International Journal of Multidisciplinary Education and Research
www.multieducationjournal.com
ISSN: 2455-4588
Received: 10-03-2025, Accepted: 09-04-2025, Published: 25-04-2025
Volume 10, Issue 2, 2025, Page No. 16-21

In silico prediction of *Mycobacterium tuberculosis* putative vaccine candidates

Prashant Sopanrao Telgad, Dr. Umesh P Mogle
Department of Botany and Biotechnology, J.E.S. College, Jalna, Maharashtra, India

Abstract
Mycobacterium cell envelope is a house of variety of biologically active molecules of different shapes. The proteins associated with cell wall are the key determinants of pathogenesis & immunogenecity.
The whole genome sequence *of Mycobacterium tuberculosis,* which is 44,03,837 nt in length was downloaded from NCBI site. The ORFs were analyzed by PSORTb v2.0 for the prediction of the subcellular localization of protein. The non-cytoplasmic sequences were used for the further analysis.
The prediction of signal peptide and transmembrane helics was, done by using the PHOBIUS. The identification of the putative virulence factors was carried by searching for homological sequences in the VFDB (Virulence factor database.) Pfam Database was used to find out most potent putative vaccine candidates.
These proteins hold the promise of being potential vaccine candidates for developing an effective vaccine against this disease.

Keywords: Mycobacterium tuberculosis, vaccine candidates, subcellular localization, virulence factors, signal peptide

Introduction
TB is one of the most leading infectious disease killing people around the world. Among all infectious diseases that affect humans, tuberculosis (TB) remains the deadliest. It was first described on March 24,1882 by Robert Koch.
M. tuberculosis is an obligate aerobe (Gram positive bacterium). While mycobacteria do not seem to fit the Gram-positive category from an empirical standpoint (i.e. they do not retain the crystal violet stain), they are classified as such due to their lack of an outer cell membrane.
M.TB. Complexes are always found in the well-aerated upper lobes of the lungs. The bacterium is a facultative intracellular parasite, usually of macrophages, and has a slow generation time, 15-20 hours, a physiological characteristic that may contribute to its virulence. (Todar's Online Textbook of Bacteriology 2005) [6]
At present, epidemiologists estimate that one-third of the world population is infected with tubercle bacilli, which is responsible for 8-10 million new cases of TB and 3 million deaths annually through out the world.
Approximately 95% of new cases and 98% of deaths occur in developing nations, the reason being few resources available to ensure proper treatment and where human immunodeficiency virus (HIV) infections are common.
Mycobacterium shows a high degree of intrinsic resistance to most antibiotics and chemotherapeutic agents due to low permeability of its cell wall. Nevertheless, the cell wall barrier alone cannot produce significant levels of drug resistance.
Identification of extra cellular ORFs for *Mycobacterium tuberculosis* H37Rv, Screening for protein motifs such as signal peptides, Transmembrane domains and membrane anchoring domains, Determination of Antigenic signal peptides, Prediction of Virulence Factors, Prediction of Putative vaccine candidates.

Materials and Methods
1. Computer System (Windows Xp).
2. Internet Connection.
3. Online Bioinformatics Softwares.
4. MTB Genome Sequence (Downloaded from NCBI).

The sequence of *Mycobacterium tuberculosis* was accessed from NCBI website. Translation product of Genome sequence was further used for analysis using the following online softwares.

1. **PSORTb v.2.0:** An expanded database of proteins of known localization and new modules using frequent subsequence-based support vector machines was introduced in to PSORTb v.2.0. The program attains a precision of 96% for Gram positive and Gram-negative bacteria and predictive coverage comparable to other tools for whole proteome analysis. (Gardy J.L., Laird M.R., M. Ester *et al.* 2005) [2]

2. **Phobius:** Phobius a combined transmembrane protein topology and signal peptide predictor. The predictor is based on a Hidden Markov Model (HMM) that models the different sequence regions of signal peptide and the different regions of a transmembrane protein in a series of interconnected state. (Lukas Kall, Anders Krogh, *et al*. 2004) [3]

3. **VFDB:** The currently released VFDB contains cumulative information of VFs for 16 important bacterial pathogens, virulence-associated genes, protein structural features, functions, mechanism and important literatures. (Lihong Chen, Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen & Qi Jin. 2005) [4]

4. **Pfam:** Pfam is a database of protein families that currently contains 7973 entries. Pfam Database was used to find out most potent putative vaccine candidates. (Robert D. Finn, Jaina Mistry *et al*. 2006) [5]

Results
1. PSORTb
2. PHOBIUS
3. VFDB
4. Pfam

1) PSORTb v2.0

1) Sequence ID	701687…702637

MIRRRGARMAALLAAAALALTACAGSDDKGEPDDGGDRGASLATTSDADWKPVADI
LGRTGKLNDGSVYKIGFARSDLSVQTKGVTVAPALSLGSWVAFARTPDGQTMLMGDL
VVTEDELASVTDAVQAGGLQQTALHKHLLEQSPPIWWTHIAGHGDAADLARAVRSAL
DATDTPPPASATSGQTSLDLDTAAIDEALGRSGTIAGGVYKFFIARRDPVTMSGMLIPPS
MGLATALNFQPTGNGRAAINGDFVMTAAEVQDVVQALRGGGIDIVAIHNHGFDEQPR
LFYMHFWAENDAVALARTLRAAVDATAAR

SeqID : 701687..702637

Analysis Report:

CMSVM+	Unknown	[No details]
CWSVM+	Unknown	[No details]
CytoSVM+	Unknown	[No details]
ECSVM+	Unknown	[No details]
HMMTOP+	Unknown	[1 internal helix found]
Motif+	Unknown	[No motifs found]
Profile+	Unknown	[No matches to profiles found]
SCL-BLAST+	Unknown	[No matches against database]
SCL-BLASTe+	Unknown	[No matches against database]
Signal+	Non-Cytoplasmic	[Signal peptide detected]

Localization Scores:

Cytoplasmic	0.00
CytoplasmicMembrane	3.33
Cellwall	3.33
Extracellular	3.33

Final Prediction:

Unknown

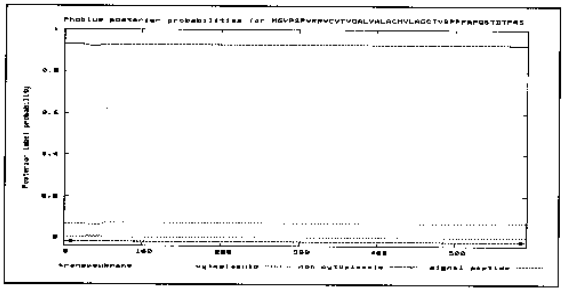
2) Sequence ID: -	**complement(3456605..3457525)**
3) Sequence ID: -	**complement(3284945..3285646)**
4) Sequence ID: -	**complement(2176312..2176791)**
5) Sequence ID: -	**2015160..2016680**
6) Sequence ID :-	**1542236..1543021**
7) Sequence ID:-	**complement(1418500..1419234)**
8) Sequence ID:-	**complement(651859..652197)**
9) Sequence ID:-	**3761335..3761724**
10) Sequence ID:-	**complement(1419447..1419794)**
11) Sequence ID:-	**1295640..1297547**
12) Sequence ID:-	**complement(2288860..2290179)**

2) PHOBIUS: -

1) 1295640-1297547

MGVPSPVRRVCVTVGALVALACMVLAGCTVSPPPAPQSTDTPRS

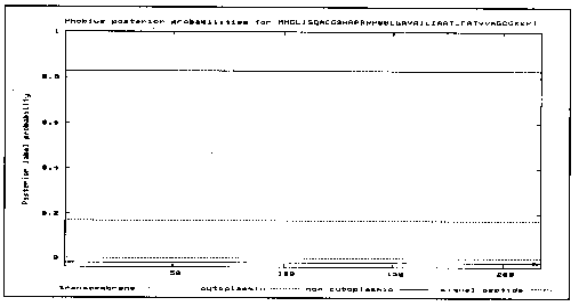
```
ID   MGVPSPVRRVCVTVGALVALACMVLAGCTVSPPPAPQSTDTPRS
FT   TOPO_DOM      1     591       NON CYTOPLASMIC.
```



2) 1542236-1543021

MNGLISQACGSHRPRRPSSLGAVAILIAATLFATVVAGCGKKPT

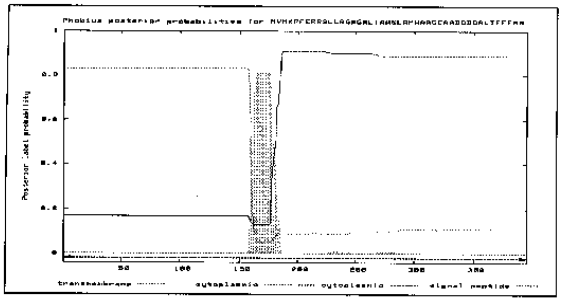
```
ID   MNGLISQACGSHRPRRPSSLGAVAILIAATLFATVVAGCGKKPT
FT   TOPO_DOM      1     217       NON CYTOPLASMIC.
```



3) 2288860-2290179

MVNKPFERRSLLRGAGALTAASLAPWAAGCAADDDDALTFFFAA

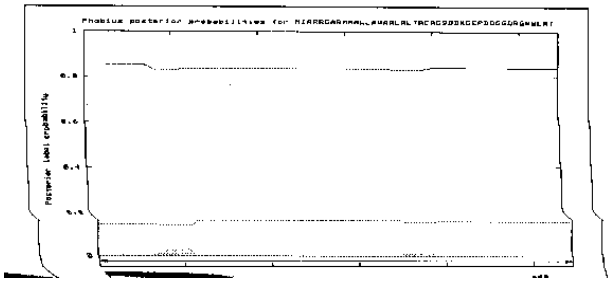
```
ID   MVNKPFERRSLLRGAGALTAASLAPWAAGCAADDDDALTFFFAA
FT   TOPO_DOM      1     395       NON CYTOPLASMIC.
```



4) 701687-702637

MIRRRGARMAALLAAAALALTACAGSDDKGEPDDGGDRGASLAT

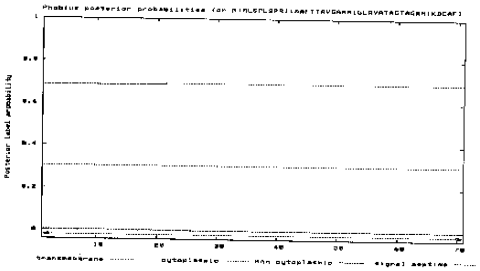
```
ID   MIRRRGARMAALLAAAALALTACAGSDDKGEPDDGGDRGASLAT
FT   TOPO_DOM      1     272       NON CYTOPLASMIC.
```



5) 1419447….. 1419794

Prediction of MTMLSPLSPRIIAAFTTAVGAAAIGLAVATAGTAGANTKDEAFI

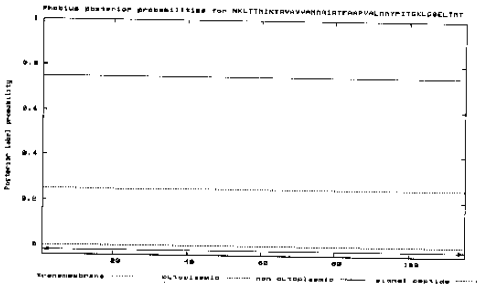
```
ID   MTMLSPLSPRIIAAFTTAVGAAAIGLAVATAGTAGANTKDEAFI
FT   TOPO_DOM    1   71    NON CYTOPLASMIC.
//
```



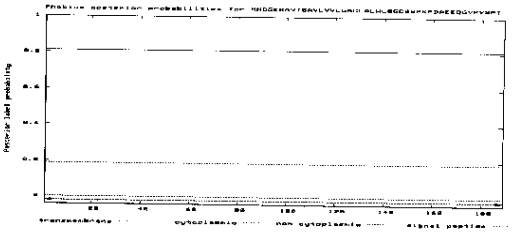
6) 2176312….,2176791

Prediction of MKLTTMIKTAVAVVAMAAIATFAAPVALAAYPITGKLGSELTMT

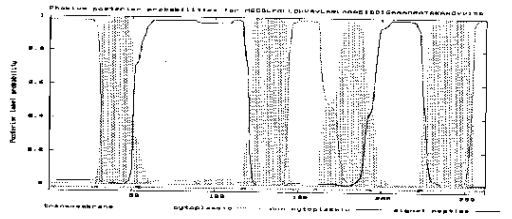
```
ID   MKLTTMIKTAVAVVAMAAIATFAAPVALAAYPITGKLGSELTMT
FT   TOPO_DOM    1   115   NON CYTOPLASMIC.
//
```



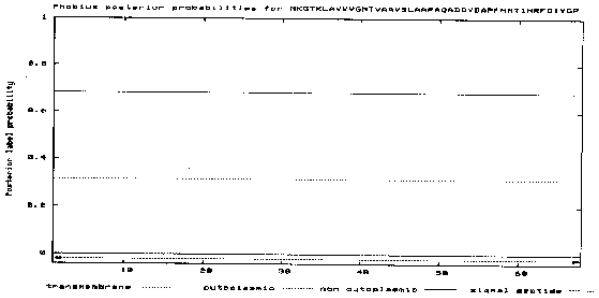
7) 3284945…..3285646
Prediction of
MNDGKRAVTSAVLVVLGACLALWLSGCSSPKPDAEEQGVPVSPT
ID MNDGKRAVTSAVLVVLGACLALWLSGCSSPKPDAEEQGVPVSPT
FT TOPO_DOM 1 189 NON CYTOPLASMIC.
//



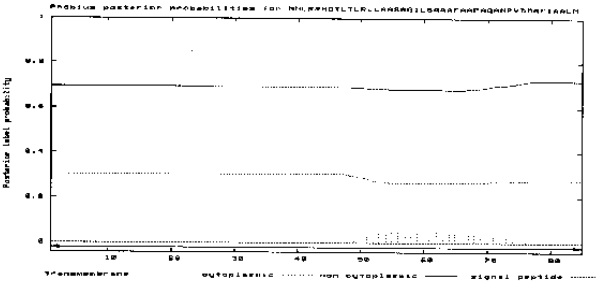
8) 3456605…3457525
Prediction of MSGGLFGLLDHVAVLARLAAASIDDIGAAAGRATAKAAGVVIDD
ID MSGGLFGLLDHVAVLARLAAASIDDIGAAAGRATAKAAGVVIDD
FT TOPO_DOM 1 31 CYTOPLASMIC.
FT TRANSMEM 32 50
FT TOPO_DOM 51 120 NON CYTOPLASMIC.
FT TRANSMEM 121 143
FT TOPO_DOM 144 163 CYTOPLASMIC.
FT TRANSMEM 164 189
FT TOPO_DOM 190 226 NON CYTOPLASMIC.
FT TRANSMEM 227 254
FT TOPO_DOM 255 262 CYTOPLASMIC.
//



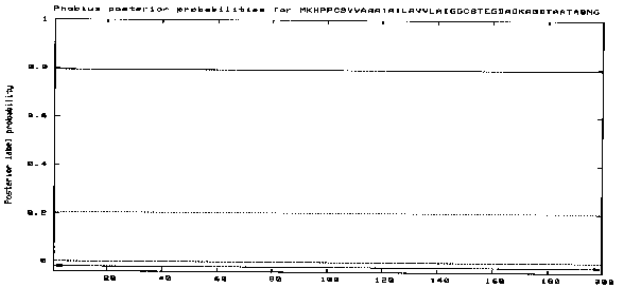
9) 651859….652197
Prediction of MKGTKLAVVVGMTVAAVSLAAPAQADDYDAPFNNTIHRFGIYGP
ID MKGTKLAVVVGMTVAAVSLAAPAQADDYDAPFNNTIHRFGIYGP
FT TOPO_DOM 1 68 NON CYTOPLASMIC.
//



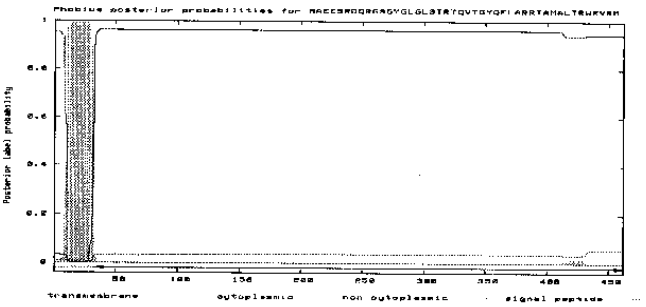
10) 3761335….3761724
Prediction of MNLRRHQTLTLRLLAASAGILSAAAFAAPAQANPVDDAFIAALN
ID MNLRRHQTLTLRLLAASAGILSAAAFAAPAQANPVDDAFIAALN
FT TOPO_DOM 1 85 NON CYTOPLASMIC.
//



11) 1418500 ……1419234
Prediction of
MKHPPCSVVAAATAILAVVLAIGGCSTEGDAGKASDTAATASNG
ID MKHPPCSVVAAATAILAVVLAIGGCSTEGDAGKASDTAATASNG
FT TOPO_DOM 1 200 NON CYTOPLASMIC.



12) 2015160…2016680
Prediction of
MAEESRGQRGSGYGLGLSTRTQVTGYQFLARRTAMALTRWRVRM
ID MAEESRGQRGSGYGLGLSTRTQVTGYQFLARRTAMALTRWRVRM
FT TOPO_DOM 1 11 CYTOPLASMIC.
FT TRANSMEM 12 32
FT TOPO_DOM 33 462 NON CYTOPLASMIC.



3) VFDB: -

1) 1419447…. 1419794

BLASTP 2.2.9

Database: Virulence Factors of Pathogenic Bacteria

2050 sequences; 747,743 total letters

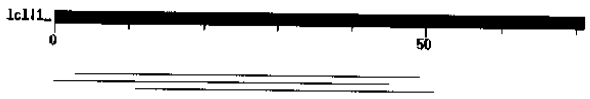
Query= MTMLSPLSPRIIAAFTTAVGAAAIGLAVATAGTAGANTKDEAFI

(71 letters)

Top of Form 1

Distribution of 3 Blast Hits on the Query Sequence

</VFs/blast/docs/newoptions.html>



Score E

Sequences producing significant alignments: (bits) Value

VFG0545 <../gene.cgi?GeneID=VFG0545> sipC - cell invasion protein [Salmonella enterica (... 27 0.11

VFG1448 <../gene.cgi?GeneID=VFG1448> kpsD - KpsD [Escherichia coli] 23 2.1

VFG1439 <../gene.cgi?GeneID=VFG1439> csnA - CS20 fimbria major subunit protein [Escheric... 22 3.5

>VFG0545 <../gene.cgi?GeneID=VFG0545> sipC - cell invasion protein [Salmonella enterica (serovar typhimurium) LT2]

Length = 409

Score = 27.3 bits (59), Expect = 0.11

Identities = 10/47 (21%), Positives = 26/47 (55%)

2) 651859….652197

BLASTP 2.2.9

Database: Virulence Factors of Pathogenic Bacteria
2050 sequences; 747,743 total letters

Query= MKGTKLAVVVGMTVAAVSLAAPAQADDYDAPFNNTIHRFGIYGP
(68 letters)

Distribution of 3 Blast Hits on the Query Sequence
</VFs/blast/docs/newoptions.html>

Score E

Sequences producing significant alignments: (bits) Value

VFG0268 <../gene.cgi?GeneID=VFG0268> hpuB - haemoglobin-haptoglobin-utilization protein ... 25 0.71

VFG1132 <../gene.cgi?GeneID=VFG1132> int - hypothetical protein [Vibrio cholerae N16961] 24 0.93

VFG1694 <../gene.cgi?GeneID=VFG1694> c3557 - ShiA homolog [Escherichia coli CFT073] 21 7.9

>VFG0268 <../gene.cgi?GeneID=VFG0268> hpuB - haemoglobin-haptoglobin-utilization protein [Neisseria meningitidis Z2491]
Length = 810

Score = 24.6 bits (52), Expect = 0.71
Identities = 11/37 (29%), Positives = 18/37 (48%)

3)2015160……2016680

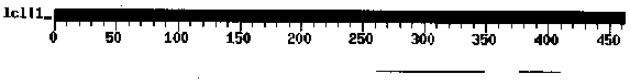
BLASTP 2.2.9

Database: Virulence Factors of Pathogenic Bacteria
2050 sequences; 747,743 total letters

Query= MAEESRGQRGSGYGLGLSTRTQVTGYQFLARRTAMALTRWRVRM
(462 letters)

Top of Form 1

Distribution of 2 Blast Hits on the Query Sequence
</VFs/blast/docs/newoptions.html>



Sequences producing significant alignments bits(value)

VFG2121 <../gene.cgi?GeneID=VFG2121> nleA - NleA encoded by cryptic prophage CP-933P [Es... 27 2.2

VFG0068 <../gene.cgi?GeneID=VFG0068> actA - actin-assembly inducing protein precursor [L... 25 8.4

>VFG2121 <../gene.cgi?GeneID=VFG2121> nleA - NleA encoded by cryptic prophage CP-933P [Escherichia coli O157:H7 EDL933]
Length = 441

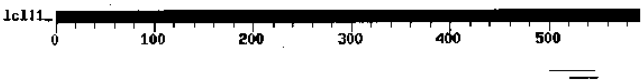
Score = 26.6 bits (57), Expect = 2.2
Identities = 13/33 (39%), Positives = 16/33 (48%), Gaps = 3/33 (9%)

4) 1295640……1297547
BLASTP 2.2.9

Database: Virulence Factors of Pathogenic Bacteria
2050 sequences; 747,743 total letters

Query= MGVPSPVRRVCVTVGALVALACMVLAGCTVSPPPAPQSTDTPRS
(591 letters)

Distribution of 2 Blast Hits on the Query Sequence
</VFs/blast/docs/newoptions.html>



Score E

Sequences producing significant alignments: (bits) Value

VFG1930 <../gene.cgi?GeneID=VFG1930> flgE2 - flagellar hook subunit protein [Campylobact... 26 5.0

VFG1933 <../gene.cgi?GeneID=VFG1933> cdtB - cytolethal distending toxin [Campylobacter j... 25 8.5

>VFG1930 <../gene.cgi?GeneID=VFG1930> flgE2 - flagellar hook subunit protein [Campylobacter jejuni NCTC 11168]
Length = 865

Score = 25.8 bits (55), Expect = 5.0
Identities = 17/47 (36%), Positives = 20/47 (42%), Gaps = 2/47 (4%)

5) 2288860…2290179
BLASTP 2.2.9

Database: Virulence Factors of Pathogenic Bacteria
2050 sequences; 747,743 total letters

Query= MVNKPFERRSLLRGAGALTAASLAPWAAGCAADDDDALTFFFAA
(395 letters)

Distribution of 3 Blast Hits on the Query Sequence
</VFs/blast/docs/newoptions.html>

Score E

Sequences producing significant alignments: (bits) Value

VFG1205 <../gene.cgi?GeneID=VFG1205> fbpA - iron(III) ABC transporter, periplasmic bindi... 29 0.37

VFG0262 <../gene.cgi?GeneID=VFG0262> fbpA - iron(III) ABC transporter, periplasmic bindi... 29 0.37

VFG1694 <../gene.cgi?GeneID=VFG1694> c3557 - ShiA homolog [Escherichia coli CFT073] 24 9.1

>VFG1205 <../gene.cgi?GeneID=VFG1205> fbpA - iron(III) ABC transporter, periplasmic binding protein [Neisseria meningitidis MC58 (serogroup B)]
Length = 331

Score = 28.9 bits (63), Expect = 0.37
Identities = 29/119 (24%), Positives = 49/119 (41%), Gaps = 12/119 (10%)

4) Pfam :-

1) 3284945-3285646. Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF1396	3	188	1	210	389.5	4.6e-114	glocal	Protein of unknown function (DUF1396)

2) 3456605—3457525 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF808	1	252	1	318	464.2	1.6e-136	glocal	Protein of unknown function (DUF808)

3) 2288860—2290179 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! SBP_bac_1	5	307	1	399	133.0	8.2e-37	glocal	Bacterial extracellular solute-binding protein

4) 2176312—2176791. Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF1942	1	113	1	130	248.7	1.2e-71	glocal	Domain of unknown function (DUF1942)

5) 2015160—2016680. Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF690	1	453	1	518	933.4	9.6e-278	glocal	Protein of unknown function (DUF690)

6) 1542236--1543021 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF1396	13	212	1	210	433.0	4e-127	glocal	Protein of unknown function (DUF1396)

7) 1419447--1419794 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF732	1	70	1	124	71.9	2.1e-18	glocal	Protein of unknown function (DUF732)

8) 1418500—1419234 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF1396	1	196	1	210	425.4	7.5e-125	glocal	Protein of unknown function (DUF1396)

Content-type: text/plain Could not open template image '/2bands.gif'.

9) 1295640—1297547 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! SBP_bac_5	61	501	1	508	217.5	3.1e-62	glocal	Bacterial extracellular solute-binding proteins, family 5 Middle

10)701687-702637 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF1529	6	129	1	127	259.5	7e-75	glocal	Domain of Unknown Function (DUF1259)
!! DUF1529	148	269	1	127	263.0	6.2e-76	glocal	Domain of Unknown Function (DUF1259)

11) 651859-652197 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF732	1	62	1	124	29.0	7.1e-06	glocal	Protein of unknown function (DUF732)

12)3761335—3761724 Pfam 21.0 (Janelia Farm)

Pfam HMM search results, glocal+local alignments merged (Pfam_ls+Pfam_fs)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! DUF732	1	68	1	124	67.4	4.5e-17	glocal	Protein of unknown function (DUF732)

Conclusion

The genome of *Mycobacterium tuberculosis* has been completely sequenced and approximately 16% of its total genome contains certain unknown proteins, these proteins might be involved in specific Mycobacterial function, and can be targeted as potential drug targets and as antigens in vaccines.

ORFs were analyzed by PSORTb v2.0 to predict sub-cellular localization of proteins, out of which non-cytoplasmic sequences were further analyzed to predict Signal Peptide & Transmembrane helics with use of PHOBIUS software.

Protein sequences showing Signal Peptides & Transmembrane helics were further used for detecting virulence factors from using VFDB software. Protein sequences showing greater virulency were used to find out most potent putative vaccine candidates using Pfam database.

The proteins which are obtained by PSORTb v2.0, PHOBIUS, VFDB and finally by using Pfam database hold the promise of being potential vaccine candidates for developing an effective vaccine against Tuberculosis disease.

The repeated use of vaccines such as BCG do not show boosting effect throughout the life and even *Mycobacterium tuberculosis* have adapted itself to resist current drugs. So, there was a need to find out the most putative vaccine candidate for this dreadful disease. (B.R. Bloom & Fine, P.E.M. 1994; Brennan P.J., Nikadio H. 1995; Jarlier V., Nikaido H. 1994) [8, 9]

References

1. www.niaid.nih.gov./Dir/labs/lhd/barry.htm.
2. Gardy JL, Laird MR, Ester M, *et al*,2005 PSORTb V2.0: Expanded prediction of bacterial protein sub cellular localization and in sights gained from comparative protein analysis. Bioinformatics,2005:21:617-23.
3. Kall L, Krogh A, *et al*,2004 A combined transmembrane topology and signal peptide prediction method. J Mol Biol,2004:338:1027-36.
4. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q,2005 VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res,2005:33:D325-8.
5. Finn RD, Mistry J, *et al*,2006 Pfam: clans, web tools and services. Nucleic Acids Res,2006:33:D247-51.
6. Todar's Online Textbook of Bacteriology. Kenneth Todar, University of Wisconsin – Madison Department of Bacteriology, 2005.
7. Bloom BR, Fine PEM. The BCG experience: implications for future vaccines against tuberculosis. In: Bloom BR, editor. Tuberculosis: Pathogenesis, Protection and Control. ASM Press, 1994.
8. Brennan PJ, Nikadio H. 1995. The envelope of mycobacteria. Annu Rev Biochem,1995:64:29-63.
9. Jarlier V, Nikaido H. 1994. Mycobacterial cell wall: Structure and role in natural resistance to antibiotics. FEMS Microbiol Lett,1994:123:11-8.