Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604
Vol 3, Issue 1, 2012, pp 238-245 http://www.bipublication.com

GENOME BASED VACCINE DEVELOPMENT IN *BURKHOLDERIA MALLEI* THROUGH *IN SILICO* IDENTIFICATION OF CELL SURFACE ANTIGEN

Mohd. Sohel Ather Shaikh
P.G.Department of Bioinformatics, Shankarlal Khandelwal College, Akola, Maharashtra (INDIA)
(sohel.bioinf@rediffmail.com /sohelather@gmail.com)

ABSTRACT:-
Burkholderia mallei are responsible for Glander disease especially in Horse and humans. Several strains are reported to be resistant to prophylactic drugs and hence demands for better prevention than cure. In view of vaccine development, genome of *Burkholderia mallei* provided us the opportunity for locating several surface antigens involving bioinformatics approach. Target identification is the first step in the drug and vaccine discovery process. The *In silico* method reduces the time as well as the cost of target screening. Although a powerful technique that can be applied to a wide range of pathogens, in my study Bio-programs like SignalP3.0, LipoP1.0, TMHMM, Fuzzpro, BLASTP and HLA Pred searched out highly conserved surface antigens as lipoprotein and cell wall anchored proteins, and highlighted 104 prominent surface antigens which are involved in subunit vaccine development programs.

KEYWORDS:- Vaccine leads, Cell surface antigen, Reverse vaccinology, Bioprograms, LPXTG motifs, HLA binders.

1) INTRODUCTION:-

Burkholderia mallei, it is non-motile, aerobic Gm-ve cocco-bacillus a Burkholderia-genus human and animal pathogen causing Glanders; the Latin name of this disease (*malleus*) gave name to the causative agent species and grows on a Macconkey agar. It is closely related to *B. pseudomallei* [8] *B. mallei* evolved from *B. pseudomallei* by selective reduction and deletions from the *B. pseudomallei* genome [15]. *B.mallei* was used in Ist and IInd world war, even though it is so highly infective and a potential biological weapon.[21].The bacteria usually infect a person through their eyes, nose, mouth, or cuts in the skin. Once a person is infected with the bacteria, they develop a fever and rigors. Eventually they will get pneumonia, pustules, and abscesses, which will prove fatal within a week to ten days if left untreated by antibiotics. If the bacteria enter through the skin, a local skin infection can result, while inhaling *B. mallei* can cause septicemia or pulmonary infections of muscles, the liver, or spleen. *B. mallei* infection has a fatality rate of 95% if left untreated, and a 50% fatality rate in individuals treated with antibiotics. There is currently no vaccine available for humans or animals to protect against *B. mallei* infection.[6]. The process begins with use of bio-programs for the identification of all putative surface proteins, which are most logical choice as vaccine leads. Several surface localized proteins predicted with precision within genomic sequences using several computer programs based on signal peptides, LPXTG motifs, trans-membrane helices and other surface protein prediction algorithms [9,10]. This approach was first successfully applied to identify vaccine leads in *Neisseria meningitides* [3,14], *Streptococcus agalactiae* [12], *Staphylococcus aureus* [4], *Porphyromonas gingivalis* [16], *Chlamydia pneumonia* [13], *Bacillus anthracis* [15], *Streptococcus suis* [11], *Echinococcus granulosus* [7], and *Streptococcus sanguinis*

[19]. With the web based bio-programs, search could be focused on both lipoproteins and cell wall anchored proteins those believed to be most likely vaccines leads via comparative genomics [20]. By applying this methodology, certainly the sum of efforts required decreases to a greater extent with respect to PCR amplification, gene cloning and protein expression and gives a greater chance of success. To screen probable vaccine leads, bioprograms such as SignalP, LipoP, FuzzPro in EMBOSS package and TMHMM were involved, which can sort lipoproteins and cell wall anchored proteins based on their signature sequence pattern available in protein sequences [2,9,10]. My study based on the Bioinformatics approach, highlighted selection of the best vaccine leads available in the genome of *Burkholderia mallei* and those could be involved in vaccine development program.

2) MATERIALS AND METHODS:-

2.1) DATA COLLECTION:-

The primary information regarding the availability of protein sequences of *B.mallei* was gathered from the website: www.genome.jp/kegg/.

2.2) BIO-PROGRAMS:-

2.2.1) SignalP 3.0:-

SignalP 3.0 server predicted the presence and location of signal peptide cleavage sites in proteins using Gram-positive prokaryotes as default setting. The method incorporated a prediction of cleavage sites and a signal peptide/non-signal peptide prediction using artificial neural networks. The website address is: www.cbs.dtu.dk/services/SignalP..

2.2.2) LipoP:-

The lipoP server predicted lipoproteins and discriminates between lipoprotein signal peptides, other signal peptides and n-terminal membrane helices in Gram positive bacteria. The result was recorded based on the positive signal for SpI as signal peptide along with cleavage sites for signal peptidase I and SpII as signal peptide along with cleavage sites for signal peptidase II and recorded negative for signals of TMH and CYT which suggests the protein was showing the possibility of spanning n-trans-membrane helix and cytoplasmic regions respectively. The web address is: www.cbs.dtu.dk/services/LipoP.

2.2.3) TMHMM:-

The TMHMM server predicted trans-membrane helices in proteins by searching hydrophobic regions. The algorithm predicted number of helices and highlighted spanning length of peptides. Sorted proteins filtered and selected further with maximum two trans-membrane helices only. Such a decision made as earlier failures were reported for protein's unsuccessful expression in *E. coli*, having more than two trans-membrane helices and so makes it difficult to study in *In vitro* conditions. The web address is: www.cbs.dtu.dk/services/TMHMM/.

2.2.4) Fuzzpro:-

The online server used to search following pattern in *B.mallei* [5].

[LY]PX[TSA][GNAST]X(0,10){DEQNKRP}{
DEQNKRP}
{DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQ
NKRP}{DE
QNKRP}{DEQNKRP}{DEQNKRP}{DEQNK
RP}X(0,15)[DEQNKRH]X(0,5)>.

Fuzzpro used PROSITE patterns to recognize cell wall anchored proteins. Given patterns were specifications of a short length of sequence to be found. They were used to specify a search for an exact sequence or they can allow various ambiguities, matches to variable lengths of sequence and repeated subsections of the sequence. The web address is:
http://bioinformatica.cecalc.ula.ve/cgi-bin/emboss/fuzzpro.

2.2.5) BLASTP:-

Selected vaccine leads of *B.mallei* were analyzed by BLASTP for homology in protein databases of fifteen *Burkholderia* sp. The

selection of highly conserved leads was based on results obtained by setting E-value threshold at 0.0001. Such an E-value ensured that proteins filtered from SignalP3.0, TMHMM, Fuzzpro and LipoP were further signatured for their highly conserved nature among genus *Burkholderia*, which is a required for any vaccine lead. Fifteen *Burkholderia* species protein databases involved in analysis were: *Burkholderia.ambifaria, B.cenocepacia, B.gladioli BSR, B.glumae BGR1, B.multivorans, B.phymatum STM815, B.phytofirmans PsJN, B.rhizoxinica HKI 454, B.sp. 383, B.sp. CCGE1001, B .sp. CCGE1002, B.sp. CCGE1003, B.thailandensis, B.vietnamiensis G4, B.xenovorans LB400.*

The web address is:
http://www.ncbi.nlm.nih.gov/sutils/genom_table .cgi [17, 18].

2.2.6) HLA Pred:-

HLA Predicted HLA binding regions from antigen sequence. Each sequence in single letter FASTA format was given as an input. As in program, 36 class I and 51 class II alleles were available for analysis, but randomly we have selected 5 class I and 4 class II alleles.

3) **OBSERVATION AND RESULTS:-**

3.1) SignalP 3.0:-

In total 3400 proteins of *B. mallei*, program sorted only 185 proteins harboring signal sequences based on positive scores. The selection of each vaccine lead was based on positive peptide signals for all five measures as: max.C, max.Y, max.S, mean S and Mean D as shown in Fig 1
(A&B)

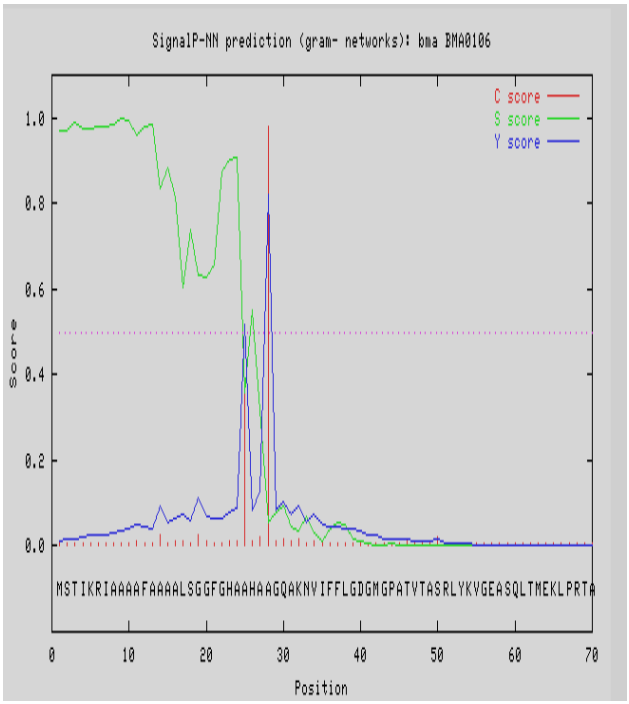

Fig 1 (A) Positive graphical view of neural network method as for C, S and Y scores in SignalP.

data

>bma_BMA0106 length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	28	0.978	0.52	YES
max. Y	28	0.822	0.33	YES
max. S	9	0.998	0.92	YES
mean S	1-27	0.831	0.49	YES
D	1-27	0.827	0.44	YES

Most likely cleavage site between pos. 27 and 28: AHA-AG

Fig 1 (B) Positive tabular data of neural network method as for C, S and Y scores in SignalP.

3.2) LipoP 1.0:-

Out of 3400 proteins of *B. mallei* screened for presence of lipoprotein, only 318 predicted to have defined signals,
collectively for SpI and SpII enzymes. The positive leads as lipoprotein were selected based on highest score obtained by either SpI or SpII as compared to score of TMH and CYT as in Fig 2 (A&B).

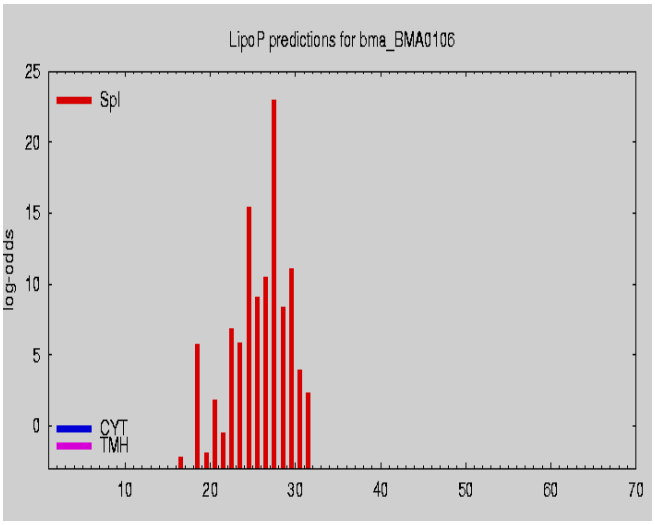

Fig 2 (A) Predicted positive signal for SpI by LipoP with probable cleavage at 27-28 in sequence.

```
# bma_BMA0106 SpI score=22.9594 margin=23.160313 cleavage=27-28
# Cut-off=-3
bma_BMA0106     LipoP1.0:Best   SpI     1       1       22.9594
bma_BMA0106     LipoP1.0:Margin SpI     1       1       23.160313
bma_BMA0106     LipoP1.0:Class  CYT     1       1       -0.200913
bma_BMA0106     LipoP1.0:Class  TMH     1       1       -1.42364
bma_BMA0106     LipoP1.0:Signal CleavI  27      28      22.9507 # HAAHA|AGQAK
bma_BMA0106     LipoP1.0:Signal CleavI  24      25      15.4318 # GFGHA|AHAAG
bma_BMA0106     LipoP1.0:Signal CleavI  29      30      11.0896 # AHAAG|QAKNV
bma_BMA0106     LipoP1.0:Signal CleavI  26      27      10.4374 # GHAAH|AAGQA
bma_BMA0106     LipoP1.0:Signal CleavI  25      26      9.07658 # FGHAA|HAAGQ
bma_BMA0106     LipoP1.0:Signal CleavI  28      29      8.32644 # AAHAA|GQAKN
bma_BMA0106     LipoP1.0:Signal CleavI  22      23      6.86327 # SGGFG|HAAHA
bma_BMA0106     LipoP1.0:Signal CleavI  23      24      5.86189 # GGFGH|AAHAA
bma_BMA0106     LipoP1.0:Signal CleavI  18      19      5.72096 # AAALS|GGFGH
bma_BMA0106     LipoP1.0:Signal CleavI  30      31      3.89548 # HAAGQ|AKNVI
bma_BMA0106     LipoP1.0:Signal CleavI  31      32      2.29008 # AAGQA|KNVIF
bma_BMA0106     LipoP1.0:Signal CleavI  20      21      1.781   # ALSGG|FGHAA
bma_BMA0106     LipoP1.0:Signal CleavI  21      22      -0.552228       # LSGGF|GHAAH
bma_BMA0106     LipoP1.0:Signal CleavI  19      20      -1.86942        # AALSG|GFGHA
bma_BMA0106     LipoP1.0:Signal CleavI  16      17      -2.17254        # FAAAA|LSGGF
```

Fig 2 (B) Tabular data for positive highest score for SpI predicted by LipoP.

3.3) TMHMM:-

By screening 3400 proteins of *B. mallei*, algorithm predicted presence of trans-membrane helices in the 250 proteins, which were further screened for number of trans-membrane helices spanned by each protein in the membrane. Hence in decision leads having more than two trans-helices were filtered out from the study and not considered as leads as in Fig 3 (A & B), Table 1 (A & B).

```
# bma_BMA0106 Length: 467
# bma_BMA0106 Number of predicted TMHs:  1
# bma_BMA0106 Exp number of AAs in TMHs: 21.26107
# bma_BMA0106 Exp number, first 60 AAs:  21.26027
# bma_BMA0106 Total prob of N-in:        0.94522
# bma_BMA0106 POSSIBLE N-term signal sequence
bma_BMA0106     TMHMM2.0        inside      1     6
bma_BMA0106     TMHMM2.0        TMhelix     7     29
bma_BMA0106     TMHMM2.0        outside     30    467
```

Table 1 A

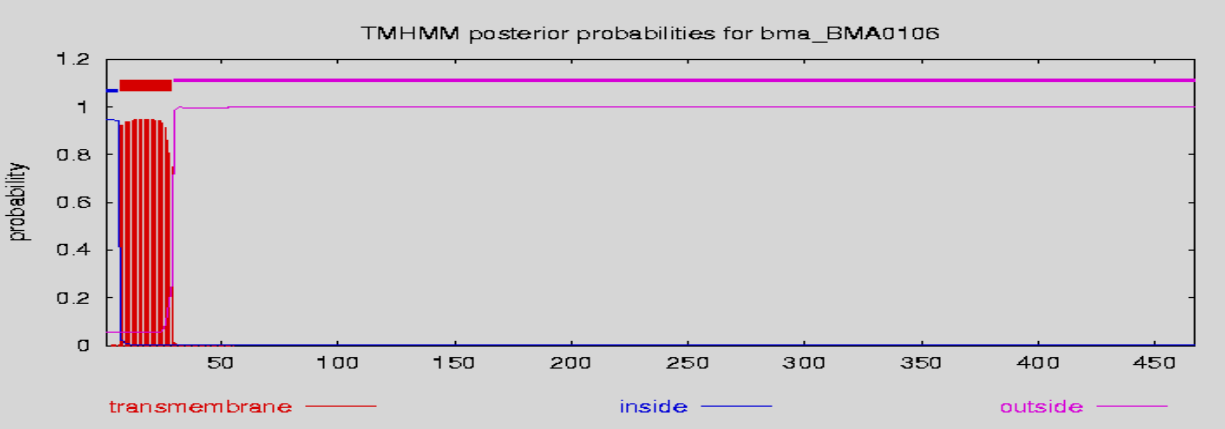

Fig 3(A) Graphical views of one trans-membrane helix predicted by TMHMM.

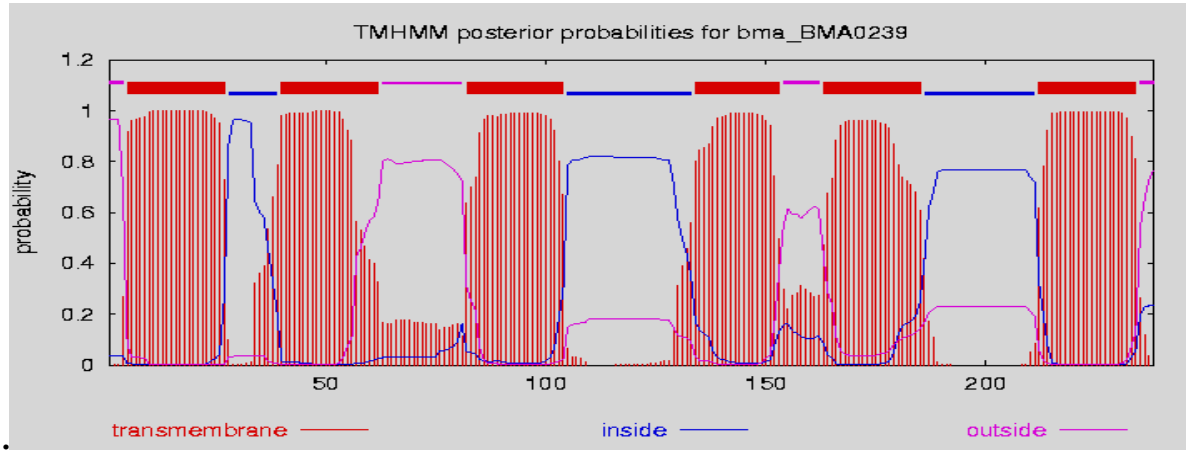

Fig 3 (B) Graphical views of more than two trans-membrane helices predicted by TMHMM

Table 1 B

```
# bma_BMA0239 Length: 238
# bma_BMA0239 Number of predicted TMHs:  6
# bma_BMA0239 Exp number of AAs in TMHs: 134.82599
# bma_BMA0239 Exp number, first 60 AAs:  44.51174
# bma_BMA0239 Total prob of N-in:        0.03488
# bma_BMA0239 POSSIBLE N-term signal sequence
bma_BMA0239     TMHMM2.0          outside     1     4
bma_BMA0239     TMHMM2.0          TMhelix     5    27
bma_BMA0239     TMHMM2.0          inside     28    39
bma_BMA0239     TMHMM2.0          TMhelix    40    62
bma_BMA0239     TMHMM2.0          outside    63    81
bma_BMA0239     TMHMM2.0          TMhelix    82   104
bma_BMA0239     TMHMM2.0          inside    105   133
bma_BMA0239     TMHMM2.0          TMhelix   134   153
bma_BMA0239     TMHMM2.0          outside   154   162
bma_BMA0239     TMHMM2.0          TMhelix   163   185
bma_BMA0239     TMHMM2.0          inside    186   211
bma_BMA0239     TMHMM2.0          TMhelix   212   234
bma_BMA0239     TMHMM2.0          outside   235   238
```

3.4) Fuzzpro:-

In Fuzzpro out of 3400, only 1 proteins predicted positive for the LPXTG pattern in which all have shown somewhat similar sequence patterns for the given LPXTG pattern as in Fig 4.

```
# Program: fuzzpro
# Rundate: Fri 24 Jun 2011 05:59:24
# Commandline: fuzzpro
#   -auto
#   -sequence
/geninf/prog/www/htdocs/tools/emboss/output/626382/.sequence
#   -pattern
"[LY]PX[TSA][GNAST]X(0,10){DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQNKRP}
{DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQNKRP}X(0,15)[DEQN
KRH]X(0,5)> "
# Sequence: BMA2937    from: 1   to: 287
# HitCount: 1
# Pattern_name Mismatch Pattern
# pattern1        0
[LY]PX[TSA][GNAST]X(0,10){DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQNKRP}
{DEQNKRP}{DEQNKRP}{DEQNKRP}{DEQNKRP}X(0,15)[DEQNKRH]X(0,5)>
 Start    End     Pattern_name        Mismatch Sequence
  259     287         pattern1        0 YPEASATVVFVIMAIVLLIRPAGLFGKER
# Total_sequences: 1   # Total_hitcount: 1
```

Fig 4 LPXTG based mismatch pattern predicted by Fuzzpro with several mismatches in the sequences.

3.5) BLASTP:-

Advanced BLASTP program with E-value threshold of 0.0001 helped to find out Streptococcus specific conserved vaccine leads obtained from four programs. BLASTP has reduced the vaccine lead number to acceptable total 104. These 104 leads were finally represented as vaccine candidates as they all qualified for conserved lipoproteins and cell wall anchored proteins which was required for vaccine success as in Table 2.

Table 2

KEGG NO:-	Protein
BMA0058	D-alanyl-D-alanine carboxypeptidase family protein
BMA0064	ABC transporter periplasmic substrate-binding protein
BMA0086	dsbA; thiol:disulfide interchange protein DsbA
BMA0091	peptide ABC transporter periplasmic peptide-binding protein
BMA0145	M48 family peptidase
BMA0147	hypothetical protein (A)
BMA0148	nitrogen regulation protein NtrY (A)
BMA0189	ubiB, aarF, yigQ, yigR; ubiquinone biosynthesis protein UbiB (EC:1.14.13.-)

3.6) HLA Pred:

The alleles considered in study were HLA-A2,HLA-A*0201,HLA- A*0202,HLA-A*0203,HLA-A*0205,HLA-DRB1*0101,HLA-DRB1*0102,HLA-DRB1*0301, HLA-DRB1*0305 .find out the **109** HLA binders.

Allele	Rank	Position	Sequence	Score	Prediction
HLA-A2 Threshold(3%)= 2.190	1	81	YLVFEALDA	6.340	Binder
HLA-A*0201 Threshold(3%)= 7.470	1	395	ILLMFSKKK	13.840	Binder
HLA-A*0202 Threshold(3%)= 9.130	1	300	YKAGQPIGT	14.280	Binder
HLA-A*0203 Threshold(3%)= 3.790	1	20	APAAAAAPT	9.560	Binder
HLA-A*0205 Threshold(3%)= 6.600	1	122	VSVHDLVYG	11.050	Binder

Allele	Rank	Position	Sequence	Score	Prediction
HLA-DRB1*0101 Threshold(3%)= 0.140	1	154	FVNMMNAEA	2.300	Binder
HLA-DRB1*0102 Threshold(3%)= 0.700	1	131	MIIQSGNDA	2.500	Binder
HLA-DRB1*0301 Threshold(3%)= 2.960	1	127	LVYGMIIQS	5.600	Binder
HLA-DRB1*0305 Threshold(3%)= 1.700	1	364	LVADGKTVA	4.600	Binder

4) **CONCLUSION:-**

With the increased sensitivity of algorithms in bio-programs along with available genome sequence information, bioinformatics helped to develop new methodology in the course of which the selection of vaccine candidates for *B.mallei* seems to be uncomplicated, and really this strategy enabled us to find out the 104 most probable *B.mallei* genome specific surface antigens. Whereas similar approach has developed vaccine for *S.sanguinis* SK36.This indicates that selection of any of the 104 vaccine leads may be a better choice as starting point for reverse vaccinology as per similar success recorded for lipoproteins and cell wall anchored proteins [4, 12,16,19]. The combined analysis by all five programs viz., LipoP, Fuzzpro, TMHMM, SignalP and BLASTP, read out probable conserved vaccine leads as important cell surface antigens. Previous antigen success suggested that ABC transporter family and some lipoproteins involved in vaccine development program were better candidates, in that concern ABC transporter proteins were reported by us could be the initial leads which may be implemented prior to other in vaccine biology [19]. The epitope prediction of antigens will allow us to develop subunit vaccine and in future may decide the success of the vaccine. The HLA Pred server allowed identification and prediction of peptides/regions from the antigenic sequence binding with HLA class 1 and/or class II alleles. The server identified the experimentally proven binders (available in MHCBN database) in query antigen sequence. The prediction of HLA binders (5 class I and 4 class II) in antigen sequence was based on quantitative matrices. This study may prove to be the better starting material for the reverse vaccinology of *B.mallei* and similar approach could be implemented for other organisms for searching probable vaccine leads.

5) **ACKNOWLEDGMENT:-**

I am immensely grateful thanks to Mr. Dilip G. Gore for their valuable guidance and intending all the possible help for the project. From bottom

of heart, thanks to C.N.Dipke Sir H.O.D. (Bioinformatics), and J. M. Saboo Sir Principal of Shankarlal Khandelwal College Akola.

6) **REFERENCES:-**

1]Ariel N, Zvi A, Grosfeld H, Gat O, Inbar Y. (2002). Search for potential vaccine candidate open reading frames in the *Bacillus anthracis* virulence plasmid pXO1: *In silico* and in vitro screening. *Infect Immun.* 70: 6817–6827.

2] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 340: 783–795.

3] Dilip G and Reecha P. (2011). *In silico* identification of cell surface antigens in *Neisseria meningitides. Biomirror* vol.2 1-5.

4] Etz H, Minh DB, Henics T, Dryla A, Winkler B.(2002). Identification of in vivo expressed vaccine candidate antigens from *Staphylococcus aureus*. *Proc Natl Acad Sci U S A* 99: 6573–6578.

5] Fiona M. Roche, Ruth Massey, Sharon J. Peacock, Nicholas P. J. Day, Livia Visai, Pietro Speziale, Alex Lam, Mark Pallen and Timothy J. Foster. (2003). Characterization of novel LPXTGcontaining proteins of *Staphylococcus aureus* identified from genome sequences. *Microbiology* 149, 6573-6578.

6] Fong, I.W., and Alibek, K. (2005). Bioterrorism and infectious Agents: A New Dilemma for the 21st Century. Springer, 99 – 145.

7] Gan W, Zhao G, Xu H, Wu W, Du W, *et al*. (2010). Reverse vaccinology approach identify an *Echinococcus granulosus* tegumental membrane protein enolase as vaccine candidate. *Parasitol Res.* 106: 873–882.

8] Godoy D, Randle G, Simpson AJ, *et al.* (2003). "Multilocus Sequence Typing and Evolutionary Relationships among the Causative Agents of Melioidosis and Glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*". *J Clin Microbiol* 41: 2068–2079.

9] Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H, et al. (2003). Prediction of lipoprotein signal peptides in Gram negative bacteria. *Protein Sci.* 12:1652–1662.

10] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J MolBiol.* 305: 567– 580.

11] Liu L, Cheng G, Wang C, Pan X, Cong Y, *et al*. (2009). Identification and experimental verification of protective antigens against *Streptococcus suis* serotype 2 based on genome sequence analysis. *Curr Microbiol.* 58: 11–17.

12] Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, *et al*. (2005). Identification of a universal Group B Streptococcus vaccine by multiple genome screen. *Science* 309: 148–150.

13] Montigiani S, Falugi F, Scarselli M, Finco O, Petracca R, *et al*. (2002) Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. Infect Immun. 70: 368–379.

14] Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, *et al*. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287: 1816– 1820.

15] PLoS Pathog.; Hwang, J; Yi, H; Ulrich, RL; Yu, Y; Nierman, WC; Kim, HS; Ochman, Howard (2010). "The early stage of bacterial genome-reductive evolution in the host". *PLoS pathogens* 6 (5): e1000922. doi:10.1371/journal.ppat.1000922. PMC 2877748.PMID 20523904.

16] Ross BC, Czajkowski L, Hocking D, Margetts M, Webb E, *et al.* (2001). Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. *Vaccine* 19: 4135–4142.

17] Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa Agrawala, Aleksandr Morgulis, Alejandro A. Schaffer, and Yi-Kuo Yu. (2005). Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* 272:5101-5109.

18] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang Zhang, Webb Miller, and David J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* 25:3389-3402.

19] Xiuchun Ge, Todd Kitten, Cindy L. Munro, Daniel H. Conrad, Ping Xu. (2010). Pooled Protein Immunization for Identification of Cell Surface Antigens in *Streptococcus sanguinis* Vol. 5.Issue 7e11666.

20] Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, *et al*. (2004).The genome of *Cryptosporidium hominis. Nature* 431: 1107–111.

21] Whitlock, G.C., Estes, D.M., and Torres, A.G. (2007). Glanders: off to the races with Burkholderia mallei. FEMS Microbiology Letters, 277(2), 115 – 122.