Netw Model Anal Health Inform Bioinforma (2014) 3:56
DOI 10.1007/s13721-014-0056-z

ORIGINAL ARTICLE

HLAB27Pred: SVM-based precise method for predicting HLA-B*2705 binding peptides in antigenic sequences

Arun Gupta · Sharat Chandra · Tiratha Raj Singh

Received: 18 December 2013 / Revised: 3 April 2014 / Accepted: 5 April 2014 / Published online: 24 April 2014

Abstract Binding of Major histocompatibility complex (MHC) peptide is a prerequisite for T cell activation in the immune system. MHC-binding peptides have shown promising results for immunodiagnostics and immunotherapeutic purposes. HLA-B*2705 is found to be associated with the development of variety of autoimmune diseases including Ankylosing spondylitis. Detecting MHC class I allele HLA-B*2705 binding peptides can reduce the number of peptides that need to be tested experimentally. This work describes the implementation of SVM algorithm, developed for the identification of HLA-B*2705 binding peptides in antigenic sequences. The specificity and sensitivity obtained during the development of this server are 85 and 86 %, respectively. Whereas average precision and average recall values were observed to be 85 and 86 %, respectively. Training on wide-scale data made this method more accurate and robust than all available other methods for the HLA-B*2705 allele and would prove its usability in the biomedical domain. A web server HLA-B27pred is available at http://www.nuccore.org/hlab27pred for academic and research purpose.

A. Gupta and S. Chandra are joint first authors and equal contributors.

A. Gupta
School of Bio-Sciences and Technology, VIT University,
Vellore, Tamil Nadu 632 014, India

S. Chandra
Molecular and Structural Biology Division, CSIR-Central Drug
Research Institute, Council of Scientific and Industrial Research,
Chattar Manzil Palace, PO Box 173, Lucknow 22600, India

T. R. Singh (✉)
Department of Biotechnology and Bioinformatics, Jaypee
University of Information Technology (JUIT), Waknaghat,
Solan, HP 173234, India
e-mail: tiratharaj@gmail.com

Keywords Ankylosing spondylitis · MHC-binding peptides · HLA-B*27 allele · Immunoinformatics · SVM

1 Introduction

The major role of immune system is to identify pathogen-infected cells and differentiate them from healthy cells. It has to be accomplished by major histocompatibility complex (MHC) class I and class II antigen processing. The presentation of peptides by MHC class I molecules and their recognition by T cell receptors are fundamental for cell-mediated immune response (Flower 2008). As T cell recognition is limited to the peptides being presented by MHC molecules, the basis for the anticipation of T cell epitopes is the prediction of peptides that can bind to MHC molecules (Flower 2003; Flower and Doytchinova 2002). Some HLA alleles occur at a much higher frequency in those suffering from certain diseases than in general population. The diseases which are associated with particular MHC alleles include autoimmune disorders, disorders of the complement system, certain viral diseases, some neurologic disorders, and several different kinds of allergies (Vyes and Tood 1996). The human MHC class I allele HLA-B27 is one of the best investigated MHC class I molecules, which is partly due to its strong association with the development of variety of autoimmune diseases, including Ankylosing spondylitis (AS) (Allen et al. 1999). AS is a human leukocyte antigen HLA-B27-associated disease that is characterized by chronic progressive inflammation in the axial joints and enthuses (the ligaments and tendon attachments in bone) (Kim et al. 2005).

An individual with HLA-B27 allele has a relative risk of 90 (means 90 times greater likelihood as compared to someone in the general population) of developing the autoimmune diseases such as pelvic and spinal arthritis and AS. A given disease may be quantified by determining the frequency of the HLA alleles expressed by individuals afflicted with the disease, then comparing these data with the frequency of the same alleles in the general population (Dale and Federman 1997). In addition HLA-B27 is known to present antigenic peptides derived from major infectious agents, such as Epstein-Barr virus, influenza virus, and human immunodeficiency virus (HIV) to cytotoxic T lymphocytes.

Among the HLA-B27 subtypes, HLA-B*2705 is the most common and exhibits a very clear association with AS (Khan et al. 2007; Acar et al. 2011; Diyarbakir et al. 2012). Study on Turkish patients of AS suggested that the frequency of HLA-B27 subtypes is not significantly different between patients and controls. Polymorphic studies have also been performed on HLA-B27 patients with juvenile and adult-onset AS in Southern China (Mou et al. 2010). A recent study suggested that the Chinese traditional herb 'lei gong teng' might be a potential drug for patients who are HLA-B27-positive (Zhao et al. 2011). For the treatment of such an autoimmune disease, it is important to determine the peptides that would bind to MHC class I molecules which will help in the treatment of these diseases. The experimented methods for identification of these peptides are both time consuming and cost-intensive, which have been widely used to select a small number of candidate epitopes for experimental verification (Purcell et al. 2007).

A number of methods have already been proposed for predicting MHC-binding peptides from an antigenic sequence. Available computational methods are usually based on different principles: (1) motif-based methods such as position-specific scoring matrix (PSSM) (Parker et al. 1991; Rammensee et al. 1999) and statistical approach based on Hidden Markov Model (HMM) (Mamitsuka 1998), (2) Machine-learning methods based on Artificial Neural Network (ANN) (Nielsen et al. 2003); Support Vector Machine (SVM) (Donnes and Kohlbacher 2006) and kernel-based methods (Salomon and Flower 2006), (3) Semi-supervised machine-learning methods (Murugen and Dai 2005; Hertz and Yanover 2006), (4) structure-based methods (Doytchinova and Flower 2001). A number of algorithms have earlier been developed over the years for the prediction of HLA-B*2705 binding peptides from an antigenic sequence. Some of these algorithms are: Propred1 (Singh and Raghava 2003), nHLAPred (Bhasin and Raghava 2007), SYFPEITHI (Rammensee et al. 1999), SVMHC (Donnes and Kohlbacher 2006), BIMAS (Parker et al. 1994) and RANKPEP (Reche et al. 2002). All of these algorithms are based on different prediction methods and are developed using limited number of binding and non-binding data sets ranging from 20 to few hundreds (not more than 300) in binding and non-binding sets each. The better performance of the algorithms is defined by the size and quality of the data used for its development or training. Compared to other existing prediction algorithms, we have used large amounts (969 peptides) of high quality data from a more accurate and suitable database resource. Immune epitope database and analysis resource (IEDB; www.iedb.org), which provides a catalog of experimentally characterized data, has been used as the data resource to extract nonamer binding peptides of HLA-B*2705 (Vita 2010). To increase the performance and also to overcome the limitations of existing algorithms, we propose a more precise SVM-based algorithm for the prediction of human-specific HLA-B*2705 binding peptides.

2 Materials and methods

2.1 Data sets

Our data set consists of 969 unique MHC-binding peptides extracted from IEDB (Vita 2010). Same number of non-binding peptides (all peptides other than binding peptides) was generated randomly from a random selection of human proteins obtained from Unitprot. The MHC-binding and non-binding peptides final ratio was kept at 1:1 to evaluate the performance of the method by considering the accuracy of kernel parameters. SVM-based prediction for binding and non-binding peptides was categorized into two classes as $+1$ and -1, respectively. The machine-learning methods like SVM and many others do not understand alphabets, but accept digits as inputs and also return results in digits again. Hence, we used binary sparse encoding to represent the peptides. Each amino acid is represented as a bit vector of 20 elements, where the combination of 1 and 0s represents a particular amino acid e.g., (Ala: 100000000 00000000000, Cys: 01000000000000000000, etc.), hence every nonameric peptide is represented by a binary vector of total length of 180 bits (9×20).

2.2 SVM predictor

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

$$f(x) = \operatorname{sign}\left[\sum_i y_i \propto_i k\left(x_{i,}x\right) + b,\right] \tag{1}$$

where the kernel function k measures the similarity of its two vector arguments. For a linear SVM, the inner product kernel function $f(x)$ is used. If $f(x)$ is positive, then the sample is predicted to be in class $+1$, otherwise in class -1. The summation is over the set of "support vectors" that defines 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 two-group classification, the SVM separates the classes with a surface that maximizes the margin between them. We have implemented SVM using SVM_light (Joachims 1999).

Optimal parameters for any classification task are not known from the beginning, it is necessary to test different parameters to find the optimal ones. This is best done by a systematic sampling of the parameter space by grid search. Here, several combinations of two parameters, c and γ, were tested, given the start, stop, and step size for each of the parameters. In general, the radial basis function (RBF) kernel is a reasonable first choice which nonlinearly maps samples into a higher dimensional space, unlike the linear kernel, and can handle the case when the relation between class labels and attributes is nonlinear. Four variables are defined and used for this purpose: true positives (TP)—the number of binders predicted as such, true negatives (TN)—the number of non-binders predicted as such, false positives (FP)—the number of predicted binders that actually are non-binders, and false negatives (FN)—the number of predicted non-binders that actually are binders. A perfect correlation of real and predicted values will result in MCC value of 1, random predictions would result in lesser values, close to 0, whereas negative correlation would result in values closer to -1. Hence, the value of MCC is another parameter that defines the quality of the model generated and used for training the SVM algorithm. Sensitivity defines the rate of prediction of true positives from the set of prediction that are truly accepted and falsely rejected whereas, specificity determines the rate of true negatives from the set of predictions that are truly rejected and falsely accepted. For evaluation of the performance of methods, we have used standard parameters.

$$\begin{aligned}\text{Sensitivity} &= \text{TP}/\text{TP}+\text{FN}\\ \text{Specificity} &= \text{TN}/\text{TN}+\text{FP}\\ \text{Accuracy} &= \text{TP}+\text{TN}/\text{TP}+\text{FP}+\text{TN}+\text{FN}\end{aligned}$$

$$\text{Mathew}'\text{s Correlation Coefficient (MCC)} = \frac{(\text{TP}\times\text{TN})-(\text{FN}\times\text{FP})}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \quad (2)$$

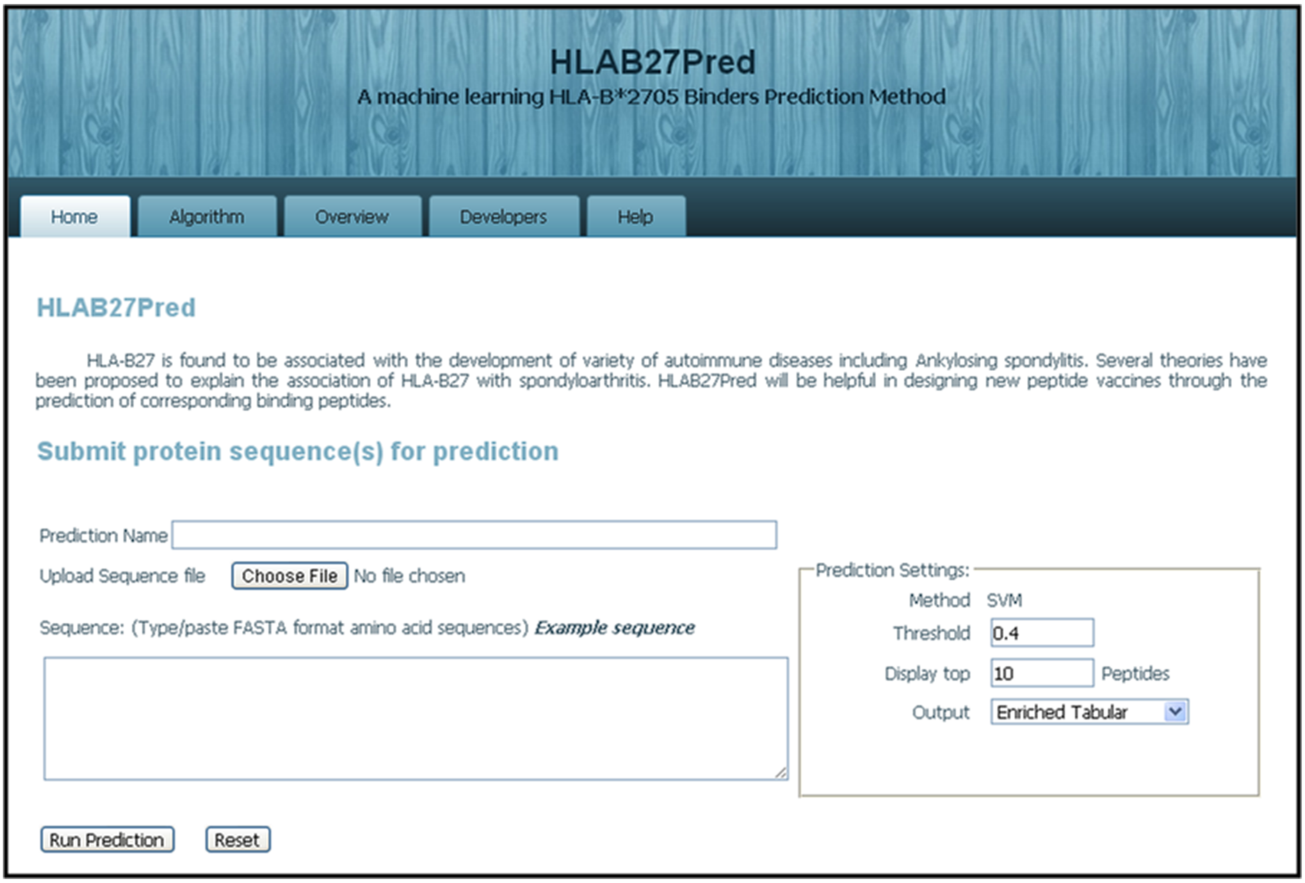
SVM-based algorithms are particularly appealing for binding and non-binding peptide prediction because of the ability of SVMs to build efficient prediction models when the dimensionality of the data is high and the number of observations is limited. A machine-learning method needs a sufficient amount of data for training, so we investigate 969 binders for training, extracted from IEDB27. We generated five different sets randomly from the original set of binding and non-binding peptides, for the purpose of training and testing the SVM algorithm for any given set of parameters, hence called five cross validation. Each set contains 20 % test and 80 % train data. In our problem for dealing with 969 peptides data, each set contains 387 peptides from test and remaining 1,551 peptides from train data sets, respectively.

3 Results and discussion

A web server HLAB27pred has been developed that allows prediction of HLAB*2705 (MHC class I alleles)-based nonamer epitopic binding peptides. This server implements SVM for the purpose of prediction. HLAB27pred is applicable for the all possible protein sequences. User can select SVM from the submission form (Fig. 1) to the prediction algorithm for their query protein sequences. They can either upload a file containing single or multiple sequences in FASTA format, or sequence can be pasted in the provided text area.

The results (Fig. 2) from this server are displayed in two different formats as: (i) Plain Tabular and (ii) Enriched Tabular. Both formats present information in different forms. The plain tabular format displays only the sequence submitted by user and the predictions made by this server in a tabular form. The enriched tabular format additionally displays a summary table containing several decisive details that might be of interest to the user (Fig. 2). For SVM method, user can set a threshold value, and the nonamer peptides above this threshold value will only be displayed. This would help the users to easily select the candidate-binding peptides. By default it is 0.4. For a prediction with higher specificity, one can choose a higher threshold values, whereas for a high sensitivity the threshold value should be low or close to the optimum. Users can also specify to display the top n predicted peptides from the submission form. In case if this number entered by users is higher than the number of binding peptides predicted and the number of binding peptides above the threshold score, users will be able to see the later number only.

The overall performance of the method was compared on the parameters sensitivity, specificity, accuracy and MCC. The final data set is divided into five sets and each of these parameters was evaluated to identify the best value of c and γ, at which we need almost equal as well as higher value of sensitivity and specificity. After whole process of optimization, the value of regulatory parameters c and γ of RBF kernel was decided. The specificity and sensitivity obtained during the development of this server are 85 and



HLAB27Pred

A machine learning HLA-B*2705 Binders Prediction Method

HLAB27Pred

HLA-B27 is found to be associated with the development of variety of autoimmune diseases including Ankylosing spondylitis. Several theories have been proposed to explain the association of HLA-B27 with spondyloarthritis. HLAB27Pred will be helpful in designing new peptide vaccines through the prediction of corresponding binding peptides.

Submit protein sequence(s) for prediction

Prediction Name

Upload Sequence file

Sequence: (Type/paste FASTA format amino acid sequences) *Example sequence*

Prediction Settings:

Method	SVM
Threshold	0.4
Display top	10 Peptides
Output	Enriched Tabular

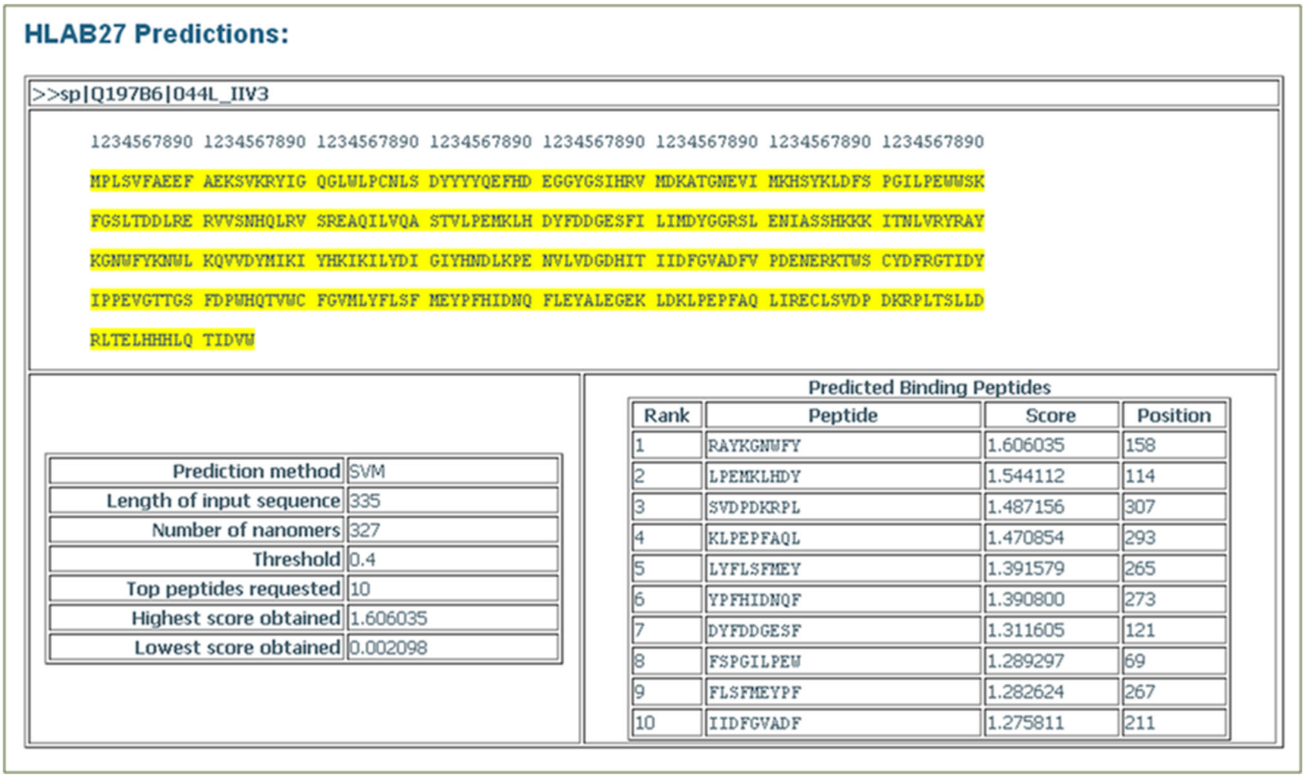
Fig. 1 An overview of the homepage of HLAB27Pred server

86 %, respectively. Whereas average precision and average recall values were observed to be 85 and 86 %, respectively.

Cross-validation techniques are being used characteristically for evaluating the performance of prediction methods. But the best way of evaluating the performance of a newly developed method is to test it on some independent data set that contains peptides other than those used in the training and testing of the algorithms. An independent blind data set has been derived from associated databases such as SYFPEITHI (Rammensee et al. 1999), MHCBN (Lata et al. 2009), MHCPEP (Brusic et al. 1996), etc., which contains 1,000 human epitopic nonamer peptides, not used in the training and testing of HLAb27pred algorithm. It has been found that method is working fine while evaluating the performance of this method on independent data set (data set is available upon request). The overall significant accuracy of 88 % of the developed method for independent data set was achieved.

We also performed a comparative study of binding peptides from proteins, Q5MH36, Q5MH37, Q5MH38, Q9MYH5, P00458, and O19193, binding to HLAB*2705. The binding nonamers and rank of these with different predictor's are shown in Table 1. All these are found at higher ranks in HLAb27pred than with other methods and prove the accuracy and rational wide usage of this method. This accuracy has been achieved due to training on larger data set than any other method available for this particular allele.

There are evidences where several studies have been performed to evaluate various immunological parameters for several disease and disorders (Chandra et al. 2010; Chandra and Singh 2012; Gupta et al. 2011) and it is believed that this study will also provide biological insights for the cure of autoimmune diseases. Main objective of the present study was to develop improvised algorithm for the prediction of HLA-B*2705 binding peptides from antigenic sequences of humans. Since this method is trained specifically on larger data set of human HLA-B*2705 binding peptides, it is more accurate and better in performance as all other previously developed methods have been trained on fewer amounts of data sets and also with multiple alleles together which can create ambiguousness in training. Our method is more specific and this specificity and training on large data set make it more robust and useful for associated autoimmune diseases like AS and



HLAB27 Predictions:

>>sp|Q197B6|044L_IIV3

```
1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
MPLSVFAEEF AEKSVKRYIG QGLWLPCNLS DYYYYQEFHD EGGYGSIHRV MDKATGNEVI MKHSYKLDFS PGILPEWWSK
FGSLTDDLRE RVVSNHQLRV SREAQILVQA STVLPEMKLH DYFDDGESFI LIMDYGGRSL ENIASSHKKK ITNLVRYRAY
KGNWFYKNWL KQVVDYMIKI YHKIKILYDI GIYHNDLKPE NVLVDGDHIT IIDFGVADFV PDENERKTWS CYDFRGTIDY
IPPEVGTTGS FDPWHQTVWC FGVMLYFLSF MEYPFHIDNQ FLEYALEGEK LDKLPEPFAQ LIRECLSVDP DKRPLTSLLD
RLTELHHHLQ TIDVW
```

Prediction method	SVM
Length of input sequence	335
Number of nanomers	327
Threshold	0.4
Top peptides requested	10
Highest score obtained	1.606035
Lowest score obtained	0.002098

Predicted Binding Peptides

Rank	Peptide	Score	Position
1	RAYKGNWFY	1.606035	158
2	LPEMKLHDY	1.544112	114
3	SVDPDKRPL	1.487156	307
4	KLPEPFAQL	1.470854	293
5	LYFLSFMEY	1.391579	265
6	YPFHIDNQF	1.390800	273
7	DYFDDGESF	1.311605	121
8	FSPGILPEW	1.289297	69
9	FLSFMEYPF	1.282624	267
10	IIDFGVADF	1.275811	211

Fig. 2 HLAb27pred screen shot for output of web page. Input sequence is highlighted in *yellow*, *lower left box* of the figure indicates various parameters and related outcomes including the threshold values selected by the user. *Lower right box* of the figure indicates predicted epitopic peptides, their respective scores and positions in the input sequence (color figure online)

Table 1 Comparative study of binding peptides from proteins, Q5MH36, Q5MH37, Q5MH38, Q9MYH5, P00458, and O19193, binding to HLAB*2705

Protein	#	HLAb27pred (SVM)	SVMHC	HLA_BIND	SYFPEITHI
Q5MH36, 37, 38	3	1, 3, 4	2, 6, 8	1, 13, 11	1, 8, 14
Q9MYH5	1	3	5	6	6
P00458	2	1, 2	115, 194	Not found	60, 170
O19193	2	6, 7	21, 141	9, 11	11, 80

others. Web server HLA-B27pred is available at http://www.nuccore.org/hlab27pred for academic and research use.

4 Conclusion

In summary, we have developed an accurate method based on SVM to predict the binding epitopic peptides. This type of analysis is important for in silico analysis of epitopes and the minor histocompatibility antigens (mi-HAgs), those have a role in autoimmune diseases. This method would mainly find applications in immunodiagnostics, immunotherapeutics and molecular understanding of autoimmune susceptibility. A user-friendly web interface as well as the good prediction performance will make HLAB27pred a very important tool for peptide-based vaccine design.

Acknowledgments Authors would like to thank Sachin Pundhir, Digvijay Singh Chauhan, and Shekhar Chandra for technical help and discussions.

References

Acar M, Cora T, Tunc R, Acar H (2011) HLA-B27 subtypes in Turkish patients with ankylosing spondylitis and healthy controls, Rheumatol Int (Epub ahead of print)

Allen R, Bowness P, McMichael A (1999) The role of HLA-B27 in spondyloarthritis. Immunogenetics 50:220–227

Bhasin M, Raghava GPS (2007) A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. J Biosci 32(1):31–42

Brusic V, Rudy G, Kyne AP, Harrison LC (1997) MHCPEP, a database of MHC-binding peptides: update 1996. Nucleic Acids Res 25:269–271

Chandra S, Singh TR (2012) Linear B-cell epitopes prediction for epitope vaccine design against meningococcal disease and their computational validations through physicochemical properties. Netw Model Anal Health Inf Bioinform 1:153–159

Chandra S, Singh D, Singh TR (2010) Prediction and characterization of T-cell epitopes for epitope vaccine design from outer membrane protein of Neisseria meningitidis serogroup B. Bioinformation 5(4):155–161

Dale DC, Federman DD (1997) SAM CD: a comprehensive knowledge base of internal medicine. Scientific American, New York

Diyarbakir E, Eyerci N, Melikoglu M, Topcu A, Pirim I (2012) HLA B27 subtype distribution among patients with ankylosing spondylitis in Eastern Turkey, Genet Test Mol Biomarkers 16(5):456–458. doi:10.1089/gtmb.2011.0183

Donnes P, Kohlbacher O (2006) SVMHC: a server for prediction of MHC-binding peptides. Nucleic Acids Res 34:W617–W622

Doytchinova IA, Flower DR (2001) Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. J Med Chem 44:3572–3581

Flower DR (2003) Towards in silico prediction of immunogenic epitopes. Trends Immunol 24:667–674

Flower DR (2008) Bioinformatics for vaccinology. Wiley-Blackwell, USA

Flower DR, Doytchinova IA (2002) Immunoinformatics and the prediction of immunogenicity, Appl Bioinformatics 1(A):167–176

Gupta A, Chaukiker D, Singh TR (2011) Comparative analysis of computational epitope predictions: proposed library of putative vaccine candidates for HIV. Bioinformation 5(9):386–389

Hertz T, Yanover C (2006) PepDist: a new framework for protein-peptide binding prediction based on learning peptide distance functions. BMC Bioinformatics 7:S3

Joachims T (1999) Making large-scale SVM learning practical. In: Scholkopf B, Burges C, Smole A (eds) Advances in Kernel methods-support vector learning. MIT Press, Cambridge, pp 169–184

Khan MA, Mathieu A, Sorrentino R, Akkoc N (2007) The pathogenetic role of HLA-B27 and its subtypes. Autoimmun Rev 6:183–189

Kim TH, Uhm WS, Inman R (2005) Pathogenesis of ankylosing spondylitis and reactive arthritis. Curr Opin Rheumatol 17:400–405

Lata S, Bhasin M, Raghava GPS (2009) MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. BMC Res Notes 2:61

Mamitsuka H (1998) Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. Proteins 33:460–474

Mou Y, Wu Z, Gu J, Liao Z, Lin Z, Wei Q, Huang J, Li Q (2010) HLA-B27 polymorphism in patients with juvenile and adult-onset ankylosing spondylitis in Southern China. Tissue Antigens 75:56–60

Murugen N, Dai Y (2005) Prediction of MHC class II binding peptides based on an iterative learning model. Immunome Res 1:6

Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci 12:1007–1017

Parker KC, Bednarek MA, Coligan JE (1991) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152:163–175

Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152:163

Purcell AW, McCluskey J, Rossjohn J (2007) More than one reason to rethink the use of peptides in vaccine design. Nat Rev Drug Discov 6:404–414

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

Reche PA, Glutting JP, Reinherz EL (2002) Prediction of MHC class I binding peptides using profile motifs. Hum Immunol 63:701–709

Salomon J, Flower DR (2006) Predicting class II MHC-peptide binding: a kernel based approach using similarity scores. BMC Bioinformatics 7:501

Singh H, Raghava GPS (2003) ProPred 1: prediction of promiscuous MHC class I binding sites. Bioinformatics 19:1009–1014

Vyes YC, Tood JA (1996) Genetic analysis of autoimmune disease. Cell 85:311–318

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2010) The immune epitope database 2.0. Nucleic Acids Res 38(Database issue):D854–D862

Zhao L, Liu CH, Yu D (2011) High-throughput screening of chemical libraries for modulators of gene promoter activity of HLA-B2705: environmental pathogenesis and therapeutics of ankylosing spondylitis. J Rheumatol 38:1061–1065