OWA-PSSM - A Position Specific Scoring Matrix based Method Integrated with OWA Weights for HLA-DR Peptide Binding Prediction

Wen-Jun Shen, Hau-San Wong*
Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
Email: wenjshen@student.cityu.edu.hk
Email: cshswong@cityu.edu.hk
**Corresponding author*

***Abstract*—The binding of foreign peptides to MHC II molecules play a vital role in stimulating CD4+ helper T lymphocytes immune response. We present a pan-specific method, OWA-PSSM, that uses 35 pocket profiles generated by the TEPITOPE method to predict MHC II/peptide binding. Additionally, we develop a novel weighting approach, incorporated with OWA (Ordered Weighted Averaging) weights, for the construction of position specific scoring matrices (PSSMs) for 879 DRB alleles. This method is evaluated on four independent benchmark datasets and is demonstrated to outperform the TEPITOPEpan method. For the DRB alleles covered by TEPITOPE, the prediction performance of OWA-PSSM is comparable with TEPITOPE whereas OWA-PSSM can make prediction for up to 879 DRB alleles, while TEPITOPE can only perform prediction for 50 DRB alleles.**

***Keywords*-MHC II; DRB allele; pan-specific; PSSM; TEPITOPE; RBF kernel; OWA weight**

I. Introduction

MHC I (Major Histocompatibility Complex Class I) and MHC II (Major Histocompatibility Complex Class II) are two general classes of MHC molecules. The physiologic function of these two classes of MHC molecules is similar, i.e. both of them present antigen fragments to T lymphocytes [1]. The predicted accuracy of MHC I/peptide binding is much higher than MHC II/peptide binding, since the binding cleft of MHC I molecules is closed at both ends and will only accommodate peptide fragments with 9 to 11 amino acids. However, the binding cleft of MHC II molecules is open at both ends and is capable of binding to longer peptide fragments with 9 to 25 amino acids. Studying and understanding the mechanisms of peptide presentation by MHC I and II molecules is very important for vaccine design and immunotherapies, and have aroused many researchers' interests in the past three decades.

Both MHC I and II molecules are highly polymorphic. The MHC I molecules are encoded by three different loci on the genome, named A, B and C locus. The MHC II molecules are encoded by an additional one, named D locus, which is partitioned into three smaller loci, called DP, DQ and DR locus. Every locus is able to produce hundreds of different allelic variants which can bind to specific peptide fragments. Because of the diverse polymorphism of MHC molecules and the high cost of MHC/peptide binding measurements based on biochemical approaches, very limited MHC alleles have experimentally-determined binding datasets. For those MHC alleles without any peptide binding information or with only a very small peptide binding dataset, we cannot use allele-specific approach to perform prediction. Pan-specific approach [2] is proposed to deal with this problem which allow accurate prediction for those alleles with rare or no peptide binding data by making use of other alleles with sufficient experimentally measured binding data.

The pioneering and most popular pan-specific method for MHC II molecules is TEPITOPE [3], which was proposed in 1999. It predicts HLA-DR/peptide binding based on 50 virtual matrices that are derived from 35 pocket profiles. Each pocket profile quantitatively describes the binding strength between 20 natural amino acids with a specific pocket. These 35 pocket profiles are generated from 11 HLA-DR alleles in vitro. The basic idea of TEPITOPE is that HLA-DR binding cleft is independent of pocket profiles. Thus, if two HLA-DR alleles have identical amino acids in the same pocket, known as pocket pseudo sequence, they should have the same pocket profile. In recent years, some pan-specific approaches, such as MHCIIMulti, MultiRTA, NetMHCIIpan-2.0, MULTIPRED2 and TEPITOPEpan, that can construct predictor for large-scale MHC II alleles have been proposed.

Our method, which we refer to as OWA-PSSM, is similar to TEPITOPEpan in that it is also built on the TEPITOPE method. However, we further introduce a Gaussian RBF kernel which is integrated with the Amino Acid Distance Matrix [4] to measure the similarity between pocket pseudo-sequences. We also develop a novel weighting approach among pocket profiles through introducing the OWA (ordered weighted averaging) weights [5], [6]. Here, these OWA weights are generated from the exponential probability density function [7]. In our experiments, we will compare

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

the prediction performance of OWA-PSSM with TEPITOPE and TEPITOPEpan. Different from other pan-specific methods mentioned above, these three methods do not rely on training data and thus are much more efficient in performing prediction.

II. Materials and Methods

Four independent benchmark datasets are introduced in Section A. The pan-specific HLA-DR method, OWA-PSSM, is presented in Section B to Section D. A PSSM is constructed by assembling 9 pocket profiles in order. We use two different approaches to generate profiles for pocket 4/6/7/9 and pocket 1/2/3/5/8, respectively.

A. Data

We downloaded the complete set of DRB (HLA-DR β chain) protein files from the IMGT/HLA Sequence Database on April 14th, 2012. After removing 21 non-expressed alleles, there remain 890 DRB alleles. On the basis of IMGT nomenclature, each residue in the DRB alleles is followed by an index number ranging between -29 to 237. In our model, we only need to consider the residues with index number in the range of 9 to 86 since all polymorphic residues used in TEPITOPE occur within this range. There are nine DRB alleles whose residue at 86 is neither Gly nor Val, and these kind of alleles are also excluded in our model. Finally, 879 DRB alleles are covered in our model and collected into a set called $\mathcal{D}$.

The HLA-DR peptide binding dataset described in the SMM-align publication is used to determine the parameters of our model. This dataset covers 14 DRB alleles and 4603 peptides with logarithmic transformed IC50 binding affinities. We use the same dataset as in the case of TEPITOPEpan to search for optimal parameters such that the prediction performance can be compared more fairly. Eight datasets containing binders and non-binders for DRB1*04:01 were downloaded from MHCBench. These eight datasets are used to perform an in-depth evaluation of OWA-PSSM compared with TEPITOPE and TEPITOPEpan. A large-scale dataset described in the NetMHCIIpan-2.0 publication containing 28 HLA-DR alleles and 1164 HLA-DR ligands is used to perform further evaluation. The final dataset consisting of 32 3D X-ray structures of MHC II/peptide binding complexes [2] was retrieved from the PDB database.

B. Generation of Profiles for Pocket 4 6 7 9

OWA-PSSM is a PSSM (position specific scoring matrix) based method. Each DRB allele is associated with a PSSM of size 20 by 9 (20 amino acids by 9 pockets). We adopt the same polymorphic pocket residue indexes as those used in TEPITOPE to generate pocket pseudo-sequences.

Given a specific pocket, pocket 4, 6, 7 or 9, the associated pocket profiles without duplicate generated by TEPITOPE, which we call raw pocket profiles, are collected into a set, called P, $P = \{p_1, p_2, \ldots, p_m\}$ and $|p_i| = 20, i = 1, 2, \ldots, m$ where m is the number of raw pocket profiles and p_i is a 20-dimensional column vector. The pocket pseudo-sequences associated with those raw pocket profiles, called raw pocket pseudo-sequences, are collected into a set R, $R = \{r_1, r_2, \ldots, r_m\}$ and $|r_i| = n, i = 1, 2, \ldots, m$, where n is the length of a pseudo-sequence.

Given a 20 by 20 substitution matrix, each pseudo-sequence is encoded into a $20n$-dimensional vector. The encoded pseudo-sequences associated with the raw pseudo-sequences are collected into a set $V, V = \{v_1, v_2, \ldots, v_m\}$.

For a predicted allele $l, l \in D$, suppose v_l and v_i are two encoded pseudo-sequences at the same pocket, we use Gaussian RBF kernel to compute the similarity score between these sequences:

$$k(v_l, v_i) = e^{-\frac{\|v_l - v_i\|^2}{2\sigma^2}}, v_i \in V. \quad (1)$$

Obviously, $0 < k(v_l, v_i) \leq 1$ and $k(v_l, v_i) = 1$ if and only if two encoded pseudo-sequences are identical. σ is set to 1 by default throughout the paper.

The pocket profile for the predicted DRB allele l is defined as a weighted average over m raw pocket profiles in the set P. OWA weights are applied to determine this weighting vector. After sorting these similarity scores in a descending order, an OWA weight will be associated with a specific ordered position. Essentially, a higher weight should be assigned to a raw pocket profile whose associated pseudo-sequence has higher similarity with that of the predicted allele. A new pocket profile is generated as illustrated in Figure 1.

An OWA operator is defined as follows. Given a k-dimensional vector $U, U = (u_1, u_2, \ldots, u_k)$, we sort the elements of U in descending order, and define the ordered vector as $U^*, U^* = (u_1^*, u_2^*, \ldots, u_k^*)$. An OWA operator is a mapping F from $R^k \rightarrow R$, such that

$$F(u_1, u_2, \ldots, u_k) = w_1 u_1^* + w_2 u_2^* + \ldots + w_k u_k^*$$

and where $\sum_{i=1}^{k} w_i = 1; w_i \in (0, 1)$.

The OWA weight w_i is associated with a particular ordered position i rather than the specific value u_i^* or u_i. The re-ordering is the most important step of the OWA operator. In this paper, we develop a new weighting approach which is inspired by the OWA operator [5].

Here, the OWA weights are defined as follows. The PDF (probability density function) of an exponential distribution is defined as:

$$f(x; \mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, x \geq 0.$$

where $\mu > 0$. The OWA weight distribution can be defined as the discretization of the exponential PDF [7] as follows:

$$F(X = i) = \frac{1}{\mu_m} e^{-\frac{i}{\mu_m}}, i = 1, 2, \ldots, m. \quad (2)$$

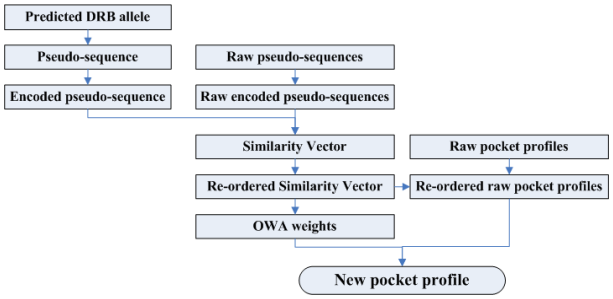

Figure 1. Schematic illustration of a new profile generation approach for pocket 4/6/7/9.

where $\mu_m = \lambda(1+m)$, $\lambda > 0$ controls the steepness of the OWA weight distribution. The OWA weight distribution will become steeper as λ gets smaller. With a small λ, a large weight will be assigned to the maximum value.

The OWA weights are defined as:

$$P(X=i) = \frac{F(X=i)}{\sum_{k=1}^{m} F(X=k)}, i = 1, 2, \ldots, m. \quad (3)$$

Suppose $w_i = P(X=i)$, it satisfies the following two constraints: $\sum_{i=1}^{m} w_i = 1; w_i \in (0,1)$.

We develop an approach similar to the OWA operator in spirit to generate pocket profiles for predicted alleles. The OWA weights generated from the exponential PDF is used as the weighting vector.

For a predicted DRB allele l, let $K_l = (k_{l1}, k_{l2}, \ldots, k_{lm})$, where $k_{li} = k(v_l, v_i), v_i \in V, i = 1, 2, \ldots, m$, and the corresponding raw pocket profiles be $P = \{p_1, p_2, \ldots, p_m\}$. We then sort the elements of K_l in descending order, denote the re-ordered vector of K_l as $K_l^* = (k_{l1}^*, k_{l2}^*, \ldots, k_{lm}^*)$ and the associated OWA weighting vector as W, $W = (w_1, w_2, \ldots, w_m)$. The raw pocket profiles corresponding to the re-ordered vector K_l^* is P^*, $P^* = \{p_1^*, p_2^*, \ldots, p_m^*\}$, and thus the pocket profile for the predicted allele is defined as:

$$\overline{p}_l = w_1 p_1^* + w_2 p_2^* + \ldots + w_m p_m^*. \quad (4)$$

We note that $\overline{p}_l$ is a general form of the pocket profile for a predicted allele. Hence, for a specific pocket u, $\overline{p}_l$ is denoted as $\overline{p}_l^u, u = 4, 6, 7, 9$, where

$$\overline{p}_l^u = \begin{cases} w_1 p_1^* + w_2 p_2^* + \ldots + w_{11} p_{11}^*, & u = 4, 7, \\ w_1 p_1^* + w_2 p_2^* + \ldots + w_6 p_6^*, & u = 6, 9. \end{cases} \quad (5)$$

C. Generation of Profiles for Pocket 1 2 3 5 8

We preserve the merits of TEPITOPE by applying a similar approach to determine the pocket profiles for pocket 1, 2, 3, 5 and 8. For pocket 1, if the 86th residue of a DRB allele is Gly, the pocket profile is determined to be 0 for aromatic amino acids (Phe, Trp, Tyr) and -1 for aliphatic amino acids (Ile, Leu, Met, Val). On the other hand, if this residue is Val, then 0 is assigned to aliphatic, and -1 to aromatic. Similar to ProPred [8], -999 is assigned to other amino acids. This is to indicate that aromatic and aliphatic amino acids are common residues at position 1 of a 9-mer binding core. For pockets 2 and 3, the pocket profile is identical among all alleles [8]. For pockets 5 and 8, since they have little influence on the binding affinity, the corresponding profiles are set to zero vectors. For a predicted DRB allele l, we denote the associated pocket profiles as $\overline{p}_l^u, u = 1, 2, 3, 5, 8$.

D. Position Specific Scoring Matrices

In sections B and C, we have defined nine pocket profiles $\overline{P} = \{\overline{p}_l^1, \overline{p}_l^2, \ldots, \overline{p}_l^9\}$ for any allele $l, l \in \mathcal{D}$.

The PSSM for any allele $l, l \in \mathcal{D}$, is defined as:

$$PSSM_l = [\overline{p}_l^1, \overline{p}_l^2, \ldots, \overline{p}_l^9]. \quad (6)$$

where $\overline{p}_l^u, u = 1, 2, \ldots, 9$ is a 20-dimensional column vector.

E. Statistical Tests

The one-tailed binomial test is used for statistical comparison. The P value is computed based on the number of times OWA-PSSM outperforms the other (omitting ties), and the comparison is deem to be statistically significant when p is less than 0.05.

III. Results

There are two types of parameters in our model: those associated with a substitution matrix and a control parameter λ. The dataset described in the SMM-align was used to study the effect on prediction performance through varying these parameters. The experiments were performed on 50 symmetric substitution matrices downloaded from the AAindex database and the control parameter λ was chosen from $\{0.01 \times n : n = 1, 2, \ldots, 100\}$. In the experiment, the Amino Acid Pair Distance Matrix [4] gives the best prediction performance corresponding to an AUC value of 0.743 at $\lambda = 0.06$. We observe that a λ value less than 0.1 leads to a better performance than a λ value larger than 0.1. This is supported in part by the observation that a larger weight should be assigned to a raw pocket profile whose pseudo-sequence is more similar to the predicted allele's pseudo-sequence. As a result, the Amino Acid Pair Distance Matrix and $\lambda = 0.06$ are used to generate PSSMs for 879 DRB alleles in $\mathcal{D}$. The prediction performance for each allele in the Nielsen dataset is given in Table I. OWA-PSSM performs slightly better than the TEPITOPE and TEPITOPEpan methods in this dataset.

The MHCBench dataset and the HLA-DR ligand dataset were then used to compare the performance of OWA-PSSM with TEPITOPE and TEPITOPEpan. The final dataset which contains 32 X-ray structures is used to evaluate the capability of OWA-PSSM to identify the binding core of binders. The prediction performance is given in Table II, III and IV,

respectively. OWA-PSSM thus outperforms TEPITOPEpan significantly ($p < 0.05$, Binomial test, excluding ties) in both MHCBench and HLA-DR ligand datasets. Furthermore, OWA-PSSM correctly identifies the binding cores of all 32 MHC II/peptide complexes whereas TEPITOPEpan misidentifies two. Table II, III and IV also show that the prediction performance of OWA-PSSM is comparable with TEPITOPE. However, OWA-PSSM can perform prediction for up to 879 DRB alleles whereas TEPITOPE can predict for only 50.

Table I
PREDICTION PERFORMANCE OF THE OWA-PSSM METHOD COMPARED TO THE TEPITOPE AND TEPITOPEPAN METHODS ON THE NIELSEN DATASET.

Method	TEPITOPE	TEPITOPEpan	OWA-PSSM
AUC*	0.736	0.732	0.743
# Best	1	5	9
p-value	0.1133	0.3953	N/A

The prediction performance is given in terms of the AUC (*average per allele) with a binding threshold of 500nM. $\#Best$ shows the number of best predicted alleles for each approach. The comparison is statistically significant when p is less than 0.05. The PSSMs of TEPITOPE were obtained from its public server ProPred (http://www.imtech.res.in/raghava/propred/page4.html). The PSSMs of TEPITOPEpan were obtained from its web server (http://www.biokdd.fudan.edu.cn/Service/TEPITOPEpan/TEPITOPEpan.html).

Table II
PREDICTION PERFORMANCE IN TERMS OF THE AUC ON THE MHCBENCH DATASET.

Method	TEPITOPE	TEPITOPEpan	OWA-PSSM
AUC*	0.730	0.718	0.724
# Best	5	0	4
p-value	0.8555	0.0039	N/A

*average per set

Table III
PREDICTION PERFORMANCE IN TERMS OF THE AUC ON THE HLA-DR LIGAND DATASET.

Method	TEPITOPE	TEPITOPEpan	OWA-PSSM
AUC*	0.801	0.756	0.778
# Best	10	5	14
p-value	0.9755	0.0178	N/A

*average per allele

Table IV
COMPARISON OF OWA-PSSM WITH TEPITOPE AND TEPITOPEPAN IN IDENTIFYING MHC II-PEPTIDE BINDING CORES.

Method	TEPITOPE	TEPITOPEpan	OWA-PSSM
# Correct	30/30*	30/32	32/32

*There are 30 out of 32 alleles covered by TEPITOPE.

IV. DISCUSSION AND CONCLUSION

In this paper, we have developed a pan-specific method, called OWA-PSSM, for MHC II/peptide binding prediction. Our method is a natural extension of TEPITOPE which is a pioneering and best known pan-specific prediction method for MHC II/peptide binding. We preserved the merits of TEPITOPE, and developed a novel weighting approach, through introducing OWA weights, to enable OWA-PSSM to perform prediction for 879 DRB alleles.

For the DRB alleles covered by TEPITOPE, the prediction performance of OWA-PSSM is demonstrated to be comparable with TEPITOPE. However, OWA-PSSM can perform prediction for up to 879 DRB alleles whereas TEPITOPE can perform prediction for only 50 DRB alleles. Compared with TEPITOPEpan which is a recently presented pan-specific approach, the prediction performance of OWA-PSSM is slightly better in quantitative prediction of peptide binding and identification of the binding cores. In particular, for the prediction of HLA-DR ligands and T-cell epitopes, OWA-PSSM significantly outperforms TEPITOPEpan.

ACKNOWLEDGMENT

The work described in this paper was supported by a grant from the City University of Hong Kong [Project No. 7002771].

REFERENCES

[1] J. Neefjes, M. Jongsma, P. Paul, and O. Bakke, "Towards a systems understanding of MHC class I and MHC class II antigen presentation," *Nature Reviews Immunology*, vol. 11, no. 12, pp. 823–836, 2011.

[2] L. Zhang, Y. Chen, H. 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, p. e30483, 2012.

[3] T. Sturniolo, E. Bono, J. Ding, L. Raddrizzani, O. Tuereci, U. Sahin, M. Braxenthaler, F. Gallazzi, M. Protti, F. Sinigaglia *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.

[4] T. Miyata, S. Miyazawa, and T. Yasunaga, "Two types of amino acid substitutions in protein evolution," *Journal of Molecular Evolution*, vol. 12, no. 3, pp. 219–236, 1979.

[5] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.

[6] D. Filev and R. Yager, "On the issue of obtaining OWA operator weights," *Fuzzy Sets and Systems*, vol. 94, no. 2, pp. 157–169, 1998.

[7] R. Sadiq and S. Tesfamariam, "Probability density functions based weights for ordered weighted averaging (OWA) operators: an example of water quality indices," *European Journal of Operational Research*, vol. 182, no. 3, pp. 1350–1368, 2007.

[8] H. Singh and G. Raghava, "ProPred: prediction of HLA-DR binding sites," *Bioinformatics*, vol. 17, no. 12, pp. 1236–1237, 2001.