BIOINFORMATICS

Vol. 19 no. 15 2003, pages 1978–1984
DOI: 10.1093/bioinformatics/btg255

Application of support vector machines for T-cell epitopes prediction

Yingdong Zhao[1], Clemencia Pinilla[2], Danila Valmori[3], Roland Martin[4] and Richard Simon[1,*]

[1]Biometric Research Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA, [2]Torrey Pines Institute for Molecular Studies, San Diego, CA 92121, USA, [3]Division of Clinical Onco-Immunology, Ludwig Institute for Cancer Research, University Hospital (CHUV), Lausanne, Switzerland and [4]Neuroimmunology Branch, National Institute of Neurological Disorder and Stroke, Bethesda, National Institutes of Health, MD 20892, USA

Received on October 24, 2002; revised on March 12, 2003; accepted on April 7, 2003

ABSTRACT

Motivation: The T-cell receptor, a major histocompatibility complex (MHC) molecule, and a bound antigenic peptide, play major roles in the process of antigen-specific T-cell activation. T-cell recognition was long considered exquisitely specific. Recent data also indicate that it is highly flexible, and one receptor may recognize thousands of different peptides. Deciphering the patterns of peptides that elicit a MHC restricted T-cell response is critical for vaccine development.

Results: For the first time we develop a support vector machine (SVM) for T-cell epitope prediction with an MHC type I restricted T-cell clone. Using cross-validation, we demonstrate that SVMs can be trained on relatively small data sets to provide prediction more accurate than those based on previously published methods or on MHC binding.

Contact: rsimon@mail.nih.gov

Supplementary information: Data for 203 synthesized peptides is available at http://linus.nci.nih.gov/Data/LAU203_Peptide.pdf

INTRODUCTION

Peptides degraded from foreign or self-proteins bind to major histocompatibility complex (MHC) molecules. The MHC–peptide complex can be recognized by T-cell receptors and trigger an immune response. Identifying characteristic patterns of immunogenic peptide epitopes can provide fundamental information for understanding disease pathogenesis and etiology, and for therapeutics such as vaccine development.

Due to the complexity of the tri-molecular complex (peptide, MHC molecule, and T-cell receptor), early studies focused on the interaction between peptide and MHC. Structural studies and systematic binding analyses have provided insight into the peptide binding patterns to MHC (Engelhard, 1994; Madden, 1995; Rothbard and Gefter, 1991; Sette *et al*., 1994). Mathematical approaches including binding motifs (Hammer *et al*., 1993; Hammer, 1995; Rammensee *et al*., 1995; Sette *et al*., 1989), quantitative matrices (Parker *et al*., 1994; Southwood *et al*., 1998; Sturniolo *et al*., 1999), artificial neural networks (ANNs) (Brusic *et al*., 1998; Gulukota *et al*., 1997; Milik *et al*., 1998), and support vector machines (SVMs) (Dönnes and Elofsson, 2002) used to model these interactions have led to an increasingly more refined understanding of MHC/peptide binding.

MHC binders are not always T-cell epitopes however. Efforts to predict candidate T-cell epitopes have been utilized ANNs (Honeyman *et al*., 1998). A full ANN with an indicator for each amino acid at each position requires 200 input nodes (20 amino acids $\times 10$ positions). Large ANNs require very large amounts of data to avoid obtaining poor predictions resulting from over-fitting a limited set of training data (Rumelhart *et al*., 1986). The number of weights for edges joining m input nodes to h hidden layer nodes is $h \times m$. Hence even with only $h = 2$, a prohibitive amount of data is required for properly training a network with 402 parameters. To generate such an extensive amount of data for a single TCR is very expensive. Accurate modeling strategies that are more efficient in use of TCR proliferation assay data and antigen recognition data are needed.

In recent years, various pattern recognition techniques have been applied in biology. SVMs are one of the most powerful new techniques and have been effective in DNA sequence analysis, protein structure prediction and gene expression pattern discovery (Brown *et al*., 2000; Furey *et al*., 2000; Guyon *et al*., 2002; Hua and Su, 2001; Zien *et al*., 2000; Ding *et al*., 2001; Zavalijevski *et al*., 2002). SVMs are particularly appealing for T-cell epitope prediction because of the ability of SVMs to build effective predictive models when the dimensionality

*To whom correspondence should be addressed.

Published by Oxford University Press

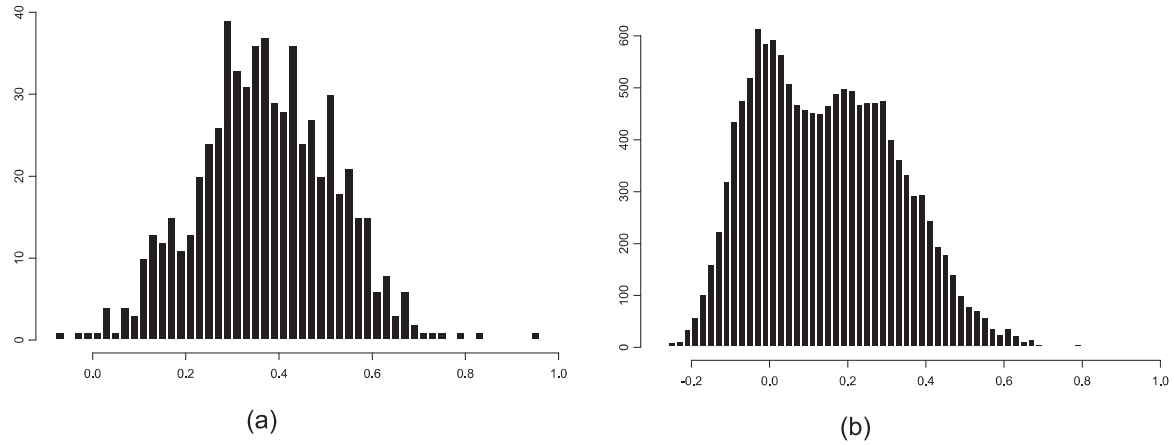

Fig. 1. Pairwise comparison of the peptides in positive (**a**) and negative (**b**) groups. Pearson correlation coefficients were calculated for all pairs in each group using 10 physical factors in 10 positions.

of the data is high and the number of observations is limited. SVMs are based on a strong theoretical foundation for avoiding over-fitting training data and they do not have the problem of the numerous local optimal that limit ANN models (Vapnik, 1995).

We analyzed our relatively small data set by building a SVM. This is the first time a SVM has been used for T-cell epitope prediction.

SYSTEM AND METHODS

T-cell clone and antigen recognition assay

Melan-A-specific CTL clone LAU203-1.5 was derived from tumor-infiltrated lymph node cells of a melanoma patient and antigen recognition was assessed using a chromium-release assay as previously described (Valmori *et al*., 1998).

Peptide synthesis and test

Peptides were synthesized by the simultaneous-multiple-peptide-synthesis methods (Pinilla *et al*., 1994) and characterized using HPLC and mass spectrometry.

LAU203-1.5 is an A*0201 restricted T-cell clone (TCC) from tumor-infiltrated lymph node cells of a melanoma patient. 203 synthetic peptides were selected based on results using single- and multiple-amino acid-substitutions and combinatorial peptide library experiments with a chromium release antigen recognition assay (Rubio-Godoy *et al*., 2002). These peptides were tested against the LAU203-1.5 clone using the same assay. A peptide with percentage specific lysis higher than 10% was considered positive.

Training and test data sets

Due to the imbalance of two classes in the data set (36 stimulatory peptides and 167 non-stimulatory peptides), we first divided the data into positive and negative groups. Then in each group random sampling was used to select 80% of the total peptides for training and 20% as a test set. Finally the positive and negative groups were combined separately in the training and test sets. This procedure was repeated independently 10 times.

Each amino acid in a peptide was encoded by ten factors. These orthogonal factors were obtained from 188 physical properties of 20 amino acids via multivariate statistical analyses by Scheraga's group (Kidera *et al*., 1985). They account for 86% of the variance of the 188 physical properties. These factors included alpha-helix or bend-structure preference, bulk, beta-structure preference, hydrophobicity, normalized frequency of double bend, normalized frequency of alpha-region, and pK-C. This encoding reduces the dimension of predictors by half while enabling structural and biophysical properties to be better represented compared to using amino acid indicator variables. Since our peptides are all 10-mers, the total number of input variables is 100.

To ensure that the peptides were sufficiently dissimilar for the cross-validation to be valid, we calculated the pairwise Pearson correlation coefficients for all positive peptides and negative peptides. Figure 1 shows the histograms of the correlation coefficients in each group. Only 5% in the positive group and 1% in the negative group have correlations larger than 0.6.

Training a support vector machine

SVM training was performed using *SVMlight* (version 4.0) (Joachim, 1999). There were 100 input variables, which represent the ten positions in the peptide. The class values were set to 1 for positive peptides and -1 for negative peptides. The threshold to predict positive or negative peptide was set to 0 by default.

For two group classification, SVM separates the classes with a surface that maximizes the margin between them. It is an approximate implementation of the Structural Risk

Minimization induction principle, which attempts to minimize with the generalization error for independent data rather than minimizing the mean square error for the training set (Vapnik, 1995).

SVM classification of a sample with a vector x of predictors is based on:

$$f(x) = \text{sign}\left(\sum_i y_i \alpha_i k(x_i, x) + b \right)$$

where the kernel function $k(\ ,\ )$ measures the similarity of its two vector arguments. For linear SVM, the inner product kernel function is used. If $f(x)$ is positive, then the sample is predicted to be in class $+1$; otherwise class -1. The summation is over the set of 'support vectors' that define the boundary between the classes. Support vector x_i is associated with a class label y_i that is either $+1$ or -1. The $\{\alpha_i\}$ and b coefficients are determined by 'learning' the data.

For each training set consisting of 80% of the observations, a fully specified linear SVM was developed. This SVM model was then applied to the 20% test set. During learning on the 80% training set, leave-one-out cross-validation was employed to automatically optimize the relative misclassification costs for the two classes and to optimize the tuning parameter that reflects the trade-off between the training error and class separation. This leave-one-out process only utilized data from the 80% training set.

Training and testing were repeated ten times for randomly determined training/test set partitions. The final indexes were averaged over the ten replicates.

ANN and Decision Tree Classifiers

The same training/test set partitions used for SVM analyses were also used for building and evaluating ANN and Decision Tree classifiers. The same input vector encoding was also used. The neural network analysis was performed using the Neuroshell 2 software package (Ward Systems Group). We chose a feed-forward architecture with three layers (single hidden layer). There were 100 neurons in the input layer and two neurons in the hidden layer. Each SVM training set was separated into training set and control set at a 9 to 1 ratio. The control set was used for controlling ANN training. The ANN production set was the same as the SVM test set. The learning rate and momentum were both set to 0.1, and the learning epoch was 2000. The threshold to predict positive or negative peptide was set to 0 by default.

The classification trees were generated using the Classification and Regression Tree approach (Breiman *et al*., 1984) implemented in S-plus 2000 software. The predictors were the same as the SVM input. The responses were set to 1 for positive peptides and 0 for negative peptides. We used the same data sets generated for SVM. Ten-fold cross-validation within each training set was used for optimally pruning the trees.

A*0201 peptide-binding based predictions

SYFPEITHI is a profile based method to predict MHC binding peptides (Rammensee *et al*., 1999). Thirteen different MHC class I types of binding peptides can be predicted. It is publicly available through a website (http://syfpeithi.bmi-heidelberg.com/Scripts /MHCServer.dll/EpPredict.htm). A*0201 was used to predict all 203 melanoma clone LAU203-1.5 peptides. The threshold between binders and non-binders was optimized.

SVMHC is based on SVM to predict the binding of peptides to MHC type I molecules (Dönnes and Elofsson, 2002). It contains prediction for 26 MHC class I type from the MHCPEP database and 6 MHC class I types from the SYFPEITHI database. It can be publicly accessed (http://www.sbc.su.se/svmhc/). A*0201 was used to predict all 203 melanoma clone LAU203-1.5 peptides. The threshold between binders and non-binders was kept as zero by default.

RESULTS

Since identifying stimulatory (positive) peptides is of greatest concern, sensitivity and positive predictive value (PPV) were used to evaluate the models. Sensitivity is the portion of all positive peptides that are correctly identified. PPV is the probability that a peptide predicted to be positive actually does stimulate the TCC. Sensitivity indicates the ability of the classifier to detect real epitopes whereas PPV reflects the efficiency of the method. A classifier with low PPV will result in the generation of numerous non-stimulatory peptides for the next rounds of testing.

Table 1 shows the performance of the SVM for the 10 test sets. The average cross-validated sensitivity and PPV were 76.3 and 71.6%, respectively, for the 10 test sets. An averaged ROC curve (Swets, 1988) was also determined based on applying SVM models to the 10 different test sets (Table 2). The area under the averaged ROC curve was 0.919 (Fig. 2).

ANN models were optimized by modifying the learning rate and momentum. The optimized models gave an average sensitivity of 55.0% and PPV of 81.7% on the 10 test sets (Table 2). Decision tree classifiers gave an average sensitivity of 46.3% and PPV of 86.7% for the same ten test sets (Table 2).

In order to compare with other MHC-binding based predictions, we applied both SVMHC and SYFPEITHI to predict all 203 synthetic peptides of melanoma clone LAU203-1.5. For SVMHC based on training data from the SYFPEITHI database, the sensitivity was 30.6% and PPV was 45.8%. For SVMHC based on training data from the MHCPEP database, the sensitivity was 38.9% and PPV was 45.1%. For SYFPEITHI, the sensitivity was 86.1% and PPV was 34.8% (Table 2).

Among the 203 synthesized peptides, 105 peptides were predicted with high scores using the score matrix based approach (Zhao *et al*., 2001) and were selected as positive peptides for synthesis. None of the remaining 98 unrelated

Table 1. Cross-validation of SVM models in the 10 test sets (42 peptides)

Test Set	Sensitivity	PPV
1	4/8	4/7
2	7/8	7/9
3	7/8	7/9
4	5/8	5/6
5	6/8	6/7
6	5/8	5/6
7	6/8	6/11
8	5/8	5/8
9	8/8	8/13
10	8/8	8/11

Table 2. Comparison of SVM performance to other methods

Method	Sensitivity	PPV
SVM	0.763	0.716
ANN	0.550	0.817
Decision tree	0.463	0.867
Score matrix[a]	1.000	0.343
SYFPEITHI	0.863	0.348
SVMHC (a)[b]	0.306	0.458
SVMHC (b)[c]	0.389	0.451

[a]The analysis was based on an approach using a Z-matrix as described (Zhao *et al*., 2001).
[b]The SVM model was trained based on A*0201 restricted MHC binding data from SYFPEITHI database.
[c]The SVM model was trained based on A*0201 restricted MHC binding data from MHCPEP database.

peptides was significantly recognized by TCR and all of them were predicted to have relative low scores. This translated the sensitivity and the PPV of the score matrix based approach to 100 and 34.3%, respectively (Table 2). The area under the ROC curve was 0.833 (Fig. 2).

DISCUSSION

The ANN and decision tree classifiers had slightly better positive predictive values than the SVM but their sensitivities were substantially lower. The ANN model had many more parameters than the SVM and probably requires a larger number of training peptides for equivalent performance. The ANN model was more difficult to fit and optimize because of the number of parameters. Although the threshold distinguishing stimulatory and non-stimulatory peptides can be shifted to increase the sensitivity, the specificity and the PPV will suffer corresponding decreases. For our data set, the ratio of the numbers of peptides in two classes is about 1 : 5. A small reduction of specificity will lead to a much larger reduction of PPV.

Decision tree classifiers are attractive because they are easily interpretable. There are many kinds of decision tree classifiers, with considerable arbitrariness in the criteria for determining the variables and cut-points used for splits. Decision tree classifiers are data-greedy; each split partitions the data into disjoint subsets which are then analyzed separately to determine

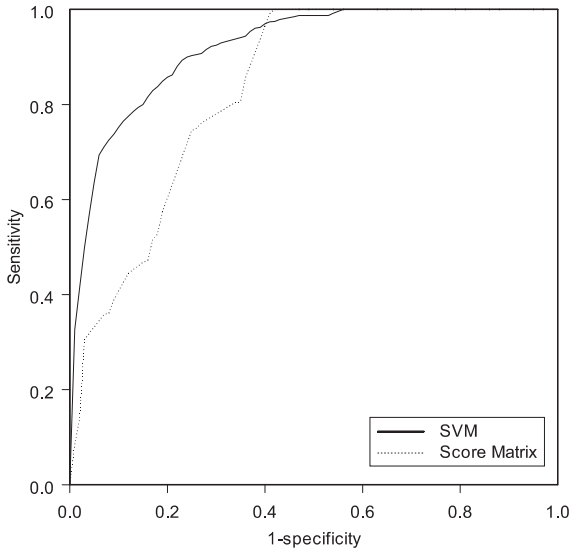

Fig. 2. ROC comparison of SVM to score matrix based approach.

their next splits. It is easy to over-fit decision tree classifiers and such models require large training sets. The Classification and Regression Tree methodology that we used employs cross-validation within the training set to optimally prune the tree and avoid over-fitting. Nevertheless, the optimal decision tree classifier had a sensitivity considerably less than the SVM.

Selection of input variables, a suitable kernel, and optimal learning parameters play key roles in developing SVMs. We examined different types of input variables for predicting TCR epitopes. We first used indicators for the 20 amino acids (1 present or 0 absent) at each position of the 10-mer peptide. The second set of variables we evaluated was based on use of the amino acid substitution matrices such as Blossum (Henikoff and Henikoff, 1992) or PAM (Dayhoff *et al*., 1978). In this case, each amino acid was encoded by numerical value representing its distance from Alanine. Both types of inputs gave considerably poorer results than using the ten principal factors as inputs.

We found that the simple linear kernel performed best in our data set, compared to the polynomial and radial basis kernel functions. This is not surprising since the VC dimension is lower with a linear kernel (Vapnik, 1995) and hence generalization performance with limited training data is likely to be better.

Often the tuning parameters for SVM learning are chosen somewhat arbitrarily. We used a leave-one-out cross-validation imbedded in the training set to select the tuning parameters optimally. Leave-one-out cross-validation was employed to optimize the two tuning parameters: (i) the relative misclassification costs for the two classes and (ii) the trade-off between the training error and class separation. This cross-validation was performed entirely within the training set.

TCR ligands are not always high affinity MHC binders, and only a fraction of the potential MHC-binding peptides is a T-cell epitope for a specified TCR. Approaches to identify T-cell epitopes based on the prediction of which peptides would be good binders for specific MHC molecules are not accurate, since a functional T-cell response requires adequate MHC–peptide binding as well as proper interaction of the MHC–peptide ligand with a specific TCR. The comparisons clearly show our SVM approach to predict T-cell epitopes is superior to the publicly available methods such as SVMHC and SYFPEITHI (Table 2).

Previously, we reported a novel approach using biometric score matrix combined with combinatorial peptide libraries to predict T-cell epitope candidates (Zhao *et al*., 2001). That approach is based on a simple linear model under the assumption of independent contribution of side chains of amino acid within the peptide whereas the assumption of largely independent contributions of individual amino acids to stimulation seems to be a reasonably good approximation. Interactions of adjacent amino acids also exist and their effect may not always be predicted on the basis of individual substitutions (Leggatt *et al*., 1998; Hemmer *et al*., 1999, 2000). In addition, physicochemical characteristics of individual amino acids can change the HLA binding register. SVMs provide a framework for more sophisticated models that can take into account the interactions among the numerous factors that may influence T-cell recognition, and thereby accelerate the process of finding T-cell epitopes. Comparison of ROCs between SVM approach and score matrix based approach clearly indicates the SVM model greatly improves the prediction accuracy (Fig. 2 and Table 2).

Antigenic synthetic decapeptide Melan-$A_{26–35}$ (EAAGIGILTV) was predicted to be a T-cell epitope by our SVM model. It was strongly stimulatory, being among the peptides with highest percentage of specific lysis. We did *in silico* single amino acid substitution at all positions and used the SVM model to predict the activity of the mutated peptides. For A*0201, position 2 and 9/10 are considered to be the putative MHC anchors (Rammensee *et al*., 1995). Substitution of Thr in position 9 with hydrophobic residues Phe, Leu and Ile yielded the highest SVM scores while substitution of Val increased the SVM score slightly. At position 10, hydrophobic residues Val, Leu, and Ile were the only ones to keep the SVM unchanged; other substitutions would lead to either a reduced positive SVM score or even a negative SVM score. This is somewhat consistent with the A*0201 binding motif. When we examined the score matrix generated from combinatorial peptide library data, the above three residues did not have the highest scores at position 9; the matrix scores of Leu and Ile were well below the average. At position 2, Cys, Arg, Glu, Met, and Thr yielded higher SVM scores while Leu kept almost the same score as Ala in the template. Some of the above residues do not appear in the MHC anchor motif at position 2. Of interest, Met yielded a higher SVM score than Thr. Previous report had showed that replacing a Thr with a Met at the second position of gp100 epitope g209-217 (ITDQVPFSV) altered the binding affinity of the peptide to the HLA-A2 molecule and led to an increased recognition of the MHC–peptide complex by the TCR (Parkhurst *et al*., 1996). On the other hand, substitution with polar residues Ser, Thr, and Asn at position 2 would yield negative SVM scores. Residues at position 4–8 were suggested to be primarily involved in TCR recognition (Parkhurst *et al*., 1996). Gly, the simplest amino acid with no side chain, was the only amino acid to be allowed at position 6 in order to keep the peptide to be predicted positive, while at the same position the nonpolar residue Proline was the least favored one and yielded the lowest negative SVM score. Hydrophobic residues were favored at position 5 (Ile, Phe, Ala, Val) and 7 (Ala, Ile, Leu, Val), while replacing Leu with hydrophobic residues Phe or Ile at position 8 would lead to 1.5- to 2-fold increase of SVM scores. At position 4, Arg, Ser, and Thr doubled the SVM scores compared to Ala in the template.

Finally, in order to help interpret the SVM predictions for the single residue substitutions of the synthetic decapeptide (EAAGIGILTV), we calculated the Pearson correlation coefficients for each of the 494 physical properties listed in the public database (http://www.genome.ad.jp/dbget/aaindex.html) against the SVM scores of 20 mutants in each position. Several physical properties were highly correlated ($|r| > 0.7$) with positions 4, 5, 6, 7, 9, and 10. For example, we found that position 9 was highly correlated with van der Waals parameter R0 and position 7 was correlated with partial specific volume. Positions 5 and 7 were all correlated with normalized frequency of beta-sheet while position 6 was negatively correlated with normalized relative frequency of helix end.

Our results suggest that SVMs can be effectively used for predicting T-cell epitopes. Using physical property factors to encode the candidate peptides enables SVM classifiers to achieve good performance with modest amounts of synthesized peptide training data. This makes for an efficient process of prediction and synthesis of additional peptides because positive peptides are most informative. The SVM predictor can be used to provide information about the nature of the tri-molecular complex of peptide, MHC molecule and TCR. Further investigations of the use of SVM for T-cell epitope prediction are warranted as a potentially efficient and powerful method for defining candidate autoantigens, finding the antigenic targets and molecular mimics in complex infectious organisms, and developing vaccines for infectious diseases and cancers.

REFERENCES

Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) *Classification and Regression Tree*. Chapman & Hall /CRC Press, New York.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. and Haussler,D. (2000) Knowledge-based

analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Brusic,V., Rudy,G., Honeyman,M., Hammer,J. and Harrison,L. (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*, **14**, 121–130.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, **5**, 345–352.

Ding,C.H.Q. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Dönnes,P. and Elofsson,A. (2002) Prediction of MHC I binding peptides, using SVMHC. *BMC Bioinformatics*, **3**, 1–8.

Engelhard,V.H. (1994) Structure of peptides associated with class I and class II MHC molecules. *Annu. Rev. Immunol.*, **12**, 181–207.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Huassler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Gulukota,K., Sidney,J., Sette,A. and DeLisi,C. (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.*, **267**, 1258–1267.

Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Hammer,J., Valsasnini,P., Tolba,K., Bolin,D., Higelin,J., Takacs,B. and Sinigaglia,F. (1993) Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell*, **74**, 197–203.

Hammer,J. (1995) New methods to predict MHC-binding sequences within protein antigens. *Curr. Opin. Immunol.*, **7**, 263–269.

Hemmer,B., Gran,B., Zhao,Y., Marques,A., Pascal,J., Tzou,A., Kondo,T., Cortese,I., Bielekova,B., Straus,S.E. *et al*. (1999) Identification of candidate T-cell epitopes and molecular mimics in chronic Lyme disease. *Nat. Med.*, **5**, 1375.

Hemmer,B., Pinilla,C., Gran,B., Vergelli,M., Ling,N., Conlon,P., McFarland,H.F., Houghten,R. and Martin,R. (2000) Contribution of individual amino acids within MHC molecule or antigenic peptide to TCR ligand potency. *J. Immunol.*, **164**, 861.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Honeyman,M.C., Brusic,V., Stone,N.L. and Harrison,L.C. (1998) Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.*, **16**, 966–969.

Hua,S. and Su,Z. (2001) A novel method of protein secondary structure prediction with segment overlap measure, support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.

Joachim,T. (1999) *Making Large Scale SVM Learning Practical. Advances in Kernel Methods-Support vector learning*. MIT Press, Cambridge.

Kidera,A., Konishi,Y., Oka,M., Ooi,T. and Scheraga,H.A. (1985) Statistical analysis of the physical properties of the 20 naturally occuring amino acids. *J. Protein Chem.*, **4**, 23–55.

Leggatt,G.R., Hosmalin,A., Pendleton,C.D., Kumar,A., Hoffman,S. and Berzofsky,J.A. (1998) The importance of pairwise interactions between peptide residues in the delineation of TCR specificity. *J. Immunol.*, **161**, 4728–4735.

Madden,D.R. (1995) The three-dimensional structure of peptide–MHC complexes. *Annu. Rev. Immunol.*, **13**, 587–622.

Milik,M., Sauer,D., Brunmark,A.P., Yuan,L., Vitiello,A., Jackson,M.R., Peterson,P.A., Skolnick,J. and Glass,C.A. (1998) Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nat. Biotechnol.*, **16**, 753–756.

Parker,K.C., Bednarek,M.A. and Coligan,J.E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chains. *J. Immunol.*, **152**, 163–175.

Parkhurst,M.R., Salgaller,M.L., Southwood,S., Robbins,P.F., Sette,A., Rosenberg,S. and Kawakami,Y. (1996) 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.

Pinilla,C., Appel,J.R. and Houghten,R.A. (1994) Investigation of antigen-antibody interactions using a soluble, non-support-bound synthetic decapeptide library composed of four trillion sequences. *Biochem. J.*, **301**, 847–853.

Rammensee,H.G., Friede,T. and Stevanoviic,S. (1995) MHC ligands and peptide motifs, first listing. *Immunogenetics*, **41**, 178.

Rammensee,H.-G., Bachman,J., Philipp,N., Emmerich,N., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: a database for MHC ligands and peptide motifs. *Immnogenetics*, **50**, 213–219.

Rothbard,J.B. and Gefter,M.L. (1991) Interactions between immunogenetic peptides and MHC proteins. *Annu. Rev. Immunol.*, **9**, 527.

Rubio-Godoy,V., Dutoit,V., Zhao,Y., Simon,R., Guillaume,P., Houghten,R.A., Romero,P, Cerottini,J.C., Pinilla,C. and Valmori,D. (2002) Positional scanning-synthetic peptide library-based analysis of self- and pathogen-derived peptide cross-reactivity with tumor-reactive Melan-A-specific CTL. *J. Immunol.*, **169**, 5696–5707.

Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Sette,A., Sidney,J., del Guercio,M.F., Southwood,S., Ruppert,J., Dahlberg,C., Grey,H.M. and Kubo,R.T. (1994) Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Mol. Immunol.*, **31**, 813.

Sette,A., Buus,S., Appella,E., Smith,J.A., Chesnut,R., Miles,C., Colon,S.M. and Grey,H.M. (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl Acad. Sci. USA*, **86**, 3296.

Southwood,S., Sidney,J., Kondo,A., Guercio,M., Appella,E., Hoffman,S., Kubo,R.T., Chesnut,R.W., Grey,H.M. and Sette,A. (1998) Several common HLA-DR types share largely overlapping peptide binding repertoires. *J. Immunol.*, **160**, 3363–3373.

Sturniolo,T., Bono,E., Ding,J., Raddrizzani,L., Tuereci,O., Sahin,U., Braxenthaler,M., Gallazzi,F., Protti,M.P., Sinigaglia,F. and Hammer,J. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.*, **17**, 555.

Swets,J. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Valmori,D., Gervois,N., Rimoldi,D., Fonteneau,J.F., Bonelo,A., Lienard,D., Rivoltni,L., Jotereau,F., Cerottini,J.C. and Romero,P. (1998) Diversity of the fine specificity displayed by HLA-A*0201-restricted CTL specific for the immunodominant Melan-A/MART-1 antigenic peptide. *J. Immunol.*, **161**, 6956–6962.

Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer, Berlin.

Zavalijevski,N., Stevens,F.J. and Reifman,J. (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.

Zhao,Y., Gran,B., Pinilla,C., Markovic-Plese,S., Hemmer,B., Tzou,A., Whitney,L.W., Biddison,W.E., Martin,R. and Simon,R. (2001) Combinatorial peptide libraries and bio-metric 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.

Zien,A., Ratsch,G., Mika,S., Scholkopg,B., Lengauer,T. and Muller,K.R. (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 815–824.