BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 11(11)

Homology modeling and assigned functional annotation of an uncharacterized antitoxin protein from *Streptomyces xinghaiensis*

Arafat Rahman Oany[1]*, Md Shahabuddin Ahmed[1], Nasreen Jahan[1], Md Abdul Latif [1], Shahin Mahmud[1], Md. Ahmed Hossain[1], Fatema Akter [2], Hasibul Haque Rakib[1], Md. Shariful Islam[1]

Received November 01, 2015; Accepted November 04, 2015; Published November 30, 2015

[1]Department of Biotechnology and Genetic Engineering, Faculty of Life Science, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh; [2]Department of Biochemistry and Molecular Biology, University of Dhaka, Bangladesh; Arafat Rahman Oany – Email: arafatr@outlook.com; Phone: +880 15 5881 9130; *Corresponding author

Abstract:
Streptomyces xinghaiensis is a Gram-positive, aerobic and non-motile bacterium. The bacterial genome is known. Therefore, it is of interest to study the uncharacterized proteins in the genome. An uncharacterized protein (gi|518540893|86 residues) in the genome was selected for a comprehensive computational sequence-structure-function analysis using available data and tools. Sub-cellular localization of the targeted protein with conserved residues and assigned secondary structures is documented. Sequence homology search against the protein data bank (PDB) and non-redundant GenBank proteins using BLASTp showed different homologous proteins with known antitoxin function. A homology model of the target protein was developed using a known template (PDB ID: 3CTO:A) with 62% sequence similarity in HHpred after assessment using programs PROCHECK and QMEAN6. The predicted active site using CASTp is analyzed for assigned anti-toxin function. This information finds specific utility in annotating the said uncharacterized protein in the bacterial genome.

Keywords: antitoxin, homology modeling, active-site residues, prediction, hypothetical protein, *Streptomyces xinghaiensis*

Background:
Streptomyces are soil-conquering gram-positive bacteria and a member of the order of Actinomycetales **[1].** *Streptomyces xinghaiensis*, a novel species of *Streptomyces*, was isolated from a marine sediment sample collected from Xinghai Bay, Dalian, China **[2]**. The *S. xinghaiensis* draft genome contains 7,618,725 bp with a GC content of 72.5%, representing approximately 92.7% of the 8.2-Mb estimated size of the genome. Analysis of the genome revealed a number of genes related to the biosynthesis of secondary metabolites. At least 15 clusters involved in secondary metabolism were identified; these include one gene cluster that highly resembles the gene cluster of ribostamycin **[3],** an amino-glycoside antibiotic. Toxin-antitoxin (TA) system was widely adopted in many genomes like bacteria and archaea and is usually recognized as a maintenance or stability mediator **[4, 5].** Although, the exact role of this system in the genome is not clear but, acts as sentinels against DNA loss and various stress management process like programmed cell death and antibiotic resistance **[6].** According to the mode of action, the TA systems have been classified into three broad classes. Namely, class I, II and Class III. Among them, class II is predominant in many organisms **[7].** The class II TA system consists of two proteins called toxin and antitoxin. The toxin is neutralized by antitoxin through direct protein-protein interaction and/or interaction with palindrome sequences within the promoters for suppressing transcription of the TA system **[8-10].**

The sequencing technology is both sophisticated and advanced in dealing with massive amount of data in recent years. Unfortunately, many of these genomes are still not fully annotated and they comprise of various genes or proteins with uncharacterized function and unknown 3D structures. This is due to several limitations, such as the cost and time necessary for experimental methodologies. Hence, an alternative method using computer aided mathematical models are frequently

used to gain insight **[11-13].** Therefore, it is of interest to study the uncharacterized proteins in the genome. An uncharacterized protein (gi|518540893|86 residues) in the bacterial genome was selected for a comprehensive computational sequence-structure-function analysis using available data and tools.

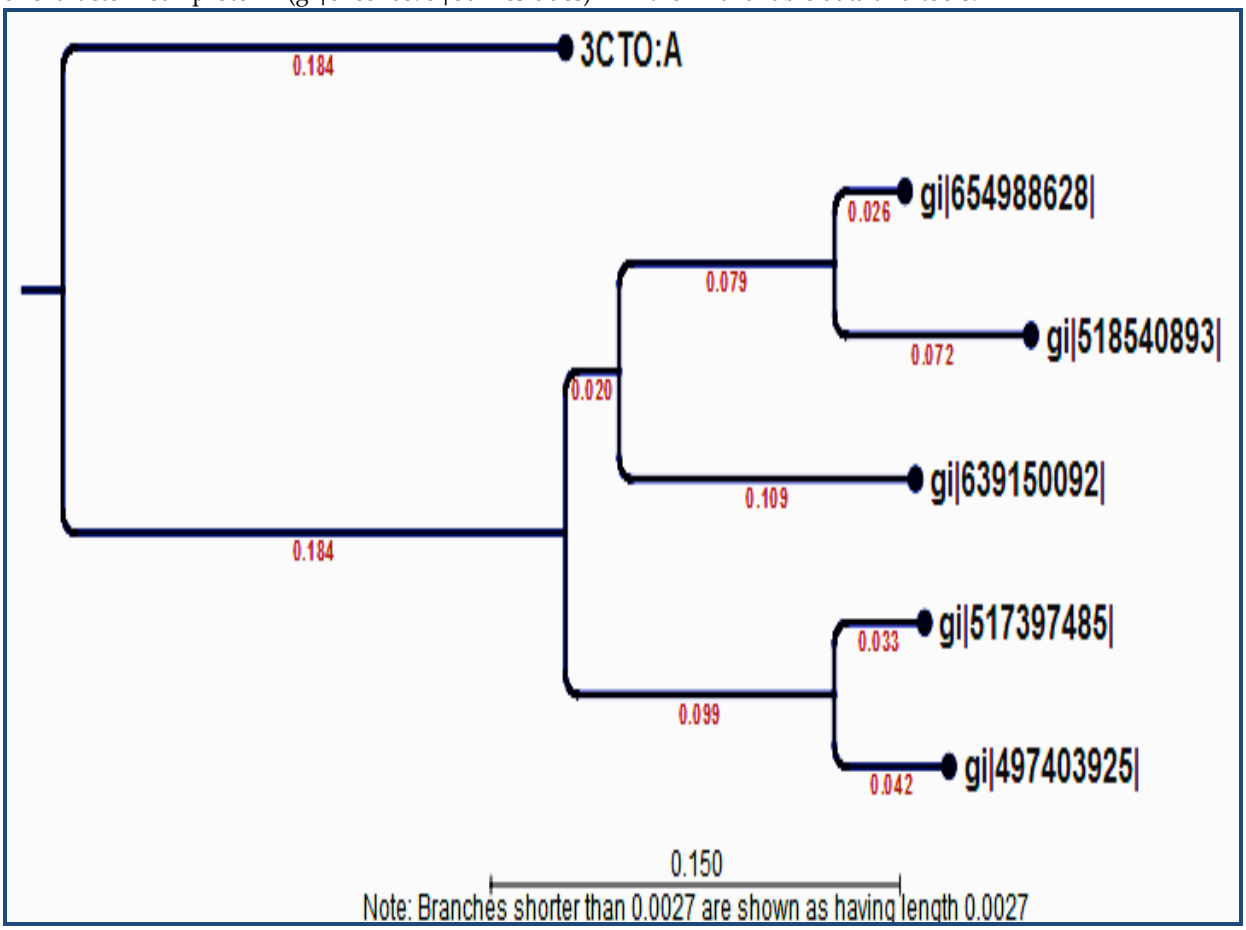

Figure 1: Phylogenetic analysis of different antitoxin protein of *Streptomyces* sp. with the target protein (gi|518540893|) having true distance (Red mark) is shown. Here, the neighbor joining method is used for the construction of the tree with bootstrap 10000. Closer distances with other annotated antitoxin proteins have placed the hypothetical protein in the same group.

Methodology:

Sequence retrieval

We inspected the NCBI (http:// www.ncbi.nlm.nih.gov/) **[14]** protein databases for proteins containing antitoxin like sequences. An uncharacterized protein (gi|518540893|) from *Streptomyces xinghaiensis* consisting of 86 amino acid residues was selected for the study and its sequence was downloaded in FASTA format for further analysis.

Analysis of physico-chemical properties

The ProtParam (http://web.expasy.org/protparam/) **[15]** tool of ExPASy was used for the analysis of the physical and chemical properties of the targeted protein sequence. The properties including aliphatic index (AI), GRAVY (grand average of hydropathy), extinction co-efficients, iso-electric point (p^I) and molecular weight were analyzed.

Sub-cellular localization prediction

Determining sub-cellular localization is crucial for understanding protein function and is also vital for genome analysis. Prediction of sub-cellular localization of the protein from *Streptomyces xinghaiensis* was completed using CELLO (version 2.5), a multiclass support vector machine classification system **[16, 17]**.

Protein family and phylogeny analysis

The BLASTp program from NCBI (http://www.ncbi.nlm.nih.gov/) **[18]** was used for searching the similarity of the protein against the non-redundant database with default parameters. Then the target protein was analyzed for the presence of conserved domains based on sequence similarity search with close orthologous family members. For this purpose, three different tools and/or databases including Proteins Families Database (Pfam), **[19]** NCBI Conserved Domains Database (NCBI-CDD), **[20]** and SUPERFAMILY **[21]** were used. Pfam is a database of protein families that includes annotations and multiple sequence alignments generated using hidden Markov models. NCBI-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. The SUPERFAMILY annotation is based on a collection of hidden Markov models, which represent structural protein domains for SCOP super-family level. The annotation is produced by scanning protein sequences from completely sequenced genomes against the hidden Markov models with these features. The phylogeny analysis was completed using the CLC Sequence Viewer v7.0.2 (http://www.clcbio.com) for understanding molecular evolution.

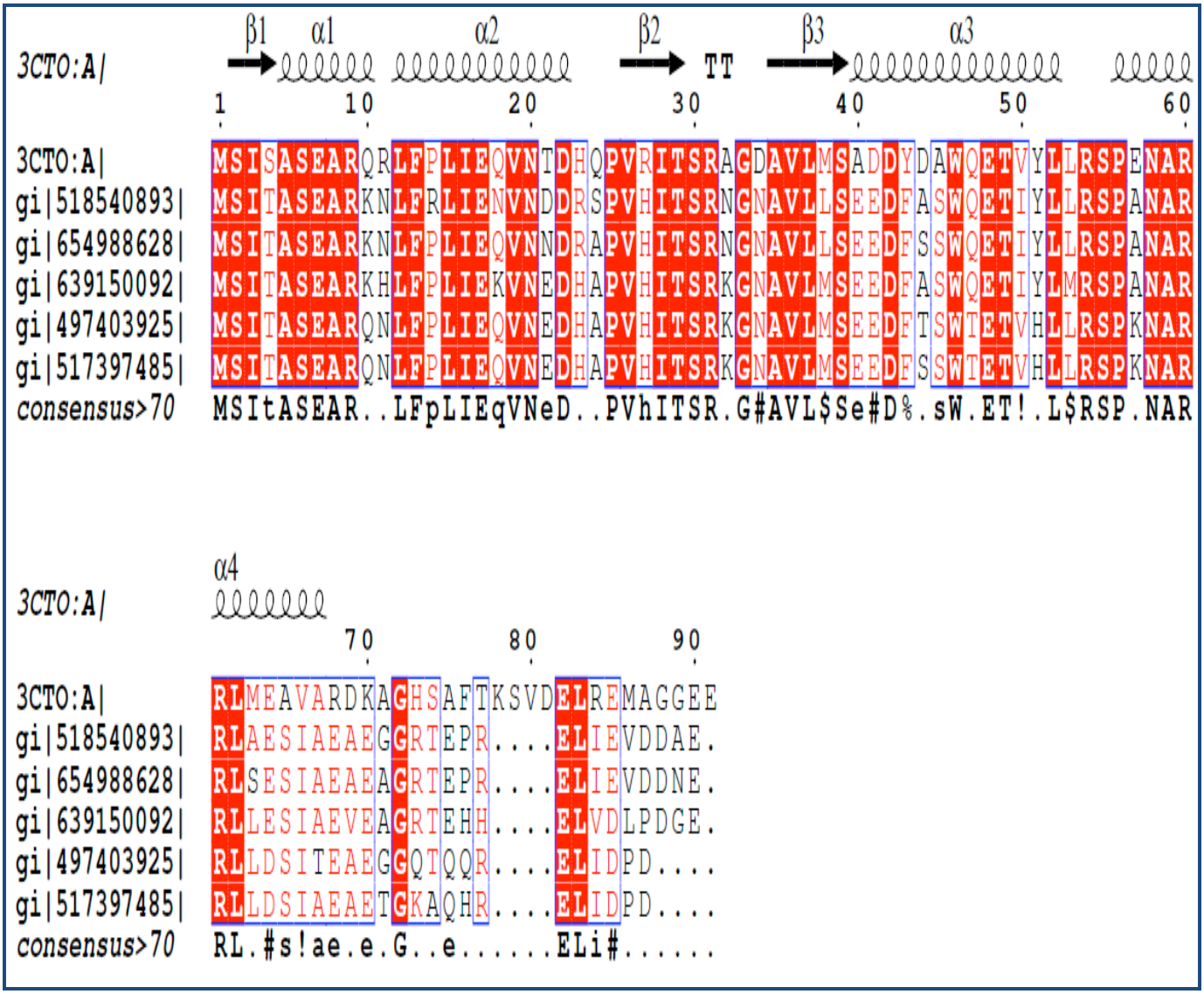

Figure 2: Multiple sequence alignment (MSA) of different antitoxin proteins with predicted secondary structure elements is shown. The sequence (gi|518540893|) for the target protein with the secondary structures (alpha helix and beta strands) is shown on the top of the alignment. The target protein shows 62% sequence similarity with the structure known template with PDB ID 3CTO:A. The rest of the sequences show 90% similarity with the target protein.

Multiple sequence alignment and Secondary structure analysis

A combined approach was used to get structural and functional insights through sequence comparison. We fetched several annotated antitoxin protein sequences of *Streptomyces* species from the NCBI protein database and the multiple sequence alignment (MSA) along with the target protein were obtained using BioEdit biological sequence alignment editor **[22].** These aligned sequences were used further for the prediction of the secondary structures using EsPript 3.0 **[23].**

Homology Modeling

Homology modeling was used to determine the three-dimensional structure of the target protein. A BLASTp **[18]** search with default parameters was performed against the Brookhaven Protein Data Bank (PDB) to find suitable templates for homology modeling. PDB ID: 3CTO: A, was identified as the best template based with 62% sequence similarity between query and template protein sequence. The tertiary structure was predicted using MODELLER **[24]** through HHpred **[25, 26]** tools of the Max Planck Institute for Developmental Biology.

Model quality assessment

The quality of the predicted structure was assessed by PROCHECK **[27]** and QMEAN6 **[28]** programs of ExPASy server of SWISS-MODEL Workspace **[29].** Furthermore, Root Mean Squared Deviation (RMSD), superimposition of query and template structure was generated by using UCSF Chimera 1.5.3 **[30].** The Z score of the template and query were also assessed by ProSA-web server **[31]**. Finally, the model and the template structure superimposed were visualized by using PyMOL **[32]** (The PyMOL Molecular Graphics System, Version 1.5.0.4, Schrödinger, and LLC).

Active site determination

Active site of the protein was determined by the computed atlas of surface topography of proteins (CASTp) **[33]** server, which provides an online resource for locating, delineating, and

measuring concave surface regions on the three-dimensional structures of proteins.

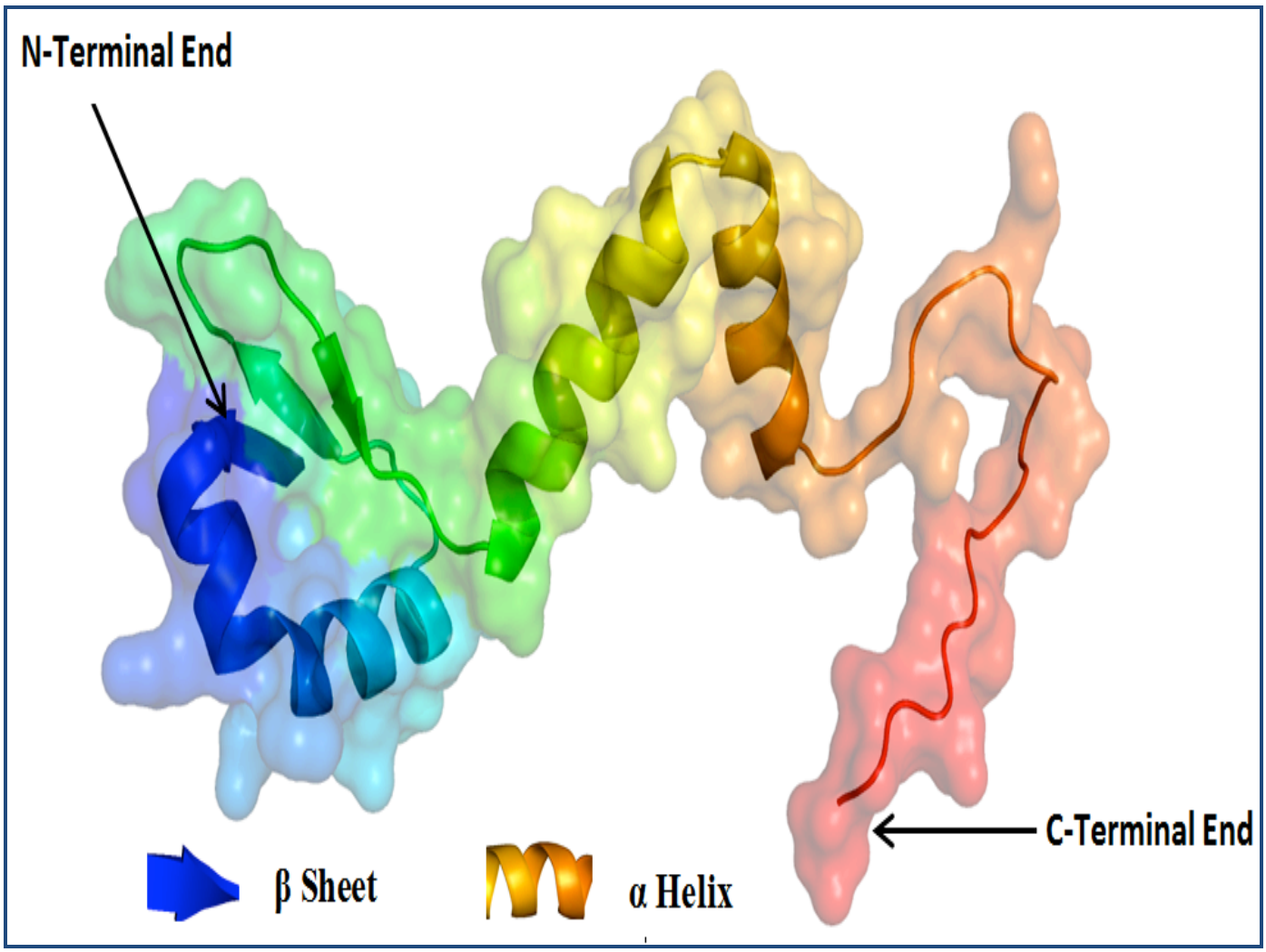

Figure 3: Predicted 3D structure of the target protein. The N-terminal end starts with beta sheet (Blue) and the C-terminal end is coiled structure (Red).

Results & Discussion:

Various physiological and chemical properties of the target protein were assessed by ProtParam tool. These include aliphatic index (AI; score of 88.60), instability index (II; score of 81.60), p^I; score of 4.61, extinction coefficient; score of 6990 and average hydro-pathicity; score of -0.573. All of these calculations are related to the stability of the protein for its function **[34].**

Sub-cellular localization is an essential feature of a protein. Cellular functions are usually localized in specific enclosed area; so, foretelling the sub-cellular localization of an unknown protein may possibly use to obtain handy information about their function. Therefore, this information is also valuable for drug designing for the target protein **[35].** Here, the sub-cellular localization of the target protein predicted by CELLO is cytoplasm. The BLASTp search against the non-redundant database showed homology (up to 90% sequence similarity) with other known antitoxin proteins from different *Streptomyces* species **Table 1 (see supplementary material).** Phylogenetic analysis is shown in **Figure 1** using the same data and their evolutionary relatedness is depicted. The output of the tree with the true distance inferred the evolutionary similarity of different antitoxin genes.

Numerous web tools were used to search for conserved domains and potential function of the target protein. Based on consensus predictions made by Pfam, NCBI-CDD and SUPERFAMILY suggested that the target protein contains PhdYeFM_antitox superfamily domains and is currently classified as antitoxin Phd_YefM in the type II toxin-antitoxin system. Pfam server predicted the Antitoxin Phd_YefM, type II toxin-antitoxin system at 1–74 amino acid residues with an e-value of 1.9e-21. The PhdYeFM_antitox super family was also found by the NCBI-CDD server at 2-81 amino acid residues with an e-value of 3.27e-20. The SUPERFAMILY server found the domain at positions 3-79 amino acid residues with an e-value of 2.49e-22. In this system, once the antitoxin protein is bound to their toxin companions, they bind DNA via the N-terminus and inhibit the expression of the operons, which contain genes encoding the TA system **[36, 37].**

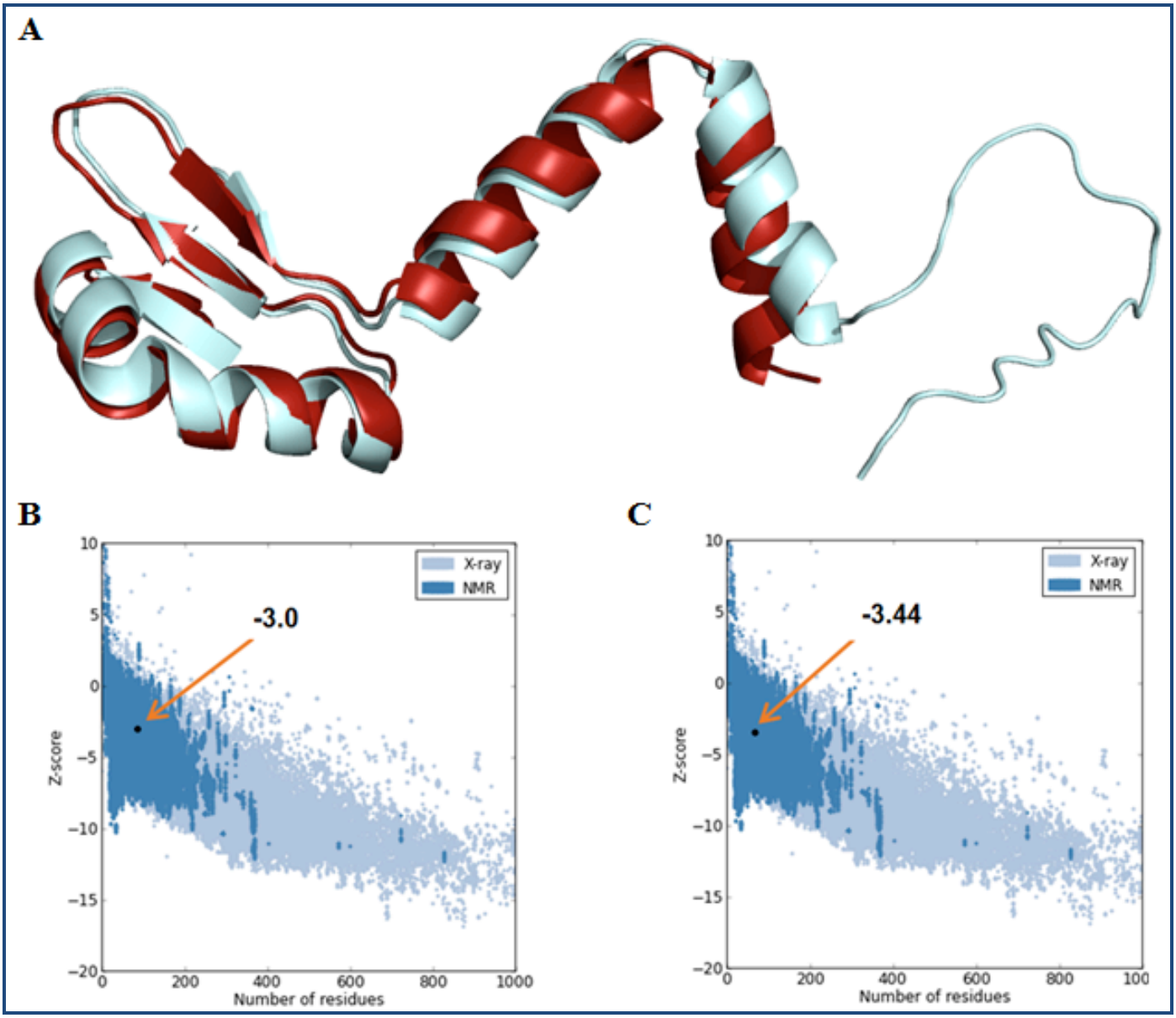

Figure 4: The 3D structure superposition of template structure and predicted model is shown. Here, in figure 5A, the template 3CTO:A (red color) and the target protein (cyan color) is shown. The RMSD value for this superposition is 0.709 Å. Figure 5B showed the Z score of the model (target protein) and Figure 5C showed the Z score of the template (3CTO:A).

The MSA of different antitoxin proteins of *Streptomyces* and the target protein (gi|518540893|) are depicted in **Figure 2.** The secondary structure of these proteins are also included in this figure and showed that they are mostly conserved throughout the alignment along with the template. Homology modeling is an important part in the recent past for the comparative modeling of various unknown structures with enormous available tools **[38, 39].** The structure for the target protein is unknown. Therefore, it is of interest to develop a homology model of the protein as shown in **Figure 3.** Here, the template (PDB ID: 3CTO: A) is *M. tuberculosis* YefM antitoxin with 62% sequence similarity with the target.

Quality assessment of the predicted 3D model was completed using PROCHECK using "Ramachandran plot" where we got 93.6% amino acid residues within the favored region. The quality of the model was further checked by QMEAN6 server where the model was placed inside the dark grey zone and considered as a good model with a QMEAN6 score of 0.608.

Superimposition between the model and the template is shown in **Figure 4A.** The RMSD value obtained from the superimposition of target and the template (3CTO: A) in UCSF Chimera was found to be 0.709 Å, suggesting a reliable three-dimensional model. The Z score evaluates the global model quality and is used to check whether the input structure is within the range of scores usually found for native proteins of similar size. The z for the model obtained from ProSA was -3 **(Figure 4B)** and for the template was -3.44 **(Figure 4C),** proposing the homology between target and the model. The active site of the protein was analyzed using the CASTp server. The identification and characterization of functional sites on proteins have increasingly become an area of interest. On account of the analysis of the active site residues for the binding of ligands provides insight towards the design of inhibitors of

an enzyme. In this study, we have also analyzed the best active site area of the protein as well as the number of amino acids involved **(Figure 5).** In most cases, class II antitoxin have two domains, one is DNA-binding domain located in the N-terminal region and other is toxin binding domain located in the C- terminal end **[40-43].** In our analysis, we have also found similar domain based active sites in the target protein model. Those were depicted using a spherical view in **Figure 5.**

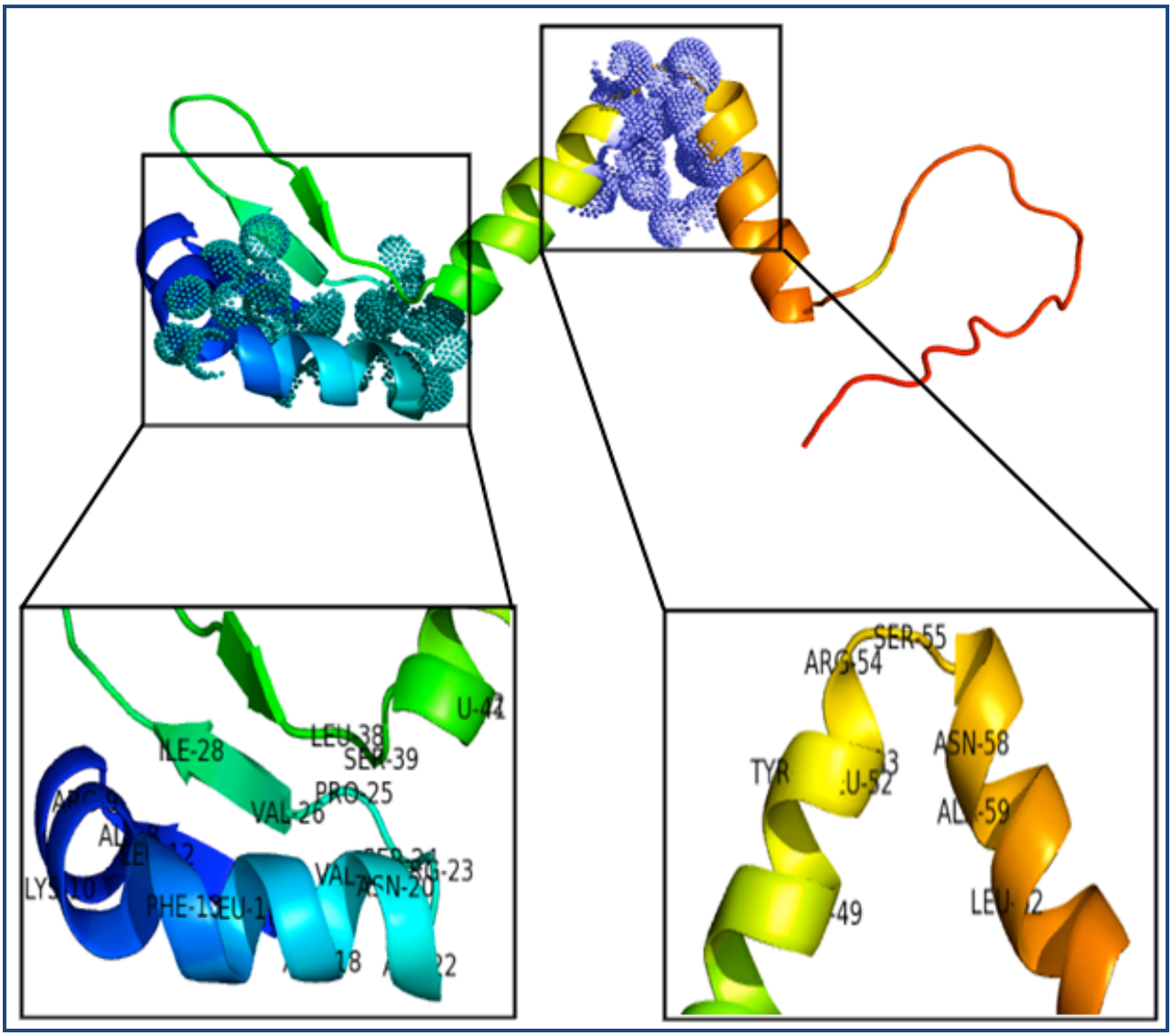

Figure 5: Active sites (spherical view) identification of the protein through the CASTp server is shown. Here, the amino acid residues in the active sites are depicted with zoomed view for better visualization. The N-terminal region starts from the left end (Blue marked) and the right end (Red coil region) is the C-terminal.

Conclusion:

We describe the homology model with possible assigned function of an uncharacterized protein from *Streptomyces xinghaiensis*. The analysis shows that target protein is antitoxin, which acts as in a type II toxin–antitoxin (TA) systems. This TA system composed of two genes encoding a labile antitoxin and a stable toxin. This data finds utility in the annotation of the target protein.

Acknowledgement:

The authors sincerely acknowledge Shah Adil Ishtiyaq Ahmad (Assistant professor of Biotechnology and Genetic Engineering Department) for providing the necessary suggestions and facilities throughout the study.

Reference:

[1] Chandra G & Chater KF, *FEMS Microbiol Rev.* 2014 **38**: 345 [PMID: 24164321]
[2] Zhao XQ *et al. Int J Syst Evol Mic robiol.* 2009 **59**: 2870 [PMID: 19628610]
[3] Subba B *et al. Mol Cells.* 2005 **20**: 90 [PMID: 16258246]
[4] Van Melderen L, *Curr Opin Microbiol.* 2010 **13**: 781 [PMID: 21041110]
[5] Lewis K, *Annu Rev Microbiol.* 2010 **64**: 357 [PMID: 20528688]
[6] Makarova KS *et al. Biol Direct.* 2009 **4**: 19 [PMID: 19493340]
[7] Bailey SE & Hayes F, *J Bacteriol.* 2009 **191**: 762 [PMID: 19028895]
[8] de la Hoz AB *et al.Proc Natl Acad Sci U S A.* 2000 **97**: 728 [PMID: 10639147]

[9] Hallez R *et al. Mol Microbiol.* 2010 **76**: 719 [PMID: 20345661]
[10] Bhatia U *et al. Science.* 1997 **276**: 1724 [PMID: 9206831]
[11] Oany AR *et al. Gene Regul Syst Bio.* 2014 **8**: 141 [PMID: 25574135]
[12] Oany AR *et al. Bioinform Biol Insights.* 2014 **8**: 65 [PMID: 24683305]
[13] Oany AR *et al. Austin J Comput Biol Bioinform.* 2014 **1**: 5
[14] Benson DA *et al. Nucleic Acids Res.* 2000 **28**: 15 [PMID: 10592170]
[15] Gasteiger E *et al. New York: Springer;* **2005**: 571
[16] Yu CS *et al. Protein Sci.* 2004 **13**: 1402 [PMID: 15096640]
[17] Yu CS *et al. Proteins.* 2006 **64**: 643 [PMID: 16752418]
[18] Altschul SF *et al. FEBS J.* 2005 **272**: 5101 [PMID: 16218944]
[19] Bateman A *et al. Nucleic Acids Res.* 2004 **32**: D138 [PMID: 14681378]
[20] Marchler-Bauer A *et al. Nucleic Acids Res.* 2015 **43**: D222 [PMID: 25414356]
[21] Gough J *et al. J Mol Biol.* 2001 **313**: 903 [PMID: 11697912]
[22] http://www.mbio.ncsu.edu/Bioedit/bioedit.html
[23] Gouet P & Courcelle E, *Bioinformatics.* 2002 **18**: 767 [PMID: 12050076]
[24] Sali A *et al. Proteins.* 1995 **23**: 318 [PMID: 8710825]
[25] Söding J, *Bioinformatics* 2005 **21**: 951 [PMID: 15531603]
[26] Söding J, *Nucleic Acids Res.* 2005 **33**: W244 [PMID: 15980461]
[27] Laskowski RA, *Journal of applied crystallography.* 1993 **26**: 283
[28] Benkert P *et al. Bioinformatics* 2011 **27**: 343 [PMID: 21134891]
[29] Arnold K *et al. Bioinformatics.* 2006 **22**: 195 [PMID: 16301204]
[30] Pettersen EF *et al. J Comput Chem.* 2004 **25**: 1605 [PMID: 15264254]
[31] Wiederstein M & Sippl MJ, *Nucleic Acids Res.* 2007 **35**: W407 [PMID:17517781]
[32] DeLano WL. The PyMOL molecular graphics system. 2002.
[33] Dundas J *et al. Nucleic Acids Res.* 2006 **34**: W116 [PMID: 16844972]
[34] Shoichet BK *et al. Proc Natl Acad Sci U S A.* 1995 **92**: 452 [PMID: 7831309]
[35] Wang J *et al. BMC Bioinformatics.* 2005 **6**: 174 [PMID: 16011808]
[36] Anantharaman V & Aravind L, *Genome Biol.* 2003 **4**: R81 [PMID: 14659018]
[37] Garcia-Pino A *et al. Cell.* 2010 **142**: 101 [PMID: 20603017]
[38] Vitkup D *et al. Nat Struct Biol.* 2001 **8**: 559 [PMID: 11373627]
[39] Chance MR *et al. Protein Sci.* 2002 **11**: 723 [PMID: 11910018]
[40] Santos-Sierra S *et al. FEMS Microbiol Lett.* 2002 **206**: 115 [PMID: 11786266]
[41] Smith JA & Magnuson RD, *J Bacteriol.* 2004 **186**: 2692 [PMID: 15090510]
[42] Bernard P & Couturier M, *Mol Gen Genet.* 1991 **226**: 297 [PMID: 2034222]
[43] Brown BL *et al. PLoS Pathog.* 2009 **5**: e1000706 [PMID: 20041169]

Edited by P. Kangueane

Citation: **Oany *et al.*** Bioinformation 11(11): 493-500 (2015)

BIOINFORMATION

open access

Supplementary material:

Table 1: Similar proteins obtained from non-redundant database.

GenBank ID	Organism	Protein name	Identity	Score	E-value
gi|518540893|	*Streptomyces sp. TAA486*	antitoxin	91%	157	2e-47
gi|654988628|	*Streptomyces sp. TAA486*	antitoxin	78%	138	4e-40
gi|639150092|	*Streptomyces himastatinicus*	antitoxin	75%	130	5e-37
gi|517397485|	*Streptomyces sp. PsTaAH-124*	antitoxin	73%	127	7e-36