INSILICO MODELING OF FabH OF BACILLUS CEREUS AND CONFORMATIONAL STUDY WITH CoA

Sameer Hassan[1] and Meenupriya J.[2]
[1]Biomedical Informatics Centre, Tuberculosis Research Centre (ICMR), Chennai, India
[2]Deparment of Biotechnology, Sathyabama University, Chennai, India
E-mail : [1]sameerh@trcchennai.in

ABSTRACT

Pathogenic bacteria are sometimes very dangerous to humans. One among them is Bacillus cereus that causes food poisoning in human. Bacillus cereus is gram-positive bacteria that infect food items like rice, fish, raw meat and improperly cooked food. This bacterium is highly heat resistant.The fatty acid biosynthesis pathway is an attractive but still largely unexploited target for the development of new antibacterial agents. FabH, an essential enzyme for bacterial viability, catalyzes the initiation of fatty acid elongation by condensing malonyl-ACP with acetyl-CoA. Inhibitors of the condensation step of fatty acid biosynthesis represent new classes of compounds with antibiotic potential.The characteristics and the biochemical activity of FabH have been analyzed. FabH in Bacillus cereus has been studied and its structure is modeled using insilico methods. Structural analysis and comparative study is done between homologous proteins and the active site of the protein is predicted. This could be a key drug target in Bacillus cereus. FabH has been proved to be a novel drug target in various bacteriae. Similarly, in Bacillus cereus too, it could be a highly potential drug target. In future, studies on drug discovery of this particular target will be of greater use to mankind to overcome the diseases caused by Bacillus cereus.

KEYWORDS : Bacillus cereus, FabH, antibacterial activity

I. INTRODUCTION

Bacteria that cause disease are called pathogenic bacteria. Bacteria can cause diseases in humans, in other animals, and also in plants. Some bacteria can only make one particular host ill; others cause trouble in a number of hosts, depending on the host specificity of the bacteria. The diseases caused by bacteria are almost as diverse as the bugs themselves and include food poisoning, toothache anthrax and even certain forms of cancer.

Bacillus species are aerobic, sporulating, rod-shaped bacteria that are ubiquitous in nature. *Bacillus* endospores are resistant to hostile physical and chemical conditions, but in addition various Bacillus species have a wide range of physiologic adaptations which enable them to survive or thrive in harsh environments, ranging from desert sands and hot springs to Arctic soils and from fresh waters to marine sediments. Because the spores of many Bacillus species are resistant to heat, radiation, disinfectants, and desiccation, they are difficult to eliminate from medical and pharmaceutical materials and are a frequent cause of contamination. *Bacillus* species are well known in the food industry as spoilage organisms.

Food borne illness is an ever-present threat that can be prevented with proper care and handling of food products. It is estimated that between 24 and 81 million cases of food borne diarrhea disease occur each year.

β-Ketoacyl-Acyl Carrier Protein Synthase III (FabH)

` It is a Determining Factor in Branched-Chain Fatty Acid Biosynthesis. Genomic research is playing a critical role in the discovery of new antimicrobial drugs. The rapid increase in bacterial and eukaryotic genome sequences allows for new and innovative ways for obtaining antimicrobial protein targets. β-Ketoacyl-ACP synthase III (FabH), an essential enzyme for bacterial viability, catalyzes the initiation of fatty acid elongation by condensing malonyl-ACP with acetyl-CoA.

The FabH protein, or 3-ketoacyl-ACP synthase III, is a member of the ß-ketoacyl synthase family of enzymes. The primary reaction of the FabH enzyme is the condensation of malonyl-ACP with acetyl-coenzyme A (CoA). It is unique among ß-ketoacyl synthase enzymes in that it utilizes acetyl-CoA as a donor and has been shown to have an acetyl-CoA-ACP transacylase activity in vitro. Despite the overall similarities in their primary amino acid sequences, the FabH proteins from various bacterial species have been shown to have very different substrate specificities.

FabH from Gram-negative Escherichia coli has been studied extensively. The E. coli FabH crystal structure has been solved in the presence and the absence of the substrate, acetyl-CoA. In the crystal structure, the close approximation of Cys112 to CoA suggests it may play an important role in catalysis. Modeling based on a bound CoA molecule has identified His244 and Asn274 as additional residues that might be involved in catalysis. Because they are essential enzymes for bacteria and differ significantly from human fatty acid synthase (FAS), various bacterial FabHs have been studied as potential anti-bacterial targets. Substrate specificity of the various FabH enzymes appears to be the determining factor in the

biosynthesis of branched- or straight-chain fatty acids of the type II fatty acid synthase. Consistent with this notion, FabH purified from Gram-negative and Gram-positive bacteria, despite their overall similar catalytic mechanism, have displayed significantly different substrate specificities

FATTY ACID SYNTHESIS

The Fatty Acid Synthesis takes place in three steps,

(i) Synthesis of malonyl-CoA via acetyl-CoA carboxylase

(ii) Fatty acid synthase

(iii) Fatty acid elongation and desaturation

II. MATERIALS & METHODS

Various Bioinformatics tools and databases have played a vital role in the molecular modeling of FabH in Bacillus cereus. A step-by-step procedure is followed for modeling the protein.

TOOLS & DATABASE

ENTREZ

The Entrez Global Query Cross-Database Search System is a powerful federated search engine, or web portal that allows users to search many discrete health sciences databases at the National Center for Biotechnology Information (NCBI) website. Entrez Global Query is an integrated search and retrieval system that provides access to all databases simultaneously with a single query string and user interface. Entrez can efficiently retrieve related sequences, structures, and references. Entrez is not a database itself, but rather is the interface through which all of its component databases can be accessed and traversed-an integrated information retrieval system. The Entrez information space includes PubMed records, nucleotide and protein sequence data, three-dimensional structure information, and mapping information.

BLAST

BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity.

MODELLER

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (3,4), and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc.

RASMOL

RasMol is a program that allows you to view molecular structures on the computer screen, and to manipulate them. RasMol was designed for viewing protein structures -- molecules so large that one would not make an ordinary molecular model by hand. However, it can also be used for small molecules. Using RasMol for small molecules is particularly useful if you do not have a set of models. If you do have models, it may be good to learn to use RasMol with small molecules, and even compare the RasMol model with the "physical" models.

CLUSTALW

Multiple alignments of protein sequences are important tools in studying sequences. The basic information they provide is the identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins and in identifying new members of protein families.

TREEVIEW

TreeView is a simple program for displaying phylogenies on Apple Macintosh and Windows PCs. TreeView provides a simple way to view the contents of a NEXUS, PHYLIP, Hennig86, Clustal, or other format tree file. While PAUP and MacClade have excellent tree printing facilities, there may be times you just want to view the trees without having to load the data set they were generated from. The PHYLIP package contains tree drawing programs which offer a greater variety of trees than TreeView, but are somewhat clumsy to use. The forthcoming PAUP* for Windows does not have a graphical interface, hence TreeView allows you to create

publication quality trees from PAUP files, either directly, or by generating graphics files for editing by other programs.

PFAM

Pfam is a database of protein families that currently contains 7973 entries (release 18.0). The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). Proteins are generally composed of one or more functional regions, commonly termed domains. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

SWISS-PROT

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Recent developments of the database include format and content enhancements, cross-references to additional databases, new documentation files and improvements to TrEMBL, a computer-annotated supplement to SWISS-PROT. The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria: (i) annotations, (ii) minimal redundancy and (iii) integration with other databases.

Protein Data Bank (PDB)

This database contains the known enzyme structures that have been deposited in the Brookhaven Protein Data Bank.The PDB structure entries, consisting of a collection of files having nondescript names, cannot be easily grasped in a biochemically meaningful context.

PDBSUM

PDBsum is a web-based database providing a largely pictorial summary of the key information on each macromolecular structure deposited at the Protein Data Bank (PDB). It includes images of the structure, annotated plots of each protein chain's secondary structure, detailed structural analyses generated by the PROMOTIF program, summary PROCHECK results and schematic diagrams of protein–ligand and protein–DNA interactions. RasMol scripts highlight key aspects of the structure, such as the protein's domains, PROSITE patterns and protein–ligand interactions, for interactive viewing in 3D.

KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals. The PATHWAY database records networks of molecular interactions in the cells, and variants of them specific to particular organisms.

The KEGG, the Kyoto Encyclopedia of Genes and Genomes, was initiated by the Japanese human genome programme in 1995. According to the developers they consider KEGG to be a "computer representation" of the biological system. The KEGG database can be utilized for modeling and simulation, browsing and retrieval of data.

PROSITE

PROSITE is an annotated collection of motif descriptors dedicated to the identification of protein families and domains. The motif descriptors used in PROSITE are either patterns or profiles, which are derived from multiple alignments of homologous sequences. This gives to these motif descriptors the notable advantage of identifying distant relationships between sequences that would have passed unnoticed based solely on pairwise sequence alignment. Patterns and profiles have both their own strengths and weaknesses, which define their area of optimum application.

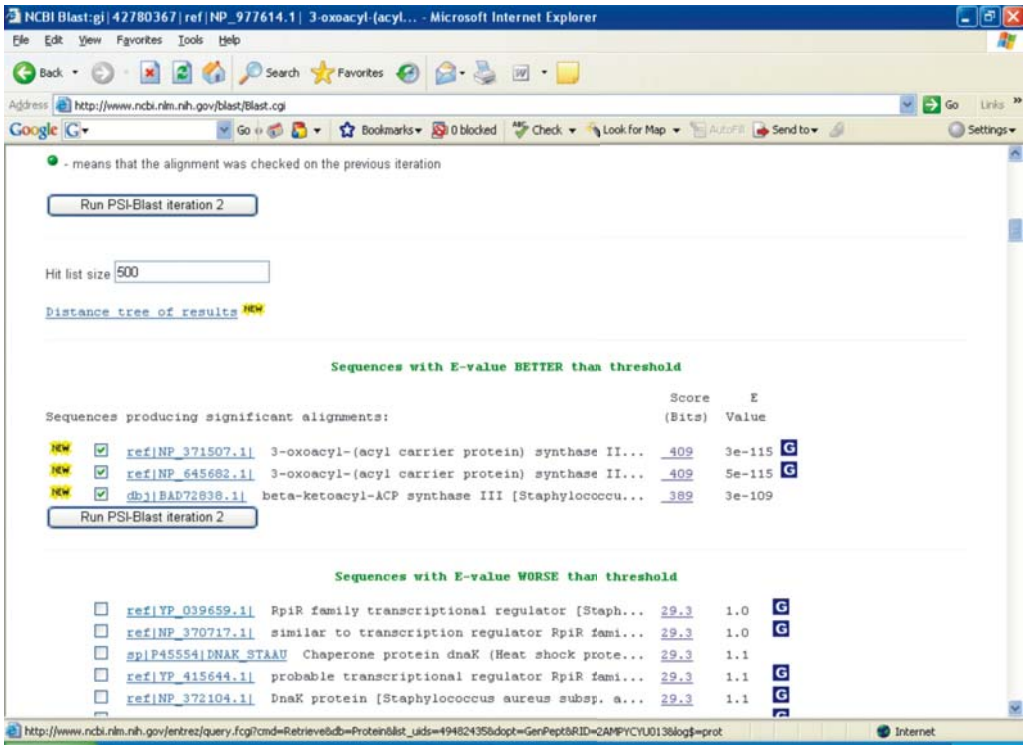
III. RESULTS AND DISCUSSION

a) Obtain the sequence of the target protein

The initial step in modeling a protein is to find its protein sequence. FabH of Bacillus cereus has to be searched by Entrez system in NCBI. The sequence with accession number NP_977614.1 was retrieved and used for analysis.

b) FabH sequence of pathogenic bacteriae

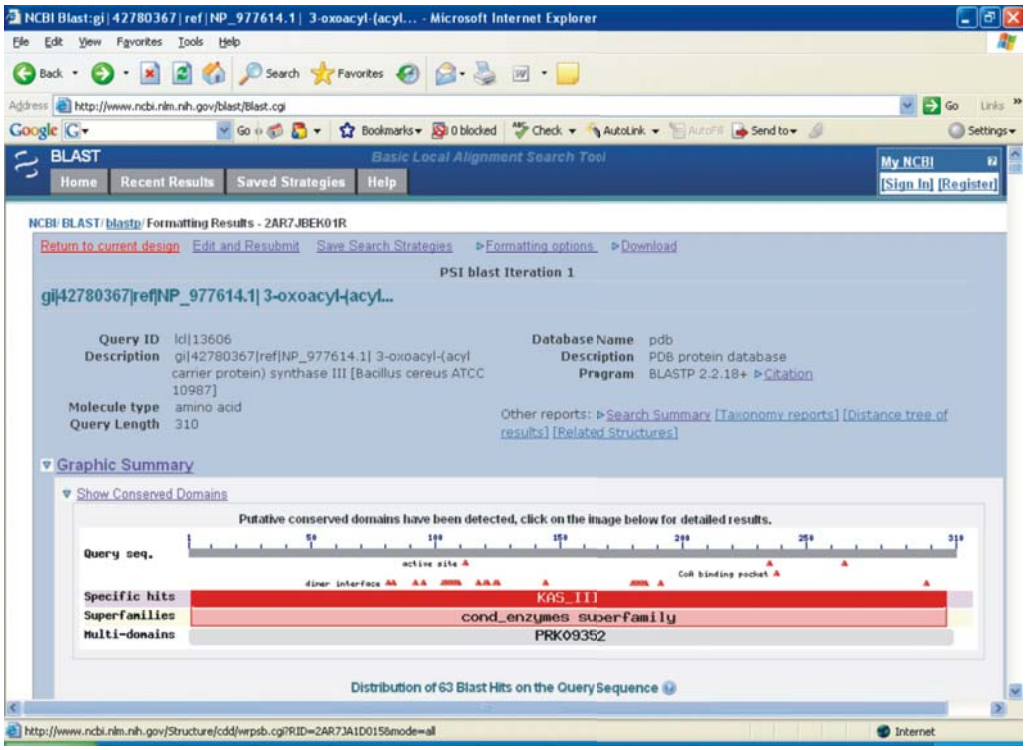
Sequence data are compared with one another using the Basic Local Alignment Search Tool (BLAST). Using PSI-BLAST obtain the FASTA sequence of various pathogenic gram-positive and gram-negative bacteria



c) Domain of FabH in Bacillus cereus

Pfam database is used to obtain the domain details of the protein.

Below is a screenshot showing the domain details of FabH of Bacillus cereus..



Pfam-A Matches

Show or hide all alignments.

Pfam-A	Description	Entry type	Sequence		HMM		Bits score	E-value	Alignment mode
			Start	End	From	To			
Thiolase_C	Thiolase, C-terminal domain	Domain	58	76	27	45	11.2	0.0025	fs
ACP_syn_III	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III	Domain	106	184	1	86	167.2	4.4e-47	ls
Gcd10p	Gcd10p family	Family	123	158	1	41	3.5	0.5	fs
ACP_syn_III_C	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III C terminal	Domain	219	308	1	90	190.3	5e-54	ls
PPV_E2_N	E2 (early) protein, N terminal	Family	242	255	1	14	5.3	0.44	fs

d) Download PDB files of similar proteins

The structurally solved proteins will have a Brookhaven format of their sequence given in a PDB file. This is one of the supported formats of the protein sequences similar to FASTA. To download PDB files and to find their ligand, go to PDB database site and search based on protein PDB id.

The Ligand information given by the PDB database is of high importance in predicting the catalytic or the active site of the protein.

Table-1. Ligand information given by PDB database

PDB Code	Percentage	Length	Ligand
1MZJ	36%	339	ACETYL GROUP, COENZYME A
2AJ9	39%	356	No ligand
2AHB	39%	356	No ligand
1U6S	40%	335	DODECYL-COA
1M1M	40%	355	No ligand
1HZP	40%	335	LAURIC ACID , GLYCEROL
1HNH	44%	317	COENZYME A
1EBL	44%	317	COENZYME A
1MZS	44%	317	1-(5-CARBOXYPENTYL)-5-(2,6-DICHLOROBENZYLOXY)- 1H-INDOLE-2-CARBOXYLIC ACID, PHOSPHATE ION
1HN9	45%	317	PHOSPHATE ION
1UB7	49%	322	GLYCEROL
2EBD	51%	309	No ligand
1ZOW	61%	313	No ligand

e) Model the target protein

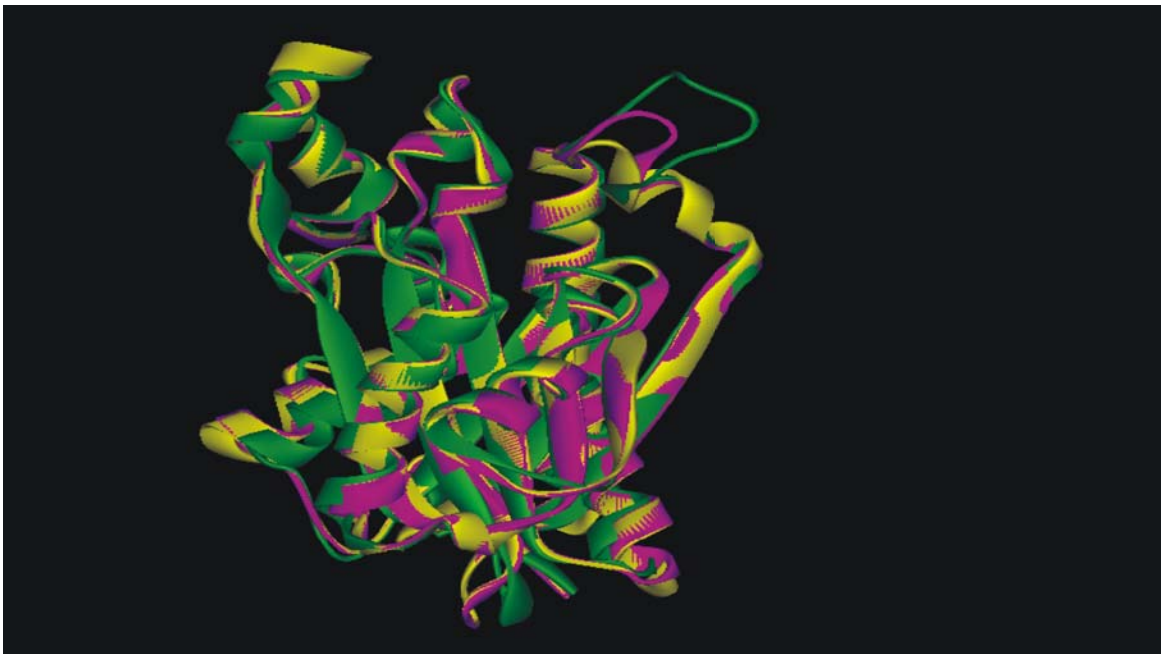
Protein can be modeled using software called modeller9v3.A given number of models are created for the target by comparing with the solved protein. The target sequence is aligned and then modeled. The energy minimization values of the models are calculated and they are given as an output of the model action. Using the most identical sequence & its PDB, model the target using Modeller. Choose the one with lowest E-value in the obtained number of models. The energy calculation output of the modeller is as follows,

Filename	Molpdf
FAB.B99990001.pdb	1741.84753
FAB.B99990002.pdb	1675.17114
FAB.B99990003.pdb	1617.06946
FAB.B99990004.pdb	1660.35107
FAB.B99990005.pdb	1690.41553

The file with lowest energy value is taken. Here the lowest value is 1617.06946. Hence FAB.B99990003.pdb is cho Asen as the best model.

f) Rasmol view of the target protein

PDB file containing the comparative sequence details that can give a comparative structure can be downloaded as an output of the previous step. This PDB file can be viewed in Rasmol that give a comparative structure view.



Target -

1EBL -

1ZOW -

g) Multiple sequence alignment of the proteins

CLUSTALW is the tool used here for multiple sequence alignment.
The three proteins 1ZOW, 1EBL and the target are done multiple sequence alignment using CLUSTALW and the output is captured.

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	gi|42780367|ref|NP_977614.1|	310	2	gi|75765832|pdb|1ZOW|A	313	61
1	gi|42780367|ref|NP_977614.1|	310	3	gi|7245808|pdb|1EBL|A	317	45
2	gi|75765832|pdb|1ZOW|A	313	3	gi|7245808|pdb|1EBL|A	317	39

PLEASE NOTE: Some scores may be missing from the above table if the alignment was done using multiple CPU mode. Please check the output.

Sort by | Sequence Number | View Output File

Alignment

Hide Colors | View Alignment File

```
CLUSTAL 2.0.5 multiple sequence alignment

gi|42780367|ref|NP_977614.1|      MNVGILGIGRYVPEKVVTNHDLEKIMDTSDEWIRTRTGIAERRIADDTID 50
gi|75765832|pdb|1ZOW|A            MNVGIKGFGAYAPEKIIDNAYFEQFLDTSDEWISKMTGIKERHWADDDQD 50
gi|7245808|pdb|1EBL|A             XYTKIIGTGSYLPEQVRTNADLEKXVDTSDEWIVTRTGIRERHIAAPNET 50
                                    . * * * * **::   *  ::*: :******* . *** **: *

gi|42780367|ref|NP_977614.1|      TSYMAVEASKKALEDAGISGEDIDLILVATVTPDRAFPAVACVIQEAIGA 100
gi|75765832|pdb|1ZOW|A            TSDLAYEASVKAIADAGIQPEDIDMIIVATATGDMPFPTVANMLQERLGT 100
gi|7245808|pdb|1EBL|A             VSTXGFEAATRAIEXAGIEKDQIGLIVVATTSATHAFPSAACQIQSXLGI 100
                                  .*  . **: :*:  ***. ::*.:*:***.:   .**:.*  :*. :*

gi|42780367|ref|NP_977614.1|      KHAAAMDLSAACAGFMYGMITAQQFIQTGTYKNILVVGSDKLSKIVDWND 150
gi|75765832|pdb|1ZOW|A            GKVASMDQLAACSGFMYSMITAKQYVQSGDYHNILVVGADKLSKITDLTD 150
gi|7245808|pdb|1EBL|A             KGCPAFDVAAACAGFTYALSVADQYVKSGAVKYALVVGSDVLARTCDPTD 150
                                     .::*  ***:** *.: .*.*::::*  :  ****:* *::  * .*

gi|42780367|ref|NP_977614.1|      RNTAVLFGDGAGAIVMGAVSEGKGVLSFELGADGSGGKHLYQ-------D 193
gi|75765832|pdb|1ZOW|A            RSTAVLFGDGAGAVIIGEVSEGRGIISYEMGSDGTGGKHLYL-------D 193
gi|7245808|pdb|1EBL|A             RGTIIIFGDGAGAAVL-AASEEPGIISTHLHADGSYGELLTLPNADRVNP 199
                                  *.* ::******* ::  .**  *::* .: :**: *: *

gi|42780367|ref|NP_977614.1|      EY---VMMNGREVFKFAVRQLGDSCLRVLDKAGLTKEDVDFLVPHQANIR 240
gi|75765832|pdb|1ZOW|A            KDTGKLKMNGREVFKFAVRIMGDASTRVVEKANLTSDDIDLFIPHQANIR 243
gi|7245808|pdb|1EBL|A             ENSIHLTXAGNEVFKVAVTELAHIVDETLAANNNDRSQLDWLVPHQANLR 249
                                  :    :   *.****.**  :..   ..:   .   .::* ::*****:*

gi|42780367|ref|NP_977614.1|      IMESARERLNLPQEKMSMTIEKFGNTSASSIPIAMVEELQNGRIQDGDLI 290
gi|75765832|pdb|1ZOW|A            IMESARERLGISKDKMSVSVNKYGNTSAASIPLSIDQELKNGKLKDDDTI 293
gi|7245808|pdb|1EBL|A             IISATAKKLGXSXDNVVVTLDRHGNTSAASVPCALDEAVRDGRIKPGQLV 299
                                  *:.:: ::*. . ::: ::::: .*****:*:* :: : :::*::: .: :

gi|42780367|ref|NP_977614.1|      ILVGFGGGLTWGAVALRWGK 310
gi|75765832|pdb|1ZOW|A            VLVGFGGGLTWGAMTIKWGK 313
```

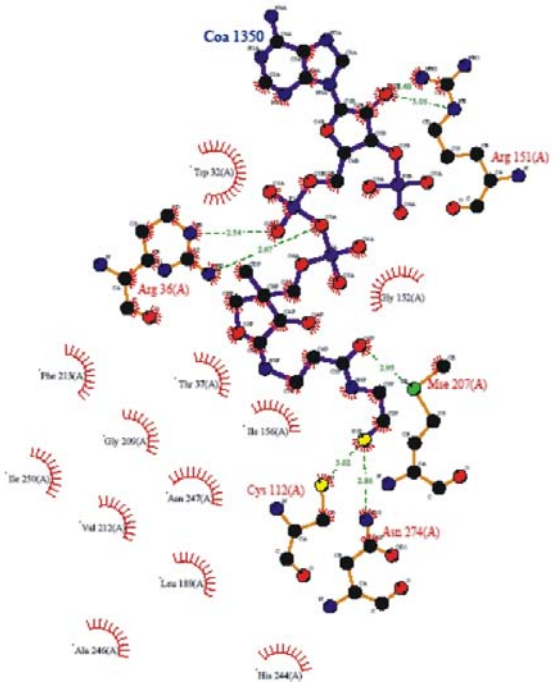
The proteins whose structures where compared are to be done a multiple sequence alignment. This may show the highly conserved regions of the protein and their alignment.

h) Obtain the active sites

As the structure and the sequence of the similar proteins are aligned already with the protein, it is necessary to confirm that their active sites are similar. PDBSum can be used to predict the active sites of the proteins whose structure are already solved. The active site and the catalytic sites are marked for the protein in PDBSum from which we can identify the sites.The already aligned sequences of the protein from CLUSTALW can now be marked with the active and catalytic site from PDBSum that can confirm the similarity in their conserved regions. The active site positions can be checked again the SSM server output for the corresponding position in the target protein for its amino acid along with its position in the chain.

I) Obtain the ligplot of the protein

The ligand binding site of the target protein can be identified from the ligand binding site of the protein with which it is compared. Ligplot is available for structurally solved proteins. This can be obtained from PDBSum which gives a clear picture of the ligand binding site of the comparing protein. Among 1ZOW and 1EBL, only the latter binds to a ligand. The ligplot of 1EBL is obtained from PDBSum. The output is captured and shown below.



The below table illustrates the same. Some of the ligand binding sites have the same amino acids with matched positions. In most of the positions, the amino acids are same but their positions in the chain vary. This shows that there are deviations in certain catalytic and active sites.

Table-2. Deviations in Catalytic and active sites

1EBL	1ZOW	FABH
Trp32	Trp32	Trp32
Arg36	Met36	Arg36
Thr37	Thr37	Thr37
Cys112	Cyst112	Cys112
Phe213	Phe207	Phe204
Gly209	Gly203	Gly200
Ile156	Leu156	Leu156
His244	Ile244	Ile241
Asn247	Asn241	Asn238
Val212	Val206	Val203
Leu189	Leu190	Leu190
Ala246	Ala240	Ala237
His244	His238	His235
Arg151	Arg151	Arg151
Gly152	Ser152	Asn152
Mse207	Met201	Met198
Asn274	Asn268	Asn265

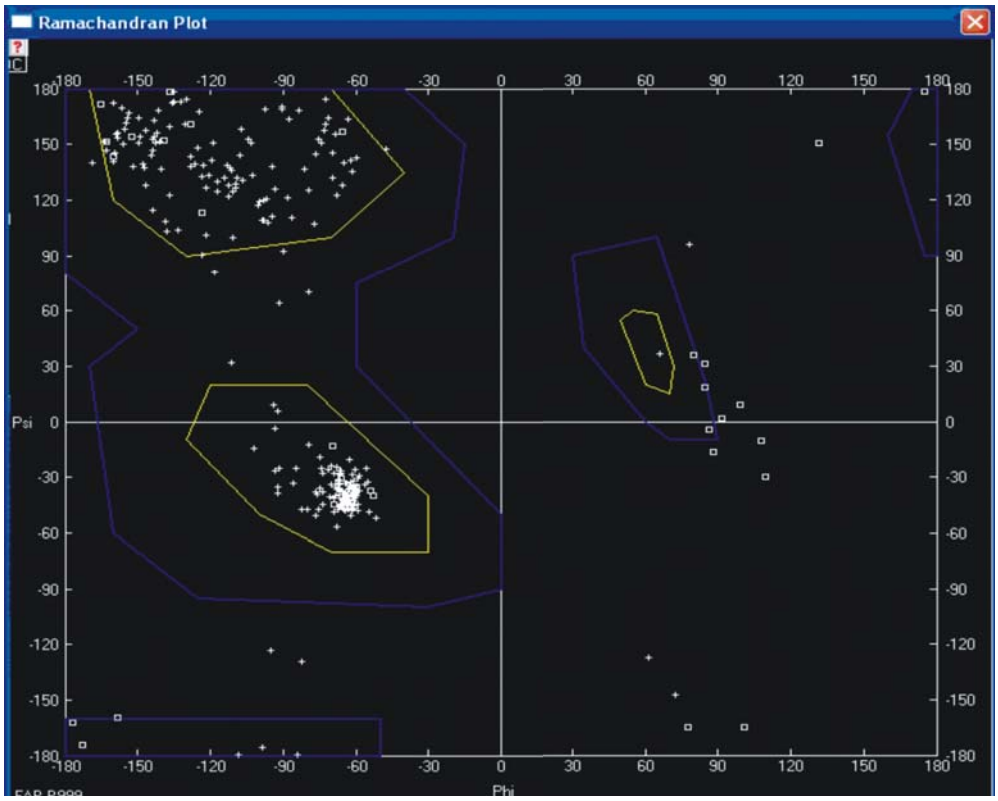
j) Calculation of deviation

The ligand binding can be viewed markedly with the Rasmol software. The exact binding positions of the ligand can be seen. The two compared proteins can be seen compared for their binding activity of the ligand. The one that does not bind to ligand shows a slightly deviating position from the one that actually binds to the ligand.The deviations can be better studied if the phi and psi angles of the proteins are compared for each position of deviation.This helps us to determine the phi and psi angles for various terminal positions. From this a comparative study of the deviations can be performed for the target protein. The Table - 3 below illustrates the phi and psi angle differences between 1ZOW and 1EBL.

Table - 3 phi and psi angle differences between 1ZOW and 1EBL.

	1ZOW		1EBL	
Arg151 (ARG151)	-49.1	-37.4	-50.9	-42.8
Phe207(Phe213)	-55	-53.6	-51	-53.1
His238(His244)	-61.4	132.4	-61.9	136.3
Ile244(Ile250)	-63.1	-47	-61.2	-50.3
Asn241(Asn247)	-175.4	134.6	-179.6	160.8

k) Obtain Ramachandran Plot



This shows that most of the residues are in the allowed region in the target protein.

I) Comparative study between various pathogenic bacteria

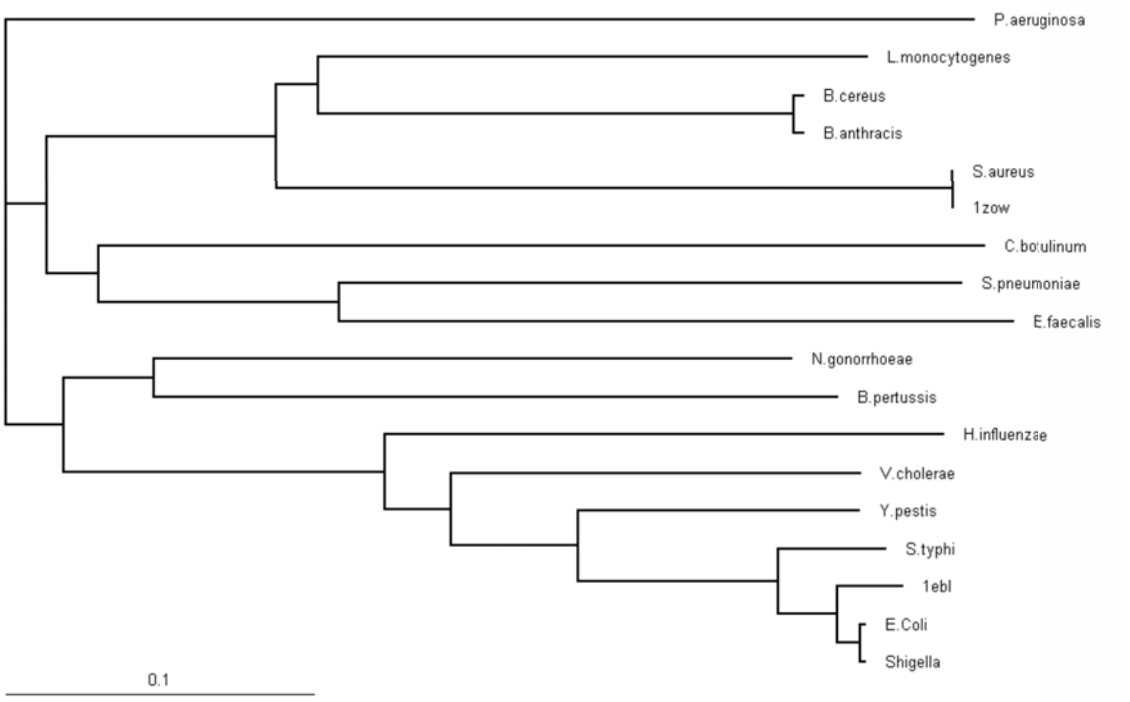
To show the similarity in the FabH sequence of various pathogenic bacteria, a multiple sequence alignment is to be done. The obtained sequences along with protein length, identity and E-value are tabled to give a clear view of the comparison.The protein sequence is blasted against the organisms and it can be detailed as below.

Table-4. Protein details in various pathogenic bacteria

Organism	Protein length	Identity %	E-Value
staphylococcus aureus	313	61%	3.00E-115
Streptococcus pneumoniae	324	44%	1.00E-74
Bacillus anthracis	308	99%	1.00E-180
Listeria monocytogenes	312	66%	4.00E-125
Enterococcus sp.	321	43%	9.00E-71
Clostridium botulinum	326	44%	3.00E-75
Neisseria gonorrhoeae	320	46%	7.00E-72
Bordetella pertussis	328	44%	3.00E-78
Haemophilus influenzae	316	44%	5.00E-76
Helicobacter pylori	331	42%	4.00E-78
Escherichia Coli	317	45%	2.00E-75

A treeview structure of the alignment is obtained and the output is captured as below.

From the treeview structure we can predict the phylogeny of the protein.



The sequences seem to be highly similar. Also the phylogeny tree shows that they are highly conserved.

IV. CONCLUSION

Various tools and databases have been utilized and the outputs are analysed. The target protein sequence was blasted against various pathogenic bacteria. Also the sequences with high similarity were found by blasting against PDB. The sequences with E-value less than 1 and percentage-of-identity greater than 40% were considered. The domain of the target sequence is obtained which has CoA binding sites. The sequence with highest similarity was taken and the protein was modeled. From the modeller results, the structure with lowest energy calculations was considered. As the most similar protein does not contain a ligand, the CoA ligand binding protein with high similarity is also considered for structural comparison. Multiple sequence alignment of the three sequences was obtained and their active sites remain conserved. Rmsd is calculated based on the alignment of the three structures. The output showed the deviations in the amino acid positions whose phi and psi angle deviations were further studied and the Ramachandran Plot was obtained. Thus from the results it shows that the obtained structure is reliable of FabH of Bacillus cereus is reliable. Also FabH sequence of various pathogenic bacteria shows their phylogeny. From the results we conclude that the obtained structure of FabH of Bacillus cereus is stable and reliable. So it can be a key for further studies on this protein. It can also be a useful target for drug discovery.

Mr Sameer Hassan is currently working as Scientist B, ICMR- Biomedical Informatics Centre at Tuberculosis Research Centre, Chennai. He has also obtained Advance Diploma in Proteomics and Molecular Modeling from GVK Bioscience, Hyderabad. He is currently involved in genome and proteome analysis of *Mycobacterium tuberculosis* and Mycobacteriophages. Comparative genomics, Sequence analysis, Protein modeling and ligand interaction are his areas of interest.