Project 3 – AlphaFold

Introduction

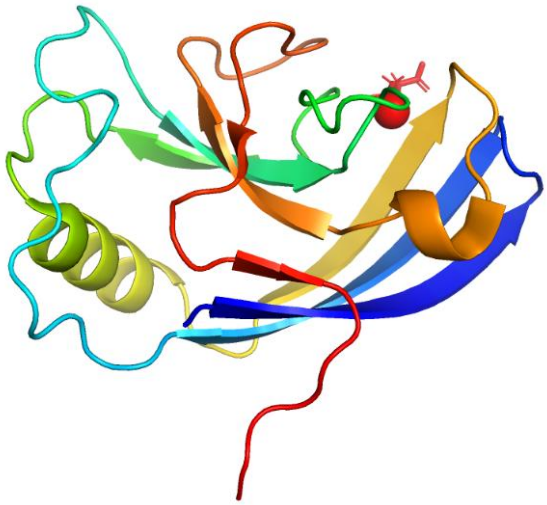
The human body can produce anywhere from 20,000 to over a million different kinds of proteins each having its own type of unique structure and sequence allowing for precise and specific functionality. Proteins also contain sites for interaction both among other proteins and other molecules via binding sites, channels and receptors. Proteins operate on the principle of structure corresponding to function understanding the structure of proteins can have beneficial implications in the field of drug design, disease understanding, protein engineering, and evolutionary research (Corvalan, 2020).

Amino acid sequences dictate how a protein folds into its specific native state however, often times it conveys limited information about how the protein actually folds. It would take an immense amount of time to find all of the possible conformations to get to find the native state however, most proteins are able to fold in very short amounts of time. This is where computer-based 3D structure prediction plays a role. In 1994 CASP was founded or the critical assessment of protein structure prediction and Google's DeepMind participated and won in 2018, in 2020 however, their AI program AlphaFold showed massive strides in the field for its highly accurate structure prediction (Corvalan, 2020).

Sequence 1

The first given sequence is for the fimbrial adhesion also called the pili. They are polymer fibers that help both gram negative and positive bacteria in attachment to surfaces, motility, DNA transfer and biofilm formation (Berne, 2015). This specific sequence was found from the gram-negative bacteria Proteus mirabilus known for its role in catheter associated urinary tract infections. In the study that the structure standard was created from it was found that the structure contained a specific site for a transition metal center with Zn2+ with three histidine residues and a ligand. Since there is this zinc dependence present in the Proteus mirabilus, by changing the levels of zinc present it can have positive implications for the treatments of catheter based urinary tract infections (Jiang, 2020).

The following is the gold standard obtained from the protein database for the chain A sequence of the fimbral adhesion protein



Gold Standard for the 6Y4F Chain A obtained from rcsb.org

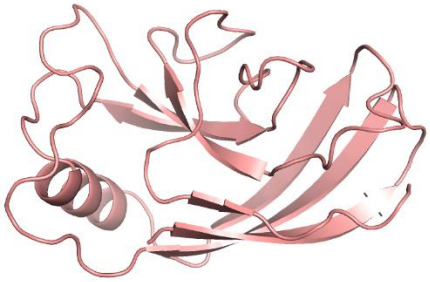
Using AlphaFold the following structure predictions were created for the 6Y4F protein

Rank 1

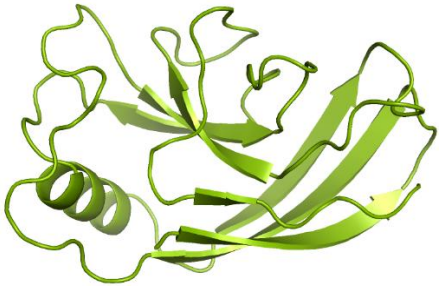
Rank 2

Rank 3

Rank 4



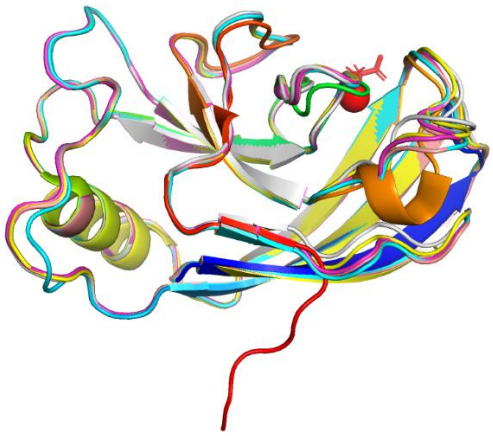
Rank 5



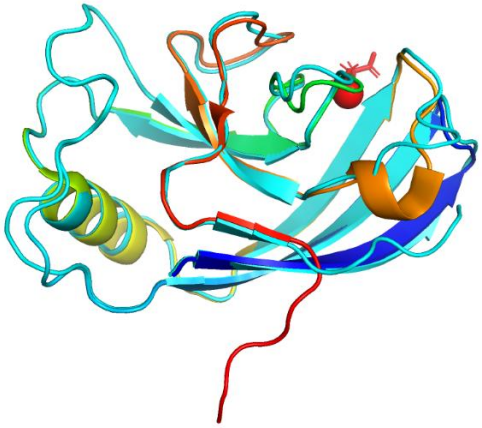
In order to compare the different alignments generated by AlphaFold, the RMSD score can be used also known as the root mean square deviation. This measures the average distance between the atoms in aligned proteins. A RMSD score of 0 suggests identical structures however, it is difficult to compare values between structures of different sizes and the same value will have different significance based on the number of residues in that sequence (Carugo, 2001).

When the RMSD scores were compared to the gold standard which was obtained from the protein data bank the following scores were obtained.

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
0.498	0.590	0.602	0.502	0.620



All generated sequences aligned with gold standard



Best generated sequence (Rank1) aligned with gold standard

Based on the scores the RMSD values are pretty close to the gold standard obtained. Each of the ranks have a very low RMSD value which means that there is less distance between the gold standard and the generated sequences. This suggests that the prediction was very close to the actual standard structure.

RMSD scores for each of the generated sequences compared to each other

Rank	Scores
1 & 2	0.234
1 & 3	0.214
1 & 4	0.186
1 & 5	0.255
2 & 3	0.204
2 & 4	0.219
2 & 5	0.224
3 & 4	0.197
3 & 5	0.209
4 & 5	0.256

Generated sequences aligned with each other

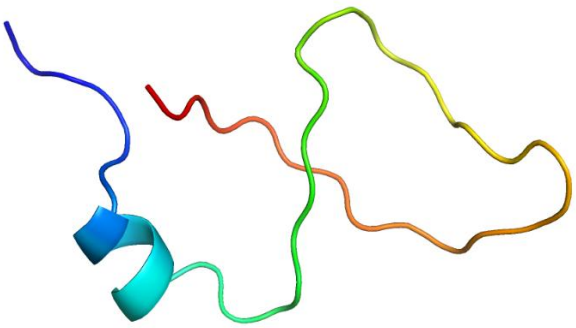
Comparing the sequences between each other their scores are very similar which suggest that they are similar in structure. Therefore, AlphaFold for this sequence was able to generate models that are close to the actual gold standard structure.

Sequence 2

The second sequence provided is found in homo sapiens and is the chain B of the Cryo-Em structure of CST bound to telomeric single stranded DNA. The CST complex is essential for telomere replication and functions as a DNA polymerase alpha primase cofactor. Telomeres are important DNA-protein structures found at the end of chromosomes that protect it from degradation, unneeded recombination, repair and intrachromosomal fusions. They are vital in protecting the information in the genome. (Shammas, 2011).

Its functions include to recover stalled replication fork and help with DNA damage repair. Mutations in the CST can lead to genetic diseases such as Coats plus syndrome along with dyskeratosis congenita (Lim, 2020). The CST complex can bind both short telomeric DNA sequences to elongate the DNA sequence along with longer ssDNA sequences. There however, is limited understanding how the mammalian complex functions (Lim, 2020).

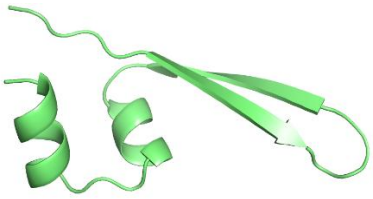
The following is the gold standard obtained from the protein database for the chain A sequence of the fimbral adhesion protein



Gold Standard for the 6W6W Chain B obtained from rcsb.org

Using AlphaFold the following structure predictions were created

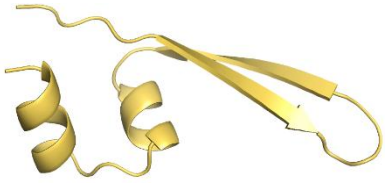
Rank 1



Rank 2

Rank 3

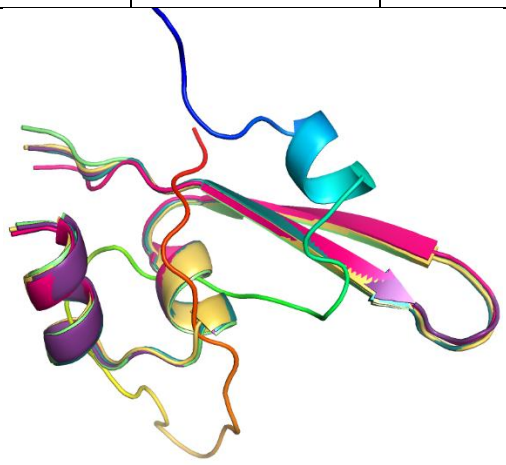
Rank 4



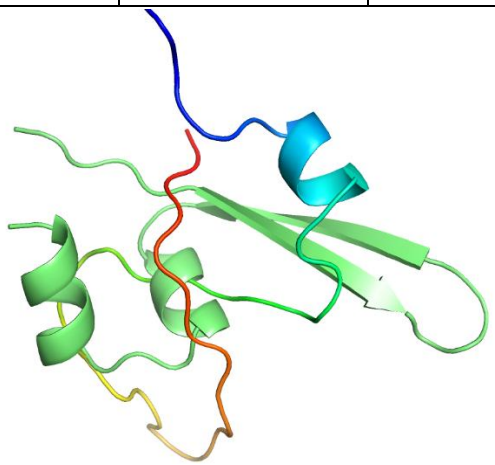
Rank 5

When the RMSD scores were compared to the gold standard which was obtained from the protein data bank the following scores were obtained.

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
12.311	12.443	12.629	12.562	12.353



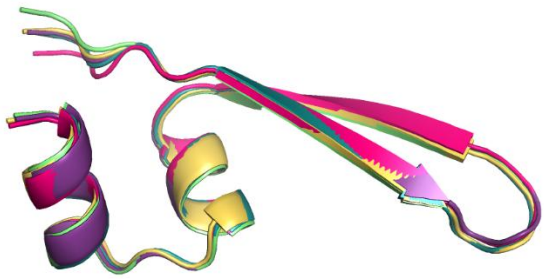
All generated sequences aligned with gold standard



Best generated sequence (Rank 1) aligned with gold standard

Rank	Scores
1 & 2	0.342
1 & 3	0.314
1 & 4	0.472
1 & 5	0.341
2 & 3	0.339
2 & 4	0.410
2 & 5	0.323
3 & 4	0.356
3 & 5	0.267
4 & 5	0.578

Comparing the RSMD scores to each other



Generated sequences aligned with each other

Analyzing this second sequence it can be seen that the RMSD values are significantly higher. Ranging in the 12s when compared to the gold standard sequence suggesting that there is a significant difference present between the generated models and the gold standard one. This suggests the generated models are maybe not as accurate and precise relative to the standard. When comparing the sequences to themselves they are relatively similar to each other as their RMSD value is less than 1.

Conclusions

In conclusion AlphaFold is a very powerful tool in terms of being able to model proteins and generate accurate models it can help scientists save a lot of time and give them a good basis for generating the hypothesis for how many proteins look like.

There was a contrast present however in the accuracy of the AlphaFold model. For the first sequence there were around 130 amino acids given so it was a longer sequence. As maybe a result of that there was a high accuracy structure produced. Comparatively for the second sequence only had about 40 amino acids so it was much smaller which is potentially why a structure with less accuracy could have been generated. Additionally, since sequence 2 was a partial component of a larger protein, the generated structure was not as accurate as the first.

There can be several other factors to consider along when comparing two proteins. One is the way by which the comparison is made in this case the RMSD value. RMSD values can be useful when applied to very similar proteins but comparing RMSD values between two different structures is less useful. The same RMSD value can mean different things when considering the residues that structure has. Therefore, it can be a good indicator for structural identity but less so for structural divergence (Carugo, 2001).

Another consideration factor is that proteins are very dynamic they rapidly changed based on the environment, temperature, binding of ligands and certain protein-protein interactions. It is hard to define a single structure for a lot of proteins as they are rapidly undergoing changes. Therefore, AI models can be very accurate to generate certain structures, however they do not yet take into account the other factors that go into protein dynamics (Terwilliger, 2024).

Overall, AI models such as AlphaFold are very useful for scientists often times as a starting point to generate protein structures, however there is still a way to go in generating super highly accurate models considering all of protein dynamics that can match the crystal structures that can be obtained.

References

Li, Y., Chen, Z., & Gong, S. (2015). The molecular mechanisms of telomerase regulation. *Protein & Cell*, *6*(5), 334–338. https://doi.org/10.1007/s13238-015-0154-5

Piip (2021, August 26). The Importance of AlphaFold in Protein Structure Prediction. https://piip.co.kr/en/blog/alphafold-protein-structure-prediction-importance-1

Shan, M., Liao, J., & Liao, M. (2020). Overexpression of TGF-β1 in murine myocardium increases susceptibility to cardiac dysfunction and Lethality in Chagas disease. *PLOS Pathogens*, *16*(9), e1008707. https://doi.org/10.1371/journal.ppat.1008707

Onuchic, J. N., & Wolynes, P. G. (2004). Theory of protein folding. *Current Opinion in Structural Biology*, *14*(1), 70–75. https://doi.org/10.1016/j.sbi.2004.01.011

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., Hassabis, D., & ... Legresley, P. (2022). Improved protein structure prediction using potentials from deep learning. *Science*, *376*(6623), eaz9649. https://doi.org/10.1126/science.aaz9649

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., & ... Senior, A. W. (2022). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Kohli, P. (2023). Highly accurate protein structure prediction with AlphaFold. *Nature Methods*, *19*(4), 384–391. https://doi.org/10.1038/s41592-022-01234-5

Chin, J. W. (2012). Expanding and reprogramming the genetic code. *Nature*, *484*(7393), 339–344. https://doi.org/10.1038/nature10946

AlphaFold (n.d.). AlphaFold. https://alphafold.ebi.ac.uk/

Hill, J. M., & Quenelle, D. C. (2018). Cardiotropic alphaherpesvirus associated with transient dilated cardiomyopathy in a healthy adult. *Virology Journal*, *15*(1), 45. https://doi.org/10.1186/s12985-018-0958-8