From the SelectedWorks of Ping Zhang

November 28, 2008

A hybrid model for prediction of peptide binding to MHC molecules

P. Zhang
V. Brusic
K. Basford

Available at: https://works.bepress.com/ping_zhang/11/

A Hybrid Model for Prediction of Peptide Binding to MHC Molecules

Ping Zhang[1], Vladimir Brusic[1,2], Kaye Basford[1]

[1]The University of Queensland, School of Land, Crop and Food Sciences, QLD 4072 ,
Australia
[2]Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston MA, USA
Emails: p.zhang2@uq.edu.au; vladimir_brusic@dfci.harvard.edu; k.e.basford@uq.edu.au

Abstract. We propose a hybrid classification system for predicting peptide binding to major histocompatibility complex (MHC) molecules. This system combines Support Vector Machine (SVM) and Stabilized Matrix Method (SMM). Its performance was assessed using ROC analysis, and compared with the individual component methods using statistical tests. The preliminary test on four HLA alleles provided encouraging evidence for the hybrid model. The datasets used for the experiments are publicly accessible and have been benchmarked by other researchers.

1 Introduction

The concept of peptide vaccines uses the basic principle that T lymphocytes recognize antigens as peptide fragments that are generated from degradation of protein antigens. Peptides are subsequently bound to MHC class I or II molecules and displayed on the surface of the antigen presenting cells. Identification of epitopes and peptides that can bind MHC molecules is important for understanding the basis of cellular immune responses and the design of vaccines. Peptides that are presented by MHC molecules and recognized by T cells are termed T-cell epitopes. Human MHC is known as human leukocyte antigen (HLA).

Prediction of MHC binding peptides and T-cell epitopes has become one of the popular areas of bioinformatics applications in immunology. Statistical modeling and machine learning techniques have been used to build the prediction models based on the available databases. Recent analysis by Lin et al [1] showed that Matrix-based Models, Artificial Neural Networks (ANN) and Support Vector Machines (SVM) produce high accuracy in predicting the binding or nonbinding peptides to some MHC molecules. Because of the large number of different HLA molecules, where more than 2000 variants have been identified in humans, and variable performance of prediction methods for different HLA molecules, it is not possible to identify the best model for this type of prediction. Hybrid models were proposed, aimed at improvement of prediction performance. Moutaftsi et al [2] performed prediction of HLA binding peptides using the combination of prediction results from multiple matrices including Udaka [3], Parker [4], ARB [5,6] and SMM [7]. Bhasin and

Raghava [8] proposed a hybrid prediction model which combined a quantitative matrix based approach (QM) and a neural network (NN) approach. This method was implemented as nHlaPred server (http://www.imtech.res.in/raghava/nhlapred) and reported that the NN and QM can complement each other, leading to reduction in false prediction.

The performance of prediction models can be evaluated by several measures, including classification rate, sensitivity (SE) and specificity (SP). A common measure for assessing prediction accuracy is using receiver operating characteristic (ROC) curve. ROC curve plots SE vs 1-SP for a full range of decision thresholds. The area under the ROC plot (AUC) gives the overall evaluation of the model. By convention the AUC values range between 1.0 (perfect skill) and 0.5 (zero skill). The bigger the value is, the better overall performance of the model is represented. More details about ROC curve and its applications can be found in [9].

In this paper, we studied a hybrid method which combines the SVM and stabilized matrix method (SMM). The performance of the models was evaluated using ROC curve and compared with the single models.

2 SVM and Matrix Techniques for Prediction of MHC Binding Peptides

2.1 Support Vector Machine (SVM)

SVM is a kernel based machine learning technique that originated in modern statistical learning theory [10]. SVM can make data sets linearly separable by transferring them into a feature space of higher dimension. SVMs have excellent generalizing capability [11]. They use kernel functions to map the data; the most popular kernels use linear, polynomial, and Gaussian function. SVM can be applied for both classification and regression tasks. Some reports state that SVMs outperformed most other systems in a wide variety of applications [12]. SVMs have been used to predict MHC-binding peptides and they were reported as highly accurate. Dönnes and Elofsson [13] developed a model SVMHC based on support vector machines to predict the binding of peptides to MHC class I molecules. They compared its performance with two profile based methods, SYFPEITHI [14] and HLA_BIND [4] and claimed a slightly better result. Bhasin and Raghava [15] applied SVM for prediction of peptides binding with the MHC class II allele HLA-DRB1*0401. Their results showed good performance of SVM compared with other classification methods including Matrices, Motifs and ANN.

Zhang et al [16, 17] have also shown the advantages of SVM for predicting the MHC binding peptides. They reported a system based on SVM that outperformed ANN and Hidden Markov Model (HMM) [17], and then improved the model by adding the data representation of peptide/MHC interaction [16]. SVM regression models (SVR) were also used for prediction of peptide binding affinity to MHC molecules [18] and for binding/nonbinding peptide classification based on certain output cutoff values [19].

These models mainly used the binary coding of peptide sequences as the input variables. Zhao et al [20] encoded each amino acid in the peptide sequences by ten factors which were obtained from 188 physical properties of 20 AAs via multivariate statistical analyses by Scheraga's group [21] for their SVM classification model. The SVM model was used for T-cell epitopes prediction and was also compared with ANN and Z-Matrix based approaches [22]. Cui et al [23] also reported high performance of SVM for a variety of HLA alleles including 18 class I and 12 class II, and prediction of newly reported epitopes with high accuracy (11 out of 15).

In general, SVM has been reported by many researchers to perform well for prediction of MHC binding peptides and epitopes regardless of the kernels which were used with the models. This demonstrated the ability of SVMs to build effective predictive models when the dimensionality of the data is high and the number of observations is limited [11].

On the other hand, it has also been claimed that the performance of SVM strongly depends on both the quality and quantity of data. It fails if an unbalanced data set is used for learning. Riedesel et al [24] used the least square optimization method (LSM) with a weighting procedure to deal with asymmetric data sets with a small number of binding and a large number of non-binding peptides. This approach was expected to yield higher prediction accuracy than SVM. However, in their paper the assessment for the model performance using ROC curve showed better performance of SVM than LSM.

There are three important algorithms for training SVM: Chunking, Sequential Minimum Optimization, and SVM^{light} [25]. SVM^{light} is an implementation of an SVM learner which addresses the problem of large tasks. It decomposes the problem with many training examples into a series of smaller tasks. The main advantage of this decomposition is that it suggests algorithms with memory requirements linear in the number of training examples and in the number of support vectors [26].

2.2 Stabilized Matrix Method (SMM)

The Matrix methods assign scores for 20 amino acids (AAs) to all positions of MHC grooves. These scores are calculated based on the frequencies of the AAs appearing at the positions within a large number of binding experiments [27]. Brief reviews of methods for building matrix models can be found in [11] and [27]. Studies have shown that simple predictions using scoring matrices yield reasonably good results when little experimental data are available (typically tens to hundreds of peptides) compared with the size of the sequence space (20^9 for 9-mers). This indicates that the relationship between sequence and affinity can be approximated by the independent binding assumption, i.e. amino acids at different positions of a peptide contribute independently to the overall binding affinity of the peptide [5].

Peters et al [5] developed a matrix-based algorithm called the Stabilized Matrix Method (SMM). First they create a matrix by using a frequency-based formula then modify the matrix to compensate for the errors contained in experimental data. This method was first tested on prediction of peptide binding HLA-A*0201 molecule using a set of 9-mer peptides. The performance of SMM was compared with three widely used matrix-based methods: BIMAS [4], SYFPEITHI [14] and the polynomial method (PM) [28]. SMM considerably outperforms the other methods on three independent test sets. This method was then les, peptide transport by the transporter associated with antigen presentation (TAP) and proteasomal cleavage of protein sequences [6]. The implementation of the algorithm successfully applied to predicting peptide binding to MHC molecufor creating the matrix has been made publicly available (http://www.mhc-pathway.net/smm).

3 Hybrid Model

3.1 Methodology

We propose a hybrid prediction model that combines SVM and SMM methods, as depicted in Figure 1. A matrix is built based on the SMM approach with the protein peptide sequences as the input. The output values of the SMM model are combined with the binary code of the sequences for the final classification of the peptides as the binding or nonbinding to the MHC-I molecules. The hybrid mode can be described in the following steps:

1) Use the training data to build the SMM matrix model and test on the test set. Save the output from both the training and test data.
2) Normalize the output from the matrix model. The normalization is performed using every output value for every peptide divided by the maximum output value of the training set.
3) Create the input vector for every peptide with the binary coding of the original sequence plus the output from the SMM matrix model. Each amino acid in the peptide can be encoded as a binary string of length 20 with one position set to “1” and the other positions set to “0”. The binary code of a nonamer peptide will be represented by the binary string of length 180. With the output from the matrix model added, the input number to the SVM will be 181.
4) Build the SVM model using the training data for final classification.
5) Test the model using test data.

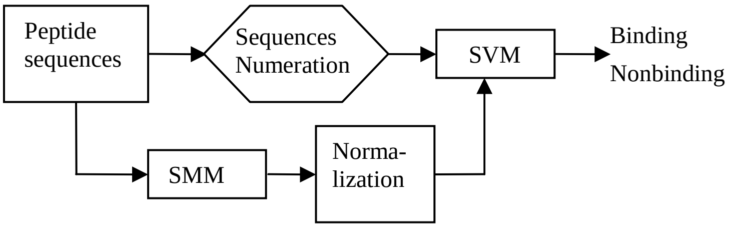

Fig. 1: The hybrid model with SMM matrix and SVM

3.2 Datasets

Peters et al [29] presented a dataset that was used for benchmarking the MHC/peptide binding prediction models. The entire dataset can be downloaded from the publicly accessible website mhcbindingpredictions.immuneepitope.org (IEDB). They used this data set to compare performances of three prediction methods (ARB, SMM and ANN) developed in their lab. They also provided training and testing data sets used in their study. In this research the same training and testing 9-mer peptide datasets of four HLA class I alleles including HLA-A*0101, HLA-A*0201 HLA-A*0202 and HLA-A*0301 have been used in our experiments. The numbers of peptides included in the datasets are 1157, 3089, 1447 and 2094 respectively.

3.3 Implementation and Experimental Results

In this research, we used SVM^{light} (http://www-ai.cs.uni-dortmund.de/svm_light) for all the SVM implementation. Additional C code to adjust the parameters was added to the original code for simulations. The SMM C code for UNIX was downloaded from the source and used to create the matrices using benchmark datasets.

For our experiments, the dataset of each allele were split into five equally sized subsets. In each partition, one subset was used for testing, while other subsets were combined and used for training. As a result, five prediction models were built for each allele.

To evaluate the performance of the models, the ROC curves were drawn for every test set (5 for each HLA variant) independently. To compare the hybrid model performance with the single SMM and SVM models, the experiments using SMM and SVM were run with the same datasets and the AUC value were calculated. Figure2 shows the AUC values for the different models for the 5 datasets for each allele. The horizontal axis represents the dataset used for each allele, and the vertical axis is the AUC value. From the figure, we can see a small improvement in the performance of the hybrid model. However, the AUC values of the hybrid model are all higher than or equal to the values from the SVM, although the differences are not large. There was only one value (out of 20) from SMM higher than the corresponding hybrid value.

Table 1 lists the average AUC values of the ROC curves for the three models for different HLA alleles. To determine whether the hybrid model performs better than the single methods, two-way analysis of variance (ANOVA) was performed for each allele. The p-values from the analysis are also shown in this table. The Least Significant Difference (LSD) was calculated to compare the model mean values. The results are listed in the same table. The p values show that for 3 variants, the hybrid models have shown statistically significant improvement, while there was no difference for HLA-A*0101. The LSDs show the significant difference of the SMM from other two models, but do not show a significant difference between SVM and the hybrid mode.

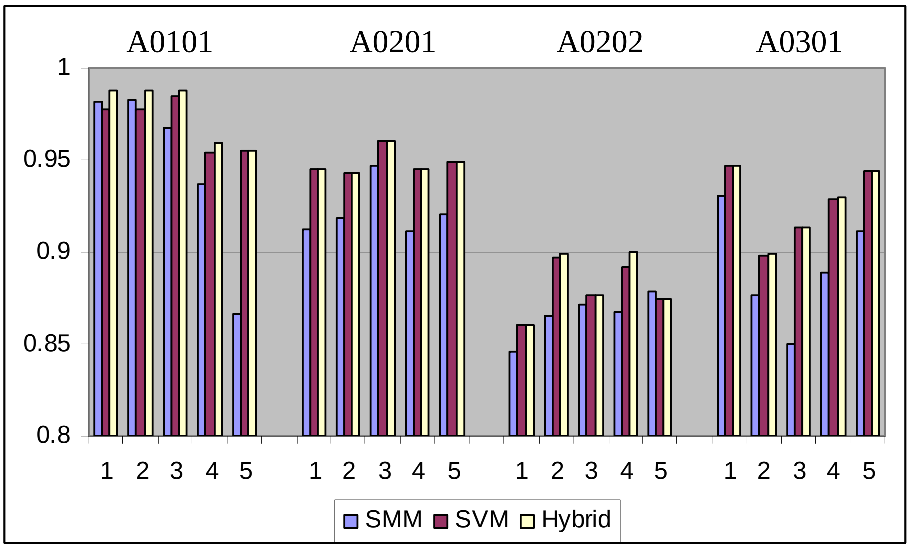

Fig. 2: AUC values of different models for each allele

Table 1: The average AUC values of each allele from the three models and ANOVA test results

Model	Allele			
	A0101	A0201	A0202	A0301
SMM	0.947 a	0.922 b	0.866 b	0.892 b
SVM	0.970 a	0.948 a	0.880 a	0.926 a
Hybrid	0.974 a	0.948 a	0.882 a	0.927 a
P-value (within allele)	0.1734	0.00003	0.0425	0.0011

Note: Within an allele, means followed by the same letter are not significantly different from one another at the 5% significance level

4 Discussion and Conclusion

This research proposed the idea of combining the Matrix approach and the SVM. It used the SVM with the traditional kernels and the most used encoding for the input peptide sequences. The preliminary test of the hybrid model, based on a limited number of alleles, showed encouraging results. The ANOVA test on the model means did not show significant difference between the SVM and the hybrid model, however, the hybrid model did not perform worse than the SVM in any single test. This indicates that there is potential for further development of the hybrid model. In future research, new kernel functions such as kernels combining biological features [30] and more efficient encoding or input vectors can be tested with the hybrid system. Different matrix approaches can also be applied in the hybrid system. As the performance of SVM is dependent on the training data, the optimization of training data can also be considered in future work.

Acknowledgements:

The authors acknowledge the support of the ImmunoGrid project, EC contract FP6-2004-IST-4, No. 028069.

References

1. Lin, HH., Ray, S., Tongchusak, S., Reinherz, EL., Brusic, V.: Evaluation of HLA Class I Peptide Binding Prediction Servers: Applications for Vaccine Research. BMC Immunol. 9:8, doi: 10.1186/1471-2172-9-8 (2008).
2. Moutafts, M., Peters, B., Pasquetto, V., Tscharke, D.C., Sidney, J., Bui HH., Grey, H. and Sette, A.: A Consensus Epitope Prediction Approach Identifies the Breadth of Murine T_{CD8+} - Cell Responses to Vaccinia Virus. Nature Biotechnology, 24(7), 817-819 (2006).
3. Udaka, K., Wiesmuller, K.H., Kienle, S., Jung, G., Tamamura, H., et al.: An Automated Prediction of MHC Class I - Binding Peptides Based on Positional Scanning with Peptide Libraries. Immunogenetics. 51, 816–828 (2000).
4. Parker, K.C., Bednarek, M.A., Coligan, J.E.: Scheme for Ranking Potential HLA-A2 Binding Peptides Based on Independent Binding of Individual Peptide Side-Chains. J Immunol. 152, 163–175 (1994).
5. Peters, B., Tong. W., Sidney, J., Sette, A. and Weng, Z.: Examining the Independent Binding Assumption for Binding of Peptide Epitopes to MHC-I Molecules. Bioinformatics 19, 1765 - 1772 (2003).
6. Peters, B. and Sette, A.: Generating Quantitative Models Describing the Sequence Specificity of Biological Processes with the Stabilized Matrix Method. BMC Bioinformatics. 6:132 (2005).
7. Bui, H.H., Sidney, J., Peters, B., Sathiamurthy, M., Sinichi, A., Purton, K.A., Mothe, B.R., Chisari, F.V., Watkins, D.I., and Sette, A..: Automated Generation and Evaluation of Specific MHC Binding Predictive Tools: ARB Matrix Applications. Immunogenetics. 57, 304–314 (2005).
8. Bhasin, M. and Raghava, G.P.S.: A Hybrid Approach for Predicting Promiscuous MHC Class I Restricted T Cell Epitopes. J. Biosci. 32:31-42 (2006)
9. Zweig, M.H. and Campbell, G.: Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Clinical Chemistry. 39(4), 561-577 (1993).
10. Vapnik, V.: Statistical Learning Theory. New York: Wiley (1998).
11. Tsurui, H., and Takahashi, T.: Prediction of T-Cell Epitope. Journal of Pharmacological Sciences, 105, 299-316 (2007).
12. Cristianini, N. and Shawe-Taylor, J.: An Introduction to Support Vector Machines and other Kernel-based Learning Methods (2000).
13. Dönnes, P. and Elofsson, A.: Prediction of MHC Class I Binding Peptides, Using SVMHC, BMC Bioinformatics. 3:25, doi:10.1186/1471-2105-3-25 (2002).

14. Rammensee, H., Bachmann, J., Emmerich, N.N., Bachor, O.A., and Stevanovic, S.: SYFPEITHI: Database for MHC Ligands and Peptide Motifs. Immunogenetics. 50: 213-219 (1999).
15. Bhasin, M., Raghava G.P.: SVM Based Method for Prediction HLA-DRB1*401 Binding Peptides in an Antigen Sequence. Bioinformatics. 20:421-423 (2004).
16. Zhang, G.L., Bozic, I., Kwoh, C.K., August, J.T. and Brusic, V.: Prediction of Supertype-specific HLA Class I Binding Peptides Using Support Vector Machines. Journal of Immunological Methods. 320(1-2), 143-154 (2007).
17. Bozic, I., Zhang, G.L. and Brusic, V.: Predictive Vaccinology: Optimisation of Predictions Using Support Vector Machine Classifiers. Lecture Notes in Computer Science. 3578, 375-381 (2005).
18. Lui, W., Meng, X., Xu, Q., Flower, D.R. and Li, T. Quantitative Prediction of Mouse Class I MHC Peptide Binding Affinity Using Support Vector Machine Regression (SVR) Models. BMC Bioinformatics. 7:182 (2006).
19. You, L., Zhang, P., Bodén, M. and Brusic, V.: Understanding Prediction Systems for HLA-Binding Peptides and T-cell Epitope Identification, Lecture Note in Bioinformatics 4774, 337-348 (2007).
20. Zhao, Y., Pinilla, C., Valmori, D., Martin, R., Simon, R.: Application of Support Vector Machines for T-Cell Epitopes Prediction. Bioinformatics. 19, 1978-1984 (2003).
21. 21 Kidera, A., Konishi, Y., Oka, M., Ooi, T. and Scheraga, H.A.: Statistical Analysis of the Physical Properties of the 20 Naturally Occuring Amino Acids. J. Protein Chem., 4, 23–55 (1985).
22. Zhao, Y., Gran, B., Pinilla, C., Markovic-Plese, S., Hemmer, B., Tzou, A., Whitney, L.W., Biddison, W.E., Martin, R. and Simon, R. Combinatorial Peptide Libraries and Biometric Score Matrices Permit the Quantitative Analysis of Specific and Degenerate Interactions Between Clonotypic T-Cell Receptors and MHC–Peptide Ligands. J. Immunol. 167, 2130–3141 (2001).
23. Cui, J., Han, L.Y, Lin, H.H., Zhang, H.L. Tang, Z.Q., Zheng, C.J., Cao, Z.W. and Chen, Y.Z.: Prediction of MHC-binding Peptides of Flexible Lengths from Sequence-derived Structural and Physicochemical Properties. Mol Immunol. 44, 866-877 (2007).
24. Riedesel, H., Kolbeck, B., Schmetzer, O. and Knapp, E.W.: Peptide Binding at Class I Major Histocompatibility Complex Scored with Linear Functions and Support Vector Machines. Genome Informatics. 15(1), 198-212 (2004).
25. Dong, J. and Suen, C. Y.: A Fast SVM Training Algorithm. International Journal of Pattern Recognition and Artificial Intelligence. 17(3), 367-384 (2003).
26. Joachims, T. (Ed.): Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge, USA (1999).
27. Yu, K., Petrovsky, N., Schonbach, C., Koh, J.Y.L. and Brusic, V.: Methods for Prediction of Peptide Binding to MHC Molecules: A Comparative Study. Mol. Med., 8, 137-148 (2002).
28. Gulukota, K., Sidney, J., Sette, A. and DeLisi, C.: Two Complementary Methods for Predicting Peptides Binding Major Histocompatibility Complex Molecules. J. Mol. Biol. 267, 1258–1267 (1997).
29. Peters, B., Bui, H.H., Frankild, S., Nielsen, M., Lundegaard, C., et al.: A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules. Plos Computational Biology. 2(6), 574–584 (2006).
30. Yang, Z.R. and Johnson, F.C.: Prediction of T-cell epitopes Using Biosupport Vector Machines. J Chem Inf Model. 45(5), 1424-1428 (2005).