Fayoum University Faculty of Engineering, 2024, Vol: 7(2), 338-346
https://fuje.journals.ekb.eg/
ISSN Online: 2537-0634
ISSN Print: 2537-0626

Protein 3D structure prediction Using Homology Method

Rania Ahmed Abdel Azeem Abul Seoud[1], Nahed Mahmoud El Gali[1], Amany Lotyef[1], Margret Ezzat[1*]

[1]Electrical Engineering Department, Faculty of Engineering, Fayoum University, Fayoum 63514, Egypt

*Corresponding author: Margret Ezzat (me1804@fayoum.edu.eg).

How to cite this paper: Abul Seoud, R,A.A. El Gali, N.M., Lotyef, A. & Ezzat, M. (2024). Protein 3D structure prediction Using Homology Method. *Journal of Fayoum University Faculty of Engineering, Selected papers from the Third International Conference on Advanced Engineering Technologies for Sustainable Development ICAETSD, held on 21-22 November 2023, 7(2), 338*-346.
https://dx.doi.org/10.21608/fuje.2024.345050

Abstract

For decades, the prediction of protein three-dimensional structure from amino acid sequence has been a magnificent challenge problem in computational biophysics. This research topic has drawn scientists from a variety of areas of study, including biochemistry and medicine, due to its inherent scientific interest as well as the numerous potential applications for reliable protein structure prediction algorithms, ranging from genome comprehension to protein function prediction. In the past decade, there has been a significant improvement in methods for protein structure prediction and design. New data-intensive and computationally demanding approaches for structure prediction have been developed as a result of increases in computing power and the rapid growth of protein sequence and structure datasets. These approaches typically begin by assuming a probability distribution of protein structures given a target sequence and then finding the most likely structure; however, computer scientists formulate protein structure prediction as an optimization problem in finding the structural solution. Homology modeling, also known as Comparative modeling of the 3D structure of a protein by utilizing structural information from other known protein structures with good sequence similarity, is employed in our study. Homology models contain significant information about the spatial organization of key residues in the protein and are frequently employed in drug design for screening large libraries using molecular docking techniques. The generic structure prediction flowchart is followed by presentations and discussions of important concepts and techniques.

Keywords

Protein structure prediction, homology modeling, *ab initio* prediction, threading methods

1. Introduction

Understanding protein structure is an essential beginning toward sensible structure-based drug development and virtual molecular library screening. The finely customized 3-D structures of naturally evolved proteins, which are determined by their genetically encoded amino acid sequences, enable the study of the wide variety of molecular functions carried out by these proteins. As a result, an analytical understanding of the association between amino acid sequence and protein structure could open up opportunities both for the rational engineering of different protein functions by designing amino acid sequences with specific forms and by predicting functions from genome sequence data. The ability to predict and generate the 3-D structures of proteins has grown significantly over the last ten years, and these developments could have significant medical and biological ramifications (Huang et al., 2023). To identify structurally interacting residues exclusively from sequence information, new machine-learning methods have been created that examine the patterns of linked mutations in protein families. For the first time, enhanced protein energy functions have made it achievable, to begin with a rough structure prediction model and develop it along energy-guided lines until it closely matches the empirically observed structure (AlQuraishi, 2021).

2. Literature review:

Protein conformational sampling and sequence optimization advancements have made it possible to construct unfamiliar protein structures and compounds, some of which have medicinal potential. In (Wang et al., 2017) Wang et al. applied the deep learning method to enhance the predicted residue contacts using CCMPred which uses a 60-layer ResNet to count all residue pairs concurrently, the suggested methodology RaptorX-Contact has excellent correctness in predicting inter-residue contacts. In 2018, Senior et al. proposed AlphaFold, producing the last structure utilizing optimization rather than sampling to accomplish protein prediction with enhanced accuracy and hardiness which is built on a potential of mean force stimulated by the predicted distance map of the goal protein using a simple gradient descent procedure. In (Mirabello & Wallner, 2019), Mirabello et al. offered raw MSA (an end-to-end prototype) using raw MSAs as input. Raw MSA uses the embedding concept from natural language processing, which plots a protein sequence into an adaptively learned uninterrupted space. (Ingraham et al., 2019), Ingraham et al. suggested an end-to-end prediction method called NEMO which presented the potential to construct multimodal predictions and confirm overview capability which comprises a neural energy function and an unfolded Monte Carlo simulator that simulates the folding process, (AlQuraishi, 2019), Al Quraishi proposed RGN, another end-to-end differentiable prototype through a neural network that enhances both local and global geometry simultaneously. In (Mao et al., 2019) Mao et al. offered GDFold, a methodology for speedy protein structure prediction utilizing a neural network. The methodology predicts inter-residue contacts with its architecture improved concluded and utilizes all of the predicted inter-residue contacts rather than counting the top-scored contacts only. (Yang et al., 2020), Yang et al. proposed the trRosetta process, which uses deep neural networks to assess inter-residue distance, dihedral torsion angles, and the relative direction of long-distance residue pairs. trRosetta builds an energy function utilizing the predicted inter-residue distances and directions and then seeks the structure with the lowest energy. In (Huang et al., 2023), Rao et al. proposed a Transformer frame that learns protein structure and function from sets of homologous sequences ordered as several sequence alignments. The pattern, called ssMSA Transformer, inserts row, and column attention to utilize the protection management of residues and correspondence of aligned residues across the input sequences. This model exhibited wonderful implementation in predicting inter-residue contacts and protein structure (Huang et al., 2023). In 2022, Lin et al. proposed a very large protein language model ESM-2 with 15 billion factors and then developed the prediction software ESMFold (Lin et al., 2022), even larger than the language models used by OmegaFold (670 million parameters).

These researches showed the superpower of the language pattern in protein structure prediction. The planners of the CASP competitions examined the prediction tactics that joined CASP competitions and ascribed the improvement in prediction accuracy to the following key technologies:

(i) The structure prediction uses segment assembles and substitution methods.
(ii) utilizing knowledge of co-evolution to forecast inter-residue interactions.
(iii) using a deep learning technique to estimate inter-residue distances, such as ResNet.

(iv) utilizing Transformer to estimate protein structure end-to-end and the spacing between residues.

Scientists from a variety of fields, such as computer science, physics, biochemistry, medicine, and mathematics, have been attracted to the study of protein structure prediction. These researchers approached the same issue using several research paradigms. To comprehend their advantages and disadvantages, we will compare different study paradigms in this article. Homology modeling is quickly replacing older methods for acquiring the 3D structure coordinates and other valuable structural and functional insights with the introduction of new modeling tools and algorithms.
Homology modeling has recently made significant advances, particularly in the areas of loop and side-chain modeling, model validation, and evaluation. Until recently, it was not possible to predict accurate models of protein structure. Improvements in sequence search/analysis, scoring systems, and tertiary structure prediction approaches enable the creation of robust, error-free models with high statistical significance. The advancement of this technique has been further aided by freely accessible, easy-to-use modeling servers and structure validation tools.

3. Methodology

The majority of the currently used methods successfully utilize the relation between sequence and structure as well as the evolutionary data contained by the target protein's homologous proteins to predict structure. The present methods can be categorized between template-based modeling (TBM), which needs template proteins, or proteins with known structures, and free modeling (FM, also known as ab-initio approaches), which does not require any templates. The homology modeling and threading TBM techniques can be further classified. Following is a detailed description of the fundamental concept and representative software implementations of various approaches.

3.1. Protein structure prediction

There are two general methods for predicting the structure of a protein of interest (the "target"): template-based modeling, in which the target's structure is modeled using the structure of a related protein that has already been determined, and template-free modeling, which does not rely on a structure's overall similarity to one in the PDB and can therefore be used for proteins with novel folds. Here, we briefly introduce template-based modeling techniques. Next, we continue to template-free modeling and demonstrate recent advancements in that field (Eisenberg et al., 1992).

3.1.1. Template-based modeling

Finding a proper structural template, alignment of the target sequence to the template structure, and molecular modeling to take into account alterations insertions, and removals that were present in the target-template alignment are the stages in usual template-based modeling. By searching the PDB sequences with single-sequence search tools like BLAST (Altschul et al., 1997), it is feasible to recognize templates that are tightly related to one another. A target sequence profile (Eddy, 1998) generated from a multiple-sequence alignment can be employed to search a database of sequence profiles for proteins with known structures by profile-to-profile comparison or to correspond to a collection of structural templates to identify compatibility between sequences and structures to find closely related templates. After developing a model, template choice techniques frequently return an original target-template alignment that can be directly altered. By performing side-chain optimization only on altered sites and reconstructing the backbone surrounding these modifications. Established methods (Brownstein et al., 2017) can be implemented to quickly construct molecular models of the target sequence provided an alignment to a template. More complicated techniques utilizing multiple templates and broad backbone structural sampling may be required to target protein sequences that are only indirectly connected to proteins with known structures (Brownstein et al., 2017).

3.1.1.1. Threading methods

The method is also known as the Fold Recognition Method. It is also known as fold recognition because the recognition of the Template is a problem in and of itself. Comparative modeling-like procedures apply to threading. The goal of all methods is to identify "folds"

from a library of folds of known protein structures. The threading algorithm for protein structure prediction matches sequences without known structure with protein folds using a library of recognized three-dimensional structures. From the database of protein structures, a collection of unique protein folds is derived. Using threading techniques, a Target sequence is compared to a collection of structural Templates. This technique benefits from being aware of the pre-existing structures in the database and the physical characteristics that stabilize them.

3.1.1.2. Homology modeling

The idea behind homology modeling is that as homology proteins, especially the close homology proteins, typically share similar structures, and as protein structures are more conserved than sequences during the evolutionary process, we can construct a structure for a target protein by referring to the structures of its homologies (Figure 1B). "Sequence-sequence" alignment is a popular technique for finding homologies between a target protein and other proteins. If the sequence alignment of two proteins reveals a substantial amount of sequence similarity, they will be regarded as homology proteins (Speed, 2002). Running structure modeling programs, such as MODELLER (Speed, 2002), allow for the construction of the target protein's structure based on the target protein's acquired alignment with a homology template. In this approach, the sequences of target proteins and templates serve as representations The MSAs, PSSM, and profile hidden Markov model sequences of homologous proteins (Altschul et al., 1997) are efficient in enhancing the sensitivity of homology detection. PSI-BLAST (Altschul et al., 1997), PDB-BLAST (Speed, 2002), SAM-T99 (Altschul et al., 1997). are examples of excellent homology detection or homology modeling software programs.

The steps of the multi-step homology modeling method can be summarized into the following:
1) Recognition and alignment of the template.
2) Alignment optimization.
3) model creation, which entails side chain, loop, and backbone generation.
4) Model improvement.
5) verification.
Obtaining the Target protein sequence from several freely accessible databases is the first and most important step. The following procedure is template identification, which involves choosing a protein template whose 3D coordinates are known and which has the greatest similarity to the target sequence. Sequence alignment is used to compare the Target and Template protein sequences. BLAST (Basic Local Alignment Search Tool) is the alignment algorithm that is used the most frequently. The Template should not have an E-value greater than 1 for the Target and Template sequences to align well. Aligning two sequences in a location with a relatively low percentage of sequence identity can occasionally be challenging. This problem is fixed by employing numerous Templates that are homologous to the Target protein and actively correcting the alignment. Model creation begins upon the achievement of an ideal alignment. The exactness of the alignment created and the validity of the Template are recognized to have a significant impact on the quality of the model produced. The likelihood of the model being wrong decreases if the alignment is precise. The coordinates of those Template residues that appear in the alignment with the model sequence can be duplicated in a model creation if the sequence alignment is strong. Only the backbone coordinates (N, Cα, C, and O) are replicated when two aligned residues are different. As a result of base or amino acid insertions and deletions, the Target-Template alignment may have spaces.

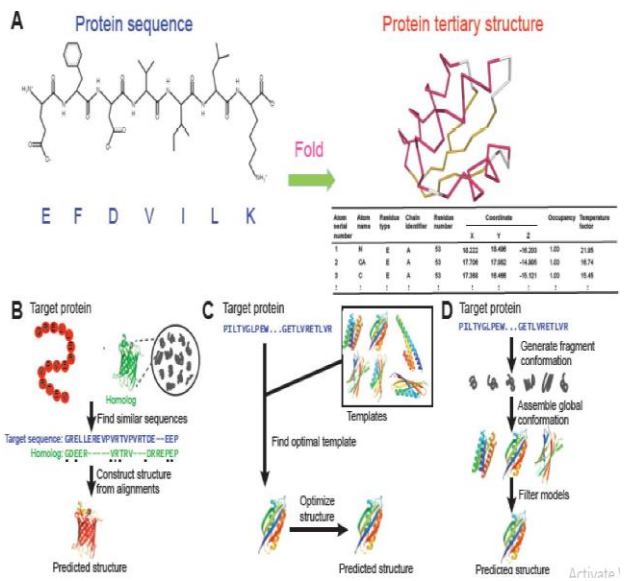


Figure 1: Protein sequence, protein structure, and protein structure prediction (Jiang et al., 2013)

A sample of protein sequence and its tertiary structure. Here, figure 1 displays a C-terminal fragment of the ribosomal protein L7/L12 from Escherichia coli (PDB ID: 1CTF), which contains a full of 74 residues connected via peptide bonds. The tertiary structure states the unique 3D coordinates of each atom in the relative position of the entire protein. Cartoon backbone illustration is extensively utilized to envision protein tertiary structure. B. Homology modeling method for protein structure prediction C. Threading method for protein structure prediction. D. Ab initio prediction approach. PDB, Protein Data Bank; ID, identification; 3D, 3-dimensional.

3.1.2. Template-free modeling

Proteins lacking a general structural connection to a protein in the PDB can use template-free modeling techniques. These approaches require a conformational sampling strategy to produce native-like conformations in the absence of a structural template. without a template, the structure prediction process. The target protein and associated sequences are first assembled into a multiple-sequence alignment. Then, local structural properties like secondary structure and backbone torsion angles are predicted using the sequences of the target and its homologs, as well as non-local structural features like residue-residue interactions or inter-residue distances across the polypeptide chain. The construction of 3D models of the target protein structure is guided by these anticipated properties, which are later improved, rated, and compared to one another to choose the final predictions.

3.1.2.1. Ab initio prediction methods

When no appropriate homolog is located in the database, ab initio prediction is used. It is predicated on Anfinsen's theory that the protein's natural state represents the world's lowest possible level of free energy. These global minima of the protein are searched for using ab initio approaches. To explore the conformational space in the free energy landscape and obtain the global minima, the correct native-like conformation must be found.

Homology modeling steps

Homology modeling is also called comparative modeling, which is used to predict the 3D structure of a protein with an unknown structure by using the known structure of a homologous protein. The process of homology modeling is run by seven classical steps. In this paper, SWISS-MODEL is used for prediction which is a fully automated protein structure homology-modeling server, accessible via the Expasy web server.

1. Identification and selection of templates

First, we start searching for a template based on sequence–sequence alignment, from Protein Data Bank Protein Data Bank (PDB) e. Second, the computation of an accurate alignment must be feasible between the Target sequence and the Template structure as shown in Figure 2 and Figure 3. Overall model accuracy can be predicted by the degree of sequence similarity between the Target and the Template. To identify templates, there are many tools to detect alignment methods e.g. HMMER, PSI-BLAST, HMM, SAM, and HHsearch. In the case of low homology (identity below 35%; the number of identical amino acids in an alignment), we use other methods are used for alignment to reduce shifts and gaps such as profile-profile alignments, Hidden Markov Models (HMMs) and position-specific iterated BLAST (psi-BLAST) to reduce gaps in the alignment.

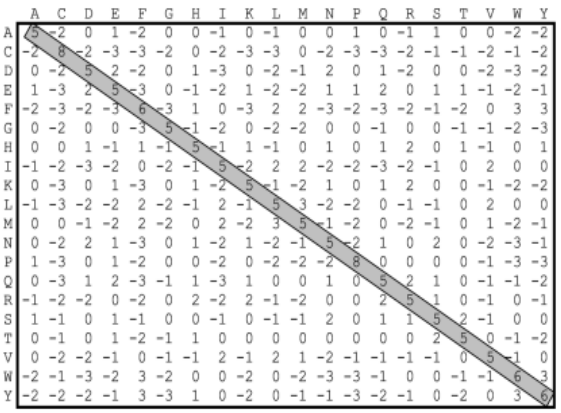


Figure 2: A typical residue exchange or scoring matrix used by alignment algorithms (Jiang et al., 2013)

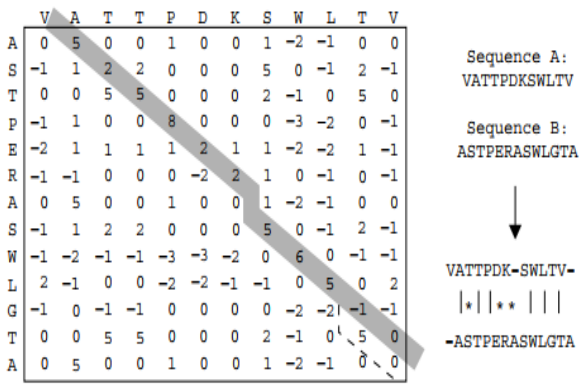


Figure 3: The alignment matrix for the sequences VATTPDKSWLTV and ASTPERASWLGTA, using the scores from Figure 2

2. Alignment of the Query Sequence with the Template:

The scoring function is properly optimized to determine an alignment. For matching residues in the Target and the Template sequence, a substitution score matrix and a gap penalty function are included in a scoring function. A program for multiple sequence alignment is used to look for structurally conserved regions. Figure 4 shows the target template 'cytoplasmic' If there are many available templates there are many factors such as sequence identity and similarity, alignment scores, and phylogenetic relationships that should be considered, as well as other information such as biological function and environmental context.

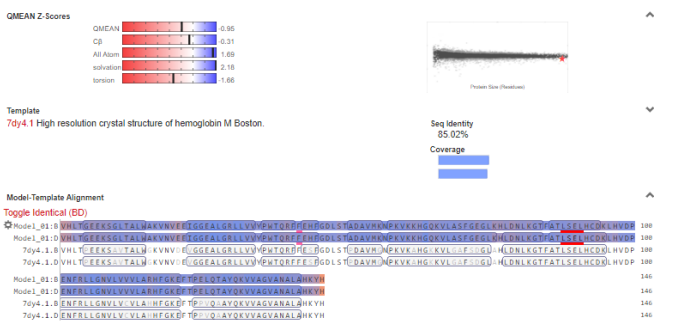
Using multiple sequence alignment, insertions and deletions can be done in very different regions of the molecule. To make gaps as tiny as possible, they must be moved. One or more empty spaces that are aligned with

3. Building a Three-Dimensional Model of the Query Protein:

Sequence alignment is followed by the creation of the target proteins' three-dimensional models. Different techniques are used to create 3D representations of proteins using their templates. Rigid-body assembly, segment matching, spatial restraint, and artificial evolution methods are the four groups into which these techniques are divided. The protein structure is divided

letters in the opposite sequence are known as gaps. To ensure a good alignment, it is recommended to choose to model only that part of the sequence that aligns without a large number of gaps in the alignment. Further, alignment errors are the main cause of deviations in comparative modeling even when the correct Template is chosen. In the alignment, gap positions should lie outside the secondary structures and in the loop areas. It is also recommended to integrate accessory information about secondary structures, conserved family residues, transmembrane helices, active site residues, etc. into the alignment to improve its accuracy.

a: Target protein



b: Alignment

Figure 4: "a" is the target protein, "b" is the alignment of the query

into basic conserved core sections, loops, and side chains during rigid-body assembly. With the aid of programs like 3D-JIGSAW, BUILDER, and SWISS-MODEL, the rigid body parts are extracted from the template protein structures and assembled as explained in Figure 5.

Figure 5: 3D structure of the target protein sequence

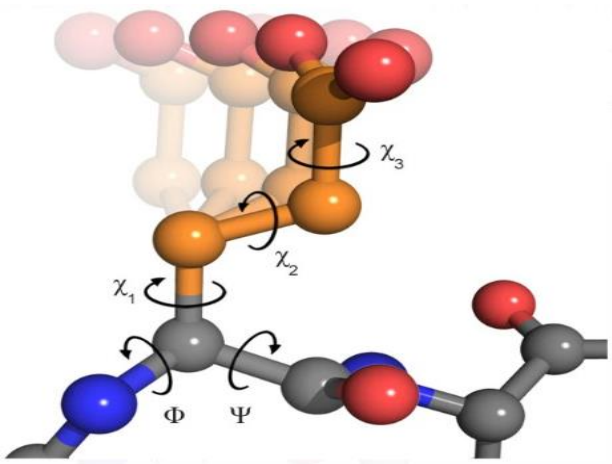


Figure 6: Example of a backbone-dependent rotamer library (Harder et al., 2010)

4. Loop Modeling

When modeling proteins, sequence alignments may have gaps or insertions. Loops are a type of gap that has undergone less structural conservation throughout evolution. Because loops are so critical in defining how a protein functions, loop modeling is a crucial stage in protein structure prediction. In the majority of cases, the alignment between the model and template sequence contains gaps. Either gap in the model sequence (deletions)or in the template sequence (insertions). In the first case, one simply omits residues from the template, creating a hole in the model that must be closed.

In the second case, one takes the continuous backbone from the template, cuts it, and inserts the missing residues. There are generally two methods of loop modeling: 1) by finding similar loops in similar proteins and 2) by generating the segment. Finding loops from similar proteins is done by taking some residues before and after the insertion as “anchor” residues and performing the loop search against databases like PDB with similar anchor residues. The best-fitting loop is copied to the model. Loop selection must be done by carefully examining it for steric overlaps, and by checking loop atoms against the rest of the protein’s atoms.

5. Side Chain Modeling

Side chain modeling is the process of predicting the conformation of the side chains of amino acids in a protein structure. This process is crucial in evaluating protein-ligand and protein-protein interactions.

This is usually done by placing the side chains onto the backbone coordinates that are derived from a parent structure or ab initio modeling simulations as explained in Figure 6. Searching every possible conformation of a side chain is computationally time-consuming and not effective so, most side chain prediction programs use the preferred conformation called rotamers. These rotamers are stored in a rotamer library, which is a collection of preferred side chain conformations ranked by their frequency of occurrence. Different energy functions and search strategies are used to select the most appropriate rotamer for each amino acid side chain based on the preferred protein sequence and the given backbone coordinates. There are various tools available for side chain modeling, such as RAMP and SCWRL.

6. Model optimization

it is used to optimize the quality of the final model. This step is done by using energy minimization utilizing molecular mechanics force fields, to reduce atomic clashes, and exclude all major and small errors. Further optimization can be done using molecular dynamics and Monte Carlo simulations.

7. Model Validation and Evaluation

To make sure that the 3D model of the query protein is suitable. Every homology model contains errors. The number of errors mainly depends on two values: The percentage sequence identity between the template and target. The number of errors in the template. The reason

may either be attributed to a low percentage sequence identity among Target and Template or an error in the Template. Therefore, the model must be evaluated for the overall correctness of the fold/structure, and stereochemical parameters like bond lengths and bond angles as shown in Figure 7.

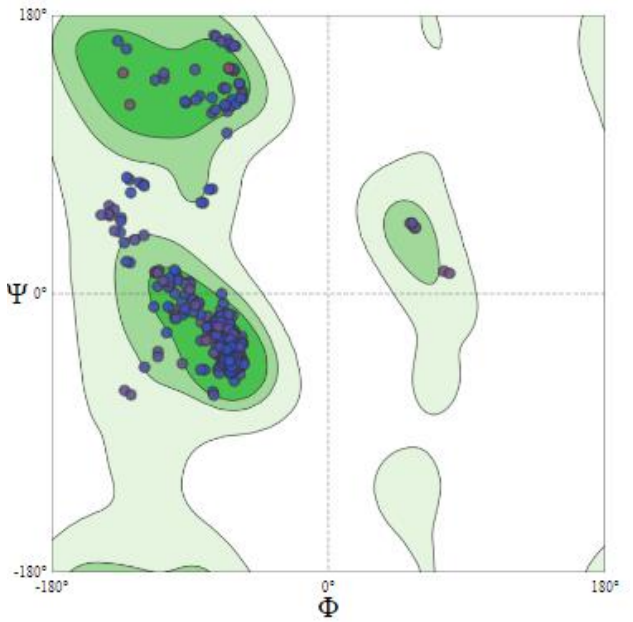


Figure 7: A pair of phi(Ø), psi(Ψ) dihedral angles that occur in a protein structure.

Other properties like the distribution of polar and nonpolar residues, bad contacts, etc. can also be compared with real structures. Validation can be of two types: Internal or External Validation. Internal Validation: Performed self-consistency checks. It deals with the assessment of the stereochemistry of the model like bond lengths, bond angles, dihedral angles, etc. It can be done by programs like WHATCHECK and PROCHECK. b. External Validation: Based on information that was not used in the calculation of the model. External validation involves the check over whether the correct Template was used (based on identity percent). This can be predicted by comparing the Z-score of the model and the Template structure. Z-score is the measure of the compatibility between a model's sequence and its structure as indicated in Figure 8

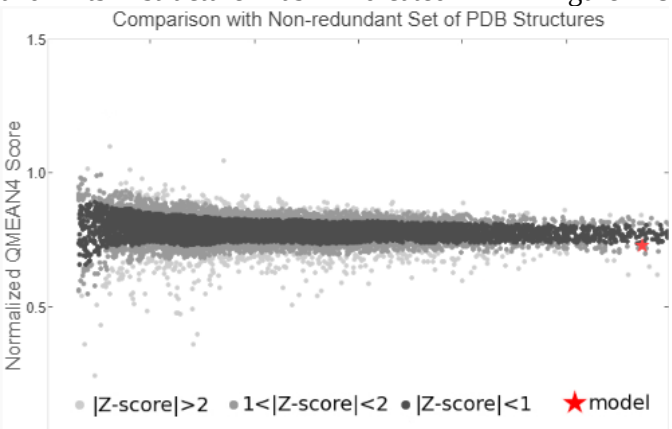


Figure 8: Comparison between the Z-score of the model and the template structure which indicates that the model is accepted

Figure 9 shows the similarity in all chains (A, B, C, D) concerning the high identity template according to the number of residues. The least similarity of the target sequence to template is greater than 40 % and the highest one is greater than 90 % which indicates the high quality of the target model.

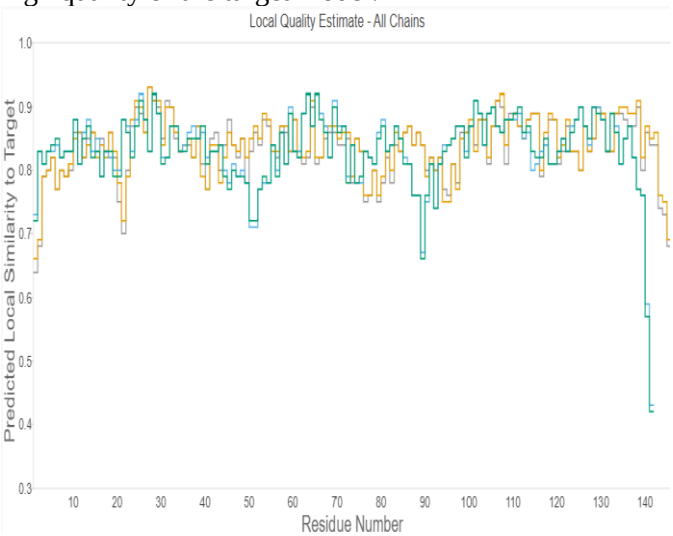


Figure 9: Similarity to target for all chains according to the number of residues

SOLUTIONS AND RECOMMENDATIONS

1. Structures with greater experimental support will yield the biggest gains. The likelihood that modeling assignments will accurately identify the fold will rise as a result of the correctness of the created model.
2. The accuracy of the structure model can also be improved by further progress in the sequence-structure

alignment. The accuracy of the model is increased by a more aligned profile with fewer gaps. Using threading-based strategies to match the sequence to structures discovered through comparative modeling may be one way to improve sequence-structure alignment.

3. The majority of concerns and problems related to determining protein structural coordinates are resolved by protein tertiary structure prediction. When there is little sequence similarity and it is possible to model the target protein using numerous protein structural templates, multiple Template-based homology modeling can be used.

4. By using molecular dynamic simulations, the modeled structure produced by homology modeling, threading, or even ab initio approaches can be improved.

8. Conclusion

In this study, important methods of protein structure prediction are introduced which are three basic approaches based on the query-template identity. We focus on homology modeling to predict the 3D structure of a target sequence using the SWISS model tool. The homology modeling is in a safe zone if the identity is greater than 30%. However, threading is used to predict protein structure when the query Template identity is less than 30%. In comparative modeling, the Target-Template alignment is used firstly to choose an appropriate Template. Once the alignment is correct, the real modeling process which includes backbone creation, side-chain modeling, and loop modeling, followed by Model validation and optimization.

The selected Template and its alignment with the Target sequence play a significant role in the model's accuracy. The target model is evaluated according to the template sequence, which is accepted due to the high identity between them. The identity reached 85.02%, which indicates the minimum error for the target model.

References

AlQuraishi, M. (2019). End-to-End Differentiable Learning of Protein Structure. *Cell Systems*, *8*(4), 292-301.e3. https://doi.org/10.1016/j.cels.2019.03.006

AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, *65*, 1–8. https://doi.org/10.1016/j.cbpa.2021.04.005

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 Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Brownstein, M. J., Khodursky, A. B., Charlie, C., & Michael, J. (2017). *Functional Genomics- Methods in Molecular Biology* (Vol. 224). https://link.springer.com/content/pdf/10.1007%2F978-1-4939-7231-9.pdf

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Eisenberg, D., Bowie, J. U., Lüthy, R., & Choe, S. (1992). Three-dimensional profiles for analysing protein sequence-structure relationships. *Faraday Discussions*, *93*, 25–34. https://doi.org/10.1039/FD9929300025

Harder, T., Boomsma, W., Paluszewski, M., Frellsen, J., Johansson, K. E., & Hamelryck, T. (2010). Beyond rotamers: A generative, probabilistic model of side chains in proteins. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-306

Huang, B., Kong, L., Wang, C., Ju, F., Zhang, Q., Zhu, J., Gong, T., Zhang, H., Yu, C., Zheng, W.-M., & Bu, D. (2023). Protein Structure Prediction: Challenges, Advances, and the Shift of Research Paradigms. *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2022.11.014

Ingraham, J., Riesselman, A., Sander, C., & Marks, D. (2019). Learning protein structure with a differentiable simulator. *7th International Conference on Learning Representations, ICLR 2019*, 1–24.

Jiang, R., Zhang, X., & Zhang, M. Q. (2013). Basics of bioinformatics: Lecture notes of the graduate summer school on bioinformatics of China. *Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China*, *9783642389511*, 1–395. https://doi.org/10.1007/978-3-642-38951-1

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Dos, A., Costa, S., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., & Ai, M. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022.07.20.500902. https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.07.20.500902v1.abstract

Mao, W., Ding, W., Xing, Y., & Gong, H. (2019). AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. *Nature Machine Intelligence*, *2*(1), 25–33. https://doi.org/10.1038/s42256-019-0130-4

Mirabello, C., & Wallner, B. (2019). RAWMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PLoS ONE*, *14*(8), 1–15. https://doi.org/10.1371/journal.pone.0220182

Speed, T. P. (2002). Biological sequence analysis. *Selected Works of Terry Speed*, 1–366. https://doi.org/10.1007/978-1-4614-1347-9_14

Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. In *PLoS Computational Biology* (Vol. 13, Issue 1). https://doi.org/10.1371/journal.pcbi.1005324

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(3), 1496–1503. https://doi.org/10.1073/pnas.1914677117