International Journal for Computational Biology (IJCB)
Vol4., No2., August 2015, pp- 21-30
ISSN: 2278-8115

Insilico Proteome Screening to Identify Prospective Drug Targets in *Bacillus anthracis*

Amajala Krishna Chaitanya[1,*], Iska Bhaskar Reddy[1], Kunal Zaveri[1]
[1] Department of Biochemistry and Bioinformatics, Institute of Science, GITAM University
Visakhapatnam – 530045, Andhra Pradesh, India

Article Info

Article history:

Received: Mar 2nd 2015
Revised: Apr 4th 2015
Accepted: Apr 7th 2015

Keyword:

Anthrax
B. anthracis
Essential genes
Drug targets

ABSTRACT

Various Insilico based genome screening methods helped us in identifying the key drug targets for a pathogen. The accuracy of the predictions are systematically based on the benchmarks at different stages of methodology and the kind of dataset which is considered for the study. In the current study, we made an effort to screen the entire proteome of *Bacillus anthracis* for identification of putative drug targets. *B. anthracis* is the causautive agent for anthrax disease. Instead of genome sequence, the metabolically classified proteome of *B. anthracis* from JCVI-CMR database was considered for the present study. The entire proteome is been categorized into 25 different metabolisms and in each sub-categorised metabolisms respective protein sequences were retrieved and subjected to screening against Database of Essential Genes (DEG) and Human-Basic Local Alignment Search Tool (H-BLAST) databases. In total 136 essential genes/proteins were obtained from the DEGp (protein) screening whereas 145 Non-Human Homologs (NHHs) were predicted. The identified 145 NHHs are further subjected to criteria based selection to identify the most suitable, functional, putative drug targets. The 8 common hits of both DEG and H-BLAST were considered to be better potential targets as they justify the criteria of being an essential gene/protein, non-human homolog, availability of the 3D structure in PDB and having a significant functional role in the cellular biochemical processes.

Corresponding Author:

Amajala Krishna Chaitanya,
Department of Biochemistry and Bioinformatics, Institute of Science, GITAM University,
Visakhapatnam – 530045, Andhra Pradesh, India
Email: chy2ak@gmail.com

How to Cite:

Chaitanya *et. al.* Insilico Proteome Screening to Identify Prospective Drug Targets in *Bacillus anthracis.* IJCB. 2015; Volume 4 (Issue 2): Page 21-30.

1. INTRODUCTION

The genome sequencing projects of various pathogens is becoming increased day-by-day, resulting the genomic data to be analyzed by diverse approaches to identify the putative anti-pathogenic drug targets and vaccine candidates. Insilico methods can be used to identify putative gene products through the comparative genomics approach [1]. One of the approach in comparative genomics is applying the subtractive genomics by which it enables the subtraction of datasets between the host and pathogen genome and provides information for a set of genes that are likely to be essential to the pathogen but absent in the host [2], that is human and should have no homolog in human, so that when a drug or a lead compound is designed considering the potential target it should only be against the mechanism and functionality of the pathogen not the host. The current studies make use of the subtractive proteomics approach to analyze the complete proteome of *Bacillus anthracis* to search for potential drug targets [3].

Bacillus anthracis, the organism that causes anthrax is a large Gram positive, aerobic, spore bearing bacillus, 1–1.5 X 3–10 µm in size, is the only obligate pathogen within the genus bacillus [4]. The disease occurs mostly in wild or domestic mammals and may occur in humans when exposed to infected animals, or upon direct exposure. Fully virulent forms of *B. anthracis* carry two large plasmids: pXO1 and pXO2 [5]. Plasmid pXOl encodes for three toxins: lethal factor, edema factor (calmodulin-dependent adenylate cyclase), and the protective antigen. These proteins are individually nontoxic; however, they act in binary combinations and produce two distinct toxic responses in the host organism: edema and cell death [6], whereas the pXO2 plasmid encodes the capsule which is composed of a high-molecular-weight polypeptide (poly-D-glutamic acid) which inhibits host phagocytosis of the vegetative form of B. anthracis.

B. anthracis is considered to be one of the most likely biological warfare agents due to the ability of its spores to be transmitted by the respiratory route, the high mortality associated with inhalation anthrax, and spore stability [5]. Its use as a potential bioweapon has accelerated the race to develop vaccines and antitoxins for this Category A bioterrorism agent as defined by the National Institute of Allergy and Infectious Diseases (NIAID). Currently drugs such as ciprofloxacin, doxycycline and Penicillin-G are used for the therapy and licensed anthrax vaccine named Anthrax Vaccine Adsorbed (or AVA) is recommended. As this is only single anthrax vaccine licensed in U.S., the new vaccines are in the underdevelopment stage. As well the drug therapies are commonly done on post-exposure prophylaxis. Requirement of booster doses frequently for vaccine and prolonged therapies with antibiotics, thus underscores the need of new drug target(s) for development of effective drug against the *B. anthracis.*

2. RESEARCH METHOD

2.1. Strain Selection and Retrieval of Metabolic Profile based Proteome

As many *B. anthracis* strains genomes are available in the genome databases such as National Center for Biotechnology Information (NCBI) Genomes, J. Craig Venter Institute-Comprehensive Microbial Resource (JCVI-CMR) [7], and PAThosystems Resource Integration Center (PATRIC) [8], the strain selection was done by the comparative analysis of specific attributes like: Review information (Recently updated), Number of protein-coding genes (5902), Status of sequencing (Completed), Plasmid coding (two plasmids coding virulence factors). Taking into account these attributes, a *Bacillus anthracis* strain CDC 684 (Accession Number: NC_012581.1) was selected for the current study.

The total number of protein-coding genes including the genome and both the plasmids in *B. anthracis* CDC 684 strain are 5902 according to the Comprehensive Microbial Resource (CMR) database. (http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi) which has the same numbers of genes data represented into respective functional metabolisms with the Gene locus information for each of the protein-coding gene. The protein coding genes were retrieved from the tool - Role Category Pie Chart in JCVI-CMR database, which shows the number and percentage of genes/proteins for every metabolic role category in the CMR (Fig. 1).

The number of genes involved in each metabolism gives an overview of the active metabolism functioning in an organism. As Swiss-prot database is consisting of highly annotated protein sequences, the protein sequences(s) for each gene of respective metabolism were retrieved by using the gene locus information as query. The retrieved sequences were further used/implied to find the probable drug targets.

Currently, the JCVI-CMR database is no longer supported and been taken offline. But all the metabolic based proteome data can be retrieved back reliably by using Gene locus IDs from Uniprot protein sequence database.

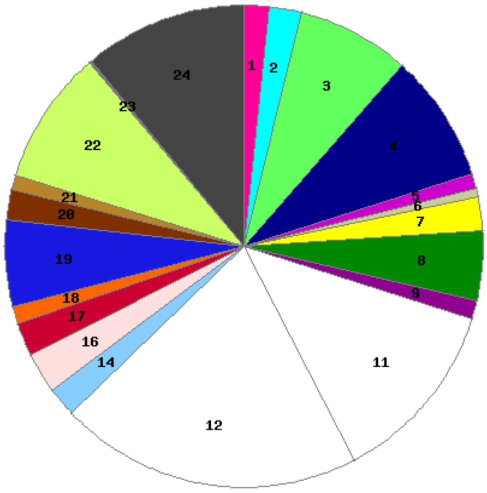
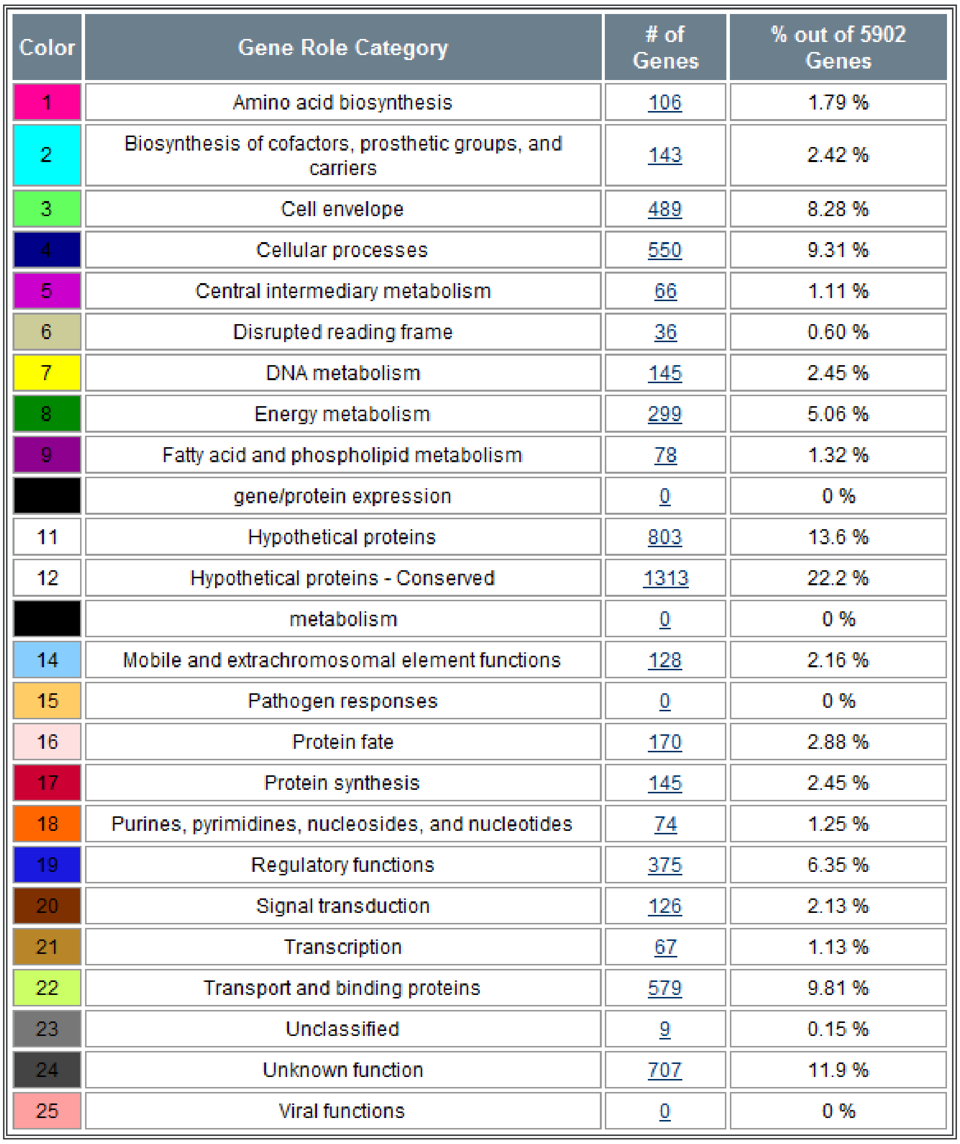


Figure 1a. Metabolic Role Category Pie chart



Color	Gene Role Category	# of Genes	% out of 5902 Genes
1	Amino acid biosynthesis	106	1.79 %
2	Biosynthesis of cofactors, prosthetic groups, and carriers	143	2.42 %
3	Cell envelope	489	8.28 %
4	Cellular processes	550	9.31 %
5	Central intermediary metabolism	66	1.11 %
6	Disrupted reading frame	36	0.60 %
7	DNA metabolism	145	2.45 %
8	Energy metabolism	299	5.06 %
9	Fatty acid and phospholipid metabolism	78	1.32 %
■	gene/protein expression	0	0 %
11	Hypothetical proteins	803	13.6 %
12	Hypothetical proteins - Conserved	1313	22.2 %
■	metabolism	0	0 %
14	Mobile and extrachromosomal element functions	128	2.16 %
15	Pathogen responses	0	0 %
16	Protein fate	170	2.88 %
17	Protein synthesis	145	2.45 %
18	Purines, pyrimidines, nucleosides, and nucleotides	74	1.25 %
19	Regulatory functions	375	6.35 %
20	Signal transduction	126	2.13 %
21	Transcription	67	1.13 %
22	Transport and binding proteins	579	9.81 %
23	Unclassified	9	0.15 %
24	Unknown function	707	11.9 %
25	Viral functions	0	0 %

Figure 1b. Number of Protein-Coding Genes involved in Various Metabolic Categories

2.2. Finding the Essentiality of Proteins

Database of Essential Genes (DEG) provides the information about the essential genes which are indispensible for the survival of an organism [9]. Protein sequences those retrieved from Swiss-Prot in FASTA format are subjected to Prokaryotic DEG-BLAST by selecting the blastp (Protein Query vs. Protein DB) program. The essential genes are short listed based on the percentage of identity. The protein sequences were selected as essential genes with the percentage of identity having more than 70%. The observed parameters for DEG pBLAST are tabulated (Table 1).

Table 1: observed parameters for DEG pBLAST

S.No	Parameter	Value
1.	Filter	ON
2.	Expect	0.0001
3.	Matrix	BLOSUM 62
4.	Perform gapped alignment	ON
5.	Descriptions	100
6.	Alignments	100

2.3. Identifying the Non-Human Homologous Proteins

All the protein coding sequences of *B. anthracis* were also analyzed for sequence similarity with human proteome using Human-BLASTP (H-BLAST) database [10] with an E-value cutoff of 10. Non-human homologous proteins were identified based on the criterion that the percent of identity should be less than 20%. The observed parameters for H-BLAST are tabulated (Table 2).

Table 2: observed parameters for H-BLAST

S.No	Parameter	Value
1.	Database	RefSeq protein
2.	Max target sequences	100
3.	Expect threshold	10
4.	Word size	3
5.	Matrix	BLOSUM 62
6.	Gap costs	Extension:11 Extension:1
7.	Compositional adjustments	Conditional compositional score matrix
8.	Filter	Low Complexity Region
9.	Algorithm	blastp (protein-protein BLAST)

2.4. Predicting Subcellular Localization of Essential and Non-Human Homologous Proteins

PSORTb v3.0.2 is the most precise bacterial localization prediction tool which accepts the fasta formatted protein sequences [11]. By selecting the following parameters: Organism-Bacteria; Gram Stain-Positive; Output format-Normal, the sequences of obtained essential genes and non-human homologs were submitted for their subcellular localization prediction.

2.5. Selecting the putative ‘Best’ Drug Targets

For consideration of ‘best’ drug targets, the following three different criterias were implemented [13].

1. The protein must be a non-human homolog.
2. The 3D structure availability in PDB and its highest percentage identity with of the same *B. anthracis* species.
3. The protein must be an essential molecule for the survival of the organism.

The protein sequences which satisfy the criteria of being non-human homologs are subjected to PDB-BLAST [14] to retrieve the most significant 3D structures. The 3D structures of the proteins were short listed based upon the cutoff percentage identity between sequence and the structure to be 60%.
The observed parameters for PDB-BLAST are tabulated (Table 3).

Table 3: observed parameters for PDB-BLAST

S.No	Parameter	Value
1.	Database	Protein Data Bank PDB
2.	Max target sequences	100
3.	Expect threshold	10
4.	Word size	3
5.	Matrix	BLOSUM 62
6.	Gap costs	Extension:11 Extension:1
7.	Compositional adjustments	Conditional compositional score matrix
8.	Filter	Low Complexity Region
9.	Algorithm	blastp (protein-protein BLAST)

3. RESULTS AND ANALYSIS

3.1 Genome, Proteome and Metabolic Profile of *B. anthracis*

From the JCVI-CMR database, the retrieved number of protein-coding genes from *genome* and the plasmids pXO1 and PXO2 of *B. anthracis* (CDC 684) genome are 5579, 206, 117 respectively resulting in total of 5902 numbers of proteins. The detailed genome and proteome components information from NCBI and UniprotKB respectively are shown in figure 2.

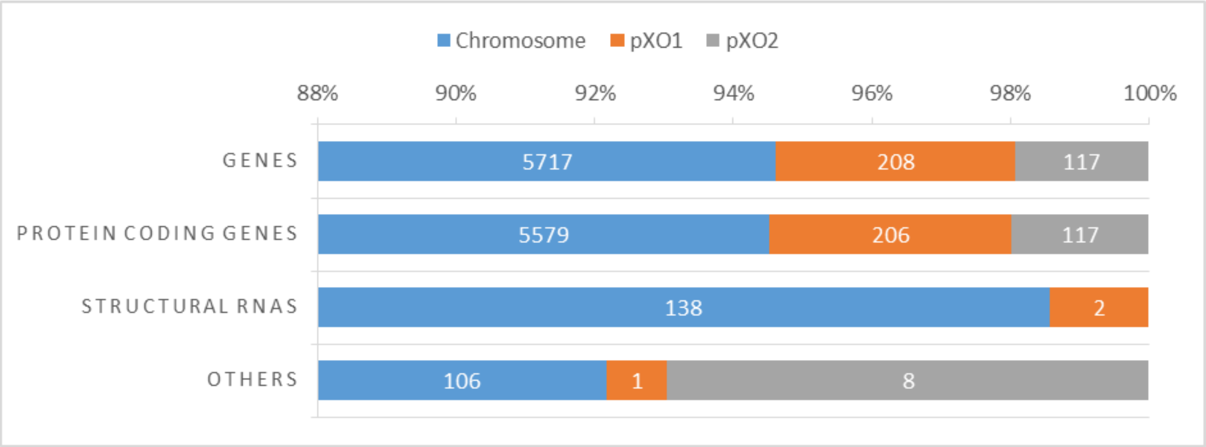


Figure 2. Genome and Proteome Components of *B. anthracis* CDC 684 Strain

All the genes/proteins in JCVI-CMR were annotated into 21 functional metabolic categories. Among those gene role categories, 35.8 % of genes were encoded from the category of Hypothetical Proteins and Conserved Hypothetical proteins. Other than hypothetical proteins, there are genes categorized as Unknown Function comprising of 11.9 %. Hence the remaining 52.3 % of genes are categorized into key metabolisms. Among these key metabolisms, the major numbers of genes are involved in the metabolisms of cell envelope, cellular processes, regulatory functions and transport and binding proteins. This study signifies that the biochemical processes are requiring more than 50 % of the total genome of the bacterium.

3.2. Defining Essential Genes/Proteins in Discovery of Minimal Gene Complement

The database of essential genes hosts records of essential genes among a wide range of organisms including both prokaryotes and eukaryotes. Essential gene products can be represented as good targets for antibacterial drugs. The screening of the each gene product of *B. anthracis* is done by considering the sequence identity to be > 70 % against the DEG prokaryotic pBLAST. From the DEG-BLAST studies, the total number of 146 essential genes were predicted. Among them the number of redundant and conserved hypothetical proteins predicted were 4 and 6 respectively. They were removed from the final count resulting in 136 numbers of essential genes. From all the metabolisms, the majority of the essential genes were predicted from protein synthesis metabolism which can be suggested that protein anabolism is highly essential for the survival of the organism. The further DEG results were summarized in the table 4. Many of predicted essential genes were exhibiting poor range of 20-40% sequence similarities with the human proteome. All the details of 136 gene products with corresponding metabolic family, gene id, common name, EC number including the subcellular localization "and" prediction of the protein were provided in the Supplementary Table 1.

Essential genes/proteins can also be treated as potential and important drug candidates as well vaccine candidates if they are membrane-based proteins or enzymes. As only 136 gene products were predicted as essential genes which comprises of only 2.3 % of the total genes, they can also serve as minimal gene complement signifying the minimum number of proteins necessitated for the organism's critical metabolic pathways and its survival [12].

Table 4. Summary of DEG (Database of Essential Genes) Screening Analysis

Essential genes predicted	146
Redundant + Conserved hypothetical proteins (4 + 6)	10
Essential genes as enzymes	73
Essential genes which are cytoplasmic	118
Total number of essential genes considered	**136**

3.3. Non-Human Homologs (NHHs) are the most preferable drug/vaccine candidates

The same previously retrieved dataset which was used to screen DEG was employed to identify the non-human homologs by performing H-BLAST (Human-specific BLAST). The sequence identities of <20% between proteomes of *B. anthracis* and human were considered as reliable drug targets. As the dataset is metabolic profile based proteome, some proteins such as toxins, family based transcriptional regulators, polysaccharide synthase family proteins, histidine kinases, DNA-binding response regulator proteins, etc which were existing in redundant numbers were not included in the JCVI-CMR database. By screening the entire proteome of B. anthracis against H-BLAST, a total number of non-human homologs predicted were 145. The redundant proteins were removed and further subjected to PDB-BLAST to retrieve the appropriate 3D strucutres. The 3D structures of the protein molecules are necessary for further studies like identification of active sites and for protein-ligand docking. The 3D strucutres which have the sequence identity percentage of >60% with respect to non-human homolog sequnce were only considered as important for further selection of suitable of drug targets. The detailed H-BLAST and PDB-BLAST results were summarized in the Table 5. All the details of 145 non-human homologs with corresponding metabolic family, gene locus, common name, EC number, protein length, PDB Id and its corresponding percent identity and the subcellular localization prediction of the non-human homolog were provided in the Supplementary Table 2.

Table 5. Summary of H-BLAST and PDB-BLAST Screening Analysis

Non-human homologs predicted	145
Hypothetical proteins / Uncharacterized proteins	11
Putative membrane proteins predicted	9
Putative lipoproteins predicted	5
Non-human homologs localized in cytoplasm	51
Non-human homologs localized in cytoplasmic membrane	55
Non-human homologs localized in cell wall and extracellular	7
3D structures obtainable for NHHs	69
Total number of Non-human homologs considered as putative drug targets (Essential + Non-human homolog + Availability of better 3D-Structure + Important Functional role)	**8**
3D Structures availability with >90% sequence/structural similarity	3

3.4. Selection of 'functional' drug targets obtained through criteria based database screening

Total number of protein coding genes that fulfill the criteria of being essential and non-human homologs genes with their respective metabolic function are graphically represented in the Figure 3. The total numbers of essential genes that are required for the *B. anthracis* for its minimum survival are 136 genes. Among them the majority of the genes (56) are involved in the protein synthesis metabolism. The total numbers of genes that are non-human homologs i.e., absent in humans or don’t play any important role in human metabolism were predicted to be 145**,** of which the genes coding for cell envelope and transport binding proteins metabolisms are more in number with a count of 35 and 32 respectively.

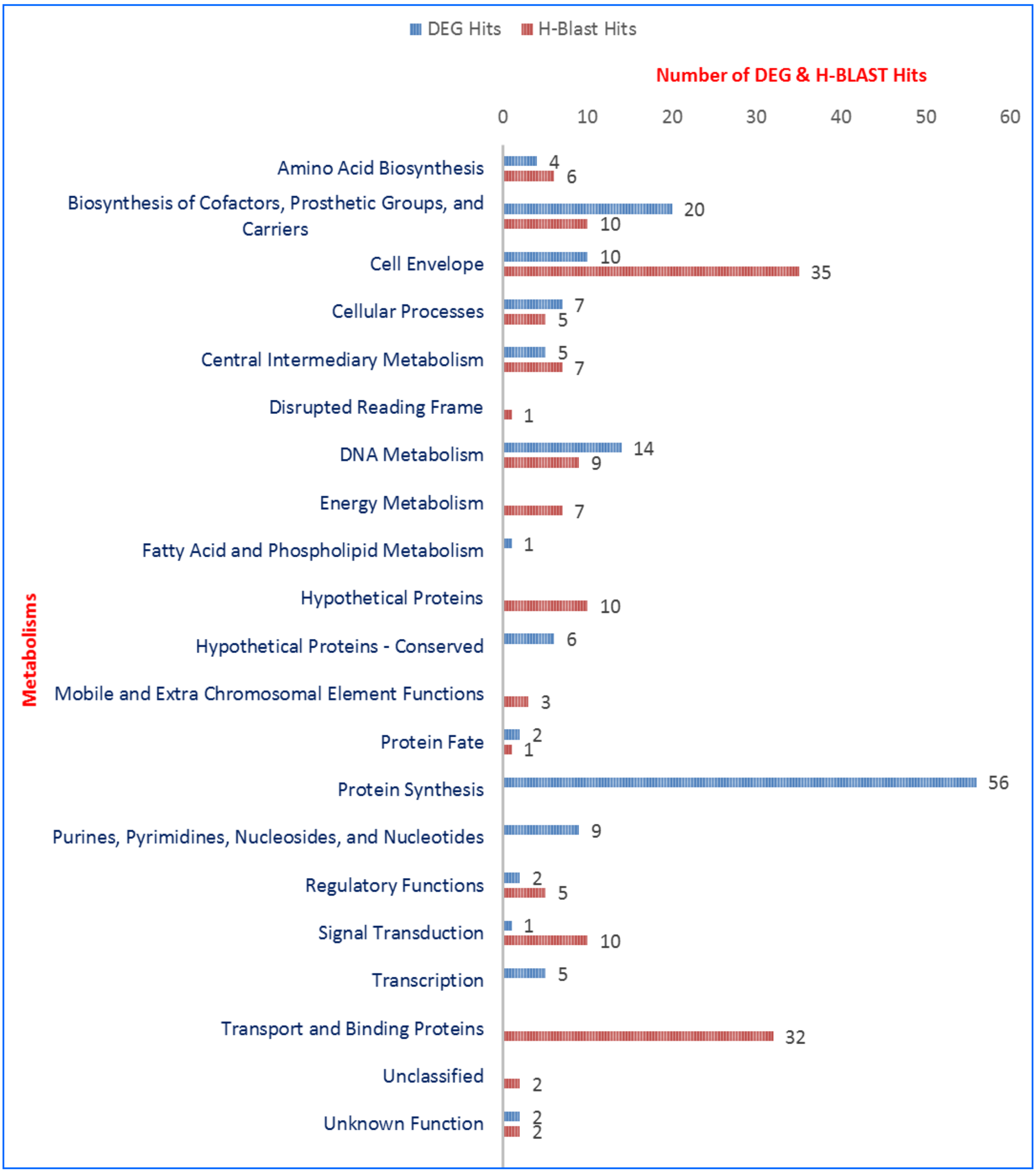


Figure 3. Number of Essential Genes and Non-Human Homologs Hits obtained

The screening of 145 NHHs against PDB-BLAST resulted in retrieval of 69 similar three-dimensional protein structures and for the remaining 76 numbers of proteins, no significant protein structures were available in PDB. On the basis of PDB-BLAST alignments, the sequences having percentage identity of >60% were considered to be reliable 3D structures for the prospective drug targets. Out of 145, 69 NHHs were possesing the 3D structures whereas for the remaining 76 NHHs 3D structures were not available in the PDB. The hypothetical / uncharacterized proteins (11), putative membrane proteins (9), and putative lipoproteins (5) were not considered for further selection as the information regarding them were poorly annotated. The cytoplasmic membrane (55), cell wall and extracellular proteins (7) can be better drug candidates than cytoplasmic proteins (51) as they are readily available to interact with the external molecules on the bacterial surface (Figure. 4).

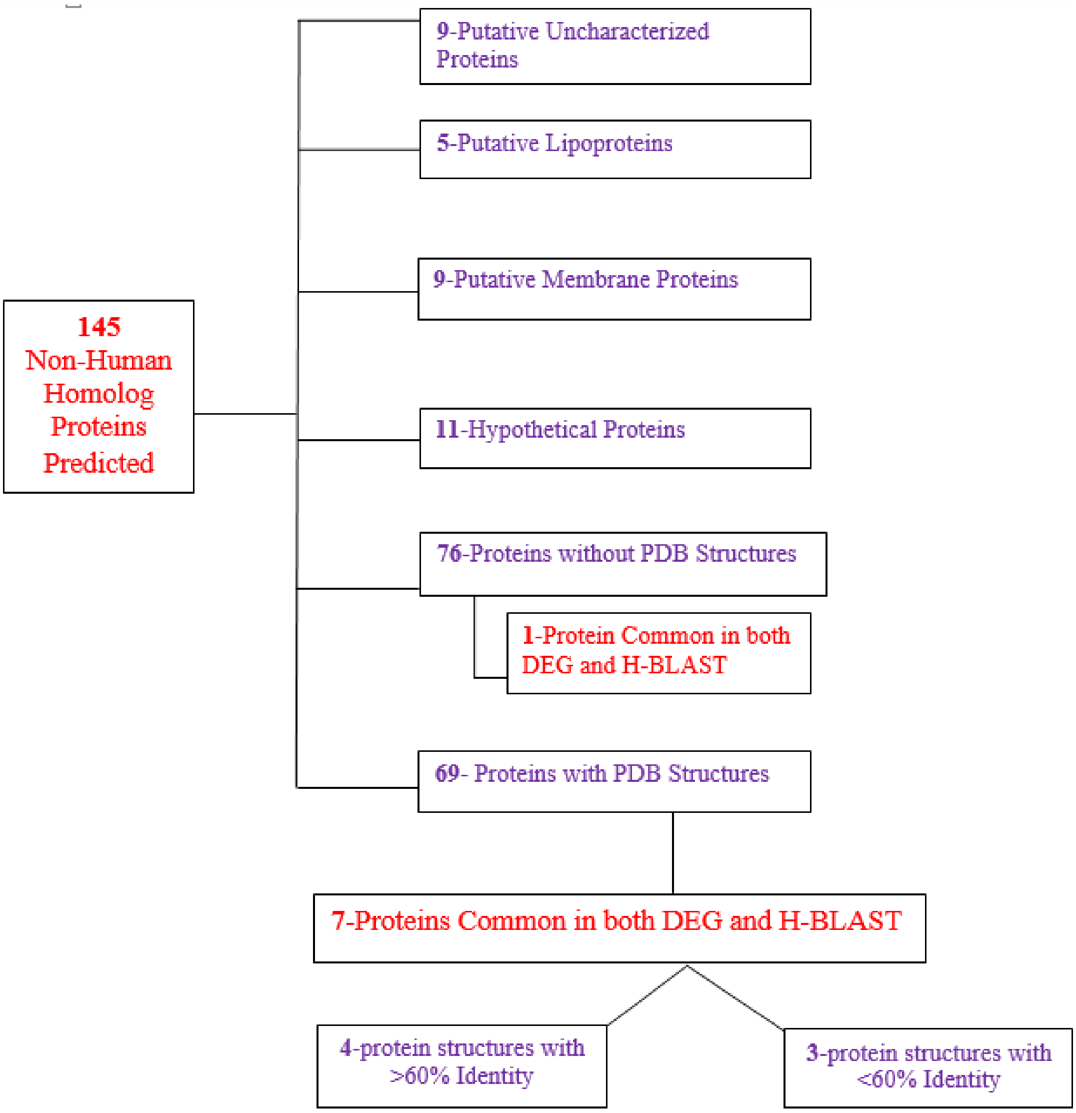

145
Non-Human Homolog Proteins Predicted
9-Putative Uncharacterized Proteins
5-Putative Lipoproteins
9-Putative Membrane Proteins
11-Hypothetical Proteins
76-Proteins without PDB Structures
1-Protein Common in both DEG and H-BLAST
69- Proteins with PDB Structures
7-Proteins Common in both DEG and H-BLAST
4-protein structures with >60% Identity
3-protein structures with <60% Identity

Figure 4. Flow Chart representing the selection of 'functional' drug targets

The eight common hits in both DEG and H-BLAST screening (Table 6) were considered to be the most putative and 'functional' drug targets as they do not resemble to any of the human proteins. It plays an essential role for the survival of the organism, and readily available with similar structural analogue. The 3D structure of a protein plays critical role in understanding the cellular processes as well as the mechanism of catalysis. These studies provide nessesary information regarding the drug target which could be helpful to the structural bioinformatician and researchers working in drug discovery processes.

Table 6. List of common protein hits obtained in DEG & H-BLAST Screening

S.No.	Protein	Metabolism	PDB ID	Identity
1	Delta-lactam-biosynthetic de-N-acetylase	Sporulation & Cellular Processes	2J13	99%
2	Iron-sulfur cluster assembly scaffold SufA	Biosynthesis of Cofactors, Prosthetic Groups & Carriers	1XJS	83%
3	2-C-methyl-D-erythritol 2,4 cyclodiphosphate synthase	Biosynthesis of Cofactors, Prosthetic Groups & Carriers	3F6M	65%
4	FeS assembly ATPase SufC	Biosynthesis of Cofactors, Prosthetic Groups & Carriers	2D2E	61%
5	Cysteine desulfurase SufS	Biosynthesis of Cofactors, Prosthetic Groups & Carriers	1T3I	52%
6	DNA topoisomerase I	DNA Metabolism	1ECL	43%
7	FeS assembly protein SufB	Biosynthesis of Cofactors, Prosthetic Groups & Carriers	2ZU0	29%
8	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	Biosynthesis of Cofactors, Prosthetic Groups & Carriers	3D Struture Not available	-

4. CONCLUSION

The free availability of the genome/proteome sequences and user friendly tools of bioinformatics has led to an enormous activity in analyzing the biological data resulting in interpretion of the biological information by making accurate inferences or predictions. In the current study, the *B. anthracis* proteome was screened against the database of essential genes, human specific-BLAST database and PDB-BLAST database to identify the probable essential genes/proteins and non-human homologous proteins. A series of functional selection critiria was implemented in retrieving and analyzing the sequences data which resulted in predicting the significant putative drug targets. The requirement of novel drug targets are principally necessitated in providing the alternative paths to combat the pathobiology of *B. anthracis.*

ACKNOWLEDGEMENTS

We would like to thank Management of GITAM University, Visakhapatnam, India for providing Bioinformatics Infrastructure facility. We are very thankful to Dr. N. Nageswara Rao Reddy for his directional ideology in the statement of research problem and outlining the research approach. We also thank M. Divya, D. Ramya, K. Kavitha, K. Praseetha and Mamtha Kumari for their assistance in collection and analyzing the relative data.

SUPPLEMENTARY FILES

1. List of Essential Genes
2. List of Non-Human Homologs

REFERENCES

[1] Fritz B, Raczniak GA., Bacterial genomics: potential for antimicrobial drug discovery. BioDrugs. 2002; 16(5):331-7. Cited in PubMed; PMID 12408737.

[2] Aditya Narayan Sarangi, Rakesh Aggarwal, Qamar Rahman, Nidhi Trivedi., Subtractive Genomics Approach for in Silico Identification and Characterization of Novel Drug Targets in *Neisseria Meningitides* Serogroup B. J Comput Sci Syst Biol Volume 2(5): 255-258 (2009) – 255.

[3] Rathi B, Sarangi AN, Trivedi N. Genome subtraction for novel target definition in *Salmonella typhi*. Bioinformation. 2009 Oct 11; 4(4):143-50. Cited in PubMed; PMID 20198190.

[4] Spencer RC. Bacillus anthracis. J Clin Pathol. 2003 Mar; 56(3):182-7. Cited in PubMed; PMID 12610093.

[5] Ariel N, Zvi A, Makarova KS, Chitlaru T, Elhanany E, Velan B, Cohen S, Friedlander AM, Shafferman A. Genome-based bioinformatic selection of chromosomal Bacillus anthracis putative vaccine candidates coupled with proteomic identification of surface-associated antigens. Infect Immun. 2003 Aug; 71(8):4563-79. Cited in PubMed; PMID 12874336.

[6] Riedel S. Anthrax: a continuing concern in the era of bioterrorism. Proc (Bayl Univ Med Cent). 2005 Jul; 18(3):234-43. Cited in PubMed; PMID 16200179.

[7] Davidsen T, Beck E, Ganapathy A, Montgomery R, Zafar N, Yang Q, Madupu R, Goetz P, Galinsky K, White O, Sutton G. The comprehensive microbial resource. Nucleic Acids Res. 2010 Jan; 38(Database issue):D340-5. Cited in PubMed; PMID 19892825.

[8] Gillespie JJ1, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, et. al., PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun. 2011 Nov; 79(11):4286-98. Cited in PubMed; PMID 21896772.

[9] Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes. Nucleic Acids Res. 2004 Jan 1; 32(Database issue):D271-2. Cited in PubMed; PMID 14681410.

[10] McGinnis S1, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004 Jul 1; 32(Web Server issue):W20-5. Cited in PubMed; PMID 15215342.

[11] N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26(13):1608-1615.

[12] Mushegian AR, Koonin EV,. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A. 1996 Sep 17; 93(19):10268-73. Cited in PubMed; PMID 8816789.

[13] K. Zaveri, A. Krishna Chaitanya, I. Bhaskar Reddy., Virtual screening and docking studies of identified potential drug target: Polysaccharide deacetylase in *Bacillus anthracis.*, International Letters of Natural Sciences, 7 (2015) 70-77.

[14] Amajala Krishna Chaitanya, I. Bhaskar Reddy, P. Minakshi, Z. Kunal, DSVGK. Kaladhar, Homology Modeling and Structural Analysis Of DNA Binding Response Regulator of Bacillus anthracis, International Journal of Scientific Research; Vol 2, Issue 8, 32-34.