ESKİŞEHİR TEKNİK ÜNİVERSİTESİ BİLİM VE TEKNOLOJİ DERGİSİ
C- YAŞAM BİLİMLERİ VE BİYOTEKNOLOJİ

Eskişehir Technical University Journal of Science and Technology C- Life Sciences and Biotechnology

2024, 13(1), pp. 31-51, DOI: 10.18036/estubtdc.1378676

RESEARCH ARTICLE

PROTEIN STRUCTURE PREDICTION: AN IN-DEPTH COMPARISON OF APPROACHES AND TOOLS

Elif ALTUNKÜLAH [1], Yunus ENSARİ [1,*]

[1] Bioengineering Department, Faculty of Engineering and Architecture, Kafkas University, Kars, Türkiye

ABSTRACT

Proteins play crucial roles, including biocatalysis, transportation, and receptor activity, in living organisms. Moreover, their functional efficacy is influenced by their structural properties. Determining the three-dimensional structure of a protein is crucial to comprehending its catalytic mechanism, identifying potentially beneficial mutations for industrial applications, and enhancing its properties, including stability, activity, and substrate affinity. Although X-ray crystallography, nuclear magnetic resonance (NMR), and electron microscopy are employed to ascertain protein structures, many researchers have turned to bioinformatics modeling tools because of the high cost and time demands of these techniques. For structure prediction, there are three basic methods: ab initio (de novo), homology-based, and threading-based modeling techniques.

In this study, 11 modeling tools belong to different approaches were compared through modeling of various proteins; *Geobacillus kaustophilus* ksilan alpha-1,2-glucuronidase, *Actinosynnema pretiosum* bifunctional cytochrome P450/NADPH-P450 reductase, human high affinity cationic amino acid transporter 1 (SLC7A), human proton-coupled zinc antiporter (SLC30A) and *Bacillus subtilis* RNA polymerase sigma factor (sigY). Generated models were validated through QMEAN, QMEANDisCo, ProSA, ERRAT and PROCHECK tools. All of the studied proteins could be successfully modeled using homology modeling techniques, while some of the proteins could not be effectively modeled using threading or ab initio-based methods. YASARA generated reliable models for proteins that contain heteroatoms, such as P450 monooxygenases, because other tools exclude heteroatoms in their produced structures. Among approaches for modeling without templates, AlphaFold is a potent tool. On the other side, well-known template-based tools like YASARA, Robetta, and SWISS-MODEL have arisen. These results will help scientists choose the best protein modeling strategy and tool to guarantee high-quality structures.

Keywords: Protein structure prediction, Homology modeling, Alphafold, YASARA, Ab initio modeling

1. INTRODUCTION

The interatomic angles, folding-loop motives, and ultimately the three-dimensional structures of the proteins are determined by their sequences and complex interactions with amino acids, which are the building blocks of proteins. [1] While being synthesized in the ribosomes of cells, enzymes acquire three-dimensional structures with distinct folding patterns. These three-dimensional structures are critical for proteins to perform their biological functions, and misfolding can result in enzyme dysfunction or structural disorders in the organism. [2] Predicting the three-dimensional structures of proteins with known sequences and revealing their patterns is critical for understanding their functional mechanisms and improving the properties such as activity and stability. [3–5] Although traditional methods such as X-ray crystallography, nuclear magnetic resonance (NMR), and electron microscopy can be used to determine protein structures, their high cost and time requirements necessitated the development of alternative methods. [6] Various approaches for protein structure prediction in bioinformatics studies have been developed, ranging from databases using basic statistical methods to artificial intelligence and deep learning algorithms. [7] The first examples of computational protein structure prediction were based on detecting amino acid affinities for folding. In the following years, more successful results in structure estimation were obtained using techniques such as the calculation of free energy levels and the use of multiple alignment methods. [8] There are three main approaches

*Corresponding Author:yunusensari@kafkas.edu.tr
Received: 20.10.2023 Published:30. 01.2024

for structure prediction: ab initio (de novo), homology-based, and threading-based modeling methods. They may also be divided into template-based modeling and template-free modeling categories. [9] With the aid of computer-assisted protein structure prediction and design techniques, which are widely used today, it is possible to understand the secondary and tertiary structures of proteins based on their amino acid sequences.

De novo or ab initio modeling is based on estimating the most likely low-energy conformation that the amino acid sequence may have. [10] Although the accuracy of the models obtained is low in ab initio modeling, which is a basic approach in which protein structure is estimated based on physicochemical properties, successful results can be obtained in modeling proteins with shorter than 100 amino acid sequences. [11] To reduce computation time, probabilistic and predictive approaches have led to the use of ab initio methods as techniques for revealing folding patterns of small proteins rather than exact structure prediction. [12] In the future, particle-based methods that provide coarse-grained predictions, such as Monte Carlo simulation, have developed to overcome these constraints. [11] In a study by Liwo et al. on optimization of potential energy functions, <6Å models were created with root mean square deviations of protein fragments up to 61 amino acids. [13] Simon et al. successfully modeled 73 of 172 target proteins up to 150 amino acid sequences using ROSETTA. [14] TOUCHSTONE II software was used to generate folding patterns for 83 of 125 target proteins with up to 174 amino acid sequences. [15] Bradley et al. used atomic ROSETTA to predict high-resolution models on sequences of less than 85 amino acids, employing a combination of structural sampling methods, methods that calculate the packaging of protein nuclei in detail, and high-resolution structure prediction. [16] In another study, the I-TASSER method was developed, which allows the accurate detection of folding regions in small proteins by iterative application of the TASSER method. [17] Different algorithms designed to improve the accuracy of methods for estimating the spatial structure with the lowest free energy allow for greater accuracy in predicting the three-dimensional structures of larger proteins. Rashid et al.'s random-start strand method, developed in 2013, greatly improved the results of single-point searches using a three-dimensional 100-center cubic lattice. [18] Studies have also been conducted to improve the quality of the fragment libraries based on the ratio of the number of segments close to the main backbone of the protein to the accuracy of the structure prediction. In a study that targeted particles recorded on templates with similar structural information in order to reduce the size of the conformational search space, the accuracy rate of the models obtained after classification was found to be 7% higher than that of standard piece-based estimators. [19] End-to-end learning and attention-based networks are used by AlphaFold; therefore, models with high confidence can be created. A study using AlphaFold2, an ab initio modeling tool, by Akdel et al. demonstrated that this tool can provide models with accuracy close to experimental data. [20] The accuracy of ab initio tools is limited and they require considerable computing resources to explore the wide range of protein conformations. This is why, up to now, only the structures of small proteins have been successfully predicted with this approach. [21]

In homology-based approaches, also known as comparative modeling, experimentally determined structures that are topologically equivalent to the target protein are used. [22] With the development of remote homology detection methods based on pairwise comparison of protein sequences, the use of tools that can obtain reliable and sensitive results in protein structure prediction has become widespread. [23] Finding the best pattern is the first step in homology modeling. Then the target and template sequences are aligned, the framework is built, and finally the model is evaluated. [24] The accuracy of homology models based on the structure of distantly related or spatially equivalent proteins is proportional to the alignment quality. [22]

Threading-based approaches are based on solvent availability and the secondary structure of proteins. These methods use the effect of different amino acid alternatives on structural coiling motives and align the questioned protein sequence with previously resolved similar template protein sequences and model them according to statistical probabilities and energy calculations. [25] Proteins acquire their three-dimensional natural structure according to distant interactions between amino acids. In threading-based approaches, the data of these interactions are transferred to a scoring system and three-dimensional models are created by overlapping the structure information with the existing template models. [26,27] These approaches can be thought of as an intersection of ab initio and homology-based approaches. Just like in ab initio design, threading-based methods use energy minimization. [13] Threading methods, like homology-based modeling, aim to create a model by predicting the curling of the query protein sequence based on previously defined patterns. However, both methods ignore the possibility of the protein folding in a random conformation with a previously unknown template. [15] The main distinction between threading-based methods and comparative methods is that threading-based methods can model without requiring structure knowledge of homologous sequences. Threading uses sequence-to-structure alignment, whereas comparative methods require sequence-to-sequence alignment. [12] Since threading-based techniques focus on structural similarity, they are successful in recognizing sequence-structure pairs with similarity in folding motifs, but may be insufficient in recognizing homologous pairs. [28] In this study, six proteins from different protein classes were modeled using eleven different modeling tools belonging to three approaches. The generated models were then evaluated and compared in terms of model quality, RMSD and visually.

2. MATERIALS AND METHODS

Geobacillus kaustophilus ksilan alpha-1,2-glucuronidase (Accession Number: KJE27682), *Actinosynnema pretiosum* subsp. auranticum bifunctional cytochrome P450/NADPH-P450 reductase (Accession Number: Q8KUI0), human high affinity cationic amino acid transporter 1 (Accession Number: P30825) (SLC7A), human proton-coupled zinc antiporter (Accession Number: Q9Y6M5) (SLC30A) and *Bacillus subtilis* RNA polymerase sigma factor (sigY) protein sequences were used for modeling.

2.1. Ab Initio (De novo) Based Modeling Tools

The Alphafold, BhageerathH+, and RaptorX tools were used for *ab initio*-based modeling of target proteins. AlphaFold is a neural network-based structure prediction tool that enables modeling based on physical and biological data when data on similar protein structures are not available. [29] AlphaFold, an artificial intelligence tool based on deep learning principles, produces modeling results that are very close to experimental studies. It predicts structure by constructing new neural network architectures within the framework of the evolutionary, physical, and geometric rules of protein structure. [30] Another prediction tool, BhageerathH+, is an energy-based application for structure prediction of small globular proteins. [31] BhageerathH+ provides modeling results based on *ab initio* modeling. [32] The sequence in FASTA format is converted to PDB format, trial models are created, energy minimization is performed after steric mismatches are eliminated by passing through biophysical filters, the models with the lowest energy are selected, filtered, and the results are ordered. [31] RaptorX is a tool that focuses on solving the therading problem with the linear integer programming method, using *ab initio* methods to generate the unaligned loop regions and the final model. [12,33] RaptorX can estimate structural elements as well as the amino acid ratios that contribute to the disordered conformational randomness of the secondary structure. [34] The model is generated by energy optimization after obtaining the set of all applicable solutions. [25]

2.2. Homology Based Modeling Tools

SWISS-MODEL, IntFOLD, Phyre^2, Robetta, ModWeb and YASARA tools were used for homology based modeling of proteins. SWISS-MODEL was the first online modeling tool. [35] The SWISS-MODEL, which performs homology-based modeling; consists of five basic steps: amino acid sequence input, pattern search, pattern selection, model building, and quality estimation. [36] Protein sequences or UniProt accession code can be used directly , as well as target-pattern sequence alignment via a manually or automatically selected template. [37]

Another homology modeling tool, Robetta, generates several alignment alternatives based on features such as the requirement of a region in the protein sequence for folding. The best models are selected by combining criteria such as alignment compromise, hydrophobic embedding measure, low and high resolution energy functions. [22] IntFOLD, which allows the estimation of amino acids in the binding sites as well as access to the structure information from the protein sequence, is one of the alternative tools that provides the revealing of the relationship between the structure and function of the protein. [38] IntFOLD generates folding libraries during analysis by using multiple templates during model building. [32,39] Phyre^2, another online application for predicting the secondary and tertiary structures of proteins using a homology-based analysis method, provides users with both *ab initio* modeling from amino acid sequences and manual pattern generation. The tool's modeling method consists of collecting homologous sequences, scanning the folding library for known folding patterns, modeling loops and side chain placement. [40]

ModWeb, yet another comparative protein structure modeling tool, works on the principle of aligning the PSI-BLAST sequence profile of the target sequence with template sequences extracted from the Protein Data Bank and comparing them to select the best model. [26] Models are created by aligning one or more FASTA-formatted sequences to the best patterns using ModPipe. In addition, ModWeb allows users to define and model all homologous sequences in the UniProtKB database by using a protein structure and sequence profile as input. [41]

YASARA, generates hybrid models, is a molecular modeling tool with a visualization algorithm that performs a lattice-based neighbour search in conjuction with unbound force calculations at each step of the simulation without generating pair lists. [42,43] YASARA; is used to display models created by other methods as well as molecular models in appropriate formats [44] and for docking to understand the protein-ligand interaction. [45]

2.3. Threading Based Modeling Tools

The C-I-TASSER and LOMETS tools were used for threading based modeling of proteins. The C-I-TASSER tool was developed by combining I-TASSER's fragment assembly simulations with inter-aminoacid contact maps from deep neural network learning for modeling the folding motifs of non-homologous proteins. [46] C-I-TASSER, which offers the opportunity to make successful models in the structure prediction of proteins -especially when there is not any template models are available- finds structure patterns through LOMETS using contact maps and atomic models collected with Monte Carlo simulations, and creates the final model. [9,46] Another threading-based application, LOMETS, is a local meta-threading application combining nine different threading servers. LOMETS scans a library of patterns at various resolutions obtained by different methods such as X-ray crystallography, electron microscopy and NMR spectroscopy. The resulting patterns are evaluated based on query-sequence-to-pattern alignment scores using threading methods. [9] LOMETS also includes C_α atom and side chain contact distance maps assembled as a result of threading alignments and guides applications such as MODELLER, ROSETTA, TASSER. [47]

2.4. Model Quality Determination

QMEAN, QMEANDisCo, ProSA, ERRAT and PROCHECK tools were used to determine the quality of the protein models. QMEAN (Qualitative Model Energy Analysis - Qualitative Model Energy Analysis) model quality detection tool of the SWISS-MODEL server is an application that compares and scores the geometric properties of the model (dual atomic distances, rotation angles, all atom interaction, solvent accessibility, etc.) with statistical data obtained from experimental structures. With QMEAN, it is possible to both measure the overall reliability of the model and determine the local quality per amino acid. [48] In QMEAN, each amino acid is scored between 0 and 1 by calculating the statistical values of the potential mean strength in terms of similarity to the natural structure. The higher the similarity, the higher the model reliability and score. [36,37] The Z-score is calculated by comparing the QMEAN score to the distributions obtained from the high-resolution structures resolved by X-ray chromatography. QMEANDisCo, another quality detection application developed over QMEAN calculations, is a tool that evaluates the distance localization of experimentally determined protein structures homologous to the model under consideration. The accuracy of the results obtained is proportional to the number of homologs of the query model. QMEANDisCo scoring cannot provide reliable results in protein models with few or no homologs. [49]

ProSA-web (Protein Structure Analysis) is another online tool used to validate protein models. It compares the query model to results from X-ray analysis, NMR spectroscopy, or theoretical calculations. The tool computes the model's structural energy and displays it as a Z-score and an amino acid energy graph. [50] The Z-score indicates the overall model quality and measures the deviation of the total energy of the structure according to an energy distribution derived from random conformations. [51,52] Positive values of the Z-score may indicate that the model is problematic or inaccurate. [50]

Model errors are caused by three major factors: misdirection of amino acids due to backbone linkages, errors in alignment or misregistration of amino acids, and side chain misplacement. [53] To detect faulty areas, various techniques are used. Ramachandran analysis of peptide dihedral angles is the first of these methods, and it is based on the classification of allowed and disallowed conformations. [54] Protein folding is defined by the φ (phi), ψ (psi) and ω (omega) angles of the backbone loops. Among them, the allowed loop options of angle ω are quite limited. [55] Ramachandran analysis is based on the principle of constructing two-dimensional scatter plots of other φ and ψ angles and comparing them with a predicted distribution. [56] By analyzing the statistics of unbound interactions between different types of atoms, ERRAT calculates the quality factor by plotting the data obtained as a result of the calculations, the value of the error function against a sliding window position of 9 amino acids. [53]

The SAVES developed by UCLA-DOE-LAB is an online verification tool that includes PROCHECK and ERRAT calculates Ramachandran plots and scores model quality respectively. The PROCHECK tool generates the graphs based on the comparison of stereochemical parameters of the given protein against similar patterns of known structure. [57] These parameters are stereochemical criteria used to determine the quality of a structure. [58] The obtained Ramachandran graphs show φ and ψ twist angles for all amino acids in the query protein structure except the chain ends. Because glycine amino acids are incompatible with other side chain types, they are depicted as independent triangles. The dark red regions shown in Ramachandran plots are identified as "nuclei". In these regions, amino acids with optimal angles are marked, and more than 90% of the amino acid sequence of an ideal model would be expected to be found. [58]

3. RESULTS AND DISCUSSION

Geobacillus kaustophilus ksilan alpha-1,2-glucuronidase, complete sequence and heme domain of *Actinosynnema pretiosum subsp. auranticum* bifunctional cytochrome P450/NADPH-P450 reductase, human high affinity cationic amino acid transporter 1 (SLC7A1), human proton-coupled zinc antiporter (SLC30A1), and *Bacillus subtilis* RNA polymerase sigma factor (sigY) protein models were created using the AlphaFold, BhageerathH+, C-I-TASSER, IntFOLD, LOMETS, ModWeb, Phyre^2, RaptorX, Robetta, SWISS-MODEL and YASARA modeling tools. The aforementioned proteins are members of different protein families. While *Geobacillus kaustophilus* ksilan alpha-1,2-glucuronidase and *Actinosynnema pretiosum subsp. auranticum* bifunctional cytochrome P450/NADPH-P450 reductase are enzymes, SLC7A1 and SLC30A1 are membrane proteins and sigY is a regulatory protein that controls the transcription. Since AlphaFold models are accessed through the database, AlphaFold models for all proteins were downloaded from UniProt except the heme domain (the catalytic domain of the enzyme) of the bifunctional cytochrome P450/NADPH-P450 reductase since the model of the full protein exist in the database. BhageerathH+ failed during fragment assembly and *ab initio* loop sampling of high affinity cationic amino acid transporter 1 and proton-coupled zinc antiporter proteins. The RaptorX built model for only the *Geobacillus kaustophilus* ksilan alpha-1,2-glucuronidase.

The models were validated using the ERRAT, PROVE, ProSA, SWISS-MODEL QMEAN, and QMEANDisCo tools. In order to compare the modeling tools, the models with the highest accuracy of the tools that output more than one model were selected based on the results of the validation. The models with the highest ERRAT quality score were selected and compared with the models obtained from the other tools. In addition, RMSD values were calculated from the model-to-model comparisons. The ERRAT quality scores of all generated models, local quality estimation tables of QMEAN and QMEANDisCo tools, Z-PLOTs, and Ramachandran plots of the best models of each tool are shown in the Supporting Information.

3.1. 3D Modeling of *Geobacillus kaustophilus* xylan alpha-1,2-glucuronidase

From the modeling of the xylan alpha-1,2-glucuronidase protein, composed of 679 amino acids, using various tools, one structure prediction was obtained from AlphaFold, ModWeb, Phyre2, and YASARA, while two were obtained from the SWISS-MODEL tool. Additionally, five structure predictions were acquired through the use of BhageerathH+, C-I-TASSER, IntFOLD, LOMETS, RaptorX, and Robetta tools. All other tools generated models with 679 residues, while the Phyre2 generated model with 677, ModWeb 675, and SWISS-MODEL 677 residues (Table 1). All generated models were visualized using PyMOL and shown in Figure 2. Additionally, Table 1 shows the ERRAT quality scores, QMEAN, QMEANDisCo values, and Z-Scores. The YASARA tool generated the best model in terms of ERRAT Quality Score (98.36), while the SWISS-MODEL tool generated the best model in terms of QMEAN and QMEANDisCo scores (0.06 and 0.94 respectively). Furthermore, most of the models have ERRAT score above 90 and all models except RaptorX have QMEANDisCo score close to 1. All models have similar Z-values and Z-Plots (Figure S3) which are within the range of scores for similarly sized native proteins. Upon examination of the QMEAN graphs, it is evident that the low confidence regions among all models are quite similar. Analysis of Ramachandran plots (Figure S4) and statistics reveals that the number of amino acids residing in disallowed regions vary across the AlphaFold, LOMETS and YASARA models was 1, IntFOLD was 2 and the model obtained by SWISS-MODEL was 4. The ratio of the residues located in the most favored regions varies between 79.8 to 93.4. All of the protein models exhibit nearly identical structures based on their topology. This is associated with the RMSD values obtained from comparing the structural characteristics of each model. Notably, the RaptorX model displays a high RMSD in comparison to the other models, which indicates that there are some notable differences in the RaptorX model.

Table 1. ERRAT Quality Scores, QMEAN, QMEANDisCo Values, and Z-Scores of *Geobacillus kaustophilus* ksilan alpha-1,2-glucuronidase modeling.

Approach	**Tool**	**Amino acid number in model**	**ERRAT Quality Score**	**QMEAN**	**QMEANDisCo**	**Z-SCORE**
Ab initio	***AlphaFold***	679	96.42	0.40	0.93	-11.90
	BhageerathH+	679	92.85	-1.18	0.90	-11.55
	RaptorX	679	87.16	-2.56	0.61	-12.07
Threading	***C-I-Tasser***	679	95.37	-2.42	0.92	-11.91
	LOMETS	679	93.89	-0.72	0.93	-11.96
Homology based	***IntFOLD***	679	91.21	-0.86	0.91	-11.62
	ModWeb	675	89.51	-0.56	0.92	-11.79
	Phyre^{2}	677	90.28	-0.18	0.91	-11.53
	Robetta	679	96.42	0.72	0.93	N.C
	SWISS-MODEL	676	93.69	0.06	0.94	-11.95
	YASARA	679	98.36	-0.26	0.89	-11.70

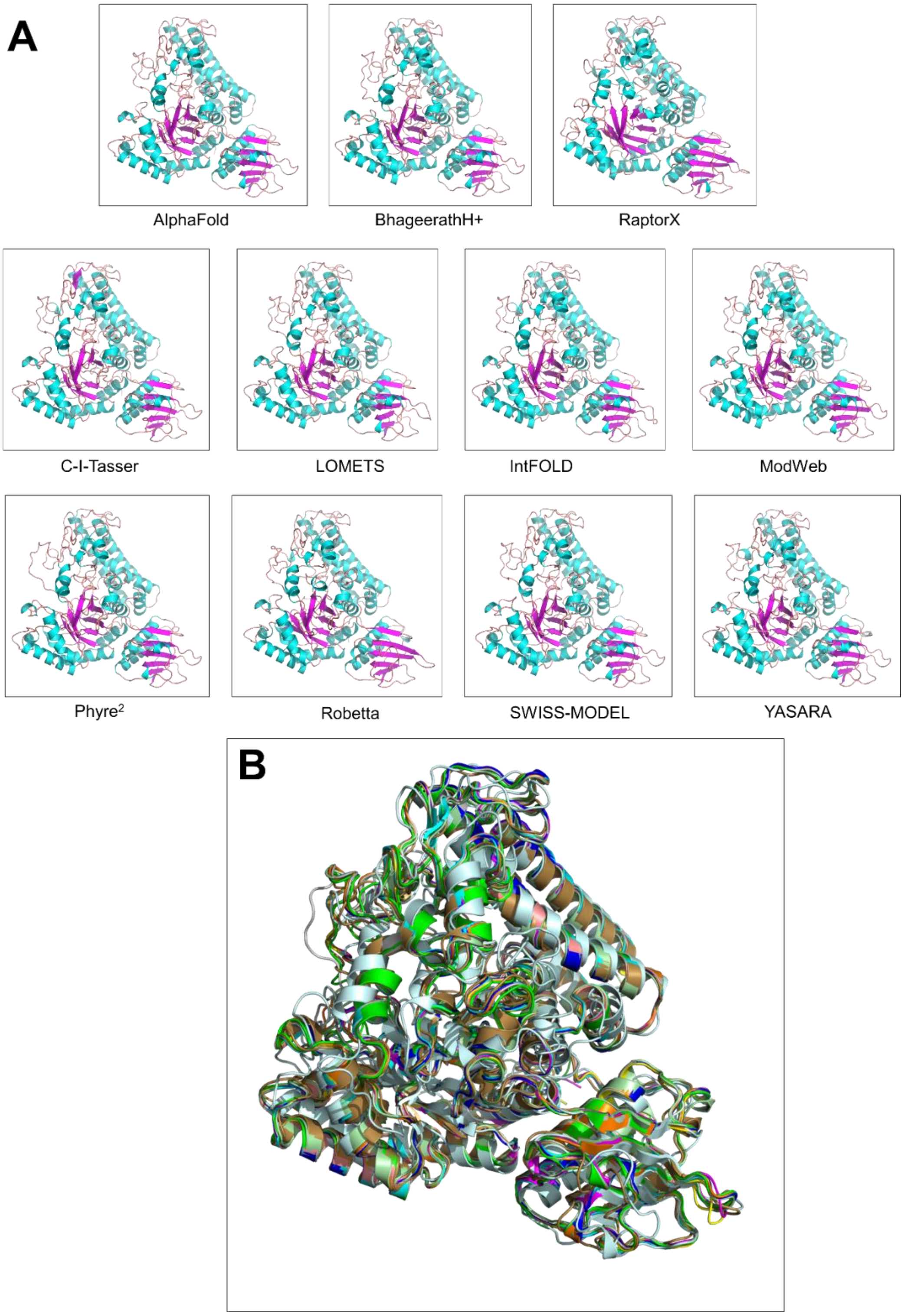

Figure 1. 3D models of *Geobacillus kaustophilus* ksilan alpha-1,2-glucuronidase protein. A. Cartoon representation of models generated by different tools (Cyan shows a-helix's, magenta shows b-sheets, and salmon color shows loops). B. Overlaid view of all generated models. Blue; AlphaFold, green; BhageerathH+, pale cyan; RaptorX, cyan; C-I-Tasser, magenta; LOMETS, yellow; IntFOLD, salmon color; ModWeb, grey; Phyre2, sand color; Robetta, orange; SWISS-MODEL, pale green; YASARA.

Table 2. RMSD values of model-to-model comparisons for *Geobacillus kaustophilus* ksilan alpha-1,2-glucuronidase protein.

Approach	**RMSD**	***AlphaFold***	***BHAGEERATHH+***	***RaptorX***	***C-I-Tasser***	***LOMETS***	***IntFOLD***	***ModWeb***	***Phyre²***	***Robetta***	***SWISS-MODEL***	***YASARA***
Ab initio	***AlphaFold***											
	BhageerathH+	1.51										
	RaptorX	4.79	5.09									
Threading	***C-I-Tasser***	1.46	2.25	5.05								
	LOMETS	1.19	2.00	5.02	1.78							
Homology based	***IntFOLD***	1.09	1.49	4.77	1.59	1.25						
	ModWeb	0.97	1.40	4.75	1.51	1.28	0.86					
	Phyre²	1.51	1.74	4.85	1.65	1.48	1.57	1.57				
	Robetta	2.10	2.69	4.90	2.40	2.38	2.15	2.07	2.23			
	SWISS-MODEL	0.84	1.29	4.77	1.48	1.15	1.02	1.00	1.41	2.13		
	YASARA	1.46	2.22	5.08	2.04	1.90	1.60	1.43	1.70	2.50	1.37	

3.2. Bifunctional Cytochrome P450/NADPH-P450 Reductase Models

Since the heme domain of P450 monooxygenases refers to the catalytic domain, both the heme domain and full protein were modeled separately. First, we modeled the heme domain of the bifunctional cytochrome P450/NADPH-P450 reductase, which consists of 482 amino acids, using 11 different tools. We retrieved the AlphaFold structures from the UniProt database. However, since the heme domain is part of the full sequence, the AlphaFold structure was not available in the database. Additionally, the RaptorX tool failed for modeling. On the contrary, one structure was derived from the utilization of Phyre2, SWISS-MODEL, and YASARA tools, whereas three were obtained from the ModWeb tool. Additionally, the remaining tools each yielded five structure predictions. All other tools output models comprising 482 amino acid sequences, while the ModWeb models comprise 466, Phyre^2 457 and SWISS-MODEL 462 residues. The models generated by C-I-Tasser, Robetta, and YASARA exhibited the highest ERRAT quality scores (Table 3). Based on the ERRAT, QMEAN and QMEANDisCo scores, homology based methods have higher quality scores compared to ab-initio and threading based methods. All models' Z-Plots are in the range of scores for similarly sized native proteins except for the model generated by BhageerathH+ (Figure S7). The Z-Score and Z-Plot for the models obtained using the Robetta tool were not calculable. The SWISS-MODEL tool produced the model with the highest QMEANDisCo score. When the QMEAN Local Quality Estimation plots (Figure S5 and S6) were examined, it was discovered that the positions of the heme domain's beginning and ending amino acids were the least reliable regions of all models. Based on the Ramachandran Plots (Figure S8), models generated by BhageerathH+, C-I-Tasser, and LOMETS possessed the lowest quality, with residues located in the most favored regions ranging between 70-85 %. When visualizing the generated models in Pymol, it was clearly seen that BhageerathH+ generated the least favorable model due to numerous secondary structure elements, including α-helix and β-sheets, not being modeled (Figure 2). However, the rest of the models share similar folding and low RMSD values except BhageerathH+ and LOMETS. Furthermore, P450 monooxygenases are heme containing enzymes and thus, structure should have heme molecule in the structure. Only the model generated by YASARA has the heme molecule and the remaining models lack the heteroatom in their final structure. As a result, YASARA yielded the best model for the heme domain of the bifunctional cytochrome P450/NADPH-P450 reductase, based on quality scores and heteroatom feature.

Table 3. ERRAT Quality Scores, QMEAN, QMEANDisCo Values, and Z-Scores of bifunctional Cytochrome P450/NADPH-P450 reductase heme domain modeling.

Approach	Tool	Amino acid number in model	ERRAT Quality Score	QMEAN	QMEANDisCo	Z-SCORE
Ab initio	***AlphaFold***	-	-	-	-	-
	BhageerathH+	482	46.51	-11.15	0.35	-4.26
	RaptorX	-	-	-	-	-
Threading	***C-I-Tasser***	482	96.19	-5.33	0.74	-10.54
	LOMETS	482	82.87	-5.72	0.63	-11.02
Homology based	***IntFOLD***	482	77.92	-2.59	0.73	-10.62
	ModWeb	458	77.11	-2.75	0.72	-10.66
	Phyre^2	457	65.70	-3.63	0.69	-10.79
	Robetta	482	95.98	0.58	0.75	NA
	SWISS-MODEL	462	91.69	-1.94	0.76	-10.98
	YASARA	482	95.72	-1.53	0.72	-10.59

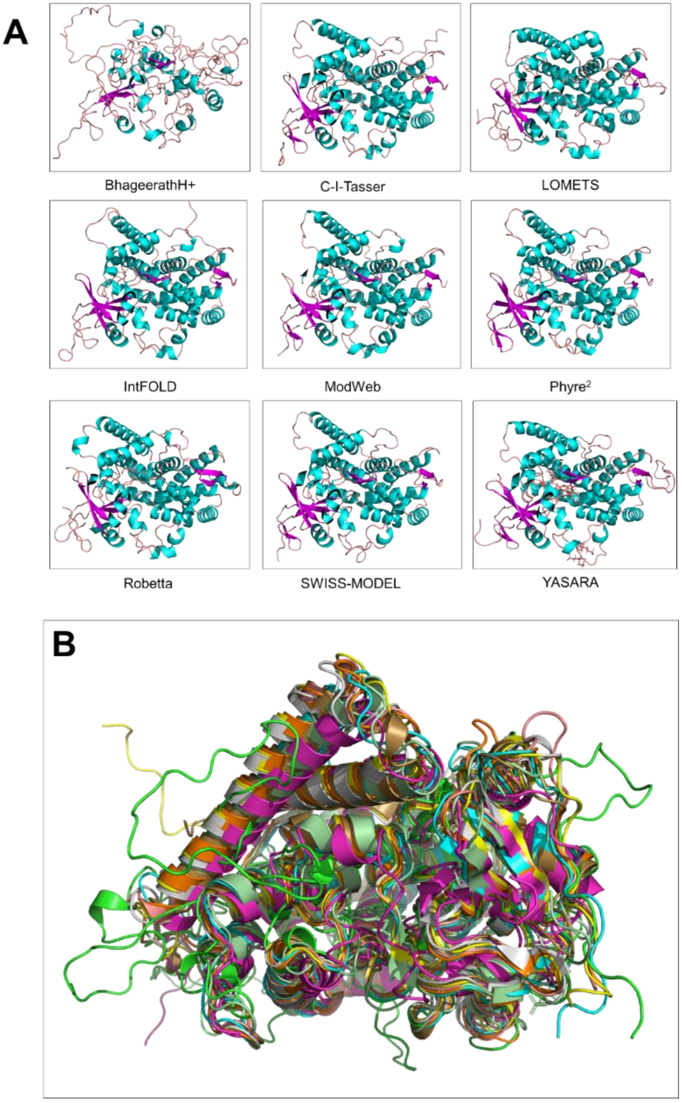

Figure 2. 3D models of bifunctional cytochrome P450/NADPH-P450 reductase heme domain protein. A. Cartoon representation of models generated by different tools (Cyan shows a-helix's, magenta shows b-sheets, and salmon color shows loops). B. Overlaid view of all generated models. Green; BhageerathH+, cyan; C-I-Tasser, magenta; LOMETS, yellow; IntFOLD, salmon color; ModWeb, grey; Phyre^2, sand color; Robetta, orange; SWISS-MODEL, pale green; YASARA.

Table 4. RMSD values of model-to-model comparisons for bifunctional cytochrome P450/NADPH-P450 reductase heme domain protein.

Approach	***RMSD***	***AlphaFold***	***BHAGEERATHH+***	***RaptorX***	***C-I-Tasser***	***LOMETS***	***IntFOLD***	***ModWeb***	***Phyre2***	***Robetta***	***SWISS-MODEL***	***YASARA***
Ab initio	***AlphaFold***											
	BhageerathH+											
	RaptorX											
Threading	***C-I-Tasser***		21.79									
	LOMETS		21.25		5.96							
Homology based	***IntFOLD***		21.44		3.95	6.43						
	ModWeb		21.48		2.69	4.19	2.24					
	Phyre^2		21.46		2.93	4.16	2.27	2.63				
	Robetta		21.55		4.35	5.18	3.99	2.87	2.95			
	SWISS-MODEL		21.53		2.41	4.24	2.16	2.31	2.43	2.78		
	YASARA		21.85		4.78	6.08	6.62	3.14	3.46	5.02	3.23	

The full sequence of the bifunctional Cytochrome P450/NADPH-P450 reductase enzyme containing 1005 residues was also modeled. Out of eleven tools, six generated models for the target sequence. However, all ab-initio and threading based tools, which are template independent tools, except AlphaFold, failed to generate a model. Usually, these tools have sequence length limitations. When analyzing the obtained six structural models, the models generated by AlphaFold, Phyre^2, Robetta and SWISSMODEL comprised the almost entire sequence on the structure. Nevertheless, ModWeb generated the model with 522 residues and YASARA generated the model with 642 residues which contains the heme and FMN binding domain. AlphaFold, Robetta, SWISSMODEL, and YASARA models have higher quality scores (Table 5) compared to ModWeb and Phyre^2. When evaluating the QMEAN scores, Phyre^2 has below -4 which indicates low quality. And similarly, ModWeb has -3,56 which is close to -4 and it has also low quality. The QMEAN score for YASARA is -1,17 this is lowered because of the lower quality of some residues around 480 which lie in the linker region between heme and FMN domain. However, aside from this linker region, the rest of the model's QMEAN score is better. However, the model obtained from AlphaFold is the only model which is out of the range in the Z-Plots (Figure S11). All generated models have high percentage of residues located in the most favored regions on their Ramachandran plots (Figure S12). Visual inspection of the models reveals low folding similarities among the models, resulting in high RMSD values. Regarding heteroatom composition, the model generated by YASARA contains both heme and FMN molecules, while SWISSMODEL generated model contains only the heme molecule. The remaining four models lack heteroatoms as observed in the heme domain modeling.

Table 5. ERRAT Quality Scores, QMEAN, QMEANDisCo Values, and Z-Scores of bifunctional Cytochrome P450/NADPH-P450 reductase full sequence modeling.

Approach	**Tool**	**Amino acid number in model**	**ERRAT Quality Score**	**QMEAN**	**QMEANDisCo**	**Z-SCORE**
Ab initio	***AlphaFold***	1005	93.139	0.40	0.71	-16.22
	BhageerathH+	-	-	-	-	-
	RaptorX	-	-	-	-	-
Threading	***C-I-Tasser***	-	-	-	-	-
	LOMETS	-	-	-	-	-
Homology based	***IntFOLD***	-	-	-	-	-
	ModWeb	522	62.840	-3.56	0.63	-10.5
	Phyre^2	1000	69.596	-4.20	0.69	-14.61
	Robetta	1005	96.3	1.04	0.70	-16.31
	SWISS-MODEL	1000	93.598	-2.57	0.73	-15.01
	YASARA	642	97.078	-1.17	0.71	-12.83

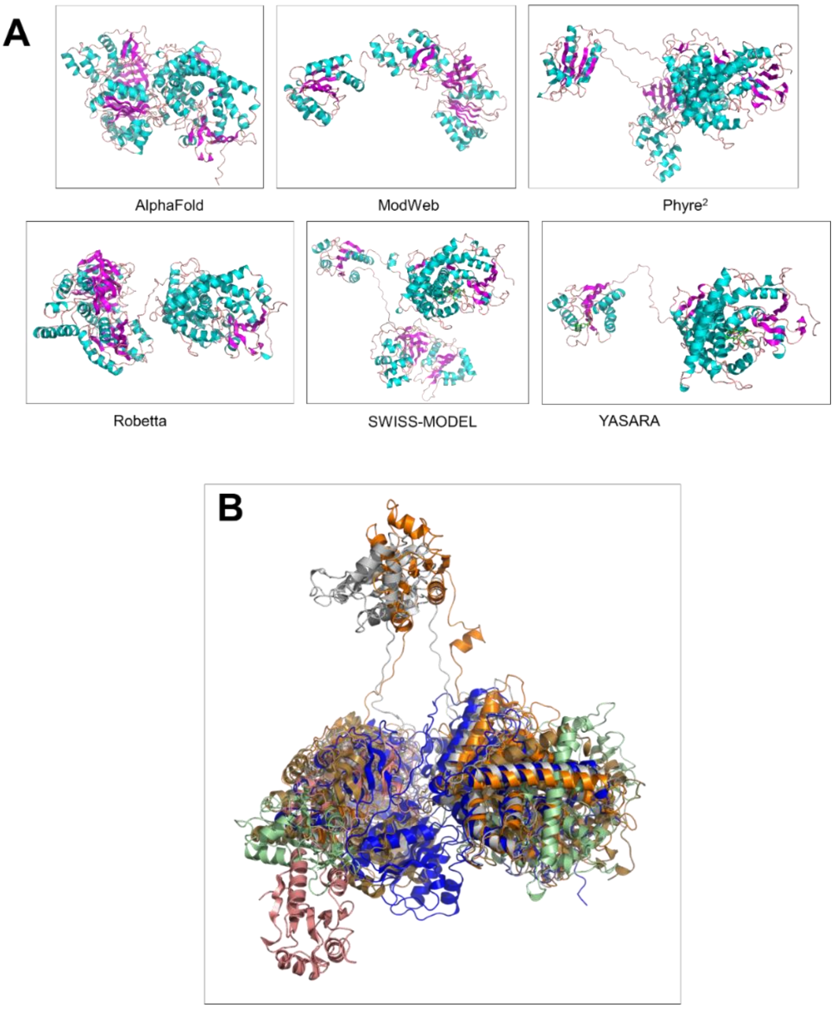

Figure 3. 3D models of bifunctional cytochrome P450/NADPH-P450 reductase full sequence protein. A. Cartoon representation of models generated by different tools (Cyan shows a-helix's, magenta shows b-sheets, and salmon color shows loops). B. Overlaid view of all generated models. Blue; AlphaFold, salmon color; ModWeb, grey; Phyre^2, sand color; Robetta, orange; SWISS-MODEL, pale green; YASARA.

Table 6. RMSD values of model-to-model comparisons for bifunctional cytochrome P450/NADPH-P450 reductase full sequence protein.

Approach	**RMSD**	***AlphaFold***	***BHAGEERATHH+***	***RaptorX***	***C-I-Tasser***	***LOMETS***	***IntFOLD***	***ModWeb***	***Phyre^2***	***Robetta***	***SWISS-MODEL***	***YASARA***
Ab initio	***AlphaFold***											
	BhageerathH+											
	RaptorX											
Threading	***C-I-Tasser***											
	LOMETS											
Homology based	***IntFOLD***											
	ModWeb	57.97										
	Phyre^2	78.83						89.57				
	Robetta	54.88						79.12	69.86			
	SWISS-MODEL	64.02						62.77	74.50	96.38		
	YASARA	63.78						6.38	67.54	99.59	8.04	

3.3. High Affinity Cationic Amino Acid Transporter 1 Models (SLC7A1)

Ten tools successfully generated models for the high affinity cationic amino acid transporter protein, which consists of 629 residues. It is difficult to determine membrane proteins structure experimentally, thus computational prediction is a promising approach. [21] While, one structure prediction was obtained from AlphaFold, Phyre^2, and YASARA tools. 2 from SWISS-MODEL, 3 from the ModWeb tool and 5 structure predictions were obtained from the remaining tools. ModWeb, Phyre^2, and SWISS-MODEL generated models consisting of 247, 451, and 589 residues, respectively. The remaining tools generated models with entire sequence. ERRAT quality scores, QMEAN, QMEANDisCo values, and Z-Scores of all models are shown in Table 7. Of the models, Robetta, YASARA, and AlphaFold had ERRAT quality scores exceeding 90 and their QMEAN values were higher than -4. Conversely, the remaining tools had QMEAN values below -4 indicating a low quality model. Furthermore, Robetta, YASARA, and AlphaFold had the highest QMEANDisCo scores which were approximately 0.6. According to Z-Plots, the scores for models generated by YASARA, and AlphaFold are out of the range of typically observed for native proteins of similar size determined by X-Ray and NMR. Additionally, the Z-Score and Z-Plot for Robetta model was not calculated. Based on Ramachandran Plots (Figure S16), more than 90% of residues in models generated by AlphaFold and YASARA are in the most favored regions. Generated ten models showed varying topology, but, AlphaFold, LOMTES, Phyre^2, Robetta, and SWISS-MODEL share more or less similar folding structures (Figure 4). This similarity was also confirmed through RMSD calculation and AlphaFold, Phyre^2, and Robetta have lowest RMSD values which indicates the folding similarity (Table 8). As a conclusion, all tested tools did not generate reliable 3D model for the SLC7A1, which is a membrane transporter protein, based on different quality parameters.

Table 7. ERRAT Quality Scores. QMEAN. QMEANDisCo Values. and Z-Scores of High Affinity Cationic Amino Acid Transporter 1 modeling.

Approach	Tool	Amino acid number in model	ERRAT Quality Score	QMEAN	QMEANDisCo	Z-SCORE
Ab initio	***AlphaFold***	629	94.79	-2.37	0.64	-6.78
	BhageerathH+	629	15.58	-14.23	0.22	0.91
	RaptorX	-	-	-	-	-
Threading	***C-I-Tasser***	629	81.48	-11.68	0.57	-3.56
	LOMETS	629	84.33	-5.93	0.55	-6.23
Homology based	***IntFOLD***	629	73.29	-9.18	0.58	-4.58
	ModWeb	247	79.83	-7.98	0.28	-1.96
	Phyre^2	451	88.18	-5.80	0.71	-3.51
	Robetta	629	99.19	-2.08	0.61	N.C.
	SWISS-MODEL	589	85.38	-7.50	0.61	-4.65
	YASARA	629	96.93	-3.83	0.60	-5.30

Table 8. RMSD values of model-to-model comparisons for High Affinity Cationic Amino Acid Transporter 1 protein.

Approach	RMSD	*AlphaFold*	*BHAGEERATHH+*	*RaptorX*	*C-I-Tasser*	*LOMETS*	*IntFOLD*	*ModWeb*	*Phyre^2*	*Robetta*	*SWISS-MODEL*	*YASARA*
Ab initio	***AlphaFold***											
	BhageerathH+	30.88										
	RaptorX											
Threading	***C-I-Tasser***	15.75	36.02									
	LOMETS	10.40	31.13		16.20							
Homology based	***IntFOLD***	19.30	28.09		25.98	18.61						
	ModWeb	28.32	28.62		27.69	28.83	28.40					
	Phyre^2	4.00	23.70		2.50	4.73	3.07	23.05				
	Robetta	4.66	30.75		15.63	10.66	19.13	28.16	3.95			
	SWISS-MODEL	19.58	26.07		23.80	19.09	19.40	29.85	2.55	19.81		
	YASARA	27.56	35.30		33.34	29.74	34.31	29.42	4.96	28.18	23.77	

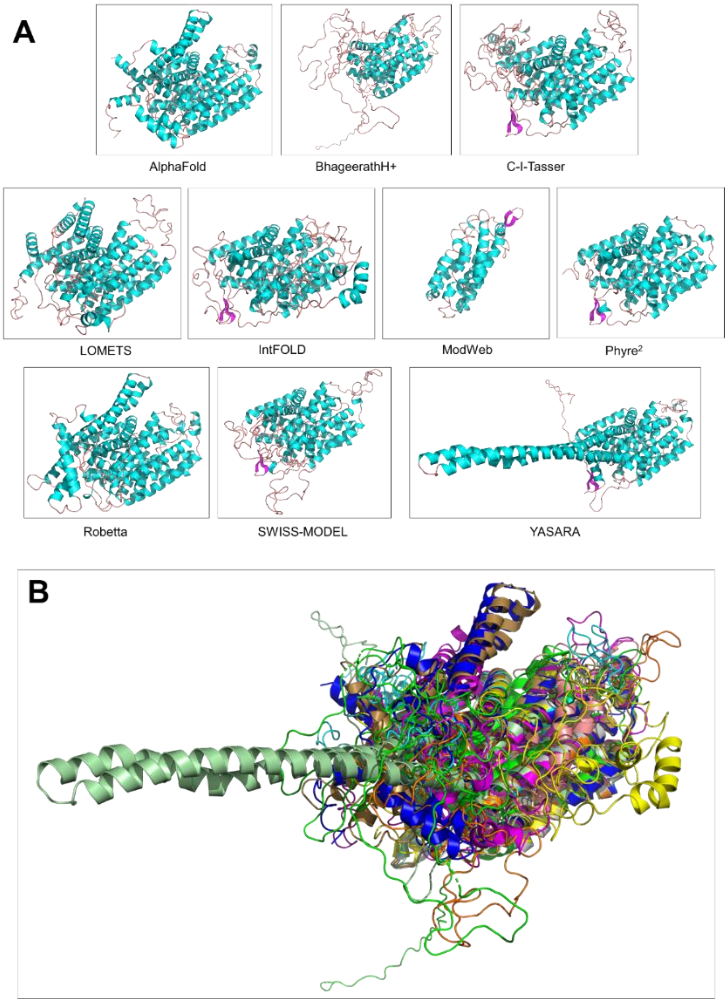

Figure 4. 3D models of High Affinity Cationic Amino Acid Transporter 1 protein. A. Cartoon representation of models generated by different tools (Cyan shows a-helix's. magenta shows b-sheets. and salmon color shows loops). B. Overlaid view of all generated models. Blue; AlphaFold. green; BhageerathH+. cyan; C-I-Tasser. magenta; LOMETS. yellow; IntFOLD. salmon color; ModWeb. grey; Phyre^2. sand color; Robetta. orange; SWISS-MODEL. pale green; YASARA.

3.4. Proton-Coupled Zinc Antiporter Models (SLC-30A)

Nine tools, apart from BhageerathH+ and RaptorX, successfully generated models of the human Proton-Coupled Zinc Antiporter, comprised of 507 amino acids. AlphaFold, C-I-TASSER, LOMETS, IntFOLD, and Robetta generated full sequence models. However, the remaining tools were unsuccessful in generating models with entire sequence (Table 9). AlphaFold, Robetta, and YASARA had the higher ERRAT quality scores of 93.38, 95.82, and 89.87, respectively, compared to other full sequence models. Only, the model generated by Robetta had the QMEAN value higher than –4, while all other models had QMEAN values below -4 which is an indicator of a model with low quality. Comparing the QMEANDisCo scores of AlphaFold, Robetta, and YASARA, Robetta had the highest score of 0.6. Moreover, 86.1 % of the residues in the model generated by Robetta were in the most favored regions on the Ramachandran Plot. On visual inspection, AlphaFold and Robetta exhibited structural similarities. The Robetta tool generated the model with the top scores for ERRAT, QMEAN, and QMEANDisCo. After analyzing the QMEAN Local Quality Estimation graphs (Figure S17-18), it was found that there were frequently occurring low confidence regions in all models. Further examination revealed that the region spanning from the 140th to 220th amino acids served as the common region with the lowest confidence score in all models. Hence, this could be a contributing factor to the varying folding patterns observed in the models. Robetta, a homology-based tool, generated a more reliable model for the SLC-30A based on the evaluations mentioned earlier. Moreover, Robetta generated better models for membrane proteins. However, it takes days to build models in the Robetta server.

Table 9. ERRAT Quality Scores. QMEAN. QMEANDisCo Values. and Z-Scores of Proton-Coupled Zinc Antiporter modeling.

Approach	Tool	Amino acid number in model	ERRAT Quality Score	QMEAN	QMEANDisCo	Z-SCORE
Ab initio	***AlphaFold***	507	93.38	-6.74	0.44	-5.87
	BhageerathH+	-	-	-	-	-
	RaptorX	-	-	-	-	-
Threading	***C-I-Tasser***	507	88.94	-12.00	0.36	-5.84
	LOMETS	507	81.19	-5.68	0.50	-6.16
Homology based	***IntFOLD***	507	59.957	-9.65	0.47	-5.63
	ModWeb	110	87.25	-4.23	0.36	-2.93
	Phyre^2	288	84.64	-6.41	0.60	-4.04
	Robetta	506	95.82	-1.26	0.60	N.C.
	SWISS-MODEL	421	70.3	-6.86	0.46	-4.99
	YASARA	507	89.87	-6.72	0.27	-4.06

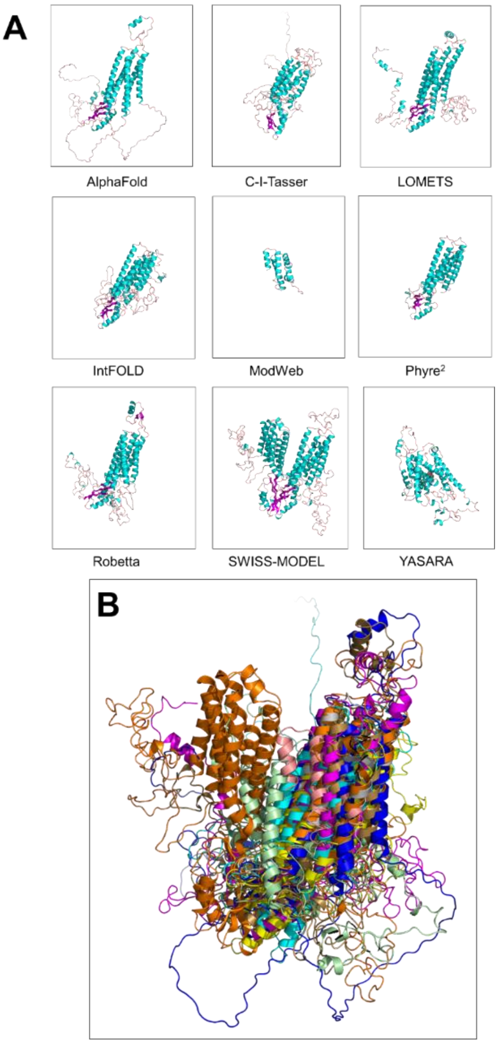

Figure 5. 3D models of Proton-Coupled Zinc Antiporter protein. A. Cartoon representation of models generated by different tools (Cyan shows a-helix’s. magenta shows b-sheets. and salmon color shows loops). B. Overlaid view of all generated models. Blue; AlphaFold. cyan; C-I-Tasser. magenta; LOMETS. yellow; IntFOLD. salmon color; ModWeb. grey; Phyre^2. sand color; Robetta. orange; SWISS-MODEL. pale green; YASARA.

Table 10. RMSD values of model-to-model comparisons for Proton-Coupled Zinc Antiporter protein.

Approach	RMSD	*AlphaFold*	*BHAGEERATHH+*	*RaptorX*	*C-I-Tasser*	*LOMETS*	*IntFOLD*	*ModWeb*	*Phyre^2*	*Robetta*	*SWISS-MODEL*	*YASARA*
Ab initio	***AlphaFold***											
	BhageerathH+											
	RaptorX											
Threading	***C-I-Tasser***	34.90										
	LOMETS	22.66			28.27							
Homology based	***IntFOLD***	27.55			28.40	21.81						
	ModWeb	17.30			17.35	16.92	18.07					
	Phyre^2	6.11			6.74	7.15	3.58	18.06				
	Robetta	18.51			32.76	16.59	22.20	16.63	6.59			
	SWISS-MODEL	15.90			24.83	14.62	18.41	17.90	3.49	13.01		
	YASARA	38.56			40.07	34.29	30.84	16.50	28.17	34.25	32.38	

3.5 *Bacillus subtilis* RNA Polymerase Sigma Factor (sigY)

The sigma factor protein of *Bacillus subtilis* RNA polymerase comprises 178 residues and is the shortest and final protein utilized for modeling. All tools except RaptorX generated model for sigY. Phyre^2, ModWeb, and SWISS-MODEL generated models of 154, 171, and 175 residues respectively (Table 11). It is worth noting that the model generated by Phyre2 lacks the region spanning between residues 90 and 109. Moreover, ModWeb and SWISS-MODEL lack the residues present at the beginning of the sequence. Alphafold, YASARA, and Robetta achieved perfect ERRAT scores of 100, while threading-based methods surpassed 95. The worst QMEAN scores were seen in BhageerathH+ and C-I-Tasser, with -10.28 and -3.78, respectively. On the other hand, IntFOLD had the highest QMEAN score, reaching 0.31, which was closest to 1.0. Except for BhageerathH+ (0.26), all models scored above 0.6 in terms of QMEANDisCo. All models had Z-scores similar to those of native proteins of similar size. Ramachandran plots indicated that over 95% of the residues in the protein models produced by AlphaFold, Robetta, IntFOLD, ModWeb, and YASARA are situated in the most favored regions. BhageerathH+, C-I-Tasser, and ModWeb generated models with differing structures from the other tools, as observed from Figure 6 and confirmed by RMSD calculations in Table 12.

Table 11. ERRAT Quality Scores. QMEAN. QMEANDisCo Values. and Z-Scores of *Bacillus subtilis* RNA polymerase sigma factor modeling.

Approach	Tool	Amino acid number in model	ERRAT Quality Score	QMEAN	QMEANDisCo	Z-SCORE
Ab initio	***AlphaFold***	178	100	-0.81	0.66	-6.88
	BhageerathH+	178	50.74	-10.26	0.26	-1.77
	RaptorX	-	-	-	-	-
Threading	***C-I-Tasser***	178	98.24	-3.78	0.63	-5.75
	LOMETS	178	95.88	-1.12	0.67	-6.52
Homology based	***IntFOLD***	178	98.23	0.31	0.67	-6.98
	ModWeb	171	76	-2.13	0.63	-5.9
	Phyre^2	154	87.67	-1.99	0.70	-6.54
	Robetta	178	100	1.46	0.67	-6.86
	SWISS-MODEL	175	88.62	-2.17	0.66	-6.2
	YASARA	178	100	-1.05	0.66	-6.13

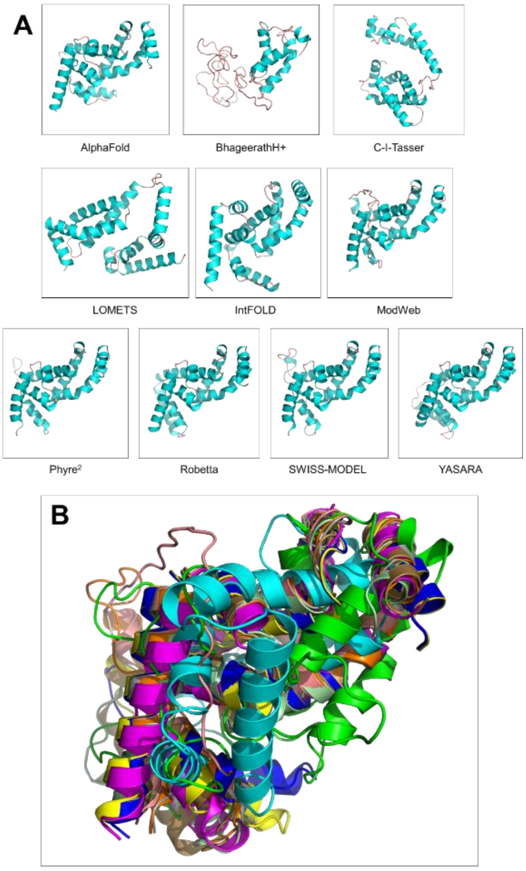

Figure 6. 3D models of *Bacillus subtilis* RNA polymerase sigma factor protein. A. Cartoon representation of models generated by different tools (Cyan shows a-helix's. magenta shows b-sheets. and salmon color shows loops). B. Overlaid view of all generated models. Blue; AlphaFold. green; BhageerathH+. cyan; C-I-Tasser. magenta; LOMETS. yellow; IntFOLD. salmon color; ModWeb. grey; Phyre^2. sand color; Robetta. orange; SWISS-MODEL. pale green; YASARA.

Table 12. RMSD values of model-to-model comparisons for *Bacillus subtilis* RNA polymerase sigma factor protein.

Approach	**RMSD**	***AlphaFold***	***BHAGEERATHH+***	***RaptorX***	***C-I-Tasser***	***LOMETS***	***IntFOLD***	***ModWeb***	***Phyre^2***	***Robetta***	***SWISS-MODEL***	***YASARA***
Ab initio	***AlphaFold***											
	BhageerathH+	16.56										
	RaptorX											
Threading	***C-I-Tasser***	14.88	19.08									
	LOMETS	5.29	17.99		13.32							
Homology based	***IntFOLD***	2.24	16.52		14.84	4.62						
	ModWeb	11.25	17.74		16.62	12.06	11.10					
	Phyre^2	5.82	13.77		14.22	11.04	6.08	6.48				
	Robetta	4.52	17.15		14.48	3.96	3.71	11.04	5.47			
	SWISS-MODEL	7.17	16.85		14.40	7.38	6.69	7.69	4.60	5.86		
	YASARA	4.71	17.06		14.47	4.82	4.47	11.51	5.67	3.88	6.79	

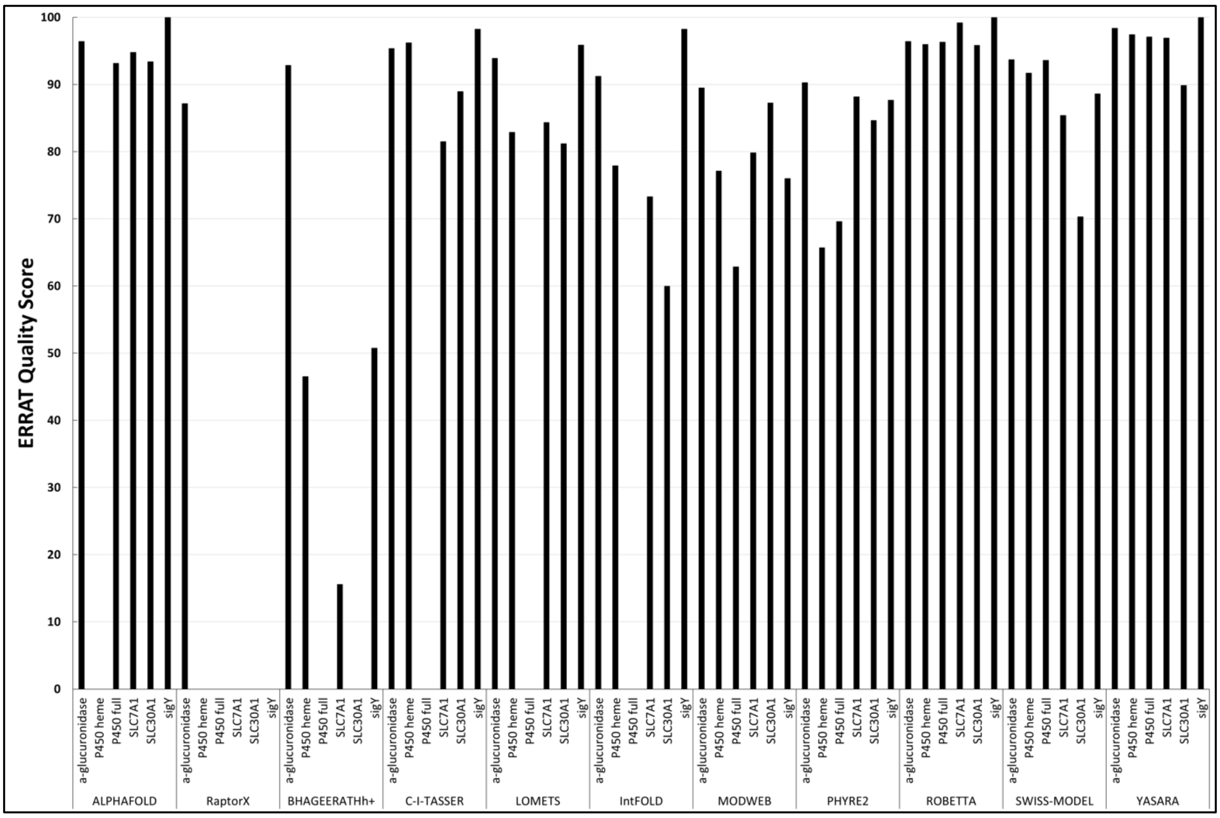

Figure 7. ERRAT quality scores of all generated models with tested eleven different tools.

Elucidating functions and mechanism of proteins is one of the primary questions in biochemistry. The structure of proteins has a major influence on the variety of activities they can perform. [59] Thus, one of the main goals of structural biology is to determine the three-dimensional structure of proteins. [21] Limited research has been conducted thus far to compare various model building tools, with a strong focus on homology modeling tools. Nikolaev and colleagues compared three homology modeling tools (Modeller, I-TASSER, and Rosetta) for predicting membrane proteins. Their findings indicate that successful modeling requires a target-template sequence identity of at least 40%. [21] Jang et al., conducted a study in order to compare multiple alignment tools and two template based model building program. SWISS-MODEL built models with better accuracy compared to Modeller. Because RMSD values of models generated both tools were below 1 and thus there is no significant quality difference between two tested programs for the modeling of soluble proteins. [60]

In our study, we have compared various tools from three different modeling approach. In addition to above mentioned results, YASARA, AlphaFold and SWISS-MODEL are the fastest tools among the all tested tools. On the other hand, other tools also had server problems and thus, they were not available time to time. Furthermore, YASARA generated high quality models since it generates hybrid models through performing hybrid modeling by separating the query protein into different units and selecting separate patterns for each unit. Major drawback of YASARA is that it is a paid tool, while all tested other tools are free of charge.

4. CONCLUSIONS

In this study, we conducted a comparison between ab-initio, threading, and homology modeling protein approaches. We tested 11 modeling tools to build models of six proteins. Template-based homology modeling tools, in particular, successfully built models for all of the tested proteins; however, threading and ab initio-based tools were unsuccessful in building models for some of the proteins. For example, ab initio and threading-based methods were unsuccessful in generating a model for the complete sequence of the Bifunctional Cytochrome P450/NADPH-P450 Reductase protein. Furthermore, RaptorX could only produce a model for *Geobacillus kaustophilus* ksilan alpha-1,2-glucuronidase. YASARA is suitable for proteins that contain heteroatoms, such as P450 monooxygenases, since most other tools do not include heteroatoms in their produced structures. AlphaFold is a powerful tool among template-free modeling methods. On the other hand, YASARA, Robetta, and SWISS-MODEL have emerged as prominent template-based tools. These findings will aid researchers in selecting the suitable protein modeling approach and tool for ensuring high-quality structures.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORSHIP CONTRIBUTIONS

Elif Altunkülah; performed the analysis, collected the data, analyzed the data, wrote the manuscript draft. **Yunus Ensari**; conceptialization, design of study, supervised the research, wrote and edited the manuscript.

REFERENCES

[1] Smith GM. The Nature of Enzymes. In: Biotechnology. 1995. p. 4–72.

[2] Benítez CMV, Lopes HS. Protein structure prediction with the 3D-HP side-chain model using a master–slave parallel genetic algorithm. J Brazilian Comput Soc. 2010;16(1):69–78.

[3] Divya M, Jain SJMN, Phadke SR, Kishore R, Kamate M, Gupta N, et al. Protein structure prediction for novel mutations in Arylsulfatase-A gene. Mol Cytogenet. 2014;7(1):P62.

[4] Alford RF, Fleming PJ, Fleming KG, Gray JJ. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane. Biophys J. 2020 Apr;118(8):2042–55.

[5] Batbat T, Öztürk C. Ayrık Yapay Arı Kolonisi Algoritması İle Protein Yapısı Tahmini. Bilişim Teknol Derg. 2016 Sep 30;9(3):260–3.

[6] Li X, Hu C, Liang J. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins. 2003 Dec;53(4):792–805.

[7] Torrisi M, Pollastri G, Le Q. Deep learning methods in protein structure prediction. Comput Struct Biotechnol J. 2020;18:1301–10.

[8] Aydin Z, Singh A, Bilmes J, Noble WS. Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. BMC Bioinformatics. 2011;12(1):154.

[9] Pearce R, Zhang Y. Toward the solution of the protein structure prediction problem. J Biol Chem. 2021;297(1):100870.

[10] ANFINSEN CB, HABER E, SELA M, WHITE FHJ. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci U S A. 1961 Sep;47(9):1309–14.

[11] Lee J, Wu S, Zhang Y. Ab Initio Protein Structure Prediction. In: From Protein Structure to Function with Bioinformatics. Dordrecht: Springer Netherlands; 2009. p. 3–25.

[12] Abbass J, Nebel JC, Mansour N. Ab Initio Protein Structure Prediction: Methods and challenges. In: Biological Knowledge Discovery Handbook. 2013. p. 703–24.

[13] Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci U S A. 1999 May;96(10):5482–5.

[14] Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol. 2001 Mar;306(5):1191–9.

[15] Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction. Biophys J. 2003;85(2):1145–64.

[16] Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science. 2005 Sep;309(5742):1868–71.

[17] Wu D, Wu T, Liu Q, Yang Z. The SARS-CoV-2 outbreak: What we know. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2020 May;94:44–8.

[18] Rashid MA, Shatabda S, Newton MAH, Hoque MT, Sattar A. A Parallel Framework for Multipoint Spiral Search in ab Initio Protein Structure Prediction. Adv Bioinformatics. 2014;2014:985968.

[19] Abbass J, Nebel JC. Customised fragments libraries for protein structure prediction based on structural class annotations. BMC Bioinformatics. 2015;16(1):136.

[20] Akdel M, Pires DE V, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, et al. A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol. 2022;29(11):1056–67.

[21] Nikolaev DM, Shtyrov AA, Panov MS, Jamal A, Chakchir OB, Kochemirovsky VA, et al. A Comparative Study of Modern Homology Modeling Algorithms for Rhodopsin Structure Prediction. ACS Omega. 2018;3(7):7555–66.

[22] Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res. 2006;34(17):e112.

[23] Battey JND, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. Proteins. 2007;69 Suppl 8:68–82.

[24] Heneghan MN, McLoughlin L, Murray PG, Tuohy MG. Cloning, characterisation and expression analysis of α-glucuronidase from the thermophilic fungus Talaromyces emersonii. Enzyme Microb Technol. 2007;41(6):677–82.

[25] Xu Y, Liu Z, Cai L, Xu D. Protein Structure Prediction by Protein Threading BT - Computational Methods for Protein Structure Prediction and Modeling: Volume 2: Structure Prediction. In: Xu Y, Xu D, Liang J, editors. New York, NY: Springer New York; 2007. p. 1–42.

[26] Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, et al. Tools for comparative protein

structure modeling and analysis. Nucleic Acids Res. 2003 Jul;31(13):3375–80.

[27] Shao M, Wang S, Wang C, Yuan X, Li SC, Zheng W, et al. Incorporating Ab Initio energy into threading approaches for protein structure prediction. BMC Bioinformatics. 2011 Feb;12 Suppl 1(Suppl 1):S54.

[28] Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol. 2001 Jun;310(1):243–57.

[29] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022 Jan;50(D1):D439–44.

[30] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.

[31] Jayaram B, Bhushan K, Shenoy SR, Narang P, Bose S, Agrawal P, et al. Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. Nucleic Acids Res. 2006;34(21):6195–204.

[32] Jabeen A, Mohamedali A, Ranganathan S. Protocol for Protein Structure Modelling. In: Ranganathan S, Gribskov M, Nakai K, Schönbach CBTE of B and CB, editors. Oxford: Academic Press; 2019. p. 252–72.

[33] Chen CC, Hwang JK, Yang JM. (PS)2-v2: template-based protein structure prediction server. BMC Bioinformatics. 2009;10(1):366.

[34] Chandra Sekhar Mukhopadhyay, Ratan Kumar Choudhary MAI. Basic Applied Bioinformatics. Wiley-Blackwell; 2017. 472 p.

[35] Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis. 2009 Jun;30 Suppl 1:S162-73.

[36] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 Jul;46(W1):W296–303.

[37] Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W252-8.

[38] Roche DB, Buenavista MT, Tetchner SJ, McGuffin LJ. The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W171-6.

[39] Roche DB, Tetchner SJ, McGuffin LJ. FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinformatics. 2011;12(1):160.

[40] Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845–58.

[41] Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, et al. ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. 2014 Jan;42(Database issue):D336-46.

[42] Krieger E, Vriend G. YASARA View - molecular graphics for all devices - from smartphones to workstations. Bioinformatics. 2014;

[43] Krieger E, Vriend G. New ways to boost molecular dynamics simulations. J Comput Chem. 2015 May;36(13):996–1007.

[44] Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011 Jan 1;39(suppl_1):D411–9.

[45] Krieger E, Vriend G. Models@Home: distributed computing in bioinformatics using a screensaver based approach. Bioinformatics. 2002 Feb;18(2):315–8.

[46] Zheng W, Zhang C, Li Y, Pearce R, Bell EW, Zhang Y. Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell reports methods. 2021 Jul;1(3).

[47] Wu S, Zhang Y. LOMETS: A local meta-threading-server for protein structure prediction. Nucleic Acids Res. 2007 May 15;35(10):3375–82.

[48] Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, Bordoli L, et al. The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Res. 2017 Jan;45(D1):D313–9.

[49] Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics. 2020 Mar 15;36(6):1765–71.

[50] Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W407-10.

[51] Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins. 1993 Dec;17(4):355–62.

[52] Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol. 1995 Apr;5(2):229–35.

[53] Colovos C, Yeates TO. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Sci. 1993 Sep 1;2(9):1511–9.

[54] Ramachandran GN, Sasisekharan V. Conformation of Polypeptides and Proteins In: Anfinsen CB, Anson ML, Edsall JT, Richards FMBTA in PC, editors. Academic Press; 1968. p. 283–437.

[55] MacArthur MW, Thornton JM. Deviations from planarity of the peptide bond in peptides and

proteins. J Mol Biol. 1996 Dec;264(5):1180–95.

[56] Hooft RWW, Sander C, Vriend G. Objectively judging the quality of a protein structure from a Ramachandran plot. Bioinformatics. 1997 Aug 1;13(4):425–30.

[57] Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993 Apr 1;26(2):283–91.

[58] Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. Proteins. 1992 Apr;12(4):345–64.

[59] Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun. 2021;12(1).

[60] Jang WD, Lee SM, Kim HU, Lee SY. Systematic and Comparative Evaluation of Software Programs for Template-Based Modeling of Protein Structures. Biotechnol J. 2020;1–21.