A SUPPORT VECTOR MACHINE APPROACH FOR PREDICTION OF T CELL EPITOPES*

LEI HUANG AND YANG DAI†

Department of Bioengineering (MC063)
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL 60607, USA
{*lhuang7, yangdai*}*@uic.edu*

Abstract A new peptide encoding scheme is proposed to use with support vector machines for the direct recognition of T cell epitopes. The method enables the presentation of information on both (1) amino acid positions in peptides and (2) the similarity between amino acids through the use of sparse indicator vectors and the BLOSUM50 matrix. A procedure of feature selection is also introduced. The computational results demonstrate superior performance over previous techniques.

Keywords: T cell epitope recognition, Support vector machine, Feature selection

1. Introduction

In Silico T cell epitope identification currently relies on the prediction of peptide binding to major histocompatibility complex (MHC) molecules. Antigens are degraded into a set of peptide fragments through the action of the proteasome and the resulting peptides presented by MHCs are recognized by one or few of a large set of T cell receptors (TCRs). Methods for the prediction of MHC binding peptides have been developed based on structural binding motifs[9,10,20,21,23,24] or quantitative matrices,[15,18] Artificial Neural Networks (ANN),[2,8,12,22] and Support Vector Machines (SVM).[5] These techniques, however, do not discriminate between T cell epitopes and non-epitopes which are both MHC binders.[7] The methods of direct prediction developed in 1980s were based on the structural analysis of T cell epitopes.[4,25,26]

Recently, methods of direct prediction of T cell epitopes based on machine learning techniques such as SVM and ANN with the use of sequence information have been proposed.[1,27] In Zhao *et al.*,[27] each amino acid in a peptide was encoded by 10 physical properties of the 20 amino acids. These 10 properties include alpha-helix or bend-structure

*This research is partially supported by National Science Foundation (EIA-022-0301) and Naval Research Laboratory (N00173-03-1-G016).
†Corresponding author.

preference, bulk, beta-structure preference, hydrophobicity, normalized frequency of double bend, normalized frequency of alpha region, and pK-C. [14] These properties represent a better class of information in comparison with the amino acid indicator variables. Bhasin and Raghava[1] encoded each peptide using only the amino acid indicator vector. That is, each amino acid of a peptide is represented by a 20-dimensional vector. They tested this encoding method by SVM and ANN. In both of the studies, SVM was demonstrated as a potential machine learning method to make relatively accurate predictions based on training with small data sets.

In this work, a new encoding scheme of peptides for the direct prediction of T cell epitopes is investigated. This encoding method combines the BLOSUM50 matrix [11] with the amino acid indicator vector. This is achieved by replacing each nonzero entry in the vector by the corresponding value appeared in diagonal entries in the BLOSUM matrix. This is different from other encoding methods using the BLOSUM matrix in which each amino acid is simply represented by its BLOSUM score.[17] The new encoding method simultaneously incorporates information on both the position and similarity of the amino acids.

The new method was evaluated on a data set employed by Zhao *et al.*[27] with the use of SVM. The computational results demonstrate the superior performance of this method in comparison with the methods using the indicator vector or the 10 physical properties of the amino acids.[27]

2. System and Method

2.1. *Training and testing data*

The data set used in Zhao *et al.*[27] was utilized in this experiment. T cell clone and antigen recognition assay Melan-A-specific CTL clone LAU203-1.5 were derived from the tumor-infiltrated lymph node cells of a melanoma patient and antigen recognition was assessed using a chromium-release assay (see details in Zhao *et al.*[27]) Among the 203 synthetic peptides, 36 were tested stimulatory (positive) and 167 were tested non-stimulatory (negative). Since the numbers of the positive and negative peptides are unbalanced, in our cross-validation, the numbers of positive and negative peptides in the training and testing sets were maintained in a similar ratio. These peptides consist of 10 amino acids.

2.2. *Peptide encoding method*

One of important elements that influence the effectiveness of a SVM model is the design of the encoding method for the training data. In this study, we introduce an encoding technique that combines the amino acid substitution matrix BLOSUM50 [11] together with the conventional 20-dimensional indicator vector (1 present or 0 absent of an amino acid) at each position. The representation of each amino acid by an indicator vector has been used extensively. It provides very precise information about the peptide, such as the position of each amino acid in the peptide. The use of BLOSUM matrix in this study, however, is different from the way it has been used before.[17,27] Usually, given a peptide, each amino

acid is simply represented by its BLOSUM score. In this case, the encoding vector is of dimension s for a peptide with length s. The BLOSUM score contains prior knowledge about which amino acids are similar or dissimilar to each other in distantly related proteins. However, it is clear that this encoding method loses some information about each amino acid. For example, the hydrophilic amino acids Arg, Asn, Gln and the hydrophobic amino acid Met all have the same BLOSUM score 7. If they appear at the same position in the peptides, this encoding method would not be able to discriminate. Our new encoding method avoids this ambiguity by replacing each non-zero value in the indicator vector by the BLOSUM score of the corresponding amino acid. For a peptide with length 10, the dimension of the encoding vector is thus 200. This scheme encodes not only the position of residues but also the similarity scores. Therefore, the entire vector provides more accurate information about a peptide.

2.3. *Feature ranking*

Since the feature vectors are relatively sparse, a feature selection procedure is used to exclude the features that appear less frequently and less discriminative. A simple procedure for feature selection is carried out based on the Fisher's score of each feature.

Let n_1 and n_2 be the numbers of vectors in the positive (Pos) and negative (Neg) training sets, respectively. Denote an encoding vector of peptide by $\boldsymbol{x}$. The Fisher's score for each feature j is defined as

$$F_j = \frac{|\mu_j^{Pos} - \mu_j^{Neg}|}{\sqrt{(s_j^{Pos})^2 + (s_j^{Neg})^2}},$$

where

$$\mu_j^{Pos} = \frac{1}{n_1} \sum_{\boldsymbol{x}_i \in Pos} x_{ij}, \quad \mu_j^{Neg} = \frac{1}{n_2} \sum_{\boldsymbol{x}_i \in Neg} x_{ij},$$

$$(s_j^{Pos})^2 = \frac{\sum_{\boldsymbol{x}_i \in Pos} (x_{ij} - \mu_j^{Pos})^2}{n_1}, \quad (s_j^{Neg})^2 = \frac{\sum_{\boldsymbol{x}_i \in Neg} (x_{ij} - \mu_j^{Neg})^2}{n_2}.$$

A feature with a higher Fisher's score is considered as more discriminative. The features are then assembled according to their scores in a descending order.

2.4. *Training with support vector machine*

Suppose that we are given a set of m points $\boldsymbol{x}_i$ $(1 \leq i \leq m)$ in an n-dimensional space. Each point $\boldsymbol{x}_i$ is labeled by $y_i \in \{1, -1\}$ denoting the membership of the point. An SVM is a learning method for binary classification. Using a nonlinear transformation ϕ, it maps the data to a high dimensional feature space in which a linear classification is performed.

It is equivalent to solving the quadratic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi_1,\ldots,\xi_m} \quad & \frac{1}{2}\boldsymbol{w}\cdot\boldsymbol{w} + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(\phi(\boldsymbol{x}_i)\cdot\boldsymbol{w}+b) \geq 1-\xi_i \quad (i=1,...,m), \\ & \xi_i \geq 0 \quad (i=1,...,m), \end{aligned}$$

where C is a parameter. The decision function is defined as $f(\boldsymbol{x}) = \phi(\boldsymbol{x})\cdot\boldsymbol{w}+b$, where $\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i\phi(\boldsymbol{x}_i)$ and α_i $(i=1,...,m)$ are nonnegative constants determined by the dual problem of the optimization defined above. Therefore, the function can be represented as

$$f(\boldsymbol{x}) = \sum_{i=1}^{m}\alpha_i\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x})+b = \sum_{i=1}^{m}\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x})+b$$

through the definition of a kernel function $K(\cdot,\cdot)$. For details of SVMs refer to Cristianini and Shawe-Taylor.[3]

According to our preliminary study, the linear kernel ($\phi(\boldsymbol{x}) = \boldsymbol{x}$) in SVM gives the best performance. This could be due to the sparsity of encoding vectors and the small number of training points. Therefore, we used the linear SVM model in the present experiment.

In order to handle the unbalancedness between the numbers of peptides in the positive and negative training sets, different parameters C_+ and C_- were associated with the positive and negative training errors respectively. That is, the objective function in the above quadratic programming is replaced by

$$\frac{1}{2}\boldsymbol{w}\cdot\boldsymbol{w} + C_+\sum_{i:y_i=1}\xi_i + C_-\sum_{i:y_i=-1}\xi_i.$$

The ratio of C_+ to C_- is bounded by the value of n_2/n_1 in general,[16] however, the best ratio is usually determined through cross-validation by searching in the range of $[1, n_2/n_1]$. Accordingly, there are two parameters associated with a SVM model:

(1) C_- : the trade-off between the negative training error and class separation;
(2) J : the ratio of C_+ to C_-.

From the above discussion, $C_+ = C_- * J$.

Since the identification of positive peptides is of great interest, the quality of the SVMs was evaluated by the precision (positive prediction value)

$$precision = \frac{tp}{tp+fp},$$

and recall (sensitivity):

$$recall = \frac{tp}{tp+fn},$$

where tp (resp. tn) is the number of predicted positive (resp. negative) peptides which are true positive (resp. negative), and fp (resp. fn) is the number of predicted positive (negative) peptide which are true negative (resp. positive).

The F-score, given by

$$F_{score} = \frac{2 * precision * recall}{precision + recall},$$

was employed as a criterion for the determination of the SVM parameters in the cross-validation. These criteria will also provide a more accurate evaluation of the classifier when dealing with unbalanced positive and negative data sets.

An experimental protocol similar to the one used in Zhao *et al.* [27] was employed in our study in order to conduct a valid comparison. A 20% fraction of the data was set aside as a testing set for positive and negative sets, respectively. The remaining 80% fraction of the data was used as the training set. The number of features and the parameters associated with the SVM were optimized through a 10-fold cross-validation on the training set. The final classifier was obtained by training a SVM on the whole training set again with the selected features and optimized parameters. The classifier was then evaluated on the reserved 20% testing set. This entire process was repeated 10 times with different random splittings of the training and testing data sets; the final results were averaged over 10 runs. Note that a leave-one-out cross-validation was used to optimize the classifier in Zhao *et al.* [27]

The details of the 10-fold cross-validation on the 80% of training data are as follows. For each training set, the order of the features was rearranged by the Fisher's scores. Selecting a small set of features from the top of the sorted feature list, the averaged F-score was calculated through the 10-fold cross-validation for a pair of parameters $(C\_, J)$. Searching through all possible values of the parameters in a given range will identify the best F-score and the corresponding pair of parameters for the fixed set of features.

In our experiment, this process was repeated from 60 to 200 features from the sorted list, each time with an incremental step of 20 features. The optimal number of features and the best parameter pair associated with the best F-score were identified. A summary of the procedure is shown as follows.

Procedure

(1) Prepare the sorted list of features according to their Fisher's scores for the given training data set.
(2) Choose the initial set of features $V(|V| = 60)$ from the top of the sorted list.
(3) Identify the best pair of parameters $(C\_, J)$ through a 10-fold cross-validation and store the corresponding F_{score}.
(4) If there is no feature left for inclusion, then go to (5). Otherwise, add the next 20 features from the sorted list and go to (3).
(5) Identify the best F_{score} and its associated feature set and parameter pair.
(6) Use the optimized feature set and parameter pair to train the full set of training data.
(7) Calculate the recall and precision of the testing set using the classifier obtained in the previous step (6).

The basis of the selection of the first 60 features from the sorted list to serve as a start is as follows. Since the encoded sequences are relatively sparse, we have to include a

sufficient number of features to guarantee that no peptide will be associated with an empty vector of features.

The parameter searches were conducted as follows:

(1) C_- : $[0.005, 0.2]$ with a step size of 0.005.
(2) J : $[0.5, 4.0]$ with a step size of 0.1.

3. Results and Discussion

The SVMLight package[13] was used in the implementation. We report the recall and the precision of testing data sets (20% of the data) for 10 runs. The results of the SVM trained with the original 200 features are shown in Table 1. The results of the SVM in conjunction with the feature selection are shown in Table 2. For comparison, the method of SVM with peptides encoded with the amino acid indicator vectors was also implemented. This result is presented in Table 3. Finally we summarize our results and compare to those reported in Zhao *et al.*[27] in Table 4.

In the comparison to the findings of Zhao *et al.*, we observe that our encoding method without feature selection improved the recall from 0.763 to 0.813 and enhanced precision from 0.716 to 0.835, both substantial improvements. It is further shown that with feature selection, a similar level of recall (0.800) and precision (0.820) can be reached with the use of an average of 96 selected features. In the comparison to the results of SVM with indicator vector, we observe that our encoding method without feature selection improved the recall from 0.750 to 0.813 and enhanced precision from 0.686 to 0.835. The ROC curve corresponding to the classifier obtained from the full set of features is shown in Figure 1. The area under the ROC was 0.9233, which compares favorably to 0.833 in Zhao *et al*. The substantial improvements in the recall, precision and area under the ROC demonstrated the effectiveness of the new encoding technique.

The top features on the sorted list imply the existence of a correlation between the peptide sequences and their stimulatory activity. For example, Gly, which appears most frequently at position 6 in positive peptides, ranks first in the sorted list. Phe, often observed at position 3 in negative peptides, was ranked in the second position in the list. Other top features include Lys (position 1 in negative peptides), Ile (positions 5 and 7), and Phe (position 4). The two latter residues are frequently seen in the positions involved in TCR recognition according to Zhao *et al.*

Our results suggest that the feature selection may extract the most important information that contributes to the stimulatory activity of T cell epitopes and non-epitopes. The combination of feature selection and SVM can be further explored in prediction of T cell epitope with a more complex encoding scheme.

4. Conclusion

A new encoding method for the direct recognition of T cell epitopes and non-epitopes through support vector machine has been developed. This encoding method combines the information in the conventional sparse encoding vector and BLOSUM scores. The

Table 1. Result of the proposed SVM without feature selection in the ten test datasets.

Test Set	Recall	Precision
1	7/8	7/8
2	7/8	7/7
3	8/8	8/10
4	5/8	5/5
5	5/8	5/8
6	8/8	8/12
7	7/8	7/7
8	6/8	6/7
9	6/8	6/7
10	6/8	6/9
avg	0.813	0.835

Table 2. Result of the proposed SVM with feature selection in the ten test datasets.

Test Set	Recall	Precision	#Features
1	6/8	6/8	160
2	7/8	7/7	120
3	8/8	8/12	60
4	6/8	6/6	60
5	4/8	4/7	160
6	8/8	8/9	80
7	7/8	7/7	80
8	6/8	6/7	80
9	5/8	5/6	60
10	7/8	7/11	100
avg	0.800	0.820	96

superiority of this encoding method and the effectiveness of the feature selection procedure were demonstrated.

Acknowledgments

The authors are thankful to Sarat Chandra Maruvda and Deepa Vijayraghavan for the assistance with computing environment. They would also like to express their gratitude to anonymous referees for useful comments.

Table 3. Result of SVM with the amino acid indicator vectors in the ten test datasets.

Test Set	Recall	Precision
1	6/8	6/10
2	6/8	6/6
3	8/8	8/10
4	7/8	7/10
5	5/8	5/9
6	5/8	5/12
7	8/8	8/15
8	5/8	5/8
9	5/8	5/8
10	5/8	5/5
avg	0.750	0.686

Table 4. Summary of the proposed methods to Zhao *et al.*

Method	Recall	Precision	#Features
method 1	0.813	0.835	200
method 2	0.800	0.820	96
method 3	0.750	0.686	200
method 4	0.763	0.716	100

method 1: proposed encoding method without feature selection.
method 2: proposed encoding method with feature selection.
method 3: encoding method with amino acid indicator vectors.
method 4: method by Zhao *et al.*. The results are taken from Zhao *et al.*[27]

References

1. M. Bhasin and G.P.S. Raghava. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine*, in press, 2004.
2. V. Brusic G. Rudy, M. Honeyman, J. Hammer and L. Harrison. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics.* 14:121-130, 1998.
3. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.
4. C. DeLisi and J.A. Berzofsky. T-cell antigenic sites tend to be amphipathic structures. *Proc. Natl. Acad. Sci.*, USA, 82:7078, 1985.
5. P. Dönnes and A. Elofsson. Prediction of MHC I binding peptides, using SVMHC. *BMC Bioinformatics,* 3:1-8, 2002.
6. V.H. Engelhard. Structure of peptides associated with class I and class II MHC molecules. *Annu. Rev. Immunol.*, 12:181-207, 1994.
7. D.R. Flower. Towards in Silico prediction of immunogenic epitopes. *Trends Immunol.*, 24:667-74, 2003.
8. K. Gulukota, J. Sidney, A. Sette and C. DeLisi. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.*, 267:1258-1267, 1997.

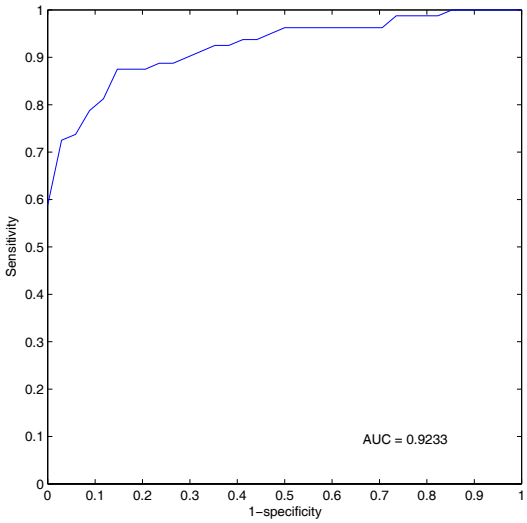

Figure 1. ROC curve of the SVM model without feature selection. AUC stands for the area under the ROC curve. The specificity is defined as $tn/(tn + fp)$.

9. J. Hammer, P. Valsasnini, K. Tolba, D. Bolin, J. Higelin, B. Takacs and F. Sinigaglia. Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell*, 74:197-203, 1993.
10. J. Hammer. New methods to predict MHC-binding sequences within protein antigens. *Curr. Opin. Immunol.*, 7:263-269, 1995.
11. S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci.*, USA, 89:10915-10919, 1992.
12. M.C. Honeyman, V. Brusic, N.L. Stone and L.C. Harrison. Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.*, 16:966-969, 1998.
13. T. Joachims. *Making Large Scale SVM Learning Practical - Advances in Kernel Methods-Support vector learning.* MIT Press, Cambridge, 1999.
14. A. Kidera, Y. Konishi, M. Oka, T. Ooi and H.A. Scheraga. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.*, 4:23-55, 1985.
15. G.E. Meister, C.G. Roberts, J.A. Berzofsky and A.S. De Groot. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences. *Vaccine*, 13:581-591, 1995.
16. K. Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring. *Proc. 16th Int'l Conf. on Machine Learning (ICML-99)*, 1999.
17. M. Nielsen, C. Lundegaard, P. Worning, S.L. Lauemϕller, K. Lamberth, S. Buus, S. Brunak and O. Lund. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Science*, 12:1007-1017, 2003.
18. K.C. Parker, M.A. Bednarek and J.E. Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chains. *J. Immunol.*, 152:163-175, 1994.
19. C. Pinilla, J.R. Appel and R.A. Houghten. Investigation of antigen-antibody interactions using a soluble, non-support-bound synthetic decapeptide library composed of four trillion sequences. *Biochem. J.*, 301:847-853, 1994.
20. H.G., Rammensee, T. Friede and S. Stevanoviic. MHC ligands and peptide motifs, first listing.

Immunogenetics, 41:178-228, 1995.
21. H.-G. Rammensee, J. Bachman, N. Philipp, N. Emmerich, O.A. Bachor and S. Stevanovic. SYFPEITHI: a database for MHC ligands and peptide motifs. *Immunogenics*, 50:213-219,1999.
22. D.E. Rumelhart, G.E. Hinton and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533-536, 1986.
23. 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.*, USA, 86: 3296-3300, 1989.
24. A. Sette, J. Sidney, M.F. del Guercio, S. Southwood, J. Ruppert, C. Dahlberg, H.M. Grey and R. T. Kubo. Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Mol. Immunol.*, 31:813, 1994.
25. J.L. Spouge, H.R. Guy, J.L. Cornette, H. Margalit, K. Cease, J.A. Berzofsky and C. DeLisi. Related Articles, Links Strong conformational propensities enhance T cell antigenicity. *J. Immunol.*, 138:204-212, 1987.
26. C.J. Stille, L.J. Thomas, V.E. Reyes, R.E. Humphreys. Hydrophobic strip-of-helix algorithm for selection of T cell-presented peptides. *Mol. Immunol.*, 24:1021-1027, 1987.
27. Y. Zhao, C. Pinilla, D. Valmori, R. Martin and R. Simon. Application of support vector machines for T-cell epitopes prediction. *Bioinformatics*, 19:1978-84, 2003.