Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

A Meta-predictor for MHC Class II Binding Peptides Based on Naïve Bayesian Approach

Lei Huang[¶], Oleksiy Karpenko[¶], Naveen Murugan[¶], and Yang Dai*

***Abstract*—Prediction of class II MHC-peptide binding is a challenging task due to variable length of binding peptides. Different computational methods have been developed; however, each has its own strength and weakness. In order to provide reliable prediction, it is important to design a system that enables the integration of outcomes from various predictors. In this paper, we introduce a procedure of building such a meta-predictor based on Naïve Bayesian approach. The system is designed in such a way that results obtained from any number of individual predictors can be easily incorporated. This meta-predictor is expected to give users more confidence in the prediction.**

I. Introduction

T cell-mediated immune responses are initiated by the activation of effector T cells. The activation process requires the recognition of the complex formed between an antigen peptide and a major histocompatibility complex (MHC) protein by the T cell receptor. The identification of peptides that bind to MHC molecules plays a crucial role in understanding the mechanisms of both humoral and adaptive immunity as well as developing epitope-based vaccines. Experiments for measuring the binding affinities of peptides to MHC molecules are time consuming and expensive. It is a prohibitive task to identify potential binding peptides from the host and pathogen proteins on a genome-wise scale. Therefore, considerable efforts have been made on the development of computational tools for the identification of MHC-binding peptides [1],[2].

Two major types of MHC molecules are involved in the peptide binding process. MHC class I molecules present endogenous antigens (e.g. viral peptides or tumor antigens synthesized within the cytoplasm of a cell) to CD8+ cytotoxic T cells. MHC class II molecules, on the other hand, present exogenously derived proteins (e.g. bacterial proteins or viral capsid proteins) through antigen presenting cells (APC) to CD4+ helper T cells [3]. Generally, antigen peptides that bind to both MHC class I and class II molecules are approximately nine amino acid residues long. However, the peptide-binding groove of a MHC class II molecule is open at both ends, which makes it capable of accommodating longer peptides of 10-30 residues [4]-[6].

The length variability complicates the prediction of peptide-MHC class II binding. However, analyses of the binding motif and the structure of peptide-MHC class II complexes have suggested that a core of 9 residues within a peptide is essential for peptide-MHC binding. Computational methods for the prediction include simple binding motifs [7], [8], quantitative matrices [9], [10], [11], [12], hidden Markov models [13], artificial neural networks [14], [15], support vector machines [16], and linear programming [17]. Some of these methods require a preprocessing step to align binding sequences with various lengths for the identification of subsequences of the binding cores. Since each method has its own strength and weakness, it is hard for an immunologist to select a single method from the pool of existing predictors. Therefore, a system that produces reliable prediction through the integration of outcomes from major prediction methods is in clear need.

In this paper, the procedure for building such a system based on the Naïve Bayesian [18] approach is presented. The Bayesian framework has the flexibility to incorporate any predictor that makes prediction from a computed score correlated with the binding affinity of MHC class II peptides. Three individual predictors, i.e., ProPred, the Gibbs sampler, and the LP model were selected based on their availability and overall performance.

ProPred, designed by Singh and Raghava [10], applied the quantitative matrices from 51 HLA-DR alleles for the prediction of MHC class II binding peptides. These matrices were generated from a pocket profile database described by Sturniolo et al. [9] and covered the majority of human HLADR specificity.

Nielsen, et al. [11] proposed an advanced motif sampler method based on the Gibbs sampling technique, which efficiently samples the possible alignment space of binder sequences. For each alignment a log-odds weight matrix was calculated for the identified binding core subsequences. This matrix serves as the position-specific scoring matrix for the computation of a score for a nonamer.

Motivated by a text mining model designed for building a classifier from labeled and unlabeled examples, Murugan and Dai [17] developed an iterative supervised learning

Manuscript received April 3, 2006. This work was supported in part by the NIH under Grant 1 R03 AI069391-01.

Y. Dai is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607 USA. (phone: 312-413-1487; fax: 312-413-2018; e-mail: yangdai@ uic.edu).

L. Huang is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607 USA. (e-mail: lhuang7@ uic.edu).

O. Karpenko is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607 USA. (e-mail: okarpe2@uic.edu).

N. Murugan is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607 USA. (e-mail: nmurug1@uic.edu).

[¶]These authors contributed equally

model for the prediction of MHC class II binding peptides. The iterative learning model, based on linear programming (LP), enables the use of non-binder information for the detection of the binding cores from a set of putative binding cores and for the construction of the predictor simultaneously. The outcome of this predictor is a position specific weight matrix that can score amino acids at each position of a nonamer.

II. Materials

In our study, datasets of binding and nonbinding peptides for specific MHC class II alleles were obtained from two databases: AntiJen [19] and MHCBN [20]. Considering the size of training set, 9 alleles were selected: HLA-DRB1*0101, HLA-DRB1*0301, HLA-DRB1*0401, HLA-DRB1*0701, HLA-DRB1*0802, HLA-DRB1*1101, HLA-DRB1*1302, HLA-DRB1*1501, and HLA-DRB5*0101. These peptides were preprocessed to reduce the redundancy. The numbers of MHC class II binding peptides and nonbinding peptides in each allele used for building the Bayesian model are shown in TABLE I. The predictors chosen were ProPred, the Gibbs sampler, and the LP model. Every predictor can score the binding ability for an individual peptide. The scores from all predictors for all peptides in a particular training set were used to build the Bayesian model.

TABLE I
Numbers (after homology reduction) of MHC class II binding and nonbinding peptides used for building the Bayesian predictor.

MHC class II alleles	Number of binding peptides	Number of non-binding peptides
HLA-DRB1*0101	395	119
HLA-DRB1*0301	394	199
HLA-DRB1*0401	891	186
HLA-DRB1*0701	222	80
HLA-DRB1*0802	86	79
HLA-DRB1*1101	365	120
HLA-DRB1*1302	164	58
HLA-DRB1*1501	493	79
HLA-DRB5*0101	366	58

III. Methods

The Bayesian predictor is trained based on the prediction outcomes obtained from each of the individual predictors on the set of training peptides. The system is flexible to incorporate results from any number of predictors and we can accommodate as many as m predictors. In general, the requirement for each predictor is the generation of a score for a given peptide sequence. This score of a peptide is designated as the highest value among all scores that are assigned to the overlapping nonamers of the peptide by a predictor. A peptide is predicted as a binder (*resp.* nonbinder) if this score is above (*resp.* below) a prescribed threshold value.

In order to build a Bayesian predictor, we first prepared a training dataset for each allele. Any peptide sequence with length less than nine residues or with undetermined residues in certain positions was discarded. Then the dataset was reduced to account for redundancy. This step was necessary to prevent overestimation of the performance of a predictor. After the reduction there were no two peptide sequences in the set with sequence identity >90% over an alignment of length at least nine residues. For each individual predictor, the predictive score for each peptide in the training set (including binding and nonbinding sequences) was obtained. These scores formed the input set from which the Bayesian predictor was built.

For each predictor a set of threshold values that produce distinct pairs of sensitivity and specificity on the training sets is determined. The sensitivity and specificity are defined as TP/(TP+FN) and TN/(TN+FP), respectively, where TP and FN are the respective numbers of predicted binders and nonbinders which are true binders; TN and FP are respective numbers of predicted nonbinders and binders which are true nonbiners. Upon the completion of this step, a set of threshold values for predictor j was obtained, say $\delta^j = (\delta_1^j,...,\delta_{t_j}^j)$, j=1,…,m, where t_j is the number of possible threshold values with the above property for predictor j.

In order to build the input data to training the Bayesian classifier, the best combination $\delta^* = (\delta^{*1},...,\delta^{*m})$ of threshold values has to be determined, where each δ^{*j} (j=1,…,m) is the selected threshold value for predictor j. This combination was determined by finding the highest average area under receiver operating characteristic curve (AROC) value for the Bayesian predictor with a k-fold cross-validation procedure. To perform the k-fold cross-validation, the ratio between the number of binders and the number of nonbinders in all k folds was kept approximately equal.

For each combination of threshold values $(\delta_{i_1}^1,...,\delta_{i_m}^m)$, a prediction outcome table was setup for the (k-1)-folds of the training peptides, where $\delta_{i_j}^j$ is the i_jth threshold value for predictor j, j=1,…,m and i_j=1,…,t_j. This table is of size $n \times m$, where n is the number of peptides in the training folds. The outcome obtained from predictor j for a peptide is denoted by a binary number f_j: f_j=1 if the peptide is predicted as binder, f_j=0 otherwise. Accordingly, the prediction outcome obtained from the m predictors for each peptide will be coded by a binary string $f_1f_2..f_m$. The probability

table for the Bayesian predictor is built from the $n \times m$ table described above. Let y_i denote the label of each peptide: y_i=1 if it is a binder, y_i=-1 if it is not a binder. The probabilities for each value f_j of the m features for the binder class and the nonbinder class were computed as follows.

$$p(f_j = 1 \mid \text{binder class}) = \frac{\sum_{i:y_i=1} I(f_{ij}=1)}{\text{total number of binders}}, \quad j = 1,...,m, \tag{1}$$

$$p(f_j = 0 \mid \text{binder class}) = \frac{\sum_{i:y_i=1} I(f_{ij}=0)}{\text{total number of binders}}, \quad j = 1,...,m, \tag{2}$$

$$p(f_j = 1 \mid \text{nonbinder class}) = \frac{\sum_{i:y_i=-1} I(f_{ij}=1)}{\text{total number of nonbinders}}, \quad j = 1,...,m, \tag{3}$$

and

$$p(f_j = 0 \mid \text{nonbinder class}) = \frac{\sum_{i:y_i=-1} I(f_{ij}=0)}{\text{total number of nonbinders}}, \quad j = 1,...,m. \tag{4}$$

where $I(\cdot) = 1$ if the condition in the parenthesis is true; $I(\cdot) = 0$ otherwise. Note that (i) the total numbers of binders and nonbinders are respectively those in the (k-1) training folds; (ii) the index i in the numerator of each formula runs through all peptides in the (k-1) training folds; and (iii) f_{ij} is the prediction by predictor j for the nonamer with the highest score from peptide i.

For each overlapping nonamer s_i of a peptide x from the testing fold, the ratio of probabilities was computed

$$R_i = \frac{p(f=1 \mid s_i)}{p(f=0 \mid s_i)} = \frac{\prod_{j=1}^{m} p(f_{ij} \mid \text{binder class})}{\prod_{j=1}^{m} p(f_{ij} \mid \text{nonbinder class})}, \tag{5}$$

and the highest one was selected as the ratio R_x of the peptide x, where f_{ij} was the prediction outcome obtained from predictor j for nonamer s_i. This formula is a straightforward application of the Bayesian rule, without the inclusion of the ratio of prior probabilities p(binder) and p(nonbinder). The influence of prior probabilities on prediction will be implicitly considered through threshold of ratio R_i. With a prescribed threshold δ_B for the Bayesian predictor, the peptide was predicted as a binder if R_x was greater than δ_B, otherwise as a nonbinder. Varying the threshold values for δ_B, the AROC value for the current testing fold was calculated.

This same procedure was repeated for the other k-1 sets of the different training and testing folds to obtain the average AROC value from the k testing folds. After obtaining the average AROC values for all possible combinations of $(\delta_{i_1}^1,...,\delta_{i_m}^m)$, the best combination $\delta^* = (\delta^{*1},...,\delta^{*m})$ was identified that corresponded to the highest average AROC value. The final Bayesian predictor was constructed by using the outcome table determined from the best combination of threshold $\delta^* = (\delta^{*1},...,\delta^{*m})$ for the entire training peptides. By varying the threshold values for δ_B, the corresponding sensitivity and specificity for the entire training set were obtained and the AROC value was computed. The general framework of building a Bayesian predictor is summarized in Fig. 1.

The threshold δ_B for the final Bayesian classifier has to be determined for the prediction of the MHC class II binding ability of an unknown peptide. The recommended value for δ_B is that the sensitivity and specificity of the predictor are approximately equal. However, it is also possible to select a value for δ_B at which the sensitivity is higher than the specificity; or conversely, choose a value for δ_B at which the specificity is higher than the sensitivity. The Bayesian predictor predicts a peptide as a binder if the highest value among the ratios $p(f=1 \mid s_i)/p(f=0 \mid s_i)$ for all overlapping nonamers s_i from the peptide is great than δ_B; otherwise predicts it as nonbinder.

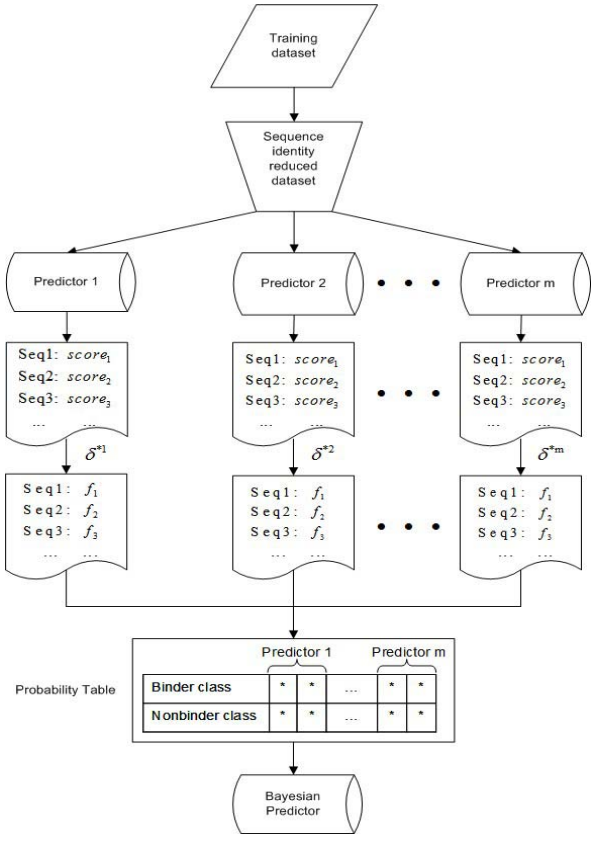

Fig. 1. Illustration of the framework for building a Bayesian predictor.

IV. Results

The performance of the Bayesian predictor to predict the training set was evaluated by using a 5 fold cross-validation technique on the training dataset, since there were no independent test set for all the alleles. The result was compared with that of the three individual predictors (i)

ProPred, (ii) Gibbs sampler, and (iii) the LP predictor. TABLE II shows the comparison of the performance of the different methods over the various alleles. Details can be found in TABLE II.

TABLE II
AROC values of the different methods

MHC class II alleles	Gibbs	LP	Propred	Bayesian
HLA-DRB1*0101	0.766	0.940	0.847	0.929
HLA-DRB1*0301	0.613	0.769	0.619	0.730
HLA-DRB1*0401	0.724	0.849	0.758	0.804
HLA-DRB1*0701	0.764	0.973	0.775	0.957
HLA-DRB1*0802	0.805	0.975	0.762	0.941
HLA-DRB1*1101	0.717	0.859	0.674	0.817
HLA-DRB1*1302	0.796	0.981	0.771	0.931
HLA-DRB1*1501	0.644	0.925	0.677	0.886
HLA-DRB5*0101	0.863	0.962	0.714	0.949

Of the 9 alleles HLA-DRB1*0101, HLA-DRB1*0301, HLA-DRB1*0401, HLA-DRB1*0701, HLA-DRB1*0802, HLA-DRB1*1101, HLA-DRB1*1302, HLA-DRB1*1501, and HLA-DRB5*0101 on which the methods were tested, the AROC values of the best performing individual method were 0.940, 0.769, 0.849, 0.973, 0.975, 0.859, 0.981, 0.925 and 0.962, while the AROC values of the Bayesian method were 0.929, 0.730, 0.804, 0.957, 0.941, 0.817, 0.931, 0.886 and 0.949 respectively. Thus the Bayesian predictor had an average AROC value of 3% less that of the best performing method. However, the Bayesian predictor is expected to give users more confidence in the prediction results, by using the combined information from different methods. The corresponding web server for the Bayesian Prediction model can be accessed at http://array.bioengr.uic.edu/cgi-in/mhc2srv/testing.web.pl.

V. Conclusion

A new MHC class II binding peptide prediction method combining the existing methods has been developed based on the Bayesian Model. The performance of the new model is comparable to that of other models. However, the model is more reliable as it combines the strengths from the various methods. A future step is the introduction of weight to each method for further improvement of performance of the Bayesian model.

References

[1] D. R. Flower, "Vaccines in silico - the growth and power of immunoinformatics," The Biochemist, 2004, vol. 26 pp. 7-20.

[2] A. S. De Groot and J. A. Berzofsky, "From genome to vaccine--new immunoinformatics tools for vaccine design", Methods, 2004, vol. 34, no. 4, pp. 425-428.

[3] P. Parham, The Immune System, New York: Garland Science, 2005.

[4] F. Castellino, G. Zhong, and R. N. Germain, "Antigen presentation by MHC class II molecules: invariant chain function, protein trafficking, and the molecular basis of diverse determinant capture", Hum. Immunol,. 1997, vol. 54, pp. 159-169.

[5] A. Sette, S. Buus, E. Appella, J. A. Smith, R. Chesnut, C. Miles, S. M. Colon, and H. M. Grey, "Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis", Proc. Natl. Acad. Sci. U. S. A., 1989, vol. 86, pp. 3296-3300.

[6] H. Max, T. Halder, H. Kropshofer, M. Kalbus, C. A. Muller, and H. Kalbacher, "Characterization of peptides bound to extracellular and intracellular HLA-DR1 molecules", Hum. Immunol., 1993, vol. 38, pp. 193-200.

[7] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovic, "SYFPEITHI: database for MHC ligands and peptide motifs", Immunogenetics, 1999, vol. 50, pp. 213-219.

[8] F. Borras-Cuesta, J. Golvano, M. Garcia-Granero, P. Sarobe, J. Riezu-Boj, E. Huarte, and J. Lasarte, "Specific and general HLA-DR binding motifs: comparison of algorithms", Hum. Immunol., 2000, vol. 61, pp. 266-278.

[9] T. Sturniolo, E. Bono, J. Ding, L. Raddrizzani, O. Tuereci, U. Sahin, M. Braxenthaler, F. Gallazzi, M. P. Protti, F. Sinigaglia, and J. Hammer, "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices", Nat. Biotechnol., 1999, vol. 17, pp. 555-561.

[10] H. Singh, and G. P. Raghava, "ProPred: prediction of HLA-DR binding sites", Bioinformatics, 2001, vol. 17, pp. 1236-1237.

[11] M. Nielsen, C. Lundegaard, P. Worning, C. S. Hvid, K. Lamberth, S. Buus, S. Brunak, and O. Lund, "Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach", Bioinformatics, 2004, vol. 20, pp. 1388-1397.

[12] O. Karpenko, J. Shi, and Y. Dai, "Prediction of MHC class II binders using the ant colony search strategy", Artif. Intell. Med., 2005, vol. 35, pp. 147-156.

[13] R. Kato, H. Noguchi, H. Honda, and T. Kobayashi, "Hidden Markov model-based approach as the first screening of binding peptides that interact with MHC class II molecules", Enzyme Microb. Technol., 2003, vol. 33, pp. 472-481.

[14] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, and L. Harrison, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network", Bioinformatics, 1998, vol. 14, pp. 121-130.

[15] M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemoller, K. Lamberth, S. Buus, S. Brunak, and O. Lund, "Reliable prediction of Tcell epitopes using neural networks with novel sequence representations", Protein Sci., 2003, vol. 12, pp. 1007-1017.

[16] M. Bhasin, and G. P. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence", Bioinformatics, 2004, vol. 20, pp. 421-423.

[17] N. Murugan, and Y. Dai, "Prediction of MHC class II binding peptides based on an iterative learning model", Immunome Res., 2005, vol. 1, pp. 6.

[18] S. Theodoridis, and K. Koutroumbas, Pattern Recognition, San Diego, CA, Academic Press, 1999.

[19] C. P. Toseland, D. J. Clayton, H. McSparron, S. L. Hemsley, M. J. Blythe, K. Paine, I. A. Doytchinova, P. Guan, C. K. Hattotuwagama, and D. R. Flower, "AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data", Immunome Res., 2005, vol. 1, pp. 4.

[20] M. Bhasin, H. Singh, and G. P. Raghava, "MHCBN: a comprehensive database of MHC binding and non-binding peptides", Bioinformatics, 2003, vol. 19, pp. 665-666.

[21] J. A. Swets, "Measuring the accuracy of diagnostic systems", Science, 1998, vol. 240, pp. 1285-1293.