A Review On Homology And Thread Modelling Of Protein Sequence Using Protein Similarity Search And Alphafold2 Colab

1st Chhote Lal Prasad Gupta , 2nd Shashank Gaur , 3rd Manish Kumar Soni

[1]Computer Science and EngineeringBansal Institute of Engineering andTechnology Lucknow, India clpgupta@gmail.comhttps://orcid.org/ 0000-0002-7202-3872
[2]Computer Science and EngineeringBansal Institute of Engineering and Technology Lucknow, India shshnk004@gmail.com https://orcid.org/ 0000-0001-6292-6632
[3]Computer Science and EngineeringBansal Institute of Engineering and Technology Lucknow, India Manish.soni.csit@gmail.com
DOI: 10.47750/pnr.2022.13.S08.140

Abstract

Predicting a protein structure using primary method like X-ray crystallography and Nuclear Magnetic Resonance (NMR) are very costly and time taking for predictinga 3D protein structure, these primary methods takes three to five years, which is too much time. So, to improve the predictiontime of the protein structure Deep Mind has developed a computational tool called alphafold colab to predict the 3D protein structure of the unknown protein sequence using computational method which are based on Artificial Intelligence(AI). In this paper we are reviewing homology modelling andthread modelling based predicted 3D protein structure of the unknown protein sequence based on AlphaFold Database (AFDB).

Keywords— AFDB, Artificial Intelligence (AI), pLDDT, Identities %, AlphaFold

I. INTRODUCTION

A protein's three-dimensional shape can be predicted fromits amino acid sequence using protein structure prediction. [1].This issue is crucial since a protein's function helps in the determination of its structure [2], but determining protein structures experimentally can be challenging. Recent years have seen significant advancements due to the use of geneticdata. Covariation in homologous sequences can be studied toidentify which amino acid residues are in contact, which helpswith the prediction of protein structures [3].

Grasp the function of a protein requires a fundamental understanding of the three-dimensional or tertiary structure ofproteins. X-ray crystallography and nuclear magnetic resonance are the primary methods used to ascertain the three-dimensional (3D) structure of proteins (NMR). Proteins are crystallised through X-ray crystallography, and their structures are subsequently ascertained through X-ray diffraction. It can take up to three to five years and is not always easy to determine 3D structure using X-ray crystallography. Another effective method to ascertain the protein structure is NMR. The protein can be investigated in an aqueous environment that may more nearly approximate its real physiological state using NMR than with X-raycrystallography. NMR's main drawback is that it can only beused for tiny proteins with less than 150 amino acids. The difference between the known protein structures and sequences is expanding dramatically. Therefore, it is necessary to develop computational methods for predicting protein structures. Computer-aided proteinconformation/tertiary structure prediction may make it easierto comprehend protein folding, predict tertiary structures forproteins with known sequences but unknown structures, editproteins to include new functionalities, and create drugs.

Three different approaches have been used to tackle the problem of predicting protein structure: i) computer simulation based on empirical energy calculations, ii) knowledge-based approaches using data derived from structure-sequence relationships from experimentally determined protein 3-D structures, and iii) hierarchical methods. Each strategy has advantages and disadvantages.

Knowledge-based approachesHomology modelling

Building an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a comparable homologous protein is known as homology modelling, also known as comparative modelling of proteins (the "template"). In orderto create an alignment that maps the residues in the query sequence to the residues in the template sequence, homologymodelling must first identify one or more known protein structures that are likely to match the structure of the query sequence. Protein structures have been found to be more conserved among homologues than protein sequences, yet sequences with less than 20% sequence identity may have highly diverse structures. [4]

Both naturally occurring homologous proteins and proteins that are related through evolution share similar amino acid sequences. It has been established that three-dimensionalprotein structure has undergone more evolution than would bepredicted based just on sequence conservation. [5]

A structural model of the target is then created using the sequence alignment and template structure. Detectable levelsof sequence similarity typically imply significant structural similarity since protein structures are more conserved than DNA sequences. [6]

One approach of modelling the structure of proteins is known as homology modelling, which bases its predictions onthe following two observations:

1. The particular arrangement of amino acids governs protein structure. Therefore, at least conceptually, knowing the order of amino acids is sufficient to determine the structureof a protein.

2. Throughout evolution, protein structures have remained more constant than DNA sequences. The 61 proteincodon triplets (excluding "stop codons") that determine the 20amino acids deserve praise! Similar sequences adopt nearly comparable structures, and distantly related sequences still fold into similar structures because the changes occur much more slowly than the associated sequence.

Therefore, by knowing the one or more known sequenceswhose protein structures we already know, the structure of query sequences can be generated. The 7 steps of homology modelling are as follows:

a. **Template identification and initial alignment:** Finding a homolog with the structure from which it earned itsname "Homology" (in other words, finding an alignment of query sequence to the database for a sequence for which the structure is already in Database with Significant Identity).

b. **Correction of alignment:** More complex techniques are required to achieve a better alignment after applying the basic screen mentioned above to identify one or more potential modelling templates.

c. **Generation of Backbone:** The construction of the real model can begin once the alignment is complete. Most ofthe model is simple to construct, including the backbone: Thecoordinates of the template residues that appear in the alignment with the model sequence are simply copied.

d. **Loop modelling:** There are gaps in the alignment ofthe model and template sequence. Gaps in the template sequence or the model sequence (insertions). In the first scenario, residues are simply left off the template, leaving a gap in the model that needs to be filled. In the second instance, the continuous backbone from the template is used, it is cut, and the missing residues are inserted.

e. **Side-Chain modelling:** In structurally related proteins, residues that are conserved frequently exhibit similarχ1-angles when we examine their side-chain conformations (rotamers). Therefore, conserved residues can be copied in their entirety from the template to the model.

f. **Optimization of model :** We require the proper backbone, which depends on the rotamers and their packing, in order to accurately forecast the side chain rotamers. Iterative modelling of the rotamers and backbone structure is a common strategy for solving this challenge. We first anticipate the rotamers, then redesign the backbone to make room for them, and last, we retrofit the rotamers to the updated backbone. You keep doing this until the solution converges. There are a number of rotamer prediction and energy minimization stages involved.

g. **Validation of model :** All protein structures have flaws, and homology models are no different. For a particular procedure, two values are mostly responsible for the amount of errors:

- The fraction of the template's sequence that matches the model sequence
- The quantity of template errors

h. **Threading modelling :** By repeating certain steps of the homology modelling procedure, errors in the model can be found and fixed. Running a shorter molecular dynamics simulation helps reduce minor inaccuracies that are introduced during the optimisation process. By selecting a different loop conformation in the loop modelling process, a loop error can be fixed. Large errors in the backbone conformation can necessitate repeating the entire procedure with a new alignment or even a different template.

Protein threading, also known as fold recognition, is a technique for modelling proteins that have the same fold as proteins with well-known structures but lack analogous proteins with well-known structures. Protein threading is usedfor proteins that do not have their homologous protein structures deposited in the Protein Data Bank (PDB), but homology modelling is used for proteins that have. This is where protein threading varies from the homology modellingapproach of structure prediction. In order for threading to function, one must have statistical understanding of the relationship between the protein sequence being modelled andthe structures that have been deposited in the PDB.

Protein threading, also known as fold recognition, is a technique for modelling proteins that have the same fold as proteins with well-known structures but lack analogous proteins with well-known structures. Protein threading is usedfor proteins that do not have their homologous protein structures deposited in the Protein Data Bank (PDB), but homology modelling is used for proteins that have. This is where protein threading varies from the homology modellingapproach of structure prediction. In order for threading to function, one must have statistical understanding of the relationship between the protein sequence being modelled and the structures that have been deposited in the PDB.

The idea of threading protein sequences through different folding patterns entails creating model structures that are purposefully misfolded by adding an erroneous sequence to the backbone of another protein. A precise alignment betweenthe amino acid sequence of the protein under consideration and the appropriate amino acid residue locations of the foldingmotif is necessary to thread a sequence through a fold. A collection of potential amino acid locations in three- dimensional space are established by the known structure. Byaligning the amino acids in the query sequence, the known structure is made to resemble the query sequence. These approaches' main goal is to identify the most likely fold for agiven sequence or the appropriate sequences that might fold into a structure. Only proteins whose amino acid sequences adopt one of the protein folds previously examined by experimental approaches are typically subject to the threadingmethod. The quantity of folds that are available with atomic-level structural knowledge determines how well threading works. When the atomic structure of a fold is known, a protein query sequence can be fitted with the fold.

II. EXPERIMENTAL SETUP

In our approach, we have taken two protein knowledge based computational tools, AlphaFold colab and AlphaFold2colab. In this, a protein Keratin has been taken, if we say more specifically then, Keratin, type I cytoskeletal 14 (P02533 . K1C14_HUMAN) form the Uniport database and after that we have modified the protein sequence. Now, by using Protein Similarity Search computational tool we predict the identity percentage of the unknown protein with the Protein Database(PDB). Experimental sample data in table1.

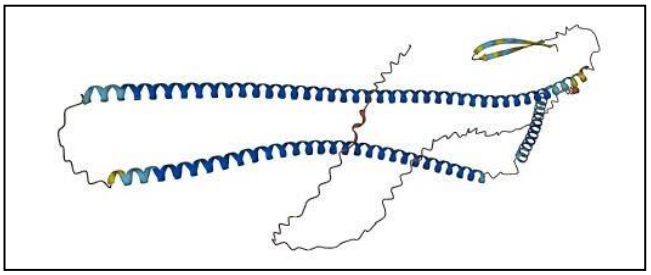
Table1: Experimental Data

Protein structure taken	Keratin, type I cytoskeletal 14 (P02533.K1C14_HUMAN)
Protein Sequence	MTTCSRQFTSSSSMKGSCGIGGGIGGGSSRISSVLAGGSCRAPSTYGGGLSVSSSRFSSGGAC GLGGGYGGGFSSSSS SFGSGFGGGYGGGLGAGLGGGFGGGFAGGDGLLVGSEKVTMQNLNDRLASYLDKVRAL EEANADLEVKIRDWY QRQRPAEIKDYSPYFKTIEDLRNKILTATVDNANVLLQIDNARLAADDFRTKYETELNLRM SVEADINGLRRVLDE LTLARADLEMQIESLKEELAYLKKNHEEEMNALRGQVGGDVNVEMDAAPGVDLSRILNE MRDQYEKMAEKNRK DAEEWFFTKTEELNREVATNSELVQSGKSEISELRRTMQNLEIELQSQLSMKASLENSLEET KGRYCMQLAQIQEM IGSVEEQLAQLRCEMEQQNQEYKILLDVKTRLEQEIATYRRLLEGEDAHLSSSQFSSGSQSS RDVTSSSRQIRTKVD VHDGKVVSTHEQVLRTKN
Unknown Protein sequence	MTTCARQFTAAAAMKGACGIGGGIGGGAARIAAVLAGGACRAPATYGGGLAVAAARFA AGGACGLGGGYGGG FAAAAAAFGAGFGGGYGGGLGAGLGGGFGGGFAGGDGLLVGAEKVTMQNLNDRLAAY LDKVRALEEANADLE VKIRDWYQRQRPAEIKDYAPYFKTIEDLRNKILTATVDNANVLLQIDNARLAADDFRTKY ETELNLRMAVEADIN GLRRVLDELTLARADLEMQIEALKEELAYLKKNHEEEMNALRGQVGGDVNVEMDAAPG VDLARILNEMRDQYE KMAEKNRKDAEEWFFTKTEELNREVATNAELVQAGKAEIAELRRTMQNLEIELQAQLAM KAALENALEETKGRY CMQLAQIQEMIGAVEEQLAQLRCEMEQQNQEYKILLDVKTRLEQEIATYRRLLEGEDAHL AAAQFAAGAQAARDVTAAARQIRTKVMDVHDGKVVATHEQVLRTKN
Protein Database	Uniprot

III. RESULT

In this approach, we are defining predicted 3D protein structure of modified protein sequence as defined in the experimental setup in three ways:

1. Using Protein Similarity Search
2. Using AlphaFold2 colab



Using Protein Similarity Search

In this approach, the unknown protein sequence has beengone through Homology modelling and because of this procedure we got number of predicted proteins which are related to AlphaFolf Database (AFDB). In these results,Identities % is the prime focus. Some of the results are:

Table 2: Predicted 3D protein structure of the unknownprotein sequence

Seq - uence	**AFDB : ID**	**Source**	**Len gth**	**Sco re (Bi ts)**	**Identities %**	**Positives %**	**E()**

1	AFDB :AF- P02533-F1	Keratin , type Icytoskeletal 14	472	387.6	89.0	100.0	7.1 E-106
2	AFDB :AF- Q6IFV1-F1	Keratin , type Icytoskeletal 14	485	333.3	78.2	91.3	1.7 E-89
3	AFDB :AF- Q61781-F1	Keratin , type Icytoskeletal 14	484	332.2	79.8	94.7	3.6 E-89
4	AFDB :AF- P08779-F1	Keratin , type Icytoskeletal 16	473	330.5	77.9	95.2	1.2 E-88
5	AFDB :AF- Q9Z2K1-F1	Keratin , type Icytoskeletal 16	469	299.9	71.6	93.2	1.9 E-79
6	AFDB :AF- Q6IFU9-F1	Keratin16	463	290.3	71.2	93.8	1.4 E-76

Therefore, as we can see in this table2 that there are differentpredictions related to our unknown protein structure, which are present in the AlphaFold Database (AFDB). Now, it can be noticed that the Identities % of AFDB: AF-P02533-F1 is higher (89.0) that means this protein structure is 89.0 % similar to our unknown protein structure.

Model Confidence:

Very high (pLDDT > 90) Confident (90 > pLDDT > 70) Low (70 > pLDDT > 50) Very low (pLDDT < 50)

Fig 1: AFDB: AF-P02533-F1 Protein Structure

Using AlphaFold2 Colab

In this approach we are predicting the 3D protein structure ofthe unknown protein sequence using the computational tool that is AlphaFold2 colab which is based on artificial Intelligence (AI) and designed as a deep Learning system.

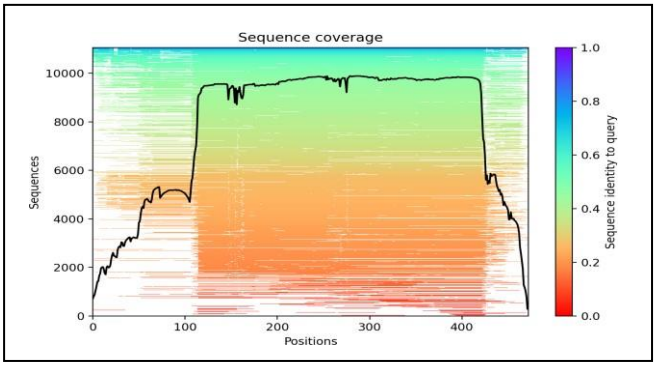


Fig 2: Sequence Alignment Graph of the unknown proteinsequence

In Fig 2, according to this graph, the Alphafold2 colab has a count of 10,000 sequences that match the unknown protein sequence. In this graph, if we see the curve line, then there are 400 residues which are related to the unknown protein sequence.

If we focus over the coverage of the unknown protein sequence, then nearly 9500 sequences have beencovered by this unknown protein sequence.
In this graph, if we look at the red colour region, so nearly initial 2000 sequences have very low identity, but if we see from the range of 2000 to 6000 protein sequences, then it canbe said that the identity of the unknown protein sequence withthe AFDB is improving, and as we go on in the upper range,the identity of the unknown protein sequence with the AFDBhas been improved with high accuracy.
In alphafold2 colab, this unknown protein sequence has beenpredicted through five model with a measure that is predictedLocal Distance Difference Test (pLDDT). On a scale from 0to 100, AlphaFold generates a per-residue estimate of its confidence. The model's anticipated score on the lDDT-C metric correlates to this confidence level, which is known aspLDDT.

Table 3: Predicted model's pLDDT score of the unknownprotein sequence using AlphaFold2 colab

Model Number	pLDDT score
Model 1	69.9
Model 2	71.9
Model 3	73.0
Model 4	70.5
Model 5	70.9

If we study the table 3 then,

- In model 1 pLDDT score is 69.9 which means that, confidence of the model is low that is the accuracy of thepredicted model with respect to the unknown protein sequence is very low.
- In model 2 pLDDT score is 71.9 which means that, confidence of the model is not so good and not so bad also means the accuracy of the predicted model with respect to the unknown protein sequence is average.
- In model 3 pLDDT score is 73 which means that, confidence of the model 3 is higher than the confidence of the model 2 and in other words the accuracy of the predicted model 3 is higher than the accuracy of the predicted model 2.
- In model 4 pLDDT score is 70.5 which means that, confidence of the model 4 is lower than, the confidence of the model 2 and 3.
- In model 5 pLDDT score is 70.9 which means the accuracy of the model 5 is higher than model 4.
- In this approach, we got the final predicted 3D protein structure whose structure are defined by processing five-model using alphafold2 colab.

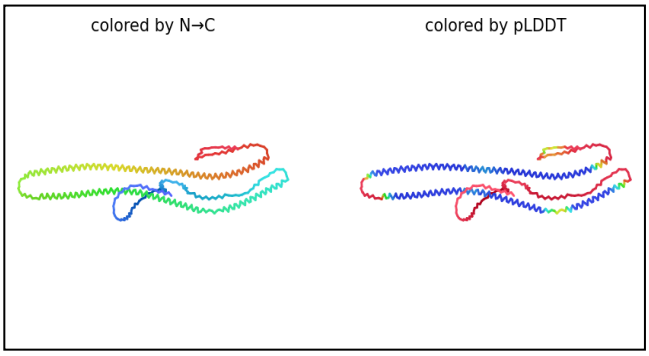


Fig 3: Predicted 3D protein structure of the unknownprotein sequence using Alphafold2 colab

In fig 3, it has been seen that the percentage of the blue regionis large that means the confidence of predicting the 3Dprotein structure of the unknown protein sequence is higher.

IV. CONCLUSION

In this paper, we have predicted an unknown protein sequence using several methods, like using protein similaritysearch and by using alphafold2 colab. In these approaches, wewere focusing on the similarity score of the unknown proteinsequence with

the parameters like identities % and pLDDT score, which are related to protein similarity search and alphafold2 colab respectively, and after analysing these all results, we can conclude that it is true that alphafold2 colab can predict the 3D protein structure of any unknown protein sequence with higher accuracy. But, if we compare the resultsof 3D protein structure prediction using protein similarity search, which gives identities % of 89.0 % and alphafold2 colab which gives pLDDT score of five models as given in table 3, then, The similarity score of the predicted protein structure using protein similarity search is higher as compared to protein structure prediction made by alphafold2 colab. In other words, we can say that protein structure predictions made using homology modelling have a higher similarity score as compared to the predictions made using thread modelling.

v. REFERENCES

[1] Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein-folding problem. Annu Rev Biophys. 2008;37:289-316. doi: 10.1146/annurev.biophys.37.092707.153558. PMID: 18573083; PMCID: PMC2443096.

[2] Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science. 2012 Nov 23;338(6110):1042-6. doi: 10.1126/science.1219021. PMID: 23180855.

[3] Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ. Assessment of contact predictions in CASP12: Co-evolution and deeplearning coming of age. Proteins. 2018 Mar;86 Suppl 1(Suppl Suppl 1):51-66. doi: 10.1002/prot.25407. Epub 2017 Nov 7. PMID: 29071738; PMCID: PMC5820169.

[4] Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986 Apr;5(4):823-6. doi: 10.1002/j.1460-2075.1986.tb04288.x. PMID: 3709526; PMCID: PMC1166865.

[5] Kaczanowski, S., Zielenkiewicz, P. Why similar protein sequences encode similar three-dimensional structures?. Theor Chem Acc 125, 643–650 (2010). https://doi.org/10.1007/s00214-009-0656-3

[6] Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000;29:291-325. doi:10.1146/annurev.biophys.29.1.291. PMID: 10940251.

[7] Gupta, C., Bihari, A. and Tripathi, S., 2022. Protein Classification Using Machine Learning and Statistical Techniques.