Prediction of Structure and Function of Metagenomic Sequence SUZ71503

Rebecca Feeley[1*], Thais Gallo Nisenbaum[2*], Lindsay A Wilhelmus[1*]

[1] Bioinformatics, Johns Hopkins University, Baltimore, MD, 21218, USA
[2] Biotechnology, Johns Hopkins University, Baltimore, MD, 21218, USA

* To whom correspondence should be addressed.
Name: Rebecca Feeley
Email: rfeeley1@jhu.edu

Name: Thais Gallo Nisenbaum
Email: tgallon1@jhu.edu

Name: Lindsay A Wilhelmus
Email: lwilhel2@jhu.edu

ABSTRACT

The analysis of SUZ71503 and a similar protein structure revealed that it may function as an oxidoreductase. Investigating the physiochemical properties, domains, localization, and predicted structure of the protein identified, SUZ71503 is a 26 kDa protein with a net negative charge at physiological pH and an isoelectric point of 4.8. The protein is localized to the cytoplasm and is comprised of both β-sheets and α-helices. It contains more β-sheets than α-helices, but the α-helices are longer in length. Further analysis using various software and web-based programs led to the discovery of a defining feature of the sequence, the Rossman-fold domain, present in many oxidoreductase proteins. SUZ71503 3D structure was modeled based on putative oxidoreductase Rv2002, a tyrosine-dependent oxidoreductase similarly structured to SUZ71503. Putative oxidoreductase Rv2002 was also identified as a close match by NCBI BLAST, demonstrating both high-sequence and high-structure alignment.

INTRODUCTION

SUZ71503 is a hypothetical 249 amino acid sequence from a marine metagenome (NCBI taxonomy ID: 408172). The sequence was translated from the CDS positions 13139 to 13888 on locus 24357 of the metagenome-assembled genome UINC01001124.1, which contains a mixture of DNA sequences from various organisms. This sequence belongs to the metagenomic project Metzyme (PRJEB26691), which analyzes the genome of marine microorganisms located in the depths of the Pacific Ocean. Using metagenomics, the goal of the Metzyme Bioproject is to characterize genes responsible for nitrification, or the oxidation of ammonia (NH_3) to nitrite (NO_2-), followed by the oxidation of NO_2- to nitrate (NO_3-) (1). Nitrification is carried out by certain bacteria and archaea, and the key enzymes involved in the process include ammonia monooxygenase, hydroxylamine dehydrogenase, and nitrite oxidoreductase (2).

The goal of this analysis was to determine the structure and function of SUZ71503. First, we calculated the protein's isoelectric point, molecular weight, and net charge at specific pH values. According to NCBI conserved domain database, SUZ71503 contains a Rossmann-fold NAD(P)-binding domain. Using protein prediction tools, our team identified the protein's subcellular localization and secondary structure and generated a 3D model of the protein structure. The identification and analysis of a similar, known protein identified by BLASTp allowed us to determine the accuracy of the prediction tools and draw further conclusions for SUZ71503. The 3D structure was compared to a similar protein from *B. megaterium*.

MATERIAL AND METHODS

I. Physicochemical Properties

The isoelectric point, molecular weight, and net charge at different pH values of SUZ71503 were calculated using R version 4.2.2 and the Peptides R package version 2.4.4. The sequence was defined according to the FASTA sequence for SUZ71503.1 in the NCBI protein database. The Peptides package in R provides several functions used to calculate the physicochemical properties of protein sequences, providing information on their classification.

II. Protein Domain Identification

The domains present were investigated using the NCBI conserved domain database and confirmed by InterPro. The web-based InterproScan employs the InterPro database, and it's overlapping information on known protein families and their domains and functional sites within UniProtKB and other resources to accurately predict the protein family and functional domains of a query sequence and integrate them into one result (3).

III. Protein Localization Prediction

The subcellular localization of SUZ71503 was predicted using PSORTb v3.0.3, a program used for predicting the protein localization of prokaryotes and archaea. PSORTb makes a prediction by analyzing different modules containing information about biological features known to affect protein localization. The results are verified according to the database PSORTdb, which contains various protein localizations that were verified experimentally (4). The analysis was done for gram-positive and gram-negative bacteria and archaea.

IV. Secondary Structure Prediction

The secondary structure of SUZ71503 was predicted using neural network models originally developed by Qian & Sejnowski. Neural network models utilize a training set to predict α-helices, β-sheets, and coil in globular proteins (5).

Two neural network web-based programs were used for the prediction: Hierarchical Neural Network (HNN) and PHD. HNN uses a combination of statistical methods and amino acid physicochemical properties to make a prediction. The PHD algorithm makes a prediction by searching the sequence in Blastp and performing multiple sequence alignment, which is then used as the input for the program. (6).

V. Protein Structure Prediction

The amino acid 3D structure was modeled using Protein Homology/analogY Recognition Engine V 2.0 (PHYRE2).

PHYRE2 works in four distinct stages. First, the software gathers homologous sequences to the query by searching for profile-profile matching with increased sensitivity compared to PSI-blast. The secondary structure is then predicted by PSI-Pred using neural networks that detect the presence of alpha helices and beta sheets with 75%-80% accuracy. Stage 2 builds on the profile and prediction generated in stage 1 and generates a hidden markov model (HMM) that is scanned against a large database of HMMs to generate a crude prediction of the backbone that may contain insertions, deletions, and lacks side chains. Stage 3, loop modeling, handles indels by filling in the gaps in the model with a library of known protein fragments. The fragments are fitted and scored based on how minimally they change the dihedral angles of the model until a top-scoring model is selected. Finally, in stage 4, the sidechains are fitted to the modeled backbone, which can be about 80% accurate if the backbone is correct (7).

VI. Identification and Analysis of Related Protein in the PDB Database

To identify the top match in the PDB database, NCBI Blastp was used with the FASTA sequence of SUZ71503 as the query, and results were restricted to the PDB database. The match was selected based on the lowest E-value result. The protein was further inspected to determine its SCOP/CATH classifications, which are provided as annotations in the PDB database.

VII. Superimposition of Related Proteins

The Vector Alignment Search Tool (VAST) was used on the 1NFF protein structure to find the related protein 3AUT. The Cn3D viewer was used to align the sequences and superimpose structures.

RESULTS

I. Physicochemical properties of SUZ71503

Isoelectric point: The isoelectric point of the protein was calculated using the Stryer pKa scale and was found to be 4.788308.

Molecular Weight: The monoisotopic molecular weight of the protein, representing the most common isotope of a given amino acid, was found to be 26039.13 Da

Net Charge: The protein's net charge was calculated using the Lehninger pKa scale at two different pH values. At a pH of 7.0, the protein was found to have a net charge of -12.81968. At a pH of 7.6, the protein was found to have a net charge of -13.36323.

A protein's isoelectric point (pI) is the pH in which the protein carries a charge of 0. As the theoretical pI is lower than the pH values of 7.0 and 7.6, the protein carries a net negative charge, as seen above. Additionally, SUZ71503 would be soluble and acidic at the given pH values.

II. SUZ71503 contains a Rossmann-fold NAD(P)-binding domain

Both the NCBI Conserved Domains database and InterProScan identified the Rossmann-fold NAD(P)-binding domain in the amino acid sequence. The domain consists of a central beta-sheet surrounded by an alpha/beta folding pattern. It is made up of a polypeptide binding site and a NAD chemical binding site. Rossman-fold domains are extremely common in proteins and make up about 20% of known protein structures found in PDB (8).

III. The subcellular localization of SUZ71503 is the cytoplasm

PSORTb v3.0.3 predicted SUZ71503 to be localized in the cytoplasm for gram-positive and gram-negative bacteria and archaea, with a confidence score of 9.97 and 9.96, respectively. The final prediction was based on results from two PSORTb analytical modules: CytoSVM and SCL-BLAST, both of which predicted SUZ71503 to be localized in the cytoplasm.

CytoSVM determines if a protein belongs to the cytoplasm based on machine-learning algorithms. SCL-BLAST searches the protein using BLASTp against a subset of the PSORTdb database for each organism selected. The SCL-BLAST search parameters contain an E-value threshold of 10e-9 and length restrictions (9).

IV. The secondary structure of SUZ71503 is classified as α-β

A protein's secondary structure is the local spatial arrangement of the protein's backbone determined by amino acid side chains. Common secondary structures include α-helices and β-sheets. The whole sequence comprises of 249 amino acids; the most common amino acids in sequence are:

- Glycine (G) 30 (12.0%) - Polar

- Alanine (A): 28 (11.2%) - Hydrophobic
- Valine (V): 22 (8.8%) - Hydrophobic
- Leucine (L): 18 (7.2%) - Hydrophobic
- Glutamic Acid (E): 18 (7.2%) - Polar, negative charge

HNN Prediction: SUZ71503 is comprised of more β-sheets than α-helices, but α-helices are longer in length. Specifically, the protein is predicted to contain 15 β-sheets and 9 α-helices. The sequence is predicted to be 22.49% (56 amino acids) α-helical and 24.90% β-sheet (62 amino acids). The remaining 52.61% of the sequence is predicted as random coils. The average length of the α-helices is 6.2 amino acids, with the longest comprising of 15 residues and the shortest comprising of 1 residue. The average length of the β-sheets is 4.1 amino acids, with the longest comprising of 9 residues and the shortest comprising of 1 residue.

PHD Prediction: SUZ71503 is comprised of more β-sheets than α-helices, but α-helices are longer in length. Specifically, the protein is predicted to contain 12 β-sheets and 7 α-helices. The sequence is predicted to be 30.52% (76 amino acids) α-helical and 23.69% β-sheet (59 amino acids). The remaining 45781% of the sequence is predicted as random coils. Although there are fewer α-helices, they appear to be much larger in length than the β-sheets. The average length of the α-helices 8.4 amino acids, with the longest comprising of 15 residues and the shortest comprising of 7 residues. The average length of the β-sheets is 5 amino acids, with the longest comprising of 7 residues and the shortest comprising of 1 residue.

V. 3D modelling of SUZ71503 predicted by PHYRE2

The results from PHYRE2 generated a model in which 236 residues (95%) of our query sequence (SUZ71503) was modeled with 100% confidence. The model was built based on the template for 1nffa_, a member of the Tyrosine-dependent oxidoreductases family that contains an NAD(P)-binding Rossman-fold domain and has a 51% identity with the query. Additional 100% confidence models were generated based on other oxidoreductases and Rossman-fold containing structures.

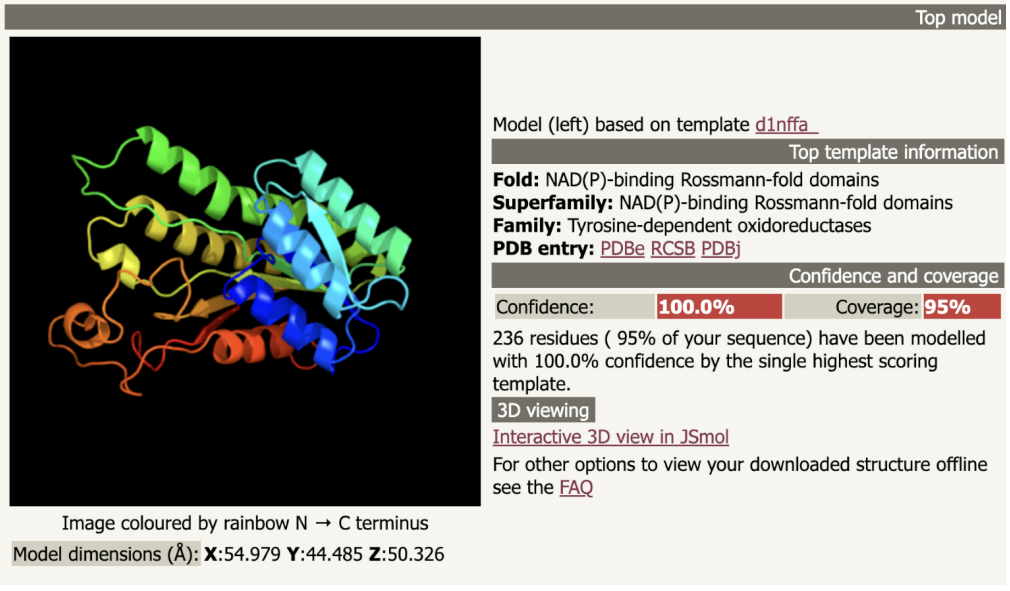

Figure 1. A depiction of SUZ71503 protein model generated by PHYRE2 based on template 1nffa

Figure 2.
An alternative view of the 3D model of SUZ71503 from PHYRE2 depicting the secondary structure. Alpha helices are shown in pink, surrounding the central beta sheets in yellow.

VI. Chain A of putative oxidoreductase Rv2002 is similar to SUZ71503

The top BLAST hit in the PDB database is the Chain A, Putative oxidoreductase Rv2002 from *Mycobacterium tuberculosis* (Accession: 1NFF_A). It is part of the "3beta17beta hydroxysteroid dehydrogenase-like, classical" conserved protein domain family, which is described by NCBI as a functionally diverse family of single-domain oxidoreductase proteins that all share a Rossman fold structure. The function of an oxidoreductase is to catalyze oxidoreduction reactions through the transfer of electrons, where oxygen acts as the acceptor molecule. This is important in the cellular function of aerobic and anaerobic organisms. This is the same molecule used as a template for the 3D model of the SUZ71503 structure.

VII. Analysis and Structure of putative oxidoreductase Rv2002 in the PDB Database

CATH, SCOP, and SCOP2 are databases that classify proteins according to their structures and domains. The CATH database was established in the mid-1990s (10). It classifies proteins into five different levels: class, architecture, topology, homologous superfamily, and sequence family (11). The SCOP database was first released in 1994, and SCOP 2 began development in 2014 (12). Both SCOP and SCOP2 rely on experts to classify proteins with structures in the Protein Data Bank. However, the protein classification in SCOP was based on a tree-like classification. As the amount of data became larger, the complex relationship between protein structures became clear. SCOP2 classifies structures in a complex network of nodes, not a tree (13).

In the CATH database, the Class level simply describes the protein as mainly alpha-helix, mainly beta-sheet, or both. The 1NFF protein is in class Alpha Beta, meaning that it has both alpha helices and beta sheets. The architecture is a 3-Layer (aba) sandwich, the most common protein architecture.

The topology, the Rossmann fold, is also the most common (14). The homology is NAD(P)-binding Rossmann-like Domain.

The SCOP database classifications of 1NFF are similar to the CATH database classification. This is an alpha and beta protein with a NAD(P)-binding Rossmann-fold domain. The SCOP database gives more information for the Family: tyrosine-dependent oxidoreductase. The SCOP2 database provides another classification: SDR, or short-chain dehydrogenase/reductase.

VIII. Superimposition of putative oxidoreductase Rv2002 onto the related structure

The VAST tool was used to find related structures to the putative oxidoreductase Rv2002. The protein 3AUT, a glucose dehydrogenase enzyme from *B. megaterium*, was selected. This protein was selected due to its relatively high sequence identity (33%) and the fact that it uses NADH as a ligand, similar to the 1NFF protein.

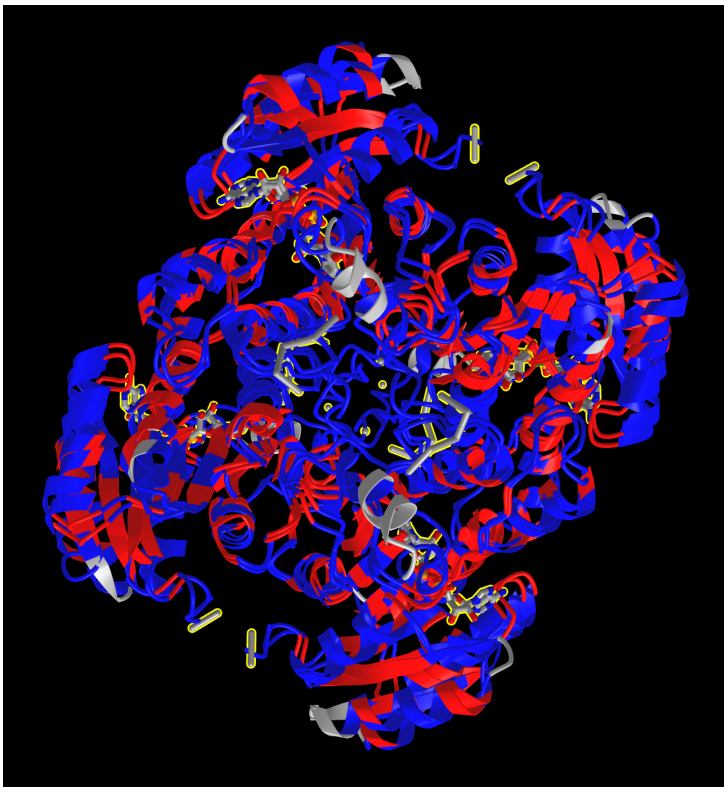

Figure 3:
3D model of the superimposition of 1NFF with 3AUT. Identical residues are in red and nonidentical residues are in blue. Unmatched residues are in grey. The yellow highlighted molecules are NADH.

DISCUSSION

As a result of the investigations on the hypothetical protein SUZ71503, we believe it is an oxidoreductase enzyme, which catalyzes oxidation-reduction reactions.

According to the physicochemical data calculated by the Peptides R package, SUZ71503 has a monoisotopic molecular weight of around 26kDa and a pI of 4.8. At the pH values of 7.0 and 7.6, the protein would be considered soluble, acidic, and negatively charged. However, different pH values would alter these characteristics.

Preliminary investigations identified SUZ71503 contains the Rossmann-fold NAD(P)-binding domain, according to the NCBI Conserved Domains database and InterProScan, which demonstrates an α-β folding pattern. The protein was predicted to localize in the cytoplasm, according to PSORTb v3.0.3. The folding pattern observed in the Rossmann-fold NAD(P)-binding domain is consistent with the secondary structure prediction programs HNN and PHD, which classified SUZ71503 as an α-β protein. SUZ71503 appears to have more β-sheets than α-helices, but α-helices are longer in length. The 3D structure predicted by PHYRE2 was modeled based on the structure identified by NCBI Blastp, a tyrosine-dependent oxidoreductase containing a NAD(P)-binding Rossmann-fold domain.

Using NCBI's Blastp against the PDB database, our team identified Chain A, Putative oxidoreductase Rv2002 from Mycobacterium tuberculosis, as the closest match to SUZ71503. According to the SCOP and CATH databases, this protein belongs to the α-β class and contains a NAD(P)-binding Rossmann-like domain, consistent with our findings above. The SCOP database also states that this protein belongs to the tyrosine-dependent oxidoreductase family. According to the PDB database, Putative oxidoreductase Rv2002 consists of two chains (A and B) and has a molecular weight of 55.53 kDa. Given that SUZ71503 was a close match to only one of the chains, the value would be consistent with SUZ71503 molecular weight of 26kDa.

The Rossmann fold is found in enzymes of many different functions, including about 20% of known structures (15). It is thought that these structures evolved before the last universal common ancestor of all life on Earth. It was first discovered in nucleotide-binding proteins that utilize NADH cofactors. According to the PDB, the 1NFF protein does use NAD as a cofactor.

The Rossmann fold is made up of a three-layer "sandwich" of a beta-sheet in between two layers of alpha helices. The sequence of the protein is made up of two sets of β-α-β-α-β units. The beta strands associate with each other to form one beta sheet, while the alpha helices connect the beta strands and form the "bread" of the sandwich on either side of it. Between the two sets of β-α-β-α-β units is a cavity which can bind ligands. There is a wide diversity of ligands that are possible for this structure to bind: nucleotides, vitamins, cofactors, amino acids, carbohydrates, and others (15).

The structure of putative oxidoreductase Rv2002 (1NFF) matches the *B. megaterium* glucose 1-dehydrogenase 4 (3AUT) closely, with a sequence identity of 33%. The 3AUT protein catalyzes the conversion of β-d-glucose to d-glucono-1,5-lactone (16). It belongs to the short-chain

dehydrogenase/reductase family (SDR), the same family that is referenced in the 1NFF SCOP2 classification. Because oxidoreductases are so diverse, it is not possible to conclude that the 1NFF protein also has a glucose substrate based solely on the sequence similarity. The directed evolution technique shows that it is likely that the 1NFF protein is involved in steroid metabolism (17).

DATA AVAILABILITY

Uncharacterized protein METZ01_LOCUS24357 [marine metagenome] in the NCBI Protein database (GenBank: SUZ71503.1)
Conserved domains on uncharacterized protein METZ01_LOCUS24357 [marine metagenome], [gi|1782051241|emb|SUZ71503|] (Rossmann-fold NAD(P)-binding domain-containing protein)
Conserved Protein Domain Family in the NCBI Structure Database (3beta-17beta-HSD_like_SDR_c)
PHYRE2 3D Model Prediction of SUZ71503 (PHYRE2 Model)
Crystal structure of Rv2002 gene product from *Mycobacterium tuberculosis* on the PDB database (1NFF)

Web-based Programs:

R Peptides package is available in GitHub (https://github.com/dosorio/Peptides/)
NCBI Structure Database https://www.ncbi.nlm.nih.gov/Structure/index.shtml
InterProScan https://www.ebi.ac.uk/interpro/search/sequence/
PSORTb v3.0.3 (https://www.psort.org/psortb/index.html)
HNN Secondary Prediction Method (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hnn.html)
PHD Secondary Prediction Method (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_phd.html)
NCBI Blastp (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins)
Protein Homology/analogY Recognition Engine V 2.0 (http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index\)
Protein Data Bank https://www.rcsb.org/

ACKNOWLEDGEMENT

We would like to acknowledge Dr. Jonathon Bennett for his time and effort in responding to our questions.

FUNDING

No funding to report.

CONFLICT OF INTEREST

No conflicts of interest to report.

REFERENCES

1. U.S. National Library of Medicine. (n.d.). *Metzyme (ID 480629)*. National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/bioproject/480629
2. Stein, L.Y. and Nicol, G.W. (2018). Nitrification. In eLS, John Wiley & Sons, Ltd (Ed.). https://doi.org/10.1002/9780470015902.a0021154.pub2
3. Quevillon E, Silventoinen V, Pillai S, et al. InterProScan: protein domains identifier. Nucleic Acids Res. 2005;33(Web Server issue):W116-W120. doi:10.1093/nar/gki442
4. Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., & Brinkman, F. S. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics (Oxford, England), 26(13), 1608–1615. https://doi.org/10.1093/bioinformatics/btq249
5. Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. Journal of molecular biology, 202(4), 865–884. https://doi.org/10.1016/0022-2836(88)90564-5
6. Rost, Burkhard; Sander, Chris: Prediction of protein structure at better than 70% accuracy. J. Mol. Biol., 1993, 232, 584-599. Rost, Burkhard; Sander, Chris: Combining evolutionary information and neural networks to predict protein secondary structure. Proteins, 1994, 19, 55-72.
7. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nature protocols, 10(6), 845–858. https://doi.org/10.1038/nprot.2015.053
8. Hanukoglu, I. (2015), Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites. Biochem. Mol. Biol. Educ., 43: 206-209. https://doi.org/10.1002/bmb.20849
9. PSORTB v.3.0: Documentation. (n.d.). Retrieved from https://www.psort.org/documentation/index.html
10. Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., & Sillitoe, I. (2016). Cath: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, *45*(D1). https://doi.org/10.1093/nar/gkw1098
11. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). Cath – a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093–1109. https://doi.org/10.1016/s0969-2126(97)00260-8
12. Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 2020 Jan 8;48(D1):D376-D382. doi: 10.1093/nar/gkz1064. PMID: 31724711; PMCID: PMC7139981.
13. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. 2014 Jan;42(Database issue):D310-4. doi: 10.1093/nar/gkt1242. Epub 2013 Nov 29. Erratum in: Nucleic Acids Res. 2014 Oct;42(18):11847. PMID: 24293656; PMCID: PMC3964979.
14. Cherkasov A, Jones SJ. Structural characterization of genomes by large scale sequence-structure threading. BMC Bioinformatics. 2004 Apr 3;5:37. doi: 10.1186/1471-2105-5-37. PMID: 15061866; PMCID: PMC419331.
15. Medvedev, K. E., Kinch, L. N., Schaeffer, R. D., & Grishin, N. V. (2019). Functional analysis of rossmann-like domains reveals convergent evolution of topology and reaction pathways. *PLOS Computational Biology*, *15*(12). https://doi.org/10.1371/journal.pcbi.1007569
16. Nishioka, T., Yasutake, Y., Nishiya, Y., & Tamura, T. (2012). Structure-guided mutagenesis for the improvement of substrate specificity of*bacillus megaterium*glucose 1-dehydrogenase IV. *FEBS Journal*, *279*(17), 3264–3275. https://doi.org/10.1111/j.1742-4658.2012.08713.x
17. Yang, J. K., Park, M. S., Waldo, G. S., & Suh, S. W. (2003). Directed evolution approach to a structural genomics project: RV2002 from *mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, *100*(2), 455–460. https://doi.org/10.1073/pnas.0137017100