Journal of
Proteomics & Bioinformatics

Research Article Open Access

In Silico Structure Modeling and Characterization of Hypothetical Protein YP_004590319.1 Present in *Enterobacter aerogens*

Ritika Gupta[1]*, Ankita Dey[2], Anu Vijan[3] and Bitu Gartia[4]

[1]*Department of Biotechnology, Meerut Institute of Engineering and Technology (MIET), Meerut, India*
[2]*Department of Biotechnology, Haldia Institute of Technology, West Bengal, India*
[3]*University Institute of Biotechnology, Chandigarh University, Punjab, India*
[4]*School of Biotechnology, KIIT University, Bhubaneswar, India*

Abstract

Transfer RNAs anticodon post-transcriptional modifications are responsible to the high fidelity of protein synthesis. In eubacteria, two genome-encoded transfer RNA (tRNA) species bear the same CAU sequence as the anticodons, which are differentiated by modified cytidines at the wobble positions. We have determined the structure model of the hypothetical protein. The structure unexpectedly reveals an idiosyncratic RNA helicase module fused with a GCN5-related N-acetyltransferase (GNAT) fold, which intimately cross interact. The stereo chemical quality of the protein model was checked by using in silico analysis with SWISS- MODEL, PyMol, PROCHECK, ProSA and QMEAN servers. These results may be helpful for further investigations for determining crystal structure of the hypotheitical protein and developing target molecules to inhibit Enterobacter aerogenes.

Keywords: Anticodon; tRNA; Cytidine acetyltransferase

Introduction

Enterobacter aerogenes, part of the *Enterobacteriaceae* family, is a rod-shaped bacterium that causes infections, and is usually acquired in a hospital or hospital-type atmospheres [1]. It usually causes opportunistic infections; i.e., it only causes a disease in a immune-compromised individual. *E. aerogenes* forms part of the endogenous human Gastrointestinal (GI) microflora. It also resides in water, soil and in dairy products. It is Gram-negative and anaerobic. Additionally, they account for nearly 50% of septicemia cases and more than 70% of urinary and intestinal tract infections. The severity of these infections thus create an importance to target, isolate, identify and test for susceptibility for the causes of these nosocomial infections [2].

Most cellular RNAs require post-transcriptional chemical modifications to be mature and functional [3]. Modified bases at the first (wobble) position of the tRNA anticodons are involved in maintaining the fidelity of genetic information transfer by stabilizing specific codon–anticodon interactions to prevent misreading of non-cognate codons, and to ensure the precise reading frame [4-6]. During translation tRNAs carrying the same anticodon loop can recognize same codons. This can lead error in protein synthesis. Therefore, modification in the anticodon loop is necessary for stabilizing the tRNA and to improve error in protein synthesis. Many of the enzymes are responsible for the modification of tRNA and one of them is tRNA^{met} cytidine acetyltransferase (TmcA) (Figure 1).

tRNA^{Met} cytidine acetyltransferase

tRNA^{Met} cytidine acetyltransferase which is also known as TmcA catalyzes the formation of N(4)-acetylcytidine (ac(4)C) at the wobble position of tRNA(Met), by using acetyl-CoA as an acetyl donor and ATP (or GTP). It discriminates between tRNA^{Met} and tRNA^{Ile} because they both recognize same codon on mRNA. It is found in cytoplasm of many organisms (Figure 2).

[elongator tRNA^{Met}]-cytidine^{34}±ATP+acetyl-CoA+H_2O → [elongator tRNA^{Met}]-N^4-acetylcytidine^{34}±ADP+phosphate+coenzyme A

Protein structure prediction

Protein structure prediction means to generate three-dimensional models from amino acid sequences using computer algorithms. In many cases the predicted three-dimensional protein models are highly useful for experimentalists guiding the design of new experiments for further investigations of protein functions (Figure 3). This section describes important conceptions of protein structure and current algorithms of protein structure prediction and analysis. There are four levels of protein structure.

The prediction of the three-dimensional structure of a protein indicates the prediction of a protein's tertiary structure from its amino acid sequence. Protein structure prediction technique widely used is Homology modeling also called comparative modelling (Figure 4). The modeling procedure can be divided into a number of steps as follows.

First, suitable template(s) (sequence identity ≥ 30%) related to the target sequence are selected from the Protein Data Bank (PDB) [7]. Second, an alignment of the target sequence to the template(s) is generated through Basic Local Alignment Search Tool (BLAST). Third, coordinates of the three-dimensional model are built based on the alignment and template structures. Last, the previous steps are repeated according to the model evaluation to refine the model until acceptable results are obtained.

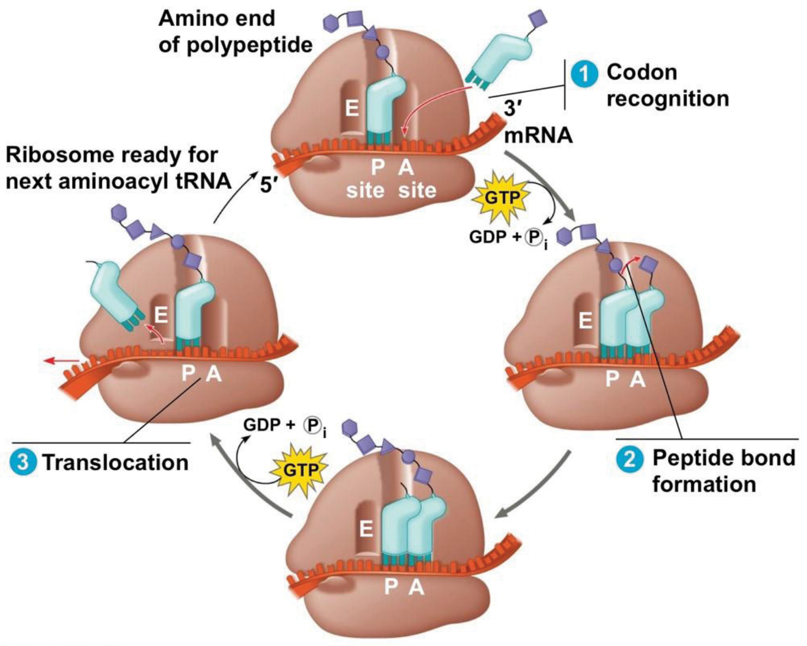
Protein structure validation

In Ramachandran plots, psi and phi angles of amino acid residues are plotted against each other, and can be constructed to evaluate a model built by homology modeling. The Ramachandran Plot helps with determination of secondary structures of proteins.

***Corresponding author:** Ritika Gupta, Department of Biotechnology, Meerut Institute of Engineering and Technology (MIET), Meerut, India, Tel: +91 9703896342; E-mail: guptaritika.27@rediffmail.com

Received May 12, 2017; **Accepted** May 31, 2017; **Published** June 02, 2017

Citation: Gupta R, Dey A, Vijan A, Gartia B (2017) *In Silico* Structure Modeling and Characterization of Hypothetical Protein YP_004590319.1 Present in *Enterobacter aerogens*. J Proteomics Bioinform 10: 152-170. doi: 10.4172/jpb.1000436



© 2014 Pearson Education, Inc.

Figure 1: Mechanism of translation.

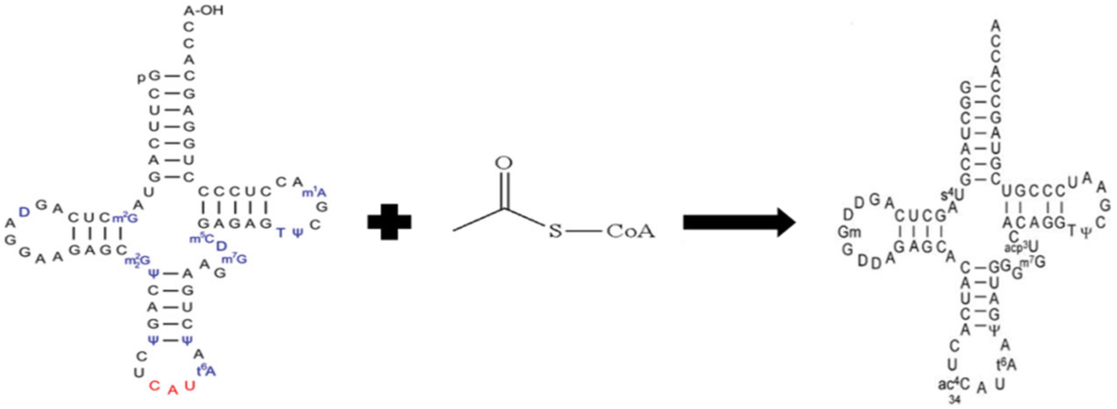


Figure 2: TmcA catalysing the formation of N(4)-acetylcytidine (ac4C) at the wobble position of tRNA(Met), by using acetyl-CoA as an acetyl donor and either ATP or GTP.

- Quadrant I shows a region where some conformations are allowed. This is where rare left-handed alpha helices lie.
- Quadrant II shows the biggest region in the graph and has the most favorable conformations of atoms. It shows the sterically allowed conformations for beta strands.
- Quadrant III shows the next biggest region in the graph. This is where right-handed alpha helices lie.
- Quadrant IV has almost no outlined region. This conformation (ψ around -180 to 0 degrees, φ around 0-180 degrees) is disfavoured due to steric clash [8].

Homology modeling

It is used when we have only the sequence and want to know the 3D structure of a target protein that has not been solved by X-ray crystallography or NMR. We will take an already determined 3D structure whose sequence will be similar to our target sequence (50% or better sequence identity), then software can be used to arrange the backbone of target sequence identically to this template. This is called "comparative modeling" or "homology modeling". In homology modeling, one or more template proteins with high sequence identity to the target sequence are identified. The target and template sequences are aligned, and a three-dimensional structure of the target protein is generated. The assembled 3D-model is then refined to bring it closer to the structure of the target protein.

A comparative modeling needs three items of input:

1. The sequence of the protein with unknown 3D structure, the "target sequence".

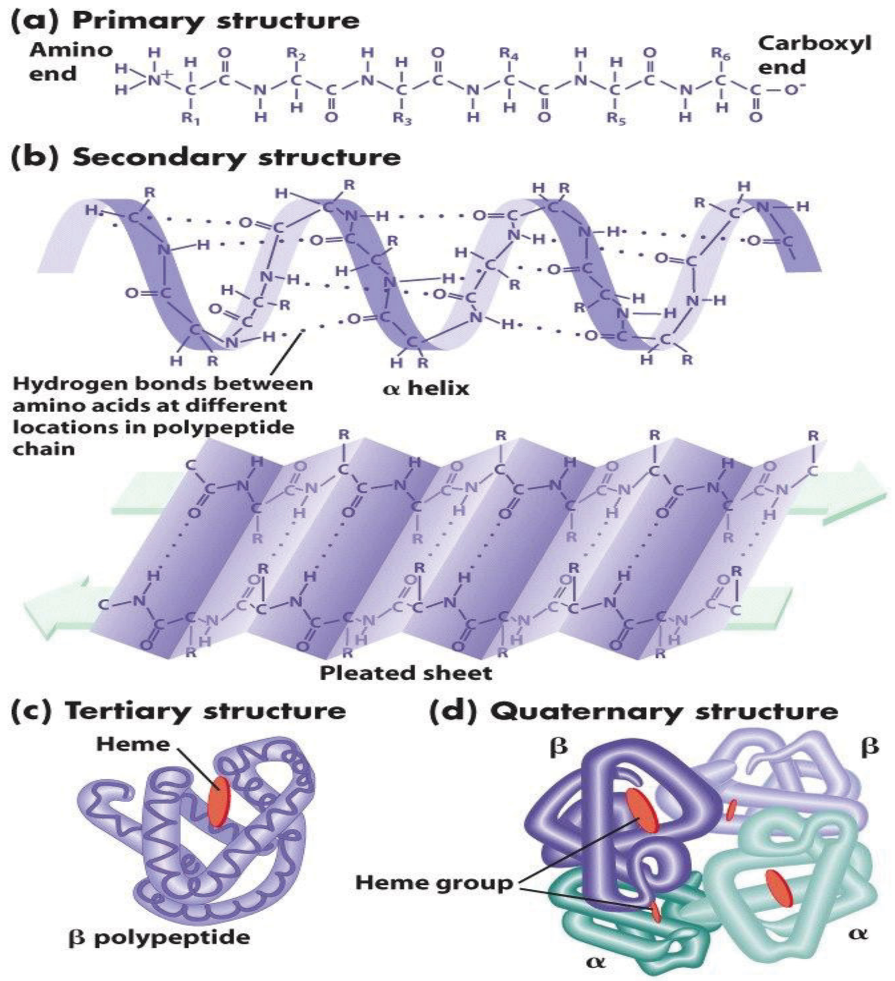


Figure 3: Four levels of protein structure.

2. On the basis of the highest sequence identity with the target sequence a 3D template is chosen. The 3D structure of the template must be determined by X-ray crystallography or NMR, and should be a published atomic coordinate "PDB" file from the Protein Data Bank.
3. An alignment between the target sequence and the template sequence.
4. Through homology modeling, backbone is arranged identically to that of the template i.e., not only the positions of alpha carbons, but also the phi and psi angles and secondary structure, are made identical to the template. It also adjusts side chain positions to minimize collisions, and may offer further energy minimization or molecular dynamics in an attempt to improve the model [9].

Material and Methods

We searched out the genome of our organism (*Enterobacter aerogenes*) and the required protein through National Center for Biotechnology Information (NCBI). The NCBI houses a series of databases relevant to biotechnology and biomedicine. Major databases include Protein Database for protein sequences, GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. Other databases include the NCBI molecular structure database. All these databases are available online through the Entrez search engine.

BLAST for identification of structural template

Basic Local Alignment Search Tool (BLAST) is an algorithm for comparing amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables us to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

Homology modeling can be done by: 1) Swiss-model; and 2) Modeller.

SWISS-model

Swiss-model is a structural bioinformatics web server dedicated to homology modeling of protein 3D structures [10,11]. It consists of three components:

1. The SWISS-model pipeline.
2. The SWISS-model Workspace.
3. The SWISS-model Repository.

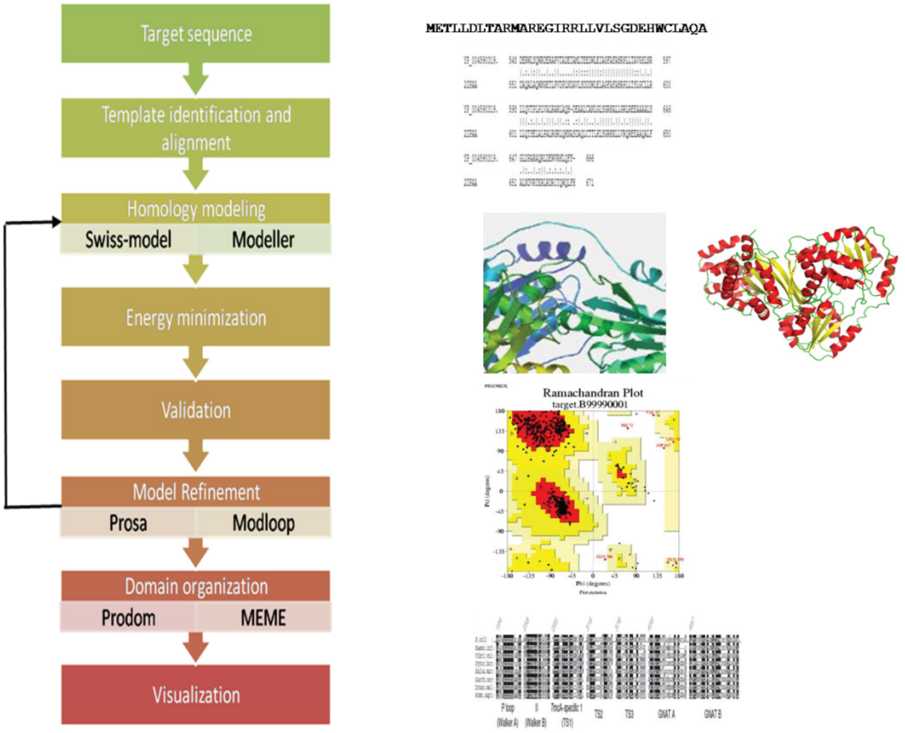


Figure 4: A simplified illustration of the modeling process.

SWISS-model pipeline comprises the four main steps that are involved in building a homology model of a given protein structure:

1. Identification of structural template. Through BLAST homologous sequences are fetched. The templates are already stored in the SWISS-MODEL Template Library (SMTL) derived from PDB.
2. Alignment of target sequence and template structure.
3. Model building and energy minimization.
4. Assessment of the model's quality using three independent methods (QMEAN, ANOLEA, GROMOS).

The SWISS-MODEL Workspace [11] is a web-based integrated service dedicated to protein structure homology modeling. It is used to build protein homology models at different levels of complexity.

To build a model at least one experimentally determined 3D structure (template) is required that shows significant amino acid sequence similarity with the target sequence. Building a homology model comprises four main steps:

1. Identification of structural template.
2. Alignment of target sequence and template structure.
3. Model building.
4. Model quality evaluation.

Model building: There are three types of modeling requests which are computed by the SWISS-MODEL server homology modeling pipeline [11]. One of them is:

Automated model: The "automated mode" is used in the case where the target-template similarity is sufficiently high to allow for fully automated modeling i.e., automated sequence alignments are sufficiently reliable when target and template share more than 50% percent of sequence identity.

The submission is of amino acid sequence or the UniProt accession code of the target protein as input data. The pipeline will automatically select suitable templates based on a Blast [12], E-value limit (which can be adjusted upon submission), experimental quality, bound substrate molecules, or different conformational states of the template.

Model assessment

Anolea: Atomic Non-Local Environment Assessment is used for the evaluation of the packing quality of the models. Energy calculations on a protein chain are performed by this program, and it evaluates the "Non- Local Environment" (NLE) of each heavy atom in the molecule. The energy for each amino acid of the protein chain is represented by the y-axis of the plot. Favourable and unfavourable energy environment for a given amino acid are represented by the negative and positive energy values respectively.

Gromos: GROMOS [13] is a molecular dynamics computer simulation package for the study of biomolecular systems applied to analyse conformations obtained by experiment or by computer simulation.

The energy for each amino acid of the protein chain is represented on the y axis. Favourable and unfavourable energy environment for a given amino acid are represented by the negative and positive energy values respectively.

QMEAN (Qualitative Model Energy Analysis)

A most important step in protein structure prediction is to estimate the quality of protein structure. There can be a set of alternative models or a single model from which the absolute quality is to be selected. QMEAN [14] estimates the quality of a protein structure. It is a composite scoring function which is able to derive both global (i.e., for the entire structure) and local (i.e., per residue) error estimates on the basis of one single model.

For each model the following data and plots are:

QMEAN score: It is the global score of the whole model and reflects the predicted model reliability. It ranges from 0 to 1.

Estimated absolute model quality: Z-score is calculated by relating the QMEAN score of the query model to the scores of a non-redundant set of high-resolution X-rays structures of similar size.

Residue error: Color gradient is uses to visualize the residue error, potentially unreliable regions, (estimated error above 3.5 Å) and the more reliable regions are shown with the red color and the blue color respectively.

Residue error plot: It models energy profile with estimated residue errors along the sequence. The QMEAN Z-score [15] provides an estimate of the absolute quality of a model by relating it to reference structures solved by X-ray crystallography. The QMEAN Z-score is an estimate of the "degree of nativeness" of the structural features observed in a model by describing the likelihood that a model is of comparable quality to high-resolution experimental structures. The three plots available for download visualize the quality of a given model with respect to these reference structures. The reference structures are a non-redundant subset of the PDB sharing less than 30% pairwise sequence identity among each other and are solved at a resolution below than 2 Å.

QMEAN returns pseudo energy of the whole model which can be used in order to compare and rank alternative models of the same target. The lower the predicted energy, the better is the model.

Modeller

It is a computer program which is used to produce homology models of protein tertiary structures [16]. It uses a technique which is related to nuclear magnetic resonance known as satisfaction of spatial restraints, by which a set of geometrical criteria are used to create a probability density function for the location of each atom in the protein. The input is the sequence alignment between the target amino acid sequence to be modeled and a template protein whose structure has been solved.

Steps to run modeller

1. It takes an input as alignment file, which is created by aligning the sequence of template and target to be modeled (target) (Figure 5) and a python file (Figure 6) and then it gives a 3D model for the target sequence containing all main chain and side chain non-hydrogen atoms as an output.
2. It gives a log file (model.log), so that errors in alignment and python file can be checked.

From the alignment of target sequence with template 3D structures many distance and dihedral angle restraints on the target sequence are calculated and from statistical analysis of the relationships between many pairs of homologous structures, these restraints were obtained. This analysis is based on a database of 105 family alignments that included 416 proteins with known 3D structure [17]. Correlations between two equivalents C_ – C_ distances, or between equivalent main chains dihedral angles from two related proteins were obtained by scanning the database. These relationships were defined as conditional Probability Density Functions (pdf's) and they can be used directly as spatial restraints. Finally, the model is obtained.

Visualization

PyMOL (www.pymol.org) is an open-source, user-sponsored, molecular visualization system for producing high quality 3D images of small molecules and biological macromolecules such as proteins. PyMOL is good for:

1. Viewing 3D molecular structures.
2. Rendering figures artistically.
3. Animating molecules dynamically.
4. Giving live 3D presentations.
5. Sharing interactive visualization.

Energy minimization

Swiss-PDB Viewer [18] is a powerful molecular modeling program used for structural alignments, homology modeling, mutating molecular models, energy minimization, and many other modeling tasks. This program can be used for various purposes such as to display, generate, analyse and also to manipulate modeling project files for the SWISS-MODEL workspace.

Validation

SAVES is a server for analyzing protein structures for validity and assessing how correct they use the following 5 programs: PROCHECK, WHAT_CHECK, ERRAT, VERIFY_3D, and PROVE.

Procheck

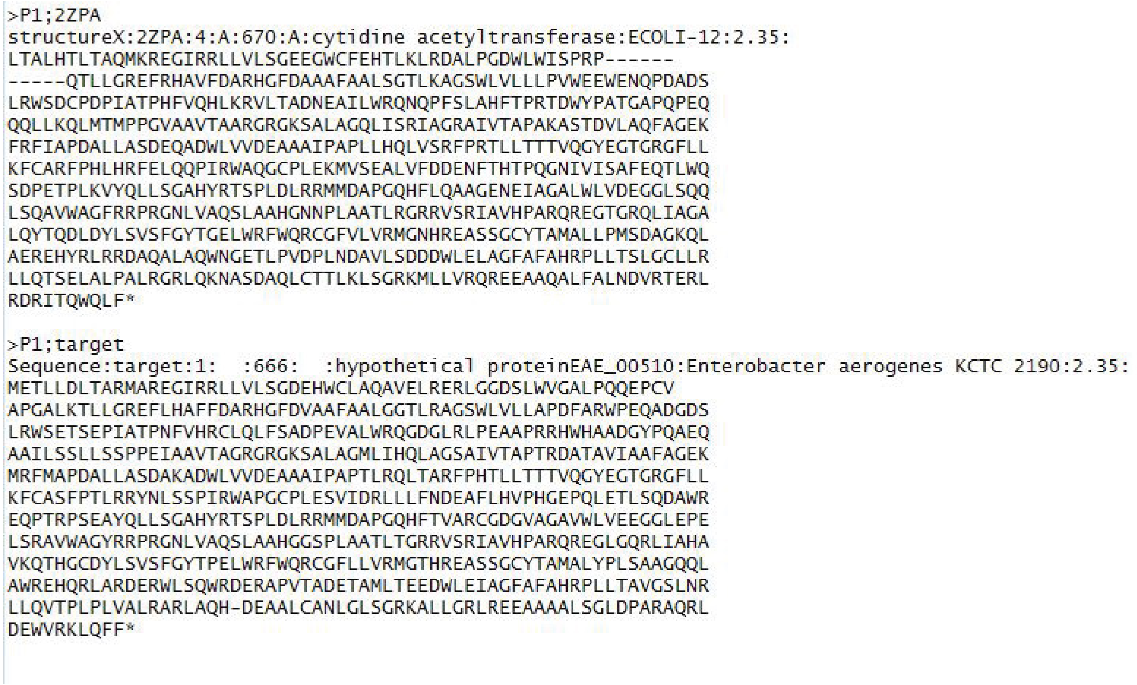
Through PROCHECK [19] the stereochemical quality of a protein structure can be known. The aim of this program is to know that how normal, or how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures.

Errat

It is a method based on characteristic atomic interaction, use for differentiating between correctly and incorrectly determined regions of protein structures. There are different types of atoms distributed non-randomly with respect to each other in proteins because of energetic and geometric effects. There will be more randomized distributions of the different atom types if there are errors in model building, which can be distinguished from correct distributions by statistical methods. This method identifies regions of error in protein crystal structures by examining the statistics of pairwise atomic interactions. This method provides a useful tool for model-building and structure verification [20].

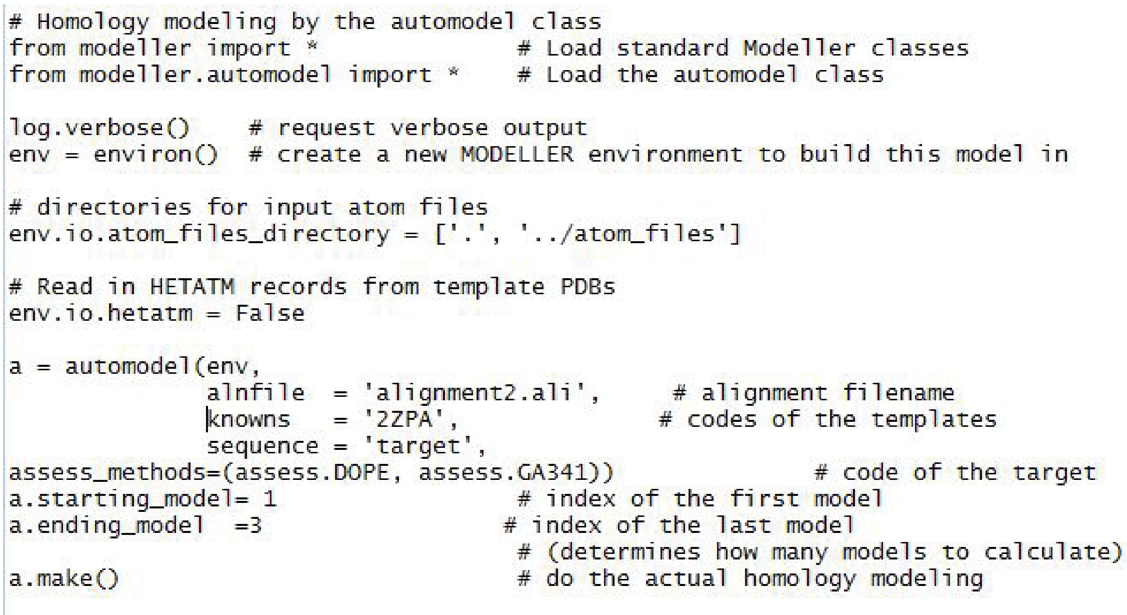
Verify 3D

To test the correctness of a 3D protein model it is necessary to check the compatibility of the model to its own amino acid sequence which is measured by a 3D profile. To check the compatibility of the sequence with the model 3D profile score S is calculated which is the sum of 3D-1D scores of overall residue positions for the amino acid sequence of the protein. A graph is plotted for the compatibility of segments of the sequence with their 3D structures, having sequence number on the x-axis and the average 3D-1D score in a window of 21 windows on the



```
>P1;2ZPA
structureX:2ZPA:4:A:670:A:cytidine acetyltransferase:ECOLI-12:2.35:
LTALHTLTAQMKREGIRRLLVLSGEEGWCFEHTLKLRDALPGDWLWISPRP------
-----QTLLGREFRHAVFDARHGFDAAAFAALSGTLKAGSWLVLLLPVWEEWENQPDADS
LRWSDCPDPIATPHFVQHLKRVLTADNEAILWRQNQPFSLAHFTPRTDWYPATGAPQPEQ
QQLLKQLMTMPPGVAAVTAARGRGKSALAGQLISRIAGRAIVTAPAKASTDVLAQFAGEK
FRFIAPDALLASDEQADWLVVDEAAAIPAPLLHQLVSRFPRTLLTTTVQGYEGTGRGFLL
KFCARFPHLHRFELQQPIRWAQGCPLEKMVSEALVFDDENFTHTPQGNIVISAFEQTLWQ
SDPETPLKVYQLLSGAHYRTSPLDLRRMMDAPGQHFLQAAGENEIAGALWLVDEGGLSQQ
LSQAVWAGFRRPRGNLVAQSLAAHGNNPLAATLRGRRVSRIAVHPARQREGTGRQLIAGA
LQYTQDLDYLSVSFGYTGELWRFWQRCGFVLVRMGNHREASSGCYTAMALLPMSDAGKQL
AEREHYRLRRDAQALAQWNGETLPVDPLNDAVLSDDDWLELAGFAFAHRPLLTSLGCLLR
LLQTSELALPALRGRLQKNASDAQLCTTLKLSGRKMLLVRQREEAAQALFALNDVRTERL
RDRITQWQLF*

>P1;target
Sequence:target:1:  :666:  :hypothetical proteinEAE_00510:Enterobacter aerogenes KCTC 2190:2.35:
METLLDLTARMAREGIRRLLVLSGDEHWCLAQAVELRERLGGDSLWVGALPQQEPCV
APGALKTLLGREFLHAFFDARHGFDVAAFAALGGTLRAGSWLVLLAPDFARWPEQADGDS
LRWSETSEPIATPNFVHRCLQLFSADPEVALWRQGDGLRLPEAAPRRHWHAADGYPQAEQ
AAILSSLLSSPPEIAAVTAGRGRGKSALAGMLIHQLAGSAIVTAPTRDATAVIAAFAGEK
MRFMAPDALLASDAKADWLVVDEAAAIPAPTLRQLTARFPHTLLTTTVQGYEGTGRGFLL
KFCASFPTLRRYNLSSPIRWAPGCPLESVIDRLLLFNDEAFLHVPHGEPQLETLSQDAWR
EQPTRPSEAYQLLSGAHYRTSPLDLRRMMDAPGQHFTVARCGDGVAGAVWLVEEGGLEPE
LSRAVWAGYRRPRGNLVAQSLAAHGGSPLAATLTGRRVSRIAVHPARQREGLGQRLIAHA
VKQTHGCDYLSVSFGYTPELWRFWQRCGFLLVRMGTHREASSGCYTAMALYPLSAAGQQL
AWREHQRLARDERWLSQWRDERAPVTADETAMLTEEDWLEIAGFAFAHRPLLTAVGSLNR
LLQVTPLPLVALRARLAQH-DEAALCANLGLSGRKALLGRLREEAAAALSGLDPARAQRL
DEWVRKLQFF*
```

Figure 5: Alignment file (alignment.aln) of target and template.



```
# Homology modeling by the automodel class
from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = ['.', '../atom_files']

# Read in HETATM records from template PDBs
env.io.hetatm = False

a = automodel(env,
              alnfile  = 'alignment2.ali',     # alignment filename
              knowns   = '2ZPA',              # codes of the templates
              sequence = 'target',
assess_methods=(assess.DOPE, assess.GA341))            # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  =3                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modeling
```

Figure 6: The python file (model.py) of target and template.

y-axis. The 3D profile score S for amino acid sequence of the model is high for correct 3D protein models [21].

Model refinement

ProSA: The ProSA program (Protein Structure Analysis) tool has a large user base and is used in the refinement and validation of experimental protein structures, structure prediction and modeling. This program exploits the advantages of interactive web-based applications for the display of scores and energy plots that highlight potential problems spotted in protein structures.

The ProSA-web service returns results instantaneously, i.e. the response time is in the order of seconds, even for large molecules [22].

ModLoop: ModLoop is a web server for automated modeling of loops in protein structures. The user has to provide the atomic coordinates of the protein structure in the Protein Data Bank format, and also the specification of the starting and ending residues of one or more segments to be modeled, containing not more than 20 residues in total. The result is the coordinates of the non-hydrogen atoms in the modeled segments. A user provides the input to the server via a simple web interface, and receives the output by e-mail. The server relies on the loop modeling routine in MODELLER that predicts the loop conformations by satisfaction of spatial restraints, without relying on a database of known protein structures [23].

Domain organisation

MEME suite: The MEME suite is a collection of tools for the discovery and analysis of sequence motifs. It is hosted at http://meme.nbcr.net/meme/. We used one of the collection tools that were MEME.

MEME: Multiple Extraction-Maximization for Motif Elicitation is a tool for discovering motifs in a group of related DNA or protein sequences. It takes a group of unaligned DNA or protein sequences as input and in output it gives as many motifs as requested by the user statistical confidence threshold [24].

ProDom: The main ProDom form consists of two parts. The first part is ProDom Browsing which allows querying of ProDom in a variety of ways: (i) By accession number (Display a ProDom entry); (ii) By the display of all proteins belonging to one or several ProDom families with logical AND/OR operators (All proteins in ProDom families); (iii) By related databases (InterPro, PROSITE, PFAM or PDB); (iv) by SWISS-PROT/TrEMBL identifier or accession number; and (v) by keyword search with AND/OR operators. The output is either cartoons displayed showing the domain arrangements of all proteins matching the query or information on a given domain family (Table 1).

The second part of the main ProDom form allows for BLAST searches in ProDom i.e., sequence is compared with the sequences within ProDom. The output is a possible domain arrangement for any query protein. When 3D structures are available for target domains, the output is directly linked to SWISS-MODEL [10,11] server for homology-based domain modeling.

Active sites: Active sites can be known by Protein Data Bank in Europe (PDBe). It aims to develop tools, services and resources that help make the wealth of data about biomacromolecular structure and function more easily accessible to the wider biomedical community [25]. It presents a slide show of images that convey important information and value-added annotation about a selected PDB entry or entries. The legend of each image contains more details as well as links to relevant web pages at PDBe or external resources.

Protein Chain A: Acetyl CoA forms hydrogen bond with ILE461, VAL463, ARG469, GLY471, and GLY473; electrostatic bond with ARG506 and van-der-waal bond with SER459, ARG460, GLN468, GLU470, ARG474, SER493, PHE494, GLU499, LEU500, ARG502, and PHE503 (Figure 7). Protein Chain B: Acetyl CoA forms hydrogen bond with ILE461, VAL463, ARG469, GLY473 and ARG474; electrostatic bond with PHE503 and van-der-waal bond with LYS301, VAL458, SER459, VAL492, SER493, PHE494, GLY471, THR472, LYS500, TRP504 and ARG506 (Figure 8).

Scoring Function Term	Raw Score	Z-score
C_beta interaction energy	-243.02	0.17
All atom pairwise energy	-18834.90	0.51
Solvation energy	-55.62	-0.21
Torsion angle energy	-104.03	-1.89
QMEAN4 Score	0.650	-1.83

Table 1: The pseudo-energies of the contributing terms together with their Z-scores.

Results

Target protein sequence

A hypothetical protein accession ID of the protein is *YP_004590319.1* was downloaded from official website of National Center for Biotechnology Information. This protein was reported from *Enterobacter aerogens* KCTC 2190. The protein has 666 numbers of amino acids. The calculated molecular weight 73016 Da. The protein has been reported to perform tRNA^{Met} cytidine acetyltransferase activity (Figure 9).

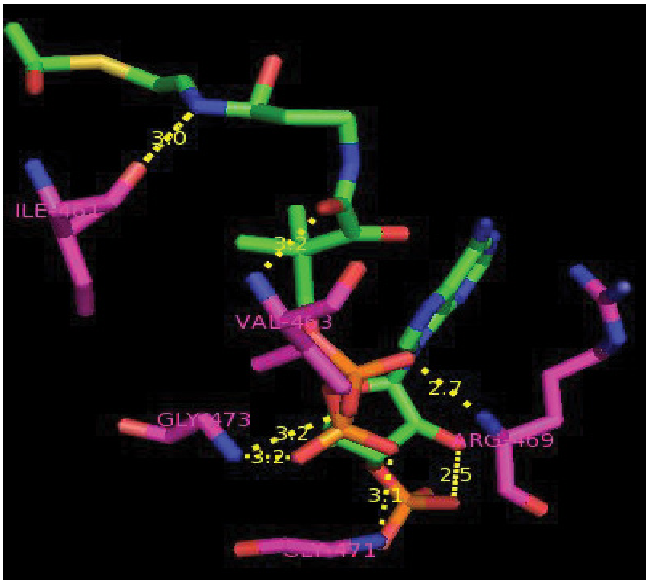


Figure 7: Hydrogen bonds between Acetyl CoA and Ile461, Val463, Arg469, Gly471, Gly473 of TmcA protein chain A.

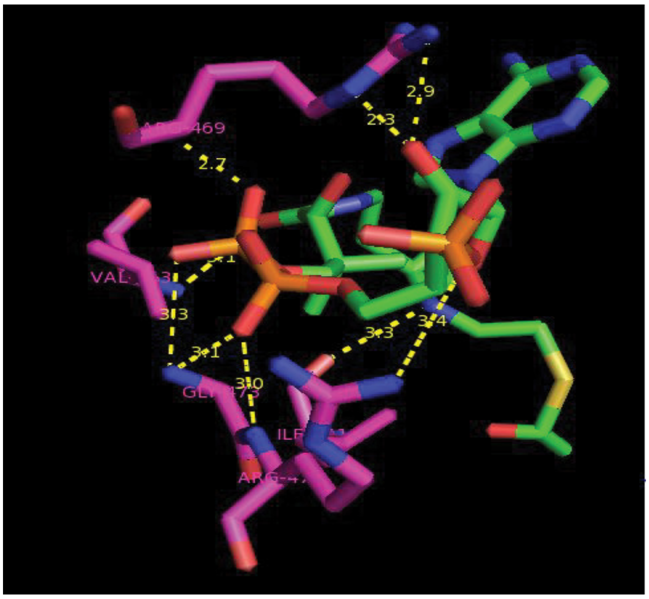
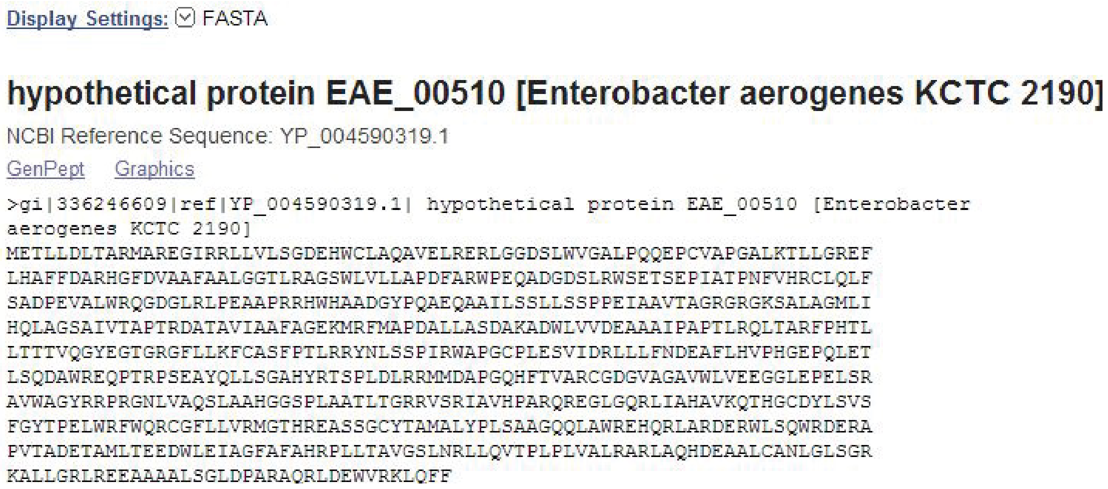


Figure 8: Hydrogen bonds between Acetyl CoA and Ile461, Val463, Arg469, Gly473 and Arg474 of TmcA protein chain B.

Protein-protein BLAST (blastp)

By giving our protein query, we got the most similar protein sequences from the protein database which we specified while performing BLAST. Here, blast against PDB database is specified. Therefore, it will find proteins similar to our target protein stored within Protein Data Bank database.

The result is shown in Figure 10. The sequence with query coverage of 100% and E value of 0.0 was of tRNA^{Met} cytidine acetyltransferase.



Display Settings: FASTA

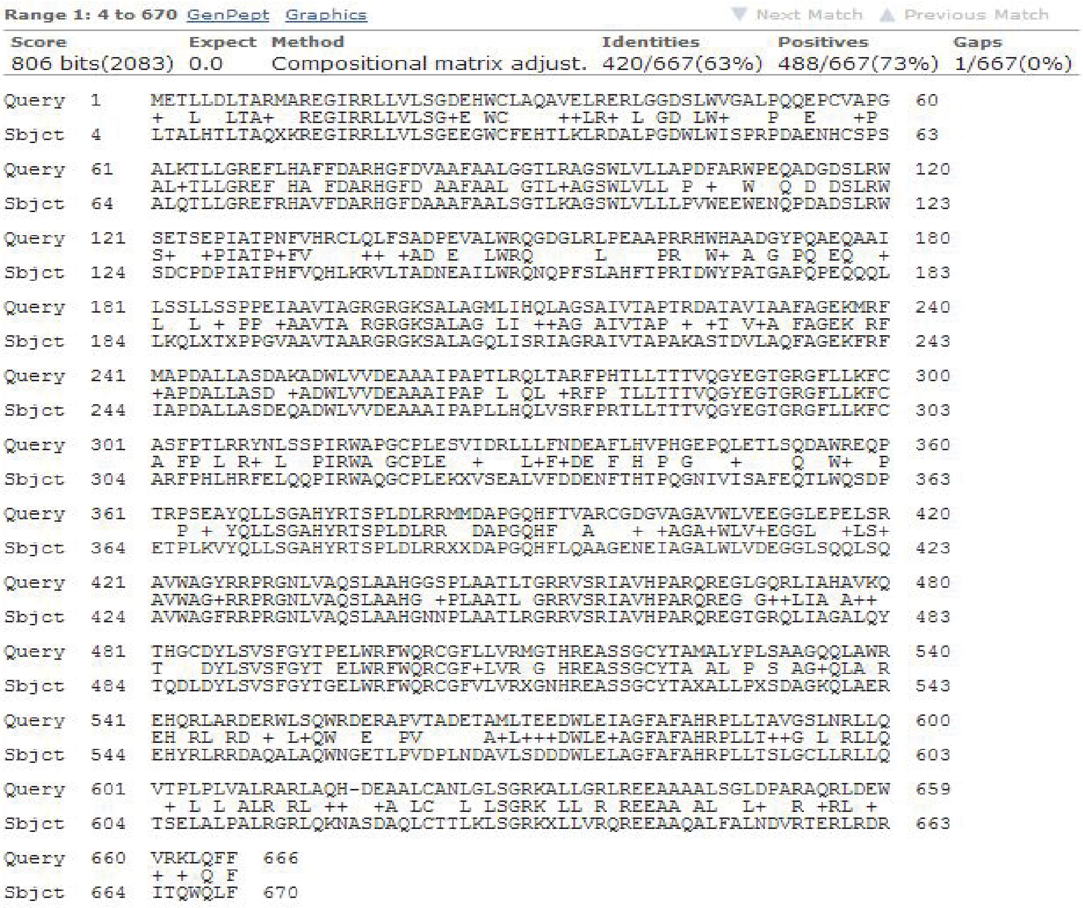
hypothetical protein EAE_00510 [Enterobacter aerogenes KCTC 2190]

NCBI Reference Sequence: YP_004590319.1

GenPept   Graphics

```
>gi|336246609|ref|YP_004590319.1| hypothetical protein EAE_00510 [Enterobacter
aerogenes KCTC 2190]
METLLDLTARMAREGIRRLLVLSGDEHWCLAQAVELRERLGGDSLWVGALPQQEPCVAPGALKTLLGREF
LHAFFDARHGFDVAAFAALGGTLRAGSWLVLLAPDFARWPEQADGDSLRWSETSEPIATPNFVHRCLQLF
SADPEVALWRQGDGLRLPEAAPRRHWHAADGYPQAEQAAILSSLLSSPPEIAAVTAGRGRGKSALAGMLI
HQLAGSAIVTAPTRDATAVIAAFAGEKMRFMAPDALLASDAKADWLVVDEAAAIPAPTLRQLTARFPHTL
LTTTVQGYEGTGRGFLLKFCASFPTLRRYNLSSPIRWAPGCPLESVIDRLLLFNDEAFLHVPHGEPQLET
LSQDAWREQPTRPSEAYQLLSGAHYRTSPLDLRRMMDAPGQHFTVARCGDGVAGAVWLVEEGGLEPELSR
AVWAGYRRPRGNLVAQSLAAHGGSPLAATLTGRRVSRIAVHPARQREGLGQRLIAHAVKQTHGCDYLSVS
FGYTPELWRFWQRCGFLLVRMGTHREASSGCYTAMALYPLSAAGQQLAWREHQRLARDERWLSQWRDERA
PVTADETAMLTEEDWLEIAGFAFAHRPLLTAVGSLNRLLQVTPLPLVALRARLAQHDEAALCANLGLSGR
KALLGRLREEAAAALSGLDPARAQRLDEWVRKLQFF
```

Figure 9: Sequence of target.



```
Score           Expect  Method                        Identities     Positives      Gaps
806 bits(2083)  0.0     Compositional matrix adjust.  420/667(63%)   488/667(73%)   1/667(0%)

Query  1    METLLDLTARMAREGIRRLLVLSGDEHWCLAQAVELRERLGGDSLWVGALPQQEPCVAPG  60
            +  L  LTA+  REGIRRLLVLSG+E WC    ++LR+ L GD LW+   P  E   +P
Sbjct  4    LTALHTLTAQXKREGIRRLLVLSGEEGWCFEHTLKLRDALPGDWLWISPRPDAENHCSPS  63

Query  61   ALKTLLGREFLHAFFDARHGFDVAAFAALGGTLRAGSWLVLLAPDFARWPEQADGDSLRW  120
            AL+TLLGREF HA FDARHGFD AAFAAL GTL+AGSWLVLL P +  W  Q D DSLRW
Sbjct  64   ALQTLLGREFRHAVFDARHGFDAAAFAALSGTLKAGSWLVLLLPVWEEWENQPDADSLRW  123

Query  121  SETSEPIATPNFVHRCLQLFSADPEVALWRQGDGLRLPEAAPRRHWHAADGYPQAEQAAI  180
            S+  +PIATP+FV    ++ +AD E  LWRQ     L    PR  W+ A G PQ EQ  +
Sbjct  124  SDCPDPIATPHFVQHLKRVLTADNEAILWRQNQPFSLAHFTPRTDWYPATGAPQPEQQQL  183

Query  181  LSSLLSSPPEIAAVTAGRGRGKSALAGMLIHQLAGSAIVTAPTRDATAVIAAFAGEKMRF  240
            L  L + PP +AAVTA RGRGKSALAG LI ++AG AIVTAP + +T V+A FAGEK RF
Sbjct  184  LKQLXTXPPGVAAVTAARGRGKSALAGQLISRIAGRAIVTAPAKASTDVLAQFAGEKFRF  243

Query  241  MAPDALLASDAKADWLVVDEAAAIPAPTLRQLTARFPHTLLTTTVQGYEGTGRGFLLKFC  300
            +APDALLASD +ADWLVVDEAAAIPAP L QL +RFP TLLTTTVQGYEGTGRGFLLKFC
Sbjct  244  IAPDALLASDEQADWLVVDEAAAIPAPLLHQLVSRFPRTLLTTTVQGYEGTGRGFLLKFC  303

Query  301  ASFPTLRRYNLSSPIRWAPGCPLESVIDRLLLFNDEAFLHVPHGEPQLETLSQDAWREQP  360
            A FP L R+ L  PIRWA GCPLE  +   L+F+DE F H P G   +    Q  W+  P
Sbjct  304  ARFPHLHRFELQQPIRWAQGCPLEKXVSEALVFDDENFTHTPQGNIVISAFEQTLWQSDP  363

Query  361  TRPSEAYQLLSGAHYRTSPLDLRRMMDAPGQHFTVARCGDGVAGAVWLVEEGGLEPELSR  420
              P + YQLLSGAHYRTSPLDLRR  DAPGQHF  A   + +AGA+WLV+EGGL  +LS+
Sbjct  364  ETPLKVYQLLSGAHYRTSPLDLRRXXDAPGQHFLQAAGENEIAGALWLVDEGGLSQQLSQ  423

Query  421  AVWAGYRRPRGNLVAQSLAAHGGSPLAATLTGRRVSRIAVHPARQREGLGQRLIAHAVKQ  480
            AVWAG+RRPRGNLVAQSLAAHG +PLAATL GRRVSRIAVHPARQREG G++LIA A++
Sbjct  424  AVWAGFRRPRGNLVAQSLAAHGNNPLAATLRGRRVSRIAVHPARQREGTGRQLIAGALQY  483

Query  481  THGCDYLSVSFGYTPELWRFWQRCGFLLVRMGTHREASSGCYTAMALYPLSAAGQQLAWR  540
            T   DYLSVSFGYT ELWRFWQRCGF+LVR G HREASSGCYTA AL P S AG+QLA R
Sbjct  484  TQDLDYLSVSFGYTGELWRFWQRCGFVLVRXGNHREASSGCYTAXALLPXSDAGKQLAER  543

Query  541  EHQRLARDERWLSQWRDERAPVTADETAMLTEEDWLEIAGFAFAHRPLLTAVGSLNRLLQ  600
            EH RL RD + L+QW  E  PV     A+L+++DWLE+AGFAFAHRPLLT++G L RLLQ
Sbjct  544  EHYRLRRDAQALAQWNGETLPVDPLNDAVLSDDDWLELAGFAFAHRPLLTSLGCLLRLLQ  603

Query  601  VTPLPLVALRARLAQH-DEAALCANLGLSGRKALLGRLREEAAAALSGLDPARAQRLDEW  659
             + L L ALR RL ++  +A LC  L LSGRK LL R REEAA AL  L+  R +RL +
Sbjct  604  TSELALPALRGRLQKNASDAQLCTTLKLSGRKXLLVRQREEAAQALFALNDVRTERLRDR  663

Query  660  VRKLQFF  666
            + + Q F
Sbjct  664  ITQWQLF  670
```

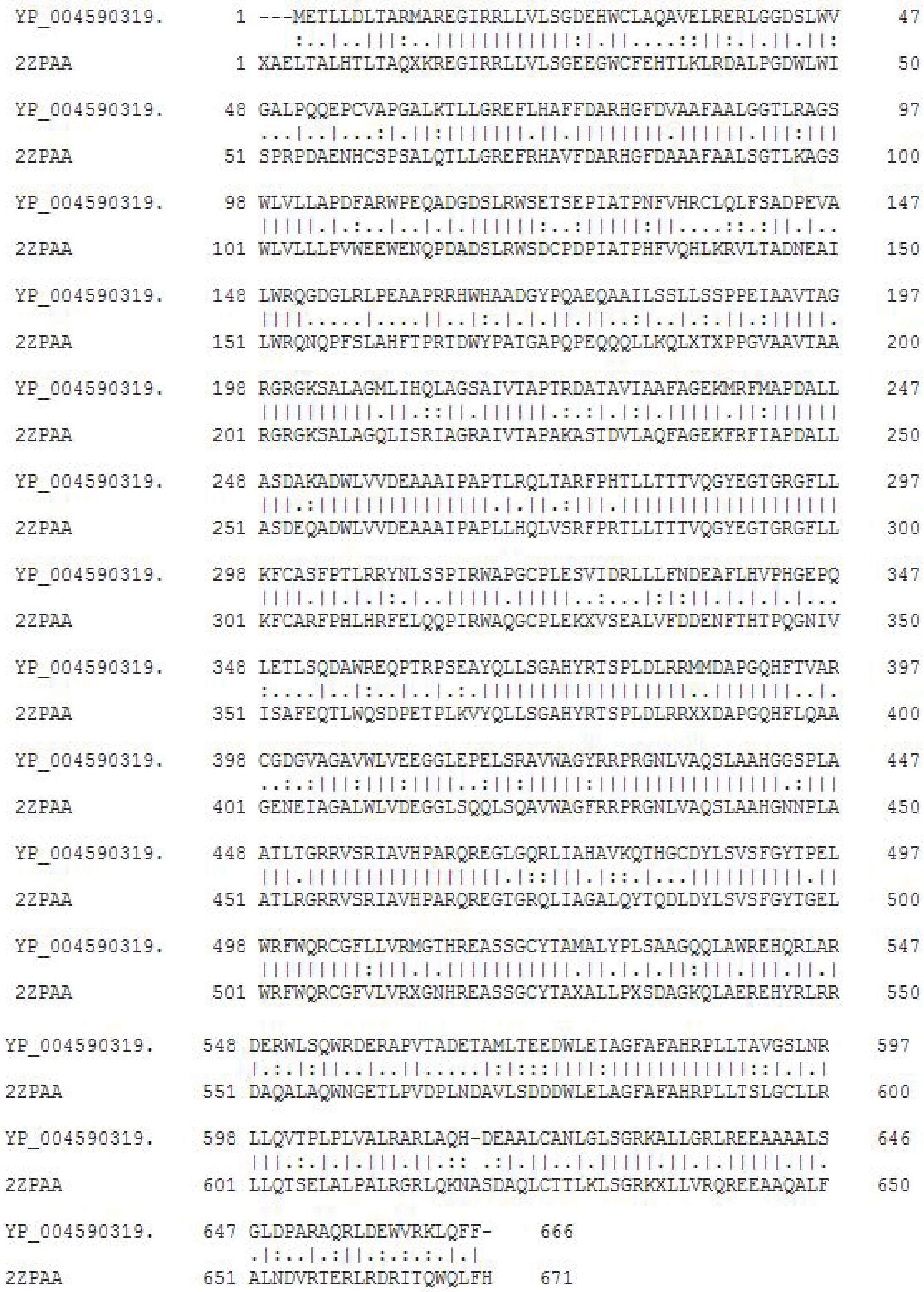
Figure 10: BLAST result of target sequence.

The PDB ID was 2ZPA. The similarity was 73%, identity was 68% and only one gap was there.

Structure functional analysis of the protein YP_004590319.1

To determine the possible function of *E. aerogens* YP_004590319.1, the sequence was subjected to comparative protein structure modeling using the target protein sequence as query for different servers described in materials and methods. For aligning both sequences we have used EBLOSUM62 matrix with a gap penalty of 10.0 and extended penalty of 0.5. We found that 420 residues out of 671 are identical showing 62.6% identity and 488 residues out of 671 are similar showing 72.7% similarity. Only 5 gaps were found and overall score was 2139.0 (Figure 11).

As per from the result from Figure 11 it is clear that target and template have similar sequence with least gaps so, we chose 2ZPA as our template.



```
YP_004590319.      1 ---METLLDLTARMAREGIRRLLVLSGDEHWCLAQAVELRERLGGDSLWV     47
                        :..|..|||:..||||||||||||:|.||....::||:.|.||.||:
2ZPAA              1 XAELTALHTLTAQXKREGIRRLLVLSGEEGWCFEHTLKLRDALPGDWLWI     50

YP_004590319.     48 GALPQQEPCVAPGALKTLLGREFLHAFFDARHGFDVAAFAALGGTLRAGS     97
                     ...|..|...:|.||:|||||||.||.||||||||.||||||.|||:|||
2ZPAA             51 SPRPDAENHCSPSALQTLLGREFRHAVFDARHGFDAAAFAALSGTLKAGS    100

YP_004590319.     98 WLVLLAPDFARWPEQADGDSLRWSETSEPIATPNFVHRCLQLFSADPEVA    147
                     |||||.|.:..|..|.|.||||||:..:|||||:||....::.:||.|..
2ZPAA            101 WLVLLLPVWEEWENQPDADSLRWSDCPDPIATPHFVQHLKRVLTADNEAI    150

YP_004590319.    148 LWRQGDGLRLPEAAPRRHWHAADGYPQAEQAAILSSLLSSPPEIAAVTAG    197
                     ||||.....|....||..|:.|.|.||.||..:|..|.:.||.:|||||.
2ZPAA            151 LWRQNQPFSLAHFTPRTDWYPATGAPQPEQQQLLKQLXTXPPGVAAVTAA    200

YP_004590319.    198 RGRGKSALAGMLIHQLAGSAIVTAPTRDATAVIAAFAGEKMRFMAPDALL    247
                     ||||||||||.||.::||.||||||.:.:|.|:|.|||||.||:||||||
2ZPAA            201 RGRGKSALAGQLISRIAGRAIVTAPAKASTDVLAQFAGEKFRFIAPDALL    250

YP_004590319.    248 ASDAKADWLVVDEAAAIPAPTLRQLTARFPHTLLTTTVQGYEGTGRGFLL    297
                     |||.:|||||||||||||||.|.||.:|||.|||||||||||||||||||
2ZPAA            251 ASDEQADWLVVDEAAAIPAPLLHQLVSRFPRTLLTTTVQGYEGTGRGFLL    300

YP_004590319.    298 KFCASFPTLRRYNLSSPIRWAPGCPLESVIDRLLLFNDEAFLHVPHGEPQ    347
                     ||||.||.|.|:.|..|||||.|||||..:...|:|:||.|.|.|.|...
2ZPAA            301 KFCARFPHLHRFELQQPIRWAQGCPLEKXVSEALVFDDENFTHIPQGNIV    350

YP_004590319.    348 LETLSQDAWREQPTRPSEAYQLLSGAHYRTSPLDLRRMMDAPGQHFTVAR    397
                     :....|..|:..|..|.:.||||||||||||||||||..|||||||..|.
2ZPAA            351 ISAFEQTLWQSDPETPLKVYQLLSGAHYRTSPLDLRRXXDAPGQHFLQAA    400

YP_004590319.    398 CGDGVAGAVWLVEEGGLEPELSRAVWAGYRRPRGNLVAQSLAAHGGSPLA    447
                     ..:.:|||:|||:||||..:||:|||||:||||||||||||||||.:|||
2ZPAA            401 GENEIAGALWLVDEGGLSQQLSQAVWAGFRRPRGNLVAQSLAAHGNNPLA    450

YP_004590319.    448 ATLTGRRVSRIAVHPARQREGLGQRLIAHAVKQTHGCDYLSVSFGYTPEL    497
                     |||.|||||||||||||||||.|::|||.|::.|...||||||||||.||
2ZPAA            451 ATLRGRRVSRIAVHPARQREGTGRQLIAGALQYTQDLDYLSVSFGYTGEL    500

YP_004590319.    498 WRFWQRCGFLLVRMGTHREASSGCYTAMALYPLSAAGQQLAWREHQRLAR    547
                     |||||||||:|||.|.|||||||||||.||.|.|.||:|||.|||.||.|
2ZPAA            501 WRFWQRCGFVLVRXGNHREASSGCYTAXALLPXSDAGKQLAEREHYRLRR    550

YP_004590319.    548 DERWLSQWRDERAPVTADETAMLTEEDWLEIAGFAFAHRPLLTAVGSLNR    597
                     |.:.|:||..|..||.....|:|:::||||:||||||||||||::|.|.|
2ZPAA            551 DAQALAQWNGETLPVDPLNDAVLSDDDWLELAGFAFAHRPLLTSLGCLLR    600

YP_004590319.    598 LLQVTPLPLVALRARLAQH-DEAALCANLGLSGRKALLGRLREEAAAALS    646
                     |||.:.|.|.|||.||.:: .:|.||..|.|||||.||.|.|||||.||.
2ZPAA            601 LLQTSELALPALRGRLQKNASDAQLCTTLKLSGRKXLLVRQREEAAQALF    650

YP_004590319.    647 GLDPARAQRLDEWVRKLQFF-    666
                     .|:..|.:||.:.:.:.|.|
2ZPAA            651 ALNDVRTERLRDRITQWQLFH    671
```

Figure 11: Pairwise alignment between target and template.

Homology modeling of YP_004590319.1 using SWISS-MODEL

We submit the sequence of our target in SWISS-MODEL workspace and requested for model. The target model was constructed and it is shown in Figure 12.

Model information

- Model residue range is 1 to 666.
- Modeling is based on the template 2ZPA.
- Sequence identity is 63.72%

Quaternary structure information

- Template is a monomer.
- Target is also a monomer.

Quality information

- QMEAN Z-score is -1.83.

Global model quality estimation

QMEAN score: It is the global score of the whole model which reflects the predicted model reliability. It ranges from 0 to 1. The QMEAN4 global score of our target is 0.65.

Estimated absolute model quality: The graph below shows the comparison of non-redundant PDB structures with our target. The areas built by the circles which are colored in different shades of grey in the plot represent the QMEAN scores of the reference structures from the PDB. It is calculated by comparing model's QMEAN score to the scores obtain for experimental structures of similar size. Z-scores are calculated for all four statistical potential terms and QMEAN Z-score of our model is -1.83 (Figure 13).

The plot below shows the density plot (based on the QMEAN score) of all reference models used in the Z-score calculation. The location of the query model w.r.t. the background distribution is marked in red. This plot basically is a "projection" of the first plot for the given protein size (Figure 14).

QMEAN4 global scores: It is a composite score consisting of a linear combination of 4 statistical potential terms (estimated model reliability between 0-1).

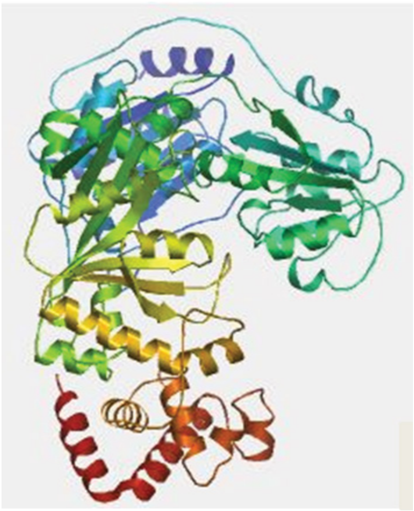


Figure 12: Target got from SWISS-MODEL.

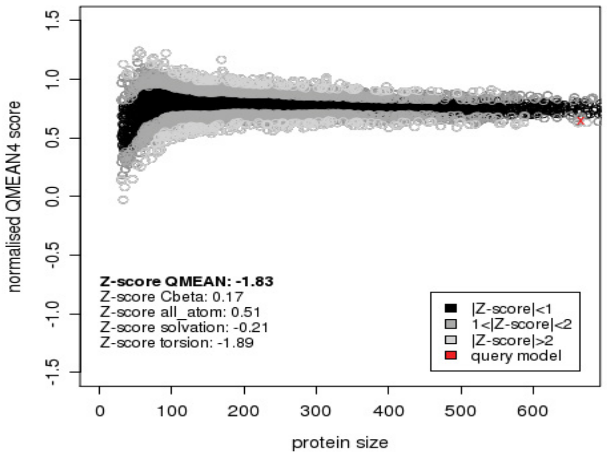


Figure 13: Comparison of our target with all PDB structures.

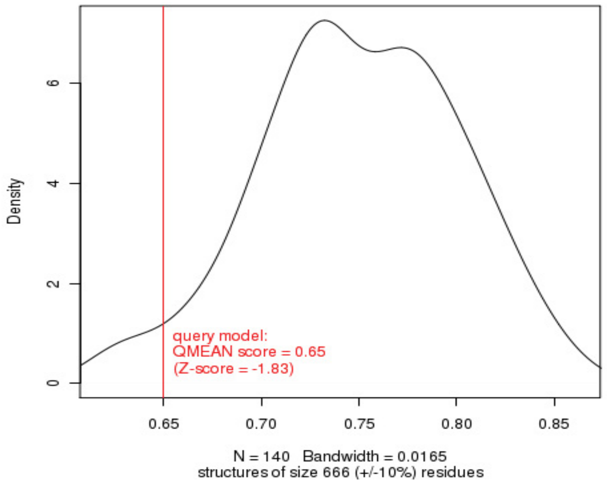


Figure 14: Density plot of all reference models. The target here is shown in red color.

Score components: By analysing the Z-scores of the individual terms, geometrical features can be identified which are responsible for large negative QMEAN Z-score. Models having strongly negative Z-scores for QMEAN are of low quality and they correspond to red regions in the color gradient. Models which slide in the light red to blue region are of good quality (Figure 15).

Local scores

Coloring by residue error: Color gradient is use to visualize per residue error. In the figure shown below, the blue color and the red color shows more reliable and potentially unreliable regions respectively (Figure 16).

Residue error plot: In the graph shown below, local model reliability with estimated per residue in accuracies along the sequence are shown (Figure 17).

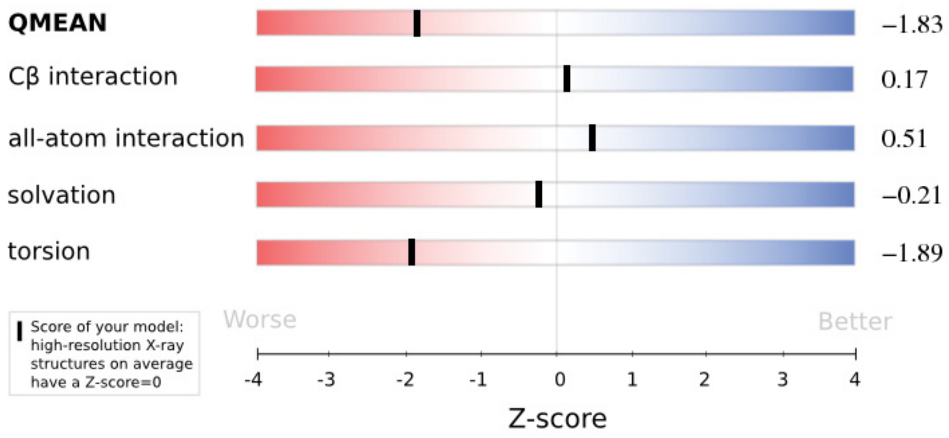


Figure 15: Z-scores of all the terms which combine to form QMEAN.

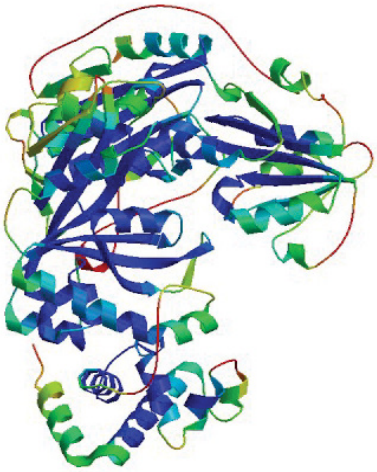


Figure 16: Structure of target from SWISS-model.

Local model quality estimation

Anolea and Gromos: In the graphs below the result of Anolea and Gromos are shown in which the green region and the red region represents the favourable and unfavourable energy environment of residues respectively (Figure 18).

Alignment: Through SWISS-MODEL we got the alignment of our target and template 2ZPA and secondary structures are also shown in the Figure 19.

Homology Modeling of YP_004590319.1 and Energy minimization: The three-dimensional structure of a hypothetical YP_004590319.1 was developed from its template TmcA of Escherichia coli (PDB ID: 2ZPA chain A, at 2.35 A ° resolution) and it was also used as template for homology modeling. The Comparative modeling of YP_004590319.1 was performed using a restrained-based approach implemented in MODELLER9.12 [16]. Target protein model was constructed (Figure 20).

The final deviations in the protein structure geometry was regularized by energy minimization with the GROMOS96 force field [13] using Deep View [18] by applying 200 steps steepest descent algorithm and 200 steps conjugate gradients algorithm. The final model was validated by using SAVES server.

Validation of Homology Model of YP_004590319.1: The quality of backbone conformation of model was assessed by PROCHECK for reliability [19]. The observed Psi-Phi pairs had, 92.5% of residues in most favored regions, 6.9% residues in additional allowed regions, 0.3% residues in generously allowed regions and 0.7% residues in disallowed regions as shown in Figure 21.

Verify 3D: Through plot we concluded that 96.55% of the residues had an averaged 3D-1D score >0.2 (Figure 22).

ERRAT: On the error axis two lines are drawn to indicate the confidence with which it is possible to reject regions that exceed the error value (Figure 23). The overall quality factor came out be 84.083. To improve the overall quality factor we used ModLoop. We will submit our target sequence and enter the residues which are exceeding the error value.

The target obtained from ModLoop was submitted in SAVES server and we got the errat result as (Figure 24).

The overall quality factor increases to 87.367. The region ranging from 160-170 is exceeding error value even after running ModLoop. We saw this region in pymol and we came to know that it is a loop region so; this is the reason that it is not getting modeled. We can resist this error as there are no active sites in this region.

ProSA: We submitted our energy minimized target in ProSA and got the following plots (Figure 25). *Z*-scores which are outside a range are characteristic for native proteins indicating erroneous structures. In order to facilitate interpretation of the *z*-score of the specified protein, its particular value is displayed in a plot that contains the *z*-scores of all experimentally determined protein chains in current PDB. This plot can be used to check whether the *z*-score of the protein in question is within the range of scores typically found for proteins of similar size belonging to one of these groups. The energy plot shows the local model quality by plotting energies as a function of amino acid sequence position *i*. In general, positive values correspond to problematic or erroneous parts of a model (Figure 26). Hence the plot is smoothed by calculating the average energy over each 40-residue fragment $s_{i,i+39}$,

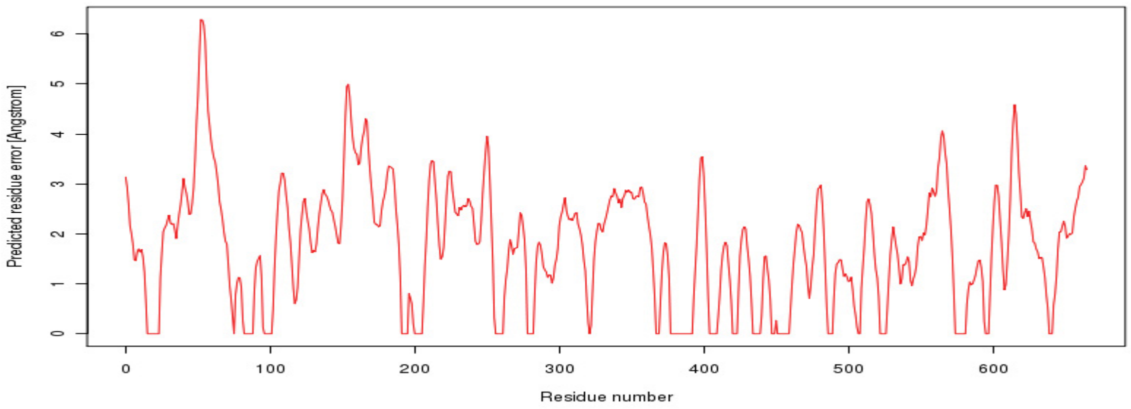
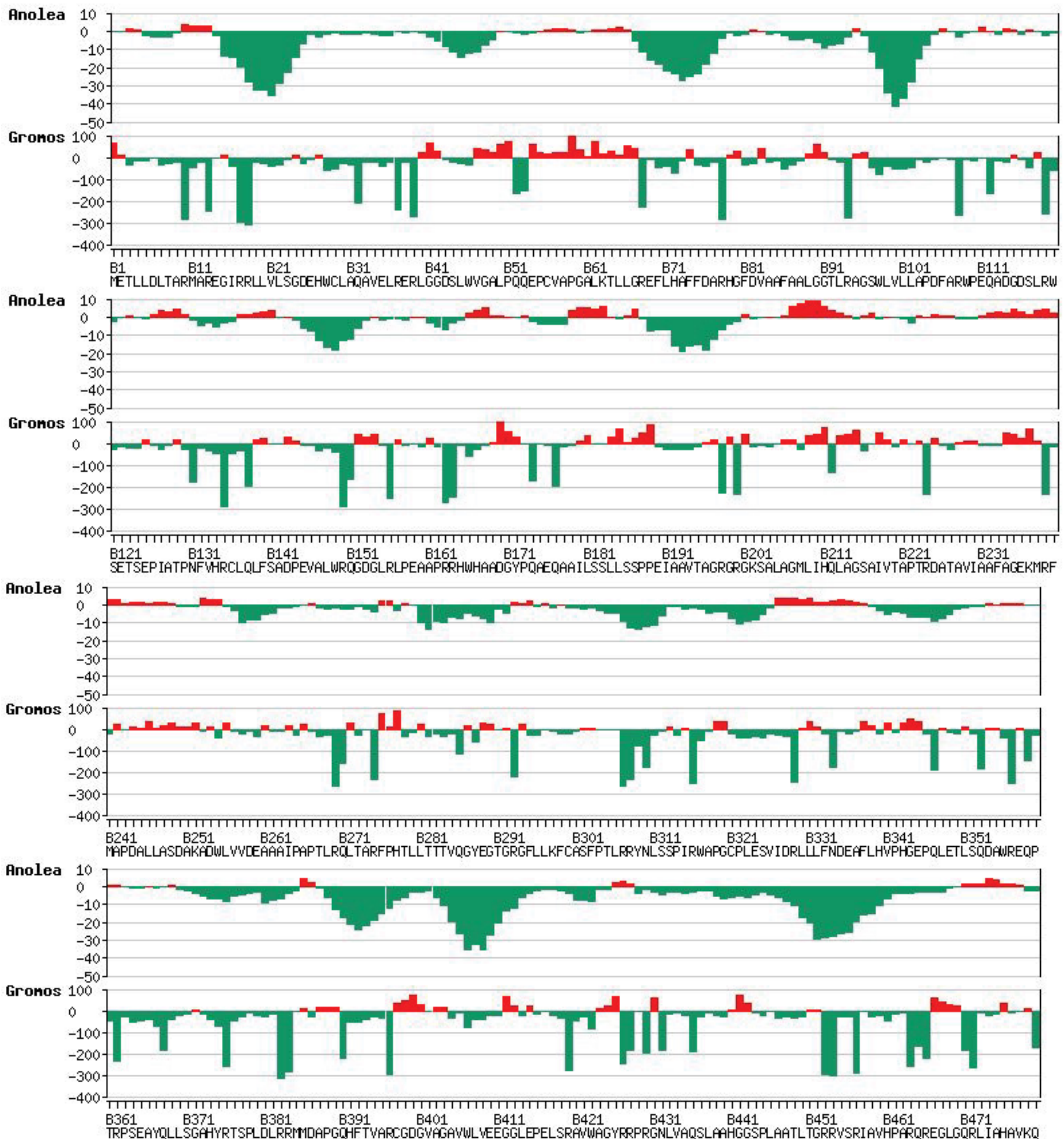


Figure 17: Predicted residue error for target structure.



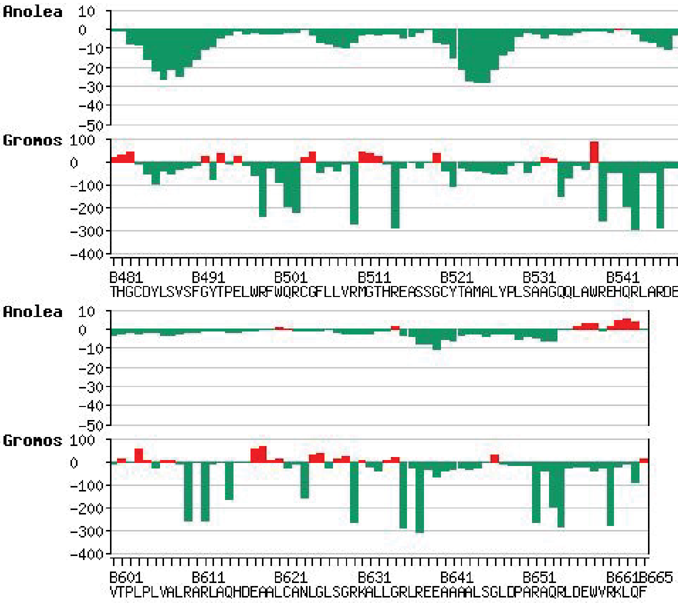
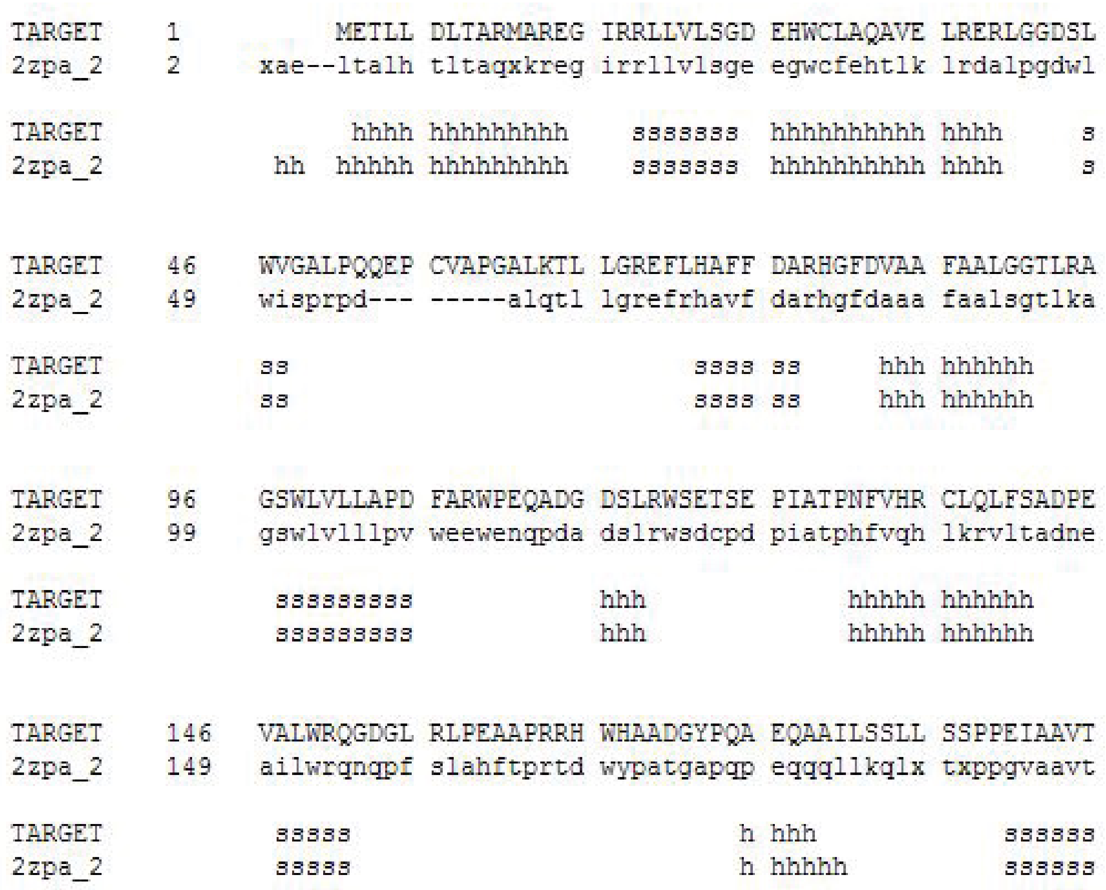
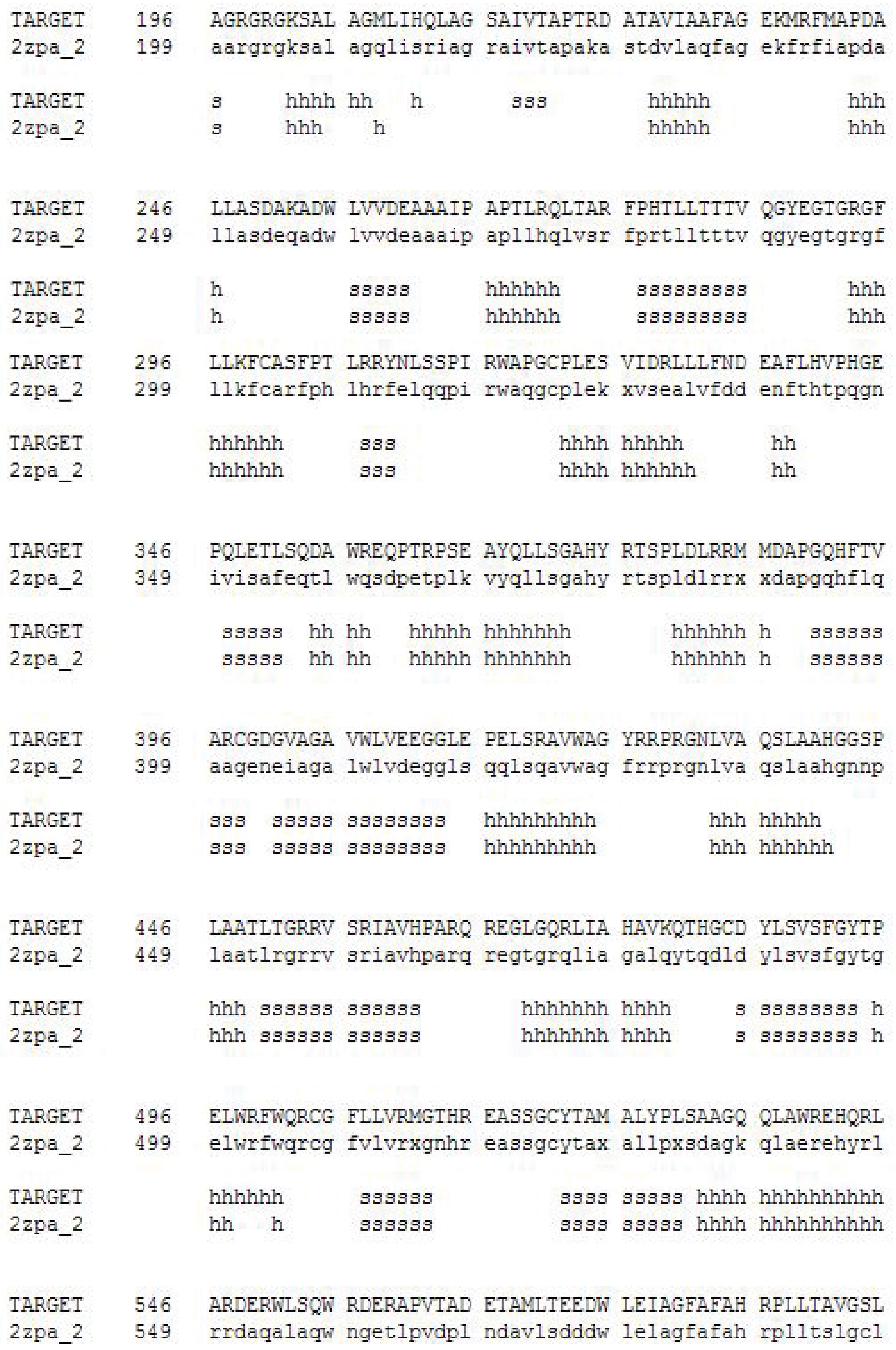


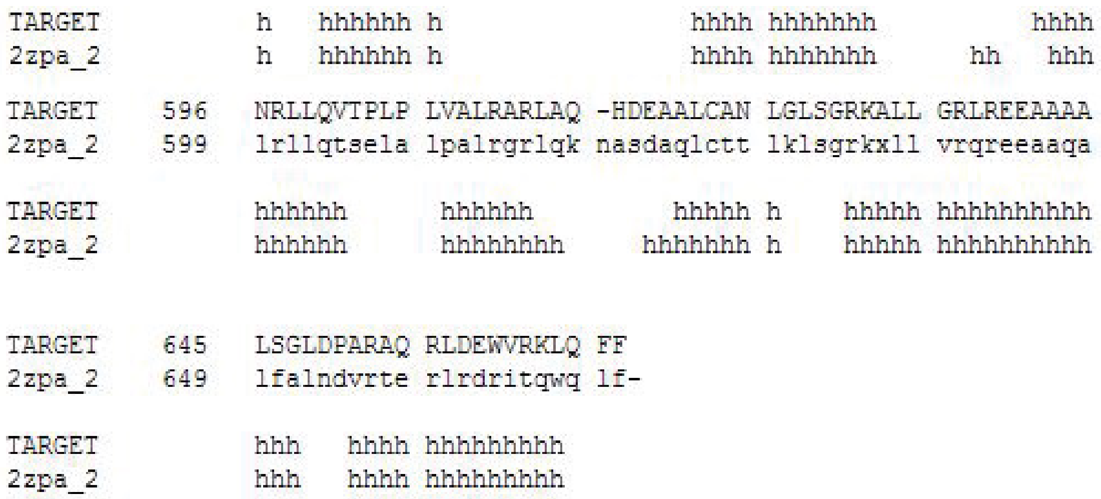
Figure 18: The graphs of Anolea and Gromos.



```
TARGET     1          METLL DLTARMAREG IRRLLVLSGD EHWCLAQAVE LRERLGGDSL
2zpa_2     2     xae--ltalh tltaqxkreg irrllvlsge egwcfehtlk lrdalpgdwl

TARGET                 hhhh hhhhhhhhh    sssssss  hhhhhhhhhh hhhh     s
2zpa_2            hh  hhhhh hhhhhhhhh    sssssss  hhhhhhhhhh hhhh     s


TARGET     46    WVGALPQQEP CVAPGALKTL LGREFLHAFF DARHGFDVAA FAALGGTLRA
2zpa_2     49    wisprpd--- -----alqtl lgrefrhavf darhgfdaaa faalsgtlka

TARGET           ss                           ssss ss      hhh hhhhhh
2zpa_2           ss                           ssss ss      hhh hhhhhh


TARGET     96    GSWLVLLAPD FARWPEQADG DSLRWSETSE PIATPNFVHR CLQLFSADPE
2zpa_2     99    gswlvlllpv weewenqpda dslrwsdcpd piatphfvqh lkrvltadne

TARGET            sssssssss             hhh              hhhhh hhhhhh
2zpa_2            sssssssss             hhh              hhhhh hhhhhh


TARGET     146   VALWRQGDGL RLPEAAPRRH WHAADGYPQA EQAAILSSLL SSPPEIAAVT
2zpa_2     149   ailwrqnqpf slahftprtd wypatgapqp eqqqllkqlx txppgvaavt

TARGET            sssss                          h hhh            ssssss
2zpa_2            sssss                          h hhhhh          ssssss
```



```
TARGET    196   AGRGRGKSAL AGMLIHQLAG SAIVTAPTRD ATAVIAAFAG EKMRFMAPDA
2zpa_2    199   aargrgksal agqlisriag raivtapaka stdvlaqfag ekfrfiapda

TARGET          s     hhhh hh   h       sss        hhhhh           hhh
2zpa_2          s     hhh    h                     hhhhh           hhh


TARGET    246   LLASDAKADW LVVDEAAAIP APTLRQLTAR FPHTLLTTTV QGYEGTGRGF
2zpa_2    249   llasdeqadw lvvdeaaaip apllhqlvsr fprtlltttv qgyegtgrgf

TARGET          h          sssss      hhhhhh      sssssssss        hhh
2zpa_2          h          sssss      hhhhhh      sssssssss        hhh

TARGET    296   LLKFCASFPT LRRYNLSSPI RWAPGCPLES VIDRLLLFND EAFLHVPHGE
2zpa_2    299   llkfcarfph lhrfelqqpi rwaqgcplek xvsealvfdd enfthtpqgn

TARGET          hhhhhh      sss             hhhh hhhhh       hh
2zpa_2          hhhhhh      sss             hhhh hhhhhh      hh


TARGET    346   PQLETLSQDA WREQPTRPSE AYQLLSGAHY RTSPLDLRRM MDAPGQHFTV
2zpa_2    349   ivisafeqtl wqsdpetplk vyqllsgahy rtspldlrrx xdapgqhflq

TARGET           sssss  hh hh   hhhhh hhhhhhh        hhhhhh h   ssssss
2zpa_2           sssss  hh hh   hhhhh hhhhhhh        hhhhhh h   ssssss


TARGET    396   ARCGDGVAGA VWLVEEGGLE PELSRAVWAG YRRPRGNLVA QSLAAHGGSP
2zpa_2    399   aageneiaga lwlvdeggls qqlsqavwag frrprgnlva qslaahgnnp

TARGET          sss  sssss ssssssss   hhhhhhhhh         hhh hhhhh
2zpa_2          sss  sssss ssssssss   hhhhhhhhh         hhh hhhhhh


TARGET    446   LAATLTGRRV SRIAVHPARQ REGLGQRLIA HAVKQTHGCD YLSVSFGYTP
2zpa_2    449   laatlrgrrv sriavhparq regtgrqlia galqytqdld ylsvsfgytg

TARGET          hhh ssssss ssssss        hhhhhhh hhhh     s ssssssss h
2zpa_2          hhh ssssss ssssss        hhhhhhh hhhh     s ssssssss h


TARGET    496   ELWRFWQRCG FLLVRMGTHR EASSGCYTAM ALYPLSAAGQ QLAWREHQRL
2zpa_2    499   elwrfwqrcg fvlvrxgnhr eassgcytax allpxsdagk qlaerehyrl

TARGET          hhhhhh      ssssss          ssss sssss hhhh hhhhhhhhhh
2zpa_2          hh   h      ssssss          ssss sssss hhhh hhhhhhhhhh


TARGET    546   ARDERWLSQW RDERAPVTAD ETAMLTEEDW LEIAGFAFAH RPLLTAVGSL
2zpa_2    549   rrdaqalaqw ngetlpvdpl ndavlsdddw lelagfafah rplltslgcl
```



```
TARGET          h   hhhhhh h                 hhhh hhhhhhh        hhhh
2zpa_2          h   hhhhhh h                 hhhh hhhhhhh     hh  hhh

TARGET     596  NRLLQVTPLP LVALRARLAQ -HDEAALCAN LGLSGRKALL GRLREEAAAA
2zpa_2     599  lrllqtsela lpalrgrlqk nasdaqlctt lklsgrkxll vrqreeaaqa

TARGET          hhhhhh      hhhhhh         hhhhh h    hhhhh hhhhhhhhhh
2zpa_2          hhhhhh      hhhhhhhh     hhhhhhh h    hhhhh hhhhhhhhhh


TARGET     645  LSGLDPARAQ RLDEWVRKLQ FF
2zpa_2     649  lfalndvrte rlrdritqwq lf-

TARGET          hhh   hhhh hhhhhhhhh
2zpa_2          hhh   hhhh hhhhhhhhh
```

Figure 19: The alignment of target and template got from SWISS-model.

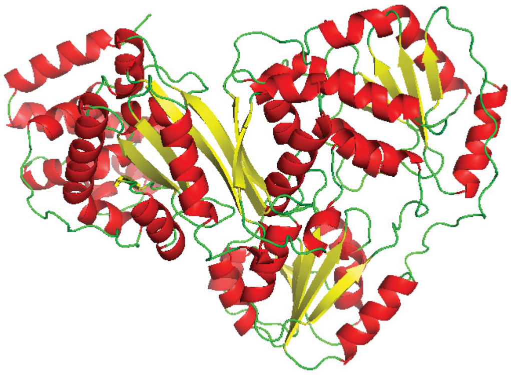


Figure 20: Target structure from modeler. The α-helices are in red color, β sheets are in yellow color and loops are in green color.

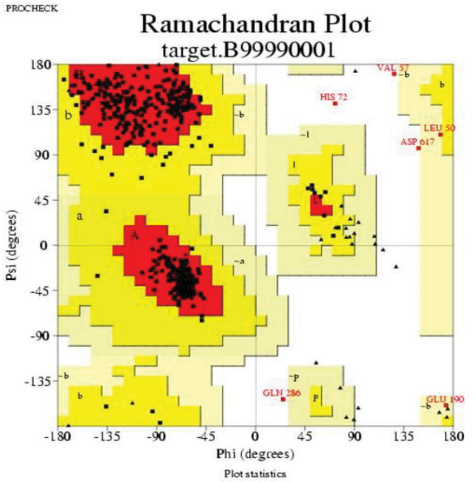
which is then assigned to the 'central' residue of the fragment at position *i* + 19.

ProSA-web visualizes the 3D structure of the protein using the molecule viewer Jmol in order to see those regions in the model that contribute to a bad overall score (Figure 27).

Domain organisation: We used ProDom and MEME to know about the domain and motifs of our protein. We submitted our target sequence and it was compared with the sequences within ProDom database. The result came out to be, mentioned below.

MEME: We submitted an unaligned sequence file in fasta format containing 5 to 6 sequences similar to our target protein (Tables 2 and 3). The result was the location of motifs in our target. Through this suite we can prove that our target contains that particular domain. The result is shown below in Figure 28.

From MEME we got to know that the region from 479-523 have acetyltransferase activity and the region from 254-295 consists of DUF699 domain (Figures 29-34).

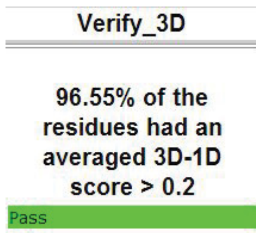


Plot statistics

Residues in most favoured regions [A,B,L]	529	92.5%
Residues in additional allowed regions [a,b,l,p]	37	6.5%
Residues in generously allowed regions [~a,~b,~l,~p]	2	0.3%
Residues in disallowed regions	4	0.7%
	----	------
Number of non-glycine and non-proline residues	572	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	53	
Number of proline residues	39	

Total number of residues	666	

Figure 21: Ramachandran plot of target.

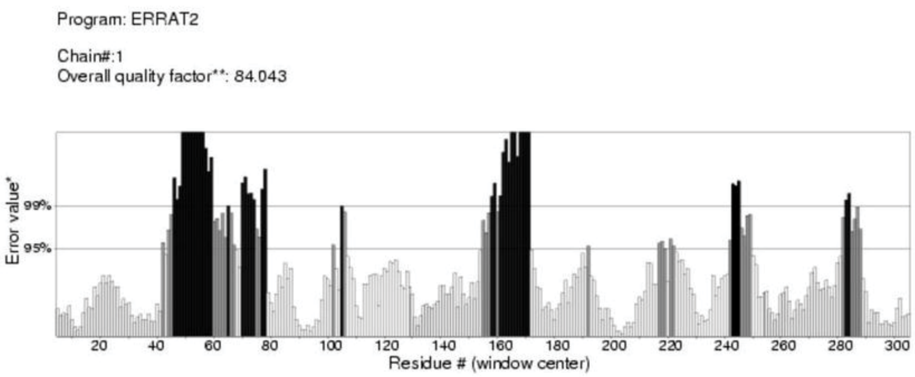


Verify_3D

96.55% of the residues had an averaged 3D-1D score > 0.2

Pass

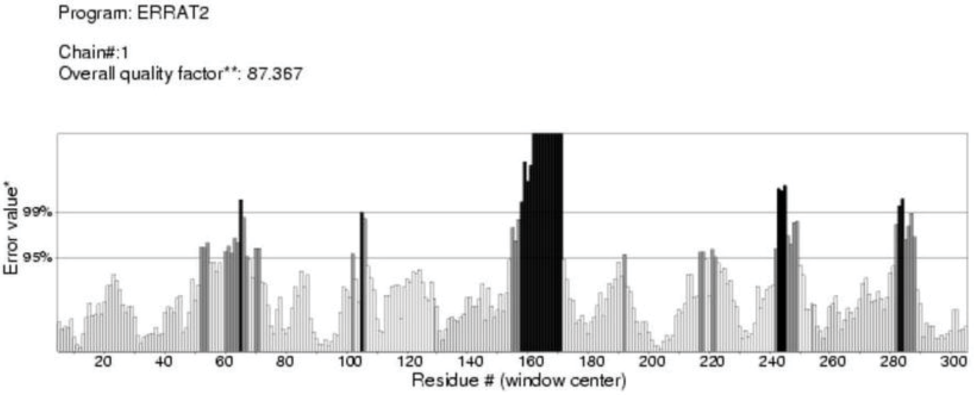
Figure 22: Verify_3D result.



Program: ERRAT2

Chain#:1
Overall quality factor**: 84.043

Figure 23: Errat result of energy minimized target**.**



Program: ERRAT2

Chain#:1
Overall quality factor**: 87.367

Figure 24: Errat result after mod loop.

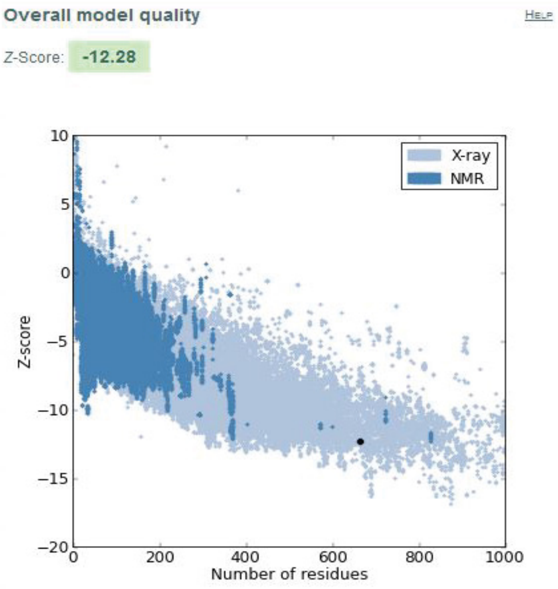


Figure 25: ProSA-web z-scores of all protein chains in PDB which are determined by X-ray crystallography (light blue) or NMR spectroscopy (dark blue) w.r.t their length. The z-score of target protein is highlighted with black dot.

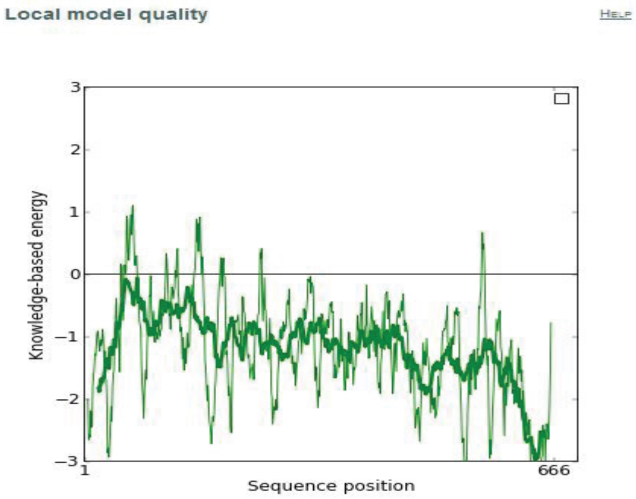


Figure 26: The energy plot of the target. Residue energies averaged over a sliding window are plotted as a function of the central residue in the window.

Conclusion

For modeling our target protein YP_004590319.1, we identified

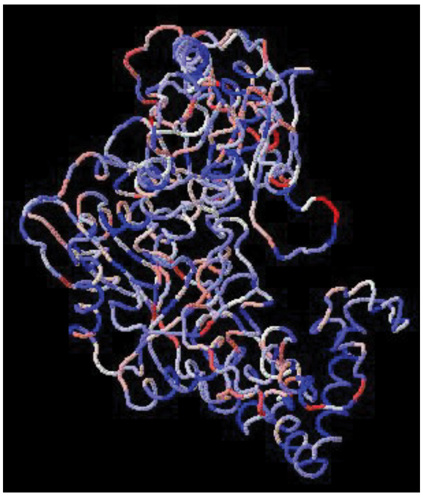


Figure 27: Residues are colored from blue to red in the order of increasing residue energy.

useful templates that were showing good similarity with our protein and it showed that our protein has tRNAmet cytidine acetyltransferase activity. Model was derived by comparative or homology modeling in SWISS-MODEL and displayed several meaningful features like the secondary structure, QMEAN score, Z-score. Model was also derived from MODELLER and then energy minimization was carried out using SWISS PDB-VIEWER (SPDBV). Model was validated through several checks such as procheck, errat, verify 3D. It was refined through ProSA (Protein Structure Analysis) and ModLoop. Domains were identified using ProDom and MEME online tools. Unexpectedly, the structure reveals an idiosyncratic RNA helicase module fused with GCN5-related N-acetyltransferase (GNAT) folds, which intimately cross-interact. The biochemical evidence from previous research articles have unravelled the function of acetyl-CoA as an enzyme-activating switch, and have proposed that an RNA helicase motor is driven by ATP hydrolysis which is then used to deliver the wobble base to the active centre of the GNAT domain. In addition, this approach can be used to guide empirical laboratory studies that would be carried out on this protein.

Position	ProDom domain	Domain name	Score	E value
1-64	#PDA1I3Q7	FULL=PREDICTED YPFI ATPASE FUSED	245	7e-20
65-112	#PD135242	FULL=PREDICTED YPFI ATPASE FUSED	189	3e-13
66-295	#PDC4L6J4	UNCHARACTERIZED FULL=PUTATIVE SUBNAME DUF699	135	5e-07
114-227	#PDA072M6	FULL=PUTATIVE UNCHARACTERIZED TRANSFERASE HYDROLASE ACETYLTRANSFERASE YPFI	327	2e-29
229-336	#PD007775	FULL=PREDICTED HYDROLASE ATPASE ACETYLTRANSFERASE YPFI	449	2e-43
367-414	#PD041580	FULL=PREDICTED ACETYLTRANSFERASE ATPASE YPFI FUSED	211	7e-16
417-487	#PD014232	FULL=PREDICTED ACETYLTRANSFERASE ATPASE YPFI FUSED	293	2e-25
489-599	#PD560856	FULL=PREDICTED YPFI ATPASE FUSED	449	2e-43

Table 2: Comparison of target sequence with sequences present in ProDom database.

Motif	Range	P-value
Motif 1	479-523	1.13e-53
Motif 2	425-474	3.16e-52
Motif 3	254-295	3.87e-47

Table 3: Location of motifs in hypothetical protein.

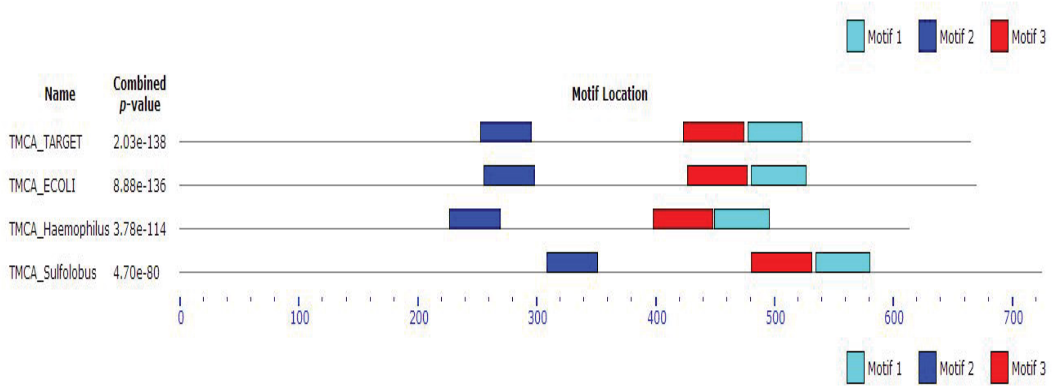
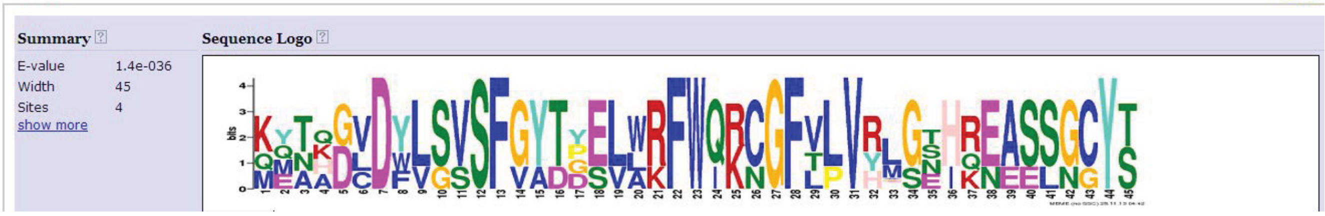


Figure 28: Comparison of location of motifs and combined p-value of TMCA from target, *Escherichia coli, Haemophilus* and *Sulfolobus.*



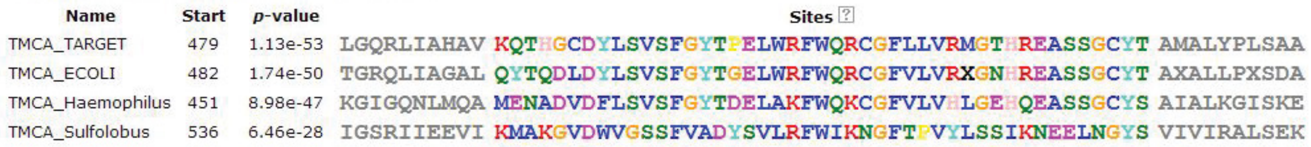
Summary ?

E-value	1.4e-036
Width	45
Sites	4

show more

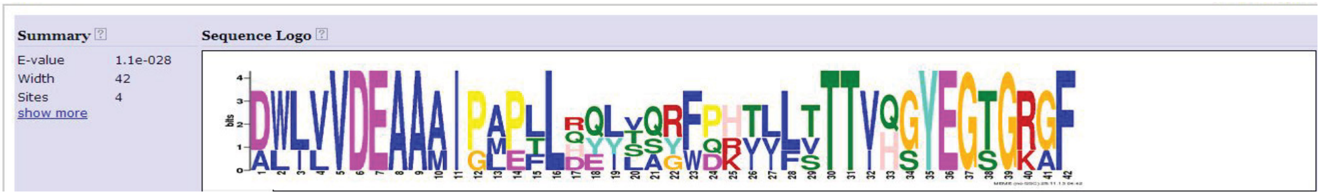
Sequence Logo ?

Figure 29: Conserved sequence of motif 1.



Name	Start	*p*-value	Sites ?		
TMCA_TARGET	479	1.13e-53	LGQRLIAHAV	KQTHGCDYLSVSFGYTPELWRFWQRCGFLLVRMGTHREASSGCYT	AMALYPLSAA
TMCA_ECOLI	482	1.74e-50	TGRQLIAGAL	QYTQDLDYLSVSFGYTGELWRFWQRCGFVLVRXGNHREASSGCYT	AXALLPXSDA
TMCA_Haemophilus	451	8.98e-47	KGIGQNLMQA	MENADVDFLSVSFGYTDELAKFWQKCGFVLVHLGEHQEASSGCYS	AIALKGISKE
TMCA_Sulfolobus	536	6.46e-28	IGSRIIEEVI	KMAKGVDWVGSSFVADYSVLRFWIKNGFTPVYLSSIKNEELNGYS	VIVIRALSEK

Figure 30: Multiple sequence alignment of motif 1.



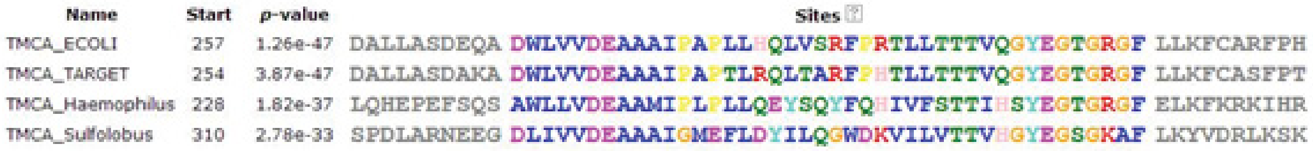
Summary ?

E-value	1.1e-028
Width	42
Sites	4

show more

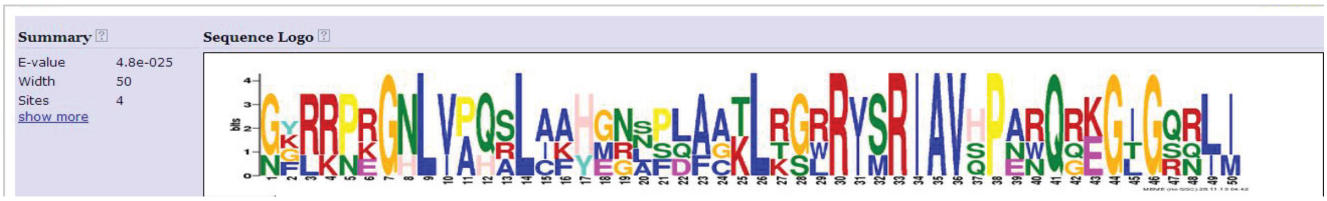
Sequence Logo ?

Figure 31: Conserved sequence of motif 2.



Name	Start	*p*-value	Sites ?		
TMCA_ECOLI	257	1.26e-47	DALLASDEQA	DWLVVDEAAAIPAPLLHQLVSRFPRTLLTTTVQGYEGTGRGF	LLKFCARFPH
TMCA_TARGET	254	3.87e-47	DALLASDAKA	DWLVVDEAAAIPAPTLRQLTARFPHTLLTTTVQGYEGTGRGF	LLKFCASFPT
TMCA_Haemophilus	228	1.82e-37	LQHEPEFSQS	AWLLVDEAAMIPLPLLQEYSQYFQHIVFSTTIHSYEGTGRGF	ELKFKRKIHR
TMCA_Sulfolobus	310	2.78e-33	SPDLARNEEG	DLIVVDEAAAIGMEFLDYILQGWDKVILVTTVHGYEGSGKAF	LKYVDRLKSK

Figure 32: Multiple sequence alignment of motif 2.



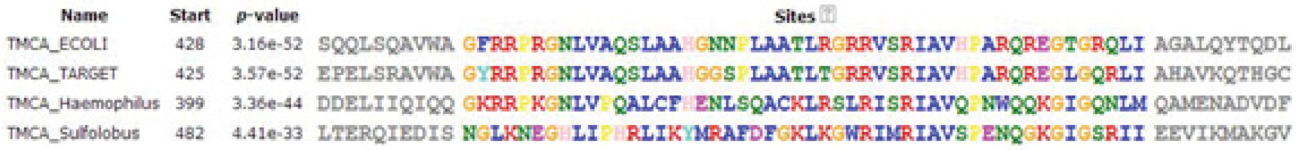
Summary ?

E-value	4.8e-025
Width	50
Sites	4

show more

Sequence Logo ?

Figure 33: Conserved sequence of motif 3.



Name	Start	*p*-value	Sites ?		
TMCA_ECOLI	428	3.16e-52	SQQLSQAVWA	GFRRPRGNLVAQSLAAHGNNPLAATLRGRRVSRIAVHPARQREGTGRQLI	AGALQYTQDL
TMCA_TARGET	425	3.57e-52	EPELSRAVWA	GYRRPRGNLVAQSLAAHGGSPLAATLTGRRVSRIAVHPARQREGLGQRLI	AHAVKQTHGC
TMCA_Haemophilus	399	3.36e-44	DDELIIQIQQ	GKRRPKGNLVPQALCFHENLSQACKLRSLRISRIAVQPNWQQKGIGQNLM	QAMENADVDF
TMCA_Sulfolobus	482	4.41e-33	LTERQIEDIS	NGLKNEGHLIPHRLIKYMRAFDFGKLKGWRIMRIAVSPENQGKGIGSRII	EEVIKMAKGV

Figure 34: Multiple sequence alignment of motif 3.

References

1. Sanders WE Jr, Sanders CC (1997) Enterobacter spp.: pathogens poised to flourish at the turn of the century. Clin Microbial Rev 10: 220-241.

2. Lederberg J, Alexander M, Bloom BR (2000) Encyclopedia of Microbiology (2nd edn.). Academic Press, San Diego, USA.

3. McCloskey JA, Crain PF (1998) The RNA modification database. Nucleic Acids Res 26: 196-197.

4. Bjork GR, Durand JM, Hagervall TG, Leipuviene R, Lundgren HK, et al. (1999) Transfer RNA modification: influence on translational frameshifting and metabolism. FEBS Lett 452: 47-51.

5. Suzuki T (2005) Biosynthesis and function of tRNA wobble modifications. In Fine-Tuning of RNA Functions by modification and editing. Springer-Verlag, NY, USA 12: 24-69.

6. Agris PF, Vendeix FA, Graham WD (2007) tRNA's wobble decoding of the genome: 40 years of modification. J Mol Biol 366: 1-13.

7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Research 28: 235-242.

8. Berg MJ, Tymoczko JL, Stryer L (2007) Chapter 2: Protein composition and structure. Biochemistry, New York, USA.

9. Martz E (2001) Comparative ("Homology") Modeling for beginners with free software.

10. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31: 3381-3385.

11. Arnold K, Bordoli L, Kopp J, Schwede T (2009) The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling. Bioinformatics 22: 195-201.

12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

13. van Gunsteren WF, Billeter SR, Hunenberger P, Riniker S, Oostenbrink S, et al. (1996) Biomolecular Simulations: The GROMOS96 Manual and User Guide. Zurich pp: 1-1042.

14. Benkert P, Tosatto SCE, Schomburg D (2008) QMEAN: A comprehensive scoring function for model quality assessment. Proteins 71: 261-277.

15. Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27: 343-350.

16. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779-815.

17. Sali A, Overington JP (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. Protein Sci 3: 1582-1596.

18. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. Electrophoresis 18: 2714-2723.

19. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26: 283-291.

20. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2: 1151-1159.

21. Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356: 83-85.

22. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35: 407-410.

23. Fiser A, Sali A (2004) ModLoop: automated modeling of loops in protein structures. Einstein College of Medicine,USA. Bioinformatics 19: 2500-2501.

24. Bailey TL (2011) DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics 27: 1653-1659.

25. Velankar S, Kleywegt GJ (2011) The Protein Data Bank in Europe (PDBe): bringing structure to biology. Acta Crystallogr 67: 324-330.