Sequence coverage and model completion using template-based protein structure prediction with and without *de novo* structure prediction protocols.

[1, 2, §]Rasheeda Sheik, [1, 2, §]Harshitha Vangalapudi and [1, 3]Ravikiran S. Yedidi*

[1]The Center for Advanced-Applied Biological Sciences & Entrepreneurship (TCABS-E), Rajahmundry, AP; [2]Department of Human Genetics, [3]Department of Zoology, Andhra University, Visakhapatnam, AP. ([§]These authors contributed equally) (*Correspondence to RSY: tcabse.india@gmail.com)

Protein 3-dimensional structure prediction is very critical in most of the biomedical research scenarios in order to understand various biological processes at the molecular level. However, the challenge has been the prediction of the accurate 3D models using various model prediction servers such as the SWISS MODEL server (builds 3D models using existing templates), ROSETTA server (builds 3D models using templates as well as *de novo* prediction wherever necessary), etc. In this study, two protein FASTA sequences were taken and were used for 3D model building with both servers. Full length 3D models were predicted by both servers for cystic fibrosis transmembrane conductance regulator (CFTR) protein due to the availability of full length templates. The SWISS MODEL server was unable to predict the full length 3D model for homeobox-B13 (Hox-B13) protein but the ROSETTA server was able to predict the full length structure of Hox-B13 protein. Our findings suggest that in the absence of a full length template, it is better to use the servers with *de novo* structure prediction capability such as the ROSETTA server. We conclude that in the presence of a full length template, no major structural deviations were observed between the predicted models of CFTR by both servers.

Keywords: Protein structure prediction, SWISS MODEL, ROSETTA, ROBETTA, computational modeling.

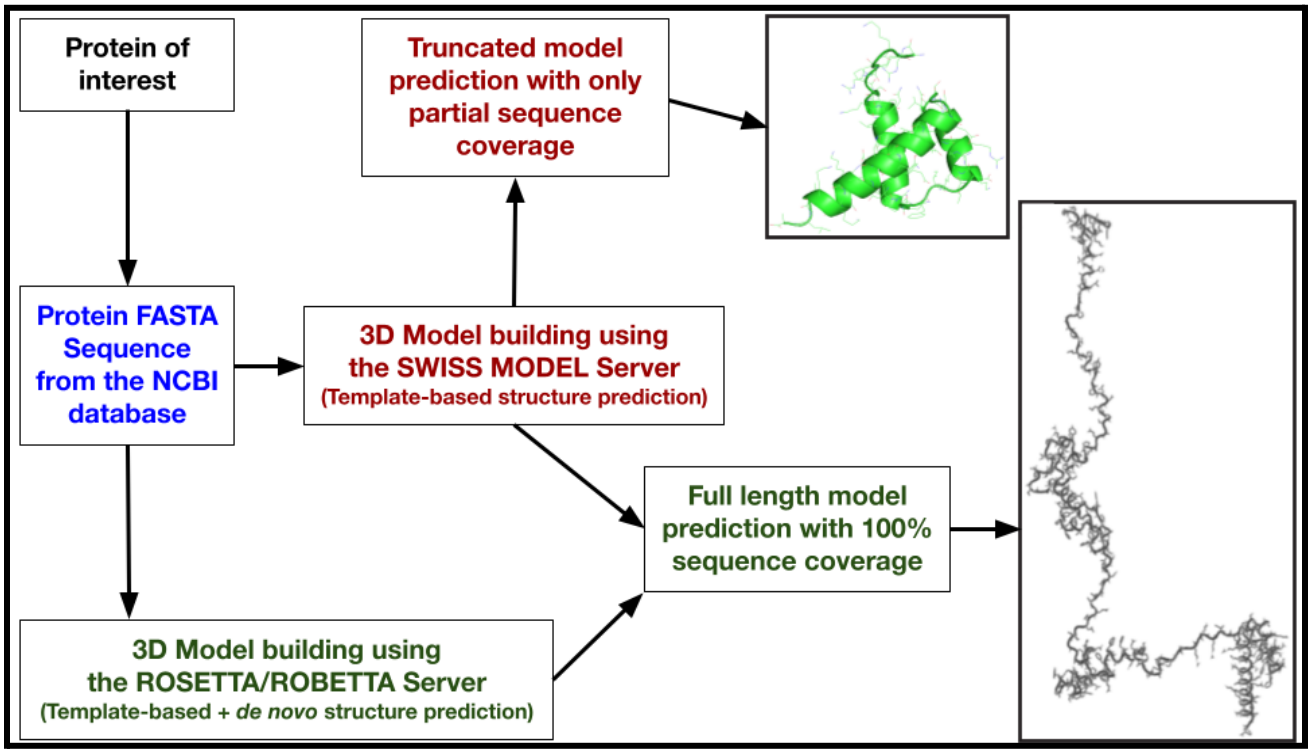

Figure 1. Overview of protein structure prediction performed in this study.

Protein homology modeling has become an integral tool for most of the biomedical research these days in spite of the rapid structure determination using various structural biology techniques. There are several online servers that can predict the 3-dimensional structures of various protein sequences using different algorithms. Some of these algorithms are computationally intensive that require cluster computing and are beyond the capabilities of a typical startup laboratory. Hence, server-based calculations became more popular where one can submit the job online and have the results delivered to the email inbox conveniently without building servers. Recently we have used the SWISS MODEL server [1] to predict the 3D models of several mutants of the SARS-CoV-2 spike protein in order to understand the mutation induced conformational changes and related consequences [2].

The accuracy in the model prediction is very important because the function of a protein is highly dependent on its 3D structure. Homology-based structure prediction servers such as the SWISS MODEL [3-19] often accompany several quality checking parameters such as the Ramachandran plots, etc. in order to evaluate the quality of the predicted structures with respect to the overall stereochemistry, amino acid side chain orientations, etc. However, sometimes, due to lack of a full length template, the predicted model could be partial that may or may not explain the biological function in detail. In such cases, *de novo* prediction is needed. In this study, we built the homology models of two proteins, the cystic fibrosis transmembrane conductance regulator (CFTR) protein and the homeobox-B13 (Hox-B13) protein (Figure 1). We used the SWISS MODEL server as well as the ROSETTA server to build the models and analyzed all four models for their completeness and overall quality of the 3D structure predicted by the servers. Full length models were built for the CFTR protein sequence by both the servers but not for the Hox-B13 protein.

Figure 2. Predicted 3D model of the CFTR protein.

Results & Discussion: As shown in Figure 2, the 3D full length predicted model of the CFTR protein looks the same irrespective of the server used for the prediction of the model. This is due to the availability of a full length template for this protein. The models were built for both the wild type and the ΔF508 mutant variants of CFTR protein as shown in Figure 3.

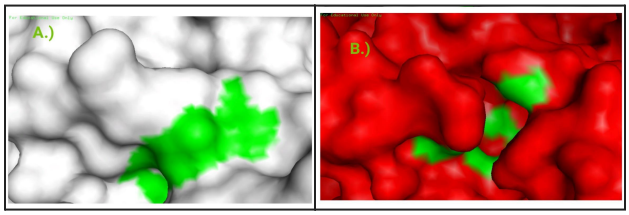

Figure 3. Wild type (panel A) vs. ΔF508 mutant (panel B) of the CFRT protein.

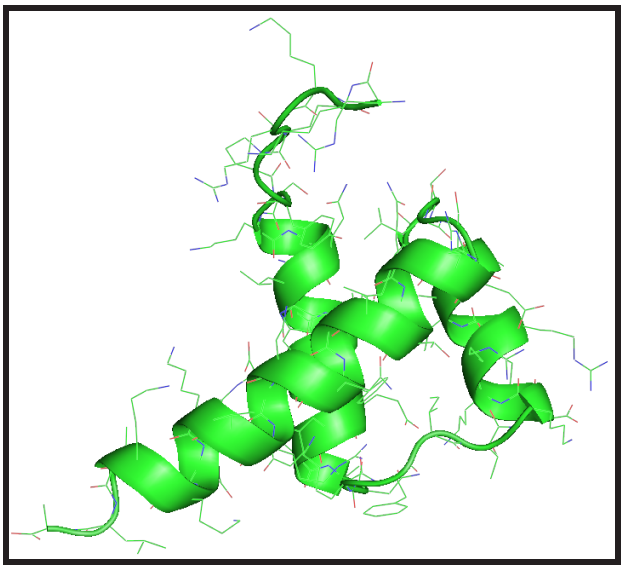

Figure 4. Partial model of Hox-B13 protein.

As shown in Figure 4, the SWISS MODEL server was unable to build the full length model for the Hox-B13 protein but yielded a partial model instead. However, when the same sequence was fed into the ROSETTA server, the full length model was built for the Hox-B13 protein (Figure 5).

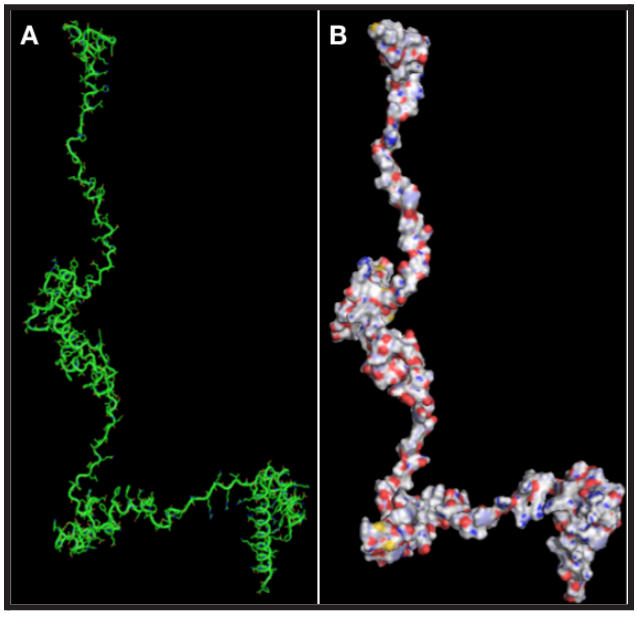

Figure 5. Full length model of the Hox-B13 shown as a cartoon model (panel A) and surface diagram (panel B). This model was obtained using the ROSETTA server.

In conclusion, our studies demonstrate that one can obtain the full length model of the desired protein by using an appropriate server that can combine both the template-based prediction and the *de novo* prediction algorithms. Partial models are often incomplete to explain the relevant biological functions of the proteins with inconclusive results.

Materials and Methods: The national center for biotechnology information (NCBI) was used to obtain the FASTA sequences of CFTR and Hox-B13 proteins with reference numbers NP_000483.3 and NP_006352.2, respectively. Three-dimensional structures were predicted by submitting the FASTA sequences to the SWISS MODEL server (https://swissmodel.expasy.org/) or ROSETTA server [20]. The final models were downloaded and analyzed using PyMol molecular graphics software.

References

1. Studer, G., Tauriello, G., Bienert, S., Waterhouse, A.M., Bertoni, M., Bordoli, L., Schwede, T., Lepore, R. Modeling of Protein Tertiary and Quaternary Structures Based on Evolutionary Information. Methods Mol. Biol. 1851, 301-316 (2019).
2. Addala S, Vissapragada M, Aggunna M, et al. Success of Current COVID-19 Vaccine Strategies vs. the Epitope Topology of SARS-CoV-2 Spike Protein-Receptor Binding Domain (RBD): A Computational Study of RBD Topology to Guide Future Vaccine Design. *Vaccines (Basel)*. 2022;10(6):841. Published 2022 May 25. doi:10.3390/vaccines10060841.
3. Peitsch, M.C., Herzyk, P., Wells, T.N., Hubbard, R.E. Automated modelling of the transmembrane region of G-protein coupled receptor by Swiss-model. Recept. Channels 4, 161-4 (1996).
4. Peitsch, M.C. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. Biochem. Soc. Trans. 24, 274-279 (1996).
5. Peitsch, M.C., Tschopp, J. Comparative molecular modelling of the Fas-ligand and other members of the TNF family. Mol. Immunol. 32, 761-772 (1995).
6. Peitsch, M.C., Jongeneel, C.V. A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. Int. Immunol. 5, 233-238 (1993).
7. Guex, N., Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714-2723 (1997).
8. Guex, N., Diemand, A., Peitsch, M.C. Protein modelling for all. Trends Biochem. Sci. 24, 364 - 367 (1999).

9. Schwede, T., Kopp, J., Guex, N., Peitsch, M.C. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 31, 3381-3385 (2003).
10. Kopp, J., Schwede, T. The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. Nucleic Acids Res. 32, D230–D234 (2004).
11. Kopp, J., Schwede, T. The SWISS-MODEL Repository: new features and functionalities. Nucleic Acids Res. 34, D315–D318 (2006).
12. Arnold, K., Bordoli, L., Kopp, J., Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22, 95-201 (2006).
13. Benkert, P., Schwede, T., Tosatto, S.C.E. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Struct. Biol. 9, 35 (2009).
14. Benkert, P., Künzli, M., Schwede, T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 37, W510-W514 (2009).
15. Kiefer, F., Arnold, K., Kuenzli, M., Bordoli, L., Schwede, T. The SWISS-MODEL Repository and associated resources. Nucleic Acids Res. 37, D387–D392 (2009).
16. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., Schwede, T. Protein structure homology modeling using SWISS-MODEL workspace. Nat. Protoc. 4, 1–13 (2009).
17. Bordoli, L., Schwede T. Automated protein structure modeling with SWISS-MODEL Workspace and the Protein Model Portal. Methods Mol. Biol. 857, 107-136 (2012).
18. Mariani, V., Biasini, M., Barbato, A., Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29, 2722-2728 (2013).
19. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L., Schwede, T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 42, W252-W258 (2014).
20. https://www.rosettacommons.org/
21. Ovchinnikov S, Kinch L, Park H, et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife*. 2015;4:e09248. Published 2015 Sep 3. doi:10.7554/eLife.09248.
22. Uchime O, Herrera R, Reiter K, et al. Analysis of the conformation and function of the Plasmodium falciparum merozoite proteins MTRAP and PTRAMP. *Eukaryot Cell*. 2012;11(5):615-625. doi:10.1128/EC.00039-12.
23. Jiang L, Kuhlman B, Kortemme T, Baker D. A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins*. 2005;58(4):893-904. doi:10.1002/prot.20347.
24. Morozov AV, Kortemme T, Tsemekhman K, Baker D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc Natl Acad Sci U S A*. 2004;101(18):6946-6951. doi:10.1073/pnas.0 307578101.
25. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol*. 2003;326(4):1239-1259. doi:10.1016/s0022-2836(03)0002 1-4.
26. Koehler Leman J, Lyskov S, Lewis SM, et al. Ensuring scientific reproducibility in bio-macromolecular modeling via extensive, automated benchmarks. *Nat Commun*. 2021;12(1):6947. Published 2021 Nov 29. doi:10.1038/s41 467-021-27222-7.
27. Alford RF, Leaver-Fay A, Jeliazkov JR, et al. Correction to "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design" [published online ahead of print, 2022 Jun 6]. *J Chem Theory Comput*. 2022;10.1021 /acs.jctc.2c00500. doi:10.1021/acs.jctc.2c0 0500.
28. Du Z, Su H, Wang W, et al. The trRosetta server for fast and accurate protein structure prediction. *Nat Protoc*. 2021;16(12):5634-5651. doi:10.103 8/s41596-021-00628- 9.
29. Norn C, Wicky BIM, Juergens D, et al. Protein sequence design by conformational landscape optimization. *Proc Natl Acad Sci U S A*. 2021;118(11):e2017228118. doi:10.1073/pnas.2 017228118.
30. Anishchenko I, Baek M, Park H, et al. Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. *Proteins*. 2021;89(12):1722-1733. doi:10.1002/prot.26194.

How to cite this article?

Sheik, R., Vangalapudi, H. and Yedidi R. S. (2022). *TCABSE-J*, Vol. 1, Issue 3:40-43. Epub: Apr 2nd, 2022.

Acknowledgements: The authors thank T*yi*DE-Toronto, Canada for helping to write this manuscript.

Funding: The authors thank TCABS-E, Rajahmundry, India and T*yi*DE-Toronto, Canada for financial support.

Conflict of interest: This research article is an ongoing project currently at TCABS-E, Rajahmundry, India. The authors invite collaborations without any conflict of interest.