Multi-Template based Homology Modeling of specificity protein 3

[1]Soni Mehul N, [2]Kumar Prasanth S, [3]ShettyShilpa S, [4]Kaid Johar SR., [5]Pandya Himanshu A

[1,2,3]Ph.D Research Scholar, [4]Associate Professor, [5]Professor
[1,2,3]Department of Botany, Bioinformatics, and Climate Change Impacts Management, University School of Science (USSC), Gujarat University, Ahmedabad-380 009, Gujarat, India.
[4]Department of Zoology, University School of Science (USSC), Gujarat University, Ahmedabad-380 009, Gujarat, India.

Abstract: **SP3 Transcription factor contains 81,925 Dalton mass, which member of the Kruppel like zinc finger protein family that is clinically relevant for many neuronal transmission diseases. Considering the functional importance and lack of X-ray crystal structure of SP3 TFs protein, present work was undertaken to build the3D structure of aprotein using homology modeling with a multi-template approach. This present study, we chose three different SP3 templates (PDB ID: 3EBT, 4M9E, and 2WBS) were used for homology modeling. Five models were developedwith the help of multiple sequence alignment respect to templates using Modeller 8.0.0 software. All models were refined and ranked as per their overall DOPE-score. The top-ranked predicted model of SP3 TFs had 93.8% of residues in favored regions as revealed by Ramachandran plot and the ERRAT score was 100% which indicated an accurate model. The results of the homology modeling study and the proposed model can be further used for understanding the structural and functional characteristics of SP3 and to gain more insights to the molecular basis of SP3 inhibition through docking and molecular dynamics simulation studies.**

IndexTerms **- Specificity Protein3 (SP3), Multi-template Homology Modeling, Modeller.**

I. INTRODUCTION

The specificity Protein 3 (SP3) TFs is a complex protein regardingbiological appearance and function. It is a member of Kruppel like zinc finger protein family. The SP3 TFs has a valuable role in regulating synaptic gene expression in neurons. Therefore, it is animportant todrug target for a neuronal transmission disease (Yamakawa et al., 2017, Johar et al., 2014). The homology modeling is computational comprising techniques to predict the three-dimensional protein structure, refer to an unavailable experimental data such as X-ray and NMR (Singh, 2016). This computational technique is also called comparative modeling or template-basedmodeling techniques of protein structure (Fiser et al., 2003). This technique used for an unavailable 3D protein structure, a drug design, protein function analysis, interactions analysis, and improved stability of rational proteins or novel functions. In this technique, to building an atomic resolution model of the query protein sequence (Target Protein) based on an available experimental 3D structure of a homologous protein structure (Template Protein). The quality of homology model structure depends on template structure selection and sequence alignments accuracy. The selection of template protein structure is based on most similarity (>50%) to a target protein sequence, and the sequence alignments are also too less presence of alignment gaps (commonly called indels), refers to missing a region in the template but present in target sequence for building a good homology model (Cavasotto and Phatak, 2009).This technique has commonly four steps: (1) To identify a template protein for a target protein; (2) To develop a sequence alignment between the template and the target protein sequence; (3) To generate a model based on the template protein structure and the alignment; (4) To evaluate and refine the model (Xiang, 2006).

The first article published by Greer J, 1980 on homology modeling approaches then after the improving and using the same concepts have been applied to the prediction of protein structures. Lesk and Chothia,1986analyzed that confirmation of protein structure is more conserved than its amino acids sequence residues, and minor changes in sequence that change the 3D protein structure and its properties. Ring et al., 1993 have identified serine and cysteine proteases inhibitors by homology models techniques. Schafferhans and Klebe, 2001 identified crystal-binding modes of bound ligands using homology models. In 2003, Vangrevelinghe et al. have used homology modeling to identify potent inhibitors from a Novartis collection of four lakhs compounds. In 2003, Tramontano and Morea had been performed a critical assessment of techniques for protein structure prediction (CASP 1-5) and published their results. In 2004-2005, Rimmer et al. have successfully developed homology model of most important target protein family GPCR (G-protein coupled receptor) and identified antagonists of GPCR. In 2000, Rychlewski et al. should improve this approaches for a better model, for instance, to use profile–profile, and in 2004, Ohlson et al. have used HMM–HMM methods, which appear to do best at identifying the template protein folds and turns. Many methods have been developed to taken to identify the best homology model from a set of predictions (Eisenberg et al., 1997; Wallner and Elofsson 2003), and it has apparently been shown in the latest CASP experiments that consensus methods and globally optimization of protein structure (Lundstro et al., 2001) using several studies.Many of programs and algorithms to developed for a homology model validation such as WHAT_CHECK (Hooft et al., 1996), SURVOL (Pontius et al., 1996), PROCHECK (Laskowski et al., 1993), Verify 3D (Luethy et al., 1992), ERRAT (Colovos & Yeates,1993), PROVE (Pontius J et al., 1996).

Now a day, some online and offline tools available to build a homology model by inputting a target protein sequence (Wallner and Elofsson, 2005). ModSeg/ ENCAD has short fragments based copies of template coordinates and bridges gaps that match the

target structure framework (Levitt, 1996). SWISS-MODEL has automated a core model by template backbone atom positions (Schwede et al., 2003). Energy minimization follows NEST tools that build a model by using an artificial evolution algorithm where changes from the template structure such as substitutions, insertions, and deletions are made one at a time and each mutation. This process is repeated until the whole query is modeled(Petrey et al., 2003). Modeller tools that most popular tools. It worked based on a statistical approach to build a homology model using python script (Sali and Blundell, 1993). RaptorX tools that give the better alignments for the most laborious 50 CASP9 target proteins compare to other servers using consensus and refinement methods. I-TASSER tools are the best server according to the 2006–2016 CASP experiments. It builds homology model structure and function prediction using a combination of ab initio folding and threading methods. Prime is commercial tools for Homology modeling. It used the physics-based energy function methods to predict homology model(Jacobson et at., 2004).

In this study, we build a homology model using multiple templates because of the increases the model accuracy as it combines information from multiple template structures. These approaches first introduced by Contreras-Moreira et al.,2003. They proposed that if it were possible to always select the better of two (or more) single-template models, the single-template performance would be superior or at least equal to the multiple-template model.In 2008, Cheng J. developed a novel multi-template algorithm to improve comparative protein modeling. This multiple-template algorithm tries to extract distance (or contact) measure from multiple templates. This measured consistent distance from multiple templates is used to guide homology modeling. In 2012, Yun-Feng XIE et al. has also developed twenty models of the angiotensin II (Ang II) type 1 (AT1) receptor (known as p30556) by multiple templates homology modeling. In 2011, Sokkar P. et al. has developed a three-different 3D-models for AT1 receptor and compared stability, quality using a multi-template homology modeling. In, 2007, Larsson P. et al. have built high-quality models of protein structure using multiple template sequence and proved that to improve quality of homology model compare to automated homology modeling.

II. RESEARCH METHODOLOGY

We build the homology modeling SP3 TFs DNA binding region which known as a Zinc Finger region using the Modeller 9.19 software.

2.1 Retrieving the Target Protein sequence and Templates selection

We retrieved DNA binding domain primary protein sequence from UniProtKB database (UniProt Consortium, 2014) which the accession number P08047 in .fasta format. After, we used Blastp (Altschul et al., 1997) homology search over RCSB PDB database (Bernstein et al., 1977) using expect threshold of 10 and BLOSUM62 scoring matrix (Henikoff and Henikoff, 1992) identified structural templates protein(s). The result of the blast to showed a 50-58% identity of the target protein sequence which was poor for an accurate homology modeling of SP3 Zinc Finger regions. Therefore, we propose a method to select when to use multiple three templates such as PDB ID: 2EBT, 4M9E, and 2WBS were retrieved a .fasta sequence from PDB online database and subjected to multiple sequence alignment for identification of conserved region between the SP3 and templates sequence using EBI T-COFFEE program (Notredame et al., 2000).

2.2 Multiple templates-based modeling

The selected three templates (PDB ID: 2EBT (Solution structure of three tandem repeats of Zinc finger-C_2H_2 domains from human Kruppel-like factor 5), 4M9E (Structure of Klf4 zinc finger DNA binding domain in complex with methylated DNA), and 2WBS (Crystal structure of the zinc finger domain of Klf4 bound to its target DNA)) are downloaded the .pdb structure from the Protein data bank (PDB) (http://www.rcsb.org). They have a helices structure. Now, we used to multiple templates based approached for homology modeling using a Modeller 9.19 software (Shaw et al., 2018). First, we download a script for multiple template-based modeling from the Modeller official site and change them according to our sequence and structure files. We ruined a Salign.py file using a mod9.19 command for constructing a multiple structure alignment of templates. This file developed a pairwise alignment file by dynamic optimization programming using a scoring function which is reliant on of the sequence and structure features. We again ruined an align2d_mult.py (implement by modeller) for setting align_block parameters and without any change of the previous alignment to also set a gap_fuction which use of a structure-dependent gap penalty and using only information of target protein sequence. After, we build the homology model based on multiple template alignment information ruined by a model_mult.py file which pre-build by the modeller. At last, we ruined an evaluate_model.py file to use the DOPE potential to an evaluate the build models.

2.3 Initial model validation

The generated models have three helical regions, and they were then analyzed by web-server SAVES (http://services.mbi.ucla.edu/SAVES/) and RAMPAGE (Anil, 2004). The SAVES is meta server which has six different programs to evaluate and validate the protein structure. It consists procheck, What_check, Errat, verify_3d, prove Ramachandran plot. We input a PDB. File only on the server and its automated run the individual programs can be selected and to evaluate its stereo-chemical quality.

III. RESULTS AND DISCUSSION

The SP3 TFs contains 781 amino acids residues with a mass is 81,925 Dalton. These are the six isomers are exist which are translated from four in-frame AUG-codons and make a difference only in the composition of their N-region which is known as a protein binding region. It has specific bind on promoter region GC-box with specific sequence 5'-GGGCGG-3'. This DNA-binding domain consists of three zinc fingers, which near the C-terminus and serine/threonine- and glutamine-rich domains in their N-terminal regions. The SP3 bind with DNA molecules, to regulate or express many biological functions such as gene expression, transcription regulation. The three Zinc- fingers are present in the SP3 Tfs of human from 621 to 703 positions of residues.

In this study, the sp3 TFs developed a model of zinc fingers of DNA binding region which not available in PDB database and any other. As describe in materials and methods, we searched a similar structure in PDB database using a blast, we got the max identity was 58% of Human Kruppel-Like Factor 5 zinc-fingers NMR solution structure (PDB ID: 2EBT), then after 55% of structure of Klf4 Zinc Finger DNA Binding Domain In Complex With Methylated DNA (PDB ID: 4M9E), and Crystal Structure of The Zinc Finger Domain of Klf4 Bound To Its Target DNA protein (PDB ID: 2WBS), and other with descending order of identity. This identity so was weak for a build an accurate homology model. Therefore, we selected a top three protein structure as a template. Now, we performed the multiple sequence alignment of a three template with zinc-fingers region sequence of SP3 TFs using T-COFFEE tools which showed in fig. 1.

Figure 1: Multiple sequence alignment in a color form using T-COFFEE tools

In this figure 1, to show the consensus sequence between the template and target sequence which needful to build an accurate homology model. We build a homology model followed by a Modeller steps with selected templates. The Modeller generated a five-model based on multi-template with a DOPE score which shows in table 1. The DOPE score to evaluate the accuracy of the model. In this table 1, we evaluate that the homo_model 3 (Fig.2)was DOPE_score: -5640.09 which is good model compare to others.

The validation of homology structure was performed by meta server SAVES, and RAMPAGE serves. In Figure 3 (A), the Ramachandran plot showed mainly the favored region, allowed region, and outlier region which our model showed 93.8%, 6.2%, and 0.0%, respectively. Therefore, we identify our homology model is an accurate and reliable model. The other validation analysis of the predicted model, they Verify_3D plot (Figure 3 (B)) was show that the 100% score of residues which determines the compatibility of an atomic model (3D) with its amino acid sequence (1D) profile. The ERRAT_plot (Figure 3(C)) showed the overall quality factor 60% of our model. The high-quality model was around 85% and higher. The PROVE analysis exposed RMS Z-scores almost equal to 1 for the high-quality model, while in our structure it was 1.553 on figure 3(D). Thus, the predicted our homology model of SP3 Zinc-finger region conformed well to the stereochemistry, indicating that it is a reasonably good quality structure.

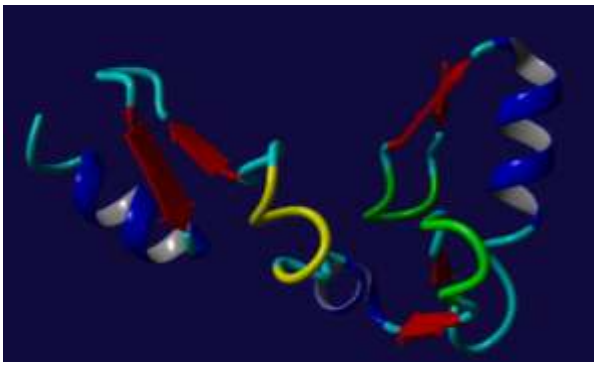

Figure 2: Homology model of Homo_model_3 File build by Modeller9.19 software

Table 1: To show the DOPE score of build Homology model

No.	Filename	DOPE_Score
1	Homo_Model_1	-5718.19
2	Homo_Model_2	-5823.64
3	Homo_Model_3	-5640.09
4	Homo_Model_4	-5880.71
5	Homo_Model_5	-5656.09

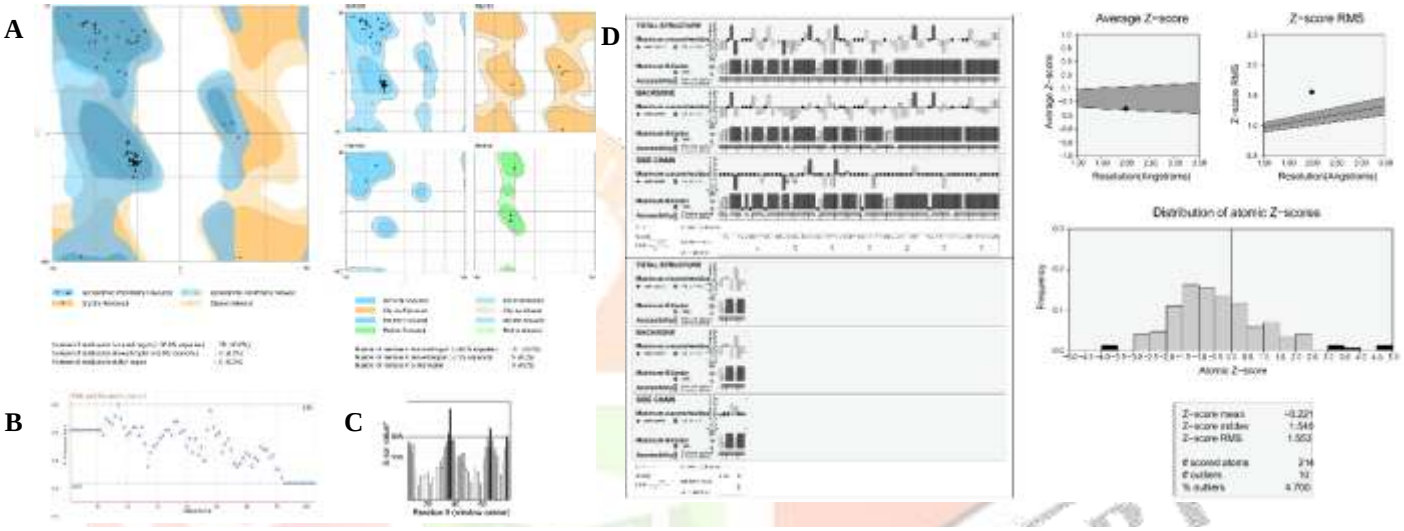

Figure 3:validation Analysis plot. (A) Ramachandran Plot, (B) Verify_3D analysis plot, (C) ERRAT analysis plot, (D) PROVE analysis plot.

IV. Conclusion

The computational *in silico* studies as molecular modeling, computer-aided drug design (CADD), etc. have very helpful to identifythe properties of the molecules like a structure, function, rate of mechanism, potent drug and them similarity molecules, etc. In this study, we investigation was carried out with significant objective to model the SP3 TFs DNA binding domain (zinc-finger region) protein using three different templates and subjecting models so obtained for structural validation using different analysis tools such as Ramachandran plot, Verify3D, PROVE, and ERRAT webserver. The SP3 model based on multi-template such as 2EBT, 4M9E, and 2WBS were found to be reliable regarding stereochemistry with 93.8% residue in the favored region of Ramachandran plot showing high accuracy of model prediction. Z score of 1.553 predicted by PROVE represented the excellent quality of the model. Moreover, the 100% score of Verify3D which determined all residues compatibility for the3D-1D profile. The predicted 3D homology protein structure will provide more insight into understanding the structure and function of the protein. Moreover, this predicted model of SP3 protein can be used for drug-developing or understand probable binding site with an interaction of ligands.

ACKNOWLEDGMENT

Mehul N. Soni and Prasanth Kumar S. acknowledges support fromDepartment of Science and Technology (DST), New Delhi, India, as INSPIRE ResearchFellowship.

REFERENCES

[1] Yamakawa, H., Cheng, J., Penney, J., Gao, F., Rueda, R., Wang, J., Yamakawa, S., Kritskiy, O., Gjoneska, E. and Tsai, L.H., 2017. The transcription factor Sp3 cooperates with HDAC2 to regulate synaptic function and plasticity in neurons. Cell reports, 20(6), pp.1319-1334.

[2] Johar, K., Priya, A. and Wong- Riley, M.T., 2014. Regulation of Na+/K+- ATPase by neuron- specific transcription factor Sp4: implication in the tight coupling of energy production, neuronal activity and energy consumption in neurons. European Journal of Neuroscience, 39(4), pp.566-578.

[3] Fiser, A. and Šali, A., 2003. Modeller: generation and refinement of homology-based protein structure models. Methods in enzymology, 374, pp.461-491.
[4] Singh, N., 2016. Molecular Modelling of Human Multidrug Resistance Protein 5 (ABCC5). Journal of Biophysical Chemistry, 7(03), p.61.
[5] Cavasotto, C.N. and Phatak, S.S., 2009. Homology modeling in drug discovery: current trends and applications. Drug discovery today, 14(13), pp.676-683.
[6] Xiang, Z., 2006. Advances in homology protein structure modeling. Current Protein and Peptide Science, 7(3), pp.217-227
[7] Greer, J., 1980. Model for haptoglobin heavy chain based upon structural homology. Proceedings of the National Academy of Sciences, 77(6), pp.3393-3397.
[8] Chothia, C. and Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. The EMBO journal, 5(4), p.823.
[9] Ring, C.S., Sun, E., McKerrow, J.H., Lee, G.K., Rosenthal, P.J., Kuntz, I.D. and Cohen, F.E., 1993. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. Proceedings of the national academy of Sciences, 90(8), pp.3583-3587.
[10] Schafferhans, A. and Klebe, G., 2001. Docking ligands onto binding site representations derived from proteins built by homology modelling. Journal of molecular biology, 307(1), pp.407-427.
[11] Vangrevelinghe, E., Zimmermann, K., Schoepfer, J., Portmann, R., Fabbro, D. and Furet, P., 2003. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. Journal of medicinal chemistry, 46(13), pp.2656-2662.
[12] Tramontano, A. and Morea, V., 2003. Assessment of homology- based predictions in CASP5. Proteins: Structure, Function, and Bioinformatics, 53(S6), pp.352-368.
[13] Rimmer, C.C., McFarland, K.P., Evers, D.C., Miller, E.K., Aubry, Y., Busby, D. and Taylor, R.J., 2005. Mercury concentrations in Bicknell's thrush and other insectivorous passerines in montane forests of northeastern North America. Ecotoxicology, 14(1), pp.223-240.
[14] Rychlewski, L., Jaroszewski, L., Li, W. and Godzik, A., 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Science, 9(2), pp.232-241.
[15] Ohlson, T., Wallner, B. and Elofsson, A., 2004. Profile–profile methods provide improved fold- recognition: A study of different profile–profile alignment methods. Proteins: Structure, Function, and Bioinformatics, 57(1), pp.188-197.
[16] Eisenberg, D.M., Davis, R.B., Ettner, S.L., Appel, S., Wilkey, S., Van Rompay, M. and Kessler, R.C., 1998. Trends in alternative medicine use in the United States, 1990-1997: results of a follow-up national survey. Jama, 280(18), pp.1569-1575.
[17] Wallner, B. and Elofsson, A., 2003. Can correct protein models be identified?. Protein science, 12(5), pp.1073-1086.
[18] Lundström, J., Rychlewski, L., Bujnicki, J. and Elofsson, A., 2001. Pcons: A neural- network–based consensus predictor that improves fold recognition. Protein Science, 10(11), pp.2354-2362.
[19] Hooft, R.W., Vriend, G., Sander, C. and Abola, E.E., 1996. Errors in protein structures. Nature, 381(6580), pp.272-272.
[20] Pontius, J., Richelle, J. and Wodak, S.J., 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. Journal of molecular biology, 264(1), pp.121-136.
[21] Lüthy, R., Bowie, J.U. and Eisenberg, D., 1992. Assessment of protein models with three-dimensional profiles. Nature, 356(6364), p.83.
[22] Colovos, C. and Yeates, T.O., 1993. Verification of protein structures: patterns of nonbonded atomic interactions. Protein science, 2(9), pp.1511-1519.
[23] Pontius, J., Richelle, J. and Wodak, S.J., 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. Journal of molecular biology, 264(1), pp.121-136.
[24] Wallner, B. and Elofsson, A., 2005. All are not equal: a benchmark of different homology modeling programs. Protein Science, 14(5), pp.1315-1327.
[25] Levitt, M., 1992. Accurate modeling of protein conformation by automatic segment matching. Journal of molecular biology, 226(2), pp.507-533.
[26] Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C., 2003. SWISS-MODEL: an automated protein homology-modeling server. Nucleic acids research, 31(13), pp.3381-3385.
[27] Petrey, D., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith- Fischman, S., Kernytsky, A., Schlessinger, A. and Koh, I.Y., 2003. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins: Structure, Function, and Bioinformatics, 53(S6), pp.430-435.
[28] Šali, A. and Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology, 234(3), pp.779-815.
[29] Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J., Honig, B., Shaw, D.E. and Friesner, R.A., 2004. A hierarchical approach to all- atom protein loop prediction. Proteins: Structure, Function, and Bioinformatics, 55(2), pp.351-367.
[30] Contreras-Moreira, B., Fitzjohn, P.W. and Bates, P.A., 2003. In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. Journal of molecular biology, 328(3), pp.593-608.
[31] Cheng, J., 2008. A multi-template combination algorithm for protein comparative modeling. BMC structural biology, 8(1), p.18.
[32] Xie, Y.F., Li, X.W., Wang, J.F., Christakos, G., Hu, M.G., An, L.H. and Li, F.S., 2012. Spatial estimation of antibiotic residues in surface soils in a typical intensive vegetable cultivation area in China. Science of the Total Environment, 430, pp.126-131.

[33] Sokkar, P., Mohandass, S. and Ramachandran, M., 2011. Multiple templates-based homology modeling enhances structure quality of AT1 receptor: validation by molecular dynamics and antagonist docking. Journal of molecular modeling, 17(7), pp.1565-1577.
[34] Larsson, P., Wallner, B., Lindahl, E. and Elofsson, A., 2008. Using multiple templates to improve quality of homology models in automated homology modeling. Protein Science, 17(6), pp.990-1002.
[35] UniProt Consortium, 2014. UniProt: a hub for protein information. Nucleic acids research, p.gku989.
[36] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), pp.3389-3402.
[37] Henikoff, S. and Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences, 89(22), pp.10915-10919.
[38] Notredame, C., Higgins, D.G. and Heringa, J., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology, 302(1), pp.205-217.
[39] Shaw, D.R. and Dussan, J., 2018. Transcriptional analysis and molecular dynamics simulations reveal the mechanism of toxic metals removal and efflux pumps in Lysinibacillus sphaericus OT4b. 31. International Biodeterioration & Biodegradation, 127, pp.46-61.
[40] Anil, B., Song, B., Tang, Y. and Raleigh, D.P., 2004. Exploiting the right side of the Ramachandran plot: Substitution of glycines by D-alanine can significantly increase protein stability. Journal of the American Chemical Society, 126(41), pp.13194-13195.