Critical Review

Prediction of T-Cell Epitope

Hiromichi Tsurui^{1,*} and Takuya Takahashi²

¹Department of Pathology, Juntendo University School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan

²Department of Bioscience and Bioinformatics, College of Information Science and Engineering, Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1, Noji-Higashi, Kusatsu, Shiga 525-8577, Japan

Received July 27, 2007

Abstract. The prevailing methods to predict T-cell epitopes are reviewed. Motif matching, matrix, support vector machine (SVM), and empirical scoring function methods are mainly reviewed; and the thermodynamic integration (TI) method using all-atom molecular dynamics (MD) simulation is mentioned briefly. The motif matching method appeared first and developed with the increased understanding of the characteristic structure of MHC-peptide complexes, that is, pockets aligned in the groove and corresponding residues fitting on them. This method is now becoming outdated due to the insufficiency and inaccuracy of information. The matrix method, the generalization of interaction between pockets of MHC and residues of bound peptide to all the positions in the groove, is the most prevalent one. Efficiency of calculation makes this method appropriate to scan for candidates of T-cell epitopes within whole expressed proteins in an organ or even in a body. A large amount of experimental binding data is necessary to determine a matrix. SVM is a relative of the artificial neural network, especially direct generalization of a linear Perceptron. By incorporating non-binder data and adopting encoding that reflects the physical properties of amino acids, its performance becomes quite high. Empirical scoring functions apparently seem to be founded on a physical basis. However, the estimates directly derived from the method using only structural data are far from practical use. Through regression with binding data of a series of ligands and receptors, this method predicts binding affinity with appropriate accuracy. The TI method using MD requires only structural data and a general atomic parameter, that is, force field, and hence theoretically most consistent; however, the extent of perturbation, inaccuracy of the force field, the necessity of an immense amount of calculations, and continued difficulty of sampling an adequate structure hamper the application of this method in practical use.

Keywords: MHC, matrix, support vector machine (SVM), empirical scoring function, molecular dynamics (MD)

Introduction

When animals of higher orders, that is, higher than *Agnatha*, are invaded by microorganisms, acquired immunity is induced to remove those infectious sources. To perform this activity, the acquired immune system has to distinguish the invaders from the host itself, and this ability is established through ontogeny. The mecha-

nism to discriminate invaders from the self is extremely complicated and still under vigorous investigation. The chief components of the acquired immune system are T cells and B cells, which can be characterized by their clone-specific receptors. A B-cell receptor, that is, an antibody, recognizes the three dimensional structure of the antigen, whereas a T-cell receptor recognizes its antigen as digested peptides, presented on MHC molecules. The production of antibodies by B cells and the cytotoxicity of T cells are sophisticatedly regulated by helper T cells and regulatory T cells; therefore, it is essential for clarifying the mechanism of this regulation to identify T cells at the epitope level. To identify T cells

*Corresponding author. tsurui@med.juntendo.ac.jp

Published online in J-STAGE

doi: 10.1254/jphs.CR0070056

based on their receptors' specificity, an MHC-tetramer was devised (1). An MHC monomer consists of an α -chain, β -chain (or $\beta 2$ microglobulin), and the presented peptide, where one of long chains is enzymatically conjugated with a biotin. Then, four monomers are tetramerized via streptoavidin. To make full use of the MHC tetramer, it will be necessary to determine the sequence of the presented peptide based not only on experimental data but on appropriate prediction methods. From the viewpoint of application in medicine, designing effective vaccines is of extreme importance, and this can be achieved through reliable predictions of the T-cell epitopes.

Here we review the recent progress on methods to predict T-cell epitope, from the rather naïve motif-matching method to sophisticated molecular dynamics (MD)-based methods. The strictest definition of an epitope is whether the corresponding T cells proliferate and take part in the immune response. The establishment of acquired immunity requires that the epitopes are properly presented as peptides on MHC molecules, and the pertinent T cells are present. As shown in Figs. 1 (2–6) and 2 (7–13), there are many intracellular enzymatic steps and inter-cellular interaction for the presentation of T-cell epitope and maturation of corre-

sponding T cells; therefore, it is unrealistic to take these steps into consideration. We do not deal with these factors in this review but describe the methods to predict MHC-binders.

Structure of peptide-MHC complex

1. T-cell receptor peptide-MHC complex

The T-cell receptor (TCR) is a member of the immunoglobulin super family and is quite similar to the F_{ab} fragment of IgG; however, it recognizes its antigen not in a soluble form, but in a complex form, as presented on the MHC molecule (Fig. 3: A and B). The binding mode is common to any TCR-peptide-MHC (TCR-pMHC) complex irrespective of class I and class II (14–17). In the case of class I, the $\nu\alpha$ domain of TCR recognizes the N-terminal side of the peptide with CDR1 and 3, and the α -helix of the $\alpha 2$ domain of MHC with CDR2; and the $\nu\beta$ domain of TCR recognizes the C-terminal side of the peptide with CDR1 and 3, and the α -helix of the $\alpha 1$ domain with CDR2. In the case of class II, the $\nu\alpha$ domain recognizes the N-terminal side of the peptide with CDR1 and 3, and the α -helix of the $\beta 1$ domain with CDR2; and the $\nu\beta$ domain recognizes the C-terminal side of the peptide with CDR1 and 3, and

Fig. 1. Pathway of antigen presentation through MHC molecules. A: Cytoplasmic proteins derived from self components or invasive microbes such as virus and bacteria are digested by proteasome, transported into the endoplasmic reticulum by TAP, and held firmly by MHC molecules. Chaperons such as calnexin, tapasin, and calreticulin are involved in this process. The peptide-MHC complex is sorted to the plasma membrane through the Golgi apparatus. Specificity of proteasome digestion and TAP transformation affect the sequence of peptides bound to the MHC molecule. B: External antigens derived from apoptotic cells and invasive microbes are engulfed and digested with cathepsins. MHC class II molecules are synthesized in the endoplasmic reticulum, being protected by the invariant chain (Ii), sorted to early endosomes, disarmed of Ii by cathepsin S, and encounters digested antigen peptides in late endosomes, where CLIP on MHC is replaced. The peptide-MHC complex is transferred to the plasma membrane.

Fig. 2. T-cell development in thymus. Bone marrow-derived lymphoid progenitor is developed through several steps; and recombination, editing, and expression of TCR are strictly regulated at each step (7, 8). In a double-positive step in the cortex, thymocytes are selected based on whether they interact moderately with MHC molecules presenting self-antigens (9–11) or not. CD4+ or CD8+ single-positive cells in the medulla interacting strongly with MHC molecules expressing self-antigens on dendritic cells and epithelial cells are depleted (12, 13). As the results of those developmental paths, a T-cell repertoire is formed.

the α -helix of the $\alpha 1$ domain with CDR2.

2. Structure of peptide-MHC complex

The MHC class I molecule consists of an α -chain and a $\beta 2$ microglobulin. The α -chain forms $\alpha 1$, $\alpha 2$, and $\alpha 3$ domains and has a transmembrane domain and a short intra-cellular portion. The class II molecule consists of an α -chain forming $\alpha 1$ and $\alpha 2$ domains and a β -chain forming $\beta 1$ and $\beta 2$ domains. Both α - and β -chains have transmembrane domains and short intracellular portions. The MHC molecule firmly holds a peptide in a groove formed with two α -helices and a β -sheet floor and presents the peptide to T cells. Peptides are fastened in the groove with hydrogen bonds (H-bonds), electrostatic interaction, and van der Waals interaction as shown in Fig. 3, C and D. H-bonds are formed between the peptide backbone and MHC-molecule, and networks around the N- and C-terminals are conspicuous in the case of class I. In some cases, an H-bond between the

peptide residue and MHC is formed (Fig. 3D, lower right). Pockets (typically positions 1, 4, 6, and 9) formed in the groove preferentially accept particular residues such as hydrophobic, basic, or acidic, depending on the amino acids (AAs) that form the pockets. Peptides that bind to MHC class I molecules are usually 8- to 11-AAs-long; the majority are 9 AAs (18). Peptides that bind to MHC class II molecules are 12- to 25-AAs-long and extend beyond the MHC groove, but contain a minimum 9-mer core (19). The H-bond network and pocket-residue interaction precisely determine the configuration of the presented peptide. All the panels are drawn using the PDB files indicated in the lower-left corner of each panel by Swiss-PDBViewer (20) and rendered by POV-Ray v3.6 (<http://www.povray.org/>).

Prediction methods

Table 1 shows the representative methods to predict the MHC-binders or T-cell epitope. The motif-matching and matrix methods are applied specifically to an MHC-peptide interaction, whereas other methods can be applied to general ligand-protein interactions. Motif matching, matrix method, and support vector machine (SVM) use only experimental binding data and do not require structural data of the MHC molecule, whereas the empirical scoring function, Monte Carlo (MC), and MD use only structural data and a general atomic parameter, that is, force field (FF); however, the latter three methods have not achieved a level of practical use in accuracy or calculation efficiency yet. These points are discussed in the following sections.

1. Motif matching

As shown in Fig. 3, E and F, a peptide bound on the MHC molecule usually has residues that are trapped in holes aligned in the groove. These holes, called pockets, preferentially accept residues with particular characteristics such as appropriate size, hydrophobicity, and electric charges; and consequently, the MHC molecule binds the peptide with the sequence satisfying the combination of the pockets' preference. For example, Table 2 lists the known binders to HLA-DR1 (21–23), which are aligned according to the N-terminal bulky aromatic or aliphatic AA residue. Newly obtained self-peptide 3 was aligned according to the putative anchors (21). Rudensky et al. purified MHC-bound peptides, determined their sequences (24), and proposed binding motifs for I-A^s, I-A^b, and I-E^b (25). Based on these proposals, Cole et al. predicted the T-cell epitopes of Sendai Virus M protein (26). Hammer et al. identified several anchors of HLA-DRB1*0101, 1101, and 0401 using the M13 phage peptide display system (27–29).

Fig. 3. Structure of presented peptide, MHC, and T cell receptor. A: The complex between TCR A6 and human Class I MHC HLA-A2 with the modified HTLV-1 TAX (Y5K-4-[3-Indolyl]-butyric acid) peptide. TCR is a member of the immunoglobulin super family and is quite similar to the Fab fragment of IgG. TCR recognizes a peptide presented in the groove of the MHC molecule together with the complex. The polymorphic portion of the α -helix is the physical substance of MHC restriction. The MHC class I consists of an α -chain and β 2 microglobulin. B: A complex of a human T cell receptor, influenza HA antigen peptide, and MHC class II molecule, HLA-DR1. MHC class II molecule consists of an α -chain and β -chain, both having 2 domains, α 1 and α 2 and β 1 and β 2, respectively. α 1 and β 1 domains form a groove that binds a peptide of 9 AAs to upward of ten AAs. The relation among TCR, peptide, and MHC molecule is quite similar to that of class I molecule. C: Complex of MHC class I H-2 D^b molecule with peptide HA₄₆₈₋₄₇₇ from influenza A virus. The peptide is located in the groove formed with 2 α -helixes and the floor of the β -sheet. The peptide is stabilized through hydrogen bonds (H-bonds, thick green dotted line) between its back bone and MHC molecule. Both amino- and carboxy-terminals are blocked with tyrosine residues of the α -helix. H-bond networks of these ends are characteristic of MHC class I. D: Complex of class II MHC molecule I-A^b bound to a peptide from I-E α -chain. The relation between the peptide and the MHC molecule is similar to that of class I. H-bond between the backbone of the presented peptide and helix is also obvious. In this complex, the H-bond network between Asp at P7 and β -chain contributes to stabilize the complex (lower right, blue: Asp at P7, Red: Lys of β -chain, Yellow: Tyr of β -chain). Note that the ends of the peptide are open, making the length of the bound peptide indeterminate. E: The same complex as panel C is shown with the surface of the MHC molecule expressed as yellow mesh. In this complex, residues corresponding to P1 and P9 almost occupy the pockets. F: The same complex as panel D is shown with the surface of the MHC molecule expressed as yellow mesh. In this complex, the rather large pockets of 1, 4, 6, and 9 are occupied with small alanine residues; therefore, the peptide is stabilized mainly through the H-bond between the peptide back bone and α -helix of the MHC molecule. Pocket-residue-specific interaction is unimportant; however, such a case is quite exceptional. All panels are drawn with Swiss-PDBViewer using the PDB data indicated at the lower-left and rendered by POV-Ray v3.6.

Table 1. Methods to estimate binding affinity between peptides and MHC molecules

Method	General physical data	MHC-specific data	Model	Mathematical procedure	Calculation efficiency	Accuracy of prediction
Motif matching	not used	eluted peptide sequence	residue-specific interaction pattern matching	pattern matching	extremely high	low
Matrix	not used	experimental peptide binding	residue-specific interaction linearity	multiplication with matrix	high	high
SVM	not used	experimental peptide binding	SVM	data encoding kernel optimization	high	high
Empirical scoring function	dielectric constant atomic size atomic charge	crystallography experimental peptide binding	continuous dielectric substance	solving Poisson-Boltzmann eq., calculating ASA	high	high (with regression to experimental data)
Thermodynamic integration using molecular dynamics	atomic parameter force field	crystallography	classical mechanics thermodynamic integration	numerical integration of motion equation	low	restricted to small perturbation

Table 2. Alignment of DR1-restricted T cell epitopes and candidate peptide (self-peptide 3)

T cell epitope	Proteinsource
K K I A K M E K A S S V F	Cicum sporozoite protein (381 – 393)
E V W R E E A Y H A A	Ragweed (51 – 61)
Q Y I K A N S K F I G I T	Tetanus toxin (830 – 842)
E A L V R Q G L A K V A	Ribonuclease (101 – 112)
R V K R G L T V A V A G	<i>M. tuberculosis</i> lpqH (3 – 14)
A K L A K L A G G V A V I K	<i>M. leprae</i> GroEL (412 – 425)
G P L K A E I A Q R L E	Influenza matrix protein (18 – 29)
P K Y V K Q N T L K L A T	Influenza hemoagglutinin (307 – 319)
A I L E F R A M A Q F S R K T D	Self-peptide SP3
i	Relative anchor position
	i + 8

This table is made based on Kropshofer et al. (ref. 21).

Binding motifs for various human and mouse MHC class I and class II molecules were reported (18). Binding motifs for the most common MHC class I molecules are relatively well-defined (30). On the other hand, binding motifs for some MHC class II molecules have been shown to be promiscuous (31 – 36). This method is simple and can be intuitively understood; however, it has become outdated due to the inadequate accuracy of prediction.

2. Matrix method

The motif-matching method pays attention to anchors, that is, key pockets of the groove, and checks the sequence with the distribution of particular AAs as to whether it satisfies the anchors' preference. Main pockets are typically located at positions 1, 4, 6, and 9,

depending on the case. The matrix method assigns scores for 20 AAs to all the positions of MHC grooves, and this set of scores is called the pocket profile; therefore, this method can be considered as the generalization of the motif-matching method. The affinity of the peptide itself is calculated as the sum of each AA's contribution, and this feature of "independence and additivity" of each position, which is a specific characteristic of the peptide-MHC interaction, is a good approximation. The affinity of a peptide to an MHC is calculated simply by applying this matrix to its sequence, as shown in Fig. 4. Figure 4A shows an additive form (37) and Fig. 4B, a multiplicative form (38). To obtain these matrix data, a large number of binding experiments or sequencing of pooled peptides is necessary, and the strategies to be taken are more

systematic and consistent than that for motif matching. Parker et al. used 154 synthesized peptides and obtained a matrix for HLA-A2 (38). Kubo et al. used not only AA sequencing of pooled peptide but polyalanine-analogue binding to refine motifs for HLA-A1, 3, 11, and 24 (39). Davenport et al. obtained a scoring matrix of HLA-DRB*1301 by sequencing the pool peptide and determining the composition of AA at each position (40). Udaoka et al. determined matrix data for H2-K^b, -D^b, and -L^d by positional scanning using combinatorial peptide libraries (41, 42). Applying a genetic algorithm, Brusic

et al. estimated the sample size to obtain a matrix with given accuracy (43). A number of binding matrices have been defined for MHC class I (44 – 46) and class II (47). Using the proliferative response of clonotype-specific T-cells as a monitor of the pocket profile, Zhao et al. directly evaluated the score of peptides as candidates for the T-cell epitope, not the binding affinity to MHC (48). This method can be rationalized by the assumption that the effect of T-cell stimulation by a peptide presented on a particular MHC molecule is the summation of the effect of each AA (49). Taking into account the structure of the stretched peptide and arrayed pockets in the groove, the additivity of each residue to peptide-MHC binding seems quite reasonable. On the other hand, additivity in T-cell stimulation is not so obvious and will require a wider spectrum of investigation. In the matrix method, the obtained pocket profiles and the matrix are considered specific to each MHC molecule. A pocket profile is common irrespective of the haplotype if the residues that form the pocket are the same (Fig. 5). Sturniolo et al. hit upon this idea, and they proposed that the matrix for any MHC is obtained by combining the pocket profiles of composite pockets (50).

Many methods to predict the T-cell epitope are available through websites, and many use matrix methods: BIMAS (http://www.bimas.cit.nih.gov/mol-bio/hla_bind/), SYFPEITHI (<http://www.syfpeithi.de/Scripts/MHCServer.dll/FindYourMotif.htm>), MHCpred (<http://www.jenner.ac.uk/MHCpred/>), ProPred (<http://www.imtech.res.in/raghava/propred1/matrices/>), and EpiPredict (<http://www.epipredict.de/Prediction/prediction.html>) promptly answer inquiries with ranked candidate sequences.

Fig. 4. Calculation of affinity score. A: Calculation of affinity score based on a matrix of the additive form. When sequence FNRIKNNVA is given, the first AA F is located at P1 and the corresponding score is 0. This MHC molecule (DRB1*0401) assumes that the first AA is aromatic or aliphatic; and for aromatic AA, i.e., F, W, and Y, assigns 0. For aliphatic AA, i.e., I, L, M, and V, it is assigned -1. When another type of AA is located at P1, the peptide is not considered to bind to the MHC molecule. Likewise, the scores corresponding to the second N and third R are 0.8 and 0.7, respectively. Total score of the sequence is the sum of these scores, i.e., 0 + 0.8 + 0.7 + 0.8 + 0.3 + 1.7 - 0.1 + 0.4 + 0 = 4.6. This figure is drawn based on Hammer et al. (37). B: Calculation of the affinity score based on a matrix of the multiplicative form. When the sequence GILGFVFTL is given, the first AA G is located at P1 and the corresponding score is 0.578. Likewise, the scores corresponding to the second I and the third L are 10.151 and 3.685, respectively. The total score for the peptide can be obtained by multiplying all the terms and taking log_e to the product, i.e., Log (0.578 × 10.151 × 3.685 × 1.000 × 6.044 × 2.588 × 6.383 × 1.000 × 2.357) = 8.519. This figure is drawn based on Parker et al. (38). The unit of the score is arbitrary.

Fig. 5. Showing the relation between motif matching, matrix, and TEPITOPE. The motif matching method focuses on key pockets, i.e., anchors, and residues that satisfy the preference of the anchor. The matrix method numerically evaluates the preference of all the pockets, i.e., assigns scores to 20 AAs (pocket profile), and determines the affinity of the peptide as the sum of each score. TEPITOPE considers that the pocket profile is the same when AAs forming the pocket are the same and constructs the matrix as the combination of these pocket profiles.

3. Perceptron and SVM method

The SVM method is a new learning method introduced by Vapnik (51). It is well-founded in terms of computational learning theory and very open to theoretical understanding and analysis. Consider the case of classifying the data of $\mathbb{R}^d$ into two classes by a hyperplane. Let w be a weighting factor vector in d -dimension space, i.e. $w \in \mathbb{R}^d$. Perceptron is a linear threshold element that can be expressed as $y = \text{sgn}[f_w(x)]$, where sgn is the sign of the argument, i.e., $y = 1$ or -1 , and $f_w(x) = w \cdot x$. Perceptron corrects its weighting factor w only when samples output an error, and corrections are made as $\Delta w \propto y_i x_i$. If the data can be linearly separable, w should converge within finite steps,

shown as $w = \sum_{i=1}^n \alpha_i x_i$, with the initial condition $w = 0$;

and all the data can be classified by this plane.

When the data are not linearly separable (Fig. 6A), they can be made linearly separable by transforming into the h -dimensional feature space, ($h > d$), i.e., $\phi(x) \in \mathbb{R}^h$; and the input/output function can be expressed as $f_w(x) = w \cdot \phi(x)$, where $w \in \mathbb{R}^h$. By applying Perceptron to those data, a hyperplane can be obtained that can classify the transformed data (Fig. 6B). In this case, error corrections are made in the same way, that is, $\Delta w \propto y_i \phi(x_i)$; and the final weighting factor is

$w = \sum_{i=1}^n \alpha_i \phi(x_i)$ with the initial condition of $w = 0$.

Substituting this relation into the input/output function, an expression using the transforming function is obtained as

$$y = \text{sgn}[\sum_{i=1}^n \alpha_i \phi(x_i) \cdot \phi(x)]$$

Let the inner product $\phi(x_i) \cdot \phi(x)$ be $k(x_i, x)$, i.e., $\phi(x_i) \cdot \phi(x) = k(x_i, x)$, and the input/output function can be written as

$$y = \text{sgn}[\sum_{i=1}^n \alpha_i k(x_i, x)]$$

This means the input/output function can be written as a linear combination of some kernel functions.

If arbitral function $k(x, y)$ may work for transformation into the feature space and for calculation of the inner product, it will be very convenient. Mercer's theorem says that the necessary and sufficient condition in which $k(x, y)$, a function of $x, y \in \mathbb{X}$, can be written as

$$k(x, y) = \sum_{j=1}^n \gamma_j \phi_j(x) \cdot \phi_j(y), \quad \gamma_j \geq 0,$$

where $\int_{\mathbb{X}} k(x, y) \phi_j(x) dx = \gamma_j \phi_j(y)$,

Fig. 6. SVM is a Perceptron in a feature space of higher dimension. A: Consider the case where original data can not be classified with a line. When data are transferred into a feature space of higher dimension, they can be classified by a plane with some margin (B). The curve in panel A is the projection of the discriminating plane in panel B. SVM has the largest margin in this transferred feature space. Support vector indicates data with the minimum distance from the discriminating plane.

is that k is a symmetric function on x and y , i.e., $k(x, y) = k(y, x)$, and is positive semidefinite, i.e., for any f ,

$$\int_{\mathbb{N}} \int_{\mathbb{N}} k(x, y) f(x) f(y) dx dy \geq 0$$

(Mercer's condition)

In this formula, the sample space is generalized from $\mathbb{R}^d$ to $\mathbb{N}$, which includes a letter string. The kernel that satisfies Mercer's condition is called the Mercer Kernel. The representative kernel function is as follows:

Sigmoidal kernel: $k(x, y) = 1 / (1 + \exp(-\beta x \cdot y))$

Radial Base Function (RBF) kernel:

$$k(x, y) = \exp(-\beta \|x - y\|^2)$$

Polynomial kernel: $k(x, y) = (x \cdot y + c)^p$.

The distance between a datum and a hyperplane can be clearly determined, and the margin of the hyperplane for the data can be defined as the minimum value of the distance from both classes; therefore, the margin can be defined for any hyperplane. Perceptron provides a discriminating hyperplane. To maximize the ability to classify the target data (test data), that is, to maximize the generalizing capability, the margin should be maximized. This problem can be replaced with a minimization of $\|w\|^2$ under the condition that $y_i w \cdot \phi(x_i) \geq 1$. This minimization can be accomplished by Lagrange's method of undetermined multipliers; therefore, to be precise, SVM is the result of training for optimization. SVM can make any data linearly separable by transferring them into a feature space of higher dimension; furthermore, it has the largest generalizing capability. SVM means a hyperplane with the maximum margin, and the points that give the minimum distance to the hyperplane are called support vectors. Many software

packages for kernel machine construction are available on websites: SVM^{light} (<http://svmlight.joachims.org/>) (52), SVM^{struct} (http://svmlight.joachims.org/svm_struct.html), LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>), mySVM (<http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html>), SVM in R (<http://cran.r-project.org/src/contrib/Descriptions/e1071.html>), and so on. A personalized SVM can be constructed by using those packages and training data. Figure 7 shows the data flow in SVM. To apply SVM to classify peptide binders or non-binders to the MHC molecule, the AA sequence should be encoded to numerical data (Fig. 8a). The most frequently used encoding is "distributed encoding" (Fig. 8b). This encoding is general but rather perfunctory. Table 3 shows representative SVMs used to predict the MHC-binder and T-cell epitope. Dönnies and Elofsson developed an SVM, that is, SVMHC (53), using this conversion and SYFPEITHI (54) and MHCPEP (30) databases. The type of kernel can be optimized and varied according to the haplotype, among linear, polynomial, and RBF. They showed that the performance of SVMHC is a little better than those of SYFPEITHI and HLA_BIND (38). Bhasin and Raghava developed SVM to predict the HLA-DRB1*0401 binding peptide using this conversion and the MHCBN database (55). They used SVM^{light} and the optimized kernel was RBF. The accuracy of the machine attained about 86%. Interaction between the peptide and MHC molecule occurs based on physicochemical forces such as electrostatic, van der Waals, and so on; therefore, if the conversion of the sequence to numeric data reflects the physicochemical features of the AA involved, the accuracy of the predic-

Fig. 7. Simplified procedure to predict target data using a support vector machine (SVM). In this case, pre-trained SVM is provided with already classified 5-dimensional (from a to e) k-points as training data and is trained, i.e., determines the discriminating plane with the largest margin. After training, target data of the same dimension are given, classified by optimized SVM, and class labels are predicted. In this figure, data scaling and optimization of kernel parameters are omitted for simplicity.

Scores used in the 11-factor encoding for 20 amino acids, after scaling to the range [0,1]

Amino Acid	Steric parameter	Hydrogen bond donors	Hydrophobicity scale	Hydrophilicity scale	Average accessible surface area	van der Waals parameter RI	van der Waals parameter epsilon	Free energy of solution in water	Average side chain orientation angle	Polarity	Isoelectric point
A	0.510	0.169	0.471	0.279	0.141	0.294	0.000	0.262	0.512	0.000	0.404
R	0.667	0.726	0.321	1.000	0.905	0.529	0.327	0.169	0.372	1.000	1.000
N	0.745	0.390	0.164	0.658	0.510	0.235	0.140	0.313	0.116	0.065	0.330
D	0.745	0.304	0.021	0.793	0.515	0.235	0.140	0.601	0.140	0.956	0.000
C	0.608	0.314	0.760	0.072	0.000	0.559	0.140	0.947	0.907	0.028	0.285
Q	0.667	0.531	0.178	0.649	0.680	0.529	0.140	0.416	0.023	0.068	0.360
E	0.667	0.482	0.092	0.883	0.602	0.529	0.140	0.561	0.163	0.960	0.056
G	0.000	0.000	0.275	0.189	0.103	0.000	0.000	0.240	0.581	0.000	0.401
H	0.686	0.554	0.326	0.468	0.402	0.529	0.140	0.313	0.581	0.992	0.603
I	1.000	0.650	1.000	0.000	0.083	0.824	0.308	0.424	0.930	0.003	0.407
L	0.961	0.650	0.734	0.081	0.138	0.824	0.308	0.463	0.907	0.003	0.402
K	0.667	0.692	0.000	0.568	1.000	0.529	0.327	0.313	0.000	0.952	0.872
M	0.765	0.612	0.603	0.171	0.206	0.765	0.308	0.405	0.814	0.028	0.372
F	0.686	0.772	0.665	0.000	0.114	0.853	0.682	0.462	1.000	0.007	0.339
P	0.353	0.372	0.012	0.198	0.411	0.588	0.271	0.000	0.302	0.030	0.442
S	0.520	0.172	0.155	0.477	0.303	0.206	0.000	0.240	0.419	0.032	0.364
T	0.490	0.349	0.256	0.523	0.337	0.235	0.140	0.313	0.419	0.032	0.362
W	0.686	1.000	0.681	0.207	0.219	1.000	1.000	0.537	0.674	0.040	0.390
Y	0.686	0.796	0.591	0.477	0.454	0.853	0.682	1.000	0.419	0.031	0.362
V	0.745	0.487	0.859	0.036	0.049	0.647	0.234	0.369	0.674	0.003	0.399

for sequence "DHW" is converted to (0.745, 0.304, 0.021, 0.793, 0.515, 0.235, 0.140, 0.601, 0.140, 0.956, 0.000, 0.686, 0.554, 0.326, 0.468, 0.402, 0.529, 0.140, 0.313, 0.581, 0.992, 0.603, 0.686, 1.000, 0.681, 0.207, 0.219, 1.000, 1.000, 0.537, 0.674, 0.040, 0.390) (11 x 3 dimensional vector)

C

Fig. 8. Encoding of sequence data to numerical data. A: Original sequence data are converted (encoded) to numerical data and applied to the SVM. B: The most simple encoding is "distributed", which has 1 at the respective position for each AA. This encoding is general but rather perfunctory and does not reflect the physical character of AAs. C: Scores consisting of 11 independent physicochemical characters of each AA are used to encode the peptide sequence into numerical data. This encoding reflects the physical nature, resulting in much better accuracy of prediction. Sequence "DHW" is converted to 33-dimensional numerical data.

Table 3. SVMs used for prediction of MHC-binder or T-cell epitope

Authors	Target	Data set	Encoding	Machine	Kernel	Performance
Dönnies & Elofsson (53)	HLA-A, B	MHCPEP, SYFPEITHI	distributed	SVM ^{Light}	lin, poly rbf	better than MHCPEP, SYFPEITHI
Zhao et al. (61)	HLA A0201 T cell epitope	Melan A26-35 derivedsynthetic peptide classified by clone LAU203-1.5	Kidera's factors	SVM ^{Light}	poly	sensitivity = 0.763, PPV = 0.716 (c.v.) sensitivity = 0.389, PPV = 0.451 (c.v.) (SVMHC trained with MHCPEP)
Bhasin & Raghava (55)	HLA-DRB1 0401	MHCBN	distributed	SVM ^{Light}	rbf	accuracy = 86.1*
Liu et al. (57)	H2-K ^k , D ^b , K ^b	MHCPred	11 factors from AA index	LIBSVM	rbf	A _{ROC} = 0.738 (H2-D ^b)** A _{ROC} = 0.834 (H2-K ^b)
Cui et al. (59)	HLA Class I (18) Class II (12)	MHCBIN, SYFPEITHI	distributed	not referred	rbf?	accuracy = 70 – 90 for binder much better than previous report for non-binder

*Accuracy is the percent of correct prediction. **A_{ROC} is an area below the ROC curve.

tion will be improved. Liu et al. proposed parameters that characterize the physicochemical aspects of each AA based on the AA index (56). They constructed an SVM (SVRMHC) predicting the quantitative affinity of peptides to MHC molecules using the AntiJen database (57, 58). The SVRMHC model construction was carried out using locally developed C and Matlab programs, where LibSVM was used for SVR-related implementation (<http://www.csie.ntu.edu.tw/~cjlin/>). Cui et al. showed high performance SVM for a variety of HLA alleles, that is, 18 class I and 12 class II, by using 3,252-4,208 binders and 168,793-234,333 non-binders, and predicted newly reported epitopes with high accuracy (11/15) (59).

The performance of SVM strongly depends on the quality and quantity of the training data. In most SVMs, the training data and the prediction are dichotomic, binder or non-binder. However, the definition of them varies from case to case. Therefore, it is difficult to give a consistent meaning to the performance among these SVMs. SVRMHC is an attempt to establish a quantitative model, not a qualitative one. It uses the AntiJen database, in which each binding experiment was represented as a (sequence: pIC₅₀) pair in the dataset. The prediction is also given as the pIC₅₀ value for query sequences. This model, even though its performance is still affected by the quality of experimental data, has clearness and consistency in itself.

Kidera et al. obtained 10 physical properties of 20 AAs from 188 physical properties through cluster analysis and factor analysis (60). Zhao et al. developed an SVM to predict T-cell epitope using these properties for encoding the AA sequence (Fig. 8C) (61). They used

peptide libraries in positional scanning format (PS-SCL) of Melan-A26-35 epitope characterized by the LAU203-1.5 T-cell clone (62). The performance of the method was quite high, indicating the validity of applying this method to predict T-cell epitopes. T-cell epitopes are not always high-affinity MHC binders, and high MHC binders are not always T-cell epitopes. Zhao's SVM showed much better performance than SVMHC trained with the SYFPEITHI database or MHCPEP database, demonstrating not only the distinction between T-cell epitopes and MHC-binders, but also the versatility of SVM.

4. Empirical scoring function method

There have been trials to express the free-energy change accompanying protein-ligand binding as the sum of independent energetic terms. In order to model the energetic terms, there are two classical and physically meaningful models, that is, an electrostatic continuum model and an all-atom model (63). The electrostatic continuum model is applied to deal with long range electrostatic interactions, including Coulomb energy *in vacuo* and solvation energy. In addition to these two methods, a simple and straightforward treatment was developed that expresses the interaction energies as specific atomic pair-wise functions such as hydrogen-bonding, ionic interaction, and lipophilic interaction. This model assigns an empirical and arbitrary interaction to these pairwise functions. Based on the physicochemical aspects, the view points that lie at the base of these models can be summarized as follows:

i) Whether to consider the structural entropy, that is, adopting structure sampling or adding entropic terms, or

not, that is, adopting only single crystal structure;

ii) How to deal with the contribution of solvent, to describe it precisely with simulations such as explicit MD and RISM, or to approximate it with electrostatic term and entropic term;

iii) How to sample structures.

These factors are not independent, and in this review, taking into account those issues, we classified the formulae into 3 groups:

Group I: System is described as the sum of component energy terms expressed as specific atomic pairwise functions such as hydrogen-bond, ionic interaction, lipophilic interaction, and so on (64–66);

Group II: System is described mainly as continuum model with the addition of some pairwise functions (67–71);

Group III: Combination of conformational sampling and continuum model. Here, structures of protein and ligand molecules are treated as flexible and energy terms are calculated by averaging the energies for sampled molecular conformations with MD and MC (72–74).

Formulae 1 shows representative empirical scoring functions classified into these 3 groups. Formulae within the same group are quite similar to each other except for some slight differences. Formulae 2 compares 3 formulae from each group, indicating this correspondence is also seen beyond the group. Figure 9 shows the derivational relation of the formulae in Formulae 1 based on this classification and the similarity of component terms. The pairwise functions, as graphed in Formulae 1 (upper right), are the functions of the distance of paired atoms and the deviation from the standard angle; and they directly evaluate the energy of atomic pairs using a PDB file. The continuum model, usually assumes the relative dielectric constant of water to be 80 and that of protein and ligand, 2–4; and it describes the system by solving the Poisson-Boltzmann equation using simulation engines such as DelPhi (finite difference method) (75) and APBS (finite element method) (76) and using PDB coordinates to determine charge distribution and electrostatic energy. The accessible surface area (ASA) is the surface area of a biomolecule (protein, ligand, etc.) that is accessible to a solvent and is useful to simply evaluate the surface free energy of a molecule. ASA is first calculated by the Lee and Richards method (77), whereas the solvent-exclude surface is calculated by the Connolly method (78). One of the most simple and understandable expressions for the change of free energy in a protein-ligand binding (for example, MHC-peptide binding) is $\Delta G_{bind} = \Delta G_{el} + \Delta G_{np} - T\Delta S_{conf}$, where ΔG_{el} represents the change of electrostatic energy consisting of Coulomb energy and solvation terms, i.e., $\Delta G_{el} = \Delta G_{coul} + \Delta G_{solv}$. ΔG_{np} repre-

sents the change of the nonpolar interaction and is thought to be proportional to the change of the accessible surface ASA, i.e., $\Delta G_{np} = \gamma_{aw} \cdot \Delta A$, where $\Delta A = A_{complex} - A_{MHC} - A_{peptide}$ (79). The typical value of γ_{aw} is 10.45 to 25 cal/mol/Å². Although the hydrophobic effect is not so prominent for the binding of soluble molecules such as crystallization compared with folding and strong binding processes (80), the effect is not negligible and affects the binding strength. Therefore, methods considering hydrophobic effects are still necessary for MHC-peptide interaction where both ligand and acceptor exist as soluble molecules. ΔS_{conf} reflects the change of the configuration space accompanying the protein-ligand (MHC-peptide) binding. Figure 10 illustrates the thermodynamic process to calculate the free energy change of a protein-ligand binding, from A to B, based on the continuum model. Considering the path through A, C, D, and B, the term ΔG_{el} and ΔG_{np} can be calculated independently. The path D→B minus path C→A gives ΔG_{el} and path C→D gives ΔG_{np} . The term $T\Delta S_{conf}$ is not included in this figure. Even though this formula is quite understandable, it is not accurate. In principle, the change of binding free energy can not be described as the sum of the elemental terms, and those terms inevitably contain some errors. Empirical function methods calculate binding free energy using atom-coordinates, and therefore the obtained values are affected by the resolution of data. As structures of molecules in solution fluctuate, so does the binding energy between them. Therefore, it is quite reasonable that the estimation averaged along the trajectory of molecular dynamics simulation would be more accurate than that based on a single crystallographic structure. Based on this idea, models, which are the combination of explicit solvent molecular dynamics simulation with implicit solvation, named molecular mechanics Poisson-Boltzmann Surface Area (MM-PBSA) (73) and molecular mechanics Generalized Born Surface Area (MM-GBSA) (74) were devised. MM-GBSA utilizes the Generalized Born approximation, which is much more efficient for evaluating hydration energy than solving the Poisson-Boltzmann equation, and still provides fairly good results. Linear Interaction Energy (LIE) was originally developed to calculate electrostatic and van der Waals potentials on an all-atom model with the explicit dealing of solvent (72) and then extended to the continuum model (81). These methods are considerably slower than scoring of single conformations, but faster than rigorous thermodynamic integration (TI) calculations (82, 83), and provide more precise prediction than scoring of single conformation methods do (73, 83–88). DynaPred also uses MD to calculate the physico-chemical features of each AA against each pocket of a

Group I

Böhm (64)
$$\Delta G_{\text{bind}} = \Delta G_0 + \Delta G_{\text{hb}} \sum_{\text{h-bond}} g(\Delta r, \Delta \alpha) + \Delta G_{\text{ionic}} \sum_{\text{ionic int}} g(\Delta r, \Delta \alpha) + \Delta G_{\text{lipo}} |A_{\text{lipo}}| + \Delta G_{\text{rot}} \text{NROT}$$

$$g(\Delta r, \Delta \alpha) = g_1(\Delta r)g_2(\Delta \alpha)$$

$$g_1(\Delta r) = \begin{cases} 1 & \text{if } \Delta r < 0.2 \text{ \AA} \\ 1 - (\Delta r - 0.2)/0.4 & \text{if } 0.2 \text{ \AA} < \Delta r < 0.6 \text{ \AA} \\ 0 & \text{if } \Delta r > 0.6 \text{ \AA} \end{cases}$$

$$g_2(\Delta \alpha) = \begin{cases} 1 & \text{if } \Delta \alpha < 30 \\ 1 - (\Delta \alpha - 30)/50 & \text{if } 30 < \Delta \alpha < 80 \\ 0 & \text{if } \Delta \alpha > 80 \end{cases}$$

Eldridge et al. (65)
$$\Delta G_{\text{bind}} = \Delta G_0 + \Delta G_{\text{hbond}} \sum_i g_1(\Delta r)g_2(\Delta \alpha) + \Delta G_{\text{metal}} \sum_{\text{aM}} f(r_{\text{aM}}) + \Delta G_{\text{lipo}} \sum_{\text{IL}} f(r_{\text{IL}}) + \Delta G_{\text{rot}} H_{\text{rot}}$$

Wang et al. (66)
$$\Delta G_{\text{bind}} = \Delta G_{\text{vdw}} + \Delta G_{\text{H-bond}} + \Delta G_{\text{deformation}} + \Delta G_{\text{hydrophobic}} + \Delta G_0$$

Group II

Froloff et al. (67)
$$\Delta G_{\text{bind}} = \Delta G_{\text{el}} + \Delta G_{\text{np}} + \Delta G_{\text{strain}} - T\Delta S_{\text{mc}} - T\Delta S_{\text{sc}} - T\Delta S_{\text{tr}}$$

$$\Delta G_{\text{el}}(\epsilon_i, \epsilon_o) = \Delta G_{\text{coul}}(\epsilon_i) + \Delta G_{\text{solv}}(\epsilon_i, \epsilon_o)$$

$$\Delta G_{\text{np}} = \gamma_{\text{aw}} \Delta(\text{CA})$$

Böhm (68)
$$\Delta G_{\text{bind}} = \Delta G_0 + \Delta G_{\text{polar}} + \Delta G_{\text{apolar}} + \Delta G_{\text{solv}} + \Delta G_{\text{flexi}}$$

$$\Delta G_{\text{polar}} = \Delta G_{\text{hb}} \sum_{\text{h-bonds}} f(\Delta R, \Delta \alpha) \times f(N_{\text{neighb}}) \times f_{\text{pcs}} + \Delta G_{\text{ionic}} \sum_{\text{ionic int.}} f(\Delta R, \Delta \alpha) \times f(N_{\text{neighb}}) \times f_{\text{pcs}} + \Delta G_{\text{esrep}} N_{\text{repulsive contacts}}$$

$$\Delta G_{\text{flexi}} = \Delta G_{\text{rot}} \text{NROT}$$

Rognan et al. (69)
$$\Delta G_{\text{bind}} = K + \alpha(\text{HB}) + \beta(\text{LIPO}) + \gamma(\text{ROT}) + \delta(\text{BP}) + \epsilon(\text{DESOLV})$$

$$\text{HB} = \sum g_1(\Delta r)g_2(\Delta \alpha)$$

$$\text{LIPO} = \sum_{\text{IL}} f(f_{\text{IL}})$$

$$\text{BP} = \sum_{\text{I,P}} f(r_{\text{IP}}) + \sum_{\text{P,L}} f(r_{\text{PL}})$$

$$\text{ROT} = 1 + (1 - 1/N_{\text{rot}}) \sum (P_{\text{p}}(r) + P_{\text{p}}(r))/2$$

$$\text{DESOLV} = \Delta G_{\text{react}} = \Delta G_{\text{react(P-L)}} - \Delta G_{\text{react(P)}} - \Delta G_{\text{react(L)}}$$

Kasper et al. (70)
$$\Delta G = \Delta G_{\text{np}} + \Delta G_{\text{el}} - T\Delta S_{\text{conf}}$$

$$\Delta G_{\text{el}} = \Delta G_{\text{coul}} + \Delta G_{\text{solv}}$$

$$\Delta G_{\text{np}} = \gamma_{\text{aw}} \Delta A$$

$$\Delta A = A_{\text{complex}} - A_{\text{DnaK}} - A_{\text{peptide}}$$

Logean & Rognan (71)
$$\Delta G_{\text{bind}} = K + \alpha(\text{HB}) + \beta(\text{LIPO}) + \gamma(\text{ROT}) + \delta(\text{BP}) + \epsilon(\text{DESOLV}) + \phi(\text{CC})$$

Group III

Aqvist et al. (101) (LIE)
$$\Delta G_{\text{bind}} \cong \Delta G_{\text{nonpolar}} + \Delta G_{\text{polar}}$$

$$= \alpha(\langle V_{\text{I-S}}^{\text{vdw}} \rangle_{\text{bound}} - \langle V_{\text{I-S}}^{\text{vdw}} \rangle_{\text{free}}) + \frac{1}{2}(\langle V_{\text{I-S}}^{\text{el}} \rangle_{\text{bound}} - \langle V_{\text{I-S}}^{\text{el}} \rangle_{\text{free}})$$

Massova & Kollman (74) (MM-PBSA/GBSA)
$$\Delta G = \Delta G_{\text{PB}} + \Delta G_{\text{np}} + \Delta E_{\text{MM}} - T\Delta S_{\text{solute}}$$

$$\Delta E_{\text{MM}} = \Delta E_{\text{es}} + \Delta E_{\text{vdw}} + \Delta E_{\text{int}}$$

$$\Delta G_{\text{bind}} = G_{\text{complex}} - (G_{\text{protein}} + G_{\text{ligand}})$$

Carlsson et al. (81) (Combination of LIE and MM-PBSA)
$$\Delta G_{\text{bind}}^{\text{LIE}} = \beta(\langle U_{\text{I-prot}}^{\text{el}} \rangle_{\text{bound}} + \langle U_{\text{I-wat}}^{\text{el}} \rangle_{\text{bound}} - \langle U_{\text{I-wat}}^{\text{el}} \rangle_{\text{free}}) + \alpha \Delta \langle U_{\text{I-S}}^{\text{vdw}} \rangle + \gamma$$

Formulae 1. Representative formulae for the empirical scoring functions are classified into 3 groups as described in the text. Group I is characterized by component pairwise functions shown in the upper right, which are the function of the distance between the pair atoms and the deviation angle from the optimum configuration. This function form is empirical and arbitrary. Formulae in Group II principally deal with the solvation effect by the continuum model. Some additional terms to take into account, like the entropic effect, hydrophobic interaction, hydrogen bond, and so on, are added. In Group III, only 3 representative formulae are shown because of their high accuracy and for simplicity.

Böhm (64) $\Delta G_{bind} = \Delta G_0 + \Delta G_{hb} \sum_{h-bond} g(\Delta r, \Delta \alpha) + \Delta G_{ionic} \sum_{ionic} g(\Delta r, \Delta \alpha) + \Delta G_{lipol} |A_{lipol}| + \Delta G_{rot} NROT$

Froloff et al. (67) $\Delta G_{bind} = \Delta G_{el} + \Delta G_{np} + \Delta G_{strain} - T\Delta S_{mc} - T\Delta S_{sc} - T\Delta S_{t,r}$

Massova & Kollman (74) $\Delta G_{bind} = \Delta G_{PB} + \Delta G_{np} + \Delta E_{MM} - T\Delta S_{solute}$

Formulae 2. Representative formulae from each group are shown. Corresponding terms are linked by solid lines.

Derivational relation of representative formulae for empirical scoring functions

Fig. 9. Derivational relations among representative empirical scoring functions are illustrated. Formulae are classified into 3 groups as shown in the text and Formulae 1. Members of these groups are affected by each other as indicated. There have been a large variety of formulae to approximate the solvent effect in Group III, we did not refer to them for simplicity. Formula of MM-PBSA/GBSA has been principally conserved, but has been modified for grid computing and high throughput calculation for screening in recent years. DynaPred is essentially an SVM, but the encoding method is a direct descendant of Aqvist's method.

particular MHC molecule (89). These calculated values are not used to estimate the binding affinity of the peptide to the MHC molecule, but to encode (convert) the peptide sequence to numeric data for SVM (Fig. 11). This SVM was trained using MHCPEP and SYFPEITHI. As a result, it showed very high performance, better than SVMHC (53) and YKW0201 (90). In Formulae 2, 3 formulae from each group are exemplified and the correspondence of energetic terms among them can be seen beyond the groups, as indicated with solid lines. ΔG_{el} in Froloff's formula represents Coulomb energy ΔG_{coul} and solvation energy ΔG_{solv} (67), and both have their corresponding terms in Kollman's formula (91); however, the hydrogen bond and ionic interaction in Böhm's formula are dependent in a peculiar way, as shown in the upper right of Formulae 1. Electrostatic energy in Froloff's formula is determined by solving the Poisson-Boltzmann equation, whereas

ΔG_{hb} and ΔG_{ionic} in Böhm's formula is calculated directly from the PDB file. ΔG_{strain} in Froloff's formula is the term for conformational energy induced by ligand binding, and this term is included in the term ΔE_{MM} of Kollman's formula. ΔS_{sc} in Froloff's formula represents the change of the conformational entropy of the peptide side-chains and is calculated using the empirical scale of Pickett and Sternberg (92). Their model assumes that a solvent-exposed side chain with a relative accessibility (RA) larger than 60% is allowed to rotate freely.

$$RA = \frac{A_{side\ chain}}{A_{side\ chain\ in\ fully\ extended\ state}}$$

Table 4 summarizes the representative empirical scoring functions mainly involved in estimating MHC-peptide interaction or TCR-p-MHC interaction (93), with their targets, classified group, and accuracy.

Fig. 10. Thermodynamic cycle to calculate the change of binding free energy. Free energy change along the path $A \rightarrow B$ is equal to the change along the path $A \rightarrow C \rightarrow D \rightarrow B$. $A \rightarrow C$ correspond to remove charges of ligand and receptors in the dissociated state, whereas $D \rightarrow B$ corresponds to replacement changes in the associated state. Electrostatic energy of C and D is 0 and that of A and B is determined by solving the Poisson-Boltzmann equation. The difference of ASA derived from complex formation along the path $C \rightarrow D$ is obtained by evaluating each ASA using an appropriate program such as MS (78) and Naccess (<http://wolf.bms.umist.ac.uk/naccess/>).

Procedure of DynaPred

Fig. 11. Procedure of DynaPred. Complexing of MHC and dummy peptide (X-Gly for P1, Gly-X for Pc, and Gly-X-Gly for other positions) was performed under MD, and physical features were averaged along the trajectory; therefore, 20 X (no. of pockets) X (no. of features)-dimensional matrix was determined. This matrix was used to encode the peptide sequence, and training and prediction were performed using encoded numerical data. SYFPEITHI and MHCPEP were used as training data sets.

5. All-atom model based TI

Like group I of the empirical function method, the TI method based on the all-atom model also uses the atomic pairwise functions that are typically expressed as the sum of coulombic interaction and Lennard-Jones potential. Because the ambiguity of atomic charge parameters and Lennard-Jones potential parameters used in the all-atom TI method is small and the agreement between calculation and experiments are relatively good, we classified the all-atom TI method from the empirical function method independently. The TI method based on the all-atom model requires only structural data (atomic coordinates) and atomic parameters, that is, a force field, and does not require experimental binding data (94, 95). This method calculates the binding energy as the difference between that of the standard path (reference path), which can be obtained by experiments or other reliable methods, and that of hypothetical paths, that is, alchemical mutation paths calculated by MD (Fig. 12). State A is a peptide with standard sequence ($\lambda = 0$) and state B is a peptide with the sequence of interest ($\lambda = 1$). The difference in free energy between the two states can be calculated as

$$G^B(p, T) - G^A(p, T) = \int_0^1 \left\langle \frac{\partial H(p, q, \lambda)}{\partial \lambda} \right\rangle_{NpT; \lambda} d\lambda,$$

where $\left\langle \frac{\partial H(p, q, \lambda)}{\partial \lambda} \right\rangle_{NpT; \lambda}$ means the ensemble average

of the differential of system's Hamiltonian with λ . Wan et al. calculated the affinity difference between HLA-A0201-peptide A6-TCR and HLA-A0201-peptide P6A-TCR using this method, and they obtained a result that well agreed with the experimental data (96). Even though this method is theoretically consistent, sampling along the ordinary trajectory is not adequate to obtain a reliable ensemble average within a practical calculation time, due to being trapped in energy local minima, especially in a case with strong interactions such as the peptide-MHC complex. To overcome this difficulty, several methods to sample sufficient structures have been devised (97–100); however, the calculation cost of the TI method to evaluate peptide-MCH binding free-energy is still large and seems to be impractical.

Conclusion

In this review, we reviewed 5 methods: motif matching, matrix, SVM, empirical scoring functions, and MD-based TI. Motif matching, matrix, and empirical scoring function methods are mainly built upon experimental binding data. MD-based TI, which uses only structural data and is theoretically the most consistent, is not yet

Table 4. Representative empirical scoring functions used for protein-ligand binding affinity

Authors	Target	Group	Description
Eldridge et al. (65)	HIV-1 protease, trypsin, thrombin, etc. and their ligands	I	$r^* = 0.42 - 0.81$
Froloff et al. (67)	H-2K ^b , HLA-A2	II	about 2 to 4 times more than experimental data
Böhm (68)	Trypsin, thrombin, renin, etc and their ligands (M.W. 66 – 1047)	II	$r = 0.82 - 0.87$
Hansson et al. (72)	Trypsin, GBP, HIVp, E pepsin	III	$r = 0.89 - 0.95$
Rognan et al. (69)	HLA-A0201, H-2K ^k	II	$r = 0.88$ (HLA-A0201) $r = 0.84$ (H-2Kk)
Kasper et al. (70)	DnaK and its small ligands	II	$r = 0.86$
Logean & Rognan (71)	HLA-A, HLA-B, HLA-Cw HLA-E, H-2K ^b , H-2D ^b	II	accuracy % = 77–85
Wang et al. (66)	200 protein-ligand complexes	I	$r = 0.77$
Wan et al. (93)	TCR-p-MHC (Tax P6A on HLA- A0201)	III	MM-PBSA: $\Delta\Delta G = -2.32$ kcal/mol** LIE: $\Delta\Delta G = -1.25$ kcal/mol (experiment: $\Delta\Delta G = -2.9$ kcal/mol)
Zoete & Michielin (88)	TC-SIYR-H-2Kb	III	$r = 0.72$ (entropy is neglected)*** $r = 0.74$ (entropy is considered)*** 20 for A1; 50 for A3, A11
Antes et al. (89)****	HLA-A		60 for A1101, A0301; 85 for A24, A2402; 90 for A2

* r is the correlation coefficient between the predicted value and experimental data. ** $\Delta\Delta G$ is the difference of binding free-energies between wild type and P6A mutant. *** r is the correlation coefficient between the predicted value and experimental T-cell stimulation activity. ****Method of Antes et al., that is, DynaPred, is essentially a SVM.

Fig. 12. Thermodynamic cycle to calculate binding energy of peptide and MHC. Free energy is a quantity of state, therefore $\Delta G_2 + \Delta G_4 = \Delta G_1 + \Delta G_3$, i.e., $\Delta G_4 = \Delta G_1 + \Delta G_3 - \Delta G_2$. ΔG_1 is the change along the standard path and is determined by experiments or other reliable methods. ΔG_2 and ΔG_3 are changes along an imaginary path (alchemic mutation) where the sequence changes from the standard sequence to that of interest. Values along these paths are calculated by integrating the derivative of Hamiltonian with respect to λ from $\lambda = 0$ to $\lambda = 1$ during MD simulation.

sufficient for screening use. MM-PBSA, MM-GBSA, and LIE are positioned as the combination of empirical scoring function and MD, and they seem promising in the view of calculation cost and accuracy. The impreciseness of peptide recognition (31–36) has been making the interpretation of peptide binding data indirect, ambiguous, and complicated. These findings urge us to reconsider the interpretation of previous experiments and stimulate us to develop methods that do not require binding experiments. The development of calculating algorithms and computer performance is quite rapid, so it is not unrealistic to predict MHC-binders based on only structural data in the near future.

References

- Altman JD, Moss PA, Goulder PJ, Barouch DH, McHeyzer-Williams MG, Bell JI, et al. Phenotypic analysis of antigen-specific T lymphocytes. *Science*. 1996;274:94–96.
- Rodriguez A, Regnault A, Kleijmeer M, Ricciardi-Castagnoli P, Amigorena S. Selective transport of internalized antigens to the cytosol for MHC class I presentation in dendritic cells. *Nature Cell Biol*. 1999;1:362–368.
- Yewdell JW, Reits E, Neefjes J. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat Rev Immunol*. 2003;3:952–961.

- 4 Honey K, Rudensky AY. Lysosomal cysteine proteases regulate antigen presentation. *Nat Rev Immunol*. 2003;3:472–482.
- 5 Paz P, Brouwenstijn N, Perry R, Shastri N. Discrete proteolytic intermediates in the MHC class I antigen processing pathway and MHC independent peptide trimming in the ER. *Mol Cell*. 2003;12:565–576.
- 6 Norbury CC, Tewalt EF. Upstream toward the “DRiP”-ing source of the MHC class I pathway. *Immunity*. 2006;24:503–506.
- 7 Ashwell JD, Klusner RD. Genetic and mutational analysis of the T-cell antigen receptor. *Annu Rev Immunol*. 1990;8:139–167.
- 8 Wagner DH Jr. Re-shaping the T cell repertoire: TCR editing and TCR revision for good and for bad. *Clin Immunol*. 2007;123:1–6.
- 9 Kisielow P, Teh HS, Bluthmann H, von Boehmer H. Positive selection of antigen-specific T cells in thymus by restricting MHC molecules. *Nature*. 1988;335:730–733.
- 10 Bill J, Palmer E. Positive selection of CD4+ T cells mediated by MHC class II-bearing stromal cell in the thymic cortex. *Nature*. 1989;341:649–651.
- 11 Benoist C, Mathis D. Positive selection of the T cell repertoire: where and when does it occur? *Cell*. 1989;58:1027–1033.
- 12 von Boehmer H, Kisielow P. Self-nonsel discrimination by T cells. *Science*. 1990;248:1369–1373.
- 13 Matsuzaki G, Yoshikai Y, Ogimoto M, Kishihara K, Nomoto K. Clonal deletion of self-reactive T cells at the early stage of T cell development in thymus of radiation bone marrow chimeras. *J Immunol*. 1990;145:46–51.
- 14 Gagnon SJ, Borbulevych OY, Davis-Harrison RL, Turner RV, Damirjian M, Wojnarowicz A, et al. T cell receptor recognition via cooperative conformational plasticity. *J Mol Biol*. 2006;363:228–243.
- 15 Teng MK, Smolyar A, Tse AG, Liu JH, Liu J, Hussey RE, et al. Identification of a common docking topology with substantial variation among different TCR-peptide-MHC complexes. *Curr Biol*. 1998;8:409–412.
- 16 Hennecke J, Carfi A, Wiley DC. Structure of a covalently stabilized complex of a human alphabeta T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1. *Embo J*. 2000;19:5611–5624.
- 17 Sant’Angelo DB, Waterbury G, Preston-Hurlburt P, Yoon ST, Medzhitov R, Hong SC, et al. The specificity and orientation of a TCR to its peptide-MHC class II ligands. *Immunity*. 1996;4:367–376.
- 18 Rammensee HG, Friede T, Stevanovic S. MHC ligands and peptide motifs: first listing. *Immunogenetics*. 1995;41:178–228.
- 19 Jardetzky TS, Brown JH, Gorga JC, Stern LJ, Urban RG, Strominger JL, et al. Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides. *Proc Natl Acad Sci U S A*. 1996;93:734–738.
- 20 Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*. 1997;18:2714–2723.
- 21 Kropshofer H, Max H, Muller CA, Hesse F, Stevanovic S, Jung G, et al. Self-peptide released from class II HLA-DR1 exhibits a hydrophobic two-residue contact motif. *J Exp Med*. 1992;175:1799–1803.
- 22 Hill CM, Hayball JD, Allison AA, Rothbard JB. Conformational and structural characteristics of peptides binding to HLA-DR molecules. *J Immunol*. 1991;147:189–197.
- 23 O’Sullivan D, Sidney J, Appella E, Walker L, Phillips L, Colon SM, et al. Characterization of the specificity of peptide binding to four DR haplotypes. *J Immunol*. 1990;145:1799–1808.
- 24 Rudensky A, Preston-Hurlburt P, Hong SC, Barlow A, Janeway CA Jr. Sequence analysis of peptides bound to MHC class II molecules. *Nature*. 1991;353:622–627.
- 25 Rudensky A, Preston-Hurlburt P, al-Ramadi BK, Rothbard J, Janeway CA Jr. Truncation variants of peptides isolated from MHC class II molecules suggest sequence motifs. *Nature*. 1992;359:429–431.
- 26 Cole GA, Tao T, Hogg TL, Ryan KW, Woodland DL. Binding motifs predict major histocompatibility complex class II-restricted epitopes in the Sendai virus M protein. *J Virol*. 1995;69:8057–8060.
- 27 Hammer J, Takacs B, Sinigaglia F. Identification of a motif for HLA-DR1 binding peptides using M13 display libraries. *J Exp Med*. 1992;176:1007–1013.
- 28 Hammer J, Valsasnini P, Tolba K, Bolin D, Higelin J, Takacs B, et al. Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell*. 1993;74:197–203.
- 29 Hammer J, Belunis C, Bolin D, Papadopoulos J, Walsky R, Higelin J, et al. High-affinity binding of short peptides to major histocompatibility complex class II molecules by anchor combinations. *Proc Natl Acad Sci U S A*. 1994;91:4456–4460.
- 30 Brusica V, Rudy G, Harrison LC. MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res*. 1998;26:368–371.
- 31 Stratmann T, Apostolopoulos V, Mallet-Designé V, Corper AL, Scott CA, Wilson IA, et al. The I-Ag7 MHC class II molecule linked to murine diabetes is a promiscuous peptide binder. *J Immunol*. 2000;165:3214–3225.
- 32 Maverakis E, van den Elzen P, Sercarz EE. Self-reactive T cells and degeneracy of T cell recognition: evolving concepts-from sequence homology to shape mimicry and TCR flexibility. *J Autoimmun*. 2001;16:201–209.
- 33 Anderton SM, Viner NJ, Matharu P, Lowrey PA, Wraith DC. Influence of a dominant cryptic epitope on autoimmune T cell tolerance. *Nat Immunol*. 2002;3:175–181.
- 34 He XL, Radu C, Sidney J, Sette A, Ward ES, Garcia KC. Structural snapshot of aberrant antigen presentation linked to autoimmunity: the immunodominant epitope of MBP complexed with I-Au. *Immunity*. 2002;17:83–94.
- 35 Seamons A, Sutton J, Bai D, Baird E, Bonn N, Kafsack BF, et al. Competition between two MHC binding registers in a single peptide processed from myelin basic protein influences tolerance and susceptibility to autoimmunity. *J Exp Med*. 2003;197:1391–1397.
- 36 Bankovich AJ, Girvin AT, Moesta AK, Garcia KC. Peptide register shifting within the MHC groove: theory becomes reality. *Mol Immunol*. 2004;40:1033–1039.
- 37 Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J Exp Med*. 1994;180:2353–2358.
- 38 Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*. 1994;152:163–175.
- 39 Kubo RT, Sette A, Grey HM, Appella E, Sakaguchi K, Zhu NZ,

- et al. Definition of specific peptide motifs for four major HLA-A alleles. *J Immunol.* 1994;152:3913–3924.
- 40 Davenport MP, Ho Shon IA, Hill AV. An empirical method for the prediction of T-cell epitopes. *Immunogenetics.* 1995;42:392–397.
 - 41 Udaka K, Wiesmuller KH, Kienle S, Jung G, Walden P. Decrypting the structure of major histocompatibility complex class I-restricted cytotoxic T lymphocyte epitopes with complex peptide libraries. *J Exp Med.* 1995;181:2097–2108.
 - 42 Udaka K, Wiesmuller KH, Kienle S, Jung G, Tamamura H, Yamagishi H, et al. An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics.* 2000;51:816–828.
 - 43 Brusic V, Schonbach C, Takiguchi M, Ciesielski V, Harrison LC. Application of genetic search in derivation of matrix models of peptide binding to MHC molecules. *Proc Int Conf Intell Syst Mol Biol.* 1997;5:75–83.
 - 44 Schonbach C, Ibe M, Shiga H, Takamiya Y, Miwa K, Nokihara K, et al. Fine tuning of peptide binding to HLA-B*3501 molecules by nonanchor residues. *J Immunol.* 1995;154:5951–5958.
 - 45 Schonbach C, Miwa K, Ibe M, Shiga H, Nokihara K, Takiguchi M. Refined peptide HLA-B*3501 binding motif reveals differences in 9-mer to 11-mer peptide binding. *Immunogenetics.* 1996;45:121–129.
 - 46 Kondo A, Sidney J, Southwood S, del Guercio MF, Appella E, Sakamoto H, et al. Prominent roles of secondary anchor residues in peptide binding to HLA-A24 human class I molecules. *J Immunol.* 1995;155:4307–4312.
 - 47 Rothbard JB, Marshall K, Wilson KJ, Fugger L, Zaller D. Prediction of peptide affinity to HLA DRB1*0401. *Int Arch Allergy Immunol.* 1994;105:1–7.
 - 48 Zhao Y, Gran B, Pinilla C, Markovic-Plese S, Hemmer B, Tzou A, et al. Combinatorial peptide libraries and biometric score matrices permit the quantitative analysis of specific and degenerate interactions between clonotypic TCR and MHC peptide ligands. *J Immunol.* 2001;167:2130–2141.
 - 49 Hemmer B, Vergelli M, Gran B, Ling N, Conlon P, Pinilla C, et al. Predictable TCR antigen recognition based on peptide scans leads to the identification of agonist ligands with no sequence homology. *J Immunol.* 1998;160:3631–3636.
 - 50 Sturniolo T, Bono E, Ding J, Radrizzani L, Tuereci O, Sahin U, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol.* 1999;17:555–561.
 - 51 Vapnik VN. *The nature of statistical learning theory.* Springer. 2000.
 - 52 Joachims T. Chapter 11. In: Schölkopf B, Burges C, Smola A, editors. *Making large-scale SVM learning practical. Advances in Kernel methods – support vector learning.* Cambridge, MA: MIT Press; 1999.
 - 53 Dönnes P, Elofsson A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics.* 2002;3:25–32.
 - 54 Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics.* 1999;50:213–219.
 - 55 Bhasin M, Raghava GP. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics.* 2004;20:421–423.
 - 56 Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* 1999;27:368–369.
 - 57 Liu W, Meng X, Xu Q, Flower DR, Li T. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics.* 2006;7:182.
 - 58 Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K, et al. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res.* 2005;1:4.
 - 59 Cui J, Han LY, Lin HH, Zhang HL, Tang ZQ, Zheng CJ, et al. Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol Immunol.* 2007;44:866–877.
 - 60 Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Protein J.* 1985;4:23–55.
 - 61 Zhao Y, Pinilla C, Valmori D, Martin R, Simon R. Application of support vector machines for T-cell epitopes prediction. *Bioinformatics.* 2003;19:1978–1984.
 - 62 Rubio-Godoy V, Dutoit V, Zhao Y, Simon R, Guillaume P, Houghten R, et al. Positional scanning-synthetic peptide library-based analysis of self- and pathogen-derived peptide cross-reactivity with tumor-reactive Melan-A-specific CTL. *J Immunol.* 2002;169:5696–5707.
 - 63 Takahashi T, Sugiura J, Nagayama K. Comparison of all atom, continuum, and linear fitting empirical models for charge screening effect of aqueous medium surrounding a protein molecule. *J Chem Physics.* 2002;116:8232.
 - 64 Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des.* 1994;8:243–256.
 - 65 Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des.* 1997;11:425–445.
 - 66 Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des.* 2002;16:11–26.
 - 67 Froloff N, Windemuth A, Honig B. On the calculation of binding free energies using continuum methods: application to MHC class I protein-peptide interactions. *Protein Sci.* 1997;6:1293–1301.
 - 68 Böhm HJ. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des.* 1998;12:309–323.
 - 69 Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem.* 1999;42:4650–4658.
 - 70 Kasper P, Christen P, Gehring H. Empirical calculation of the relative free energies of peptide binding to the molecular chaperone DnaK. *Proteins.* 2000;40:185–192.
 - 71 Logean A, Rognan D. Recovery of known T-cell epitopes by computational scanning of a viral genome. *J Comput Aided Mol Des.* 2002;16:229–243.
 - 72 Hansson T, Marelus J, Aqvist J. Ligand binding affinity prediction by linear interaction energy methods. *J Comput Aided Mol*

- Des. 1998;12:27–35.
- 73 Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J Am Chem Soc.* 1998;120: 9401–9409.
 - 74 Massova I, Kollman PA. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect Drug Discov Design.* 2000;18:113–135.
 - 75 Nicholls A, Honig B. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J Comput Chem.* 1991;12:435–445.
 - 76 Baker N, Holst M, Wang F. Adaptive multilevel finite element solution of the Poisson-Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems. *J Computat Chem.* 2000;21:1343–1352.
 - 77 Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol.* 1971;55:379–400.
 - 78 Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. *Science.* 1983;221:709–713.
 - 79 Chothia C. Hydrophobic bonding and accessible surface area in proteins. *Nature.* 1974;248:338–339.
 - 80 Takahashi T, Nagayama K, Hogyoku M. Inter-subunit electrostatic interactions in ferritin molecule: Comparison with inter-molecular interactions in crystals. *J Crystal Growth.* 1996;168: 130–137.
 - 81 Carlsson J, Ander M, Nervall M, Aqvist J. Continuum solvation models in the linear interaction energy method. *J Phys Chem B.* 2006;110:12034–12041.
 - 82 Aqvist J, Luzhkov VB, Brandsdal BO. Ligand binding affinities from MD simulations. *Acc Chem Res.* 2002;35:358–365.
 - 83 Almlof M, Aqvist J, Smalas AO, Brandsdal BO. Probing the effect of point mutations at protein-protein interfaces with free energy calculations. *Biophys J.* 2006;90:433–442.
 - 84 Huo S, Massova I, Kollman PA. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J Computat Chem.* 2002;23:15–27.
 - 85 Massova I, Kollman P. Computational alanine scanning to probe protein-protein interactions: a novel application to evaluate binding free energies. *J Am Chem Soc.* 1999;121:8142–8143.
 - 86 Li X, Zhang W, Qiao X, Xu X. Prediction of binding for a kind of non-peptic HCV NS3 serine protease inhibitors from plants by molecular docking and MM-PBSA method. *Bioorg Med Chem.* 2007;15:220–226.
 - 87 Zoete V, Michielin O, Karplus M. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. Application to HIV-1 protease inhibitors. *J Comput Aided Mol Des.* 2003;17:861–880.
 - 88 Zoete V, Michielin O. Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein-protein association using MM-GBSA: application to the TCR-p-MHC complex. *Proteins.* 2007;67: 1026–1047.
 - 89 Antes I, Siu SW, Lengauer T. DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. *Bioinformatics.* 2006;22:e16–e24.
 - 90 Yu K, Petrovsky N, Schonbach C, Koh JY, Brusica V. Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med.* 2002;8:137–148.
 - 91 Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res.* 2000;33:889–897.
 - 92 Pickett SD, Sternberg MJ. Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol.* 1993;231:825–839.
 - 93 Wan S, Coveney PV, Flower DR. Peptide recognition by the T cell receptor: comparison of binding free energies from thermodynamic integration, Poisson-Boltzmann and linear interaction energy approximations. *Philos Trans: Math Phys Eng Sci.* 2005;363:2037–2053.
 - 94 Kollman P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem Rev.* 1993;93:2395–2417.
 - 95 Simonson T, Archontis G, Karplus M. Free energy simulations come of age: protein-ligand recognition. *Acc Chem Res.* 2002;35:430–437.
 - 96 Wan S, Coveney PV, Flower DR. Molecular basis of peptide recognition by the TCR: affinity differences calculated using large scale computing. *J Immunol.* 2005;175:1715–1723.
 - 97 Nakajima N, Nakamura H, Kidera A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J Phys Chem B.* 1997;101: 817–824.
 - 98 Shimizu H. Structural phase transition of Di-block poly-ampholyte. *Mol Simulation.* 1999;22:285–301.
 - 99 Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett.* 1999;314:141–151.
 - 100 Nguyen PH, Stock G, Mittag E, Hu CK, Li MS. Free energy landscape and folding mechanism of a beta-hairpin in explicit water: a replica exchange molecular dynamics study. *Proteins.* 2005;61:795–808.
 - 101 Aqvist J, Medina C, Samuelsson JE. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* 1994;7:385–391.