BIOINFORMATICS APPLICATIONS NOTE

Vol. 20 no. 3 2004, pages 421–423
DOI: 10.1093/bioinformatics/btg424

SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence

Manoj Bhasin and G. P. S. Raghava*

Institute of Microbial Technology, Sector 39A, Chandigarh, India

Received on April 21, 2003; revised on June 18, 2003; accepted on August 8, 2003
Advance Access Publication January 22, 2004

ABSTRACT

Summary: Prediction of peptides binding with MHC class II allele HLA-DRB1*0401 can effectively reduce the number of experiments required for identifying helper T cell epitopes. This paper describes support vector machine (SVM) based method developed for identifying HLA-DRB1*0401 binding peptides in an antigenic sequence. SVM was trained and tested on large and clean data set consisting of 567 binders and equal number of non-binders. The accuracy of the method was 86% when evaluated through 5-fold cross-validation technique.

Available: A web server HLA-DR4Pred based on above approach is available at http://www.imtech.res.in/raghava/hladr4pred/ and http://bioinformatics.uams.edu/mirror/ladr4pred/ (Mirror Site).

Contact: raghava@imtech.res.in

Supplementary information: http://www.imtech.res.in/raghava/hladr4pred/info.html

In the past, few HLA-DR proteins were found to be associated with autoimmune diseases, e.g. HLA-DRB1*0401 with rheumatoid arthritis. It is important for the treatment of autoimmune diseases to determine which peptides bind to MHC class II molecules (HLA-DR) that will help in treatment of these diseases. The experimental methods for identification of these peptides are both time-consuming and cost-intensive. Computational methods thus, provide a cost effective way to identify these peptides. However, it is difficult to predict the peptides binding with HLA-DR molecules as compared to MHC class I molecules due to (i) variable length of binding peptides, (ii) undetermined core for each binding peptides and (iii) range of amino acids occupying anchor position for each MHC allele (Brusic *et al*., 1998).

Over the years, a number of methods have been developed for the prediction of HLA-DR4 binding peptides from an antigenic sequence, beginning with, early motif based methods (Chicz *et al*., 1993; Sette *et al*., 1993; Hammer *et al*., 1993, Max *et al*., 1993; Rammensee *et al*., 1995); to different scoring matrices based methods (Marshal *et al*., 1995; Southwood *et al*., 1998; Sturniolo *et al*., 1999; Reche *et al*., 2002). The artificial neural network has also been applied for the prediction of HLA-DRB1*0401 binding peptides (Brusic *et al*., 1998; Honeyman *et al*., 1998). Recently, few complex tools for identifying the HLA-DRB1*0401 binding peptides have also been designed, i.e. an iterative algorithm to optimize MHC class II binding matrix based stepwise discriminant analysis (Mallios, 1999). To improve the prediction accuracy, here we have developed a support vector machine (SVM) based method for predicting HLA-DRB1*0401 binding peptides.

HLA-DRB1*0401 binding peptides of nine or more than nine amino acids were obtained from MHCBN database (Bhasin *et al*., 2003). The binding cores (nine amino acids) from these peptides were obtained by using matrix optimization techniques (MOTs) package without considering MHC binding motifs (Singh and Raghava, unpublished data). HLA-DRB1*0401 non-binders of nine or more than nine amino acids were also obtained from the same database. Non-binders were chopped into overlapping peptides of nine amino acids. All the duplicate peptides and peptides with unnatural amino acids were removed from the binders and non-binders data set. The final data set consisted of 567 unique MHC binders and equal number of unique non-binders (randomly chosen from non-binders). The final ratio of MHC binders and non-binders was kept 1 : 1 so that the performance of the method can be evaluated by considering the single parameter accuracy at a cutoff score where sensitivity and specificity are nearly same.

SVM based method was developed using package SVM_LIGHT. (Joachims, 1999; Cristianini and Shawe-Taylor, 2000). In this study, we represented amino acids by 20 dimension vector, so input was a vector of dimension $9 \times 20 = 180$ for binders/non-binders of length nine. Performance of SVM is kernel-dependent, so we tried all types of kernels i.e. RBF, Polynomial, Sigmoid and LINEAR, and identified the kernel which gave best performance for our study. It was observed that RBF kernel is the best in classifying the data of

*To whom correspondence should be addressed.

Table 1. The performance of different MHC class II prediction algorithms on our data set

Category	Algorithm	Predictive measures Sensitivity[a]	Specificity[a]	PPV[a]	NPV[a]	Accuracy[a]
Motifs	Rammensee *et al*., 1995	55.38	53.4	54.33	54.5	54.4
	Sette *et al*., 1993	45.8	71.7	61.9	57.0	58.8
	Hammer *et al*., 1993	57.6	44.0	50.7	51.0	50.8
	Max *et al*., 1993	30.1	76.9	56.3	52.2	53.5
	Chicz *et al*., 1993	39.6	70.7	57.5	53.9	55.2
Matrices	Marshal *et al*., 1995	38.4	73.9	59.5	54.5	56.7
	Struniolo *et al*., 1999	5.9	59.3	51.7	58.2	55.0
	Brusic *et al*., 1998	50.6	57.5	54.3	53.8	54.0
	Southwood *et al*., 1998	31.0	88.1	72.4	56.1	59.6
	Borras-Cuesta *et al*., 2000	59.4	58.7	59.0	59.1	59.0
	Reche *et al*., 2002[b]	52.8	65.3	60.3	58.1	59.0
ANN	Bhasin and Raghava	80.2	77.4	77.9	79.8	78.8
SVM	Bhasin and Raghava	87.1	85.0	85.3	86.8	86.1

[a] Sensitivity is the percent of correctly predicted binders; Specificity is the percent of correctly predicted non-binders; PPV is Positive Probability Value (the probability that a predicted binder will be a binder); NPV is Negative Probability Value (the probability that a predicted non-binder will be a non-binder); accuracy is the percent of correct predictions (both binders and non-binders).

[b] In case of Reche *et al*. (2002), the binding threshold 5 is used instead of 22, as suggested by the authors of this paper. The threshold 5 is used to bring the sensitivity and specificity of prediction nearly equal.

DRB1*0401 binders and non-binders. In this study, we used the cutoff value where sensitivity and specificity were nearly equal for evaluating and developing SVM based method. The regulatory parameters *c* and *g* of RBF kernel were optimized to 5 and 0.1, respectively.

We also developed ANN based method using publicly available free simulation package SNNS4.2 (Zell and Mamier, 1997) in order to test performance of ANN based method on our data set. The training was carried out by using the standard feed-forward back propagation network with single hidden layer. The ultimate value of learning rate, learning cycles and hidden nodes were determined by monitoring the error on the training set. The values of hidden nodes, learning rate and cycles were optimized to 1, 0.01 and 300, respectively. Both SVM and ANN were provided with binary inputs (supplementary material).

In this study, 5-fold cross-validation was used to develop and assess the performance of both SVM and ANN based methods. The data set was randomly divided into the five sets, each consisting of equal number of binders and non-binders. The methods were trained on four sets and tested on remaining single set; this process was repeated five times so that each set was used as test set once. The SVM based method achieved an accuracy of ∼86% at a cutoff score where the sensitivity (∼85%) and specificity (∼87%) were nearly equal. The accuracy of the ANN based method developed in the present study is 78%. We also tested the performance of previously published methods on our data set. The performance of existing algorithms and our methods is shown in Table 1. One of the reasons of poor performance of existing methods is that these methods were designed from limited amount of data and in most of the cases non-binders were not used. Another inference obtained from the results is that SVM is superior to ANN in classifying data of MHC binders and non-binders.

In order to evaluate performance of our methods rigorously, we evaluated them on the data sets used in previous studies. All the binders used by previous methods were obtained from the literature and predicted by our methods on default threshold. As shown in Table S1 (supplementary material), our methods correctly predict all binders as binders for number of data sets. This demonstrates that our methods are not only suitable for data sets used in this study but also perform equally well on previously used data sets.

All the methods reported in literature as well as our methods were further evaluated on a blind data set created by Singh and Raghava (unpublished data; http://www.imtech.res.in/raghava/mhcbench/). This blind data set consisted 1017 peptides [694 binders and 323 non-binders, all unique peptides (duplicate peptides) and peptides having no-natural amino acids were removed from data set] of length nine that have been experimentally verified as binders or non-binders to HLA-DRB1*0401. These peptides were obtained from MHCPEP, MHCBN and the published literature (see Supplementary material). The performance of all methods on this data set is shown in the Table S2 (Supplementary material). The table illustrates that SVM based method developed in this study, outperformed all previously developed methods. The accuracy of ANN based method developed in this is study drastically drops to 42%.

The cross-validation by dividing data set into training and testing subset is the most common way to evaluate the performance of method. Infact, this is not true blind test for newly developed methods because same data is ultimately used to develop the methods. As shown in Tables S1 and S2, the performance of ANN dropped drastically when evaluated on blind/independent data set (data set not used either for testing or training). This indicates that commonly used cross-validation technique is not true validation. The developers should evaluate the performance of their method on data set, which is neither used, for training nor for testing. It was also observed that the performance of linear methods (statistical methods) is as good as non-linear methods (e.g. ANN). The SVM based method developed in this study performs well on all data sets.

A web server HLADR4Pred has been developed for predicting HLA-DRB1*0401 binding peptide. The server allows user to feed sequence in any standard sequence format (e.g. PIR, FASTA, EMBL). The server provides options of varying cutoff score to vary the stringency of prediction. These observations, on DRB1*0401 is universal so, it can be extended to other alleles for which the sufficient amount of data is available. In conclusion, these methods would find application in cellular immunology, transplantation, vaccine design, immunodiagnostics, immunotherapeutics and molecular understanding of autoimmune susceptibility.

ACKNOWLEDGEMENTS

We are thankful to Dr Vladimir Brusic, IIR, Singapore for providing their data for evaluation. We appreciate the anonymous referees for valuable suggestions. The authors are thankful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial assistance.

REFERENCES

Bhasin,M., Singh,H. and Raghava,G.P.S. (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, **19**, 666–667.

Borras-Cuesta,F., Golvano,J., Garcý'a-Granero,M., Sarobe,P., Riezu-Boj,J.I., Huarte,E. and Lasarte,J. (2000) Specific and general HLA-DR binding motifs: comparison of algorithms. *Hum. Immunol.*, **61**, 266–278.

Brusic,V., Rudy,G., Honeyman,G., 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.

Chicz,R.M., Urban,R.G., Gorga,J.C., Vignali,D.A., Lane,W.S. and Strominger,J.L. (1993) Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J. Exp. Med.*, **178**, 27–47.

Cristianini,N. and Shawe-Taylor,J. (2000) *Support Vector Machines and Other Kernel-Based learning methods*. Cambridge University Press, Cambriddge England The Edinburg Building, Cambridge, CB2 2RU, UK.

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.

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.

Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. (eds) *Advances in Kernel Methods—Support Vector Learning*. MIIT Press, Cambridge, MA, London, England.

Mallios,R.R. (1999) Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics*, **15**, 432–439.

Marshal,K.W., Wilson,K.J., Liang,J., Woods,A., Zaller,D. and Rothbard,J.B. (1995) Prediction of peptide affinity to HLA-DRB1*0401. *J. Immunol.*, **154**, 5927–5933.

Max,H., Halder,T., Kropshofer,H., Kalbus,M., Muller,C.A. and Kalbacher,H. (1993) Characterization of peptides bound to extracellular and intracellular HLA-DR1 molecules. *Hum Immunol.*, **38**, 193–200.

Rammensee,H.G., Friede,T. and Stevanovic,S. (1995) MHC ligands and peptide motifs: first listing. *Immunogenetics*, **41**, 178–228.

Reche,P.A., Glutting,J. and Reinherz,E.L. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.*, **63**, 701–709.

Sette,A., Sidney,J., Oseroff,C., del Guercio,M.F., Southwood,S., Arrhenius,T., Powell,M.F., Colon,S.M., Gaeta,F.C. and Grey,H.M. (1993) HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions. *J. Immunol.*, **151**, 3163–3170.

Southwood,S., Sidney,J., Kondo,A., del Guercio,M.F., 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., Tuercei,O., Sahin,U., Braxenthaler,M., Gallazzi,F., Protti,M.P., Sinigaglia,F. and Hammer,J. (1999) Generation of tissue-specific and promiscuous HLA ligand database using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.*, **17**, 555–561.

Zell,A. and Mamier,G. (1997) Stuttgart Neural Network Simulator version 4.2 University of Stuttgart.