Bioinformation Volume 5

open access

www.bioinformation.net Issue 9

Hypothesis

Comparative analysis of epitope predictions: proposed library of putative vaccine candidates for HIV

Arun Gupta[1, 2], Dinesh Chaukiker[1], Tiratha Raj Singh[3, 4]*

[1]School of Computer Science & IT, DAVV, Indore, India;[2]Computational Biology Group, Abhyudaya Technologies, India; [3]Bioinformatics Sub-Centre, School of Biotechnology, DAVV, Indore, India; [4]Department of Biotechnology and Bioinformatics, JUIT, Waknaghat, Solan, India; Tiratha Raj Singh - Email: tiratharaj@gmail.com; *Corresponding author

Received June 28, 2010; Accepted November 20, 2010; Published February 07, 2011

Abstract:
Designing a vaccine for a disease is one of the crucial tasks that involve millions and billions of dollars, several decades and yet there is no guarantee of successful results. Several pharmaceutical companies are investing their money and time in such activities. Computational biology could be of great help in these activities by proving a library of plausible candidates that might actually show some positive responses. MHC binding peptide prediction is one such area where the immense power of computers could be used to get a breakthrough. In this direction several databases and servers have been developed by many labs to predict the MHC binding peptides. These short peptides on the antigen surface are recognized by the MHC molecule and are presented to the receptors of T-cells for further immune response. Peptides that bind to a given MHC molecule share sequence similarity. Here we present a comparative study of servers that can predict the MHC binding peptides in a given protein sequence of the antigen. Based on this comparative analysis on HIV data, we are able to propose a library of putative vaccine candidates for the *env GP-160* protein of HIV-1.

Keywords: MHC, MHC binding peptides, HIV-1, epitopic library, putative vaccine candidates.

Background:
With the progression and success rates of several genome projects, we are provided with exponentially increasing number of proteins. This protein sequence data also include vital hidden information of pathogenic response of an organism. Sequences from pathogens provide a huge amount of potential vaccine candidates. The first step in the development of peptide vaccines is the identification of the immunodominant peptides along with proteins sequence. Theoretically every sub-sequence along with the protein could be antigenic. The experimental identification of peptide binding affinity to Major Histocompatibility Complex (MHC) **[1]** molecules requires a binding assay of each peptide, which is a time consuming and costly process. Therefore, a number of alternative research efforts have been carried out in an attempt to discover the laws of binding peptide sequence patterns. One could perform several computational analyses to screen and develop libraries of such peptides. Such libraries could help the research and development process of several pharmaceutical companies saving them money and time and will also insure less hazardous situation to handle the pathogens.

One of the major roles of immune system is to recognize and destroy cells expressing non-self or mutated proteins. The cascade of immune response begins when the cytotoxic T lymphocytes (CTLs) recognizes the specific antigen and trigger the specific immune response against them **[2].** If the antigens aren't recognized properly on the early onset of the infection, might even lead to lethal diseases. MHC molecules play an important role in the immune system. MHC helps T-Lymphocyte cells to identify the antigens and to trigger the immune response. The complex of bound peptide and MHC complex induces the naïve T cells to proliferate and differentiate into armed effector T cells that help to remove the antigens. There is a great diversity in the selectivity of peptides by MHC molecules which makes it difficult for pathogens to escape the immune response. Each different MHC molecule can bind to a set of different peptides **[3].** This brings into the consideration, the several supertypes of MHC molecules with different alleles in many supertypes.

The *env* gene in HIV encodes a single protein, *gp160*. When *gp160* is synthesized in the cell, cellular enzymes add complex carbohydrates and turn it from a protein into a glycoprotein - hence the name gp160 rather than 'p160'. gp160 is found on the outer surface or envelope of the HIV. It is composed of gp120, which protrudes from the envelope, and gp41, which is embedded in the envelope. gp160 is the unit that helps the virus to adhere and to interact with the surface proteins of host. The majority of HIV subunit vaccines are based on the envelope proteins of HIV namely gp120 and gp41, which form the gp160 glycoprotein complex, or on selected epitopes identified within these proteins **[4; 5]**

HLA-A2 is a human leukocyte antigen *serotype* within HLA-A "A" serotype group. The serotype is determined by the antibody recognition of α^2 subset of HLA-A α-chains. For A2, the alpha "A" chain are encoded by the HLA-A*02 allele group and the β-chain are encoded by B2M locus [6]. The serotype identifies the gene products of many HLA-A*02 alleles, including HLA-A*0201, *0202, *0203, *0206, and *0207 gene products. A2 is the most diverse serotype, showing diversity in Eastern Africa and Southwest Asia. While the frequency of A*0201 in Northern Asia is high, its diversity is limited to A*0201 the less common Asian variants A*0203, A*0206. Due to its diversity, it is an interesting entity to study the antigen-antibody interactions.

This study deals with a comparative analysis of epitope and MHC binding peptide prediction on envelope proteins of HIV-1. The potential value of a

Bioinformation Volume 5 open access
www.bioinformation.net Issue 9 Hypothesis

preventative and cost-effective vaccine stratagem to protect against HIV is inevitable. Based on this analysis a putative vaccine candidate library has been fabricated which focuses mainly on the performance of servers used and their predicted accuracy with their respective parameters. Generated library will be a useful resource in the process of vaccine design for HIV-1 and it will also help in the generation of similar libraries for other pathogens.

Material and methods:

Here in this study, we tested the same sequences over several servers with the similar prediction conditions and compared the results obtained. The protein sequences were fetched from the National Center for Biotechnology Information (NCBI) through their enterz search engine. We considered HIV-1 surface and envelope proteins in our study.

In our study, we compared 8 comprehensive servers **(Table 1 see Supplementary material),** based on their results and scores they assigned to different peptides predicted. Additional information provided by every server has also been analyzed along with the results. Analysis has been performed on the basis of final score given by the server. Basic description and working principles of servers is given in the text following.

ANNpred **[7]** is based on Artificial Neural Networks (ANNs) for 30 MHC alleles. *ComPred* is a comprehensive method for prediction of MHC binding peptides or CTL epitopes of 67 MHC alleles. The prediction for 30 alleles is based on the hybrid approach of Artificial Neural Networks (ANNs) and Quantitative Matrices (QM). The prediction for rest 37 MHC alleles is based on the quantitative matrices. The predicted MHC binders in ComPred and in ANNpred, both are filtered to potential CTL epitopes by using Proteasomal matrices.

The *Predep* **[8]** algorithm uses the pair-wise potential table of Miyazawa & Jernigan **[9],** and is able to identify good binders only for MHC molecules with hydrophobic binding pockets. This server returns the peptide "energy score" value. The peptides are ranked according to their energy score (the lower the better).

RankPep **[10]** server predicts MHC-I and MHC-II peptide binders from protein sequence or sequence alignments using Position Specific Scoring Matrices (PSSMs) or profiles from set aligned peptides known to bind to a given MHC molecule as the predictor of MHC-peptide binding. In addition, it predicts those MHC-I ligands with a C-terminal end is that likely to be the result of proteasomal cleavage. Profiles basically consist of a table listing the observed sequence-weighted frequency of all amino acids in every column of a sequence alignment. This server includes a selection of 102 and 80 PSSMs for the prediction of peptide binding MHCI and MHCII molecules, respectively.

HLA-BIND **[11]** allows users to locate and rank peptides that contain peptide-binding motifs for HLA class I molecules. By default, this server predicts 9-mer peptides using 20-by-9 coefficient matrix for the selected HLA molecule for the scoring. The estimated numerical score for the subsequence in case of HLA-A2 is calculated based upon the half-time of dissociation of complexes containing the peptide at 37^{o}C at pH 6.5. For other molecules, the estimate is based on the observed anchor residue preferences.

ProPred1 **[12]** is a matrix based method that allows the prediction of MHC binding sites in an antigenic sequence for 47 MHC class-I alleles. The matrices used in ProPred1 have been obtained from BIMAS **[11]** server and matrices described by Toes **[13].** ProPred1 also allows the prediction of the standard proteasome and immunoproteasome cleavage sites in an antigenic sequence. It allows filtering of MHC binders, who have cleavage site at C terminus. The most of matrices in Propred1 is multiplication type where the score of each predicted peptide is calculated by multiplying scores of each position.

We carried out the multiple sequence alignment (MSA) of various HIV envelop protein sequences through ClustalX, to select one representative sequence. We selected CAQ63623.1, being the most consistent based on the blocks appeared among all the sequences in their MSA. Resulted and computed epitopes were ranked according to comparative and statistical analysis system. Additionally all the epitopes have been compared with the LANL immunological database collection of T-cells **[14]**. As the protein is an envelope protein that helps the virus to attach with the host cell, our hypothesis is that, if the CTL can trigger an immune response against the antigen virus targeting the envelop proteins, then virus will not be able to cause any infection.

Results and discussion:

Three-tier screening was performed on the results obtained from various servers. Pre-screening: to collect the top 25 scoring peptides from all the servers. In the second round, out of the top 25 scoring peptides we selected 20 most consistent and common peptide sequences and compared the ranks and their respective positions results from every server against all other servers. This yielded a comprehensive list of 158 peptides (data not shown). Out of the comprehensive list, we selected the peptides with at least 50% occurrence and calculated their average rank **(Table 2 see Supplementary material).** Following this approach, we were able to analyze the selected sequence through various approaches viz, ANN, SVM, PSSM, etc and against various database as well. This three tier scanning increases the chances of accuracy and reduces the false-positive hits if any. Hence the obtained library is the collective favorable results of all the servers together.

An interesting aspect of this analysis was extraction of few interesting plausible epitopes while compared with LANL immunological database collection of T-cells **[14]**. Four epitopic peptide sequences viz. 'WLWYIKIFI', 'NVWATHACV', 'LLDTIAIAV', 'YIKIFIMIV', have been found in LALN database. The most important looks 'WLWYIKIFI', as it has been ranked 2nd in our analytical system and is for HLA A*0201, A2 and A2.1. Besides that we also propose another plausible epitopic sequence 'TLFNNSWTL', as it has been ranked one in our system and can be verified experimentally and might be a part of immunological databases in future. Other sequences can also included in the immunological databases after successful experimental verification.

Conclusion and Future Prospects:

Design and development of an effective HIV vaccine is exigent because of complex host-virus pathogenesis. From the study carried out and results obtained it is proposed that the given library contains the most putative vaccine peptide candidates for HIV-1. If we can target these peptide candidates, we might be able to have substantial vaccine for the cure of this disease. The putative vaccine candidates obtained might support our proposed hypothesis and will help in hindering the antigenic effect of the virus. Further, similar study can be carried out to design a consolidate library over the whole proteome level which might provide many more putative candidates, increasing the success rate of winning over the disease analysis. This study also proposes a model for carrying analyses on specific data with diversified techniques to extract something attention-grabbing and decisive.

We would like to extend this work towards the integration of various features of many servers used and would also like to apply more statistical techniques to make results more significant and better. Structural aspects of epitopes will also be incorporated to make results more robust and significant.

Acknowledgement:

Authors would like to thank the organizing committee of INBIX'10, Bioclues organization and Bioinformatics organization to give us an opportunity to present our work at INBIX'10.

References:

[1] CA.Janeway *et al Immunobiology: The Immune System in Health and Disease*. Garland Publishing, New York. (2001)
[2] PF Robbins & Y Kawakami *Curr Opin Immuno* **8**(5): 628 (1996)

[3] B Peters *et al. Bioinformatics* (2003) **19**(14): 1765 [PMID: 14512347]
[4] KB Bond et al. AIDS Res Hum Retroviruses. **17**(8):703 (2001) [PMID: 11429111]
[5] P Kangueane et al Designing HIV gp120 peptide vaccines: rhetoric or reality for neuroAIDS Chapter 9. In K. Goodkin (2008)
[6] P Shapshak & A Verma (ed.) The Spectrum of Neuro-AIDS Disorders: Pathophysiology, Diagnosis, and Treatment, ASM Press, Washington, DC 105
[7] B Arce-Gomez *et al. Tissue Antigens* **11**(2)**:** 96 (1978) [PMID: 77067]
[8] S Lata *et al. Methods in Mol Bio* **409**: 201 (2007) [PMID: 18450002]
[9] O Schueler-Furman *et al.Protein Sci* Protein Sci **9**(9): 1838 [PMID: 11045629]
[10] S Miyazawa RL Jernigan *J Mol Biol* **256**(3): 623 (1996) [PMID: 8604144]
[11] PA Reche *et al. Human Immunology* **63**: 701 (2002) [PMID: 12175724]
[12] KC Parker *et al. J. Immunol* **152**(1): 163 (1994) [PMID: 8254189]
[13] H Singh & GP Raghava *Bioinformatics* **19**(8): 1009 (2003) [PMID: 12761064]
[14] RE Toes *et al. J Exp Med* **194**(1): 1 (2001) [PMID: 11435468]
[15] http://www.hiv.lanl.gov/content/immunology/

Edited by P. Kangueane
Citation: Gupta *et al.* Bioinformation 5(9): 386-389 (2011)

Bioinformation Volume 5 open access

www.bioinformation.net Issue 9 **Hypothesis**

Supplementary material:

Table1: Comprehensive list of 8 available online servers for MHC peptide prediction used in this study.

Server	*MHC Class*	*MHC Method*
BIMAS (HLABind)	I	Quantitative matrices (QM)
CTLPRED	I	Combination of Artificial Neural Networks (ANN), and Support vector machines (SVM)
MHCPred	I	QM
nHLAPred (ANNpred and ComPred)	I	ANN
PREDEP	I	Motif matrices (MM)
PROPRED1	I	MM, also proteasomal cleavage and promiscuous peptiptides
RANKPEP	I and II	Position Specific Scoring Matrices (PSSM)
SYFPEITHI	I and II	MM, Only a few MHCII molecules

Table 2: Comprehensive library of putative vaccine candidates for HIV-1 with their computed ranks, Epitopic sequences have been highlighted in bold which are part of LALN immunological database collection of T-cells **[14]**.

		Ranks										
Peptide	**Position**	**ANNPred**	**COMPred**	**HLABind**	**MHCPEP**	**Predep**	**Propred1**	**RankPep**	**Syfpeithi**	**% of Occurrence**	**Average rank**	**Final Rank**
TLFNNSWTL	*416*	--	--	2	6	4	2	--	15	62.50%	5.80	***1***
WLWYIKIFI	*704*	19	13	1	2	1	1	--	--	75.00%	6.17	***2***
NLWRWGTMI	*12*	1	1	7	18	--	7	--	--	62.50%	6.80	***3***
SLFYRVDIV	*198*	11	7	5	15	--	5	--	4	75.00%	7.83	***4***
KMHSLFYRV	*195*	10	6	3	11	--	3	16	--	75.00%	8.17	***5***
TMILGMIII	*18*	2	2	--	--	12	--	18	--	50.00%	8.50	***6***
ALYRVATQL	*369*	16	10	10	5	--	10	1	12	87.50%	9.14	***7***
SLAEKEIRI	*293*	14	9	15	--	2	15	2	9	87.50%	9.43	***8***
SLLDTIAIA	*846*	--	--	9	21	18	9	4	1	75.00%	10.33	***9***
NVWATHACV	*66*	6	3	13	--	23	13	--	--	62.50%	11.60	***10***
QLQARVLAM	*601*	17	11	--	--	7	--	12	--	50.00%	11.75	***11***
QLTPLCVTL	*120*	--	--	--	3	9	25	17	7	62.50%	12.20	***12***
RLVQGFLAL	*773*	--	25	16	7	3	16	7	--	75.00%	12.33	***13***
LLDTIAIAV	*847*	--	19	14	8	19	14	6	--	75.00%	13.33	***14***
QMHADIISL	*102*	7	4	18	16	--	18	--	18	75.00%	13.50	***15***
AVLAIINRV	*726*	22	15	12	--	--	12	8	16	75.00%	14.17	***16***
YIKIFIMIV	*707*	20	--	--	--	15	--	14	8	50.00%	14.25	***17***
GLRIVFAVL	*720*	21	14	--	9	22	--	--	6	62.50%	14.40	***18***
RLRDFVLIA	*796*	24	16	--	19	6	--	--	10	62.50%	15.00	***19***
VLLYWGREL	*832*	--	18	20	12	14	20	--	--	62.50%	16.80	***20***