Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1402011

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 7 - Issue 2: 165-174 / March 2024

PROTEIN HOMOLOGY MODELING IN THE LOW SEQUENCE SIMILARITY REGIME

Sebnem ESSIZ[1]*

[1]Kadir Has University, Faculty of Engineering and Natural Sciences, Department of Molecular Biology and Genetics, 34083, İstanbul, Türkiye

Abstract: Predicting the 3-D structure of a protein from its sequence based on a template protein structure is still one of the most exact modeling techniques present today. However, template-based modeling is heavily dependent on the selection of a single template structure and the sequence alignment between target and template. Mainly when the target and template sequence identity is low, the error from the alignment introduces larger errors to the model structure. An iterative method to correct such alignment mistakes is used in this study with a benchmark set from CASP in the extremely low sequence-identity regime. This is a protocol developed and tested before and it evaluates the alignment quality by building rough 3-D models for each alignment. Then by using a genetic algorithm it iteratively creates a new set of alignments. Since the method evaluates models, not sequence alignments, structural features are automatically incorporated into the alignment protocol. In the current study, models from structural alignment have been built by Modeller program to show the maximum possible quality of the model that can be obtained from that template structure with the iterative modeling protocol. Then the results and correctly aligned segments from the iterative modeling protocol are analyzed. Finally, it has been shown that if a good local fragment assessment scoring function is developed, the correctly aligned segments exist in the pool of alignments created by the protocol. Thus, the improvement of modeling in the low sequence identity regime is conceivable.

Keywords: Homology modeling, Sequence-sequence alignment, Genetic algorithm, Molecular modeling, Structural alignment

***Corresponding author:** Kadir Has University, Faculty of Engineering and Natural Sciences, Department of Molecular Biology and Genetics, 34083, İstanbul, Türkiye
E mail: sebnem.gokhan@khas.edu.tr (S. ESSIZ)
Sebnem ESSIZ https://orcid.org/0000-0002-5476-4722

Received: December 09, 2023
Accepted: January 15, 2024
Published: March 15, 2024

Cite as: Essiz S. 2024. Protein homology modeling in the low sequence similarity regime. BSJ Eng Sci, 7(2): 165-174.

1. Introduction

The three-dimensional (3-D) structure of the protein dictates its biological function, consequently understanding protein structure at the molecular level is essential in terms of understanding the function and malfunctions of proteins. Experimental techniques such as X-ray crystallography, NMR spectroscopy, and Cryo-Electron Microscopy are golden standards for determining protein 3-D structure, however, they have certain limitations in terms of time, resolution, and size of the systems. Mainly NMR is limited in terms of the size of the protein, X-ray has limitations rooted in difficulties in the crystallization process under the natural physiological environment and cryo-EM has still limitations with the resolution of the maps. Due to these pitfalls, the number of protein structures resolved fell very behind the number of known unique sequences, namely protein structures in PDB Databank compared to distinct sequences obtained in UniProtKB are still 1000-fold smaller in numbers as of January 2023 (Nassar et al., 2021; Bertoline et al., 2023)

Consequently, to fill this large gap between known structures and sequences, computational studies have proven to be valuable for the prediction of protein structure (Pieper et al., 2006; Gromiha et al., 2018). In general, there are two mainstream categories in computational protein structure prediction. The first one is free modeling and the second one is template-based modeling. Ab-initio modelling is free modeling and used when there is no known structure close to the unknown target sequence. As the name indicates, ab initio modeling uses first-principal forces which drives protein to go into their native state, namely physical forces, and the potential of chemical interactions in between particles are used to simulate the protein sequence into its folded state (Bonneau and Baker, 2001; Hardin et al., 2002; Pearce et al., 2022). The main advantage of the method is that since it is not dependent on the library of known structures, it can successfully predict novel folds; however, computation time and cost are the main drawbacks.

Template-based or comparative modeling, on the other hand, doesn't use physical or chemical interactions as an all-atom force field but a less-detailed knowledge or information-based driving forces. Threading and homology modeling is both template-based modeling and they require a similar known template structure to base the model target sequence on. In threading all possible 3-D structures are tried on the given unknown target sequence while in homology modeling a single 3-D

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

structure is picked based on sequence similarity. GenTHREADER (Jones, 1999) is one of the most widely used threading software which uses a fit function by trying a target sequence to all known folds that exist in the database. Hybrid methods in between ab-initio and threading methods also exist and these methods start modeling from known smaller structure fragments and they became one of the standard ways for protein structure prediction such as I-TASSER (Yang and Zhang, 2015), Robetta (Kim et al., 2004), Rosetta@home (Rohl et al., 2004), Quark (Xu and Zhang, 2012).

Swiss-Model (Guex and Peitsch, 1997) and Modeller (Webb and Sali, 2016) are examples of homology-based modeling which are dependent on the selection of a single template structure and the sequence alignment between target and template. After the alignment of the target and template sequences, they collect geometric restraints from the known template structure and impose those restraints on the model structure of the unknown target sequence's model.

Very recently there has been a breakthrough in the protein structure prediction area. In the critical assessment of protein competition, CASP 13 and CASP 14, artificial-intelligence-based programs Alphafold and Alphafold2 (Jumper et al., 2021a; Jumper et al., 2021b). Alphafold and Alphafold2 both use machine learning by training on the distance between amino acids of known structures in PDB databank. They then create distograms, which are like the histograms of the distribution of distances in the structure. Then they use neural networks for predicting distograms from multiple sequence alignment for the target sequence. They provided suburb results compared to its competitors in recent CASP competitions. However, it is still a hybrid method in terms of machine learning from previously known structures and it inherits the problems of predicting loop segments, multi-subunit complexes, or different conformations of the proteins (Bertoline et al., 2023) Mainly using multiple sequence alignment and template structures is not a new technique, however, their training and neural network are very successful in terms of covering all the knowledge we have in the PDB databank.

In general, when there is no close template structure like for the loop segments and disordered proteins, all methods still have limitations. However, it is widely known that structure is known to be conserved better than sequence (Sauder et al., 2000). Thus, if the structure information is somehow incorporated into the sequence alignment, then it has been shown to improve modeling in terms of the problems coming from sequence alignments. Again, please note that when the sequence identity between the target and template is above %30 percent, this alignment problem doesn't exist (Sauder et al., 2000).

Homology modeling as being still the most exact out of all these solutions suffers from sequence alignment errors when the target sequence is novel. As well as the Alphafold, there have been many advances to correct alignment errors such as PSI-BLAST in the past decades. They all benefit from aligning multiple sequences as a profile. This way it becomes possible to incorporate structural features from other sequences with structures into alignment substitution matrices. There are various profile-profile alignment methods with slight differences in the ways they create the profiles and alignments. (Wang and Dunbrack, 2004; Kahsay et al., 2005; Soding, 2005; Zhou and Zhou, 2005; Dunbrack, 2006).

Moulder is such a technique developed for correcting the alignment errors in modeling by Modeller (John and Sali, 2003). It is an iterative alignment and modeling approach. First, a reliable multi-sequence alignment profile is created with close homologs of the target sequence, which contains 25 different alignments for the target sequence-template construct. In normal homology modeling, a single best-scoring alignment is moved to the modeling step. In Moulder, the best 15 of the 25 alignments are selected and moved to the structural modeling step. The resulting 15 protein structures are then subjected to a simple model scoring. The main difference between the two steps is that sequence similarity is used to create the 25 alignments. After 15 model structures are created, GA341 structure scoring will be used which is a 3-D structure assessment score. Then, genetic algorithm moves, crossover and mutation operators are applied to these alignments to reach 300 different alignments. For the newly formed alignment population, protein structures are obtained using the rapid modeling technique. The evaluation of these alignments is performed by scoring the 3-D models. It has been supported by many studies that the structures of proteins are much better preserved than their sequences. In this respect, it is important to evaluate the 300 different alignments based on models. The resulting alignment of the 10 models with the best structure score forms the family alignments in the next iteration of the genetic algorithm. Moulder algorithm runs for multiple iterations until it reaches a certain number of alignments. This protocol has already been successfully implemented and its performance in the twilight zone, the area with low sequence similarity, is limited in terms of reaching to full potential improvement provided by template structures.

In the present work, the CASP8 benchmark set is used to show the model quality in terms of profile-profile alignment first. Then for the same targets, the models based on structure-structure alignment will be evaluated. For one of the hard targets, Moulder method will be tested. Then, the populations created by Moulder genetic algorithm step will be analyzed by comparing to the ideal structural alignment. This will help to show how independently the scoring and sampling parts of the resulting populations work and how this information can be used to eliminate alignment errors in homology modeling.

The genetic algorithm generally relies on advantageous features dominating the pool of sequences over time, leading to the correct alignment. The important point here is that sampling alignments and scoring alignments cannot be considered in isolation from each other. The goal of this paper is to test and show, why Moulder, which uses a genetic algorithm for alignment, with a very low sequence similarity benchmark set is not getting the results guaranteed by templates. By analyzing this, the alignments created by the genetic algorithm will be tested in terms of the percentage of correct alignment segments. If the correct alignments exist in the population but not picked by scoring, it will be shown that if sampling and scoring algorithms can be separated, the modeling problems caused by errors in alignment can be solved by bringing correct alignment segments together.

2. Materials and Methods

2.1. MOULDER: An iterative Alignment Technique

The moves of Moulder below are directly adopted from previous studies (John and Sali, 2003; Eramian, 2008) with modified parameters to fit to the computer usage in terms of number of nodes available.

Step 1: Initial Alignment

In the modeling steps, there is a target unknown sequence and a known template structure.

First target and template sequences are obtained and then for each sequence, profiles are built by Modeller's profile.build() command with default parameters for global dynamic alignment and Uniprot-90 sequence database. Then these profiles are aligned by using Modeller's Alignment.salign() function with global dynamic scoring (Marti-Renom et al., 2004). In addition to the best alignment, 5 suboptimal alignments have been created in the alignment command by changing weight matrix values. Suboptimal alignments are created by shifting alignment parameters, in Modeller.salign() n_subopt = 5, subopt_offset = 15 was used. In the low sequence identity range, the suboptimal alignments have been shown to include many correct segments, sometimes even more than the optimal alignment (Chen and Kihara, 2011)

Step 2: Initial Models from the alignment(s)

For the input alignment, Modeller's automodel class is used with default parameters. Optimization level is set to very fast and a total of 2500 models are obtained. The reason for creating this many models is to have a good quality assessment for the initial model.

Step 3: Distribute alignment(s) to 10 nodes and apply genetic algorithm.

For the initial round, there are only 6 parent alignments. After that, there are 100 alignments and nodes receive at least 2 different alignments. This way guarantees that crossover moves can be performed. On each node, child alignments are created by applying genetic algorithm operators to the alignments.

There are five different operators:

Single-point crossover: It requires two alignments as input. Each alignment is divided into two blocks and the second part of the first alignment is swapped with the second part of the second alignment.

Double-point crossover: It requires two alignments as input. The alignments are divided into three blocks and the middle block of the first alignment is swapped with the middle block of the second alignment.

Gap Insertion: It requires a single alignment. A position is selected in the randomly picked location of the randomly selected sequence of the alignment and a random length of gaps is inserted into that sequence. To end meets, the same number of gaps are inserted at the end of the second sequence. A random number is selected from 1 to 7 for the number of gaps to be inserted.

Gap Deletion: A random gap position is selected in the first sequence, then a random amount of gap is deleted from the total length of that gap in that sequence. The same number of gaps are added to the random position of the second sequence. The amount of gap deleted is determined by a random number from 1 to the total number of gaps in the selected gap.

Gap Shift: A single gap is selected from one sequence, and it is shifted to a random position.

In the genetic move step, the weight of each move is as follows: Single-point crossover is chosen as %40, double-point cross-over is chosen as %20, gap insertion is chosen as %10, gap deletion is chosen as %10 and gap shift is set to %20.

After this step, a redundancy check is applied to the child alignments created from the parent alignments. This step is stopped when there are valid 2500 alignments in the pool of alignments.

Step 4: Model Building and picking 250 best alignments.

Like the initial alignment model, one rough modeling step is carried out, mainly again by using Modeller's automodel class. The restraints in the Modeller's automodel class were reduced to 10 Ångstrom from 14 Ångstrom. Model randomization is turned off. No molecular dynamics refinement has been carried out. And finally, the optimization cycle is reduced from 200 to 50. This way a very quick modeling step has been carried out.

Scoring function used in this step has the following Modeller scoring components:

1. Sequence identity
2. Percentage of gaps in the target/template alignment
3. Z-PAIR: Cα- and Cβ-based Distance-Dependent
4. Z-SURFACE: Cβ-based Accessible Surface Score:
5. Z-COMBINED: Combined Distance and Surface Potential Score
6. GA341: Fold Assessment Score

Step 5: Model in detail and pick the best alignments.

For the best alignments, 3 models will be modeled for each alignment and evaluated by a more detailed scoring function namely DOPE, which is the Discrete Optimized Protein Energy score developed by Modeller. It is an atomistic, distance dependent scoring function developed by Modeller (Shen and Sali, 2006). It is a more detailed scoring function than the ones present in Step 5.

Step 6: Pool and sort alignments.

Once each node has built 100 valid, non-redundant child alignments, the alignments are then ranked according to their DOPE scores. Please note that DOPE is not a 2-D sequence alignment score. Mainly alignments are evaluated from the 3-D models that have been created.

Moulder steps are summarized in Figure 1.

2.2. Benchmark Data Set

The targets of CASP 8, https://predictioncenter.org/casp8/index.cgi are downloaded from the template-based modeling class. The PDB structure of template structures is also obtained. The best templates with single chains are selected. The targets are selected from the low sequence identity range, sequence identity between target and template ranks from 1.99 % to 34.91 %. The target list with their corresponding sequence identities and sequence lengths are displayed in Table 1. The twilight zone is defined as the zone of 20–35% sequence identity. Here the selected targets are even lower than twilight zone with low sequence identities to template structures.

2.3. Profile-Profile Alignment

Target and template sequences are obtained and then for each sequence profiles are built by Modeller's profile.build() command with default parameters for global dynamic alignment and Uniprot-90 sequence database. Then these profiles are aligned by using Modeller's Alignment.salign() function with global dynamic scoring (Marti-Renom et al., 2004).

2.4. Structure-Structure Alignment

For the target list in Table 2, both target and template structures are also downloaded. Modeller's Alignment.salign() command with (0,3) 3D gap penalties are used for aligning two structures. Namely, this is the ideal test case. If both target and template structures were known, then the correct alignment would have been obtained.

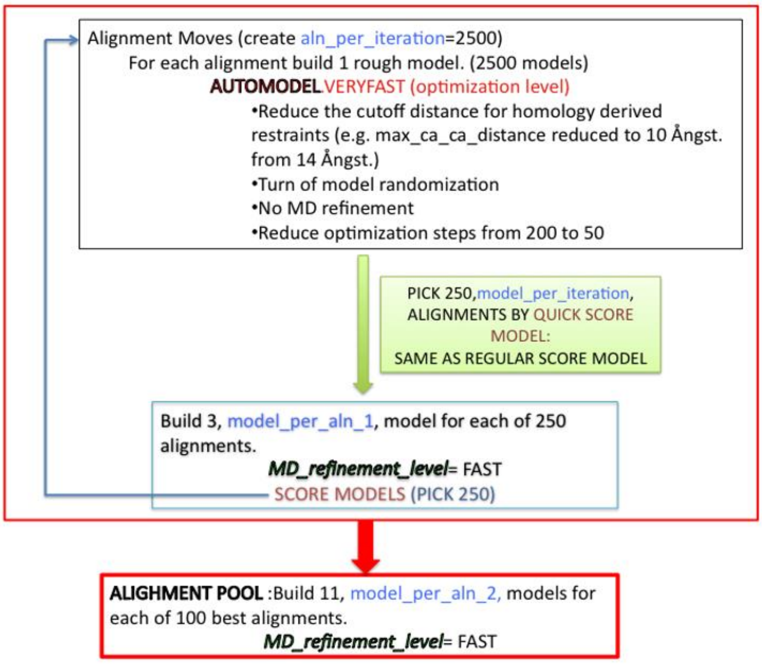
Alignment Moves (create aln_per_iteration=2500)
For each alignment build 1 rough model. (2500 models)
AUTOMODEL VERYFAST (optimization level)
•Reduce the cutoff distance for homology derived restraints (e.g. max_ca_ca_distance reduced to 10 Ångst. from 14 Ångst.)
•Turn of model randomization
•No MD refinement
•Reduce optimization steps from 200 to 50
PICK 250,model_per_iteration, ALIGNMENTS BY QUICK SCORE MODEL: SAME AS REGULAR SCORE MODEL
Build 3, model_per_aln_1, model for each of 250 alignments.
MD_refinement_level= FAST
SCORE MODELS (PICK 250)
ALIGHMENT POOL :Build 11, model_per_aln_2, models for each of 100 best alignments.
MD_refinement_level= FAST

Figure 1. Schematic of the Moulder steps of one iteration.

Table 1. Target list

Target	Seq. Length	Seq. ID (%)[1]	Target	Seq. Length	Seq. ID (%)[1]
T0414-D1	127	7.09	T0490-D1	361	14.40
T0408-D1	98	17.35	T0494-D1	345	29.28
T0409-D1	62	4.84	T0497-D1	124	30.65
T0412-D1	166	16.27	T0501-D1	213	11.11
T0420-D1	168	17.26	T0501-D2	126	13.15
T0423-D1	143	32.87	T0502-D1	93	23.66
T0424-D1	175	21.71	T0503-D1	144	14.58
T0424-D2	84	25.00	T0504-D1	62	6.45
T0436-D1	405	16.05	T0504-D2	90	11.11
T0445-D2	107	7.48	T0505-D2	104	14.42
T0477-D1	106	34.91	T0506-D2	78	20.51
T0478-D1	126	9.52	T0507-D1	124	17.74
T0481-D1	135	12.59	T0509-D1	209	20.57

[1]= The sequence identity is divided by the total length of the longer sequence.

Table 2. Comparison of models from profile-profile sequence and structure-structure alignments

			Profile-Profile Alignment			Structural Alignment		
Target	Seq. Length	Seq. Identity	Native Overlap	RMSD (Ångst)	Z-Dope	Native Overlap	RMSD (Ångst)	Z-Dope
T0414-D1	127	7.09	0.087	12.214	0.799	0.772	4.479	0.120
T0408-D1	98	17.35	0.878	3.572	-0.640	0.939	1.960	-0.463
T0409-D1	62	4.84	0.161	11.566	1.748	0.903	2.252	-0.465
T0412-D1	166	16.27	0.836	3.457	-0.618	0.880	3.066	-0.824
T0420-D1	168	17.26	0.637	12.557	0.573	0.946	2.064	-1.145
T0423-D1	143	32.87	0.958	2.095	-0.766	0.972	1.730	-0.967
T0424-D1	175	21.71	0.766	9.881	0.293	0.794	3.714	-0.415
T0424-D2	84	25.00	0.845	2.636	-0.693	0.988	1.776	-0.912
T0436-D1	405	16.05	0.691	6.997	-0.392	0.798	3.769	-0.541
T0445-D2	107	7.48	0.056	14.750	1.096	0.907	2.288	-0.749
T0477-D1	106	34.91	0.783	9.781	0.467	0.915	2.479	-0.104
T0478-D1	126	9.52	0.040	14.153	1.279	0.802	2.985	-1.247
T0481-D1	135	12.59	0.556	7.313	-1.066	0.815	3.380	-1.115
T0485-D1	207	16.43	0.647	7.923	-0.550	0.725	4.243	-0.792
T0490-D1	361	14.40	0.795	3.560	-0.184	0.850	2.631	-0.178
T0494-D1	345	29.28	0.812	4.188	-0.703	0.872	3.196	-0.947
T0497-D1	124	30.65	0.879	2.362	-1.227	0.960	1.991	-1.391
T0501-D1	213	11.11	0.643	11.347	0.330	0.831	10.418	0.015
T0501-D2	126	13.15	0.151	13.848	1.069	0.825	2.915	-0.517
T0502-D1	93	23.66	0.871	3.250	-0.523	0.968	2.255	-1.115
T0503-D1	144	14.58	0.729	7.931	0.815	0.861	4.240	0.384
T0504-D1	62	6.45	0.742	4.713	0.359	0.887	3.70	0.073
T0504-D2	90	11.11	0.111	10.972	0.901	0.750	8.830	0.894
T0505-D2	104	14.42	0.077	15.436	1.538	0.827	4.146	-1.195
T0506-D2	78	20.51	0.821	3.748	-0.396	0.936	2.228	-0.430
T0507-D1	124	17.74	0.105	11.870	0.746	0.847	2.537	-0.193
T0509-D1	209	20.57	0.828	3.274	-0.546	0.933	2.009	-1.114

Thus, structure alignment results will display how much template-based modeling can be improved if the errors from the sequence alignment are minimized (Sauder et al., 2000).

2.5. Model Assessment Parameters

Native overlap is the number of Cα atoms in the model within 3.5 Å of the corresponding atoms in the native structure divided by total number of Cα atoms. It is calculated after superposition of target and native structure. Root mean square displacement (RMSD) is proportional to the displacements of atom coordinates of the model structure from native structure by the following formula (Equation 1):

$$RMSD\ (x, y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|^2} \qquad (1)$$

In equation 1, x is the coordinates of model structure atoms, y is the coordinates of native structure atoms and i runs from 1 to number of atoms. It shows deviation from the ideal structure, while native overlap shows overlap with the ideal.

Finally, Z-Dope is the normalized DOPE scoring function. It is a distance-dependent statistical potential based on the separation between atoms and developed by MODELLER (Marti-Renom et al., 2004). Mainly Z-Dope is a statistical z-score that displays how your model's DOPE score is better than the average model. The more negative the-Z-Dope the better the model quality is.

3. Results and Discussion

3.1. Comparison of Structure-Structure and Profile-Profile Alignments

In Table 2, model qualities are displayed for profile-

profile and structure-structure alignment results. Native overlap, RMSD, and Z-Dope score values are displayed. These are 3 different model assessment scores, explained in the Methods 2.5 section, used to evaluate the quality of the models. After models created based on profile-profile and structure-structure alignments, each model is compared to the native structure via these parameters. These are direct~~ly~~ simple models from Modeller's automodel class, no iterations. In Figure 2, each of these parameters is plotted against the sequence identity between target and template. Mainly Figure 2 plots results tabulated in Table 2 according to sequence identity.

In Figure 2A, native overlap value of the models from profile-profile alignments is shown and in the low sequence identity regime, model quality is also low. When the sequence identity between target and template gets better than 15 %, the models are getting better both for profile-profile and structure-structure alignment in terms of native overlap. For the models that have very low model quality (shown in blue in Figure 2A), the native overlap values get directly above 0.6 from 0.1 when structural alignment is used. This means for this low-sequence identity regime, the template-based modeling can have much better results. However, due to the alignment errors, the models from profile-profile alignments have much lower native overlap values.

In Figure 2B, the model quality is displayed via RMSD values. Lower RMSD means the models are closer to the native structure. The model quality for profile-profile alignment is again very low in the low sequence identity, however, even for sequence identity greater than %15, there are problems with very high RMSD values for models from profile-profile alignments.

Z-Dope is a predictive scoring function, not a parameter for comparing a model to a native structure, unlike the first two parameters in Figure 2. It is the least predictive one for model assessment out of 3 parameters. Although structure-structure alignment results seem better than profile-profile alignment in Figure 2C, the predictive property of Z-Dope is still low for all models. Namely, it cannot differentiate between good models and bad models. Please note that in general, all the targets in the benchmark set are hard in terms of modeling. They all have sequence identity lower than %35, so even the structure conservation for better target and template pairs is low.

According to the results in Table 2, one of the worst profile-profile alignment results has been picked as a test case to analyze the results of Moulder. That target is T0409-D1 with a very low sequence identity of 4.84. According to profile-profile alignment results, the model's native overlap value is 0.161, RMSD is 11.566 Ångstrom and Z-Dope is 1.748.

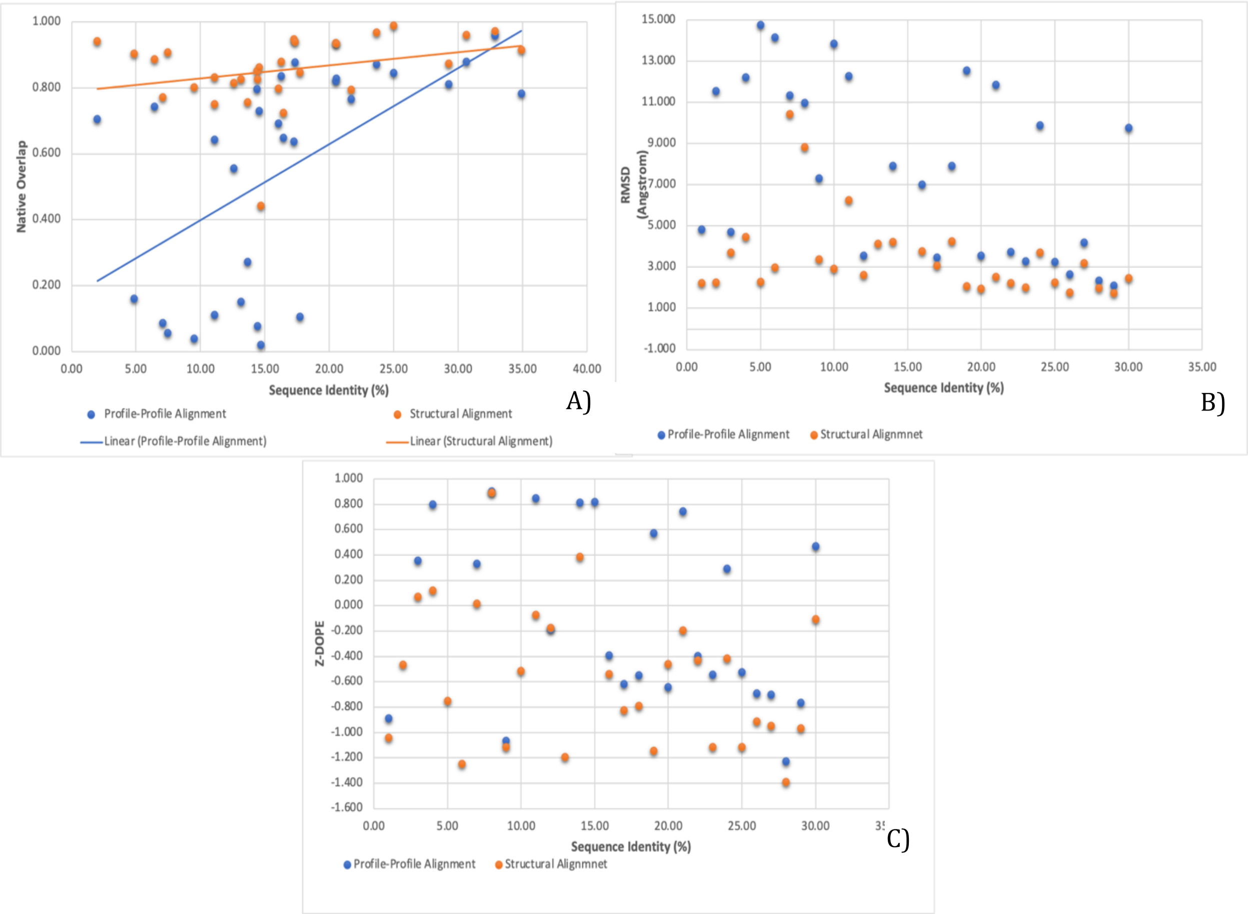

Figure 2. Benchmark set Model Quality A) Native overlap versus sequence identity B) RMSD versus sequence identity C) Z-Dope versus sequence identity.

Z-Dope is very positive and non-native like. However, when the model is built according to the structural alignment, the model has a native overlap of 0.903, a low RMSD of 2.252 Ångstrom. and finally, Z-Dope gets to a negative value. Now this target is selected for the analysis with Moulder since the low model quality comes from the errors in the sequence-sequence alignment. Thus, it has room for improvement. Overall, 30 Moulder iterations have been carried out for this target.

3.2. Analysis of the Genetic Algorithm Moves

During the iterations of Moulder the genetic algorithm moves have been analyzed for their weights. The single-point crossover is set to %40, double-point cross-over to %20, gap insertion to %10, gap deletion to %10, and gap shift is set to %20 in Moulder. In Figure 3, the moves for the initial 15 iterations are displayed. In the initial iterations number of gap deletions, gap shift, and gap insertions are higher than their expected weights because initially these mutation type changes are needed to create diverse alignment. Then the cross-overs get higher by adjusting to the final weights. Cross-overs are larger whole segment changes, thus the population needs gap moves to diverge from the starting alignment initially. The movement weights agree with the ones in Figure 3. Each color shows an iteration number starting from 1 to 15, and the y-axis is the number of moves.

3.3. Is the Correct Model in the Pool of Models?

In Figure 4A, best 100 alignments obtained after 30 iterations of Moulder are displayed. For T0409-D1, the profile-profile alignment results give 0.161 native overlap as the initial model quality. At the end of Moulder iterations, 100 best-scored alignments are obtained, and corresponding models are built by Modeller. The best model native overlap values get as better as 0.360 native overlap value.

In Figure 4B, the model picked at each iteration is shown with red lines, while the best model present in the pool of alignments at each iteration is shown with green lines. It is clear from the results in this low model quality, mainly because of the very low sequence identity, that the scoring function is not picking the best model present in the pool of models. If the scoring function could have picked the model with the better native overlap value, the model quality would have reached to 0.46 range.

In Figure 4C, RMSD of the final models from the 30 iterations of Moulder is displayed. Again, the initial RMSD is 11.566 Ångstrom in Table 3, and the best model RMSD does not even reach 9.6 Ångstrom. However, when the best models versus best-scored models during each iteration are compared in Figure 4D, there is a model that has a RMSD as low as 5 Ångstrom at iteration 24. Namely that model is in the pool of models at the end of each iteration but not picked by the scoring function. If the alignment belonging to this model was picked by Moulder for the next iteration, there would have been an improvement. But RMSD value stays, the green line in Figure 5D, very flat during 30 iterations.

Here the aim is not to benchmark Moulder one more time. In the original method paper (John and Sali (2003), 19 hard modeling targets that shared 4-27% sequence identity with their template structures were used for benchmarking, and the average alignment accuracy increased from 37% to 45% relative to the initial alignment at the end of Moulder iterations. However here in Table 2, it is shown that the modeling accuracy can be improved more than this if the alignment incorporates structural information into the sequence alignment. Please see Figure 2A for the possible improvement of models when the sequences are aligned based on structure (orange points in Figure 2A). As shown in Figure 4, Moulder iterations are sampling better models than the models picked by scoring function in the iterations of Moulder. This result is very important in terms of the development of similar methods. All sampling methods by iterations can be separated into two parts. One of them is sampling the different alignments, while the second one is picking the best one when the native structure is unknown. Here since the native structures of CASP 8 benchmark set are known, it is shown that although better alignments exist, Z-Dope scoring function was not able to pick some of them.

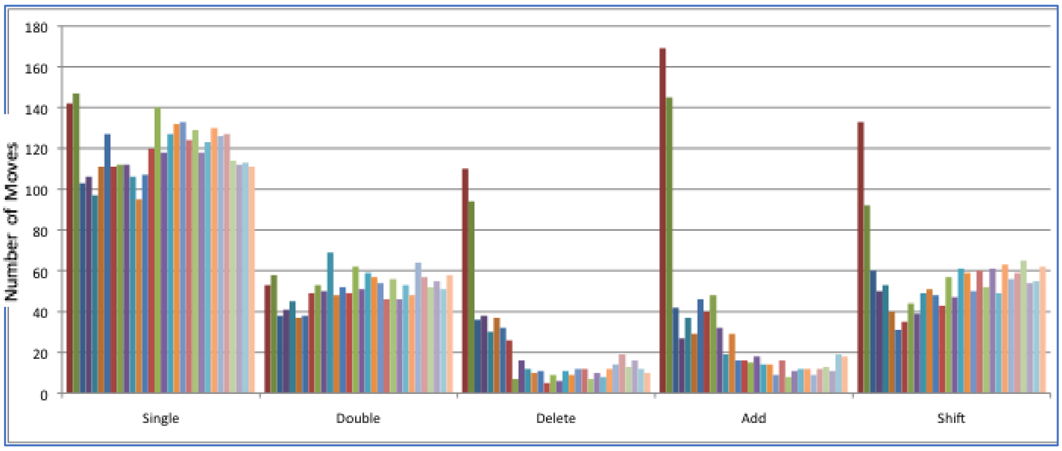

Figure 3. Genetic algorithm moves: The number of moves in each iteration is displayed with a different color. Single= Single-point cross-over, Double= Double-point cross over, Delete = Gap Delete, Add = Gap Insertion and Shift = Gap Shift.

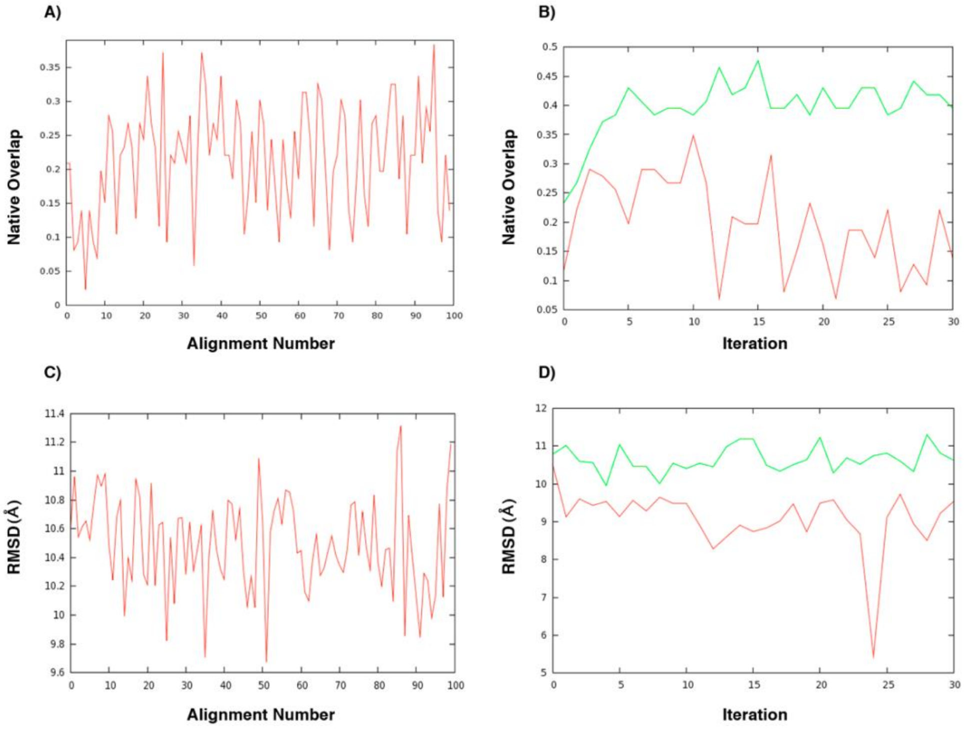

Figure 4. A) Native-overlap scores of models. B) Native overlap of models picked by DOPE (red), models with the best native overlap value (green). C) RMSD of models D) RMSD of models picked by DOPE (green), models with the best native overlap value (red).

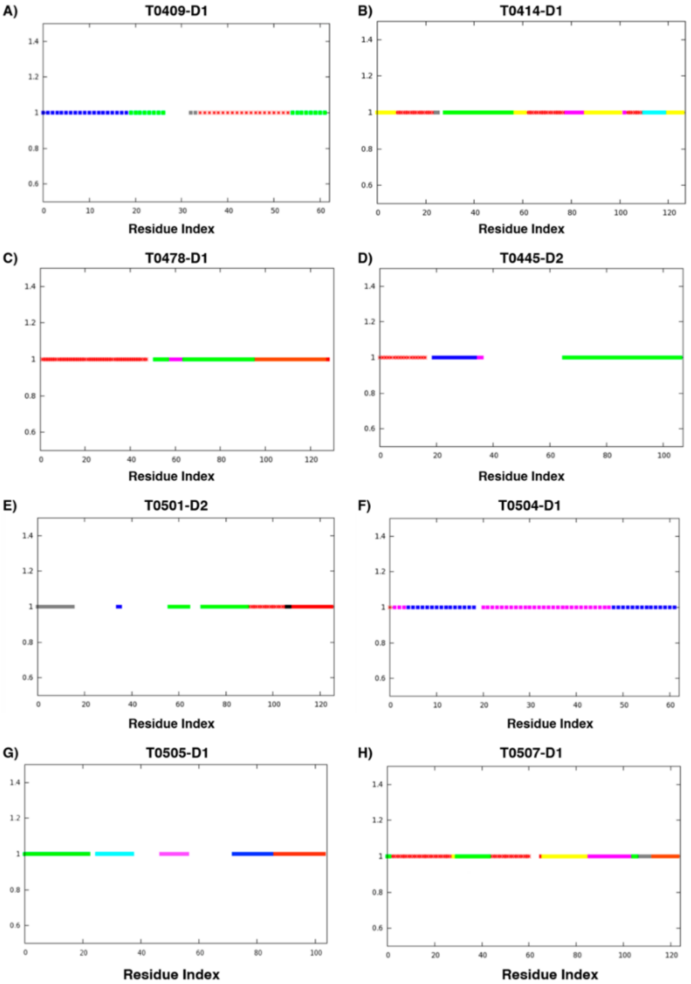

Figure 5. Coverage of suboptimal alignments. The red color shows the correctly alignment segments of the profile-profile alignment. The different colors show the correctly aligned segments from sub-optimal alignments.

3.4. Are Correct Alignment Segments in the Pool of Alignments?

The existence of correct alignment segments is checked by comparing the alignment segments to the structural alignment within the pool of suboptimal alignments. For this, the targets that have very low sequence identity and very low model quality from profile-profile alignments are selected in Table 2. Those targets are: T0409-D1, T0414-D1, T0445-D2, T0478-D1, T0501-D1, T0504-D2, T0505-D2 and T0507-D1. Their sequence identity ranges from 4.84 to 17.74. They all have very low native-overlap values with profile-profile alignments, but the models are getting much better when the modeling is done with structural alignment. That means they all have good template structures, however due to the alignment errors they are not modeled from the correct segment of the template.

To show the correctly aligned segments exist, all these targets are taken to the suboptimal alignment step. Namely, not only single best solution from dynamic programming alignment algorithm is considered, but all the additional suboptimal alignments are collected from Modeller’s suboptimal algorithm. Figure 5 from A to H displays the results for different targets and the red dots shows the correctly aligned segments in the single optimal alignment. This is what is collected from a regular profile-profile alignment. Again, the alignments are compared to the structural alignment. The different colors in Figure 5 other than red, show the correct alignment segments in the pool of sub-optimal alignments. The number of suboptimal alignments obtained depends on the Modeller’s dynamic programming alignment. For each target, the correctly aligned segments cover much more than the correctly aligned segments exist in the single optimal alignment. Thus, in the genetic algorithm moves such as single-point crossovers or double-point crossovers, if these segments can be evaluated by a more detailed and localized scoring function and brought together, the models will have a much higher number of correctly aligned positions.

4. Conclusion

In this work, one of the biggest challenges in homology modeling, modeling errors originated from alignment mistakes is evaluated with a benchmark set in the twilight zone, low sequence identity region of the template-based modeling. The profile-profile alignments produced low-quality models. When the structural features are incorporated into the alignment, the model qualities are improved. However, generally, when the target structure is not known which the standard is, incorporating structural features is not an easy task. The iterative modeling protocol, Moulder, tries to incorporate structural features by evaluating alignments with their 3-D models is used in this study. Its sampling method is shown to be successful, while scoring function used is not sufficient to pick the correct models or alignments segments. Finally, alignment of CASP 8 targets which has extremely low sequence identity but good structural overlap with the template structures is analyzed in terms of suboptimal alignments. For each target, it has been shown that the correct alignment segments exist in the pool of sub-optimal alignments. Thus, if improved scoring functions to pick these fragments are developed, the sampling with genetic algorithm moves such as crossovers will be capable of bringing the correctly aligned segments together.

Finally, the importance of the model assessment scoring functions which are good in the local quality assessment is emphasized with the results here. If the reliability of different regions of a predicted structure is measured correctly, those segments can be brought together to form a single correct overall alignment.

Author Contributions

The percentage of the author contributions is presented below. The author reviewed and approved the final version of the manuscript.

	S.E.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	100

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The author declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

Acknowledgements

The author is thankful to Prof. Dr. Andrej Sali for technical discussions about the method, Dr. David Eramian for providing previous version of the software, and Dr. Burak Servili for the support in the preparation of the figures.

References

Bertoline LMF, Lima AN, Krieger JE, Teixeira SK. 2023. Before and after AlphaFold2: An overview of protein structure prediction. Front Bioinform, 3: 1120370.

Bonneau R, Baker D. 2001. Ab initio protein structure prediction: Progress and prospects. Annu Rev Biophys Biomol Struct, 30: 173-189.

Chen H, Kihara D. 2011. Effect of using suboptimal alignments in template-based protein structure prediction. Proteins: Structure, Function and Bioinformatics, 79(1): 315-334.

Dunbrack RLJ. 2006. Sequence comparison and protein structure prediction. Curr Opin Struct Biol, 16(3): 374-384.

Eramian DD. 2008. Assessment and Prediction of Protein Structures. PhD thesis, University, University of California at San Franciso, San Francisco, pp: 252. URL: https://escholarship.org/uc/item/3k41q2cq (accessed date: June 12, 2023).

Gromiha MM, Nagarajan R, Selvaraj S. 2018. Protein structural bioinformatics: An overview. In Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 2: 445-459.

Guex N, Peitsch MC. 1997. Swiss PDB Viewer - References. Electrophoresis, 18(15): 2714-2723.

Hardin C, Pogorelov TV, Luthey-Schulten Z. 2002. Ab initio protein structure prediction. Curr Opin Struct Biol, 12(2): 176-181.

John B, Sali A. 2003. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res, 31(14): 3982-3992.

Jones DT. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol, 287(4): 797-815.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Hassabis D. 2021. Applying and improving AlphaFold at CASP14. Prot Struct Functi Bioinformat, 89(12): 1711-1721.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596: 583-589.

Kahsay RY, Wang G, Gao G, Liao L, Dunbrack R. 2005. Quasi-consensus-based comparison of profile hidden Markov models for protein sequences. Bioinformatics, 21(10): 2287-2293.

Kim DE, Chivian D, Baker D. 2004. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res, 32: W526-W531.

Marti-Renom MA, Madhusudhan MS, Sali A. 2004. Alignment of protein sequences by their profiles. Protein Sci, 13(4): 1071-1087.

Nassar R, Dignon GL, Razban RM, Dill KA. 2021. The Protein Folding Problem: The Role of Theory. J Mol Biol, 433(20): 167126.

Pearce R, Li Y, Omenn GS, Zhang Y. 2022. Fast and accurate Ab Initio Protein structure prediction using deep learning potentials. PLoS Comput Biol, 18(9): e1010539.

Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, Khuri N, Spill YG, Weinkam P, Hammel M, Tainer JA, Nilges M, Sali A. 2006. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res, 34: D291-5.

Rohl CA, Strauss CEM, Misura KMS, Baker D. 2004. Protein structure prediction using rosetta. Meth Enzymol, 383: 66-93.

Sauder JM, Arthur JW, Dunbrack RLJ. 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins, 40(1): 6-22.

Shen MY, Sali A. 2006. Statistical potential for assessment and prediction of protein structures. Protein Sci, 15(11): 2507-2524.

Soding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics, 21(7): 951-960.

Wang G, Dunbrack RLJ. 2004. Scoring profile-to-profile sequence alignments. Protein Sci, 13(6): 1612-1626.

Webb B, Sali A. 2016. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinformatics, 20(54): 5.6.1-5.6.37.

Xu D, Zhang Y. 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Prot Struct Funct Bioinformat, 80(7): 1715-1735.

Yang J, Zhang Y. 2015. I-TASSER server: New development for protein structure and function predictions. Nucleic Acids Res, 43(W1): W174-W181.

Zhou H, Zhou Y. 2005. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins, 58(2): 321-328.