Journal of Bioinformatics and Computational Biology

Direct Prediction of T-cell Epitopes Using Support Vector Machines with Novel Sequence Encoding Schemes

Lei Huang

*Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA**
lhuang7@uic.edu

Yang Dai†

*Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA**
yangdai@uic.edu

New peptide encoding schemes are proposed to use with support vector machines for the direct recognition of T cell epitopes. The methods enable the presentation of information on (1) amino acid positions in peptides, (2) neighboring side chain interactions, and (3) the similarity between amino acids through a BLOSUM matrix. A procedure of feature selection is also introduced to strengthen the prediction. The computational results demonstrate competitive performance over previous techniques.

Keywords: T cell epitope recognition; Support vector machine; Feature selection; Side-chain interaction.

1. Introduction

Computational T cell epitope identification currently relies on the prediction of peptide binding to major histocompatibility complex (MHC) molecules. In the pathway of antigen processing and presentation, 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[13,14,24,25,26,27] or quantitative matrices,[19,22] Artificial Neural Networks (ANNs),[3,12,16] and Support Vector Machines (SVMs).[7] On the other hand, research has indicated that T cell epitopes are not always high affinity MHC binders, and only a few of the potential MHC-binding peptides are T cell epitopes for a specified T cell receptor. Identification of good binders for specific MHC molecules may not provide accurate information on T cell epitopes, since a functional T cell

*Department of Bioengineering (MC063), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607, USA.
†Corresponding author.

response requires adequate MHC-peptide binding as well as proper interaction of the MHC-peptide ligand with a specific T cell receptor. The approaches mentioned above do not discriminate between T cell epitopes and non-epitopes which are both MHC binders.[11] The methods of direct prediction developed in 1980s were based on the structural analysis of T cell epitopes.[6,10,28,29]

Recently, techniques for the direct prediction of T cell epitopes based on machine learning techniques such as SVMs and ANNs with the use of sequence information have been proposed.[1,31] Bhasin and Raghava[1] compared the performance of ANNs and SVMs with the use of the amino acid (AA) indicator vector in which each amino acid of a peptide is represented by a 20-dimensional vector. Zhao *et al.*[31] employed 10 physical properties of the 20 amino acids to encode each residue of a peptide. These 10 properties include alpha-helix or bend-structure preference, bulk, beta-structure preference, hydrophobicity, normalized frequency of double bend, normalized frequency of alpha region, and pK-C.[18] This encoding represents a better class of information in comparison with the AA indicator vector for a peptide. Both studies have demonstrated that the SVMs have the potential to make relatively accurate prediction based on training with small data sets.

In this work, a new encoding scheme of peptides that combines the BLOSUM matrix[15] with the AA indicator vectors for the direct prediction of T cell epitopes was first investigated. Our method replaces each nonzero entry in the AA indicator vector by the corresponding value appeared in the diagonal entries in a BLOSUM matrix. This is different from other encoding methods in which each amino acid is simply represented by its BLOSUM score.[21] The characteristic of the new encoding method is the joint representation of information on both the position and similarity of the amino acids.

The above approach is based on the assumption of independent contribution of amino acids within a peptide to the TCR recognition. Interactions of the side chains of adjacent amino acids also exist and their effect was investigated by extending our new encoding scheme. Specifically, interaction indicator vectors of the side chains of adjacent amino acids are formed first, followed the extended Free-Wilson's additive concept of possible interactions between amino acid side chains.[30] Then each nonzero entry in the vector is replaced by the sum of values of the two corresponding residues appeared in diagonal entries in the BLOSUM matrix. It is noted that the additive concept has been explored in the affinity analysis of MHC binders based on quantitative structure-activity relationship (QSAR) studies by Doytchinova and Flower.[8,9]

The new encoding methods combined with SVMs were evaluated on two data sets: one used in Zhao *et al.*[31], and the other derived from the MHCBN database.[2] The computational results demonstrated competitive performance of the new methods in comparison with those using the indicator vector and the 10 physical properties of the amino acids.[31]

2. System and Method

2.1. *Training and testing data*

Two data sets were utilized in this experiments. The first data set is the one used in Zhao *et al.*[31]. As described,[31] the Melan-A-specific Cytotoxic T lymphocytes (CTL) clone LAU203-1.5 was derived from the tumor-infiltrated lymph node cells of a melanoma patient and the antigen recognition was assessed using a chromium-release assay (see details in Zhao *et al.*[31]) Among the 203 synthetic peptides, 36 were tested stimulatory (positive) and 167 were tested non-stimulatory (negative). These peptides have 10 amino acids. The second data set was extracted from the MHCBN database developed by Bhasin *et al.*[2] The set comprises 80 HLA-A*0201 restricted T cell epitopes and 140 peptides that are neither MHC binders nor T cell epitopes with respected to HLA A2 supertype (A*0201-07, A*0209, A*6802). These peptides consist of 9 amino acids. Since the numbers of the positive and negative peptides are unbalanced in both data sets, the numbers of positive and negative peptides in the training and testing sets were maintained in a similar ratio in our cross-validation procedure. It is noted that although the T cell epitopes in the two data sets are both HLA-A*0201 restricted, they are mutually distinctive, i.e., any epitope from the first data set does not overlap with any epitope from the second data set.

2.2. *Peptide encoding methods*

One of the important elements that influence the effectiveness of a SVM model is the design of the encoding method for training data. The AA indicator vector (1 present or 0 absent of an amino acid at a particular position) for a peptide has been used intensively. It provides very precise information on the position of each amino acid in a peptide. In order to consider the evolutionary relationship between residues occurring at the same position in peptides, Nielsen *et aI.*[21] have employed the BLOSUM score to represent each amino acid. More specifically, 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. For example, the hydrophilic amino acids Arg, Asn, Gln and the hydrophobic amino acid Met all have the same BLOSUM50 score 7. If they appear at the same position in the peptides, the method would not be able to discriminate them, although they may have different contribution to the recognition.

We introduce novel encoding techniques that combine the amino acid substitution matrix BLOSUM with the indicator vector at each position or indicator vector representing interactions of the adjacent or second adjacent residues. The first new encoding method replaces each non-zero value in the AA indicator vector by the BLOSUM score of the corresponding amino acid. By doing so it avoids the ambigu-

ity of the BLOSUM encoding method mentioned above. For a peptide with length s, the dimension of the encoding vector is thus $20s$. Obviously, 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.

The second encoding method is based on the model of Doytchinova and Flower[8,9] in which the additive concept with terms accounting for the possible interactions between the side chains of adjacent amino acids was considered for the contribution to the MHC-peptide binding affinity, in addition to the independent contribution from each residue's backbone. Each adjacent pair of amino acids is represented by a 400-dimensional vector with each entry corresponding to one of the combinations of the 20 residue pairs. This vector is concatenated to the $20s$-dimensional encoding vector in the first method, resulting in a total number of $20s + 400(s-1)$ for each peptide. In order to combine information from the BLOSUM matrix, each "1" in the vector representing the side chain interaction is replaced by the sum of the BLOSUM scores of the two corresponding residues.

The third method is to consider interactions of every second side chains in the model of Doytchinova and Flower,[8,9] i.e., the interactions between residues at position i and $i+2$, $i = 1, ..., s-2$. Use the same 400 combinations of amino acids, this part of information can be encoded with a $400(s-2)$-dimensional vector for a peptide. Therefore, the overall dimension of the vector in this encoding method is $20s + 400(s-2)$. Likewise, each "1" in the vector is replaced by the sum of two corresponding BLOSUM scores.

2.3. *Feature ranking*

Since the feature vectors obtained from the encoding methods are sparse, a feature selection procedure is used to exclude the features that appear less frequently and less discriminative. Here a simple procedure for feature selection is carried out based on 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 $\boldsymbol{x}$. 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 machines*

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:

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

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 $(i = 1,\ldots,m)$ are 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 please refer to Cristianini and Shawe-Taylor.[4]

According to our preliminary study, the linear kernel $(\phi(\boldsymbol{x}) = \boldsymbol{x})$ in SVMs 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,[20] however, the best ratio is usually determined through cross-validations by searching in the range of $[1, n_2/n_1]$. Accordingly, two parameters associated with a SVM model need to be optimized:

(1) C_+ : the trade-off between the positive training error and class separation;
(2) J : the ratio of C_- to C_+, i.e., $J = C_-/C_+$.

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

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

and recall (sensitivity):

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

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 (resp. negative) peptides which are true negative (resp. positive).

The F-score, given by

$$F\text{-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}},$$

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

Procedures. The experimental protocol is a 5-fold cross-validation (CV). Each time one fold was set aside as a testing set and the remaining 4 folds were used as the training set. The features were ranked according to their Fisher's scores in a descending order based on the training set. For a fixed number of features, a 10-fold cross-validation on the training set was used to optimize the parameters C_+ and J in terms of F-scores. To obtain the final classifier for the training set, the training was carried out again by using the optimal parameters on the entire training set and evaluated on the testing set for recall and precision. This process was repeated 5 times with different testing and training folds; the average values of recall and precision were calculated. The flowcharts of the procedures are shown in Fig. 1.

A greedy strategy was designed for the search of the best F-score and the corresponding pair of parameters within the given range of C_+ and J in the 10-fold CV on a training fold. More precisely, the search starts with a randomly selected pair (C_+, J) on a coarsely defined grid for the area determined by the ranges of C_+ and J. The neighboring pair of the current parameters (C_+, J) is chosen if it is associated with a better F-score. The search stops if no any neighboring pair can improve the F-score. A smaller area consisting of the current stopping grid point and its four neighboring grid points is identified. Then a refined grid, usually by taking half of the current grid size, is constructed for the new area. The above local search is repeated until a prescribed grid size is reached. The entire local search procedure is run for a few hundred times to ensure the identification of the best F-score. It is noted from our preliminary study that this local search can successfully find a F-score very close to the one obtained from an exhaustive search on all grid points with the prescribed grid size.

Experimental setting. In our experiment, the number of random starts for the local search is 250. The ranges of C_+ and J are [0.0001, 5] and [1, 5], respectively; the final grid sizes are 0.0001 for C_+ and 0.05 for J, respectively. Denote v_0 and v_1 the numbers of minimum and maximum features selected, respectively; and Δ the incremental step size for the number of features. The values of v_0, v_1, and Δ for the first encoding are respectively 60, $20s$ (the number of total features), and 20.

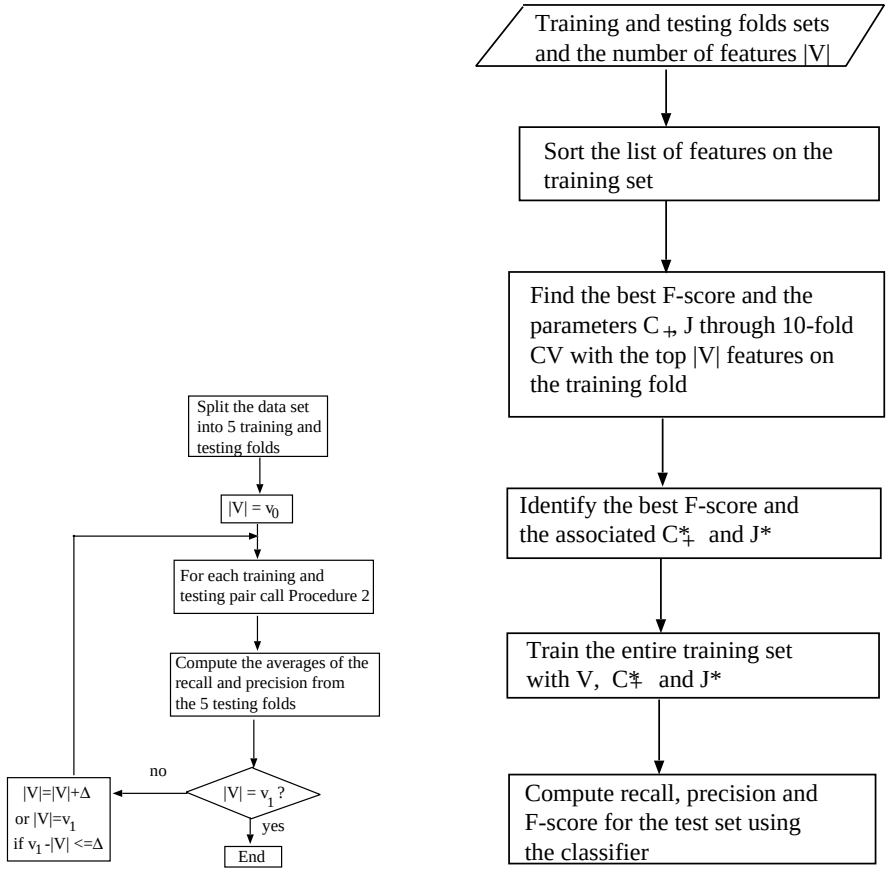

Fig. 1. **Left**. Procedure 1: the overall procedure of the training and testing. The numbers v_0 and v_1 are the minimum and maximum numbers of features selected respectively; Δ is the incremental step size for the number of features. All these numbers are prescribed. **Right**. Procedure 2: the flowchart of the training and testing with a fixed number features. The 10-fold cross-validation for the identification of the best pair of parameters (C_+^*, J^*) is carried out by a local optimization procedure.

The values of v_0 and v_1 for the second and third encoding methods are respectively 60 and the number of features remained after the execution of the feature removal procedure described below; and the value of Δ is 40.

For the second and third encoding methods, many pairs of residues only occur rarely due to the small number of training peptides. Therefore, those features that appeared less than 3 times in the training set were removed to avoid overfitting. This procedure reduced the feature numbers to approximately 270 for both sets. The BLOSUM50 matrix and the SVMLight package[17] were used in our experiments. The other BLOSUM matrices were also attempted, however, no significant difference in performance has been observed (results not shown).

In order to provide performance statistics of the proposed methods, the experiments were repeated 5 times for the 5-fold CV. The average values of recall, precision and the corresponding standard deviations were calculated.

3. Results

For each encoding method, the results with and without feature selection are reported. Only the best results with feature selection are presented here. For comparison, the experiment with the AA indicator vector was also performed. The notations for the different encoding methods are summarized in Table 1; and results are shown in Tables 2 and 3.

Table 1. The legends used in Table 2.

Legend	Encoding Method
AA-B	AA indicator vector with BLOSUM scores
AA-B-s	AA-B with feature selection
SI1-B	adjacent interaction indicator vector with BLOSUM scores
SI1-B-s	SI1-B with feature selection
SI2-B	second adjacent interaction indicator vector with BLOSUM scores
SI2-B-s	SI2-B with feature selection
AA	AA indicator vector
AA-s	AA indicator vector with feature selection
Zhao	encoding method by Zhao *et al.*[31]

Table 2. Summary of the comparison of proposed methods to existing methods for Zhao's data set.

Method	Precision	Recall	#Features
AA-B	0.798 (0.038)	0.775 (0.033)	200
AA-B-s	0.835 (0.057)	0.790 (0.051)	140
SI1-B	0.857 (0.033)	0.800 (0.038)	263
SI1-B-s	0.873 (0.028)	0.825 (0.028)	220
SI2-B	0.764 (0.047)	0.775 (0.041)	220
SI2-B-s	0.788 (0.043)	0.775 (0.064)	100
AA	0.707 (0.033)	0.730 (0.029)	200
AA-s	0.714 (0.076)	0.745 (0.026)	140
Zhao	0.716	0.763	100

[a] Zhao : The results are taken from Zhao *et al.*[31]
[b] The values in parentheses stand for standard deviations.

It is observed from Table 2 that all three new encoding methods with or without feature selection outperform Zhao's and the AA indicator vector methods. Feature selection has positive effect on performance. In comparison to the findings of Zhao

Table 3. Summary of the comparison of proposed methods to existing methods for the second data set.

Method	Precision	Recall	#Features
AA-B	0.764 (0.009)	0.770 (0.039)	180
AA-B-s	0.779 (0.021)	0.813 (0.029)	100
SI1-B	0.756 (0.028)	0.733 (0.045)	250
SI1-B-s	0.751 (0.038)	0.778 (0.034)	100
SI2-B	0.768 (0.046)	0.765 (0.031)	235
SI2-B-s	0.768 (0.017)	0.828 (0.031)	100
AA	0.759 (0.016)	0.775 (0.039)	180
AA-s	0.784 (0.015)	0.825 (0.034)	100

et al., the AA-B-s encoding method improved the recall from 0.763 to 0.790 and enhanced precision from 0.716 to 0.835, a substantial improvement.

The encoding methods with the additive contribution from the side chain interaction effect positively to prediction with different magnitude. The SI1-B-s method demonstrated superior performance; it improved recall from 0.763 to 0.825 and precision from 0.716 to 0.873. Nevertheless, the SI2-B-s performed only slightly better than the AA-B method. These observations imply that interactions of side chains of the adjacent residue pairs may be more important than those from the second adjacent residue pairs.

However, the new encoding methods did not demonstrate evidence of the improved performance over the 0-1 encoding methods for the second data set (see Table 3). Nevertheless, the evidence of improvement made through feature selection can be observed. For example, with the use of 100 selected features for each encoding method, it reaches a similar or an enhanced performance.

In order to further compare the performance with other direct T cell prediction methods, we obtained predicting results from the web server CTLPred developed by Bhasin and Raghava,[1] which is the only publicly available direct T cell predictor. This system was trained based on all 9 amino acid long CTL epitopes extracted from MHCBN database[2] and an equal number of CTL non-epitopes. It has several predictors including the QM, ANN, and SVM. The consensus method constructed from the ANN and SVM models was selected in our test, since it has the best reported performance.[1] In this framework, peptides which are predicted as epitopes by both methods are considered as epitopes; otherwise they are considered non-epitopes. We have tested several combinations of threshold values for the ANN and SVM models. For the first data set, the best performance indicates a precision value of 0.24 and a recall value of 0.5 for Zhao's data set; both are significantly lower in comparison with those obtained from the proposed methods in this paper. For the second data set, the predictor produced a precision value of 0.8 and a recall value of 0.65. Note that these peptides were actually used in training. This fact may contribute positively to the performance.

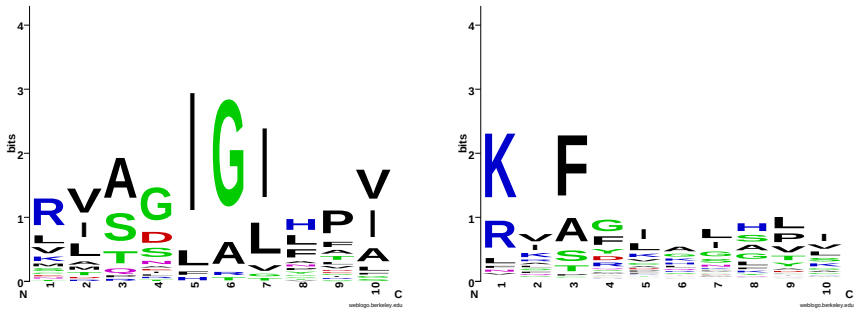

Fig. 2. The sequence logos obtained for the 36 positive peptides (left) and the 167 negative peptides (right) of Zhao's data.

4. Discussion

The sequence logos of peptides for each data set are respectively provided in Figs. 2 and 3 to facilitate the discussion.[5] It is obvious that the sequence logos for these two datasets are very different. The new encoding methods appear to be more suitable for Zhao's data set, which comprises synthetic peptides. Note that the T cell epitopes are not necessarily MHC binders. Conversely, the new methods appear to have limited impact on the performance for the second data set. It may be explained by the nature of the T cell epitopes. They are MHC binders and possess strong motifs at the anchor positions P2 and P9. The signals are so intense and discriminative so that the BLOSUM encoding methods could not enhance the prediction further. However, the feature selection still contributes positively to the prediction.

The detail analysis of the top 140 features for Zhao's data set suggested that Fisher's scores do provide raking of the importance of positions on peptides and the hydrophobic residues occurred in these positions. 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.*. It was also suggested residues at positions 4-8 to be primarily involved in TCR recognition.[23]

In order to analyze the contribution from each amino acid at different positions, the average of the weight vectors $\boldsymbol{w}$ in the SVMs for the optimized classifiers was calculated for the AA-B encoding method (see Fig. 4). It is observed that Ile at positions 5 and 7, and Gly at position 6 have very high positive values; and Phe at positions 3 and 4 contribute negatively. These weights match well to the sequence

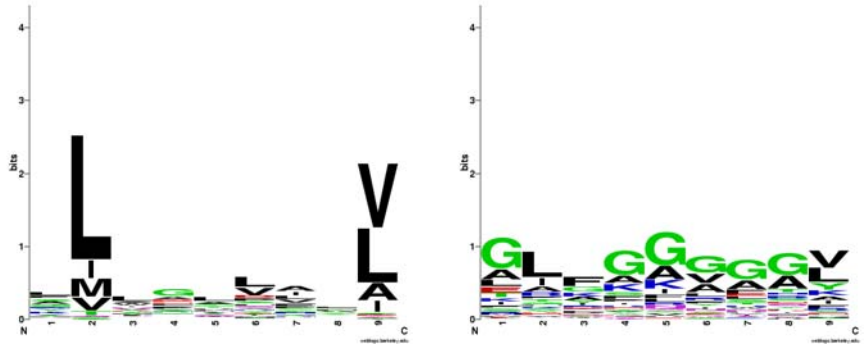

Fig. 3. The sequence logos obtained for the 80 positive peptides (left) and the 140 negative peptides (right) of the second data set.

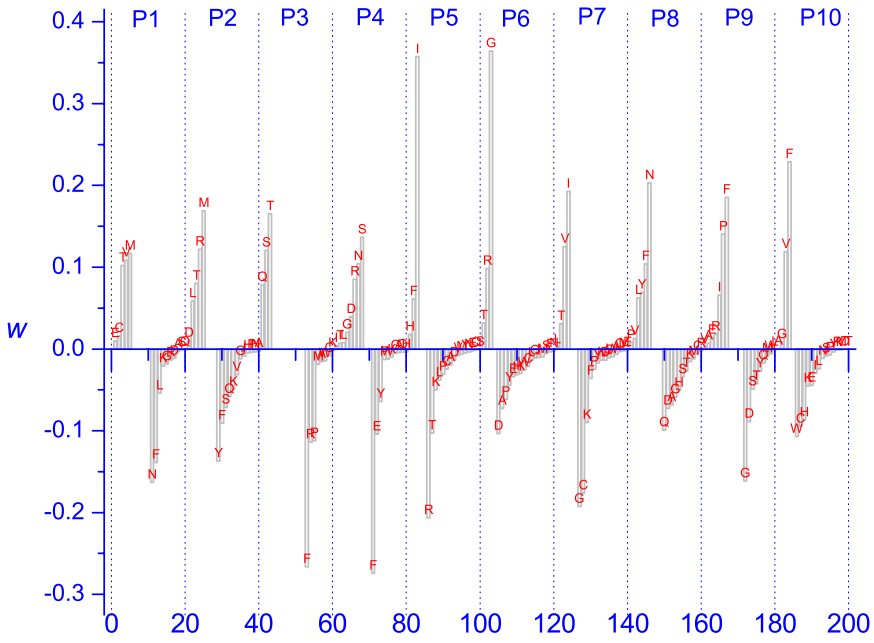

Fig. 4. The average weight values w_i $(i = 1, ..., 200)$ determined from the optimal classifiers with the AA-B method for Zhao's data set.

logos. The comparison of our findings with those of Zhao's is detailed in Table 4. Specifically,

Position 9: our method reaches a similar conclusion as that of Zhao's, except Leu has a negative contribution in our case.

Position 2: our method found no weight for Cys and Glu. This seems to be consistent to the sequence logos, where no significant information is presented for Cys and Glu. In addition, Thr was found with positive weight in our method, but was reported having negative contribution in Zhao's method. Examination of the logos indicates that Thr appears more often in positive set than in the negative set, which seems to support our results.

Position 4: both methods agree on the positive contribution from Arg, Ser and Thr.

Positions 5 and 7: both methods found Ile and Phe contribute positively. Particularly, Ile has the highest weight from our method, consisting perfectly with the sequence logo. However, Ala was reported with a negative weight in our method and positive in Zhao's method.

Position 6: the highest positive contribution from Gly and the negative contribution of Pro were confirmed from both methods.

Position 8: both methods agree on Phe's positive contribution; however, our method also indicates a positive contribution from Asn.

Further investigation of the weight vector obtained from the SI1-B-s method reveals that the weights for Ile (position 5) and Gly (position 6) pair, Gly (position 6) and Ile (position 7) pair, Gly (position 4) and Ile (position 5) pair are 0.637, 0.393, and 0.239, respectively. These features are also among the top features ranked by Fisher's score. It is not surprising that these features have large weights since their positions are involved in the TCR recognition. Furthermore, for pairs contain amino acids at position 2 and position 9, most of the corresponding features have be ranked relatively high compared to those in other positions (e.g. pairs of Lys (position 1) and Lys(position 2), Lys (position 2) and Phe (position 3), and Phe (position 8) and Pro (position 9); with weights -0.0222, -0.0253, and 0.2017 respectively). These findings imply that the combination of feature selection and the SVM may capture the important information for the prediction of T cell epitopes.

5. Conclusion

New encoding methods for the direct recognition of T cell epitopes and non-epitopes through support vector machines have been developed. These encoding methods combine information in the conventional sparse encoding vector and BLOSUM scores. The superiority of the encoding methods and the effectiveness of the feature selection procedure were demonstrated for the data set comprising synthetic peptides. Our results suggest that the feature selection may extract the most important information that contributes to the stimulatory activity of T cell epitopes

Table 4. Comparison of the analysis of residue contributions at different positions.

Zhao *et al.*	Ours
Substitution of Thr in position 9 with hydrophobic residues Phe, Leu and Ile yielded the highest SVM scores.	Phe(0.185) and Ile(0.066) have positive weights; while Leu(-0.001) has negative weight.
At position 2, Cys, Arg, Glu, Met, and Thr yielded higher SVM scores while Leu kept almost the same score as Ala. Substitution with polar residues Ser, Thr, and Asn at position 2 would yield negative SVM scores.	Met(0.169), Arg(0.122), Thr(0.080), and Leu(0.059) have positive weights, while Ala(-0.002) negative. No weights for Cys and Glu. Ser (-0.071), Thr(0.080), Asn (-0.003)
At position 4, Arg, Ser, and Thr doubled the SVM scores compared to Ala in the template.	At Position 4, Arg(0.085), Ser(0.137) and Thr(0.007) are positive; Ala(-0.019) negative.
Gly was the only amino acid to be allowed at position 6 in order to keep the peptide to be predicted positive. The nonpolar residue Pro is the least favored one.	At position 6, Gly(0.364) has the highest positive weight; Pro(-0.062) negative.
Hydrophobic residues were favored at position 5 (Ile, Phe, Ala, Val) and 7 (Ala, Ile, Leu, Val).	At position 5, Phe(0.061) and Ile(0.357) have positive weights; Ala(-0.019) negative. No weight for Val. At position 7, Leu(0.002) and Val(0.125) have positive weights, and Ala(-0.006) negative.
Replacing Leu with hydrophobic residues Phe or Ile at position 8 leads to increase of SVM scores.	At position 8, no weight for Ile. Leu(0.62), Phe(0.62), and Asn(0.20) have positive weights.

[a] Zhao's findings are based on the SVM prediction of the single amino acid substitutions of a synthetic decapeptide EAAGIGILTV, a predicted T cell epitope by their method. The SVM score means the value of the decision function.
[b] Positions 2, 9, and 10 are considered to be the putative MHC anchors.[24]
[c] Residues at positions 4-8 were suggested to be primarily involved in TCR recognition.[23]

and non-epitopes.

Acknowledgments

This research is partially supported by National Science Foundation (EIA-022-0301) and Naval Research Laboratory (N00173-03-1-G016). The authors are thankful to Deepa Vijayraghavan for her assistance with computing environment. They would also like to express their gratitude to anonymous referees for useful comments.

References

1. M. Bhasin, G. P. S. Raghava. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine*. **22**:3195-3204, 2004.
2. M. Bhasin, H. Singh, G. P. S. Raghava. MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*. **19**:665-666, 2003.
3. 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.
4. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.
5. D. E. Crooks, G. Hon, J. M. Chandonia and S. E. Brenner. A sequence logo generator. *Genome Research*, **14**:1188-1190, 2004.
6. C. DeLisi and J. A. Berzofsky. T-cell antigenic sites tend to be amphipathic structures. *Proc. Natl. Acad. Sci.*, USA, **82**:7048-7052, 1985.
7. P. Dönnes and A. Elofsson. Prediction of MHC I binding peptides, using SVMHC. *BMC Bioinformatics,* **3**:1-8, 2002.
8. I. A. Doytchinova and D. R. Flower. Quantitative approaches to computational vaccinology. *Immunol Cell Biol.* **80**:270-279, 2002.
9. I. A. Doytchinova and D. Flower. The HLA-A2-supermotif: a QSAR definition. *Org. Biomol. Chem.* **1**(15):2648-2654, 2003.
10. V. H. Engelhard. Structure of peptides associated with class I and class II MHC molecules. *Annu. Rev. Immunol.*, **12**:181-207, 1994.
11. D. R. Flower. Towards in Silico prediction of immunogenic epitopes. *Trends Immunol.*, **24**:667-74, 2003.
12. 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.
13. 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.
14. J. Hammer. New methods to predict MHC-binding sequences within protein antigens. *Curr. Opin. Immunol.*, **7**:263-269, 1995.
15. S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci.*, USA, **89**:10915-10919, 1992.
16. 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.
17. T. Joachims. *Making Large Scale SVM Learning Practical - Advances in Kernel Methods-Support vector learning*. MIT Press, Cambridge, 1999.
18. 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.
19. 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.
20. 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.
21. M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemøller, 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.
22. 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.
23. M. R. Parkhurst, M. L. Salgaller, S. Southwood, P. F. Robbins, A. Sette, S. Rosenberg and Y. Kawakami. Improved induction of melanoma-reactive CTL with peptides from

the melanoma antigen gp100 modified at HLA-A.0201-binding residues. *J. Immunol.*, **157**:2539-2548, 1996.

24. H. G. Rammensee, T. Friede and S. Stevanoviic. MHC ligands and peptide motifs, first listing. *Immunogenetics*, **41**:178-228, 1995.
25. 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.
26. 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.
27. 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.
28. J. L. Spouge, H. R. Guy, J. L. Cornette, H. Margalit, K. Cease, J. A. Berzofsky and C. DeLisi. Strong conformational propensities enhance T cell antigenicity. *J. Immunol.*, **138**:204-212, 1987.
29. 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.
30. S. Tomic, L. Nilsson and R. C. Wade. Nuclear receptor-DNA binding specificity: A COMBINE and Free-Wilson QSAR analysis. *J. Med. Chem.*, **43**:1780-92, 2000.
31. 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.