In silico Structure Prediction and Functional Annotation of *Ananas comosus* Hypothetical Protein OAY63476.1

Zainab Bibi[1], Aqsa Khalid[1*], Iqra Iftikhar[1], Muhammad Rizwan[1], Azhar Mehmood[1], Sajid Khan[1], Anum Munir[1, 2]

[1]*Department of Bioinformatics*
Government Postgraduate College Mandian
Abbottabad 22010, Pakistan
E-mails: zainabtanoli1996@gmail.com, aqsa.khalid30@yahoo.com, iqraiftikhar081@gmail.com, mrizwanhu@gmail.com, mabbasi71@gmail.com, mrsajidk@gmail.com

[2]*Department of Bioinformatics and Biosciences*
Capital University of Science and Technology
Islamabad, Pakistan
E-mail: anummuir786@yahoo.com

[*]*Corresponding author*

Received: December 10, 2018 **Accepted: December 16, 2019**

Published: December 31, 2020

Abstract: *Hypothetical proteins (HPs) are those whose sequence is present but nothing is known about their structural and functional annotations. Elucidation of structural and functional insights would be effective enough in perceiving the protein interactions and their involvement in various pathways and networks. Ananas comosus is considered as third most essential tropical fruit and is given more consideration from the commercial point of view; therefore, in this research, it is taken for analysis using different approaches of bioinformatics. On the basis of subcellular localization and secondary structure analysis, it was suggested that HP is the nuclear protein consisting of α helices and more abundant coils. Homology modeling has been done through Swiss Model to determine their template structure but template identity score is not enough which uncover the fact that the protein OAY63476.1 is unpredicted in-vitro. In order to predict the 3D structure of protein, Phyre 2 server is utilized. Results are validated by different strategies revealing the stability of the developed model. The quality factor of this protein is 45.28. Functional analysis and domain identification have been performed by using NCBI-CDD and Pfam which suggested that protein contains ring variant domain. Comparative genome analysis has been performed with other proteins of plants which demonstrate that the selected protein has the highest similarity. This study paves the way in order to determine the structural and functional annotations of other uncharacterized proteins and effective in exploring other novel proteins along with their functions in the same way as that of this research study.*

Keywords: *Ananas comosus, Homology modeling, Ab initio method, Functional annotation.*

Introduction

Those proteins which are predicted from the nucleic acid sequences and the protein sequences but their function is still unknown are called hypothetical proteins (HPs) [3]. Generally, they cover round about half of the protein-coding regions in many genomes [17]. In order to clearly understand about the biology and the genome of the organisms, it is important to uncover the functions of the HPs [16]. Structural and the functional annotations of particular HPs may result in the identification of new structures along with the others aiding in introducing new protein pathways and cascades [15].

Elucidation of structural and functional insights would be significant in perceiving the PPI interactions and their association in different pathways [15, 17]. It would also help in analyzing new conformational orientations and to evaluate new domains and motifs. These domains will serve as a pharmacological target [19]. Prediction of HPs also helps in the identification of the novel targets for screening and drug discovery procedures [3].

Pineapple (*Ananas comosus*) is a perennial herbaceous fruit which belongs to *Bromeliaceae* family. From the commercial point of view, *Ananas comusus* is considered to be the third most essential tropical fruit and it is cultivated in tropical and subtropical regions after banana and mango [19, 26]. It is of the most valuable crop possessing crassulacean acid metabolism and it is most valuable tropical fruit. The genome of pineapple possesses 38 reputed genes that are required for the carbon fixation module of CAM [13, 14].

Presently, variety of hypothetical proteins has been found in the genome of various organisms. On account of some restrictions including the cost and time for the innovative approaches, overall genomes annotation has not been attained so far [15]. *In silico* strategies in order to explain HPs are cost-effective and less time-consuming in order to explore their functionality [16]. Computational approaches involving several databases and various algorithms to determine protein functions are effective alternative approach than the laboratory methods [17, 19].

A multitude of computational methods have been developed for protein functional annotations varying from template-based approaches in which a template having a known structure and function is used in order to predict the function of the sequence in the query [9].

In this study *Ananas comusus*, an uncharacterized protein was selected. The primary sequence of the above-mentioned protein was available however the structural details were not accessible. Therefore, this study intended to examine the physiochemical and the structural attributes in order to produce the first 3D model of the HP, through *ab initio* methods, and then to execute the functional annotations.

Materials and methods

Selection of hypothetical protein

Hypothetical proteins were retrieved from the protein database of NCBI (https://www.ncbi.nlm.nih.gov/) using the keyword "hypothetical proteins" and uncharacterized proteins of *Ananas comosus* [22] was selected. Blast analysis was performed to check the similarity score with other reported hypothetical proteins.

Sequence retrieval

The amino acid sequence of *Ananas comosus* hypothetical protein was retrieved from UniProt database (http://www.uniprot.org/) in FASTA format. UniProt is a database of non-redundant protein sequences. It is a freely accessible database containing a large amount of information about biological functions of proteins [1].

Physiochemical analysis

The physiochemical analysis was performed by checking out various characteristic of protein involving molecular weight, amino acid composition, composition at the atomic level, instability index along with hydrophobicity (GRAVY). For this purpose ProtParam tool (https://web.expasy.org/protparam/) was used to perform a theoretical evaluation of

physiochemical characteristic. ProtParam is a tool that allows computation of various physical and chemical parameters obtained from protein [2, 17].

Secondary structure analysis

In order to predict the secondary structure (alpha helices and beta sheets) of uncharacterized protein server SOPMA [10] was used. Alternatively PSIPREDv3.3 [8] (http://bioinf.cs.ucl.ac.uk/psipred/) and Predict Protein (https://www.predictprotein.org/) servers were also employed in order to validate the result that was achieved from SOPMA server [15, 23].

Sub-cellular localization

Subcellular localization of this protein was predicted by Plant-mPLOc [6] (http://www.csbio.sjtu. edu.cn/ bioinf/plant-multi/). Plant-mPLOc is a server for the prediction of sub-cellular localization of proteins which aids in revealing their functions as well as interactions [4, 5]. The results of subcellular localization obtained by SOSUI [11] were cross checked with the Predict Protein servers for the confirmation of results [23].

Homology modeling

The possible 3D structure of the protein was determined through protein structure homology server SWISS-MODEL (https://swissmodel.expasy.org) by entering the amino acid sequence in FASTA format. SWISS-MODEL is a structural bioinformatics web server. Its main purpose is homology modeling of protein 3D structure based on the knowledge of templates [18].

Structure prediction

The 3D structure of hypothetical protein was predicted through Phyre 2 server. A Phyre 2 server predicts the 3D structure of protein sequence relying on the techniques of homology modeling [12]. In addition, the I-TASSER [24, 25] was used to cross check the results obtained from Phyre 2 server.

Quality assessment and structure validation

The generated model was checked for recognition of error in the 3D structure by ERRAT and Verify 3D program in verification server SAVES [7]. It regulates the compatibility of the model with its amino acid sequences based on alpha helices, beta sheets, loops and coils [20]. The quality and compatibility of the model were checked by structural assessment method, i.e. Ramachandran plots. The reliability of model was observed through Ramachandran plot generated by RAMPAGE server (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) to determine the percentage of residues in allowed, favored and outlier region [21].

Functional annotation

In order to determine conserved domains for Hypothetical protein, NCBI conserved domain database was accessed. Pfam database has also been utilized for cross-checking the results that have been computed from Conserved Domain Database. For the prediction of similarity score of the protein (comparative genome analysis) with other uncharacterized proteins of plants, comparative genome analysis was performed by using Blast program.

Submission of model in protein model database

The predicted model of protein OAY63476.1 of *Ananas comosus* has been submitted in protein model database (http://bioinformatics.cineca.it/PMDB/) with id PM0081523.

Results and discussion

Physiochemical characteristics of OAY63476.1

Protparam tool was used to analyze the physiochemical properties of all the hypothetical proteins. Protein was predicted to be comprised of 266 amino acids having molecular weight 27893.84 Dalton and isoelectric point (PI) of 6.33 which indicates negative charge protein. The instability index of the protein was computed to be 57.56 which reveal that it is unstable. The GRAVY index of protein OAY63476.1 is -0.024 which indicates hydrophilic and soluble protein. A low value of GRAVY index indicates better interaction and having higher interaction with water.

The most abundant amino acid residue was observed in Glycine (12.4%) followed by Serine (10.2%) and the lowest amino acid was observed to be Methionine (0.8%). The sequence has 27 negatively charged residues (Aspartic Acid, Glutamic Acid) and a total number of positively charged residues (Arginine, Lysine) are 26.

The molecular formula of protein was found to be $C_{1216}H_{1958}N_{352}0_{373}S_{13}$ and the total number of the atoms was 3912. The extinction coefficient was 32595 having all Cysteine residues from Cys measured in the unit of $M^{-1}cm^{-1}$ at 280 nM in water. The calculated aliphatic index was 90.90.

Subcellular localization

Subcellular localization of hypothetical protein provides information about their cellular function. The subcellular localization of query protein was characterized to be nuclear protein observed by Plant-mPLoc. The results were validated by SOSUI and Predict Protein server.

Secondary structure analysis

For the prediction of secondary structure, SOPMA server was used. The most abundant amongst them was random coil, i.e. 54.14%, alpha helix was found to be 28.20% and the extended strand was 15.04% whereas, beta turns was found as 2.63% as shown in Fig. 1. The results were validated by PSIPRED and are shown in Fig. 2.

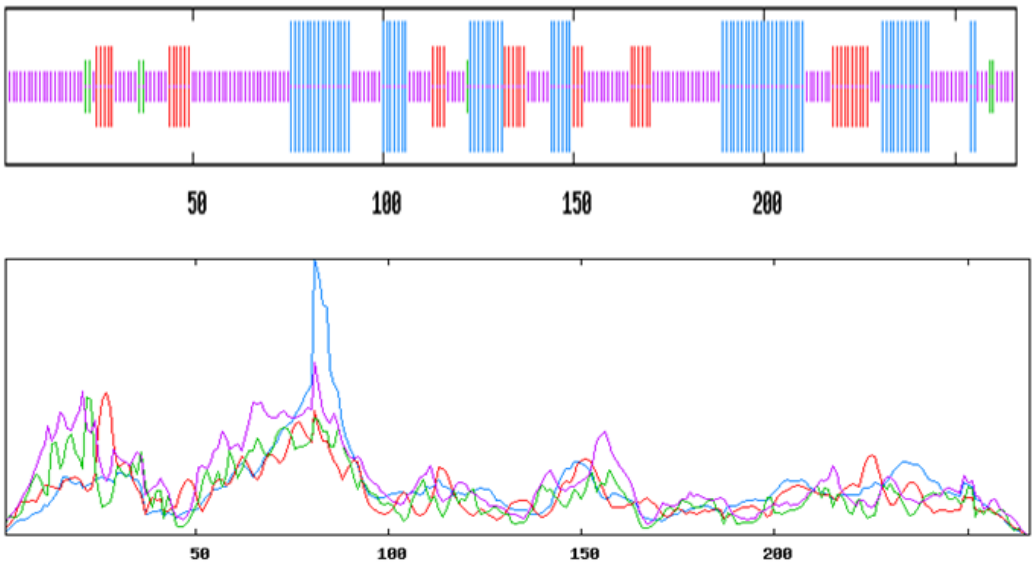

Fig. 1 Predicted secondary structure of *Ananas comosus* hypothetical protein OAY63476.1 by SOPMA

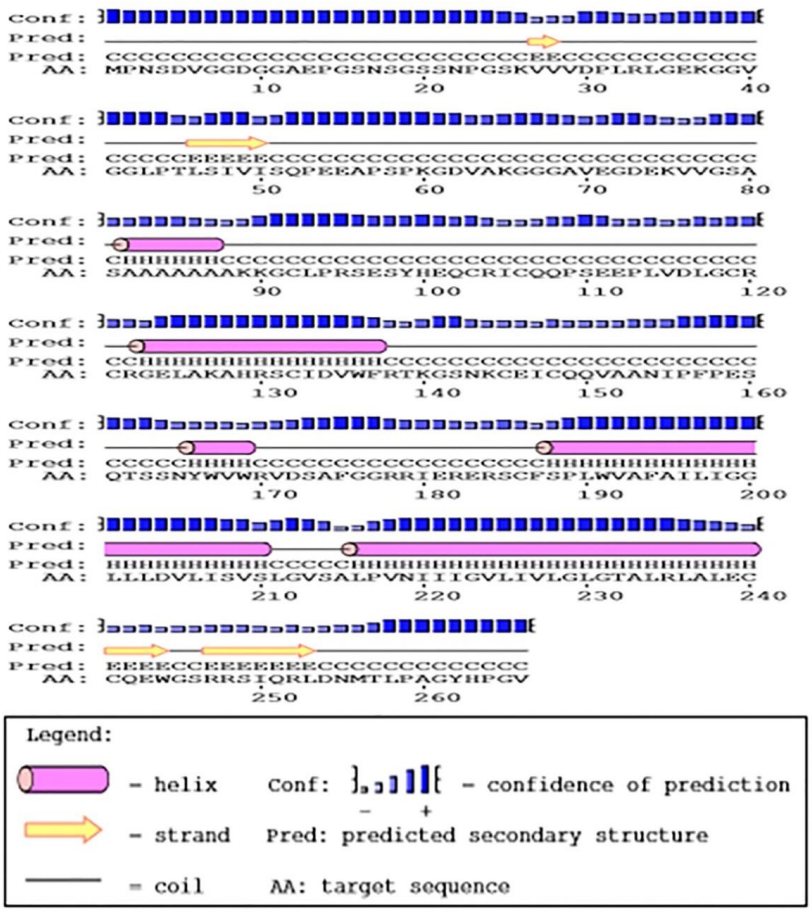

Fig. 2 Validation of predicted secondary structure of *Ananas comosus* hypothetical protein OAY63476.1 by PSIPRED server

Homology modeling

In silico analysis of hypothetical proteins aids in the determination of their biological functions. This can be performed for predicting the 3D structure of the targeted protein. In order to perform the homology modeling, the sequence was given to the server, where it performs Blast for the query sequence to identify template. The highest template identity score was found to be 28% which reveals the novelty of protein because there was no similar template structure present in any database.

3D structure prediction

The 3D Structure of a protein was predicted by an *ab initio* method with the help of Phyre 2. It gives a confidence level of 99.4% and results were validated through I-Tasser. Visualization of the 3D model was performed through Discovery Studio is shown in Fig. 3.

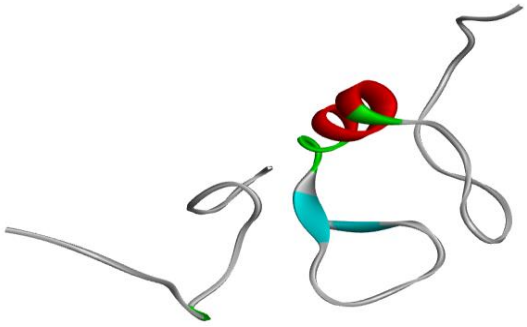

Fig. 3 Structural analysis of *Ananas comosus* hypothetical protein OAY63476.1

Quality assessment and structure validation

Reliability and accuracy of the predicted model were checked by ERRAT, which analyzed the non-bonded interaction between different types of atom relying on the atomic interaction. The quality factor was found to be 45.28 which is good enough to implement this protein in drug designing and target identification. Whereas value obtained from VERIFY 3D was 17.33% has an average 3D-1D score greater than equal to 0.2 means that structure is compatible and genuinely good. The stereo-chemical quality of the model was determined through Ramachandran plot using RAMPAGE server. it was observed that 69% of the residues are in the favored region, 22.5% is in the allowed region and 8.5% residues is in outlier region which shows reliability and efficiency of the model. The Ramachandran plot is shown in Fig. 4.

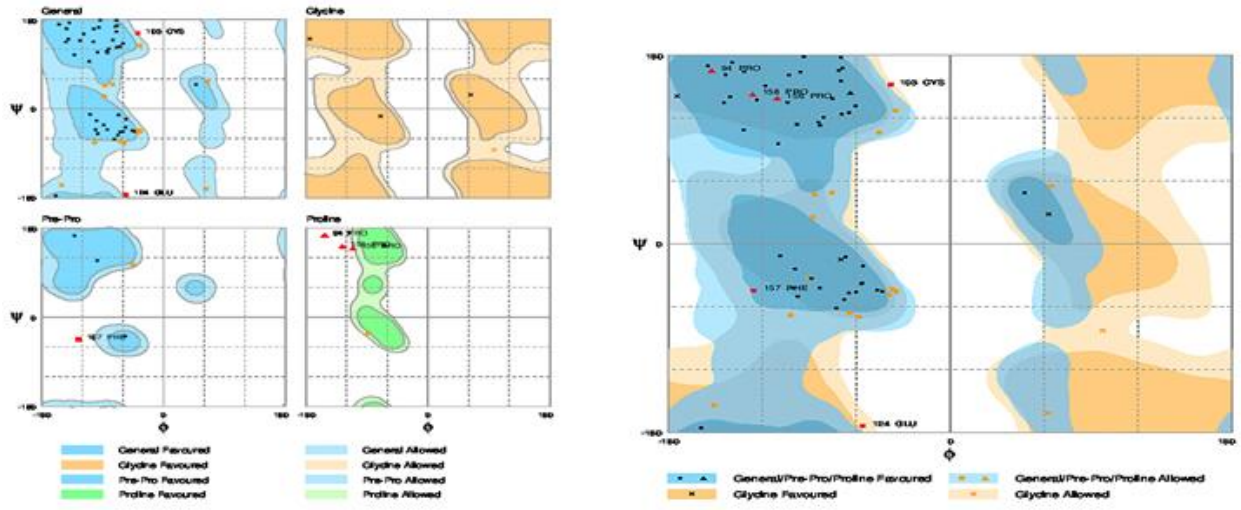

Fig. 4 Ramachandran plot for the 3D model of studied hypothetical protein OAY63476.1 by RAMPAGE server

Functional annotation

Results obtained from different databases revealed that the protein contains RING-variant domain which is considered as C4HC3 zinc finger involved in various cellular processes. Ring domains serve for binding with the ubiquitination enzymes along with their substrates and therefore possess ligase activity, i.e. binding function. Similar results have been obtained from Pfam database and results were significant, i.e. Pfam A match with the query sequence.

Comparative genome analysis

Results of Blast search uncover the fact that the hypothetical protein has the highest similarity count with another uncharacterized protein plant.

Conclusion

This study was intended to create first 3D model of uncharacterized protein OAY63476.1 of *Ananas comosus*. 3D model of protein was created by using computational method, i.e. *ab initio* method and validated through different quality assessment method. Validation results reflect the accuracy of our protein. Results revealed that protein contain ring variant domain found to be associated with various cellular processes. In future, these results will uncover further mechanism and this could be effective enough in identifying others novel proteins in a same way as we have adopted for OAY63476.1.

Acknowledgements

Authors greatly acknowledge the support provided by the Department of Bioinformatics Government Post Graduate College Mandian for conducting this research work. Authors are thankful to Professor Ghulam Rasool, Principal at Government Post Graduate College Mandian Abbottabad for encouragement and inspiration until the success and completion of this research work.

References

1. Apweiler R. (2004). UniProt: The Universal Protein Knowledgebase, Nucleic Acids Research, 32(90001), D115-D119.
2. Artimo P., M. Jonnalagedda, K. Arnold, D. Baratin, G. Csardi, E. de Castro, H. Stockinger (2012). ExPASy: SIB Bioinformatics Resource Portal, Nucleic Acids Research, 40(W1), W597-W603.
3. Bharat S. V. P., Y. B. Adimulam, S. Kodukula (2015). *In silico* Functional Annotation of a Hypothetical Protein from *Staphylococcus aureus*, Journal of Infection and Public Health, 8(6), 526-532.
4. Chou K.-C., H.-B. Shen (2007). Large-scale Plant Protein Subcellular Location Prediction, Journal of Cellular Biochemistry, 100(3), 665-678.
5. Chou K.-C., H.-B. Shen (2010). Cell-PLoc 2.0: An Improved Package of Web-servers for Predicting Subcellular Localization of Proteins in Various Organisms, Natural Science, 02(10), 1090-1103.
6. Chou K.-C., H.-B. Shen (2010). Plant-mPLoc: A Top-down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization, PLoS ONE, 5(6), e11335, https://doi.org/10.1371/journal.pone.0011335.
7. Colovos C., T. O. Yeates (1993). Verification of Protein Structures: Patterns of Nonbonded Atomic Interactions, Protein Science, 2(9), 1511-1519.
8. Cuff J. A., G. J. Barton (2000). Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction, Proteins: Structure, Function, and Bioinformatics, 40(3), 502-511.
9. Dorden S., P. Mahadevan (2015). Functional Prediction of Hypothetical Proteins in Human Adenoviruses, Bioinformation, 11(10), 466.
10. Geourjon C., G. Deleage (1995). SOPMA: Significant Improvements in Protein Secondary Structure Prediction by Consensus Prediction from Multiple Alignments, Bioinformatics, 11(6), 681-684.
11. Hirokawa T., S. Boon-Chieng, S. Mitaku (1998). SOSUI: Classification and Secondary Structure Prediction System for Membrane Proteins, Bioinformatics (Oxford, England), 14(4), 378-379.
12. Kelley L. A., S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg (2015). The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis, Nature Protocols, 10(6), 845-858.
13. Ming R., R. VanBuren, C. M. Wai, H. Tang, M. C. Schatz, et al. (2015). The Pineapple Genome and the Evolution of CAM Photosynthesis, Nature Genetics, 47(12), 1435-1442.
14. Moyle R. L., M. L. Crowe, J. Ripi-Koia, D. J. Fairbairn, J. R. Botella (2005). PineappleDB: An Online Pineapple Bioinformatics Resource, BMC Plant Biology, 5, 21, https://doi.org/10.1186/1471-2229-5-21.
15. Munir A., A. Mehmood, S. Azam (2016). Structural and Function Prediction of *Musa acuminata* subsp. *Malaccensis* Protein, International Journal Bioautomation, 20(1), 19-30.
16. Naveed M., S. Tehreem, M. Usman, Z. Chaudhry, G. Abbas (2017). Structural and Functional Annotation of Hypothetical Proteins of Human Adenovirus: Prioritizing the

Novel Drug Targets, BMC Research Notes, 10(1), https://doi.org/10.1186/s13104-017-2992-z.
17. Paul S., M. Saha, N. C. Bhoumik, S. N. Talukdar (2015). *In silico* Structural and Functional Annotation of *Mycoplasma genitalium* Hypothetical Protein MG_377, International Journal Bioautmation, 19(1), 15-24.
18. Schwede T. (2003). SWISS-MODEL: An Automated Protein Homology-modeling Server, Nucleic Acids Research, 31(13), 3381-3385.
19. Singh A., R. Chaube (2014). Bioinformatic Analysis, Structure Modeling and Active Site Prediction of Aquaporin Protein from Catfish *Heteropneustes fossilis*, https://www.researchgate.net/publication/269094386_Bioinformatic_Analysis_Structure_Modeling_and_Active_Site_Prediction_of_Aquaporin_Protein_from_Catfish_Heteropneustes_fossilis.
20. Taylor C. C., K. V. Mardia, M. Di Marzio, A. Panzera (2012). Validating Protein Structure Using Kernel Density Estimates, Journal of Applied Statistics, 39(11), 2379-2388.
21. Wang W., M. Xia, J. Chen, F. Deng, R. Yuan, X. Zhang, F. Shen (2016). Genome-wide Analysis of Superoxide Dismutase Gene Family in *Gossypium raimondii* and *G. arboreum*, Plant Gene, 6, 18-29.
22. Wheeler D. L. (2004). Database Resources of the National Center for Biotechnology Information: Update, Nucleic Acids Research, 32(90001), D35-D40.
23. Yachdav G., E. Kloppmann, L. Kajan, M. Hecht, T. Goldberg, T. Hamp, et al. (2014). PredictProtein – An Open Resource for Online Prediction of Protein Structural and Functional Features, Nucleic Acids Research, 42(W1), W337-W343.
24. Yang J., R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang (2015). The I-TASSER Suite: Protein Structure and Function Prediction, Nature Methods, 12(1), 7.
25. Yang J., Y. Zhang (2015). Protein Structure and Function Prediction Using I-TASSER: Protein Structure and Function Prediction Using I-TASSER, Current Protocols in Bioinformatics, Bateman A., W. R. Pearson, L. D. Stein, G. D. Stormo, J. R. Yates, Eds., Hoboken, New Jersey, USA, John Wiley & Sons, Inc., 5.8.1-5.8.15, http://doi.wiley.com/10.1002/0471250953.bi0508s52.
26. Zhou L., T. Matsumoto, H.-W. Tan, L. W. Meinhardt, S. Mischke, B. Wang, D. Zhang (2015). Developing Single Nucleotide Polymorphism Markers for the Identification of Pineapple (*Ananas comosus*) Germplasm, Horticulture Research, 2, 15056, https://doi.org/10.1038/hortres.2015.56.

Zainab Bibi
E-mail: zainabtanoli1996@gmail.com

Ms. Zainab Bibi is a student of Bachelor of Science in Bioinformatics at Government Post Graduate College Mandian (GPGCM), Abbottabad, Pakistan. Her areas of expertise are basic biology, cell biology, proteomics, genomics, biotechnology, system biology and botany. She has command over many bioinformatics tools and software used in drug design and in other activities related to bioinformatics. Her major interest is in drug discovery, drug design and phylogenetic analysis. Zainab Bibi is also participating in many ongoing research projects. Her ongoing project is network based functional enrichment of genes in multiple pathways.

Aqsa Khalid
E-mail: aqsa.khalid30@yahoo.com

Ms. Aqsa Khalid is a B.Sc. student in Bioinformatics at Government Post Graduate College Mandian (GPGCM), Abbottabad, Pakistan. She has completed her intermediate from Army Burn Hall College (ABHC), Abbottabad, Pakistan. Aqsa Khalid has presented her research work on drug design at national and international conferences in Pakistan. She has already worked on drug design, TSA identification, pharmacophore modeling and drug sensitivity analysis. Her ongoing project is on model development using system biology approaches. Her research interest areas are in drug design, system biology, protein structure prediction and model development.

Iqra Iftikhar
E-mail: iqraiftikhar081@gmail.com

Ms. Iqra Iftikhar is a B.Sc. student in Bioinformatics at GPGCM, Abbottabad, Pakistan. Her areas of expertise are genomics, proteomics, system biology and biotechnology. Her ongoing project is on functional enrichment analysis of genes involved in multiple pathways. Her research interest areas are in drug design, system biology, pharmacophore modeling and protein structure prediction. She has presented her research work on drug design at national and international conferences in Pakistan.

Muhammad Rizwan, M.Sc.
E-mail: mrizwanhu@gmail.com

Mr. Muhammad Rizwan has done his M.Sc. in Bioinformatics from Hazara University Mansehra Abbottabad, Pakistan. He is a lecturer of Bioinformatics at GPGCM, Abbottabad, Pakistan. His major research interests are in drug discovery, drug design, translational research and phylogenetic analysis. He has expertise in drug design, network and pathways analysis.

Prof. Azhar Mehmood, Ph.D.
E-mail: mabbasi71@gmail.com

Dr. Azhar Mehmood has done his Ph.D. in Phyto-Sociology. He has a M. Phil. in Tissue Culture. He is working as a Professor and he is a Head of Department of Bioinformatics at GPGCM, Abbottabad, Pakistan. Azhar Mehmood is also a team member of Bioinformatics Research Club in GPGCM. Prof. Azhar Mehmood has a teaching experience of 21 years and his major interests are in phyto-sociology, genetics, botany and tissue culture.

Sajid Khan, M.Sc.
E-mail: mrsajidk@gmail.com

Mr. Sajid Khan has done his M.Sc. in Bioinformatics at Mohammad Ali Jinnah University (MAJU), Islamabad, Pakistan. He is a lecturer of Bioinformatics at GPGCM, Abbottabad, Pakistan. His major interest is in drug discovery, drug design, phylogenetic analysis, networks and pathways analysis. He has completed many research projects. Sajid Khan also has command on various programming languages such as R, MATLAB, Java and C++.

Anum Munir, M.Sc. Student
E-mail: anummunir786@yahoo.com

Ms. Anum Munir is doing M.Sc. in Bioinformatics from Department of Bioinformatics and Biosciences in Capital University of Science and Technology, Islamabad, Pakistan. She is a lecturer of Bioinformatics at GPGCM, Abbottabad, Pakistan. She has expertise in drug designing, pharmacophore modeling, vaccine designing, drug repositioning and repurposing, mathematical modeling and simulation, system biology, NGS data analysis, PBPK modeling, PK/PD modeling, Pathways analysis and many other bioinformatics related research areas.