ISSN: 2319-8753

International Journal of Research in Science,

Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

Function and Structure Prediction of Rv2004c, a Hypothetical Protein from *M.tuberculosis*

V.G.Shanmuga Priya [1*],Dr.U.M.Muddapur[1], Megha Mehta[2]

R &D Center, Assistant Professors [1*,2], Department of Biotechnology, KLE Dr.M.S.Shesgiri college of Engineering and Technology,Belgaum-8, Karnataka, India.[1, 2]

Professor, Department of Biotechnology, KLE Dr.M.S.Shesgiri college of Engineering and Technology,Belgaum-8, Karnataka, India [1]

Abstract: The cell envelope of Mycobacterium tuberculosis (Mtb) is composed of a matrix of peptidoglycan, mycolic acids, lipids, and carbohydrates along with some proteins. The characterization of proteins associated with the cell wall is presently needed to understand the bacteria in terms of its survival, immune modulation in the host, resistant to drugs etc ,so as to find a way to combat tuberculosis disease. In this study, a cell wall hypothetical protein Rv2004c was taken for further characterization and functional annotation using various Bioinformatics tools like FASTA, STRING, InterproScan etc. Also insilico modeling of this protein was carried out with Robetta, an automated modeling tool and SAVS, a validation server and finally a valid model was obtained.

Keywords: Rv2004c, Hypothetical protein, FASTA, STRING, Robetta

I. INTRODUCTION

Though the TB mortality rate has decreased 41% since 1990 and the world is on track to achieve the global target of a 50% reduction by 2015, there were an estimated 8.7 million new cases of TB (13% co-infected with HIV) and 1.4 million people died from TB in 2011. Also TB is one of the top killers of women. . In the two countries with the largest number of cases, India and China, scale-up of MDR-TB cases is expected in the next three years.(WHO – Global Tuberculosis report,2012).Despite the availability of effective chemotherapy and a moderately protective vaccine, this health crisis is exacerbated by the emergence of MDR(Multidrug- resistant), XDR(extensively drug resistant) and TDR(Totally drug-resistant) strains and this has urged for the necessity of finding new drug targets and new anti-tuberculosis agents to decrease the incidence of tuberculosis disease. This necessitates the need to thoroughly study the genome and proteome content of the causative agent M. tuberculosis. Despite ever-increasing amounts of biological data, including primary data, such as genomic sequences, and functional genomic data from high-throughput experiments, there is a deficiency in functional annotation for many proteins and they are labeled as hypothetical proteins. If the protein is homologous to proteins of unknown function in other organisms they are typically referred to as "conserved hypothetical" proteins. Specifically, about 40% of the Mycobacterium tuberculosis genome is made up of proteins of unknown functions thus limiting our understanding of virulence and pathogenicity of this organism.[1]

Rv2004c is a conserved hypothetical protein of M.tuberculosis.[2](http://tuberculist. epfl.ch).Rv2004c gene transcription was observed in all of Mycobacterium complex strains except Mycobacterium bovis and Mycobacterium microti. Its expression is also noted in many other bacterium species. Rv2004c protein expression in Mycobacterium tuberculosis was confirmed by using antibodies which were able to recognize a 54-kDa molecule by immunoblotting, and its location was detected on the M. tuberculosis surface by transmission electron microscopy, suggesting that it is a mycobacterial surface protein.[3] Also in another study, Rv2004c was identified in the cell wall fraction of M. tuberculosis H37Rv using 2DLC/MS.[4] In this study, characterization and functional annotation of a hypothetical protein RV2004c of Mycobacterium tuberculosis is carried out using bioinformatics tools and also insilico modeling of this protein is done.

ISSN: 2319-8753

International Journal of Research in Science, Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

II.MATERIALS AND METHODS

A. *Retrieval of protein sequence*

UniProt is a database of protein sequence and functional information, many entries being derived from genome sequencing projects. UniProt provides four core databases: UniProtKB (with sub-parts Swiss-Prot and TrEMBL), UniParc, UniRef, and UniMes.[5] Rv2004c protein sequence is retrieved from this database.

B. *Homology-based annotation transfer*

1) Blastp tool: BLAST tool, a database searching tool under NCBI, finds regions of local similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. It follows Heuristic algorithm.[6]The sequenceof Rv2004c was submitted to this tool for search of homologous protein in database for function prediction.

2) FASTA tool: FASTA does a database search.The FASTA program follows a largely heuristic method which contributes to the high speed of its execution. It initially observes the pattern of word hits, word-to-word matches of a given length, and marks potential matches before performing a more time-consuming optimized search using a Smith-Waterman type of algorithm.[7]The Protein sequence of Rv2004c was submitted as query and Swiss prot and NCBI RefSeq database search was done with default parameter values.

C. *Characterisation and functional annotation of the protein*

Sequence of Rv2004c is submitted to the these tools given below for characterization.

1) Characterisation of physico-chemical properties*:* ProtParam tool under Swiss Institute of Bioinformatics, computes various physico- chemical properties that can be detected from a protein sequence. The parameters computed by ProtParam include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). [8]

2) Predicting the presence of Transmembrane spanning region: TMHMM is a membrane protein topology prediction method based on a hidden Markov model. It can discriminate between soluble and membrane proteins with both specificity and sensitivity better than 99%, although the accuracy drops when signal peptides are present.[9]

3) Domain prediction from SMART database: SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signalling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Here, Transmembrane segments as predicted by the TMHMM2 program ,coiled coil regions determined by the Coils2 program ,segments of low compositional complexity determined bythe SEG program, Signal peptides determined by the SignalP program. In Normal SMART, the database contains Swiss-Prot, SP-TrEMBL and stable Ensembl proteomes.[10]

4) Domain prediction from CDD : CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases -Pfam, SMART, COG, PRK,TIGRFAM. (http://www.ncbi.nlm.nih.gov/cdd)

5) InterproScan search: InterProScan is a tool that combines different protein signature recognition methods into one resource.A number of different protein sequence applications are launched. These applications search against specific databases and have preconfigured cut off thresholds.The databases searched are TMHMM,Signal peptides, ProDom, PRINTS, HMMPIR, HMMPfam, HMMSmart, HMMTigr, PROSITE profiles, HAMAP profiles.[11]

6) Assessing Protein–protein associations: The STRING database[12] integrates known and predicted protein- protein associations derived from high-throughput experimental data, the mining of databases and literature, and from predictions based on genomic analysis for a large number of organisms. Functional interactions from the STRING database are used with confidence scores as defined by the STRING schemes. These include conserved genomic neighbourhood, gene fusion events, phylogenetic profile, or gene cooccurrence across multiple genomes, text mining, experiments, and other databases (http://string-db.org/).

ISSN: 2319-8753

International Journal of Research in Science, Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

D. *Deriving the model of the protein*

1) Modeling by automated tool: The Robetta server (http://robetta.bakerlab.org) provides automated tools for protein structure prediction and analysis. For structure prediction, sequences submitted to the server are parsed into putative domains with the Ginzu protocol and structural models are generated using either comparative modeling or de novo structure prediction methods. If a confident match to a protein of known structure is found using BLAST, PSI-BLAST, FFAS03 or 3D-Jury, it is used as a template for comparative modeling. If no match is found, structure predictions are made using the de novo Rosetta fragment insertion method. Robetta was among the top performers in CASP-5, and CAFASP-3 for 'Fully Automated experiments' assessments.[13]The sequence of Rv2004c was submitted to it.

2) Verifying the validity of the Templates:The template structure database used by SWISS-MODEL (SMTL or ExPDB library) is derived from the Protein Data Bank. In order to allow sequence-based template searches, each PDB entry is split into individual chains. The separated template chains are annotated with information about experimental method, resolution , ANOLEA mean force potential, Gromos96 energy and PQS quaternary state assignment to allow for rapid retrieval of the relevant structural information during template selection. Theoretical models, structures only consisting of C alpha atoms and irregularly formatted database entries are removed. To detect distantly related template structures, a target sequence can be searched against a Hidden Markov Model based template library(HHSearch). Each HMM of the library is based on a multiple sequence alignment of the template sequence built by PSI-BLAST search (against nr90 & nr70) enriched with secondary structure assignment. Only alignments which score more than a given P-value cut-off are reported. [14]The templates used by Robetta server was validated ,by doing template search using this tool.

3) Energy minimizatio*n:* Swiss-PdbViewer, a visualization tool includes a version of the GROMOS 43B1 force field. This force field allows to evaluate the energy of a structure as well as repair distorted geometries through energy minimization. All computations are done in vacuo, without reaction field. During an energy minimization (200 cycles of Steepest Descent), the geometry is repaired to reach the minimum. [15] Five Models constructed from Robetta tool are analysed, selected and submitted for energy calculation and minimization and energy minimized structures were saved.

4) Model validation : SAVS(Structural Analysis and Verification Server (http://nihserver.mbi.ucla.edu/SAVES/) carries out the validation of the Model with 5 programs. Procheck and What_check programs checks the stereochemical quality of a protein structure. ERRAT analyzes the statistics of non-bonded interactions between different atom types in the model and compares with statistics from highly refined structures. Verify_3D Determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D). PROVE Calculates the volumes of atoms in macromolecules and calculates a statistical Z-score deviation from refined PDB-deposited structures.All 5 models predicted from Robetta server and their 5 energy minimized conformations are all submitted to this tool for validation.

III. RESULTS AND DISCUSSIONS

A. *Retrieval of protein sequence –Uniprot database*

From UniprotKB database the protein is retrieved with Id: P0A5F9 (Y2004_MYCTU) .The gene names of this protein are Rv2004c and MT2060.The length of the protein sequence is 498 AA.The NCBI Id of this protein is GI:15609141 (NP_216520).

B. *Homology-based annotation transfer*

1) Blastp tool: When the sequence alone was submitted for BlastP search, against non redundant database the hits were mostly hypothetical proteins . The hits were ranked based on decreasing bit score and increasing e-value. First few hits were hypothetical protein from Mycobacterium tuberculosis strains CDC1551, H37Rv, CCDC5079 etc.The other hypothetical protein were from many Mycobacterium species. The proteins which have max.identity of 50% and above are hypothetical proteins from Mycobacterium canettii, Mycobacterium marinum , Mycobacterium parascrofulaceum, Mycobacterium abscessus, Mycobacterium fortuitum, Mycobacterium thermoresistibile, Mycobacterium gilvum PYR-GCK, Mycobacterium smegmatis, Mycobacterium vanbaalenii ,Mycobacterium hassiacum and Mycobacterium vaccae. All these proteins seem to have phosphotransferase domain of Pfam id:01636, AAA domain and regions belonging to COG 2187 and COG 0645 which are Uncharacterized protein conserved in

ISSN: 2319-8753

International Journal of Research in Science,

Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

bacteria with unknown function and Predicted kinase respectively.Many proteins also had "NK" region and is known that Nucleoside/nucleotide kinase (NK) is a protein superfamily consisting of multiple families of enzymes that share structural similarity and are functionally related to the catalysis of the reversible phosphate group transfer from nucleoside triphosphates.

The other hits were, possible Adenylate kinase in Mycobacterium liflandii , AAA ATPase central domain protein in Mycobacterium rhodesiae, and a Mycobacterium rhodesiae protein with PFAM: Zeta toxin domain.

The other proteins which have identity less than 50% were also mostly hypothetical proteins from Mycobacterium species and Frankia sp, Thermomonospora curvata etc.Also these hypothetical protein seems to have similar properties as to the above said hypothetical protein.The other main hit was gluconate kinase from Dietzia cinnamea, Rhodococcus ruber, Rhodococcus pyridinivorans, Thermomonospora curvata etc. These gluconate kinases have NK region which was discussed earlier.

2) FASTA search: Adenylate Kinase from E.coli (214 aa) was the first hit in FASTA search. Adenylate Kinase catalyzes the reversible transfer of the terminal phosphate group between ATP and AMP. This small ubiquitous enzyme involved in the energy metabolism and nucleotide synthesis, is essential for maintenance and cell growth. (ATP + AMP = 2 ADP). Region 326-449 of the query is aligned with 5-123 of this E.coli protein with Smith-Waterman score: 120; 31.8% identity (57.6% similar) in 132 aa overlap (322-449:1-123).

The next hit was Nif specific regulatory protein (615 aa). This protein is required for activation of most nif operons, which are directly involved in nitrogen fixation, interacts with sigma-54. Though the query sequence is aligned to this protein to a stretch of 324 aa , the alignment is parse with many gaps.This may not be a signifant hit as it is a cytoplasmic protein.

One notable hit is Annexin A4 of pig and also Bovine Annexin 4. Annexins are a family of proteins that bind to phospholipids in a calcium-dependent manner. More than a thousand proteins of the annexin superfamily have been identified in major eukaryotic phyla, but annexins are absent from yeasts and prokaryotes. The unique annexin core domain is made up of four similar repeats approximately 70 amino acids long, each of which usually contains a characteristic 'type 2' motif for binding calcium ions. and they can act as atypical calcium channels. A pair of annexin repeats may form one binding site for calcium and phospholipid. Two of the 4 repeats of Pig Annexin 4 aligns to the query with Smith-Waterman score: 93; 23.9% identity (55.7% similar) in 176 aa overlap (88-259:5-166),while bovine anexin with Smith-Waterman score: 91; 24.7% identity (57.5% similar) in 146 aa overlap (88-232:5-141)

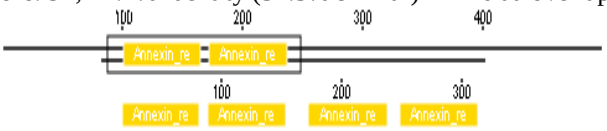


FIG 1: QUERY SEQUENCE(RV2004C) WITH PIG ANNEXIN IV ALIGNMENT.

The other hits were Helicases, Glutamyl-tRNA reductase from different organisms. But the E-value is comparatively high and score low, covering a short segment of similarity to the query sequence.

C. *Charaecterisation and functional annotation of the protein*

1) Characterisation of physico-chemical properties-Protparam Tool:The ProtParam tool calculated the molecular weight as 54422.5 Dalton and Theoretical pI to be 5.89.The Protein was found to contain more of alanine(14%),Arginine(10.2%) and valine(9.2%) residues. The number of negatively charged residues (Asp + Glu) is 73 and that of positively charged residues (Arg + Lys) is 62 which indicate the acidic nature of the protein.

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, for this protein it is found to be greater than 10 hours in Bacterial organism .The instability index of the protein was calculated to be 30.42, which indicates that the protein will be stable in the test tube as the predicted value is below 40. The average hydropathicity value of the amino acids in the protein was found to be -0.157 indicating that the protein is hydrophilic in nature. These parameters will be applicable in wet lab experiments.

2) Predicting the presence of Transmembrane spanning region-TMHMM: In the query sequence, the tool predicted no transmembrane helices. In the prediction, the expected number of amino acids in the transmembrane helices was 1.33878, where 18 amino acids are required to be predicted as a transmembrane protein. This indicates the absence of transmembrane helices in the protein.

3) Domain prediction from SMART database: In Normal SMART, the database contains Swiss-Prot, SP-TrEMBL

ISSN: 2319-8753

International Journal of Research in Science,

Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

and stable Ensembl proteomes. The query had no hits from SMART database. When Pfam search was also included, the sequence is predicted to have a hit from Pfam database, the Zeta_toxin from position 308-451 with E-value 1.70e-02 and a low complexity region from 456 to 469 comprising of 14 aa- ATAEIAAALAARQA. The GO function annotation for this Pfam hit is kinase activity (GO:0016301) and ATP binding (GO:0005524). GO:0016301 stands for 'Catalysis of the transfer of a phosphate group, usually from ATP, to a substrate molecule.'GO:0005524 stands for ' Interacting selectively and non-covalently with ATP, adenosine 5'-triphosphate, a universally important coenzyme and enzyme regulator.'

In molecular biology, the protein domain Zeta (ζ) toxin refers to a protein domain found in prokaryotes. Its function is to inhibit cell wall biosynthesis and it may act as a bactericidein nature. Zeta toxin is thought to be part of a postsegregational killing (PSK) system involved in the killing of plasmid-free cells. It relies on antitoxin/toxin systems that secure stable inheritance of low and medium copy number plasmids during cell division and kill cells that have lost the plasmid.The Zeta Toxin is like a phosphotransferase because the P-loop is located between β-strand 1 and the adjacent helix B, similarly found in protein kinases and bacterial phosphotransferases.. This domain features an α/β structure and the central twisted β-sheet contains six β-strands. The first 5 strands are parallel but β-strand 6 is antiparallel and connected by a short loop to β-strand 5. α-Helices are inserted between and flank the β-strands. Several observations lead us to propose that the ζ phosphotransferase toxin induces a set of protective responses that facilitate entry into dormancy.[16],[17]

TABLE I: HITS FROM SMART DATABASE

Name	Begin	End	E-value
Pfam:Zeta toxin	308	451	1.70e-02
Low complexity	456	469	-

Name	Begin	End	E-value
Blast:AAA	323	446	0.00e+00
SCOP:d2ak3a1	323	477	3.00e-10

4) Domain prediction from CDD : In CDD(Conserved domain database) ,on the query sequence three regions were identified. They are COG2187 and COG0645 multidomain regions with 2.21e-157 and 3.93e-53 E-values respectively and AAA_17 with E-value 3.74e-05 as shown in the fig(2).COG2187 belongs to uncharacterized protein conserved in bacteria, whose function is unknown.COG0645 belongs to predicted kinase, which is a general function prediction only.AAA -17 belongs to Pfam 13207-AAA domain which belongs to ATPase family associated with various cellular activities including membrane fusion, proteolysis and DNA replication.

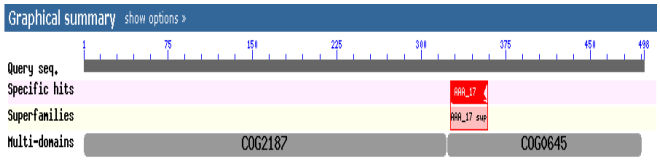


FIG.2:PREDICTION FROM CDD TOOL

5) InterproScan search:This tool has predicted two regions from the Interpro database and one from Pfam database.

ISSN: 2319-8753

International Journal of Research in Science, Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

TABLE II: HITS FROM INTERPROSCAN

InterPro	Molecular Function	location	E-Value
IPR011009 Protein kinase-like domain	transferase activity, transferring phosphorus-containing groups (GO:0016772) \	T[23-286]	3.4e-12
IPR027417 P-loop containing nucleoside triphosphate hydrolase	-	T[323-451]	7e-20
(HMMPfam prediction) PF13671 AAA_33	-	T[325-469]	4.3e-17

6) Assessing Protein –protein associations:In string when the raw sequence of Rv2004c was submitted, hit was from M.bovis and M.tuberculosis strains H37Rv, H37Ra,1435,CDC1551 and F11.Analysing the interaction network in these hits it is known that Rv2005c, Rv2006, MT2611,ahcY, ctpH are the proteins with known type which are associated to Rv2004c .Other proteins were hypothetical proteins.

TABLE III: *PROTEIN ASSOCIATED WITH RV2004C WITH THEIR NATURE OF RELATION*

Protein	**Related by:**
Rv2005c -Universal stress protein Rv2006 -trehalose-6-phosphate phosphatase otsB1 Rv2003c-hypothetical protein	gene neighbourhood in organisms of Actinobacteria class
Mtap- 5'-methylthioadenosine phosphorylase ahcY- adenosylhomocysteinase sahH	gene neighbourhood in organisms of chloroflexi class (neighbourhood of achy is also found in archaea)
ctpH – metal cation transporting P-type ATPase	Co occurance pattern
Rv2560(325 aa),MT2780(324 aa),MT3047(255 aa)-hypothetical proteins	Textmining

CtpH is known to be related with Rv2004c by cooccurance pattern. The phylogenetic display provides the phylogenetic profile of the genes. CtpH is present in all class of organism,while Rv2004c is absent in few. The 2 genes have similar expression pattern in organisms of different suborders. Both the genes seem to have high sequence conservation in 6 taxa of suborder: Corynebacterineae with similar pattern of cooccurance.

ISSN: 2319-8753

International Journal of Research in Science,

Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

When similar proteins to Rv2004c were searched, the closest was Mb2027c, a hypothetical protein from M.bovis 1173P2. When the interacting proteins were searched for it, proteins similar to above said hypothetical proteins were detected and were found to be putative transmembrane proteins. They are Mb2993c-putative membrane or secreted protein(255 aa),Mb2726-putative transmembrane alanine and leucine rich protein(324 aa), Mb2590 putative proline and glycine rich transmembrane protein(325 aa).

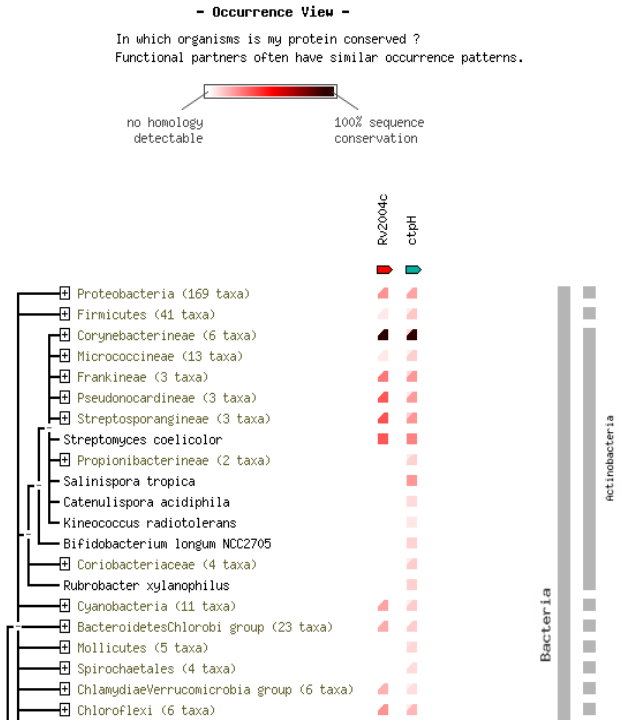


FIG 3: CO-OCCURANCE PATTERN OF RV2004C AND CTPH PROTEINS

D. *Deriving the model of the protein*

1) Modeling by automated tool-Robetta: This automated modeling tool has used two templates-3dxpA :Crystal structure of a putative aminoglycoside phosphotransferase from ralstonia eutropha jmp134(Span:1-297) and 1m8pA:Crystal Structure of P. chrysogenum ATP Sulfurylase in the T-state(span:298-498)and predicted 5 models (job ID :39437) for the query protein.The models were retrieved and saved asRobetta-1upto Robetta-5.

2) Verifying the validity of the Templates-Swiss Model-Template identification: Many templates were identified in the swiss Model template search (Workspace:P2000005). Interpro Scan has identified two regions. First is protein kinase like domain(IPR011009) from position 23 -286.The second region is identified from 322-464 as P-loop containing nucleoside triphosphate hydrolases(noIPR). And also HHsearch template library search has identified many templates. When analysed it was found that templates 3dxpA(82nd Hit) and 1m8pA(12th Hit) had better E-value and P-value while being aligned with a large extent of the query. This confirms the validity of the templates used in Robetta server.

3) Energy minimization - Swiss-PdbViewer: In SPDBV the energy calculations for the models were carried out and Energy minimization procedure was done (20steps of steepest descent) and again total energy is computed . The first calculated total energy of the models were high with positive values. After energy minimization the energy values were less indicating the structure stability. The energy minimized structures were saved as Robetta-1EM upto Robetta-5EM and their energy values are -20196.090,-24837.486,-23707.998,-23588.916 and -24659.938 orderly.

4) Model validation-SAVS :SAVS validation report of the 10 models - Robetta-1 to Robetta-5 got from Robetta prediction and their energy minimized structures from SPDBV, Robetta-1EM to Robetta-5EM were compared and analysed. It was found that of them Robetta-1EM is a good quality model with only few irregularities. From the Procheck results- In Ramachandran plot Thr28, Glu230, His427, Ser496 lie in generously allowed regions. Hence the

ISSN: 2319-8753

International Journal of Research in Science,

Engineering Innovative and Technology

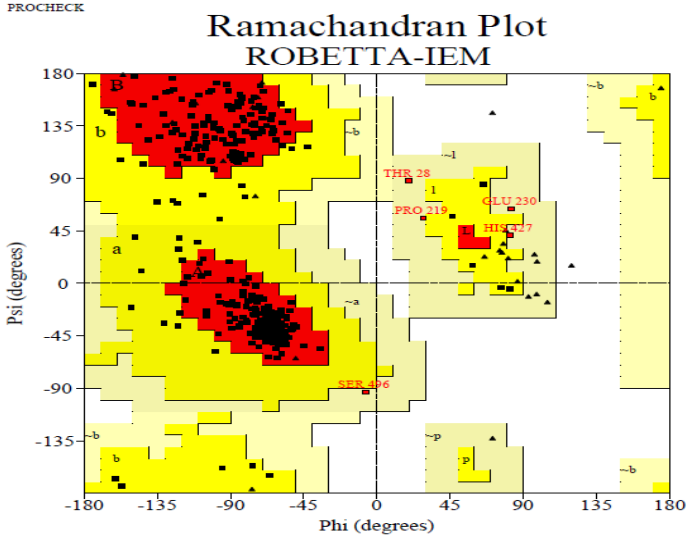
(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

warning message is given. From the Ramachandran plot for all residue types-Ala271, Ala205,Asp36, Gln297, Glu226, Glu230, Glu56, Gly452, His427, Leu234, Phe49, Pro219, Ser87, Ser496, Thr28 and Val322 have unfavorable confirmations. Regarding main chain bond angles, 16 are in off graph. Of them residues Asp229, Thr426 and Ser496 have many deviations(from the ideal) in their angles.The planarity of residues Asp48, Phe49, Asp206, Asp229 and His429 are deviated from the mean values. Of this Phe49 shows a higher planarity deviation. This is the reason for error message for planar group parameter.

Verify_3D predicts that 83.17% of the residues have the 3D-1D score greater than 0.2 and also in Errat the overall quality factor is 94.280 which indicates it as a good model and also What_check and PROVE evaluation analysis predicts this model to be reliable one.

PROCHECK



Plot statistics

Residues in most favoured regions [A,B,L]	402	90.7%
Residues in additional allowed regions [a,b,l,p]	37	8.4%
Residues in generously allowed regions [~a,~b,~l,~p]	4	0.9%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	443	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	39	
Number of proline residues	14	
Total number of residues	498	

FIG 4:RAMACHANDRAN PLOT FOR ENERGY MINIMIZED FIRST MODEL

ISSN: 2319-8753

International Journal of Research in Science, Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

FIG 5:ENERGY MINIMIZED FIRST MODEL FROM ROBETTA SERVER

TABLE IV: SAVES RESULTS FOR ROBETTA-1EM

Parameters	**Model-1**
1) Ramachandran plot:	90.7% in core region ,8.4% in allowed region, 0.9% in generously allowed region 0.0% in disallowed region (Warning)
2) No.of Residues outside the favourable regions in individual 20 AA-Ramachandra plot :	16 labelled residues(out of 496) (ERROR)
3) No.of residues outside the favourable regions in Chi1-Chi2 plots showing the chi1-chi2 sidechain torsion angle combinations for all residue types:	0 labelled residues (out of 275) (Good)
4) No.of residues that deviate from mean for Main-chain parameters*:	6 better 0 inside 0 worse (Good)
5) No.of residues that deviate from mean for side-chain parameters*:	5 better 0 inside 0 worse (Good)
6) Residue properties: (residues that show more than 2.0 standard deviations away from the "ideal" mean value are considered.)	Max.deviation: 8.4, Bad contacts: 03 Bond len/angl:28.8, Morris etal class*: 1 1 1 + 1 (ERROR)
7) G-factors*:	Dihedrals:0.27 Covalent: 0.16 Overall: 0.25 (Good)
8) M/c bond lengths:	99.4% within limits ,0.6% highlighted (Good)
9) M/c bond angles:	97.7% within limits ,2.3% highlighted 16 off graph (Error)
10)Planar groups:	89.6% within limits,10.4% highlighted 1 off graph (Error)

Main-chain parameters

ISSN: 2319-8753

International Journal of Research in Science, Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

a. Ramachandran plot quality. -measured by the percentage of the protein's residues that are in the most favoured, or core, regions of the Ramachandran plot.
b. Peptide bond planarity-measured by calculating the standard deviation of the protein structure's omega torsion angles.
c. Bad non-bonded interactions. This property is measured by the number of bad contacts per 100 residues. Bad contacts are defined as contacts where the distance of closest approach is less than or equal to 2.6Å. etc.
Side-chain parameters:
a. Standard deviation of the chi-1 gauche minus torsion angles.
b. Standard deviation of the chi-1 trans torsion angles.
c. Standard deviation of the chi-1 gauche plus torsion angles.
d. Pooled standard deviation of all chi-1 torsion angles.
e. Standard deviation of the chi-2 trans torsion angles.
***The G-factor** for Torsion angles:- phi-psi combination ,chi1-chi2 combination ,chi1 torsion for those residues that do not have a chi-2 ,combined chi-3 and chi-4 torsion angles ,omega torsion angles. For Covalent geometry:- main-chain bond lengths, main-chain bond angles*
***Morris** et al. deals with Phi-psi distribution, Chi-1 st.dev, H-bond energy st. dev.*

IV. CONCLUSION

In a study by Wolfe et al,528 proteins were extracted from M.tuberculosis cell wall by various extraction procedure. A query of the 528 total protein identifications against Neural Network or Hidden Markov model algorithms predicted secretion signals in 87 proteins. Classification of these 528 proteins also demonstrated that 35% are involved in small molecule metabolism and 25% are involved in macromolecule synthesis and degradation building upon evidence that the Mtb cell wall is actively engaged in mycobacterial survival and remodeling.[18] Some annotation predictions of cell wall proteins indicate that they are expressed based on the changing micro-environments encountered by the pathogen and play an important role in survival and multiplication of MTB in their chosen environment, and even in mediating mycobacterium-host cell interactions [19]–[21].
Here ,the *M.tuberculosis* cell wall protein-Rv2004c is characterized and analysed for its function with various reliable bioinformatics tools. Database similarity search with Blast tool confirmed it is a conserved protein in Mycobacterium genus and also seen in some lineage of Bacteria and it has a Phospho transferase domain and act as a kinase enzyme. FASTA search also revealed the same function and in addition has detected the possibility of two annexin-4 type repeats from residue 89-259. The FASTA hit, Annexin IV (ANX4) belongs to the annexin family of calcium-dependent phospholipid binding proteins. several members of the annexin family have been implicated in membrane-related events along exocytotic and endocytotic pathways. In vitro studies has shown it's possible interactions with ATP and a ability to inhibits phospholipase A2 activity.[22] Protparam tool characterizes the protein and has also predicted its nature to be a acidic, hydrophilic protein.The query sequence is found to have no transmembrane helices and SignalP regions. The absence of these two regions is a common characteristic of many cell wall-outer membrane protein.The zeta-toxin domain hit from SMART, just indicates its phosphotransferase and kinase activity and is not related specifically to the function of the zeta toxin protein as it is a cytoplasmic-plasmid protein. But it can be assumed that Rv2004c will have structural similarity to the zeta-toxin protein.Analysing the protein –protein association from STRING database, the similar cooccurance pattern of Rv2004c with ctpH(a metal cation transporting P-type ATPase) indicate that it might be having a similar function of transporting molecules.Rv2004c is seemed to be expressed more in dormant state.It is predicted to have nucleotide triphosphate hydrolase and kinase activity by binding with ATP.From the above analysis, its function can be narrowed down to inhibition of phospholipase A2 activity and membrane-related events along exocytotic and endocytotic pathways.
To predict the 3D structure of the protein, the sequence of it was submitted to Robetta tool. The energy of the 5 models predicted by it was calculated and energy minimization procedure was carried out with SPDBV tool. The models were validated with SAVS and the output was analyzed. From this the energy minimized first model was validated as a good quality model. But the residues Thr28, Glu230, His427, Ser496,Phe49, Asp229,Thr426 show irregularities. In future, if the essentiability of the protein is known, this model can be taken for further studies provided care has to be taken for the conformation of above said residues.

ISSN: 2319-8753

International Journal of Research in Science, Engineering Innovative and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

REFERENCES

1. Enault, F.,Suhre, K..,Claverie., J.M. Phydbac "Gene Function Predictor:A gene annotation tool based on genomic context analysis". BMC Bioinformatics. 6, doi:10.1186/1471-2105-6-247, 2005.
2. TubercuList--- Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList - 10 years after.Tuberculosis (Edinb). (1): 1–7. doi: 10.1016/j.tube.2010.09.008, 1991.
3. Forero M, Puentes A, Cortés J, Castillo F, Vera R, Rodríguez LE, Valbuena J, Ocampo M, Curtidor H, Rosas J, García J, Barrera G, Alfonso R,Patarroyo MA, Patarroyo ME "Identifying putative Mycobacterium tuberculosis Rv2004c protein sequences that bind specifically to U937 macrophages and A549 epithelial cells." Protein Science. 2005 Nov;14(11):2767-80. Epub 2005 Sep 30.
4. Mawuenyega KG, Forst CV, Dobos KM, Belisle JT, Chen J, Bradbury EM, Bradbury AR, Chen X,."Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling."Mol Biol Cell ,16(1):396-404), 2005
5. The UniProt Consortium-"Update on activities at the Universal Protein Resource (UniProt) in 2013"Nucleic Acids Res. 41: D43-D47 (2013).
6. Altschul, , Gish, W., Miller.W., Myers. E., Lipman, D., "Basic local alignment search tool". Journal of Molecular Biology 215 (3) 403–410, October 1990.
7. Pearson WR, Lipman DJ" Improved tools for biological sequence comparison."Proc Natl Acad Sci U S A. 1988 Apr; 85(8):2444-8
8. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., BairochA.;Protein Identification and Analysis Tools on the ExPASy Server; (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press pp. 571-607, 2005.
9. A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer."Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes". Journal of Molecular Biology, 305(3):567-580, January 2001.
10. Letunic I.Doerks T.Bork P.,"SMART7:recent updates to the protein domain annotation resource",Nucleic Acids Res;doi:10.1093/nar/gkr931, 2012.
11. Zdobnov E.M. and Apweiler R. "InterProScan - an integration platform for the signature-recognition methods in InterPro", Bioinformatics 17: 847-848, 2001.
12. Von Mering C,Jensen LJ,Snel B.et al. "STRING: known and predicted protein –protein associations, integrated and transferred across organisms"., Nucleic Acid Research.:33:D433-D437,2005.
13. Chivian D, Kim DE, Malmström L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. "Automated prediction of CASP-5 structures using the Robetta server". Proteins. 53 Suppl 6:524-33, 2003.
14. Arnold K., Bordoli L., Kopp J., and Schwede T. "The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling". Bioinformatics, 22,195-201, 2006.
15. W.F. van Gunsteren et al. "The GROMOS96 manual and user guide", Vdf Hochschulverlag ETHZ. in Biomolecular simulation,(1996)
16. Virginia S. Lioy, Cristina Machon, Mariangela Tabone, José E. Gonzalez-Pastor, Rimantas Daugelavicius,Silvia Ayora, and Juan C. Alonso " The ζ Toxin Induces a Set of Protective Responses and Dormancy", PLoS One. 7(1): e30282. Published online 2012 January 25. doi: 10.1371/journal.pone.0030282 PMCID: PMC3266247, 2012.
17. Anton Meinhart, Juan C. Alonso, Norbert Sträter,and Wolfram Saenger "Crystal structure of the plasmid maintenance system ε/ζ: Functional mechanism of toxin ζ and inactivation by $\varepsilon_2\zeta_2$complex formation", Proc Natl Acad Sci U S A. 100(4): 1661–1666, 2003.
18. Wolfe L.M., Mahaffey S.B., Kruh N.A., Dobos K.M." Proteomic definition of the cell wall of Mycobacterium tuberculosis" J. Proteome Res. 9:5816-5826, 2010.
19. Brennan, M.J.; Delogu, G.; Chen, Y.; Bardarov, S.; Kriakov, J.; Alavi, M.; Jacobs, W.R, Jr. "Evidence that Mycobacterial PE PGRS Proteins are cell surface constituents that influence interactions with other cells". Infect. Immun. 69, 7326–7333, 2001.
20. Banu, S.; Honore, N.; Saint-Joanis, B.; Philpott, D.; Prevost, M.C.; Cole, S.T. Are the PE-PGRS.,Proteins of Mycobacterium tuberculosis variable surface antigens? Mol. Microbiol., 44, 9–19, 2002.
21. Huang, Y.;Wang, Y.; Bai, Y.;Wang, Z.G.; Yang, L.; Zhao, D. "Expression of PE PGRS 62 protein in Mycobacterium smegmatis decrease mRNA expression of proinflammatory cytokines IL-1_, IL-6 in macrophages". Mol. Cell Biochem. ,340, 223–229, 2010.
22. Tait JF, Smith C, Frankenberry DA, Miao CH, Adler DA, Disteche CM "Chromosomal mapping of the human annexin IV (ANX4) gene".Genomics 12 (2): 313–8. doi:10.1016/0888-7543(92)90379-7.PMID 1346776, Mar 1992.

BIOGRAPHY

V.G.Shanmuga Priya from Vellore,Tamil Nadu, India has done B.Sc (zoology) at Auxilium college, Vellore under Madras University and secured University First Rank(1995) and M.Sc(Bioinformatics) at Auxilium College under Thiruvalluvar University and secured University Third Rank(2005) and M.Phil(Bioinformatics) at Bharathidasan University, Tamil Nadu(2007). Is Presently working as Assistant Professor in Dept of Biotechnology, KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belgaum, Karnataka, India.