A Comparison of Several Feature Encoding Techniques for MHC Class I Binding Prediction

Murat Gök

***Abstract*—Deciphering the understanding of T cell epitopes is critical for vaccine development. As recognition of specific peptides bound to Major Histocompatibility Complex (MHC) class I molecules, cytotoxic T cells are activated. This is the major step to initiate of immune system response. Knowledge of the MHC specificity will enlighten the way of diagnosis, treatment of pathogens as well as peptide vaccine development. So far, a number of methods have been developed to predict MHC/peptide binding. In this paper, several encoding schemes were performed to predict MHC/peptide complexes. The tests have been carried out on comparatively large HLA-A and HLA-B allele peptide three binding datasets extracted from the Immune Epitope Database and Analysis resource (IEDB). Experimental results show OETMAP encoding technique leads to better classification performance than other amino acid encoding schemes on a standalone classifier.**

***Index Terms*—Epitope prediction, major histocompatibility complex class I, feature encoding, peptide classification.**

I. Introduction

MHC class-I and II antigens are of immense importance to the immune system. MHC molecules (also known as the Human Leukocyte Antigen (HLA) molecules in humans) undertake the key dialogs between T cells and other cells of the body. Firstly, antigenic peptides are bound in an extended conformation within the grooves of MHC molecules, which feature pockets into which anchoring peptide side chains can fit, in the cytoplasm [1]. Secondly, MHC molecules present peptides to T Helper Lymphocytes (THL) and Cytotoxic T Lymphocytes (CTL) on the cell surface. The recognition of presented peptides by CTL cells triggers an immune response and is termed T-cell epitopes. In this way, virally infected cells, pathologically mutated cells and tumor cells are discriminated from healthy cells. The activation of CTL in the immune system requires presentation of endogenous antigenic peptides by MHC class-I molecules [2]. Identification of epitopes and peptides that can bind MHC molecules evoke the design of peptide based vaccine and immunotherapy [3]. Occurrence of MHC/peptide binding that initiates an immune response is in the range of 0.1-5% for any given protein of which some 20% remain functionally relevant [4]. Hence, computational prediction of MHC/peptide binding can save experimental efforts and time.

In the prediction of MHC specificity, sequence based and structure based methods were used for classification. If the experimental data is sufficient, sequence-based methods are more efficient than structure-based methods. The core binding motif of both MHC I and MHC II is composed of almost nine amino acids [5]. Therefore, the specificity of an MHC I molecule can be analyzed from a set of 9-mer peptides known to bind to a given allele.

In this paper, five encoding techniques have been evaluated with linear support vector machines (LSVM) algorithm for MHC binding specificity. Also, for the first time, area under receiver operating characteristic curve (AUC) performances of three feature encoding techniques (orthonormal encoding + frequency based, residue couples, Taylor's venn diagram) for each allele have been given in detail.

Manuscript received October 24, 2012; revised January 26, 2013.

Murat Gök is with the Computer Engineering, Yalova University, 77100, Yalova, Turkey (e-mail: muratgok@ gmail.com, murat.gok@yalova.edu.tr).

II. Methods

A. Feature Encoding Methods

Feature extraction process defines a mapping from the original representation space into a new space where the classes are more easily separable. The goal of feature extraction is to distill the pattern data into a more concentrated and manageable form. This will reduce the classifier complexity, increasing in most cases classifier accuracy [6]. The evaluated feature extraction methods are given in brief terms below.

The first is OE which is a common encoding technique. According to OE, each amino acid symbol P_i in a peptide is replaced by an orthonormal vector $d_i = (\delta_{i1}, \delta_{i2}, ..., \delta_{i20})$ where is the Kronecker delta symbol. Then, each P_i is then represented by a 20-bit vector, 19 bits are set to zero and 1 bit is set to one based on alphabetic order of amino acids. Each d_i vector is orthogonal to all other d_i vectors and P_i can be any one of the twenty amino acids [7]. Each nonamer thereby is represented by a vector of 180 bits. The main drawback of OE technique is that OE binary feature vectors result in information loss.

Another common approach is the frequency based (FB) method. In this method, weight of each amino acid P_i in a peptide is determined and then combined by OE. In this way, vector d_i is multiplied by the weight of amino acid P_i. FE preserves the original number of attributes.

Zvelebil [8] proposed a new encoding method based on Taylor's Venn-diagram (TVD) [9] which describes the membership of an amino acid to one of ten classes as a binary vector. The Zvelebil-encoding technique utilizes

physicochemical properties of amino acids without high dimensionality.

In [10], authors inspired by Chou's quasi-sequence-order model and Yuan's Markov chain model and developed Residue-Couple (RC) encoding technique. RC model takes into account not only the amino acid consecutive pairs but also the gapped amino acid pairs corresponding, respectively.

Last encoding technique we have re-implemented is OETMAP [11] which combines the sequence order of the residue composition based on OE and the representation of various relationships of residue based on Taylor's venn-diagram (TVD). In other words, OETMAP is a conjunction of OE and TVD methods which are complementary to each other.

B. Support Vector Machines

SVM is an effective discriminative classification method of statistic learning theory and in recent times, it is successively applied by a number of other researchers. SVM aims to find the maximum margin hyperplane to separate two classes of patterns. A transform to map nonlinearly, the data into a higher dimensional space allows a linear separation of classes which could not be linearly separated in the original space. The objects that are located on these two hyperplanes are the so-called support vectors. The maximum margin hyperplane, which is uniquely defined by the support vectors, gives the best separation between the classes [12].

III. Results

A. Experimental Results

We conducted our tests on three up-to-date data sets (F, I, S) composed of sequences of a set of 9-mer peptides known to bind to a given allele. Dataset F includes all available binders and non-binders in IEDB, dataset I includes only weak binders (50 nM to 500 nM binding affinity) and non-binders (500 nM to 1000 nM binding affinity), and dataset S included only strong binders (less than 10 nM binding affinity) and very clear non-binders (greater than 10,000 nM binding affinity) as outlined in [13].

10-fold cross validation (10-fold CV) testing protocol is applied to evaluate the performance of the methods in terms of area under ROC Curve (AUC) averaged over 10 experiments on datasets. In a cross-validation run, the 10 folds are randomly created [14]. In 10-fold CV, the encoding scheme methods are trained using 90 % of the data and the remaining 10 % of the data are used for testing of the methods. This process is repeated 10 times so that each peptide in datasets is used once. The 10 folds used in the training are different from the 10 folds used in the testing. Then the average AUC of the each method over these 10 turns are obtained. The performance of proposed feature encoding methods on dataset F, dataset I, and dataset S is shown in Table 1, Table 2 and Table 3, respectively, by means of AUC which is defined as the area under the receiver operating characteristic curve (ROC) where a ROC curve is plotted as the number of true positives as a function of false positives for varying classification thresholds to describe the performance of a model across the entire range of classification thresholds [15]. OE, combining the OE representation with the frequency based method (OE+FB), RC, TVD and OETMAP methods have been evaluated.

TABLE I: Predictive AUC Performance Of Encoding Techniques on Dataset F

	Allele	OE	OE+FB	RC	TVD	OETMAP
1	A0101	0.945	0.824	0.804	0.921	0.94
2	A0201	0.947	0.886	0.831	0.929	0.947
3	A0202	0.895	0.821	0.76	0.876	0.896
4	A0203	0.891	0.824	0.726	0.876	0.89
5	A0206	0.911	0.842	0.781	0.886	0.913
6	A0301	0.922	0.803	0.75	0.906	0.924
7	A1101	0.764	0.744	0.687	0.8	0.788
8	A2402	0.765	0.741	0.688	0.799	0.794
9	A2601	0.823	0.758	0.647	0.806	0.819
10	A3101	0.91	0.808	0.749	0.908	0.912
11	A3301	0.888	0.752	0.725	0.884	0.886
12	A6801	0.842	0.734	0.658	0.816	0.848
13	A6802	0.866	0.807	0.753	0.844	0.871
14	B0702	0.94	0.917	0.802	0.945	0.942
15	B0801	0.829	0.748	0.665	0.863	0.876
16	B1501	0.9	0.818	0.663	0.887	0.91
17	B2705	0.934	0.91	0.742	0.946	0.933
18	B3501	0.839	0.777	0.721	0.835	0.843
19	B4001	0.905	0.847	0.765	0.913	0.9
20	B4402	0.757	0.659	0.71	0.787	0.782
21	B4403	0.681	0.676	0.653	0.677	0.691
22	B5101	0.793	0.793	0.756	0.828	0.784
23	B5301	0.846	0.751	0.748	0.865	0.858
24	B5801	0.914	0.799	0.667	0.91	0.929
Average AUC		**0.863**	**0.793**	**0.727**	**0.863**	**0.87**

Table I reports that OETMAP outperforms the competing encoding techniques considered for dataset F with the value of 0.87. We notice that OETMAP combines the both effectiveness of OE and TVD. The remedy of discerning between binding and non-binding peptides is increased with the classifier thereof. Note that RC encoding technique obtained the worst performance.

The predictions on dataset I were poor (the highest average AUC value achieved was 0.583). It is obvious that intermediate binders were difficult to classify. Table III points out TVD achieved the best results. However, once again RC encoding obtained the worst performance as is dataset F.

Dataset S includes certain 9-mer peptides (i.e. strong binders and clear non-binders) and therefore, the best performance has been obtained when dataset S used. OETMAP has achieved the best result with the AUC value of 0.951.

TABLE II: PREDICTIVE AUC PERFORMANCE OF FEATURE ENCODING TECHNIQUES ON DATASET I

	Allele	OE	OE+FB	RC	TVD	OETMAP
1	A0201	0.655	0.63	0.563	0.618	0.663
2	A0202	0.5	0.504	0.464	0.554	0.512
3	A0203	0.59	0.618	0.494	0.661	0.603
4	A0206	0.653	0.622	0.643	0.616	0.662
5	A0301	0.56	0.555	0.528	0.617	0.59
6	A1101	0.587	0.596	0.519	0.604	0.603
7	A3101	0.567	0.533	0.501	0.578	0.574
8	A3301	0.534	0.495	0.517	0.562	0.559
9	A6801	0.513	0.459	0.497	0.549	0.551
10	A6802	0.516	0.527	0.493	0.585	0.53
11	B1501	0.559	0.53	0.533	0.541	0.57
Average AUC		**0.567**	**0.552**	**0.523**	**0.59**	**0.583**

TABLE III: PREDICTIVE PERFORMANCE OF FEATURE ENCODING SCHEMES FROM THE POINT OF AUC VALUES ON DATASET S

	Allele	OE	OE+FB	RC	TVD	OETMAP
1	A0101	0.976	0.886	0.8	0.926	0.968
2	A0201	0.98	0.95	0.922	0.975	0.979
3	A0202	0.984	0.966	0.908	0.979	0.982
4	A0203	0.971	0.943	0.868	0.972	0.973
5	A0206	0.973	0.953	0.911	0.98	0.972
6	A0301	0.942	0.865	0.834	0.953	0.942
7	A1101	0.975	0.856	0.809	0.968	0.97
8	A2402	0.797	0.762	0.735	0.796	0.786
9	A2601	0.936	0.891	0.593	0.958	0.962
10	A3101	0.963	0.866	0.855	0.963	0.969
11	A3301	0.905	0.802	0.744	0.902	0.932
12	A6801	0.953	0.827	0.741	0.939	0.954
13	A6802	0.958	0.937	0.879	0.936	0.955
14	B0702	0.964	0.905	0.803	0.924	0.967
Average AUC		**0.948**	**0.886**	**0.814**	**0.941**	**0.951**

IV. CONCLUSION

In this paper, we have studied the problem of whether given a nonamer peptide of any MHC allele is binding or non-binding by means of five feature encoding techniques with LSVM machine learning algorithm on three up-to-date MHC class I datasets. OETMAP technique, which is a conjunction of OE and TVD methods, in comparison with the other feature encoding techniques re-implemented on a standalone classifier approaches obtained higher AUC scores nearly for each allele in the challenge. Because independent and accurate classifiers make errors on different regions of the feature space, they can be ensemble. Hence, future works will involve the ensemble of classifiers with OETMAP encoding scheme.

ACKNOWLEDGMENT

This work was supported by Yalova University, BAP Project (Grant 2012 / 046).

REFERENCES

[1] J. D. Hayball and R. A. Lake, "The immune function of MHC class II molecules mutated in the putative superdimer interface," *Mol Cell Biochem*, vol. 273, no. 1-2, pp. 1-9, 2005.
[2] B. L. Buttgereit and R. Tampe, "The transporter associated with antigen processing: function and implications in human diseases," *Physiol Rev*, vol. 82, no. 1, pp. 187-204, 2002.
[3] L. F. Wang and M. Yu, "Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics," *Curr Drug Targets*, vol. 5, no. 1, pp.1-15, 2004.
[4] J. W. Yewdell, "Confronting complexity: real-world immunodominance in antiviral CD8+ T cell responses," *Immunity*, vol. 25, no. 4, pp. 533-543, 2006.
[5] H. G. Rammensee, T. Friede, and S. Stevanoviic, "MHC ligands and peptide motifs: first listing," *Immunogenetics*, vol. 41, no. 4, pp. 178-228, 1995.
[6] A. J. P. Jimenez and J. C. P. Cortes, "Genetic algorithms for linear feature extraction," *Pattern Recognition Letters*, vol. 27, no. 13, pp. 1508-1514, 2006.
[7] T. Rognvaldsson and L. You, "Why neural networks should not be used for HIV-1 protease cleavage site prediction," *Bioinformatics*, vol. 20, no. 11, pp. 1702-1709, 2004.
[8] M. J. Zvelebil, G. J. Barton, W. R. Taylor, and M. J. Sternberg, "Prediction of protein secondary structure and active sites using the alignment of homologous sequences," *J Mol Biol*, vol. 195, no. 4, pp. 957-961, 1987.
[9] W. R. Taylor, "The classification of amino acid conservation," *J Theor Biol*, vol. 119, no. 2, pp. 205-218, 1986.
[10] J. Guo and Y. L. Lin, "A Novel Method for Protein Subcellular Localization: Combining Residue-Couple Model and SVM," *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference*, pp. 117-129, 2005.
[11] M. Gök and A. T. Özcerit, "OETMAP: A New Feature Encoding Scheme for MHC Class I Binding Prediction," *Mol. and Cell. Biochemistry*, vol. 359, no. 1-2, pp. 67-72, 2012.
[12] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.
[13] K. Roomp, I. Antes, and T. Lengauer, "Predicting MHC class I epitopes in large datasets," *BMC Bioinformatics*, vol. 11, pp. 90, 2010.
[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd edn. Wiley, New York, 2000.
[15] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," Technical Report, HP Laboratories, Palo Alto, California, 2004.

Murat Gök, Ph.D. performed a Master in Computer Sciences at Mugla University (Turkey). After his Master thesis on the decision support systems, he began in 2006 a PhD in Computer Sciences at Sakarya University (Turkey). In June 2011, he defended his PhD thesis untitled "Prediction of HIV-1 Protese Cleavage Sites with New Techniques". Having completed his PhD, he became an assistant professor at the department of computer engineering on Yalova University (Turkey). His research interests are bioinformatics, machine learning algorithms and theories, computer programming. He has several papers on bioinformatics. He currently has several master students.