0.546	0.639	0.489	0.541	0.538	0.009	0.015	0.877	0.914

Table 2 SCSA descriptors, coefficients, and VIPs of the GA-PLS model

Number	Position	Description of SCSA descriptor	Standardized coefficient	VIP
1	P_1	Steric contact	0.1088	0.325
2		Hydrophobic interaction	0.1896	0.784
3		Hydrogen bond	0.2014	0.883
4		Conformational entropy	0.1289	0.355
5	P_2	Electrostatic interaction	−0.0280	0.784
6		Steric contact	−0.0563	0.454
7		Hydrophobic interaction	0.2656	1.678
8		Hydrogen bond	0.2113	1.253
9		Conformational entropy	0.2414	1.102
10	P_3	Hydrophobic interaction	0.0914	0.670
11		Hydrogen bond	0.1064	0.824
12		Conformational entropy	−0.1851	0.418
13	P_4	Electrostatic interaction	−0.0820	0.114
14		Hydrophobic interaction	−0.1438	0.564
15		Hydrogen bond	0.0785	0.102
16	P_5	Electrostatic interaction	0.1321	0.746
17		Hydrophobic interaction	−0.0512	0.119
18	P_6	Electrostatic interaction	0.1784	0.986
19		Steric contact	−0.0843	0.226
20		Conformational entropy	−0.0914	0.183
21	P_7	Electrostatic interaction	0.2238	1.346
22		Hydrophobic interaction	0.0938	0.298
23		Hydrogen bond	−0.0687	0.188
24		Conformational entropy	0.0720	0.342
25	P_8	Steric contact	−0.1238	0.544
26		Hydrophobic interaction	−0.0463	0.138
27		Hydrogen bond	0.0938	0.346
28	P_9	Electrostatic interaction	−0.0238	0.270
29		Steric contact	0.0738	0.415
30		Hydrophobic interaction	0.2463	1.113
31		Hydrogen bond	0.2138	1.498
32		Conformational entropy	0.1987	0.996

hydrophobic interaction of P_2, hydrogen bond of P_2, conformational entropy of P_2, hydrogen bond of P_3, electrostatic interaction of P_5, electrostatic interaction of P_7, hydrophobic interaction of P_9, hydrogen bond of P_9 and conformational entropy of P_9, have the relatively larger VIP values. Figure 5 shows the schematically representation of hydrophobic interaction and hydrogen bond of peptides in interaction with the HLA-A*0201 protein. We can see that the many positions in peptide are encircled by hydrophobic residues, and hydrogen bonds are intensive at the two ends.

Anchor residues, P_2 and P_9. Large hydrophobic side chains are preferred at P_2. They fall into hydrophobic pocket of the peptide-binding site on HLA-A*0201 molecule. This pocket has a polar rim and hydrophobic inner walls made up of Phe9, Met45 and Val67, Leu, Ile, and Met are well-accommodated in this pocket [73]. Moreover, P_2 is shown to have a high negative electrostatic potential environment [17], whereas large amounts of hydrophobic residues are assembled and may form some hydrogen bonds (Fig. 5). At P_2, five SCSA descriptors were introduced in the GA-PLS model by GA, and amongst the hydrophobic interaction, hydrogen bond and conformational entropy possess large VIP values (>1), indicating these nonbonding interactions have remarkable effects on HLA-A*0201-peptide binding. P_9 is often occupied by Leu and Val, yet allowing replacement by other hydrophobic residues such as Ile and Ala. Amino acids with hydrophobic short side chains such as Ala and Val are required for P_9 [74]. It is revealed in Fig. 5 an intensive hydrogen network formed between the HLA-A*0201 and the P_9. The VIP value of hydrogen bond at P_9 is relatively large (1.498). Besides, the hydrophobic interaction and conformational entropy also have significant contributions to binding affinity (Table 2). Note that the conformational entropy contributions of P_2 and P_9 are significant according to the VIPs and standardized coefficients of GA-PLS model. This can be considered as the anchor residues form close contacts with receptor, and the conformational space of their side chains are greatly constrained during the binding process.

Second anchor residues, P_1, P_3 and P_7. These positions were defined as the secondary anchor residue by Ruppert et al. [75], also contributing significantly to peptide affinities. In our study, 4, 3 and 4 SCSA descriptors were selected at P_1, P_3 and P_7, respectively. Sapper and Bjorkman [76] suggested that around P_1, several Tyr residues were in presence, forming a complicated hydrogen bond network, and the receptor residue Tyr159, a hydrogen bond donor, interacts with carboxyl oxygen at this position [51]. Hence the Phe and Tyr are preferred in this position to form π-π conjugations, thus hydrogen bonding with surrounding residues. In Fig. 5, a series of hydrogen bonds and hydrophobic interactions are presented. As shown in Table 2, the VIPs of hydrophobic interaction and hydrogen bond at P_1 have significant contributions to binding affinity. In P_3, amino acids with hydrophobic aromatic rings will enhance the binding affinity of the peptides. This pocket is mostly hydrophobic, accommodating Phe and Trp well [77, 78]. Figure 5 implicates at P_3, hydrophobic stacking and hydrogen bond are critical nonbonding interactions, and the VIP values of P_3 also demonstrate this point. P_7, although served as the secondary anchor residue, has fewer contributions to binding in comparison with P_1 and P_3. Sapper et al. [76] suggested that most regions around P_7 are overlapped by hydrophobic residues, and meanwhile one of its sides is occupied by strongly polar Arg97, so P_7 is amphipathic, preferred residues with small hydrophobic side chains (e.g., Val and Ala). In the GA-PLS model, the electrostatic interaction of P_7 is significant, inferred that the amphipathic environment around P_7 led to a notable electrostatic nonuniformity.

Non-anchor residues, P_4, P_5, P_6 and P_8. These positions loosely contact with HLA-A*0201 protein, and no intensive nonbonding interactions are in presence between the both [17]. In the GA-PLS model, few SCSA descriptors were introduced at these positions, and their VIP contributions were also low. Although less significant than anchor residues, they also partially affect the peptide–HLA-A*0201 binding [79].

Conclusions

In physiological environment, ligand–receptor recognition and binding are a very complex dynamic process, involved in multi conformational interactions. However, most present QSAR approaches merely take single conformation (e.g., low-energy conformation or active conformation) to create the statistical models, failing to reflect dynamic features of the complex systems. In view of that, the concept of multi conformation-based QSAR (MCB-QSAR) is proposed here, and based upon side chain rotamer library and self-consistent mean field theory, side chain conformational distributions of ligand peptides in the protein–peptide complex are modeled, and subsequently, the statistical average of multi-conformational interactions between the protein and its peptide ligands are determined. We called this new MCB-QSAR method as the side-chain conformational space analysis (SCSA). In contrast with traditional QSAR, multi conformations of peptide side chains have received much attention in the SCSA, and concomitantly, conformational entropy is also introduced definitely to construct the statistical model. In this study, SCSA was employed to perform systematical MCB-QSAR study on 419 HLA-A*0201-restricted CTL epitopes, and

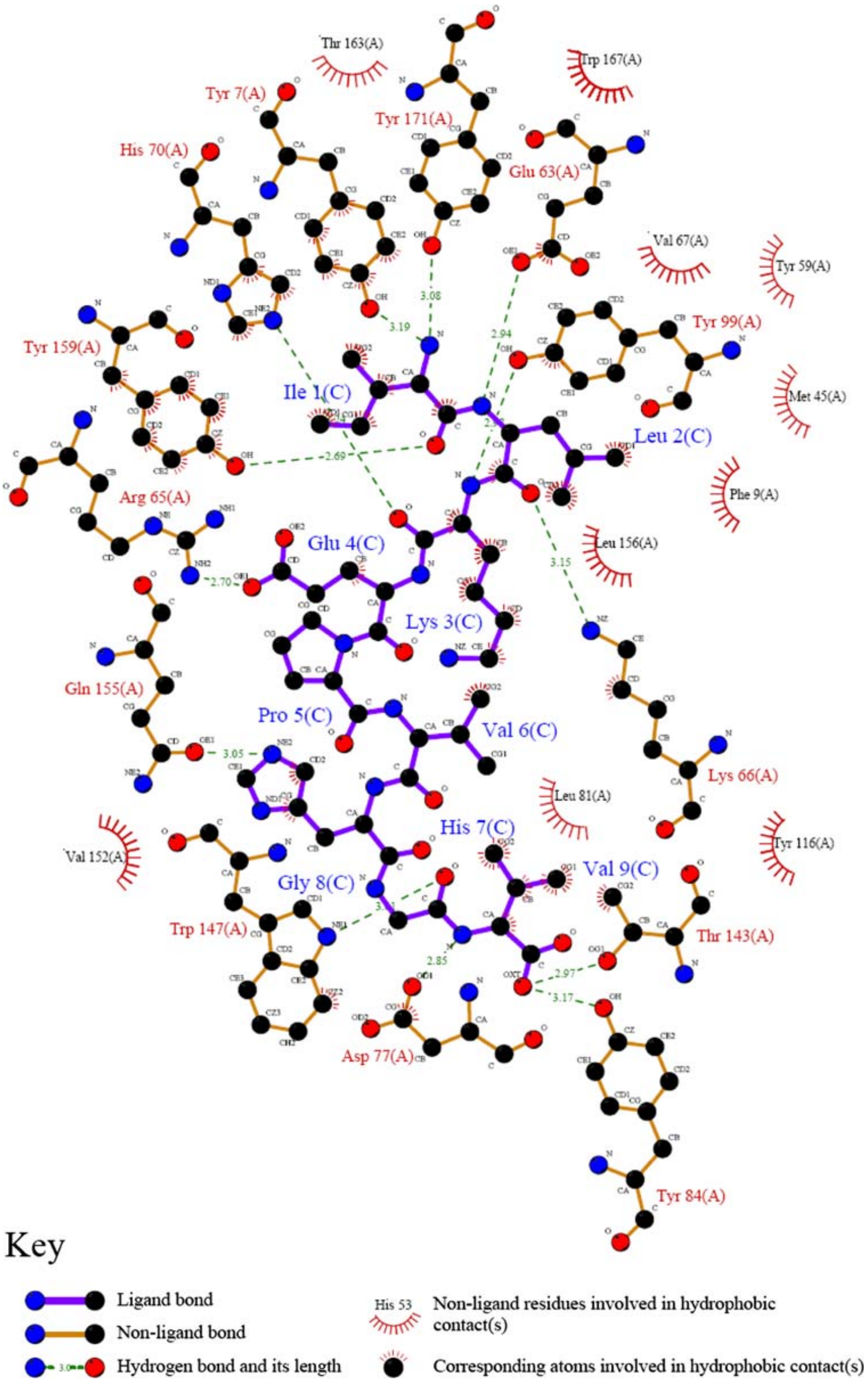

Fig. 5 A schematic representation of the hydrophobic interactions and hydrogen bonds for HLA-A*0201–peptide complex. Fringy and broken lines indicate hydrophobic interactions and hydrogen bonds, respectively (Template: 1AKJ, produced using LIGPLOT [80])

by rigorous internal and external double validations, the resulted GA-PLS model was confirmed of favorable stability and predictive power. The results show that different properties of the residues in nonapeptide may contribute distinctly to the binding. Particularly, hydrophobicity and hydrogen bond are more significant during the binding process. In analyses of the VIP values of GA-GP model and molecular graphics exhibitions, the anchor residues P_2 and P_9 of the peptides are the key positions for the peptide binding. In addition, the second anchor residues (P_1, P_3 and

P_7) also have relatively large contributions to the binding, and other positions partially affect the interactions between the HLA-A*0201 protein and its peptide ligands. The hydrogen bonds are essential to the binding specificity, and van der Waals contacts and hydrophobic interactions ensure the binding strength.

References

1. Winkler DA (2002) The role of quantitative structure-activity relation-ships (QSAR) in biomolecular discovery. Brief Bioinform 3:73–86. doi:10.1093/bib/3.1.73
2. Fujita T (1997) Recent success stories leading to commercializable bioactive compounds with the aid of traditional QSAR procedures. Quant Struct-Act Relat 16:107–112. doi:10.1002/qsar.19970160202
3. Hansch C, Fujita T (1964) ρ-σ-π analysis: a method for the correlation of biological activity and chemical structure.J Am Chem Soc 86:1616–1626. doi:10.1021/ja01062a035
4. Free SM, Wilson JB (1964) A mathematical contribution to structure-activity studies. J Med Chem 7:395–399. doi:10.1021/jm00334a001
5. Winer H (1947) Structural determination of paraffin boiling point. J Am Chem Soc 69:2636–2641. doi:10.1021/ja01203a022
6. Randic M (1975) On characterization of molecular branching. J Am Chem Soc 97:6609–6615. doi:10.1021/ja00856a001
7. Balaban AT (1982) High discrimination distance-based topological index. Chem Phys Lett 89:399–404. doi:10.1016/0009-2614(82)80009-2
8. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis(CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967. doi:10.1021/ja00226a005
9. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146. doi:10.1021/jm00050a010
10. Silverman BD, Platt DE (1996) Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. J Med Chem 39:2129–2140. doi:10.1021/jm950589q
11. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S (2000) GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. J Med Chem 43:3233–3243. doi:10.1021/jm000941m
12. Hopfinger AJ, Wang S, Tokarski JS, Jin BQ, Albuquerque M, Madhav PJ et al (1997) Construction of 3D-QSAR models using 4D-QSAR analysis formalism. J Am Chem Soc 119:10509–10524. doi:10.1021/ja9718937
13. Vedani A, Dobler M (2002) Multidimensional QSAR: moving from three- to five-dimensional concepts. Quant Struct-Act Relat 21:382–390. doi:10.1002/1521-3838(200210)21:4<382::AID-QSAR382>3.0.CO;2-L
14. Wade RC, Oritz AR, Gago F (1998) Comparative binding energy analysis. Perspect Drug Discov Des 9:19–34. doi:10.1023/A:1027247618908
15. Pouplana R, Lozano JJ, Pérez C, Ruiz J (2002) Structure-based QSAR study on differential inhibition of human prostaglandin endoperoxide H synthase-2 (COX-2) by nonsteroidal anti-inflammatory drugs. J Comput-Aid Mol Des 16:683–709. doi:10.1023/A:1022488507391
16. Santos-Filho OA, Hopfinger AJ (2006) Structure-based QSAR analysis of a set of 4-hydroxy-5, 6-dihydropyrones as inhibitors of HIV-1 protease: an application of the receptor-dependent (RD) 4D-QSAR formalism. J Chem Inf Model 46:345–354. doi:10.1021/ci050326x
17. Zhou P, Tian F, Li Z (2007) A structure-based, quantitative structure-activity relationship approach for predicting HLA-A*0201-restricted cytotoxic T lymphocyte epitopes. Chem Biol Drug Des 69:56–67. doi:10.1111/j.1747-0285.2007.00472.x
18. Walters DE, Hinds RM (1994) Genetically evolved receptor models: a computational approach to construction of receptor models. J Med Chem 37:2527–2536. doi:10.1021/jm00042a006
19. Hahn M (1995) Receptor surface models. 1. Definition and construction. J Med Chem 38:2080–2090. doi:10.1021/jm00012a007
20. Chen H, Zhou J, Xie G (1998) PARM: a genetic evolved algorithm to predict bioactivity. J Chem Inf Comput Sci 38:243–250. doi:10.1021/ci970004w
21. Frederick KK, Marlow MS, Valentine KG, Wand AJ (2007) Conformational entropy in molecular recognition by proteins. Nature 448:325–329. doi:10.1038/nature05959
22. Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. Proteins 40:389–408. doi:10.1002/1097-0134(20000815)40:3<389::AID-PROT50>3.0.CO;2-2
23. Koehl P, Delarue M (1994) Application of a self consistent mean field theory to predict protein side-chain conformations and estimate their conformational entropy. J Mol Biol 239:249–275. doi:10.1006/jmbi.1994.1366
24. Koehl P, Delarue M (1996) Mean-field minimization methods for biological macromolecules. Curr Opin Struct Biol 6:222–226. doi:10.1016/S0959-440X(96)80078-9
25. Stanfield RL, Wilson IA (1995) Protein-peptide interactions. Curr Opin Struct Biol 5:103–113. doi:10.1016/0959-440X(95)80015-S
26. Bhat TN, Sasisekharan V, Vijayan M (1979) An analysis of side chain conformations in proteins. Int J Pept Protein Res 13:170–184
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H et al (2000) The protein data bank. Nucleic Acids Res 28:235–242. doi:10.1093/nar/28.1.235
28. Dunbrack RL Jr (2002) Rotamer libraries in the 21st century. Curr Opin Struct Biol 12:431–440. doi:10.1016/S0959-440X(02)00344-5
29. Sánchez R, Šali A (1997) Advances in comparative protein-structure modelling. Curr Opin Struct Biol 7:206–214. doi:10.1016/S0959-440X(97)80027-9
30. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications 3. Nat Rev Drug Discov 3:935–949. doi:10.1016/S0959-440X(97)80027-9
31. Davidson E (1993) Molecular mechanics and modeling: overview. Chem Rev 93:2337–2350. doi:10.1021/cr00023a600
32. Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106:1589–1615. doi:10.1021/cr040426m
33. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285:1735–1747. doi:10.1006/jmbi.1998.2401
34. Cole C, Warwicker J (2002) Side-chain conformational entropy at protein-protein interfaces. Protein Sci 11:2860–2870. doi:10.1110/ps.0222702
35. Jorgensen WL, Tirado-Rives J (1988) The OPLS potential functions for proteins. Energy minimization for crystals of cyclic peptides and crambin. J Am Chem Soc 110:1657–1666. doi:10.1021/ja00214a001

36. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118:11225–11236. doi:10.1021/ja9621760
37. Chowdry AB, Reynolds KA, Hanes MS, Voorhies M, Pokala N, Handel TM (2007) An object-oriented library for computational protein design. J Comput Chem 28:2378–2388. doi:10.1002/jcc.20727
38. Hasel W, Hendrikson TF, Still WC (1988) A rapid approximation to the solvent accessible surface areas of atoms. Tetrahedron Comp Methods 1:103–116. doi:10.1016/0898-5529(88)90015-2
39. Juffer AH, Eisenhaber F, Hubbard SJ, Walther D, Argos P (1995) Comparison of atomic solvation parametric sets: applicability and limitations in protein folding and binding. Protein Sci 4:2499–2509
40. Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. Nature 319:199–203. doi:10.1038/319199a0
41. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) Semiempirical free energy force field with charge-based desolvation. J Comput Chem 28:1145–1152. doi:10.1002/jcc.20634
42. Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 28:849–857. doi:10.1021/jm00145a002
43. Boobbyer DNA, Goodford PJ, McWhinnie PM, Wade RC (1989) New Hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. J Med Chem 32:1083–1094. doi:10.1021/jm00125a025
44. Creamer TP (2000) Side-chain conformational entropy in protein unfolded states. Proteins 40:443–450. doi:10.1002/1097-0134(20000815)40:3<443::AID-PROT100>3.0.CO;2-L
45. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. doi:10.1162/153244303322753616
46. Polanski J, Bak A, Gieleciak R, Magdziarz T (2006) Modeling robust QSAR. J Chem Inf Model 46:2310–2318. doi:10.1021/ci050314b
47. Schefzick S, Bradley M (2004) Comparison of commercially available genetic algorithms: GAs as variable selection tool. J Comput Aided Mol Des 18:511–521. doi:10.1007/s10822-04-5322-1
48. Hoskuldsson P (1988) PLS regression methods. J Chemom 2:211–228. doi:10.1002/cem.1180020306
49. Wold S, Sjöström M, Eriksson L (2001) PLS regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130. doi:10.1016/S0169-7439(01)00155-1
50. Sewald N, Jakubke H-D (2002) Peptides: chemistry and biology. Wiley-VCH, Weinheim
51. Madden DR (1995) The three-dimensional structure of peptide-MHC complexes. Annu Rev Immunol 13:587–622. doi:10.1146/annurev.iy.13.040195.003103
52. Peters B, Bui H-H, Frankild S, Nielsen M, Lundegaard C, Kostem E et al (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLOS Comput Biol 2:574–584. doi:10.1371/journal.pcbi.0020065
53. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K et al (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res 1:4. doi:10.1186/1745-7580-1-4
54. Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chain. J Immunol 152:163–175
55. Parker KC, Shields M, DiBrino M, Brooks A, Coligan JE (1995) Peptide binding to MHC class I molecules: implications for antigenic peptide prediction. Immunol Res 14:34–57
56. Hagmann M (2000) Computers aid vaccine design. Science 290:80–82. doi:10.1126/science.290.5489.80
57. Brusic V, Flower DR (2004) Bioinformatics tools for identifying T-cell epitopes. Drug Discov Today BioSilico 2:18–23. doi:10.1016/S1741-8364(04)02374-1
58. Peoples GE, Goedegebuure PS, Smith R, Linehan DC, Yoshino I, Eberlein TY (1995) Breast and ovarian cancer-specific cytotoxic T lymphocytes recognize the same HER2/neu-derived peptide. Proc Natl Acad Sci USA 92:432–436. doi:10.1073/pnas.92.2.432
59. McMichael AJ, Parham P, Brodsky FM, Pilch JR (1980) Influenza virus-specific cytotoxic T lymphocytes recognize HLA-molecules. Blocking by monoclonal anti-HLA antibodies. J Exp Med 152:195–203
60. Parkhurst MR, Fitzgerald EB, Southwood S, Sette A, Rosenberg SA, Kawakami Y (1998) Identification of a shared HLA-A*0201-restricted T-cell epitope from the melanoma antigen tyrosinase-related protein 2 (TRP2). Cancer Res 58:4895–4901
61. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and inerpretation of QSPR models. QSAR Comb Sci 22:69–77. doi:10.1002/qsar.200390007
62. Golbraikh A, Tropsha A (2002) Beware of q^2!. J Mol Graph Model 20:269–276. doi:10.1016/S1093-3263(01)00123-1
63. Aptula AO, Jeliazkovab NG, Schultzc TW, Cronin MTD (2005) The better predictive model: high q2 for the training set or low root mean square error of orediction for the test set?. QSAR Comb Sci 24:385–396. doi:10.1002/qsar.200430909
64. Baroni M, Clement S, Cruciani G, Kettaneh-Wold S, Wold S (1993) D-optimal designs in QSAR. Quant Struct-Act Relat 12:225–231. doi:10.1002/qsar.19930120302
65. de Aguiar PF, Bourguignon B, Khots MS, Massart DL, Phan-Than-Luu R (1995) D-optimal designs. Chemom Intell Lab Syst 30:199–210. doi:10.1016/0169-7439(94)00076-X
66. Zhou P, Tian F, Wu Y, Li Z, Shang Z (2008) Quantitative sequence-activity model (QSAM): Applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids. Curr Comput-Aided Drug Des (in press)
67. McLachlan AD (1982) Rapid comparison of protein structures. Acta Crystallogr A38:871–873
68. Doytchinova IA, Flower DR (2002) Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three-dimensional quantitative structure-activity relationship study. Proteins 48:505–518. doi:10.1002/prot.10154
69. Khan AR, Baker BM, Ghosh P, Biddison WE, Wiley DC (2000) The structure and stability of an HLA-A*0201/octameric Tax peptide complex with an empty conserved peptide-N-terminal binding site. J Immunol 164:6398–6405
70. Madden DR, Garboczi DN, Wiley DC (1993) The antigenic identity of peptide/MHC complexes, a comparison of the conformations of five viral peptides presented by HLA-A2. Cell 75:693–708. doi:10.1016/0092-8674(93)90490-H
71. Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88:486–494. doi:10.2307/2290328
72. DeLano WL (2002) The PyMOL molecular graphics system. DeLano Scientific, San Carlos, CA, USA
73. Doytchinova IA, Flower DR (2001) Toward the quantitative prediction of T-Cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. J Med Chem 44:3572–3581. doi:10.1021/jm010021j
74. Falk K, Rötzschke O, Stefanovic S, Jung G, Rammensee H-G (1991) Allele specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290–296. doi:10.1038/351290a0
75. Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A (1993) Prominent role of secondary anchor residues in peptide binding to

HLA-A*0201 molecules. Cell 74:929–937. doi:10.1016/0092-8674(93)90472-3

76. Sapper MA, Bjorkman PJ (1991) Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. J Mol Biol 219:277–319. doi:10.1016/0022-2836(91)90567-P

77. Madden DR, Garboczi DN, Wiley DC (1993) The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. Cell 75:693–708. doi:10.1016/0092-8674(93)90490-H

78. Sarobe P, Pendleton CD, Akatsuka TD, Engelhard VH, Feinstone SM, Berzofsky JA (1998) Enhanced in vitro potency and in vivo immunogenicity of a CTL epitope from hepatitis C virus core protein following amino acid replacement at secondary HLA-A2.1 binding positions. J Clin Invest 102:1239–1248. doi:10.1172/JCI3714

79. Kubo RT, Sette A, Grey HM, Appella E, Sakaguchi K, Zhu NZ et al (1994) Definition of specific peptide motifs for four major HLA-A alleles. J Immunol 152:3913–3925

80. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8:127–134. doi:10.1093/protein/8.2.127