Mol Cell Biochem (2012) 359:67–72
DOI 10.1007/s11010-011-1000-5

OETMAP: a new feature encoding scheme for MHC class I binding prediction

Murat Gök · Ahmet Turan Özcerit

Received: 6 May 2011 / Accepted: 19 July 2011 / Published online: 30 July 2011
© Springer Science+Business Media, LLC. 2011

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 article, a novel feature amino acid encoding scheme is proposed to predict MHC/peptide complexes. In the proposed method, we have combined orthonormal encoding (OE) and Taylor's Venn-diagram, and have used Linear support vector machines as the classifier in the tests. We also have compared our method to current feature encoding scheme techniques. The tests have been carried out on comparatively large Human leukocyte antigen (HLA)-A and HLA-B allele peptide three binding datasets extracted from the Immune epitope database and analysis resource. On three datasets experimented, the IC50 cutoff a criteria is used to select the binders and non-binders peptides. Experimental results show that our amino acid encoding scheme leads to better classification performance than other amino acid encoding schemes on a standalone classifier.

M. Gök (✉)
Sakarya University Science Institute, Sakarya, Turkey
e-mail: muratgok@gmail.com

A. T. Özcerit
Computer Engineering, Sakarya University, Sakarya, Turkey

Keywords Epitope prediction · Major histocompatibility complex class I · Feature encoding scheme · Peptide classification

Abbreviations

MHC	Major histocompatibility complex
LSVM	Linear support vector machines
OE	Orthonormal encoding
TVD	Taylor's Venn-diagram
HLA	Human leukocyte antigen
IEDB	Immune epitope database and analysis resource
CTL	Cytotoxic T lymphocytes
RC	Residue-couple
(F, I, S)	Dataset (Full, intermediate, strong)
10-fold CV	10-fold cross validation
ROC	Receiver operating characteristic
AUC	Area under ROC curve

Introduction

MHC class-I and II antigens are of immense importance to the immune system. MHC molecules (also known as the HLA molecules in humans) undertake the key dialogs between T cells and other cells of the body. First, 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]. Second, MHC molecules present peptides to T Helper lymphocytes 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- 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 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.

There are a number of data stores describing the binding specificities for MHC molecules. The Immune epitope database and analysis resource (IEDB) [6] and SYFPEITHI database [7] are the main data repositories and services. Apart from IEDB and SYFPEITHI, MHCPEP [8], and MHCBN [9] are other data stores widely used for MHC alleles. But, IEDB has more entries than others and more up-to-date relatively.

The traditional feature encoding model is primarily based on the amino acid composition model. However, the amino acid composition model alone ignores a certain amount of information of the protein sequence. Unfortunately, the information about the sequence order effect cannot be easily incorporated into a pattern recognition model [10].

In this article, eight encoding schemes are evaluated to predict MHC/peptide binding. 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 $d_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{i20})$ where δ_{ij} 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 [11]. 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 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 [12]. Frequency based technique preserves the original number of attributes.

Amino acids of homologous sequences which are frequently substituted by each other over time are regarded as similar and the relationships are portrayed by substitution matrices, like the BLOSUM50 and the BLOSUM62 matrices [13]. In [14], authors described a new encoding scheme named BLOMAP which utilizes a non-linear projection method to recognize the similarity information in the BLOSUM62 matrix. The BLOMAP is an improved method of the Sammon-projection mapping.

Another encoding scheme technique is n-Grams or k-tuples [15], a pair of values (v_i, c_i), where v_i is the feature and c_i is the counts of this feature in a protein sequence. These features are all the possible combinations of n amino acids from 20 amino acids.

Zvelebil et al. [16] proposed a new encoding method based on Taylor's Venn-diagram (TVD) [17] 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 [18], 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.

In [19], four sequence-based approaches, DynaPred$^{\mathrm{POS}}$, NetMHC, SVMHC, and YKW have been experimented for predicting peptide binding to MHC class I molecules.

DynaPred$^{\mathrm{POS}}$ prediction method uses two feature matrices derived from structural calculations as basis for support vector machine training: a local, position-dependent (DynaPred$^{\mathrm{POS}}$) and a global, position-independent (DynaPred) matrix [20]. SVMHC is a SVM-based prediction technique whose kernels were optimized by systematic variation of the parameters [21]. YKW method is based on data-derived matrices [22]. Predictions of NetMHC[1] method based on artificial neural networks trained on data from 55 MHC alleles (43 Human and 12 non-human), and position-specific scoring matrices for additional 67 HLA alleles [23]. NetMHC is the state-of-the-art predictive model for MHC/peptide binding [19]. Their prediction performance was evaluated on three up-to-date datasets.

In this article, we have investigated a new feature encoding method that combines the sequence order of the residue composition based on OE and the representation of various relationships of residue based on TVD. This encoding scheme, termed OETMAP, has been applied to LSVM for MHC binding predictions. The computational

[1] http://www.cbs.dtu.dk/services/NetMHC.

results demonstrate higher performance of OETMAP technique in comparison with the feature encoding methods re-implemented on a standalone classifier approaches. Having compared the performance of eight encoding methods, we have conducted another comparison between OETMAP and four featured MHC class prediction methods namely DynaPred^{POS}, NetMHC, SVMHC, and YKW as implemented in [19].

Methods

Datasets

We conducted our tests on three up-to-date datasets[2] (Full (F), Intermediate (I), Strong (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 (50–500 nM binding affinity) and non-binders (500–1000 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 [18].

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 [24]. In the tests, LSVM algorithm was applied by OSU Toolbox [25].

Proposed encoding scheme

OE is a common method for the representation of sequences. It provides that all the vectors, obtained as binary from amino acids sequences, are mutually orthogonal and all of unit length [11]. But OE lacks of knowledge and sequence homology about proteins. We believe that amino acids sequences that are co-localized must share some similarity from the point of amino acid physicochemical properties. TVD shows the relationship of the 20 naturally occurring amino acids to a selection of physicochemical properties which are important in the determination of protein tertiary structure. TVD shown in Fig. 1 was based on the 2-D arrangement derived from Dayhoff's mutation matrix. Amino acids were then displaced from this arrangement to form groups of residues related by common physicochemical properties [17]. We believe that TVD extrapolates from physicochemical relationship of amino acids for prediction of MHC class I bindings. That is, similar physicochemical types of amino acids occur at a position for a binding peptide. But TVD is not enough itself for the amino acid representation. Consequently, OETMAP we developed consists of a conjunction of OE and TVD methods which are complementary to each other.

Let P be an amino acid sequence in MHC binding dataset. P_i be the ith amino acid in P (for $i = 1, 2, \ldots, L$ where L is the length of the amino acid sequence).

Corresponding to each P_i, we build feature vectors $\{\vec{y}\}_i^1$ and $\{\vec{y}\}_i^2$ as follows: $\{\vec{y}\}_i^1$ is the OE vector (20-bit) for P_i. The value of i represents the 20 different amino acids (briefly denoted as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V).

$\{\vec{y}\}_i^2$ is a binary feature vector composed of ten bits obtained from TVD, shown in Table 1.

The built feature vectors $\{\vec{y}\}_i^1$ and $\{\vec{y}\}_i^2$of each of P_i is concatenated:

$$\{\vec{y}\}_i = \left(\{\vec{y}\}_i^1 \parallel \{\vec{y}\}_i^2\right)$$

Finally the feature vector $\vec{\chi}$ of P, which has a dimension of $30 \times L$, is revealed in succession as follows:

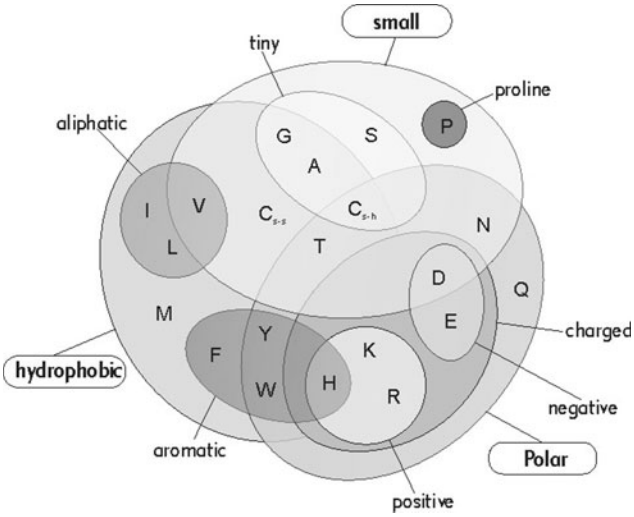

Fig. 1 TVD; amino acids are classified with respect to ten synthesis of physicochemical properties and mutation data (Adopted from Ref. [14])

[2] Publicly available on http://www.mpi-inf.mpg.de/~roomp/benchmarks/list.htm.

Table 1 Binary code vectors for $\{\vec{y}\}^2$

PC. Properties	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Hydrophobic	1	0	0	0	1	0	0	1	1	1	1	1	1	1	0	0	1	1	1	1
Positive	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Negative	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Polar	0	1	1	1	0	1	1	0	1	0	0	1	0	0	0	1	1	1	1	0
Charged	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
Small	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	1	0	0	1
Tiny	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Aliphatic	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
Aromatic	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
Proline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

$$\vec{\chi} = (\{\vec{y}\}_1 \parallel \{\vec{y}\}_2 \parallel \cdots \parallel \{\vec{y}\}_L)$$

To explain the new measure, we demonstrate an example of computation below for the sequence ALDFEQEMT in MHC binding dataset. For residue D in the peptide sequence, we have computed OE mapping as follows:

$$\{\vec{y}\}_3^1 = [0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0]$$

Next, we have computed $\{\vec{y}\}_3^2$:

$$\{\vec{y}\}_3^2 = [0\,0\,1\,1\,1\,1\,0\,0\,0\,0]$$

$\{\vec{y}\}_3$ is then computed as:

$$\{\vec{y}\}_3 = [0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0]$$

These computation procedures are repeated for each amino acid of the peptide sequence to obtain the feature vector $\vec{\chi}$.

Results

Performance of feature representation method

10-fold cross validation (10-fold CV) testing protocol is applied to evaluate the performance of the methods in terms of area under receiver operating characteristic (ROC) curve (AUC) averaged over ten experiments on datasets. In a cross-validation run, the 10-folds are randomly created [26]. 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 ten 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 ten turns are obtained. The performance of proposed feature encoding methods on datasets F, I, and S is shown in Tables 2 and 3, respectively, by means of AUC which is defined as the area under the ROC curve 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 [27]. OE, combining the OE representation with frequency based method (OE + Freq.), the substitution matrix BLOSUM50 (OE + B50), 2g, RC, BLOMAP, and TVD methods have been evaluated.

Table 2 reports that OETMAP outperforms the competing encoding techniques considered for dataset F with the value of 0.87. Note that n-grams obtained the worst performance. The predictions on dataset I were poor (the highest average AUC value achieved was 0.59). It is obvious that intermediate binders were difficult to classify. Table 2 points out TVD achieved the best results on dataset I. However, once again *n*-grams and RC methods 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 on dataset S. According to average values, the highest performance are achieved by OETMAP with the value of 0.801 compared other encoding schemes. We notice that OETMAP combines the both effectiveness of OE and TVD.

Table 3 reports that NetMHC is the best among the five predictive models particularly on dataset S. It may arise that some of the data used to train NetMHC is probably identical to that extracted from IEDB for this study as NetMHC was only accessible via a web interface [19]. YKW and OETMAP followed NetMHC where dataset I and dataset F were used, respectively. DynaPred$^{\mathrm{POS}}$, SVMHC, YKW, and OETMAP have been drawn in case of dataset S.

Table 2 Prediction performance of amino acid encoding schemes according to AUC values on datasets

	OE	OE + Freq.	OE + B50	n-grams	RC	BLOMAP	TVD	OETMAP
Dataset F	0.863	0.793	0.849	0.675	0.727	0.861	0.863	**0.87**
Dataset I	0.567	0.552	0.561	0.527	0.523	0.589	**0.59**	0.583
Dataset S	0.948	0.886	0.941	0.744	0.814	0.936	0.941	**0.951**
Average	0.793	0.744	0.784	0.649	0.688	0.795	0.798	**0.801**

Table 3 OETMAP is compared with other prediction models according to AUC performance on three datasets

	DynaPredPOS[a]	NetMHC[a]	SVMHC[a]	YKW[a]	OETMAP
Dataset F	0.86	0.94	0.86	0.86	0.87
Dataset I	0.56	0.71	0.52	0.59	0.58
Dataset S	0.95	0.99	0.95	0.95	0.95

[a] Results were obtained from [19]

Conclusion

In this article, we have studied the problem of whether given a nonamer peptide of any MHC allele is binding or non-binding by means of a new encoding scheme method. It is revealed from the experimental results that the new encoding scheme can accurately predict the MHC/peptide binding with high sensitivity on a standalone classification algorithm (LSVM) according to three up-to-date MHC class I datasets. Our proposed method can be used for other machine learning methods and can be used for any kind of peptide classification problems as well. 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.

Reproducibility material

We reported some MatLab code used for obtaining the empirical results in this article are available at: http://www.sakarya.edu.tr/aozcerit/codeMHC.rar.

Acknowledgments This study was supported by Sakarya University. BAP Project (Grant 2010-50-02-007).

References

1. Hayball JD, Lake RA (2005) The immune function of MHC class II molecules mutated in the putative superdimer interface. Mol Cell Biochem 273(1–2):1–9
2. Lankat-Buttgereit B, Tampe R (2002) The transporter associated with antigen processing: function and implications in human diseases. Physiol Rev 82(1):187–204. doi:10.1152/physrev.00025.2001
3. Wang LF, Yu M (2004) Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. Curr Drug Targets 5(1):1–15
4. Yewdell JW (2006) Confronting complexity: real-world immunodominance in antiviral CD8+ T cell responses. Immunity 25(4):533–543. doi:10.1016/j.immuni.2006.09.005
5. Rammensee HG, Friede T, Stevanoviic S (1995) MHC ligands and peptide motifs: first listing. Immunogenetics 41(4):178–228
6. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A (2005) The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol 3(3):e91. doi:10.1371/journal.pbio.0030091
7. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50(3–4):213–219
8. Brusic V, Rudy G, Harrison LC (1998) MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res 26(1):368–371
9. Lata S, Bhasin M, Raghava GP (2009) MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. BMC Res Notes 2:61. doi:10.1186/1756-0500-2-61
10. Chou KC (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278(2):477–483. doi:10.1006/bbrc.2000.3815S0006-291X(00)93815-4
11. Rognvaldsson T, You L (2004) Why neural networks should not be used for HIV-1 protease cleavage site prediction. Bioinformatics 20(11):1702–1709. doi:10.1093/bioinformatics/bth144
12. Orsenigo C, Vercellis C (2007) Predicting HIV protease-cleavable peptides by discrete support vector machines, evolutionary computation, machine learning and data mining in bioinformatics. In: EvoBIO'07 Proceedings of the 5th european conference on evolutionary computation, machine learning and data mining in bioinformatics. Springer-Verlag, Berlin, pp 197–206
13. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919
14. Maetschke S, Towsey M, Boden M (2005) Blomap: an encoding of amino acids which improves signal peptide cleavage prediction. In: Proceedings of the 3rd Asia-Pacific bioinformatics conference, London, pp 141–150

15. Wu C, Whitson G, McLarty J, Ermongkonchai A, Chang TC (1992) Protein classification artificial neural system. Protein Sci 1(5):667–677. doi:10.1002/pro.5560010512
16. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 195(4):957–961
17. Taylor WR (1986) The classification of amino acid conservation. J Theor Biol 119(2):205–218
18. Guo J, Lin YL (2005) A novel method for protein subcellular localization: combining residue-couple model and SVM. In: Proceedings of the 3rd Asia-Pacific bioinformatics conference, Singapore, pp 117–129
19. Roomp K, Antes I, Lengauer T (2010) Predicting MHC class I epitopes in large datasets. BMC Bioinformatics 11:90. doi:10.1186/1471-2105-11-90
20. Antes I, Siu SW, Lengauer T (2006) DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. Bioinformatics 22(14): e16–e24. doi:10.1093/bioinformatics/btl216
21. Donnes P, Elofsson A (2002) Prediction of MHC class I binding peptides, using SVMHC. BMC Bioinformatics 3:25
22. Yu K, Petrovsky N, Schonbach C, Koh JY, Brusic V (2002) Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol Med 8(3):137–148
23. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M (2008) NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. Nucleic Acids Res 36(Web Server issue):W509–W512. doi:10.1093/nar/gkn202
24. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167
25. Junshui Ma, Yi Zhao (2002) OSU SVM Toolbox for MATLAB. http://sourceforge.net/projects/svm/. Accessed 10 May 2011
26. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York
27. Fawcett T (2004) ROC graphs: notes and practical considerations for researchers. Technical Report, HP Laboratories, Palo Alto