A Novel Peptide Binding Prediction Approach for HLA-DR Molecule Based on Sequence and Structural Information

Zhao Li, Yilei Zhao, Gaofeng Pan, Jijun Tang, and Fei Guo

1School of Computer Science and Technology, Tianjin University, 92 Weijin Road, Nankai District, Tianjin 300072, China
2School of Computational Science and Engineering, University of South Carolina, Columbia, SC, USA

Correspondence should be addressed to Fei Guo; fguo@tju.edu.cn

Received 10 March 2016; Accepted 4 May 2016

MHC molecule plays a key role in immunology, and the molecule binding reaction with peptide is an important prerequisite for T cell immunity induced. MHC II molecules do not have conserved residues, so they appear as open grooves. As a consequence, this will increase the difficulty in predicting MHC II molecules binding peptides. In this paper, we aim to propose a novel prediction method for MHC II molecules binding peptides. First, we calculate sequence similarity and structural similarity between different MHC II molecules. Then, we reorder pseudosequences according to descending similarity values and use a weight calculation formula to calculate new pocket profiles. Finally, we use three scoring functions to predict binding cores and evaluate the accuracy of prediction to judge performance of each scoring function. In the experiment, we set a parameter $\alpha$ in the weight formula. By changing $\alpha$ value, we can observe different performances of each scoring function. We compare our method with the best function to some popular prediction methods and ultimately find that our method outperforms them in identifying binding cores of HLA-DR molecules.

1. Introduction

Histocompatibility refers to the degree of antigenic similarity between the tissues of different individuals, which determines the acceptance or rejection of allografts. Transplantation antigen or histocompatibility antigen is the cause of rejection of allografts [1, 2]. MHC (Major Histocompatibility Complex) is present on the chromosome encoding a major histocompatibility antigen, mutual recognition between control cells, and the regulation of immune response.

MHC molecule plays a key role in immunology, and the molecule binding reaction with peptide is an important prerequisite for T cell immunity induced [2, 3]. By detecting a wide variety of microbial pathogens, the immune system protects host against diseases. Because of this, the binding prediction of MHC molecules with peptides has always been a hot topic in bioinformatics. Many researches in this field not only help us to understand the process of immune but also develop the work of vaccine design assisted by computers.

MHC genes produce two different types of molecules, which are MHC I molecules and MHC II molecules [1, 2]. MHC I molecules contain two separate polypeptide chains: the MHC $\alpha$ chain encoded by MHC genes and the MHC $\beta$ chain encoded by non-MHC genes [4, 5]. MHC I class molecules are expressed in almost all eukaryotic cell surfaces, recognized by CD8+ cells. MHC II class molecules consist of two non-covalently linked polypeptide chains, namely, $\alpha$ chain and $\beta$ chain. MHC II class molecules are expressed on antigen-presenting cells in general. Foreign MHC II antigens only capture and present on the surface of antigen-presenting cells (APC) TH cell [6]. After that, APC secretes large amounts of cytoplasm, activating cell invasion defense behavior. Only the binding of antigen peptides and MHC II class molecules can activate CD4+ TH cells (helper T cells) [7]. Then, the activated TH cells would differentiate into effector cells and activate the immune response.

The structures of MHC I molecules and MHC II molecules slightly differ in the binding grooves [5]. Close grooves form on the binding of MHC I molecules and antigenic peptides. On the other hand, MHC II molecules do not have conserved residues, so they appear as open grooves. As a consequence, this will increase the difficulty in
predicting MHC II molecules binding peptides [7]. In this paper, we aim to solve more difficult problem of predicting MHC II binding peptides.

The pioneering and most popular pan-specific approach for MHC II binding prediction is the TEPITOPE method [8], and basic idea is the HLA-DR allele having identical pseudosequence. The same pocket will share the same quantitative profile. By using multiple instance learning, the MHCIIMulti method [9] can predict more than 500 HLA-DR molecules. Transforming each DRB allele into a pseudosequence with 21 amino acids and using the SMM-align method to identify binding cores, the NetMHCIIpan method [5] gets an accurate prediction by using an artificial neural network algorithm [10, 11]. Combining NN-align and NetMHCpan with NetMHCIIpan [9, 12], the MULTIPRED2 method [13–15] can get a perfect prediction for 1077 HLA-I and HLA-II alleles and 26 HLA supertypes.

In this paper, we propose a novel prediction method for predicting MHC II molecules binding peptides. First, we calculate sequence similarity and structural similarity between different MHC molecules [13, 16]. Then, we reorder pseudosequences according to descending similarity values and use a weight calculation formula to calculate new pocket profiles. Finally, we use three scoring functions to predict binding cores and evaluate the accuracy of prediction to judge performance of each scoring function [17, 18]. In the experiments, we set a parameter $\alpha$ in the weight formula. By changing $\alpha$, we can observe different performances of each of the scoring function. We compare our method with the best function to some popular prediction methods and ultimately find that our method outperforms them in identifying binding cores of HLA-DR molecule [19]. The work would suggest a novel computational strategy for special protein identification instead of traditional machine learning based methods [20, 21].

2. Materials and Methods

2.1. Data Sets. We find 39 MHC molecules and peptides binding complexes from Protein Data Bank (http://www.rcsb.org/pdb/search/), which constitutes the data set used in this paper. In this data set, lengths are between 11 and 23, and we can find polypeptide-binding sites, namely, binding cores. Table 1 lists the details of these 39 MHC molecules and peptide binding complexes [14, 22, 23].

In Table 1, the first column is PDB ID of 39 complexes from PDB; the second column is the name of corresponding alleles from 39 complexes; the third column is the corresponding polypeptide sequences, in which the enlarged nine positions are the binding cores.

2.2. Methods. There are thousands of allele variants in nature [2, 4]. It is absolutely impossible to measure the binding specificity one by one. Motivated by this perspective, we propose a new computational method to predict the binding specificity of peptides without any biochemical experiment, which combines the sequence and structural information of these known specificity-binding MHC molecules, as showed in Figure 1. We evaluate the method on all general HLA-DRB data sets, and results indicate that our method is close to the state-of-the-art technology and our approach can predict all sequence-known MHC molecules and cost little time, extending the prediction space compared with other time-consuming approaches.

2.3. Crucial Pockets relative to Binding Specificities of HLA-DR Molecules. We mainly use Position Specific Scoring Matrix (PSSM) [13, 24] in our approach, which is a popular technology in the problem of MHC binding. Roughly speaking,
there are nine amino acids in MHC binding cores, and each position is a specific pocket as showed in Table 2. We use PSSM to quantify the binding affinity between twenty basic amino acids with these nine pockets.

There are five anchor sites (1, 4, 6, 7, and 9) at the binding core for MHC II molecules, which determine the binding strength of peptides with MHC II molecules. Because site 1 of MHC II is consistent with different MHC I molecules and peptides, it is important to identify the precise quantification of its binding core in site 1, yet we use weights of four anchor sites (4, 6, 7, and 9) to define profiles. For other sites, the same approach, such as TEPITOPE, is to specify their quantitative profiles.

2.4. Computing Similarity between Different MHC Molecules

2.4.1. Sequence-Based Similarity. Sequence-based similarity can be calculated by alignment results. Here, pocket pseudosequences and associated profiles refer to raw pocket pseudosequences and raw pocket profiles, respectively. These raw pseudosequences are composed of several amino acids, whose associated residue indices are shown in Table 3. Eleven representative HLA-DR alleles are adopted to specify different profiles for anchor pockets 4, 6, 7, and 9. These eleven alleles are DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0402, DRB1*0404, DRB1*0701, DRB1*0801, DRB1*1101, DRB1*1302, DRB1*1501, and DRB5*0101.

If two alleles have identical pseudosequences in the same pocket, they will have identical profiles. For a given pocket, we collect all the different raw pocket pseudosequences into one set $R^x$, $R^x = \{r_1, r_2, \ldots, r_m\}$, and $|r_i| = n$, where $i = 1, 2, \ldots, m$, $x \in \{4, 6, 7, 9\}$, $m$ is the number of unique pseudosequences, and $n$ is the number of amino acids contained in a pseudosequence. Meanwhile, we collect all different raw profiles into one set $P^x$, $P^x = \{p_1, p_2, \ldots, p_m\}$, and $|p_i| = 20$, where $i = 1, 2, \ldots, m$. There is a one-to-one correspondence between $p_i$ and $r_i$. We use BLOSUM to calculate the sequence similarity between different MHC molecules, defined as $\text{BLOSUM} = (S_x - S_i)$. Then, we can get encoded pseudosequence, which is a $20n$-dimensional real vector $V^x = \{V_1, V_2, \ldots, V_m\}$. We use Radial Basis Function (RBF) to measure the similarity between encoded predicted pseudosequences $V_a$ and a raw encoded pseudosequence:

$$K_{\text{seq}}(V_a, V_j) = \text{BLOSUM}(V_a, V_j) = S_x - S_i, \quad V_j \subseteq V^x. \quad (1)$$

2.4.2. Structure-Based Similarity. Using MHC II HLA-peptide complex structure from Protein Data Bank (PDB), we can get the residues 3D-coordinate of the pocket in each MHC molecule, $h(p_x, p_y, p_z)$. We define vector $H^x = [h_1, h_2, \ldots, h_n]$, where $n$ is the number of amino acids in the pseudocontained sequence; meanwhile, we collect a set $S^x$, $S^x = [H_1, H_2, \ldots, H_m]$, $m$ is the number of different pseudosequences, and there is also one-to-one correspondence between $H_i$ and $r_i$. 

![Figure 1: The architecture of our approach to MHC II and peptide binding problem.](image-url)
### Table 2: 30 HLA-complexes binding pockets.

| PDB ID | Pocket 1 | Pocket 2 | Pocket 3 | Pocket 4 | Pocket 5 | Pocket 6 | Pocket 7 | Pocket 8 | Pocket 9 |
|--------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1AQD   | 82N 85V 86G | 77T 78Y 81H 82N | 78Y | 13F 74A 78Y | 13F 71R | 11L | 47Y 6L1W 6L7Q 71R | 60Y 61W | 9W 57D 61W |
| 1PYW   | 82N 85V 86G 89F | 77T 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 13F 71R | 11L | 11L 28E 6L1W 71R | 60Y 61W | 57D 61W |
| 1KLG   | 82N 85V | 78Y 81H 82N | 78Y | 13F 71R 78Y | 13F 71R | 11L | 61W | 60Y 61W | 57D 61W |
| 2FSE   | 82N 85V 86G 89F | 77T 78Y 82N | 78Y | 13F 71R 71R 74A 78Y | 13F 71R | 71R | 28E 47Y 6L1W 6L7Q 71R | 61W | 57D |
| 1KLU   | 82N 85V | 78Y 81H 82N | 78Y | 13F 71R 78Y | 13F 71R | 11L | 61W | 60Y 61W | 57D 61W |
| 1SJH   | 82N | 78Y 81H 82N | 78Y | 13F 26L 70Q 71R 74A 78Y | 71R | 11L | 61W | 60Y 61W | 57D 61W |
| 1SJF   | 82N | 78Y 81H 82N | 78Y | 13F 26L 70Q 71R 74A 78Y | 71R | 11L | 61W | 60Y 61W | 57D 60Y 61W |
| 1T5W   | 82N 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 13F 71R | 11L | 61W 71R | 60Y 61W | 9W 57D 61W |
| 1T5X   | 82N 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 71R | 11L | 61W 71R | 61W | 57D 61W |
| 2IAN   | 82N 85V | 78Y 81H 82N | 78Y | 13F 70Q 74A 78Y | 13F 70Q 71R | 11L | 61W 71R | 61W | 57D 61W |
| 2IPK   | 82N 85V 86G 89F | 77T 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 71R | 11L | 47Y 6L1W 6L7Q 71R | 60Y 61W | 9W 57D 61W |
| 1FYT   | 82N 85V 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 13F 71R | 11L | 28E 47Y 6L1W 6L7Q 71R | 60Y 61W | 9W 57D 61W |
| 1RSI   | 82N 85V 86G 89F | 77T 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 70Q 71R | 11L | 47Y 6L1W 6L7Q 71R | 61W | 9W 57D 61W |
| 1HXY   | 82N 85V 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 71R | 11L | 28E 47Y 6L1W 6L7Q 71R | 60Y 61W | 9W 57D 61W |
| 1JWM   | 82N 85V 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 71R | 11L | 28E 47Y 6L1W 6L7Q 71R | 61W | 57D 61W |
| 1JWS   | 82N 85V 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 13F 71R | 11L | 47Y 6L1W 6L7Q 71R | 61W | 9W 57D 61W |
| 1JWU   | 82N 85V 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 13F 71R | 11L | 28E 47Y 6L1W 6L7Q 71R | 61W | 9W 57D 61W |
| 1L05   | 82N 85V 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 78Y | 13F 71R | 11L | 47Y 6L1W 6L7Q 71R | 61W | 9W 57D 60Y 61W |
| 2ICW   | 82N 85V 86G 89F | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 13F 71R | 11L | 28E 47Y 6L1W 6L7Q 71R | 61W | 9W 57D 61W |
| 2OJE   | 82N 85V 86G | 77T 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 70Q 71R | 11L | 28E 47Y 6L1W 6L7Q 71R | 61W | 9W 57D 61W |
| 2G9H   | 82N 85V 86G 89F | 77T 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 71R | 11L | 13F | 28E 47Y 6L1W 6L7Q 71R | 60Y 61W | 9W 57D 61W |
| 2IAM   | 82N | 78Y 81H 82N | 78Y | 13F 70Q 71R 74A 78Y | 70Q 71R | 11L | 6L1W 71R | 60Y 61W | 57D 61W |
| 1A6A   | 82N 85V 86V | 77T 78Y 81H 82N | 78Y | 13S 26Y 74R 78Y | 71K 74R | 11S 30Y | 30Y 47Y 6L1W 6L7Q 71R | 61W | 9E 30Y 57D 61W |
| PDB ID | Pocket 1 | Pocket 2 | Pocket 3 | Pocket 4 | Pocket 5 | Pocket 6 | Pocket 7 | Pocket 8 | Pocket 9 |
|--------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1J8H   | 82N 85V 86G 89F | 77T 78Y 81H 82N | 78Y      | 13H      | 26F 70Q 74A | 13H 70Q 71K | 11V 13H 30Y | 30Y 47Y 61W 67L | 60Y 61W | 37Y 57D 61W |
| 2SEB   | 82N      | 77T 78Y 81H 82N | 78Y      | 13H      | 26F 71K 78Y | 13H 71K    | 30Y      | 30Y 47Y 61W | 60Y 61W | 61W      |
| 1BX2   | 82N 85V | 77T 78Y 81H 82N | 78Y      | 13H      | 26F 70Q 74A | 70Q        | 13R      |          |          | 57D 60Y 61W |
| 1YMM   | 82N      | 77T 78Y 81H 82N | 78Y      | 13R      | 26F 70Q 74A | 70Q        | 13R      | 61W 67I   | 61W      | 57D 61W |
| 1FV1   | 82N 85V 86G 89F | 78Y 81H 82N | 78Y      | 13Y 71R 78Y | 71R      | 13Y       | 61W 67L 71K | 61W      |          | 57D      |
| 1H15   | 82N 89F  | 77T 78Y 81H 82N | 78Y      | 13Y 71R 78Y | 71R      | 11V 13Y 30D | 61W      |          |          | 57D 60Y |
| 1ZGL   | 82N 85V 89F | 77T 78Y 81H 82N | 13Y 26F 71R 78Y | 13Y      | 13Y 28H 61W 71R | 61W      |          |          |          | 57D 60Y 61W |
Table 3: Important positions at the binding core for MHC II molecules.

| Pocket | Important positions |
|--------|---------------------|
| Pocket 1 | 82 85 86 89 |
| Pocket 2 | 77 78 81 82 |
| Pocket 3 | 78 |
| Pocket 4 | 11 13 26 28 70 71 74 78 |
| Pocket 5 | 11 13 28 70 71 74 |
| Pocket 6 | 11 13 28 70 71 74 |
| Pocket 7 | 11 28 30 47 61 67 70 71 |
| Pocket 8 | 60 61 |
| Pocket 9 | 9 30 37 57 60 61 |

Next, we need to estimate the similarity of three-dimensional structures between a measured MHC molecule and five MHC molecules with known pseudosequence PSSM. Rigid transformation is to compare three-dimensional substructures of two proteins [25, 26].

Intuitively, we fix one of the structures, A, move (transformation and rotation) the other structure, B, and find the best movement in three-dimensional space, with two atoms to the nearest structure. We calculate the Euclidean distance between two structures, defined as RMSD = |C_a - C_j|. We can get encoded pseudosequence V^x = {V_1, V_2, …, V_m} and calculate the similarity between 3D structures of encoded predicted pseudosequences V_a and a raw encoded pseudosequence:

\[ K_{spa}(V_a, V_j) = \text{RMSD}(V_a, V_j), \quad V_j \subseteq V^x. \]  

2.4.3. Overall Similarity. After that, we have obtained sequence similarity and structural similarity. We calculate final similarity score functions according to the following three formulas:

\[ K_1(V_a, V_j) = \sqrt{ \frac{K_{seq}(V_a, V_j)^2 + K_{spa}(V_a, V_j)^2}{2} }, \]  

\[ K_2(V_a, V_j) = \frac{K_{seq}(V_a, V_j) + K_{spa}(V_a, V_j)}{2}, \]  

\[ K_3(V_a, V_j) = \sqrt{ K_{seq}(V_a, V_j) + K_{spa}(V_a, V_j) }. \]  

2.5. Weights Calculation for New Pocket Profiles. We reorder all pseudosequences according to descending similarity values and use a weight calculation formula to calculate new pocket profiles. A new pocket profile is generated as a weighted average over m raw pocket profiles in P^x. Next, we use the gamma distribution to generate the weights. The gamma PDF distribution is defined as follows:

\[ g(x; k, \theta) = \frac{1}{\theta^k \gamma(k)} x^{k-1} e^{-x/\theta}, \]  

where \( x > 0 \) and \( k, \theta > 0 \), and \( \gamma(k) \) denotes the gamma function.

The weight distribution is generated to discretize the gamma PDF as follows:

\[ G(X = i) = \frac{1}{\theta^k \gamma(k)} i^{k-1} e^{-i/\theta}, \quad i = 1, 2, \ldots, m, \]  

where \( m \) is the dimension of the weights and \( k \) and \( \theta \) are the shape and scale parameters, respectively. The gamma distribution generates the weight vector to give a higher weight for more similarity pseudosequences.

After normalizing, the weight vector is defined as follows:

\[ P(X = i) = \frac{[G(X = i)]^{\alpha}}{\sum_{k=1}^{m} [G(X = k)]^{\alpha}}, \quad i = 1, 2, \ldots, m. \]  

3. Result

First, we design an experiment to choose appropriate scoring function to combine sequence similarity and structural similarity. Then, we compare with other state-of-the-art technologies, which are TEPITOPE, MultiRTA, NetMHCIIpan-2.0, and NetMHCIIpan-1.0. The result indicates that our approach can obtain better prediction and effectively extend current prediction methods. Finally, we test on more data sets.

3.1. Evaluation of Different Scoring Functions. Here, we use 30 of 39 MHC molecules and peptide complexes as test set and get the appropriate scoring functions as showed above. The value of the parameter \( \alpha \) is set to 1, 2, 3, 4, 5, 10, 15, and 20, followed by results shown in Figure 2. We find that no significant changes can be found by \( K_1(V_a, V_j) \); for \( K_2(V_a, V_j) \) and \( K_3(V_a, V_j) \), when \( \alpha = 1 \) prediction error number is 10 and 9 and when \( \alpha = 3 \) prediction errors reduced to 8, we set the value of \( \alpha \) to 3. Comparing these three functions, the least numbers of errors by three functions are 4, 8, and 8. Details are shown in Tables S1, S2, and S3 in the Supplementary Material available online at http://dx.doi.org/10.1155/2016/3832176.

3.2. Compared with Conventional Well-Known Methods. From the above experimental results, \( K_3(V_a, V_j) \) obtains the most accurate prediction, so we will select \( K_3(V_a, V_j) \) with \( \alpha = 3 \) as our final approach. We compare our current
Table 4: Comparison of our binding prediction with other approaches. The 5th column is the result of our method, and 6th to 8th columns are results of TEPITOPE, MultiRTA, and NetMHCIIpan. The bold cell means one error.

| PDB ID | Allele | Peptide | Core | Ours | TEPITOPE | MultiRTA | NetMHCIIpan-2.0 |
|--------|--------|---------|------|------|----------|----------|-----------------|
| 1AQD   | DRB1*0101 | VGSDWRFRLGRYHQYA | WRFRLGRYHQ | WRFRLGRYHQ | WRFRLGRYHQ | WRFRLGRYHQ | WRFRLGRYHQ |
| 1PYW   | DRB1*0101 | XFKQNAAAALX | FVKQNAAL | FVKQNAAL | FVKQNAAL | FVKQNAAL | FVKQNAAL |
| 1KLG   | DRB1*0101 | GELGILNAAKVPAD | IGLNAAKV | IGLNAAKV | IGLNAAKV | IGLNAAKV | IGILNAAK |
| 2FSE   | DRB1*0101 | GELGILNAAKVPAD | IGLNAAKV | IGLNAAKV | IGLNAAKV | IGLNAAKV | IGLTNAAKV |
| 1KLU   | DRB1*0101 | AGFKGEQGPKGEPGD | FKGEQGPKG | FKGEQGPKG | FKGEQGPKG | FKGEQGPKG | FKGEQGPKG |
| 1SJH   | DRB1*0101 | PVPITPMSALSEG | VIPMFSALS | VIPMFSALS | VIPMFSALS | VIPMFSALS | VIPMFSALS |
| 1SJE   | DRB1*0101 | PEPITPMSALSEGATP | VIPMFSALS | VIPMFSALS | VIPMFSALS | VIPMFSALS | VIPMFSALS |
| 1T5W   | DRB1*0101 | AAYSQTAPLLISPR | YSQTAPLL | YSQTAPLL | YSQTAPLL | YSQTAPLL | YSQTAPLL |
| 1T5X   | DRB1*0101 | AAYSQTAPLLISPR | YSQTAPLL | YSQTAPLL | YSQTAPLL | YSQTAPLL | YSQTAPLL |
| 2IAN   | DRB1*0101 | GELGILNAAKVPAD | IGLNAAKV | IGLNAAKV | IGLNAAKV | IGLNAAKV | IGLNAAKV |
| 2IPK   | DRB1*0101 | GELGILNAAKVPAD | IGLNAAKV | IGLNAAKV | IGLNAAKV | IGLNAAKV | IGLNAAKV |
| 1FYT   | DRB1*0101 | XPKWVKQNTLKLAT | WVKQNTLKL | WVKQNTLKL | WVKQNTLKL | WVKQNTLKL | WVKQNTLKL |
| 1R5I   | DRB1*0101 | PKYVKQTNLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 1HXV   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 1JWM   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 1JWS   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 1JWU   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 1LO5   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 2ICW   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 2EJE   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 2G9H   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 2IAM   | DRB1*0101 | PKYVKQNTLKLAT | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 1A6A   | DRB1*0301 | PVSKMRMAATPLLMQA | MRMATPPLL | MRMATPPLL | MRMATPPLL | MRMATPPLL | MRMATPPLL |
| 1J8H   | DRB1*0401 | PVSKMRMAATPLLMQA | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL | YVKQNTLKL |
| 2SEB   | DRB1*0401 | AYMRADAAGAGGA | MRAADAAGG | MRAADAAGG | MRAADAAGG | MRAADAAGG | MRAADAAGG |
| 1BX2   | DRB1*1501 | ENPVPVFHFKNIVTPR | VHFKNIVT | VHFKNIVT | VHFKNIVT | VHFKNIVT | VHFKNIVT |
| 1YMM   | DRB1*1501 | ENPVPVFHFKNIVTPRGGGGG | VHFKNIVT | VHFKNIVT | VHFKNIVT | VHFKNIVT | VHFKNIVT |
| 1FV1   | DRB5*0101 | NPVPVFHFKNIVTPRPSSPQ | FKNIVTPRT | FKNIVTPRT | FKNIVTPRT | FKNIVTPRT | FKNIVTPRT |
| 1H5    | DRB5*0101 | GGYVHFVKKHIVHES | YFFVKKHVH | YFFVKKHVH | YFFVKKHVH | YFFVKKHVH | YFFVKKHVH |
| 1ZGL   | DRB5*0101 | VHFKNIVTPRPGG | FKNIVTPRT | FKNIVTPRT | FKNIVTPRT | FKNIVTPRT | FKNIVTPRT |

Results: 4 errors 0 errors 4 errors 6 errors
Table 5: Other prediction results of nine MHC molecules. This table shows the prediction result of our method on 9 MHC molecules. The 5th column is the result. There is only one error result, which is shown using bold font.

| PDB ID | Allele  | Peptide          | Core          | Ours        |
|--------|---------|------------------|---------------|-------------|
| 4E41   | DRB1*0101 | GELIGILNAKVPAD  | IGILNAAKV     | IGILNAAKV   |
| 1DLH   | DRB1*0101 | PKYVKQNTLKLA    | YVKQNTKLKL   | YVKQNTKLKL |
| 1KG0   | DRB1*0101 | PKYVKQNTLKLA    | YVKQNTKLKL   | YVKQNTKLKL |
| 3L6F   | DRB1*0101 | APPAYEKLSAEQSSP | YEKLSAEQA     | YEKLSAEQA   |
| 3PDO   | DRB1*0101 | KPVSKMRMATPLLQMQLPM | MRMATPLLLM   | MRMATPLLLM  |
| 3SGD   | DRB1*0101 | KMRMATPLLQMQLPM | MRMATPLLLM   | MRMATPLLLM  |
| 3S5L   | DRB1*0101 | PKYVKQNTLKLA    | YVKQNTKLKL   | YVKQNTKLKL |
| 1HQR   | DRB5*0101 | VHFFKNIVTPRTP   | FKNIVTPRTP   | FKNIVTPRTP  |

| Results | 1 error |

Figure 2: Predicted results by different score functions. x-axis represents different α values, and the y-axis refers to predicted results of different score functions.

3.3. Other Prediction Results. When compared with other methods on the above experiments, we only use 30 of 39 MHC molecules and peptide complexes as test set. In this section, we test on the remaining nine MHC molecules. In this experiment, we choose \( K_1(V_a, V_i) \) and set the parameter \( \alpha = 3 \). As seen in Table 5, eight of nine predictions are accurate. Therefore, our approach produces a considerably great performance.

4. Conclusion

In this paper, we try to solve the problem of predicting MHC II binding peptides with a novel metric and strategy. Sequence similarity and structural similarity between different MHC molecules are calculated to reorder pseudosequences according to descending similarity, and then a weight calculation formula is used to calculate new pocket profiles. Finally, we use three scoring functions to predict binding cores and evaluate the accuracy of prediction to judge performance of each scoring function. In the experiment, we set a parameter \( \alpha \) in the weight formula. By changing \( \alpha \) value, we can observe different performances of each scoring function. Then, we compare our method with the best function to some popular prediction methods and ultimately find that our method outperforms them in identifying binding cores of HLA-DR molecules.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.
Figure 3: Comparison of different methods by sequence logos of peptides on HLA-DRB1*0101.

Acknowledgments
This work is supported by a grant from the National Science Foundation of China (NSFC 61402326) and Peiyang Scholar Program of Tianjin University (no. 2016XRG-0009).

References
[1] R. M. Zinkernagel and P. C. Doherty, “Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system,” Nature, vol. 248, no. 5450, pp. 701–702, 1974.
[2] P. I. Terasaki, “A brief history of HLA,” Immunologic research, vol. 38, no. 1–3, pp. 139–148, 2007.
[3] K. Maenaka and E. Y. Jones, “MHC superfamily structure and the immune system,” Current Opinion in Structural Biology, vol. 9, no. 6, pp. 745–753, 1999.
[4] J. Robinson, M. J. Waller, S. C. Fail et al., “The IMGT/HLA database,” Nucleic Acids Research, vol. 37, supplement 1, pp. D1013–D1017, 2009.
[5] M. Nielsen, C. Lundegaard, T. Blicher et al., “Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan,” PLoS Computational Biology, vol. 4, no. 7, Article ID e1000107, 2008.
[6] H.-G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanović, “SYFPEITHI database for MHC ligands and peptide motifs,” Immunogenetics, vol. 50, no. 3, pp. 213–219, 1999.
[7] R. N. Germain, “MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation,” Cell, vol. 76, no. 2, pp. 287–299, 1994.
[8] T. Sturniolo, E. Bono, J. Ding et al., “Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices,” Nature Biotechnology, vol. 17, no. 6, pp. 555–561, 1999.
[9] N. Pfeifer and O. Kohlbacher, “Multiple instance learning allows MHC class II epitope predictions across alleles,” Algorithms in Bioinformatics, vol. 5251, pp. 210–221, 2008.
[10] T. J. Kindt, R. A. Goldsby, B. A. Osborne, and J. Kuby, Kuby Immunology, WH Freeman & Company, New York, NY, USA, 2007.
[11] A. Sette, L. Adorini, E. Appella et al., “Structural requirements for the interaction between peptide antigens and I-E molecules,” Journal of Immunology, vol. 143, no. 10, pp. 3289–3294, 1989.
[12] M. Nielsen, C. Lundegaard, T. Blicher et al., “Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan,” PLoS Computational Biology, vol. 4, no. 7, Article ID e1000107, 2008.
[13] A. J. Bordner and H. D. Mittelmann, “MultiRTA: a simple yet reliable method for predicting peptide binding affinities for multiple class II MHC allotypes,” BMC Bioinformatics, vol. 11, article 482, 2010.
[14] P. A. Reche, H. Zhang, J.-P. Glutting, and E. L. Reinherz, “EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology,” Bioinformatics, vol. 21, no. 9, pp. 2140–2141, 2005.
[15] P. A. Reche, J.-P. Glutting, H. Zhang, and E. L. Reinherz, “Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles,” Immunogenetics, vol. 56, no. 6, pp. 405–419, 2004.
[16] A. Sette, L. Adorini, E. Appella et al., “Structural requirements for the interaction between peptide antigens and I-E.
molecules,” The Journal of Immunology, vol. 143, no. 10, pp. 3289–3294, 1989.

[17] M. Nielsen, S. Justesen, O. Lund, C. Lundegaard, and S. Buus, “NetMHCIIpan-2.0—improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure,” Immunome Research, vol. 6, no. 1, article 9, 2010.

[18] M. Nielsen and O. Lund, “NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction,” BMC Bioinformatics, vol. 10, article 296, 2009.

[19] G. L. Zhang, D. S. DeLuca, D. B. Keskin et al., “MULTIPRED2: a computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles,” Journal of Immunological Methods, vol. 374, no. 1-2, pp. 53–61, 2011.

[20] X.-Y. Cheng, W.-J. Huang, S.-C. Hu et al., “A global characterization and identification of multifunctional enzymes,” PLoS ONE, vol. 7, no. 6, Article ID e38979, 2012.

[21] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, and Z. Lin, “An approach for identifying cytokines based on a novel ensemble classifier,” BioMed Research International, vol. 2013, Article ID 686090, 11 pages, 2013.

[22] W. Shen, S. Zhang, and H. Wong, “An effective and efficient peptide binding prediction approach for a broad set of HLA-DR molecules based on ordered weighted averaging of binding pocket profiles,” Proteome Science, vol. 11, p. S15, 2013.

[23] L. Zhang, Y. Chen, H.-S. Wong, S. Zhou, H. Mamitsuka, and S. Zhu, “TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules,” PLoS ONE, vol. 7, no. 2, Article ID e30483, 2012.

[24] P. A. Reche and E. L. Reinherz, “Prediction of peptide-MHC binding using profiles,” Methods in Molecular Biology, vol. 409, pp. 185–200, 2007.

[25] F. Guo, S. C. Li, L. Wang, and D. Zhu, “Protein-protein binding site identification by enumerating the configurations,” BMC Bioinformatics, vol. 13, article 158, 2012.

[26] F. Guo, S. C. Li, and L. Wang, “Protein-protein binding sites prediction by 3D structural similarities,” Journal of Chemical Information and Modeling, vol. 51, no. 12, pp. 3287–3294, 2011.