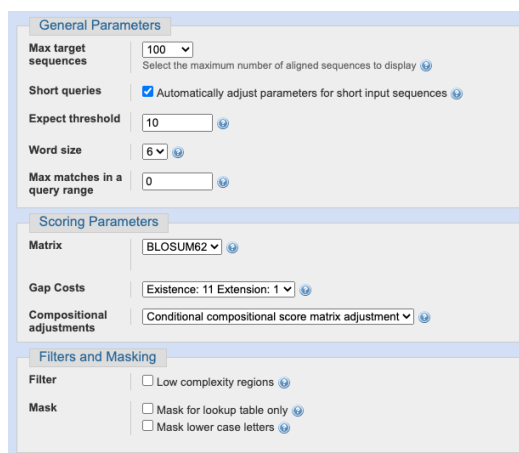


Protein Modelling Assignment by Aravind Venkateswaran z5208102

Functional characterization of a protein sequence is one of the most frequent problems in biology. Hence, protein structure prediction is crucial in the field of biology to discover essential functions that sustain life.

Finding a template

To find a template to start with the modelling, I used blast-P at <https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>. I searched using the pdb database using the following parameters:



I had only 2 hits with:

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	Chain A_C protein [Escherichia virus P2]	32.7	32.7	37%	0.014	42.11%	2L49_A
<input checked="" type="checkbox"/>	Chain Q_RNA polymerase I-specific transcription initiation factor RRN7 [Saccharomyces cerevisiae]	26.2	26.2	51%	7.6	34.78%	5NSY_Q

To search for more homologous templates, I tried using different parameters (changing matrix, etc) but didn't achieve any significant hits. I also tried to search using psi blast for distant homologues (built a PSSM from nr database), but couldn't iterate after 1 cycle since most of the hits had very large E-values (except the one above, 2L49). I finally decided to have 2L49_A (from above) as a template for my protein. I also searched for templates using Swiss Model at the ExPASy server <https://swissmodel.expasy.org/interactive>. Results:

Sort	Name	Title
<input checked="" type="checkbox"/>	2xcj.1.A	C PROTEIN CRYSTAL STRUCTURE OF P2 C, THE IMMUNITY REPRESSOR OF TEMPERATE E. COLI PHAGE P2
<input type="checkbox"/>	2i49.1.B	C protein The solution structure of the P2 C, the immunity repressor of the P2 bacteriophage
<input type="checkbox"/>	2i49.1.A	C protein The solution structure of the P2 C, the immunity repressor of the P2 bacteriophage
<input type="checkbox"/>	3mif.1.B	transcriptional regulator Putative transcriptional regulator from Staphylococcus aureus.
<input type="checkbox"/>	3mif.1.A	transcriptional regulator Putative transcriptional regulator from Staphylococcus aureus.
<input type="checkbox"/>	3zkc.1.A	HTH-TYPE TRANSCRIPTIONAL REGULATOR SINR Crystal structure of the master regulator for biofilm formation SinR in complex with DNA.
<input type="checkbox"/>	6jq1.1.A	Transcriptional regulator, XRE family Crystal Structure of DdrO from Deinococcus geothermalis
<input type="checkbox"/>	6jq1.1.B	Transcriptional regulator, XRE family Crystal Structure of DdrO from Deinococcus geothermalis

The top templates matched exactly with the blast results. When I examined the top hit 2XCJ.1.A, it had the same sequence as 2L49. Both of them were the same proteins but the method used to determine structure were different (2XCJ -> crystal, 2L49 -> solution). This was found using <http://www.rcsb.org/>.

Template details:

PBD-ID: 2XCJ chain A (chosen since it produced better results, highlighted in validation section)

Name: Crystal structure of P2 C, the immunity repressor of temperate E. coli phage P2

Description: Viral protein from the virus [Escherichia virus P2](#). [TaxID: [543](#)]

Function: DNA binding protein.

Resolution: 1.80 Å

Swissprot: Q83VS7

Why I chose this method:

Blast was my go-to for finding homologous sequences. Since I found a template greater than 25% sequence identity with a relatively good e-value, I went for homology modelling. Since swiss model found relevant templates that agreed with blast results, I am confident that this method is appropriate.

Modelling using template

I used the target-template alignment at <https://swissmodel.expasy.org/interactive#alignment> to build a model. Did a multiple alignment at clustal Omega <https://www.ebi.ac.uk/Tools/msa/clustalo/>. The alignment being:

```
>pf4mutant
MSTPADRARLLIKKIG---PKKVSLHGGDYERWKS VSKGAIRVSTEEIDV-----LVKI
FPNYALWIASGSIAPVVGQTSPDYDEANLNLSN-----QNAG
>2XCJ_1|Chain A PROTEIN|ENTEROBACTERIA PHAGE P2
MSNTISEKIVLMRKSEYLSRQQLADLTGVPGTLSYYESGR---STPPTDVM MNILQTPQ
FTKYTLWFMTNQIAPESGQIAPALAHFGQNETTSPHSGQKTG
```

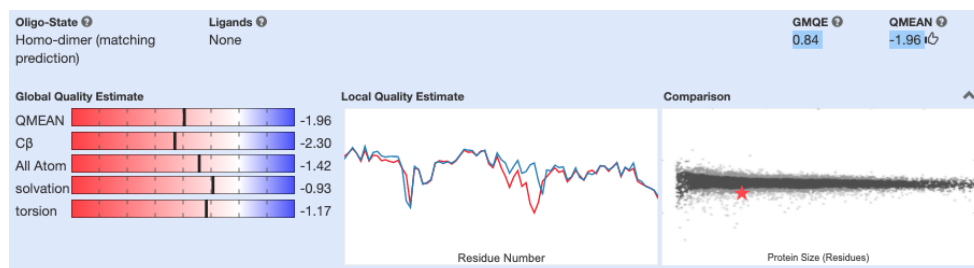
This (above) was my input for the target-template alignment at Swiss model. It produced a model for me automatically. This model was then used again to run a second iteration of Swiss model with the target alignment, which improved the results. Further iteration decreased PROVE score(z-score deviation from high-resolution protein structures on pdb) which I think is due to overfitting.

Why I chose this method:

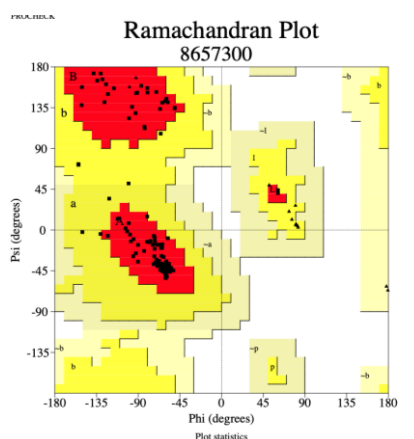
Due to time constraints, I had to quickly get a model. I had submitted multiple templates to different services (using template 2 days ago) but they are taking a long time to complete. Swiss model was quick and it did a target-template model and did several processes automatically such as Backbone generation, loop modelling, side chain modelling and optimisation of model.

Validating predicted model

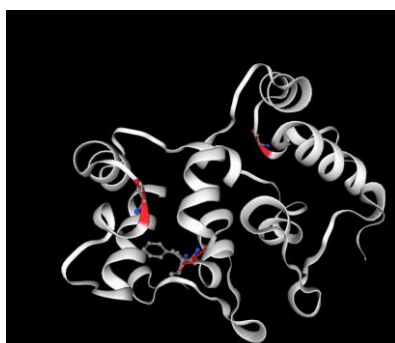
For validating my predicted model, I used PROCHECK, PROVE, WHATCHECK, Verify3D. All of these programs can be accessed through <https://servicesn.mbi.ucla.edu/SAVES/>. Swiss model shows only 4 Ramachandran outliers and $\frac{9}{1658}$ bad angles. The overall result below:



The Ramachandran plot from PROCHECK shows that most of the residues are in favoured regions.



Structure showing Ramachandran outliers



The overall result from WHATCHECK is a Pass suggesting that most of the residues have the right stereochemical properties.

49. Note: Overall summary report Pass

This is an attempt to create an overall summary of the quality of the structure. We do not recommend anyone to look at these numbers, please look at the complete report instead.

Structure Z-scores, positive is better than average:
 1st generation packing quality : -1.748
 Ramachandran plot appearance : 0.287
 chi-1/chi-2 rotamer normality : 0.206
 Backbone conformation : -34.427 (bad)

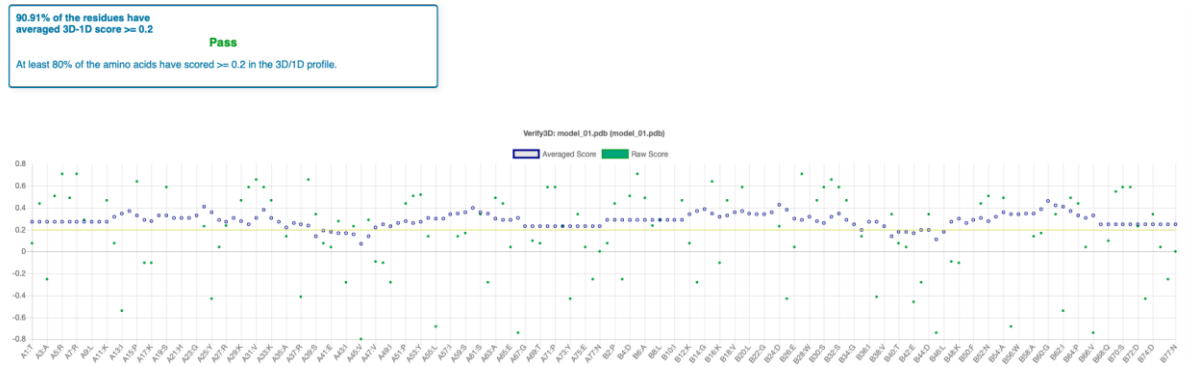
RMS Z-scores, should be close to 1.0:
 Bond lengths : 0.742
 Bond angles : 1.077
 Omega angle restraints : 1.299 (loose)
 Side chain planarity : 1.365

There were 2 structural errors out of the numerous tests

- One bond between amino acids had a very short distance, specifically:
47 VAL (49) A CG1 -- 49 ILE (51) A CG1 0.004 3.196 INTRA
- Backbone fold is unusual

Suggesting that the structure should be further optimised or manually edited to take this into account

Verify3D produced this result:



Suggesting that the compatibility of my model with its sequence is proper and sound.

My prove score was 5% (Warning), suggesting a realistic model that should be further optimized.

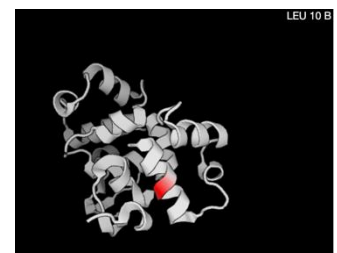
Why I chose these methods:

These methods were held in high standards to validate protein models. Hence, I used them to validate my protein.

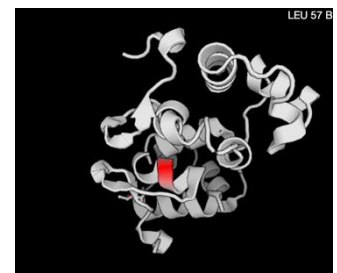
Discussion

I am fairly confident that the basic structure of the protein is similar to my predicted model. Although not exactly, the most essential structures in the target should resemble my model. This is because the template used is similar (~25%) enough to produce a realistic model closely related structure to target in homology modelling. I am confident in a few regions where the alignment of the template and the target is similar.

First region I am confident in is (3) PADRARLLIKK(11) in the target sequence. Structure shown to the right.



The second region I am confident in is the region is (53) PNYALWIAS (61). Structure shown to the right. Since most of the sequence and the type of amino acids match, I think this region should be fairly similar to the target.



In conclusion to my confident regions, I think that both the sequence regions above should have an alpha helix. I mainly came to this decision since both of these regions has similar sequence of amino acids compared to the template.

References

- Laskowski R A, MacArthur M W, Moss D S, Thornton J M (1993). PROCHECK - a program to check the stereochemical quality of protein structures. *J. App. Cryst.*, **26**, 283-291.
- Laskowski R A, Rullmann J A, MacArthur M W, Kaptein R, Thornton J M (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, **8**, 477-486.
- Pontius J, Richelle J, Wodak S.J. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol.* 1996;264(1):121-136. doi:10.1006/jmbi.1996.0628.
- WHAT IF: A molecular modeling and drug design program. G.Vriend, J. Mol. Graph. (1990) 8, 52-56.
- Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. 1991;253(5016):164-170. doi:10.1126/science.1853201.
- Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature*. 1992;356(6364):83-85. doi:10.1038/356083a0.
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296-W303 (2018).
- Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., Schwede, T. The SWISS-MODEL Repository - new features and functionality. *Nucleic Acids Res.* 45, D313-D319 (2017).
- Guex, N., Peitsch, M.C., Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* 30, S162-S173 (2009).
- Studer, G., Rempfer, C., Waterhouse, A.M., Gumienny, G., Haas, J., Schwede, T. QMEANDisCo - distance constraints applied on model quality estimation. *Bioinformatics* 36, 1765-1771 (2020).
- Benkert, P., Biasini, M., Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27, 343-350 (2011).
- Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Reports* 7 (2017).
- Massad T, Skaar K, Nilsson H, et al. Crystal structure of the P2 C-repressor: a binder of non-palindromic direct DNA repeats. *Nucleic Acids Res.* 2010;38(21):7778-7790. doi:10.1093/nar/gkq626
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.* 215:403-410.
- Gish, W. & States, D.J. (1993) "Identification of protein coding regions by database similarity search." *Nature Genet.* 3:266-272.
- Madden, T.L., Tatusov, R.L. & Zhang, J. (1996) "Applications of network BLAST server" *Meth. Enzymol.* 266:131-141.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.* 25:3389-3402.
- National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US).
- H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank *Nucleic Acids Research*, **28**: 235-242.
- Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*. 2019 Jul;47(W1):W636-W641. DOI: 10.1093/nar/gkz268.