IJCIT

International Journal of Computational Intelligence Techniques
ISSN: 0976–0466 & E-ISSN: 0976–0474, Vol. 2, Issue 1, 2011, PP-22-25
Online available at http://www.bioinfo.in/contents.php?id=26

PROFILE HIDDEN MARKOV MODEL FOR PREDICTING T CELLS EPITOPES

MUTHU KUMAR M.[1] AND SENTHAMARAI KANNAN K.[2]
[1]Department of Mathematics, Noorul Islam Centre for Higher Education, Kanyakumari, Tamilnadu, India.
[2]Department of Statistics, Manonmaniam sundaranar University, Tirunelveli, Tamilnadu, India
*Corresponding author. E-mail: muthukar2003@yahoo.co.in

Received: Received: July 17, 2011; Accepted: August 06, 2011

ABSTRACT - Prediction methods for identifying binding peptides could minimize the number of peptides required to be synthesized and assayed, and thereby facilitate the identification of potential T-cell epitopes. We developed a bioinformatics method for the prediction of peptide binding to T-cell molecules. The major T-cell contributors are selected for the dataset preparation due to its availability and originality. We used a profile hidden Markov Model (HMM) for the prediction. Sensitivity (96%) and Specificity (~100%) are evaluated for the T cells epitope and nonepitopes from the test data set. The method promises 98 % accuracy and useful for vaccine development.
Keywords: Major Histocompatibility Complex, Epitopes, Hidden Markov Model, T-cells , ROC

1. Introduction

Cells are the fundamental units of life and are sometimes called the building blocks of life. The immune system is the second most complex body system in humans. The immune system is the body's defense against infectious organisms and other foreign agents. The first line of defense is innate immunity. It has rapid nonspecific responses, which allow recognition of conserved signature structures present in many microorganisms. The second line of defense is the adaptive immune response, tailored to an individual threat. An infected host mounts an immune response specific to an infectious agent; after the infection is resolved, memory cells persist that enable a more rapid and potent response if the infectious agent is encountered again [1,2,3].

The immune system is composed of many interdependent cell types, organs, and tissues that jointly protect the body from infections (bacterial, parasitic, fungal, or viral) and from the growth of tumor cells. The immunity is a memory system of organism cell, visualized by epitopes. An epitope is the part of a protein which is recognized by the immune system. They are recognized by specific T-cells, B-cells, and the antibody produced by B-cells. Antibodies generally recognize intact proteins. When these cells recognize and are activated by specific epitopes, they begin mounting an immune response. Most epitopes are derived from proteins that the immune system classifies as non-self, meaning the proteins are part of a foreign organism such as a virus or bacterium.

T-cell immune responses are driven by the recognition of peptide antigens (T-cell epitopes) that are bound to Major Histocompatibility Complex (MHC) molecules [4].

T-cell epitope immunogenicity is thus contingent on several events, including appropriate and effective processing of the peptide from its protein source, stable peptide binding to the MHC molecule, and recognition of the MHC-bound peptide by the T-cell receptor (see Figure 1). Of these three hallmarks, MHC-peptide binding is the most selective event that determines T-cell epitopes [5].

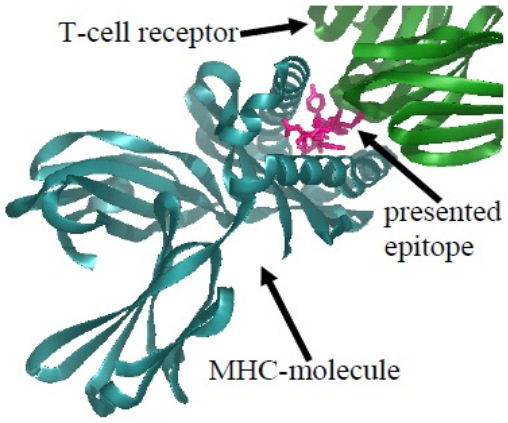

Fig. 1-T-cell Receptor binds to MHC molecule

One of the principal goals of informatics research in immunology is the development of algorithms to assist in the creation of new vaccines. Reliable epitope identification via computational means would lessen the burden required for laboratory analysis of viral, bacterial and parasitic gene products. An informatics driven approach would allow the immunologist to greatly reduce the experimental work, providing a valuable starting point for the exploration of potential binding sites.

Two main categories of specialized bioinformatics tools are available for prediction of MHC binding peptides. They are, methods based on identifying patterns in sequences of binding peptides, and those that employ three-dimensional (3D) structures to model peptide/MHC interactions. The first group includes procedures based on binding motifs, quantitative matrices, decision trees, artificial neural networks (ANNs), hidden Markov models (HMMs) and support vector machines (SVMs). HMM is considered as the most power full tool for T-cell epitope prediction interms of specificity and sensitivity [6,7,8]. In contrast, the second category corresponds to techniques with distinct theoretical lineage and includes the use of homology modeling, docking and 3D threading techniques. An unequal amount and variety of techniques have explored for the two categories in the published reports, far fewer for structure-based approach due to higher complexity in development and longer computational time. A stochastic model like HMM were used in this study for T-cell prediction from a set of amino acid sequences. In this paper we present a paired profile Hidden Markov Model (HMM) for predicting T-cell epitopes from its alleys. Our approach gives a comparative specificity, sensitivity and accuracy when compared with other similar methods.

2. Dataset preparation

The performance of method for few alleles was also evaluated on blind or independent datasets. The blind dataset was generated for each MHC allele. The binders for each allele were obtained from the published literature (Table 1). Equal number of non-binders for each allele were obtained either from MHCBN database (wherever available) or generated randomly from proteins of SWISS-PROT database. All the binders and non-binders which were used for testing and training of this method were removed from these blind datasets. We used a combination of dataset obtained from the data source in Table.1

3. Profile HMM

Hidden Markov Model (HMM) describes a probability distribution over a set of possible hidden states [9,10,11]. Profile HMMs are linear left-right models where the underlying directed graph is acyclic, with the exception of loops, hence supporting a partial order of the states. The profile HMM architecture consists of three classes of States: the Match state (M), the Insert state (I) and the Delete state (D); and two sets of parameters: transition probabilities, and emission probabilities. The match and insert states always emit a symbol, whereas the delete states are silent states without emission probabilities.

Profile analysis has long been a useful tool in finding and aligning distantly related sequences and in identifying known sequence domains in new sequences. Basically, a profile is a description of the consensus of a multiple sequence alignment. It uses a position-specific scoring system to capture information about the degree of conservation at various positions in the multiple alignments. Profile HMMs have several advantages over standard profiles. Profile HMMs have a formal probabilistic basis and have a consistent theory behind gap and insertion scores, in contrast to standard profile methods which use heuristic methods. HMMs apply a statistical method to estimate the true frequency of a residue at a given position in the alignment from its observed frequency while standard profiles use the observed frequency itself to assign the score for that residue.

3.1 Parameter Estimation

Let the transition probability of going from state k to state l a_{kl} be equal to:

$$\frac{\text{number of times go from state k to state I}}{\text{the number of times go from state k to any other state}}$$

So, to calculate the probability of transition from match state 1 to match state 2, we count the number of times we get a match (=6) in the second column, as well as the number of gaps (=1). (Note, using the initial alignment and our model, we only have insertions after the third match state.)

$$a_{M1M2} = \frac{6}{7}$$

Again, using the add-one rule, we correct our probabilities to be:

$$a_{M1M2} = \frac{6+1}{7+3} = \frac{7}{10}$$

The rest of the parameters are calculated analogously.

4. Results

4.1 Epitope region predictions

A sliding window of size 12 ± 3 amino acids is considered because mean length of the T-cell epitope is nine. A paired profile HMM was used in this study with positive and negative models. Viterbi algorithm is a common technique to find the most promising transition path in HMM with the probability [12]. Viterbi algorithm is a dynamic programming algorithm which reduced the complexity of computations drastically [13]. Viterbi probability of the fixed length sequence that represents the epitope region can be evaluated. Taking the difference of the Viterbi score generated by positive and negative HMM. If the value is above the threshold, then it is a true epitope, otherwise considered as the non epitope. The threshold is fixed as the average score of true epitopes. The results suggest that the feature selection may extract the most important information that contributes to the stimulatory activity of T-cell epitopes and non-epitopes. The architecture of the prediction system is shown in Fig(2).

4.2 Prediction system assessment

The mathematical definitions of sensitivity and specificity clarify the inherent problems of attempting to benchmark T-cell epitope prediction against irrelevant empirical results; so far as the design of peptide-based vaccines is

concerned, the most obvious of these results are from experiments that do not even simulate vaccination with peptides (e.g., where antibodies are never elicited by peptides in the first place). The designation of such results as either positive or negative is meaningless; at worst, it leads to erroneous appraisal of computational results that translates to miscalculation of both sensitivity and specificity. Consequently, methods for T-cell epitope prediction can be either under rooted or over rooted, thereby compromising efforts to assess their performance and address their limitations accordingly.

For classification type problems, a prediction can be either positive or negative [14,15,16]. These counts falls into four categories: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). These contents are used to calculate sensitivity (true positive rates), specificity (1- false positive rates) and total prediction accuracy for assessment of the prediction system are given. Various quantitative variables were employed to measure the effectiveness of the profile HMM model for predicting linear T-cell epitopes:

(i) TP, true positives - the number of correctly classified epitopes.

(ii) FP, false positives - the number of incorrectly classified non-epitopes.

(iii) TN, true negatives - the number of correctly classified non-epitopes.

(iv) FN, false negatives - the number of incorrectly classified epitopes.

Receiver Operator Characteristic (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance [16]. The ROC analysis is widely used to analyze the machine learning classifiers and generally useful as performance graphing method. ROC is a graph with true positive rates (sensitivity) on X-axis and false positive rate (1-specificity) on the Y-axis.

$$Sensitivity = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$Positive\ Predicitive\ Value = \frac{true positives}{true positives + false positives}$$

$$Parameter\ accuracy = \frac{true positives + true negative}{total no. of testing dataset}$$

$$Specificity = \frac{true\ negatives}{true\ negatives + false\ positives}$$

5. Discussions

For the external validation of this method, a set of 160 epitopes, and their source proteins were collected from AntiJen (http://www.jenner.ac.uk/AntiJen/). The epitopes were not been used to develop any of the models included in this model. To reduce the number of non-epitopes, only proteins consisting of less than 1000 amino acids were considered in the study. As the number of non-epitopes generated from one protein was significantly higher than the number of epitopes, only two parameters – sensitivity and positive predictive value (PPV) were used for the assessment of program performance. Normal Receiver Operator Characteristics (ROC) curve without fixing any cutoff (threshold) is shown in Fig(3).

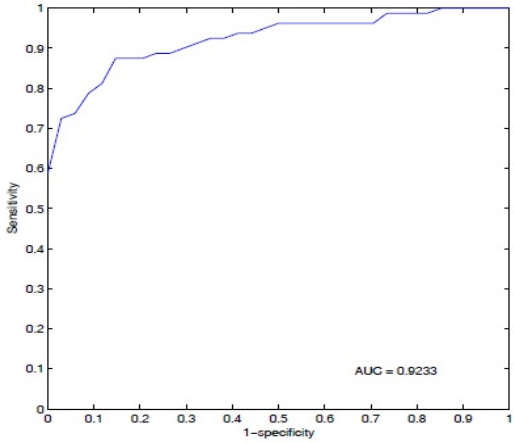

Fig. 3-ROC curve for T-cell epitope prediction

As the number of nonepitopes generated from each protein was significantly higher than the number of epitopes, only two parameters – sensitivity and positive predictive value (PPV) were used for comparison. Parameters accuracy and Specificity could be misleading. If 98% of the peptides in one source protein are non-epitopes, a model that simply predicts everything as non-epitope will not be very useful, yet it will nonetheless have an overall accuracy of 98% and a specificity of 100%. The true positives were 141 (5% cutoff), 132 (4% cutoff), 123 (3% cutoff) and 114 (2% cutoff). False negatives were 25, 34, 43 and 52, while the false positives decreased from 2743 to 2173, 1618 and 1060, respectively. The parameter sensitivity varies from 69% (at 2% cutoff) to 85% (at 5% cutoff) Fig (4). The parameter PPV diminishes from 10% (at 2% cutoff) to 5% (at 5% cutoff). Thus, the tests indicate that a 5% threshold at the final epitope selection step is sufficient to generate an 85% epitope prediction. This means that by using profile HMM, one need only test 5% of the whole sequence in order to predict 85% of available epitopes.

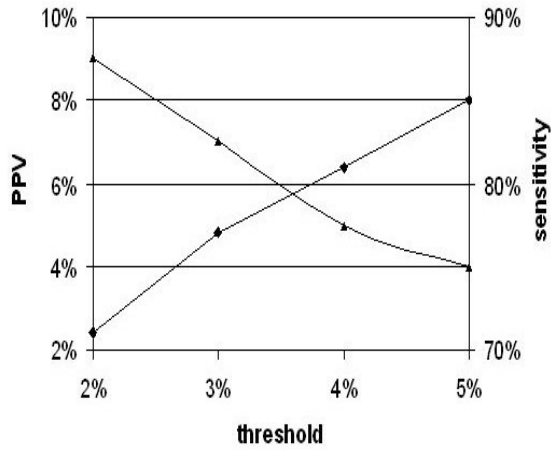

Fig. 4-ROC for T-cell epitope prediction with various cutoffs

6. Conclusion

A new encoding method for the direct recognition of T-cell epitopes and non-epitopes through Profile HMM has been developed. Our method gives much result in terms of specificity and sensitivity. Our method aims to rationalize the process of epitope searching and accelerate epitope-based vaccine design. They possess significant potential for improving the predictive ability of in silico epitope identification by adding more features and new high quality experimental data.

References

[1] Zinkernagel M., Doherty P.C. (1974) *Nature*, 248: 701-702.

[2] Terasaki P.I. (2007) *Immunol Res*, 2007; 38: 139-48.

[3] Craiu A., Akopian T., Goldberg A., Rock K.L. (1997) *Proc Natl Acad Sci US*, 94:10850-10855.

[4] Jensen P.E. (2007) *Nat Immunol*, 8: 1041-8.

[5] Maenaka K., Jones E.Y. (1999) *Curr Opin Struct Biol.*, 9: 745-53.

[6] Mamitsuka H. (1998) *Proteins,* 33: 460-74.

[7] Noguchi H., Kato R., Hanai T., Matsubara Y., Honda H., Brusic V., etal. (2002) *J. Biosci Bioeng*., 94: 264-70.

[8] Zhang C., Bickis M.G., Wu F.X., Kusalik A.J. (2006) *J. Bioinform Comput Biol*, 4: 959-80.

[9] Achuthsankar S. Nair, Vrinda V. Nair and Vinod Chandra S. S. (2007) *Proceedings of Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07), Hyderabad, India*.

[10] Vinod Chandra S.S., Achuthsankar S. Nair, Vrinda V. Nair and Mahalekshmi T. (2007) *Journal of Computer Society of India*, 37: 8-11, 2007.

[11] Rabiner L.R. (1989) *Proceedings of the IEEE*, 77:.257-286.

[12] Viterbi A.J. (1967) *IEEE Transction on Information Theory*, 13:260-269.

[13] Forney G.D. (1973) *Proceedings of IEEE*, 61: 268-278.

[14] Swets J. A. (1988) *Science*, 240(4857). 1285–1293.

[15] Lund O., Nielsen M., Lundegaard C., Kesmir C. and Brunak S. (2005) *Immunological Bioinformatics, MIT Press, Cambridge, Mass, USA, 1st edition.*

[16] Baldi P., Brunak S., Chauvin Y., Andersen C.A. and Nielsen H. (2000) *Bioinformatics* 16: 412-424.

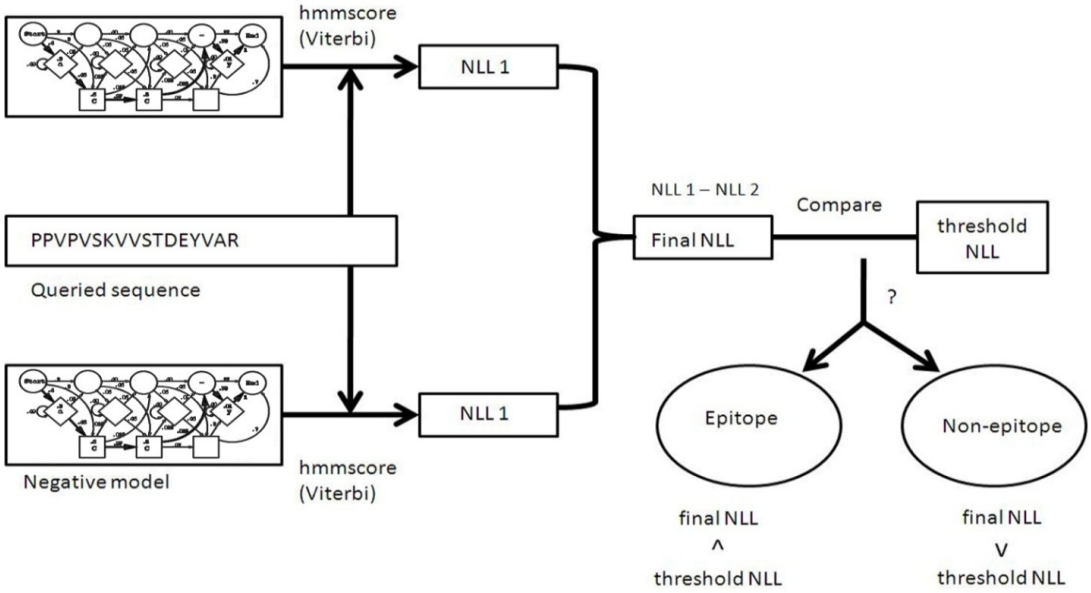

Fig. 2-Prediction system architecture

Table 1- Databases of MHC-Binding Peptides and T-Cell Epitopes

Database	URL	Description
SYFPEITHI	www.syfpeithi.de	MHC ligands and peptide motifs
HIV Database	www.hiv.lanl.gov/ content/index	HIV T-cell epitopes
EPIMHC	immunax.dfci.harvard. du/epimhc/	MHC ligands
MHCBN	http://www.imtech.res.in/raghava/mhcbn	MHC binding non binding peptides
ANTIJEN	www.antijen.org	MHC TAP binding peptides
IEDB	www.immuneepitope.org	MHC ligands and MHC binding peptides