Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Implication from predicted HLA-DRB1 binding peptides in the membrane proteins of *Corynebacterium diphtheriae*

Febin Prabhu Dass[1] and Vemuri Lakshmi Deepika[1, *]

[1]Bioinformatics Division, School of Biotechnology, Chemical and Biomedical Engineering; VIT University, Vellore, Tamilnadu, India; Vemuri Lakshmi Deepika* - Email: v.l.deepika@gmail.com; * Corresponding author

received September 18, 2008, accepted October 21, 2008; published November 02, 2008

Abstract:
The aerobic gram positive bacterium *Corynebacterium diphtheriae* causes diphtheria, a respiratory tract illness characterized by symptoms such as sore throat, low fever, and an adherent membrane on the tonsils, pharynx, and/or nasal cavity. Therefore, it is important to develop preventive vaccines for diphtheria. The availability of the 2,488,635 bp long complete sequence for the *C. diphtheriae* genome provides an opportunity to understand cell mediated immune response using Computational Biology tools from the bacterial proteome sequence data. We selected 355 membrane proteins from the *C. diphtheriae* proteome using annotation data to identify potential HLA-DRB1 binding short peptide using modeling, simulations and predictions. This exercise identified 30 short peptides in membrane proteins showing binding capability to HLA-DRB1 alleles. These peptides serve as outline for the understanding of cell mediated immune response to *C. diphtheriae*. It should be noted that the predicted data to be verified using binding assays for further consideration.

Keywords: peptide epitopes; vaccines; HLA-DRB; membrane proteins; *C. diphtheriae*

Background:
Diphtheria is an epidemic disease that remains as a threat to health in the developing world **[1]**. The diphtheria epidemic had caused more than 157,000 cases and 5000 deaths according to WHO reports **[2]**. Diphtheria is caused by *Corynebacterium diphtheriae*, a non-sporulating, non-encapsulated, non-motile, pleomorphic gram-positive bacillus. The complete genome of the living organism *Corynebacterium diphtheriae* was sequenced **[3]**. The sequence and annotation of genome is available in DDBJ/EMBL/GenBank databases with the accession number BX248353. Immunization leads to the disappearance of toxigenic strains, but toxigenicity can be rapidly conferred on non-toxigenic strains via phage conversion. This makes the return of epidemic diphtheria a real threat when there is insufficient immunity. Hence, it is crucial and critical develop vaccines for diphtheria. Identification of potential peptide candidates for vaccine development have been shown using computational methods for pathogenic organisms elsewhere **[4]**. The application of proteome sequence data and Bioinformatics tool for the development of peptide vaccines for infectious diseases has gained momentum in recent years **[5]**.

C. diphtheriae virulence factors, other than diphtheria toxin can be the potential cause to the disease. Recently, surface proteins of 67 and 72kDa, named 67-72p were isolated and related to the attachment of *C. diphtheria* to the human erythrocytes **[1]**. The pathogenicity island of the organism encode vast majority of fimbria and iron uptake system. The fimbrial system show similarities to sortases bound to surface proteins and are considered as pathogenicity factor **[3]**. It can be drawn that membrane related proteins could have pathogenicity property. The efficacy of immune response is associated with antigen specificity, diversity and human leukocyte antigen (HLA) alleles. **[6]**. There are two classes of HLA molecules, namely class I and class II that are recognized for two distinct sets of T cells, the $CD8^+$ and $CD4^+$ T cells, respectively **[7, 8]**. There are evidences suggesting the association of diphtheriae antigen with class II alleles. **[9]** Nonetheless, the HLA molecules are highly polymorphic and there are more than 800 class II HLA alleles known till date **[10]**. Here, we describe the identification of potential peptides in the membrane proteins of *C. diphtheriae* binding to HLA DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0701, DRB1*1801, DRB1*1101 and DRB1*1501 using predictions, modeling and simulations. It should be noted that these alleles have an allelic frequency of 20-50% **[11]**

Methodology:

***C. diphtheriae* membrane protein dataset**

The whole genome sequence of *C. diphtheriae* with annotated protein sequences is available at the EMBL/GenBank database

lSSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 111-113 (2008)

with accession number BX248353. The genome is 2,488,635 bp long containing 2320 CDS with about 53.5% G+C content **[3]**. The annotated protein sequences for membrane or associated proteins that are annotated as putative ABC transport membrane protein, integral membrane protein, membrane protein, fimbrial protein, surface anchored protein and hypothetical membrane protein. Thus, we created a membrane protein dataset from *C. diphtheriae* with 355 sequences (list available from authors upon request).

Prediction of antigenic regions in membrane proteins

It is important to identify antigenic regions in the 355 membrane or associated proteins in *C. diphtheriae*. The antigenic subsequences from the membrane proteins were identified using the Antigenic server in EMBOSS GUI version 1.12 **[12]**. The server uses a semi-empirical method for the prediction of antigenic regions in a protein based on the physiochemical properties of amino acids and their frequency of occurrence in experimentally known segmental epitopes as described elsewhere **[13]**. The stringency level is enhanced by keeping 2.3 as a cut-off value. Peptide sequences with less than 9 residues were ignored as class 2 HLA molecules bind peptides of length 12-15 residues long **[14]**. Thus, we identified 1448 antigenic regions from the 355 membrane proteins. However, we further reduced the set to 30 based on antigenic score obtained from the server (Table 1 under supplementary material). We used this dataset for HLA binding peptides as described in the next section.

.

Prediction of T-cell peptides

PROPRED server **[15]** was used for the prediction of MHC class II HLA DRB1 allele binding regions in the identified antigenic regions (Table 1 under supplementary material). The server employs amino acid position coefficient tables deduced from literature by linear prediction model **[16]**. A threshold of 3% was fixed to reduce the rate of false positives. Prediction were performed for alleles DRB1*0101, 0401 and 1501.We restricted our analysis to these alleles due to availability of structures determined by X-ray crystallography at the protein databank (PDB).

Discussion:

The availability of the completely sequenced genome of *C. diphtheriae* provides an opportunity to understand pathogenicity using a genome wide scanning and analysis. The annotated genome of *C. diphtheriae* contains 2320 CDS. We identified 355 membrane proteins using annotated protein information the dataset. This accounts for only 15% of the genome CDS. The corresponding protein sequences of 355 proteins were used as a dataset for the identification of peptide antigenic regions using ANTIGEN and PROPRED servers. The ANTIGEN server uses physiochemical properties of amino acids to determine antigen region regions in the membrane protein dataset. This exercise resulted in about 1448 short segments of length 9 or more from the 355 membrane proteins. The least score for the predicted antigenic peptide sequence is found to be 0.92 and the maximum score is 3.09. The minimum length of the peptide is kept as 9 residues in the parameter selection. The antigenic peptide sequences up to a maximum length of 54 residues were obtained.

The data on immunogenic to *C. diphtheriae* antigen is limited except for some sporadic data. We further reduced this number to 30 segments using antigenicity score obtained from ANTIGEN (Table 1 in supplementary material). This set of peptides is used for the screening of potential binding to HLA DRB1*0101, 0401 and 1501 using PROPRED. These alleles have coverage of about 20-150% among different populations **[11]**.The screening of peptide bindings to HLA DRB1*0101, 0401 and 1501 using PROPRED identified a list of peptides with positive binding scores (Table 1, see supplementary material). This list of peptides provides a framework for further investigation of *C. diphtheriae* towards the development of vaccine candidates. Subunit vaccines consisting of diphtheriae protein antigens present a safe and specific tool for the prevention of Diphtheria. The identification of promiscuous binding to HLA is an ideal prerequisite for the design of subunit vaccines. It should be noted that these are predicted data requiring validation using binding assays. The current challenge in synthetic vaccine design is the development of a methodology to identify and test short antigen peptides as potential T-cell epitopes. Data driven statistical methods such as PROPRED are generally available for class II DR1, and DR4. Reliable predictions of immunogenic peptides can reduce the experimental effort needed to identify new epitopes, and though reliable predictions of the MHC binding can be used to rank the possible epitopes very accurately. In this paper we have tried to accurately predict antigenic peptides using PROPRED which is a tool used to predict the MHC binding peptides. Class II MHC binding predictions need to develop a greater accuracy level, but new tools have emerged that deliver significantly improved predictions not only in terms of accuracy, but also in MHC specificity coverage. The future development of advanced tools using structured based features will significantly improve the efficiency of prediction for high true positives considerations.

Conclusion:

The scanning of immunologically relevant regions of the bacterial proteome sequence of *C. diphtheriae* is essential in the identification of specific HLA binding peptides for potential design of vaccine candidates for diphtheria. Here, we identified 30 segments using ANTIGEN at EMBOSS for further screening of these peptides to PROPRED for DRB1*0101, 0401 and 1501 specific binding. The specific segments showing positive binding score with each of these alleles is presented for further investigations using modeling towards *in vitro* investigations.

References:

[01] A. L. Mattos-Guaraldi *et al.*, *Mem Inst Oswaldo Cruz.*, 98: 987 (2004) [PMID: 15049077]

lSSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 111-113 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

[02] S. Dittmann *et al.*, *J Infect Dis*., 181: S10 (2000)
[03] A. M. Cerdino-tarraga *et al.*, *Nucleic Acids Res*., 31: 6516 (2003) [PMID: 14602910]
[04] D. N. Chakravarti *et al.*, *Vaccine*, 19: 601 (2001)
[05] N.Petoskey *et al.*, *In Silico Biology*, 3: 411 (2003) [PMID: 12954084]
[06] J. J. Kuhns *et al.*, *J. Biol. Chem*., 274: 36422 (1999) [PMID: 10593938]
[07] C. Watts, *Immunol*., 15: 821 (1997) [PMID: 9143708]
[08] R. N. Germain, *Cell*, 76: 287 (1994) [PMID: 8293464]
[09] N. D. Iushchuk *et al.*, *–Zn Mikrobial Epidermiol Immunobiol*., (1997) [PMID: 9460869]
[10] http://imgt.cines.fr/textes/IMGTrepertoireMHC/LocusGenes/nomenclatures/human/MHC/hla_serology.html#DR
[11] M. Panigada *et al.*, *Infection and immunity*, 70: 79 (2002) [PMID: 11748166]
[12] T. Carver and A. Bleasby, *Bioinformatics*, 19: 1837 (2003) [PMID: 14512356]
[13] A. S. Kolaskar and P. C. Tongaonkar, *FEBS Letters*, 276: 172 (1990) [PMID: 1702393]
[14] P. Kangueane *et al.*, *Human Immunol.*, 62: 539 (2001) [PMID: 11334679]
[15] H. Singh and G. P. S. Raghava, *Bioinformatics*, 17: 1236 (2001) [PMID: 11751237]
[16] T. Sturniolo *et al.*, *Nat. Biotechnology*, 17: 555 (1999)[PMID:10385319]

Edited by P. Kangueane
Citation: Dass & Deepika, Bioinformation 3(3): 111-113 (2008)

Supplementary material

	Allele	GenBank ID	Peptide*	PROPRED Binding Score
1	DRB*0101	CAE49048	LRILLAVAV	76.67
2		CAE50605	FVALIGVSI	68.33
3		CAE49825	VRMLVATYV	64.83
4		CAE49471	FVLLLAKLM	64.67
5		CAE50256	VVLIVAVPL	63.33
6		CAE50851	YVLAIAVLA	63.33
7		CAE50475	FLINVPVVI	62.33
8		CAE48920	LVLIAAIVL	61.67
9		CAE49353	IVLLPCVVA	61.67
10		CAE48741	LVMLYPVTS	60.00
11	DRB*0401	CAE50577	VVLSVVLLS	60.47
12		CAE50281	FVVSIVLIA	58.14
13		CAE50864	VFLTTTLLS	55.81
14		CAE49353	VIIIASLVS	55.81
15		CAE49135	VRLIDVIVS	54.42
16		CAE48741	LVILGVLVM	53.49
17		CAE50652	VLLSVTVGS	52.09
18		CAE50802	WVFLSACIS	51.16
19		CAE48779	FIGLNVLLS	51.16
20		CAE48675	VVILGVIVA	50.93
21	DRB*1501	CAE48779	VVAYPRLPL	81.63
22		CAE50888	VVYFLLLLV	78.57
23		CAE50670	VVIFAASTI	71.43
24		CAE49665	IVFFVAMVA	71.43
25		CAE48521	VILYFGMQA	69.39
26		CAE50218	VVIFHCFVG	67.35
27		CAE48598	LVNYVGHDV	66.33
28		CAE50340	IVLLAGLSL	66.33
29		CAE50810	LRFIVGLGL	66.33
30		CAE50855	LVLLVGLAV	64.49

Table 1: Predicted peptides binding to HLA DRB1*0101, 0401 and 1501. (*The anchoring region in class 2 HLA molecules lies within 9 residues window. These peptides were selected by both ANTIGEN and PROPRED).

lSSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(2): 111-113 (2008)