Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Prediction of HLA-A2 binding peptides using Bayesian network

Vadim Astakhov[1] and Artem Cherkasov[2]*

[1]Experimental Medicine Program, Department of Medicine, University of British Columbia, Vancouver, Canada; [2]Division of Infectious Diseases, Department of Medicine, Faculty of Medicine, University of British Columbia, 2733 Heather street, Vancouver, BC, Canada V5Z 3J5;

Artem Cherkasov* - Email: artc@interchange.ubc.ca; * Corresponding author

received October 5, 2005; revised October 10, 2005; accepted October 10, 2005; published online October 11, 2005

Abstract:

Prediction of peptides binding to HLA (human leukocyte antigen) finds application in peptide vaccine design. A number of statistical and structural models have been developed in recent years for HLA binding peptide prediction. However, a Bayesian Network (BNT) model is not available. In this study we describe a BNT model for HLA-A2 binding peptide prediction. It has been demonstrated that the BNT model allows up to 99% accurate identification of the HLA-A2 binding peptides and provides similar prediction accuracy compared to HMM (Hidden Markov Model) and ANN (Artificial Neural Network). At the same time, it has been shown that the BNT has that advantage that it allows more accurate performance for smaller sets of empirical data compared to the HMM and the ANN methods. When the size of the training set has been reduced to 40% from the original data, the identification of the HLA-A2 binding peptides by the BNT, ANN and HMM methods produced ARoc (area under receiver operating characteristic) values 0.88, 0.85, 0.85 respectively. The results of the work demonstrate certain advantages of using the Bayesian Networks in predicting the HLA binding peptides using smaller datasets.

Keywords: HLA; antigen presentation; peptides, Bayesian networks; machine learning

Background:

The recognition of foreign antigen peptides by the host HLA (human leukocyte antigen) molecules is critical for T mediated immune response. **[1,2]** There are two major classes of the HLA: class I molecules bind peptides originating from endogenous pathogenic proteins while the class II bind peptides derived from exogenous antigens. **[1,2,3]** The HLA class I and class II usually bind antigenic peptides consisting of 8-11 and 13-23 residues, respectively (some class II HLA bind > 40 residues long peptides). The binding is usually characterized by very high selectivity achieved through the interaction of the HLA with several critical (anchoring) residues of a peptide. Thus, despite the fact that biodegradation of antigenic proteins can theoretically produce a very large diversity of peptides, the actual number of them selectively bound to a specific HLA allele is limited. **[4,5]** This makes it not trivial and a very important goal to theoretically identify those specific fragments of protein sequences that are capable of selective interaction with specific HLA allele. It is believed that the ability of predicting HLA binding can not only provide a valuable insight into adaptive immunity but is also an essential step of *'in silico'* vaccine development. **[3]** In recent years, a number of theoretical methods for predicting HLA binding have been reported. **[1-4]** Conventionally, these tools could be divided into three major groups: profile- and matrix-based approaches **[6]**, methods utilizing the Artificial Neural Networks (ANN) **[6,7]** and those using the Hidden Markov Model (HMM). **[8]**

The profile- and matrix-operating models are based on the notion (derived by the analysis of crystal structures of peptide-HLA complexes) that the binding energy for individual residue within a peptide does not depend on the effects of the neighbouring amino acids. **[6, 9]** Thus, the corresponding methods operate by various additive scoring schemes to evaluate the likelihood of a given peptide to bind to a particular HLA allele. Such simplification allows fast processing of large amounts of data but sometimes accounts for the low accuracy compared to non-linear approximations such as Artificial Neutral Networks (ANN) or the Hidden Markov Model (HMM). **[6, 8]**

In the current study, we evaluate the previously unreported Bayesian Network (BNT) approach as another suitable method for modeling the HLA binding and compare the performance of the BNT with the results from the ANN and HMM solutions. The successful applications of the ANNs for prediction of the HLA peptides have been demonstrated in numerous studies. **[6-8]** Usually, the ANN based methods produces up to 80% accuracy in distinguishing the HLA binders from non-binders. At the same time, the ANN approximation is not suitable for processing peptides of varying length.

On another hand, this problem can be resolved by utilizing the Hidden Markov Model which allow up to 85% accuracy in discriminating the HLA binding peptides. **[8]** Being somewhat more accurate than the ANN-based methods, the HMM has the disadvantage of being more computationally demanding. Moreover, the Hidden Markov Model can only consider the mutual influence of adjacent residues but cannot account for possible distant interaction between non-neighbouring residues in a peptide. Therefore, the existing computational tools allow rather accurate theoretical prediction of the HLA binding peptides (for certain HLA

Bioinformation, an open access forum
© 2005 Biomedical Informatics Publishing Group

alleles) given that there is enough experimental data available to train the corresponding machine learning models. Here, we describe a BNT model using smaller empirical datasets compared to training requirements for ANN and HMM.

Methodology:

Dataset:

Binders: We derived a set of 244 HLA-A2 binding peptides from MHCPEP and SYFPEITHI databases. **[6, 8, 10]** Non-binders: A set of 464 non-binding peptides required for adequate model training have been randomly generated from a human albumin sequence. The size of the non-binders was chosen so as to keep the binder – non-binder ratio to 1:2. Over training: Peptides used for training were carefully curated to avoid over training by eliminating redundant peptides such that no two peptides shared more than 4 residues.

Software used in the analysis:

Several open source and commercial software products have been used in this study. The ANN (a fully connected 3-layer back-propagation configuration trained on the generalized delta rule) has been built and manipulated within the Stuttgart Neural Network Simulator (SNNS) package. The input and output layers consisted of 180 nodes and 1 node, respectively. The number of the hidden layer nodes has been tested in the range of 2 to 50. The Bayesian Network has been built with the WEKA machine-learning software and the Hidden Markov Model was created with the MATLAB (simulation and modelling software) HMM toolkit.

ANN (Artificial Neural Networks):

An application of ANN for prediction of class I HLA binding has been described by Brusic and co-authors. **[6]** The developed ANN-based method uses the machine learning algorithm to train the HLA binding patterns in peptide residues. The typical configuration of the ANN adopted for the HLA binding prediction represents a three layer Neural Network operating on the binary input. Within this approximation, each HLA binding peptide consisting of nine residues (typical for class I HLA) is represented as a string of 180 binary numbers (zeros and ones) serving as the input of the ANN. This corresponding 180-elements vector is formed by 9 blocks of 20 numbers where each block represents a consequent position on a peptide and every number in the block of 20 designates the presence or absence of specific the amino acid residue. The hidden nodes of the ANN play the role of free optimisation storing the inferred patterns emerging from the input data. Number of hidden nodes can usually be optimised during the ANN training and the output of the three layers ANN is constituent of a single node providing the binder/non-binder discrimination information.

HMM (Hidden Markov Model):

The HMM approach describes an abstract statistical system as a number of hypothetical states connected by the transition probabilities. Thus, the problem of formalization of the HLA binding is a 'natural' task for the HMM which treats a string of residues in a peptide as a Markov Chain terminated by its START and END Markov states. A peptide represented in HMM includes 20 Matching states (reflecting possible variations of amino acids) as well as 20 Deletion and 20 Insertion states altogether instructing the HMM algorithm to extract the binding patterns from the empirical HLA binding data. The HMM defines the probabilities included into the matching states on the basis of the experimental frequencies of particular residue in a given peptide position during training of empirical inputs (sets of peptides with experimentally pre-determined binding or non-binding character). The Insertion state of the Markov Chain represents a logical operation for introducing an additional residue into the construction of a pattern with uniform probabilities and the Deletion states are defined within the HMM without assigning any probabilistic properties. HMM utilizes the Baum-Welch **[8]** algorithm optimizing the transition probabilities from one state to another beginning from the Start state of the Markov Chain and then chooses the next state of the system depending on the transition probabilities of the consequent chain edges. This process repeats until the transition reaches the End state which leads to the generation of multiple patterns (sequences of states) each reflecting the probability of the studied peptide to be a HLA binder. More detailed description of the HMM method for predicting HLA binding peptides is described elsewhere. **[8]**

BNT (Bayesian Network):

The Bayesian Network method processes experimental information differently compared to conventional statistical approaches. Instead of using pre-defined analytical functions the BNT attempts to establish an optimal statistical model to fit experimental data. The Bayesian approach has found a broad application in those areas of data analysis where there is a need for extracting complex patterns from sizable amounts of information with significant levels of noise. The Bayesian method has been successfully employed for the SAGE data analysis, for modeling genetic regulatory interactions **[11]**, for solving some protein folding problems **[12]** and for text processing and diagnostics. **[13]** One of the basic Bayesian definitions is prior information P(H) where H is a model/hypothesis and P(H) is probability of a model to be true. In the context of predicting the HLA binding, as a prior information P(H) we can consider the assumption that any peptide can theoretically be a HLA binder. In other words, the initial probability P(H) for an arbitrary peptide to bind to a particular HLA allele is 50% (a chance probability which will change in a recurrent manner during the Bayesian optimisation). Another definition of the Bayesian analysis is a likelihood function P(D|H) reflecting the probability of obtaining the observed experimental data (D). This function is not pre-set prior the analysis but is estimated during the BNT optimisation. The third Bayesian category is the degree of plausibility P(H|D) (sometimes called posterior probability of initial hypothesis) which can be calculated using the Bayesian Theorem stated as follows in equation 1. Equation 1 is p(H|D) = p(D|H) p(H)/p(D), where p(D) is the normalization factor.

Bioinformation, an open access forum
© 2005 Biomedical Informatics Publishing Group

The Bayesian Network represents an application of the Bayesian theory which formalizes the joint distribution over a set of random variables $X = \{X_1,..., X_n\}$ as a product of conditional probabilities. An abstract Bayesian network can be defined by a graphical structure M combining a family F of conditional probability distributions F= {P (Xi | q)}, in turn depending on the vector of parameters q = {pa[Xi]}. The graphical structure M can be illustrated as set of nodes V and directed edges E which can connect any pair of nodes where the nodes V correspond to random variables and the edges indicate conditional dependence relations among them (Figure 1A). Here, we describe the use of Bayesian Network for the prediction of HLA binding peptides. Peptides consisting of nine residues can be described by 180 variables each reflecting a probability to have a defined residue type at a defined position of a HLA binding peptide. The relationships between these 180 variables can be optimized within the BNT methodology for peptides in the training set to yield the posterior probabilities p(H|D) according to equation 1. Figure 1A illustrates that the BNT can capture mutual influences among amino acids in a HLA binding peptide by representing it as a directed graph consisting of edges X_i. The BNT can operate on such graph on the basis of the observed frequencies of certain amino acids at defined positions of the peptides capable of binding to the HLA molecule. Accordingly, the joint probability for a given peptide to be a binder can be estimated by the BNT as given in equation 2. Equation 2 is P(X1, X2, . . . ,Xn) = P(X1 | pa[X1]) *P(X2 | pa[X2]) *.....*P(Xn | pa[Xn]) where, where P(Xi |pa[Xi]) = K_1* P(Xi | pa_1 [Xi]) + K_2* P(Xi | pa_2 [Xi]) + K_3* P(Xi | pa_3 [Xi])+… represents a sum of conditional probabilities and K_j is a weight coefficient. The entity P(Xi|pa[Xi]) in equation 2 corresponds to the conditional probability which represents the influence of variable Xi on a peptide binding ability. It should be noted that the vector pa[Xi]={pa_1[Xi], pa_2[Xi], ..} is represented in Figure 1A by the graph edges.

Results and Discussion:

Performance of BNT, ANN and HMM in original set:

A total of 708 peptides are separated into training and testing groups (in the proportion of 9:1) each containing both binding and non-binding peptides (the corresponding sets are given in Appendix 1). The very same training and testing sets are used to train and evaluate the ANN, HMM and BNT models for distinguishing the HLA-A2 interacting peptides. A constant cutoff values 1 and 0 to the HLA-A2 binding and non-binding peptides in the training set was assigned. It has been observed that both ANN and the HMM required less than 200 training cycles to achieve maximal predictive accuracy. It has also been established that by gradually changing the number of ANN hidden nodes from 2 to 50, the predictive ability of the network (the ANN learning rate was kept 0.2 with the 0.02 shift) is not significantly influenced. The processing of the data by BNT for each peptide in the training set yielded the resulting probability value $P(X_1, X_2, . . . , X_n)$. The corresponding parameters estimated by equation 2 can be found in Table 1 (see Additional file 1). Corresponding outputs from ANN and HMM are also given. Each peptide in Table 1 has been classified as the HLA-A2 binder if the corresponding BNT joint probability $P(X_1, X_2, . . . , X_n)$ exceeded 50%. The outputs from HMM and ANN have also been characterized by applying 50% cut-off. The predictive power of all three methods has been assessed by processing the testing set (67 peptides) through the pre-trained models. Subsequently, FP (false positives), FN (false negatives), TP (true positives) and TN (true negatives) were estimated. Then, sensitivity, specificity, percentage of correct predictions and Matthews Correlation Coefficients ((Mc) = $\{TP*TN-FP*FN\}/\{(TN+FN)(TN+FP)(TP+FN)(TP+FP)\}^{1/2}$) were also calculated (Table 1). Results show that ANN, HMM and BNT produced prediction accuracy for testing and training sets. However, BNT outperformed ANN by 8% for HLA-A2 peptide binding prediction. Results of the ARoc analysis for the three models are also presented in Table 1.

Bioinformation, an open access forum
© 2005 Biomedical Informatics Publishing Group

Prediction parameter	Set 1 (73 peptides)			Set 2 (635 peptides)		
	ANN	**HMM**	**BNT**	**ANN**	**HMM**	**BNT**
True Positives	17	21	24	191	193	194
True Negatives	42	44	45	386	390	392
False Positives	6	4	3	30	26	24
False Negatives	8	4	1	28	26	25
Matthew Coefficient	0.73	0.85	0.86	0.78	0.86	0.89
Specificity	0.91	0.92	0.94	0.93	0.94	0.94
Sensitivity	0.68	0.84	0.96	0.87	0.88	0.89
Correct predictions	0.89	0.93	0.95	0.91	0.93	0.95

ARoc performance

Training/Testing set separation	**ANN**	**HMM**	**BNT**
0.4 / 0.6	0.856	0.860	0.880
0.5 / 0.5	0.873	0.880	0.901
0.6 / 0.4	0.932	0.920	0.940
0.7/ 0.3	0.962	0.950	0.960
0.8 / 0.2	0.985	0.992	0.980
0.9 / 0.1	0.992	0.998	0.990

Table 1: Performance of different models in varying datasets is given

Performance of BNT, ANN and HMM in reduced set

The ARoc performances of the three models for varying proportions of training and testing set are also presented in Table 1 and Figure 1B. The ARoc for all the three models are very high for large proportions of training sets (80% and 90%). However, the ARoc for BNT is higher than ANN and HMM for low proportions of training sets (40%-60%). This suggests that BNT out performs ANN and HMM when lower proportions of training set are used and is therefore suitable for modeling when dataset size is limited (as low as 40%). The corresponding dependence between prediction accuracies and proportions of training to testing datasets is given in Figure 1C. Figure 1C suggests that the performance of all the methods is comparable when the size of the training sets is sufficiently large. However, when the training set is reduced the BNT provide more accurate predictions. It should be noted however, that the BNT is computationally intensive and may be less applicable for processing very large amounts of data.

Conclusion:

HLA binding peptide prediction finds application in vaccine design. However, the prediction of HLA binding peptides is not trivial. Here, we discussed the performance of ANN, HMM, BNT models for HLA-A2 binding peptide prediction. The prediction accuracy of ANN, HMM, BNT are similar when large training sets are used. Nonetheless, the BNT model performed better than ANN and HMM even when the training set is reduced to 40% of the original size.

Bioinformation, an open access forum
© 2005 Biomedical Informatics Publishing Group

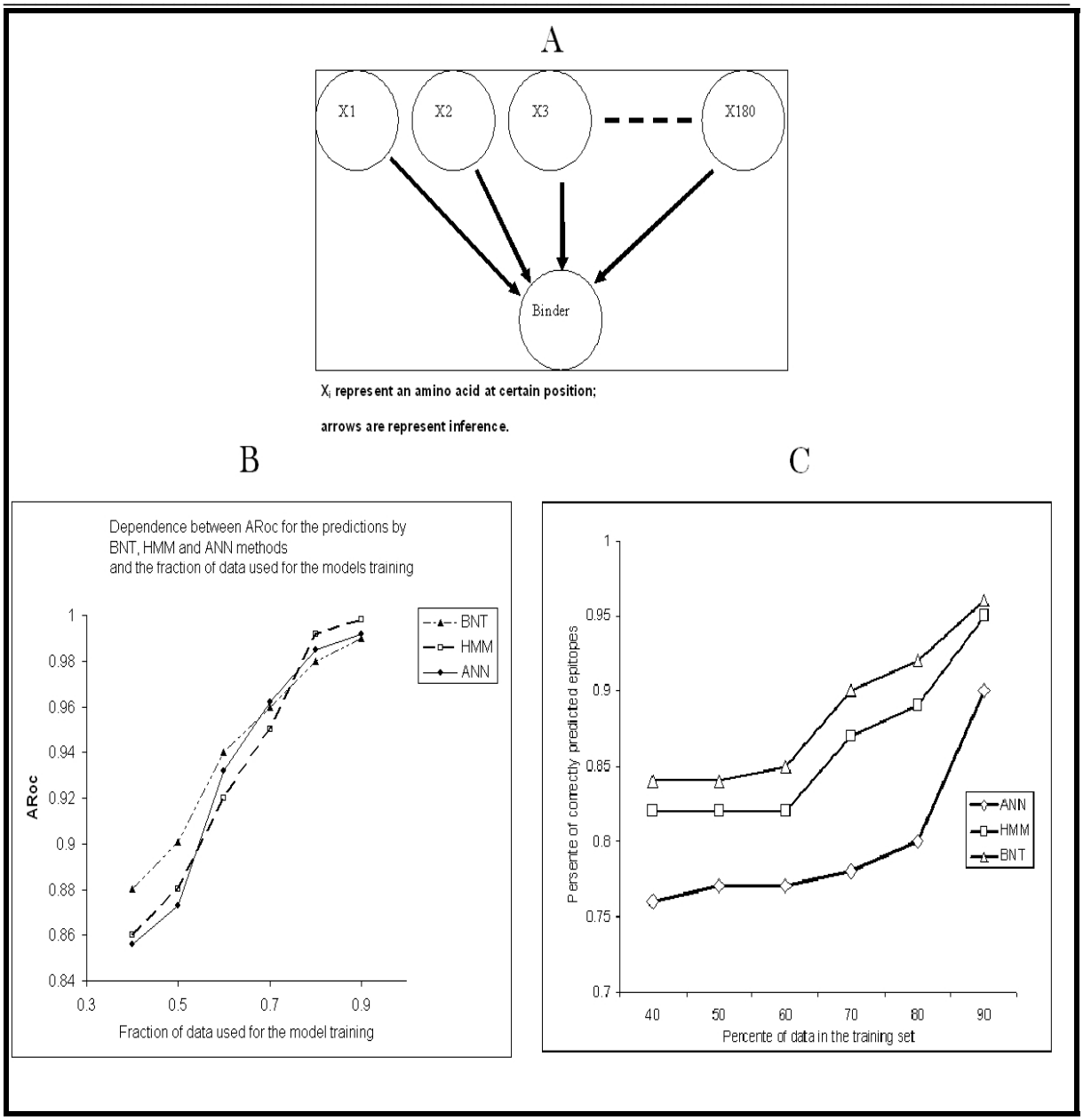

Figure 1: (A) Representation of a nine residue long peptide by the Bayesian Network. (B) Dependence between the estimated ARoc parameter by ANN, HMM and BNT with training and testing set. (C) Dependence between the prediction accuracy estimated by ANN, HMM and BNT with training set size.

References:

[1] V. E. Reyes, *et al.*, *Molecular Immunology*, 25*:* 867 *(*1988) [PMID: 3264884]

[2] J. B. Rothbard & W. R. Taylor, *EMBO Journal*, 7:93 (1988) [PMID: 2452085]

[3] A. Sette, *et al.*, *Proc Natl Acad Sci USA*, 86:3296 (1989) [PMID: 2717617]

[4] C. DeLisi & J. A. Berzofsky, *Proc Natl Acad Sci USA*, 82:7048 (1985) [PMID: 2413457]

[5] J. L. Spouge, *et al.*, *J Immunology*, 138:204 (1987) [PMID: 2431054]

[6] K. Yu, *et al.*, *Molecular Medicine*, 8:137 (2002) [PMID: 12142545]

[7] D. Hammerstrom, *IEEE Spectrum*, 26 (1993)

Bioinformation, an open access forum
© 2005 Biomedical Informatics Publishing Group

[8] H. Mamitsuka, *PROTEINS, Structure, Function, and Genetics*, 33:460 **(**1998) [PMID: 9849933]

[9] D. R. Madden, *et al.,, Cell*, 75:693 (1993) [PMID: 7694806]

[10] V. Brusic, *et al., Nucleic Acids Research*, 26:368 (1994) [PMID: 7937075]

[11] D. Husmeier, *Bioinformatics*, 19:2271 **(**2003) [PMID: 14630656]

[12] A. Raval, *et al.,, Bioinformatics*, 18:788 **(**2002) [PMID: 12075014]

[13] D. Nikovski, *IEEE Transactions on Knowledge and Data Engineering*, 12:509 (2000)

Edited by P. Kangueane

Citation: Astakhov & Cherkasov, Bioinformation 1(2): 58-63 (2005)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

© 2005 Biomedical Informatics Publishing Group