

Research Article

Computational Structure Analysis and Function Prediction of an Uncharacterized Protein (I6U7D0) of *Pyrococcus furiosus* COM1

Oany AR*, Ahmad SAI, Siddikey MAA, Hossain MU and Ferdoushi A

Department of Biotechnology and Genetic Engineering, Mawlana Bhashani Science and Technology University, Bangladesh

***Corresponding author:** Oany AR, Department of Biotechnology and Genetic Engineering, Mawlana Bhashani Science and Technology University, Santosh-1902, Tangail, Bangladesh, Tel: +8801558819130; Email: arafatr@outlook.com

Received: October 24, 2014; **Accepted:** December 18, 2014; **Published:** December 22, 2014

Abstract

Pyrococcus furiosus is a hyperthermophilic Archaea. An uncharacterized protein of this Archaea, I6U7D0 (UniProt accession) containing 349 residues was selected for *in silico* analysis. Various bioinformatic tools were used to predict the structure and function of this protein. Sequence similarity was searched through UniProt and non-redundant database using BLASTp program of NCBI and homology was found with methyltransferases. Multiple sequence alignment was used to locate the conserved residues. The secondary and three dimensional structures were predicted. The validation of the three dimensional structure was obtained through PROCHECK, Verify3D and ERRAT program. CASTp server was used to predict the active site of the protein. Molecular docking with the ligand ACY (Acetic Acid) was performed using Molegro Virtual Docker to visualize the interactions between the ligand and amino acid residues in the protein. Finally, all the accumulated results suggested the biological function of the target protein to be a methyltransferase.

Keywords: Sequence alignments; Molecular docking; Protein-Ligand interactions; Active site

Introduction

Pyrococcus furiosus is a hyperthermophilic Archaea. It is considered a model organism to study the hyperthermophilic extremophiles, mostly due to its rapid growth at 100°C and the already sequenced genome [1]. Studying *P. furiosus*, as well as other extremophiles might possess a lot of potentials because of its unique genomic and physiological features. Its thermostable enzymes and other unique proteins might be used in various applications.

With the advancement in sequencing technologies, it is now considerably easier to obtain the whole genome sequence of such single cell organisms. Still, there are protein sequences with functions yet to be discovered or experimentally confirmed. We presume these uncharacterized proteins a vast unexplored field with numerous opportunities, both as medical and industrial tools. *In silico* analysis might assist in determining the biological functions of such uncharacterized proteins. This can partly be facilitated by predicting the three-dimensional (3D) structure of the targeted protein. When the experimentally obtained structure is unavailable, comparative or homology modelling can sometimes provide a useful 3D model for the protein of interest that is related to at least one known protein structure. Comparative or homology modelling predicts the 3D structure of a given protein sequence based primarily on its alignment to one or more proteins of known structure. Over the past few decades the number of sequences in the comprehensive public sequence databases, such as Swiss-Prot/TrEMBL [2] and GenPept, [3] have increased to a greater extent compared to the number of experimentally determined structures deposited in the Protein Data Bank (PDB) [4]. As a result, a gap has formed between the number of

known sequences and confirmed functions. The *in silico* prediction methods for 3D structure and biological function of proteins might assist in reducing this gap [5].

Prediction methods are based on fold assignment, target-template alignment, model building and model evaluation [6]. However, *in silico* predicted 3D structures are confirmed only by experimental methods such as X-ray crystallography and NMR spectroscopy [7]. Homology-based gene annotation has been the standard method for allocating a function to a novel uncharacterized protein during the last decades. With the development of new algorithms and bioinformatic tools, now a day various other methods can complement the classical homology search. These methods are designed to detect presumed functional constraints on genome evolution that is known as 'genomic context' approaches [8]. Some recent study has also followed the above analysis in order to propose the function of protein, which exists in protein level [9]. In this study, an attempt has been made to predict the structure and biological function of an uncharacterized protein (I6U7D0) using various bioinformatic tools.

Materials and Methods

Sequence retrieval

Initially we searched the UniProtKB protein (www.uniprot.org/) [10] and UniProt entry I6U7D0 of Archaea, *pyrococcus furiosus* consist of 349 amino acid residues was selected by targeted selection. Then the sequence was stored as a FASTA format sequence.

Physico-chemical properties analysis

The ProtParam (<http://web.expasy.org/protparam/>) [11] tool of ExPASy was used for the analysis of the proteins physiological

Table 1: Physico-chemical properties analysis of the hypothetical protein.

No of Amino acid	MW	pI	(Asp + Glu)	(Arg + Lys)	Ext. coefficient	Aliphatic index (AI)	Instability index (II)	Grand average of hydropathicity (GRAVY)
349	40326.8	4.68	67	43	49975	96.07	39.36	-0.323

Table 2: Similar protein obtained from UniProt database.

Entry Name	Organism	Protein name	Identity	Score	e-value
I3RDM7_9EURY	<i>Pyrococcus sp. ST04</i>	Putative methyltransferase	93%	1,720	0.0
Q9UZ33_PYRAB	<i>Pyrococcus abyssi</i>	Predicted methyltransferase, DUF43 family	91%	1,711	0.0
F0LHU2_THEBM	<i>Thermococcus barophilus</i>	Predicted methyltransferase	84%	1,576	0.0
H3ZM66_THELI	<i>Thermococcus litoralis</i>	Methyltransferase	81%	1,524	0.0
C6A0P9_THESM	<i>Thermococcus sibiricus</i>	Predicted methyltransferase	78%	1,478	0.0

Table 3: Similar protein obtained from Non-redundant UniProt KB/SwissProt sequences.

Protein ID	Organism	Protein name	Identity	Score	e-value
ref WP_014733863.1	<i>Pyrococcus sp.</i>	Methyltransferase	93%	673	0.0
ref WP_013467640.1	<i>Thermococcus barophilus</i>	Methyltransferase	84%	615	0.0
ref WP_012766155.1	<i>Thermococcus sibiricus</i>	Methyltransferase	78%	577	0.0
ref WP_014806245.1	<i>Anaerobaculum mobile</i>	Methyltransferase	49%	328	7e-107
ref WP_011026055.1	<i>Thermoanaerobacter tengcongensis</i>	Methyltransferase	45%	316	3e-102

and chemical properties deduced from our protein sequence. The properties including aliphatic index, GRAVY (Grand Average of Hydropathy), Extinction coefficients, isoelectric point (pI), molecular weight etc. were analyzed through this tool.

Homology identification

To get the preliminary prediction about the function of the targeted protein, similarity search was performed with the NCBI protein Database (<http://www.ncbi.nlm.nih.gov/>) against non-redundant and SwissProt [12] database to find out the proteins that might have structural similarities with that of the uncharacterized protein by using BLASTp program [13].

Structure prediction

The retrieved sequence was used for the prediction of the Secondary structure of the protein by SABLE server (<http://sable.cchmc.org/>) [14] and the tertiary structure was predicted by (PS)²v2 server (<http://ps2v2.life.nctu.edu.tw/>) of the Molecular Bioinformatics Center, National Chiao Tung University [15]. The three dimensional structure was predicted one the basis of best scoring template for higher accuracy.

Model quality assessment

Finally the quality of the predicted three dimensional structure was assessed by PROCHECK [16], Verify3D (http://nihserver.mbi.ucla.edu/Verify_3D/) [17] and ERRAT Structure Evaluation server [18].

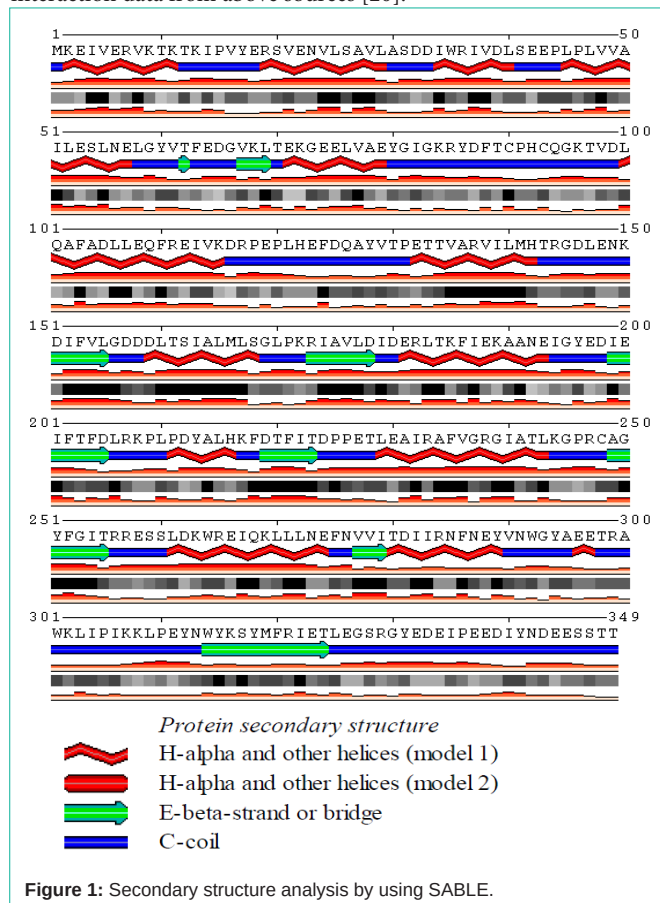
Multiple sequence alignment and phylogeny analysis

Multiple sequence alignment was carried between the uncharacterized protein and the proteins that had structural similarity with uncharacterized protein by BioEdit biological sequence alignment editor tool [19]. The phylogeny analysis was done by CLC Sequence Viewer v7.0.2 (<http://www.clcbio.com>).

Protein-Protein Interaction Analysis

Protein residues are interacting with each other for their accurate

functions. Here we used STRING (<http://string-db.org/>) a database of known and predicted protein interactions, works through physical and functional associations. That derived from Genomic Context, high-throughput experiments, (Conserved) Co-expression, and Previous Knowledge. This database is quantitatively integrates interaction data from above sources [20].



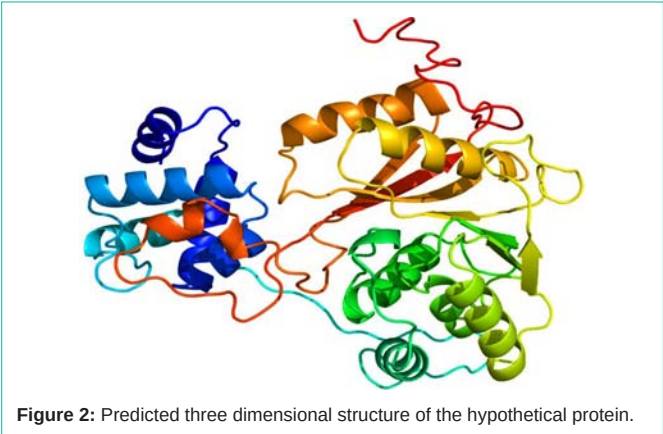


Figure 2: Predicted three dimensional structure of the hypothetical protein.

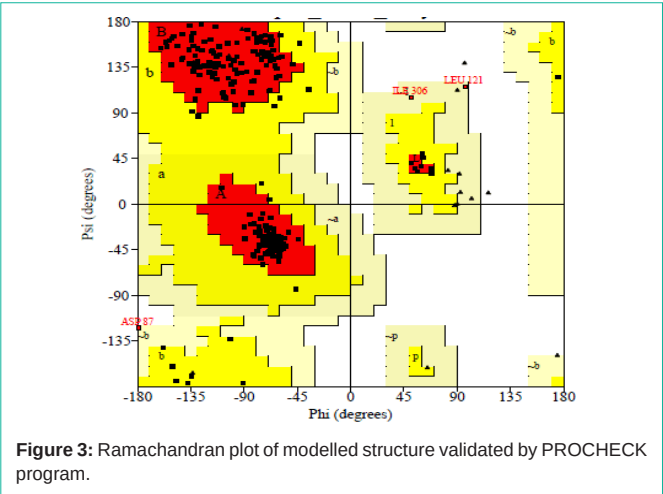


Figure 3: Ramachandran plot of modelled structure validated by PROCHECK program.

Active site detection

The active site of the protein was determined by the Computed Atlas of Surface Topography of Protein (CASTp) (<http://sts.bioengr.uic.edu/castp/>) [21] provides an online resource for locating, delineating and measuring concave surface regions on three-dimensional structures of proteins.

Table 4: Ramachandran plot statistics of the hypothetical protein.

Ramachandran plot statistics		(%)
Residues in the most favored regions [A, B, L]	290	92.7%
Residues in the additional allowed regions [a, b, l, p]	20	6.4%
Residues in the generously allowed regions [a, b, l, p]	3	1%
Residues in the disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	313	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown in triangles)	18	
Number of proline residues	16	
Total number of residues	342	

Table 5: Dock score, number of hydrogen bonds, interacting residues of 2QM3 and I6U7D0 with ACY (ligand).

Acetic acid (ACY)	Dock Score[GRID] (kcal/mol)	No. of H bonds	Interacting Residues
2QM3	-59.3778	6	Ala 295A, Trp 292(2)A, Trp 314A, Gly 293A, Tyr 294A
I6U7D0	-53.9562	4	Ala 295, Tyr 294, Gly 293, Trp 292, Glu 296

Comparative docking analysis

Further, docking studies were initiated by using Molegro Virtual Docker (MVD) [22]. However, Molegro Virtual Docker (MVD) is an integrated environment for studying and predicting how ligands interact with macro-molecules, it is usually work within specific grid line is defined by a position (x, y, z), for the hypothetical protein the grid line incorporating ligand were, X=56.69; Y=16.86; and Z=36.17. Before the docking study was perform, we fetched the ligand molecule from the (PDB: 2QM3), a methyltransferase protein of *p. furiosus*. Then docking both of the protein was performed to validate it.

Results and Discussion

The physiological and chemical properties of the hypothetical protein are described in Table 1. The BLASTp results against non-redundant and SwissProt database are shown in Table 2 and 3. Blastp analysis of the FASTA sequence of the targeted protein against non-redundant and SwissProt databases revealed an average of 80%

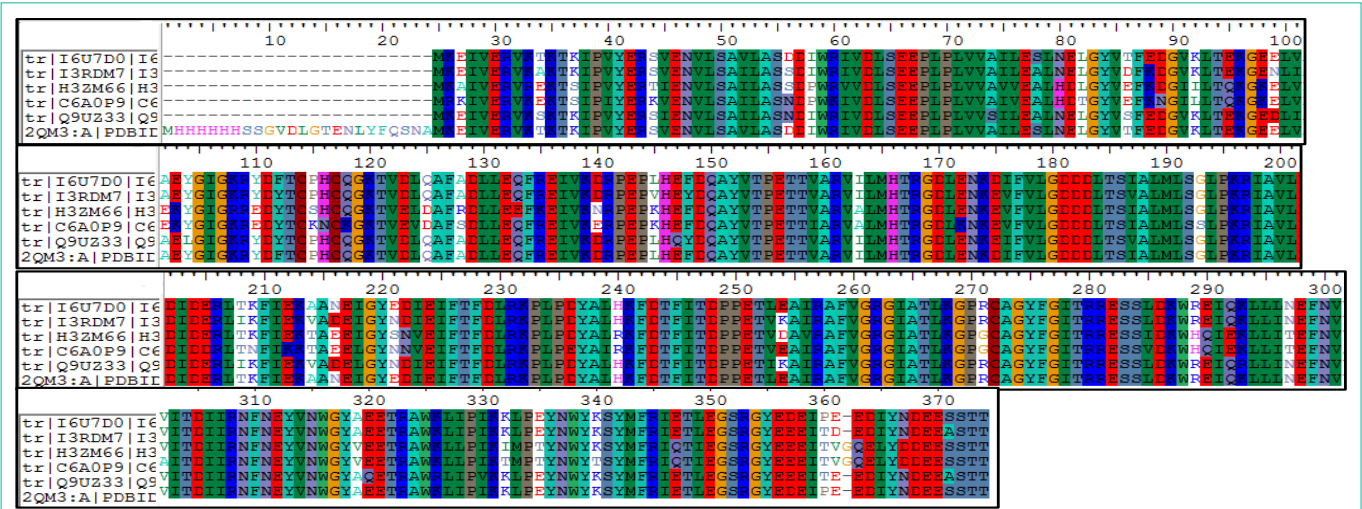


Figure 4: Multiple sequence alignment of different homologous protein.

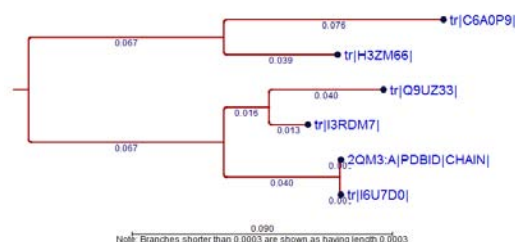


Figure 5: Phylogenetic trees with true distance of different methyltransferase proteins.

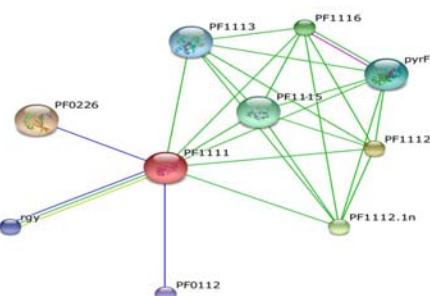


Figure 6: String network analysis of the hypothetical protein, indicates as PF1111.

A.



B.

Chain 0

```

1: MKEIVERVKT KTKIPYERS VENVLSAVLA SDDIWRIVOL SEEPFLVYA
51: ILESNELGY VTFEDGVKLT EKEELVAEY GIGKRYDFTG PHCGGKTVOL
101: QAFADLLEGF REIVKORPEP LKKEEDAAVYT PETTVAVIL MHTRODLENK
151: DITVLDGDDG SIALMLLSOL PRRIALDID ERLTKFIEKA ANEISGYEDIE
201: IFTFDKXPL PDYALHKFDT FIDPPPTLE ATRAYVORG ATLKSPRCAG
251: YGSTRRDS LDKWREIQKL LLNEFNVT DIIKNFNV NWGYASTRA
301: WKLIPIKKLP EYNWTFYF RIETLEGSRG YEDEIPEEDI YNDEESSTT
  
```

Figure 7: Active site of the hypothetical protein. (A) Here the green sphere indicates the active site of the protein. (B) The amino acid residues in the active site (Green color).

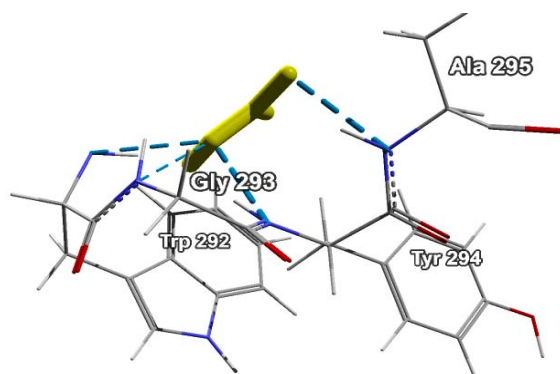


Figure 8: Molecular Docking (Targeted protein-ligand interaction).

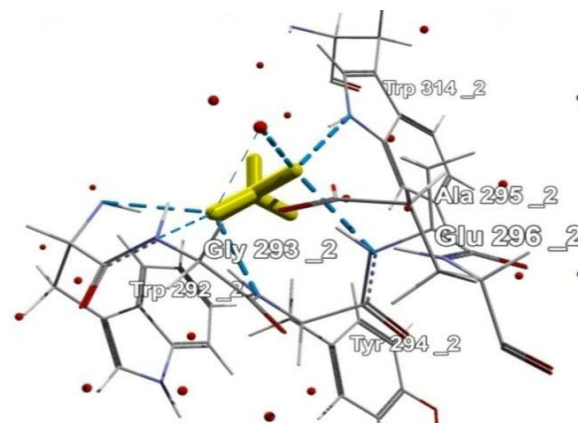


Figure 9: Molecular Docking (2QM3 protein-ligand interaction) with water complex (red).



Figure 10: Complete protein-ligand interactions.

homology with other methyltransferase proteins. SABLE server predicted the secondary structure of the protein having a good confidence of the prediction (Figure 1) and the (PS)² server predicted the three dimensional structure of the protein with 96.49% identity with the highest scoring template (PDB ID: 2QM3A) depicted in Figure 2. Validation of the predicted three dimensional model was assessed by PROCHECK through Ramachandran plot, where shows the distribution of ϕ and ψ angle in the model within the limits (Figure 3 and Table 4). Residues in the most favored regions covered 92.7%, which is the quality of a valid model. Finally the established model of 3D structure for the target sequence was verified by structure validation server verifies 3D and ERRAT. The highest score of 0.72 in the Verify 3D graph indicates that the environmental profile of the model is good and the overall quality factor predicted by the ERRAT server was 87.574 indicates a good model.

Multiple sequence alignment (Figure 4) was considered the FASTA sequences of the uncharacterized protein (I6U7D0) and the homologous annotated proteins. In order to confirm homology assessment between the proteins, down to the complex and subunit level, phylogenetic analysis was additionally performed. Phylogenetic tree was constructed based on the alignment and BLAST result give the similar concept about the protein is shown in Figure 5. The distances between branches are also included.

STRING protein-protein interaction network revealed that our hypothetical protein strongly interacts with the reverse gyrase (rgy) protein, this interaction gives us some insights that the protein could act as DNA/RNA methyltransferase (Figure 6). The predicted active site with their amino acid residues of the protein were depicted in Figure 7. Finally the comparative molecular docking study with the ligand ACY that was bounded in the active site of the protein that ensures us the function of the hypothetical protein is methyltransferase which displayed binding energies of - 59.3778 and -53.9562 kcal/mol for 2QM3 and I6U7DO proteins, respectively (Figure 8 and Table 5).

Conclusion

The study was designed to predict the three dimensional structure and biological function of I6U7D0, an uncharacterized protein of *P. furiosus* COM1. All the above findings suggested that the function of the target protein is methyltransferase. Hence, the computational approach followed in this study in predicting the function of an unknown protein envisage the utility of bioinformatics tools in predicting functional aspects, thereby assisting experimental studies on a protein.

References

- Robb FT, Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, et al. Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology. *Methods in Enzymol.* 2001; 330: 134-157.
- Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 2003; 31: 365-370.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. *Nucleic acids research.* 2002; 30: 17-20.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000; 28: 235-242.
- Baker D, Sali A. Protein structure prediction and structural genomics. *Science.* 2001; 294: 93-96.
- Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, et al. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* 2006; 34: D291-295.
- Brenner SE, Levitt M. Expectations from structural genomics. *Protein Sci.* 2000; 9: 197-200.
- Doerks T, von Mering C, Bork P. Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. *Nucleic Acids Res.* 2004; 32: 6321-6326.
- Oany AR, Jyoti TP, Ahmad SA. An In Silico Approach for Characterization of an Aminoglycoside Antibiotic-Resistant Methyltransferase Protein from *Pyrococcus furiosus* (DSM 3638). *Bioinformatics and biology insights.* 2014; 8: 65-72.
- Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. *Database.* 2011; 2011: bar009.
- Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A, et al. Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*: Springer. 2005: 571-607.
- Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 2003; 31: 365-370.
- Johnson M, Zaretskaya I, Raytselis Y, Merezukh Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. *Nucleic Acids Res.* 2008; 36: W5-9.
- Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics.* 2005; 59: 467-475.
- Chen CC, Hwang JK, Yang JM. (PS)2-v2: template-based protein structure prediction server. *BMC Bioinformatics.* 2009; 10: 366.
- Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography.* 1993; 26: 283-291.
- Eisenberg D, Lüthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* 1997; 277: 396-404.
- Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* 1993; 2: 1511-1519.
- Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Paper presented at: Nucleic acids symposium series. 1999.
- Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 2013; 41: D808-815.
- Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic acids research.* 2006; 34: W116-W118.
- Thomsen R, Christensen MH. MolDock: a new technique for high-accuracy molecular docking. *J Med Chem.* 2006; 49: 3315-3321.