Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Structural and Functional Annotation of Uncharacterized Protein in Bacillus anthracis A2012

Sharma Nikita[1*] and Goyal Anchal[2]

[1*]Assistant professor, Department of Biotechnology and Bioinformatics (Guru Nanak Girls College, Model Town, Ludhiana, Punjab) Pin code: - 141002
E-mail ID:- niksshrma3@gmail.com Mb:-9915973535
[2]Research Scholar, Department of Biotechnology and Bioinformatics (Guru Nanak Girls College, Model Town, Ludhiana, Punjab) Pin code: - 141002
E-mail ID:- goyalanchal081@gmail.com Mb:-7347200171

ABSTRACT

An extensive interpretation of *Bacillus anthracis* A2012 proteomics data showed that it contains hypothetical proteins (HPs), i.e. proteins of unknown function. *Bacillus anthracis* A2012 causes **Anthrax** which is of great concern due to its outbreaks across globe that may lead to death and used as **bioterrorism** agent. Computational analysis of 'Hypothetical' proteins provides general prediction of structure and function of an organism to a more profound understanding of their machinery. The bacterium consists of 2 plasmids, **pX01** and **pX02**. In our current work, the sequence of 129 HPs from pX01 plasmid and 59 HPs from pX02 plasmid were collected from NCBI database. Several bioinformatics tool for identification of: **motif and domain search, transmembrane analysis, membrane protein topology, sequence similarity, protein solubility, physical and chemical parameters, protein localization, fold recognition and Ramachandran plot analysis** were run to annotate HPs. It is conferred that these HPs belong to various classes of proteins such as enzymes, transporters, receptors, signal transducers and other proteins. From these HPs, **4** hypothetical proteins from both plasmids (pX01 and pX02) are **identified** with proper structure and function. Among these 4 proteins, **ONE** protein is **best annotated** with having confidence score **99%**. This research aims to produce a better understanding of mechanism of drug discovery for treatment of *Bacillus anthracis* infections.

KEYWORDS: Annotation, Anthrax, *Bacillus anthracis*, Hypothetical proteins, Plasmids.

*Corresponding author:

Sharma Nikita

Assistant professor

Department of Biotechnology and Bioinformatics

(Guru Nanak Girls College, Model Town, Ludhiana, Punjab)

Pin code: - 141002

E-mail ID:- niksshrma3@gmail.com Mb:-9915973535

INTRODUCTION

Bacillus anthracis is a gram-positive, aerobic, rod-shaped, non-motile bacteria having spore bearing bacillus with 1-1.5×3-10-µm size and the order Bacillales with single chromosomal circular DNA[1]. This research deals with *Bacillus anthracis* A2012, one of the descendents of *Bacillus anthracis* Ames Ancestor strain. The potential use of B. anthracis spores is in **bioterrorism** and warfare agent [2]. The disease is most prevalent among cattle, sheep, horses and goats. Animals contract the disease while grazing and the majority of naturally occurring human cases of Anthrax are due to either agricultural or industrial exposure. The bacterium exists in the environment as a spore and can remain viable in the soil for decades. Spores germinate inside the macrophage within an organism to produce the virulent vegetative forms that replicate and eventually kill the host. Products like meat and hides from infected animals serve as a reservoir for human disease [3]. The major virulence factors of B anthracis are encoded on two virulence plasmids **pXO1** and **pXO2**. The tri-toxin bearing plasmid pXO1 is 184.5 kbp in size and encodes for three toxins, lethal factor, oedema factor and protective antigen that allows entry of the toxin into the host cell. The smaller capsule bearing plasmid pXO2 is 95.3 kbp in size and encodes three genes i.e. cap B, cap C, and cap A, involved in the synthesis of the polyglutamyl capsule that inhibits host phagocytosis of the vegetative form of B anthracis. Both plasmids are necessary for full virulence, loss of either result in an attenuated strain [4].

Anthrax is caused by ingestion, inhalation or cutaneous forms of B. anthracis spores. **Inhalational** anthrax occurs when spores enter the human body, gets deposited into alveolar spaces. **Cutaneous** anthrax is the most common, results when spores settle into abrasions on the skin. **Gastrointestinal** anthrax results after ingestion of raw infected food containing a large number of vegetative bacteria[5]. **Injectional** anthrax was developed under soft tissue resulting from a subcutaneous drug injection. Spores germinate, multiply and disseminate throughout their host, causing septicemia, toxemia, and meningitis. Intensive antibiotic therapy is often ineffective against the lethal development of anthrax [6]. The disease has a global distribution including India but incidence in livestock and humans varies with local ecology, implementation of control strategies from animals to humans [7]. In 1979, the accidental release of aerosolized anthrax spores from a military microbiology facility in Sverdlovsk resulted in nearby 68 deaths. 22 cases of bioterrorism-related anthrax were detected in the United States in Oct 2001[8].

Antibiotic treatment of anthrax is vital in early stages, as delay decreases a victim’s chance for survival. Ciprofloxacin, levofloxacin, doxycycline, and penicillin are currently the only FDA-approved antibiotics for the treatment of anthrax. BioThrax is the only anthrax vaccine licensed by the FDA. Anthrax Immune Globulin (AIG) antitoxin is derived from the plasma of individuals previously immunized with the anthrax vaccine [9].

MATERIALS AND METHODS

Retrieval of Sequence for Analysis (NCBI): The genome of *Bacillus anthracis A2012* was retrieved in fasta format from NCBI (https://www.ncbi.nlm.nih.gov/) having Accession no. AE011190.1 for pX01 plasmid and Accession no. AE011191.1 for pX02 plasmid [10].

Motif identification (MOTIF FINDER): Motif finder (https://www.genome.jp/tools/motif/) classify a protein by searching for a protein query sequence against Motif libraries, align the sequence by PROSITE or HMMER search with a profile against protein sequence databases, search a protein sequence pattern against sequence databases, generate a profile from a set of multiple aligned sequences[11]. The goal of motif finder is the detection of novel, generate a profile from a set of multiple aligned sequences , Motifs carry out and regulate various functions, and the presence of specific motifs may help to classify a protein[12.]

Detection of conserved Domains (CDD): Conserved Domain Database is a protein annotation (https://www.ncbi.nlm.nih.gov/Structure/cdd/docs/cdd_search.html) resource. CDD record the location of functional motifs on protein domain models and map on protein sequences and facilitate the interpretation of sequence conservation and variation[13.]

Trans membrane helices prediction (TMHMM): Prediction of transmembrane helices in (http://www.cbs.dtu.dk/services/TMHMM/) integral membrane proteins. It is based on HMM approach and predict the full topology of the protein. It incorporates hydrophobicity, charge bias, helix lengths and grammatical constraints into the model. The resultant graph shows vertical red lines beneath the curves indicate portions of the sequences that match a TMHMM and are likely to enter or cross a membrane. The blue line represents the probability that a given portion of sequence lies inside the cytoplasm and the pink line represents the probability of being external to the plasma membrane or outer membrane [14, 15].

Membrane protein topology analysis (OCTOPUS): OCTOPUS (http://octopus.cbr.su.se/) represents the correct topologies for an uncharacterized sequence. It combines ANN-predicted residue scores with an HMM-based global prediction algorithm. This methods defines residues to be situated in the membrane (M), on the inside (cytoplasmic side, i) or on the outside (non-cytoplasmic side, o) [16.]

Sequence Similarity Identification (BLAST): Basic Local Alignment Search Tool is a sequence similarity search program (https://blast.ncbi.nlm.nih.gov/Blast.cgi) . BLAST uses a scoring matrix BLOSUM, block substitution matrix or PAM, percent accepted mutation to determine all high-scoring matching words from the database for each word in the query sequence [17].

Solubility Analysis in Proteins (SOSUI): SOSUI is used (http://harrier.nagahama-i-bio.ac.jp/sosui/sosui_submit.html) for discrimination of membrane proteins and soluble ones. Four physicochemical parameters are present- the hydropathy index of Kyte and Doolittle, an amphiphilicity index, an index of amino acid charges, and the length of each sequence [18].

Computation of Physical and Chemical Parameters (PROTPARAM): ProtParam (https://web.expasy.org/protparam/) computes various physicochemical properties from a protein sequence. The parameters computed are molecular weight, theoretical pI, amino acid and atomic compositions, extinction coefficient, half-life, instability index, aliphatic index, GRAVY value. The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell. The instability index provides an estimate of the stability of your protein in a test tube. The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains as alanine, valine, isoleucine, and leucine [19, 20].

Prediction of Subcellular Localization (PSORTb): PSORTb (https://www.psort.org/psortb/) is bacterial protein subcellular localization (SCL) tool. It predicts results for five major localizations for Gram-negative bacteria like cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular and four localizations for Gram-positive bacteria-cytoplasmic, cytoplasmic membrane, cell wall and extracellular [21].

Protein Fold Recognition (PHYRE2): **Phyre2** (http://www.sbg.bio.ic.ac.uk/phyre2/) is a tool to predict and analyze protein structure, function and mutations. Phyre2 uses advanced remote homology detection methods to build 3D models and predict ligand binding sites, analyze the effect of amino-acid variants for a user’s protein sequence [22].

Ramachandran Plot (SAVES5.0): The Ramachandran plot is a fundamental tool (http://servicesn.mbi.ucla.edu/SAVES/) in the analysis of protein structures. The Ramachandran plot is the 2d plot of the φ-ψ torsion angles of the protein backbone. It provides a simple view of the conformation of a protein [23].

RESULTS AND DISCUSSIONS

In-silico approach is used in the present study, to characterize the HPs from *Bacillus anthracis* A2012. 129 HPs from pX01 (AE011190.1) and 59 HPs from pX02 plasmid (AE011191.1) were chosen for functional and structural characterization.

Evaluation to Determine Motifs and Domains: The genome is retrieved from NCBI database. Motifs are determined by **Motif finder** whereas Domain analysis is carried out by **CDD**. E- Value of motifs obtained is noted along with the regions obtained in CDD. Motifs and Domains are predicted in 129 HPs from pX01 plasmid and 59 HPs from pX02 plasmid. Among these **59** proteins from **pX01** and **27** from **pX02** are predicted with conserved motifs and domains.

Identification of trans membrane helices, signal peptide and identity score: Next, identification of trans membrane helices is done by **TMHMM**, signal peptide is detected by **OCTOPUS** and **BLAST** is used for sequence identity. Out of these, **18** HPs from **pX01** and **11** HPs from **pX02** are predicted to have trans membrane helices and have sequence similarity above 90%.

Solubility examination: Solubility of protein is determined by **SOSUI**. The Average of hydrophobicity is determined along with solubility. If hydrophobicity exists, protein is termed as trans membrane. From earlier mentioned proteins, **15** proteins from **pX01** and **10** proteins from **pX02** are identified as membrane proteins.

Physiochemical characterization: **PROTPARAM** is used for analysis of no. of amino acids, molecular weight, is oelectric point and stability index. To determine the stability, tool is run and **9** proteins from both plasmids are identified to be stable as shown in table 1 and table 2:-

Table: 1 Prediction by PROTPARAM from pX01 plasmid

S.No.	Accession no.	No. of amino acids	Molecular weight	Is electric point	Stability index
1	AAM25959.1	37	4337.40	10.17	Stable
2	AAM25960.1	212	24914.01	9.28	Unstable
3	AAM25961.1	127	14563.57	10.52	Stable
4	AAM25964.1	195	22010.14	9.58	Unstable
5	AAM25975.1	230	26891.69	8.64	Stable
6	AAM25976.1	1320	156170.48	5.21	Stable
7	AAM26161.1	47	5494.58	5.94	Stable
8	AAM26016.1	108	12740.05	9.48	Stable
9	AAM26017.1	62	6903.39	9.87	Stable
10	AAM26036.1	53	6111.22	10.07	Stable
11	AAM26044.1	67	7513.80	9.22	Unstable
12	AAM26059.1	116	13464.21	9.90	Unstable
13	AAM26063.1	151	16286.12	9.08	Stable
14	AAM26071.1	160	17989.50	7.67	Unstable
15	AAM26152.1	66	7506.95	9.22	Unstable

Table 2. Prediction by PROTPARAM from pX02 plasmid

S. No.	Accession no	No. of amino acids	Molecular weight	Is electric point	Stability index
1	AAM26171.1	104	11976.38	9.67	Unstable
2	AAM26172.1	67	8075.56	9.75	Stable
3	AAM26173.1	84	9976.86	9.85	Unstable
4	AAM26178.1	92	9960.85	7.77	Stable
5	AAM26180.1	120	13642.00	7.81	Stable
6	AAM26181.1	249	28998.72	9.56	Stable
7	AAM26182.1	277	32156.65	9.39	Stable
8	AAM26184.1	130	14587.58	8.75	Stable
9	AAM26186.1	124	13944.52	8.83	Stable
10	AAM26194.1	60	6323.51	9.43	Stable

Subcellular localization: Furthermore concerning about subcellular localization, PSORTb was run and it was predicted that **7** proteins from **pX01** and **4** proteins from **pX02** are localized in cytoplasmic membrane. Based on the data obtained from these parameters, **4 proteins from pX01 and 2 proteins from pX02 are functionally annotated as shown in table 3 and table 4:-**

Table 3. Results of PSORTb from pX01 plasmid

S.No.	Accession no	Subcellular localization
1	AAM25959.1	Cytoplasmic membrane
2	AAM25961.1	Cytoplasmic membrane
3	AAM25975.1	Unknown
4	AAM25976.1	Cytoplasmic
5	AAM26161.1	Cytoplasmic membrane
6	AAM26016.1	Cytoplasmic membrane
7	AAM26017.1	Cytoplasmic membrane
8	AAM26036.1	Unknown
9	AAM26063.1	Cytoplasmic membrane

Table 4. Results of PSORTb from pX02 plasmid

S.No.	Accession No	Subcellular localization
1	AAM26172.1	Cytoplasmic membrane
2	AAM26178.1	Cytoplasmic membrane
3	AAM26180.1	Unknown
4	AAM26181.1	Cytoplasmic membrane
5	AAM26182.1	Cytoplasmic membrane
6	AAM26184.1	Unknown
7	AAM26186.1	Unknown
8	AAM26194.1	Unknown

Structure prediction: For structural annotation, phyre2 was run on these proteins. The structures with more than 30% confidence score is selected having role in bacteria. 3-D structure of 4 hypothetical proteins from **pX01** plasmid has been predicted, out of which **2 HPs** are selected on the basis of **role** in bacteria, with Accession no.AAM25976.1 and AAM26063.1. Similarly, **2 HPs** from

pX02 plasmid have been predicted with Accession no.AAM26172.1 and AAM26178.1. Table 5 and Table 6 are listed below:-

Table 5. Prediction by Phyre2 from pX01 plasmid

S.no	Accession no	Confidence Score
1	AAM25961.1	35.1%
2	AAM25976.1	99.0%
3	AAM26161.1	36.8%
4	AAM26063.1	35.6%

Table 6. Prediction by Phyre2 from pX02 plasmid

S.no	Accession No	Confidence score
1	AAM26172.1	37.4%
2	AAM26178.1	33.0%

Ramachandran plot: Ramachandran Plot analysis using **SAVES 5.0** showed that 77.8% residues were present in **allowed region** from **pX01** plasmid, and 1 protein with 77.8% and other with 94.1% residues was present in **allowed region** from **pX02** plasmid. No protein residue is found in disallowed region as shown in table 7 and table 8:-

Table 7. Prediction by SAVES 5.0 from pX01 plasmid

S.no	Accession no	Residues in most favoured regions	Residues in disallowed regions
1	AAM25976.1	77.8%	0%
2	AAM26063.1	77.8%	0%

Table 8. Prediction by SAVES 5.0 from pX02 plasmid

S.no	Accession no	Residues in most favoured regions	Residues in disallowed regions
1	AAM26172.1	77.8%	0%
2	AAM26178.1	94.1%	0%

Out of these four proteins, ONE protein (AAM25976.1) is best annotated with high confidence score of 99% by phyre 2.

Accession no. **AAM25976.1** from **pX01 plasmid** has a SIR2 protein family with Pfam ID PF13289. SIR2 is a NAD- dependent his tone deacetylase whose activity is required to promote chromatin silencing at the telomeres, chromosome segregation, DNA recombination and the determination of life span. SIR2 stands for Silent Information Regulator 2. The *SIR2* gene is broadly conserved in organisms ranging from bacteria to humans. It is involved in transcriptional repression and rDNA silencing. Sir2 is a limiting component of longevity, deletions of SIR2 shorten life span and an extra copy of this gene increases life span.

Accession no. **AAM26063.1** from **pX01 plasmid** with Pfam ID PF00939 has Na_sulph_symp protein family. The sodium symporters are integral membrane proteins that mediate the intake of molecules with the uptake of Na^+. Plasma membrane transporters use energy from the movement of Na^+ down its electrochemical gradient to transport a variety of dicarboxylates and inorganic anions across the membrane. Proteins encoded by SLC genes are divided into: the Na (+)-sulphate (NaS) co transporters and the Na (+)-carboxyl ate (NaC) co transporters.

Accession no. **AAM26172.1** from **pX02** plasmid has protein family STT3with ID PF002516. The central subunit of the OST complex is the highly conserved catalytic STT3, found in eukaryotes, archaea, and eubacteria. Bacterial OST contains single-subunit PglB protein that indicates the locations of the transmembrane helices, the connecting loops and conserved residues forming the active site of PglB. STT3 is multispanning membrane proteins, with 10-13 predicted TM segments by hydrophobicity plot involved in protein folding, intracellular trafficking, regulation of protein turnover, or cell-cell recognition.

Accession no. **AAM26178.1** from **pX02** plasmid contains TrbC protein family. Conjugal transfer protein, TrbC is a subunit of the pilus precursor in bacteria. The protein undergoes three processing steps before gaining its mature cyclic structure. This family contains several VirB2 type IV secretion proteins. The virB2 gene encodes a putative type IV secretion system. The pilus subunit, the pilin, of conjugative IncP pili is encoded by the *trbC* gene. Several amino acid exchanges in the TrbC core sequence allow prepilin cyclization but disable the succeeding pilus assembly.

CONCLUSION

In the present study, the sequence of Bacillus anthracis A2012 has been retrieved from NCBI. Functional and Structural analysis of these hypothetical proteins is carried out by using online Bioinformatics tool. Anthrax is one of the major diseases causing death. So these hypothetical proteins may have role in causing disease or related infections. Functional analysis of hypothetical proteins is necessary for the characterization of proteins and selecting novel therapeutic targets. From given hypothetical proteins, only 4 proteins are determined which execute various properties like presence of transmembrane helix, solubility of protein, subcellular localization, structure prediction and analysis of Ramachandran plot. Out of these 4 Hypothetical proteins, 1 protein with Accession no. AAM25976.1 has confidence score 99%. These properties help in improving target identification during drug discovery process. This protein contains SIR-2 domain which is NAD-dependent deacetylase, participating in a wide range of cellular events including chromosome silencing, chromosome segregation, DNA recombination and the determination of life span. Thus, the structural and functional prediction of such hypothetical proteins made it easier to determine the drug target sequence. Hence such annotation is effective for curing the disease and developing medicines or vaccines against such dreadful disease.

REFERENCES

1. Baillie L, Read T.D. Bacillus anthracis: a bug with attitude. Curr Opin Microbiol. 2001; 40:78–81.
2. Read TD, Peterson SN, Baillie L. W, Paulsen I. T et al. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. Nature. 2003; 423: 81–86.
3. Carr KA, Lybarger SR, Anderson EC, Janes BK, Hanna PC. The role of *Bacillus anthracis* germinant receptors in germination and virulence. Mol Microbiol. 2010; 75: 365–375.
4. Makins S, Ucluda I, Terakads N, et al. Molecular characterisation and protein analysis of the cap region, which is essential for encapsulation in Bacillus anthracis. J Bacteriol. 1989; 171:722–30.
5. Hanna P. Anthrax pathogenesis and host response. Curr Top Microbiol Immunol. 1998; 225:13-35.
6. Lanska, D. J. Anthrax meningoencephalitis. Neurology. 2002; 59:327–334.
7. Laforce FM. Anthrax. Clin Infect Dis. 1994; 19:1009-1013.
8. Meselson M, Guillemin J, Hugh-Jones M, Langmuir A, Popova I, et al. The Sverdlovsk Anthrax Outbreak of 1979. Science. 1994; 226:1202-1208.
9. Turnbull PC. Anthrax vaccines: past, present and future. Vaccine 1991; 9:533–539.
10. David L. Wheeler, Deanna M. Church, Edgar Ron, Federhen Scott, Helmberg Wolfgang, Thomas L. Madden et al. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Research. 2004; 32:35-40.
11. Bauer-Marchler A, Chitsaz, John B. Anderson et al. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 2009; 37: 205–210.
12. Aron Marchler-Bauer, Shennan Lu, John B. Anderson, Chitsaz Farideh , Myra K. Derbyshire, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res, 2011; 39:225–229.
13. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, et. al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acid Research, 2017; 45:200-203.
14. Churchill, G. A. Hidden markov chains and the analysis of genome structure. Computers and Chemistry, 1992; 16:107–115.
15. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 1989; 77:257–286.
16. Jones D. Improving the transmembrane protein topology prediction using evolutionary information. Bioinformatics, 2007; 23:538-544.

17. Altschul, S.F., Gish,W., Miller,W., Myers,E.W. and Lipman, D.J. Basic local alignment search tool. J. Mol. Biol, 1990; 215:403–410.
18. Kyte,J. and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol, 1982; 157:105–132.
19. Boeckmann, B., Bairoch, A., Apweiler, R., et al. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res, 2003; 31: 354–370.
20. E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. and Bairoch, A. ExPASy - the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res, 2003; 31:3784–3788.
21. Gardy JL, Brinkman F.SL. Methods for predicting bacterial protein subcellular localization. Nat. Rev. Microbiol, 2006; 4:741–751.
22. Lewis TE, et al. Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. Nucleic Acids Res, 2013; 41:499–507.
23. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol, 1963; 7:95-99.